BIT Numerical Mathematics (2021) 61:237–273
https://doi.org/10.1007/s10543-020-00826-z

BIT

Rational Krylov for Stieltjes matrix functions: convergence and pole selection

Stefano Massei[1] · **Leonardo Robol**[2]

Received: 10 June 2020 / Accepted: 22 July 2020 / Published online: 4 August 2020

Abstract
Evaluating the action of a matrix function on a vector, that is $x = f(\mathcal{M})v$, is an ubiquitous task in applications. When $\mathcal{M}$ is large, one usually relies on Krylov projection methods. In this paper, we provide effective choices for the poles of the rational Krylov method for approximating x when $f(z)$ is either Cauchy–Stieltjes or Laplace–Stieltjes (or, which is equivalent, completely monotonic) and $\mathcal{M}$ is a positive definite matrix. Relying on the same tools used to analyze the generic situation, we then focus on the case $\mathcal{M} = I \otimes A - B^T \otimes I$, and v obtained vectorizing a low-rank matrix; this finds application, for instance, in solving fractional diffusion equation on two-dimensional tensor grids. We see how to leverage tensorized Krylov subspaces to exploit the Kronecker structure and we introduce an error analysis for the numerical approximation of x. Pole selection strategies with explicit convergence bounds are given also in this case.

Keywords Rational Krylov · Function of matrices · Kronecker sum · Zolotarev problem · Pole selection · Stieltjes functions

Mathematics Subject Classification 65E05 · 65F60 · 30E20

Communicated by Lothar Reichel.

The work of Stefano Massei has been partially supported by the SNSF research project *Fast algorithms from low-rank updates*, Grant Number: 200020_178806, and by the INdAM/GNCS project "Analisi di matrici sparse e data-sparse: metodi numerici ed applicazioni". The work of Leonardo Robol has been partially supported by a GNCS/INdAM project "Giovani Ricercatori" 2018.

✉ Stefano Massei
s.massei@tue.nl

Leonardo Robol
leonardo.robol@unipi.it

[1] TU Eindhoven, Eindhoven, The Netherlands

[2] Dipartimento di Matematica, Università di Pisa, Pisa, Italy

1 Introduction

We are concerned with the evaluation of $x = f(\mathcal{M})v$, where $f(z)$ is a Stieltjes function, which can be expressed in integral form

$$f(z) = \int_0^\infty g(t,z)\mu(t)\,dt, \qquad g(t,z) \in \left\{ e^{-tz}, \frac{1}{t+z} \right\}. \tag{1}$$

The two choices for $g(t,z)$ define *Laplace–Stieltjes* and *Cauchy–Stieltjes* functions, respectively [8,31]. The former class is a superset of the latter, and coincides with the set of completely monotonic functions, whose derivatives satisfy $(-1)^j f^{(j)} \geq 0$ over $\mathbb{R}_+$ for any $j \in \mathbb{N}$.

We are interested in two instances of this problem; first, we consider the case $\mathcal{M} := A$, where $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite with spectrum contained in $[a,b]$, $v \in \mathbb{C}^{n\times s}$ is a generic (block) vector, and a rational Krylov method [18] is used to approximate $x = f(\mathcal{M})v$. In this case, we want to estimate the Euclidean norm of the error $\|x - x_\ell\|_2$, where x_ℓ is the approximation returned by ℓ steps of the method. Second, we consider

$$\mathcal{M} := I \otimes A - B^T \otimes I \in \mathbb{C}^{n^2\times n^2}, \tag{2}$$

where $A, -B \in \mathbb{C}^{n\times n}$ are Hermitian positive definite with spectra contained in $[a,b]$, $v = \mathrm{vec}(F) \in \mathbb{C}^{n^2}$ is the vectorization of a low-rank matrix $F = U_F V_F^T \in \mathbb{C}^{n\times n}$, and a tensorized rational Krylov method [8] is used for computing $\mathrm{vec}(X) = f(\mathcal{M})\mathrm{vec}(F)$. This problem is a generalization of the solution of a Sylvester equation with a low-rank right hand side, which corresponds to evaluate the function $f(z) = z^{-1}$. Here, we are concerned with estimating the quantity $\|X - X_\ell\|_2$, where X_ℓ is the approximation obtained after ℓ steps.

1.1 Main contributions

This paper discusses the connection between rational Krylov evaluation of Stieltjes matrix functions and the parameter dependent rational approximation (with the given poles) of the kernel functions e^{-tz} and $\frac{1}{t+z}$.

The contributions of this work are the following:

1. Corollary 3 provides a choice of poles for the rational Krylov approximation of $f(\mathcal{M})v$, where $f(z)$ is Laplace–Stieltjes, with an explicit error bound depending on the spectrum of A.
2. Similarly, for Cauchy–Stieltjes functions, we show (in Corollary 4) how leveraging an approach proposed in [14] allows to recover a result previously given in [4] using different theoretical tools.
3. In Sect. 3.5, we obtain new nested sequences of poles by applying the approach of equidistributed sequences to the results in Corollary 3–4.
4. In the particular case where $\mathcal{M} := I \otimes A - B^T \otimes I$ we extend the analysis recently proposed in [8] to rational Krylov subspaces. Also in this setting,

Table 1 Summary of the convergence rates for rational Krylov methods with the proposed poles

Function class	Argument	Error bound	Reference
Laplace–Stieltjes	$\mathcal{M} := A$	$\|x - x_\ell\|_2 \sim \mathcal{O}(\rho_{[a,b]}^{\frac{\ell}{2}})$	Cor. 3
	$\mathcal{M} := I \otimes A - B^T \otimes I$	$\|X - X_\ell\|_2 \sim \mathcal{O}(\rho_{[a,b]}^{\frac{\ell}{2}})$	Cor. 5
Cauchy–Stieltjes	$\mathcal{M} := A$	$\|x - x_\ell\|_2 \sim \mathcal{O}(\rho_{[a,4b]}^{\ell})$	Cor. 4
	$\mathcal{M} := I \otimes A - B^T \otimes I$	$\|X - X_\ell\|_2 \sim \mathcal{O}(\rho_{[a,2b]}^{\ell})$	Cor. 7

The convergence rate $\rho_{[\alpha,\beta]}$ is defined by $\rho_{[\alpha,\beta]} := \exp(-\pi^2/\log(4\frac{\beta}{\alpha}))$

we provide explicit choices for the poles and explicit convergence bounds. For Laplace–Stieltjes functions a direct consequence of the analysis mentioned above leads to Corollary 5; in the Cauchy case, we describe a choice of poles that enables the simultaneous solution of a set of parameter dependent Sylvester equations. This results in a practical choice of poles and an explicit error bound given in Corollary 7.

5. Finally, we give results predicting the decay in the singular values of X where $\mathrm{vec}(X) = f(\mathcal{M})\mathrm{vec}(F)$, F is a low-rank matrix, and $f(z)$ is either Laplace–Stieltjes (Theorem 6) or Cauchy–Stieltjes (Theorem 7). This generalizes the well-known low-rank approximability of the solutions of of Sylvester equations with low-rank right hand sides [5]. The result for Laplace–Stieltjes follows by the error bound for the rational Krylov method and an Eckart–Young argument. The one for Cauchy–Stieltjes requires to combine the integral representation with the ADI approximant for the solution of matrix equations.

The error bounds obtained are summarized in Table 1.

We recall that completely monotonic functions are well approximated by exponential sums [11]. Another consequence of our results in the Laplace–Stieltjes case is to constructively show that they are also well-approximated by rational functions.

1.2 Motivating problems

Computing the action of a matrix function on a vector is a classical task in numerical analysis, and finds applications in several fields, such as complex networks [7], signal processing [29], numerical solution of ODEs [20], and many others.

Matrices with the Kronecker sum structure as in (2) often arise from the discretization of differential operators on tensorized 2D grids. Applying the inverse of such matrices to a vector is equivalent to solving a matrix equation. When the right hand side is a smooth function or has small support, the vector v is the vectorization of a numerically low-rank matrix. The latter property has been exploited to develop several efficient solution methods, see [28] and the references therein. Variants of these approaches have been proposed under weaker assumptions, such as when the smoothness is only available far from the diagonal $x = y$, as it happens with kernel functions [23,25].

In recent years, there has been an increasing interest in models involving fractional derivatives. For 2D problems on rectangular grids, discretizations by finite differences or finite elements lead to linear systems that can be recast as the solution of matrix equations with particularly structured coefficients [12,24]. However, a promising formulation which simplifies the design of boundary conditions relies on first discretizing the 2D Laplacian on the chosen domain, and then considers the action of the matrix function $z^{-\alpha}$ (with the Laplacian as argument) on the right hand side. This is known in the literature as the *matrix transform method* [32]. In this framework, one has $0 < \alpha < 1$, and therefore $z^{-\alpha}$ is a Cauchy–Stieltjes function, a property that has been previously exploited for designing fast and stable restarted polynomial Krylov methods for its evaluation [27]. The algorithm proposed in this paper allows to exploit the Kronecker structure of the 2D Laplacian on rectangular domains in the evaluation of the matrix function.

Another motivation for our analysis stems from the study of exponential integrators, where it is required to evaluate the $\varphi_j(z)$ functions [20], which are part of the Laplace–Stieltjes class. This has been the subject of deep studies concerning (restarted) polynomial and rational Krylov methods [17,27]. However, to the best of our knowledge the Kronecker structure, and the associated low-rank preservation, has not been exploited in these approaches, despite being often present in discretization of differential operators [30].

The paper is organized as follows. In Sect. 2 we recall the definitions and main properties of Stieltjes functions. Then, in Sect. 3 we recall the rational Krylov method and then we analyze the simultaneous approximation of parameter dependent exponentials and resolvents; this leads to the choice of poles and convergence bounds for Stieltjes functions given in Sect. 3.4. In Sect. 4 we provide an analysis of the convergence of the method proposed in [8] when rational Krylov subspaces are employed. In particular, in Sect. 4.4 we provide decay bounds for the singular values of X such that $\mathrm{vec}(X) = f(\mathcal{M})\mathrm{vec}(F)$. We give some concluding remarks and outlook in Sect. 5.

2 Laplace–Stieltjes and Cauchy–Stieltjes functions

We recall the definition and the properties of Laplace–Stieltjes and Cauchy–Stieltjes functions that are relevant for our analysis. Functions expressed as Stieltjes integrals admit a representation of the form:

$$f(z) = \int_0^\infty g(t, z)\mu(t)\, dt, \tag{3}$$

where $\mu(t)dt$ is a (non-negative) measure on $[0, \infty]$, and $g(t, z)$ is integrable with respect to that measure. The choice of $g(t, z)$ determines the particular class of Stieltjes functions under consideration (*Laplace–Stieltjes* or *Cauchy–Stieltjes*), and $\mu(t)$ is called the density of $f(z)$. $\mu(t)$ can be a proper function, or a distribution, e.g. a Dirac delta. In particular, we can restrict the domain of integration to a subset of $(0, \infty)$ imposing that $\mu(t) = 0$ elsewhere. We refer the reader to [31] for further details.

2.1 Laplace–Stieltjes functions

Laplace–Stieltjes functions are obtained by setting $g(t, z) = e^{-tz}$ in (3).

Definition 1 Let $f(z)$ be a function defined on $(0, +\infty)$. Then, $f(z)$ is a *Laplace–Stieltjes* function if there is a positive measure $\mu(t)dt$ on $\mathbb{R}_+$ such that

$$f(z) = \int_0^\infty e^{-tz}\mu(t)\, dt. \tag{4}$$

Examples of Laplace–Stieltjes functions include:

$$f(z) = z^{-1} = \int_0^\infty e^{-tz} t\, dt, \qquad f(z) = e^{-z} = \int_1^\infty e^{-tz}\, dt,$$
$$f(z) = (1 - e^{-z})/z = \int_0^\infty e^{-tz}\mu(t)\, dt, \qquad \mu(t) := \begin{cases} 1 & t \in [0, 1] \\ 0 & t > 1 \end{cases}.$$

The last example is an instance of a particularly relevant class of Laplace–Stieltjes functions, with applications to exponential integrators. These are often denoted by $\varphi_j(z)$, and can be defined as follows:

$$\varphi_j(z) := \int_0^\infty e^{-tz} \frac{[\max\{1 - t, 0\}]^{j-1}}{(j-1)!}\, dt, \qquad j \geq 1.$$

A famous theorem of Bernstein states the equality between the set of Laplace–Stieltjes functions and those of *completely monotonic functions in* $(0, \infty)$ [10], that is $(-1)^j f^{(j)}(z)$ is positive over $(0, \infty)$ for any $j \in \mathbb{N}$.

From the algorithmic point of view, the explicit knowledge of the Laplace density $\mu(t)$ will not play any role. Therefore, for the applications of the algorithms and projection methods described here, it is only relevant to know that a function is in this class.

2.2 Cauchy–Stieltjes functions

Cauchy–Stieltjes functions form a subclass of Laplace–Stieltjes functions, and are obtained from (3) by setting $g(t, z) = (t + z)^{-1}$.

Definition 2 Let $f(z)$ be a function defined on $\mathbb{C}\backslash\mathbb{R}_-$. Then, $f(z)$ is a *Cauchy–Stieltjes* function if there is a positive measure $\mu(t)dt$ on $\mathbb{R}_+$ such that

$$f(z) = \int_0^\infty \frac{\mu(t)}{t + z}\, dt. \tag{5}$$

A few examples of Cauchy–Stieltjes functions are:

$$f(z) = \frac{\log(1+z)}{z} = \int_1^\infty \frac{t^{-1}}{t+z}\,dt, \qquad f(z) = \sum_{j=1}^{h} \frac{\alpha_j}{z-\beta_j}, \qquad \alpha_j > 0, \quad \beta_j < 0,$$

$$f(z) = z^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi} \int_0^\infty \frac{t^{-\alpha}}{t+z}\,dt, \qquad \alpha \in (0,1).$$

The rational functions with poles on the negative real semi-axis do not belong to this class if one requires $\mu(t)$ to be a function, but they can be obtained by setting $\mu(t) = \sum_{j=1}^{h} \alpha_j \delta(t-\beta_j)$, where $\delta(\cdot)$ is the Dirac delta with unit mass at 0. For instance, z^{-1} is obtained setting $\mu(t) := \delta(t)$.

Since Cauchy–Stieltjes functions are also completely monotonic in $(0,\infty)$ [9], the set of Cauchy–Stieltjes functions is contained in the one of Laplace–Stieltjes functions. Indeed, assuming that $f(z)$ is a Cauchy–Stieltjes function with density $\mu_C(t)$, one can construct a Laplace–Stieltjes representation as follows:

$$f(z) = \int_0^\infty \frac{\mu_C(t)}{t+z}\,dt = \int_0^\infty \int_0^\infty e^{-s(t+z)}\mu_C(t)\,ds\,dt = \int_0^\infty e^{-sz} \underbrace{\int_0^\infty e^{-st}\mu_C(t)\,dt}_{\mu_L(s)}\,ds,$$

where $\mu_L(s)$ defines the Laplace–Stieltjes density. In particular, note that if $\mu_C(t)$ is positive, so is $\mu_L(s)$. For a more detailed analysis of the relation between Cauchy- and Laplace–Stieltjes functions we refer the reader to [31, Section 8.4].

As in the Laplace case, the explicit knowledge of $\mu(t)$ is not crucial for the analysis and is not used in the algorithm.

3 Rational Krylov for evaluating Stieltjes functions

Projection schemes for the evaluation of the quantity $f(A)v$ work as follows: an orthonormal basis W for a (small) subspace $\mathcal{W} \subseteq \mathbb{C}^n$ is computed, together with the projections $A_{\mathcal{W}} := W^*AW$ and $v_{\mathcal{W}} := W^*v$. Then, the action of $f(A)$ on v is approximated by:

$$f(A)v \approx x_{\mathcal{W}} := Wf(A_{\mathcal{W}})v_{\mathcal{W}}.$$

Intuitively, the choice of the subspace $\mathcal{W}$ is crucial for the quality of the approximation. Usually, one is interested in providing a sequence of subspaces $\mathcal{W}_1 \subset \mathcal{W}_2 \subset \mathcal{W}_3 \subset \ldots$ and study the convergence of $x_{\mathcal{W}_j}$ to $f(A)v$ as j increases. A common choice for the space $\mathcal{W}_j$ are Krylov subspaces.

3.1 Krylov subspaces

Several functions can be accurately approximated by polynomials. The idea behind the standard Krylov method is to generate a subspace that contains all the quantities of the form $p(A)v$ for every $p(z)$ polynomial of bounded degree.

Definition 3 Let A be an $n \times n$ matrix, and $v \in \mathbb{C}^{n\times s}$ be a (block) vector. The *Krylov subspace of order* ℓ generated by A and v is defined as

$$\mathcal{K}_\ell(A, v) := \text{span}\{v, Av, \ldots, A^\ell v\} = \{p(A)v : \deg(p) \le \ell\}.$$

Projection on Krylov subspaces is closely related to polynomial approximation. Indeed, if $f(z)$ is well approximated by $p(z)$, then $p(A)v$ is a good approximation of $f(A)v$, in the sense that $\|f(A)v - p(A)v\|_2 \le \max_{z\in[a,b]} |f(z) - p(z)| \cdot \|v\|_2$.

Rational Krylov subspaces are their rational analogue, and can be defined as follows.

Definition 4 Let A be a $n \times n$ matrix, $v \in \mathbb{C}^{n\times s}$ be a (block) vector and $\Psi = (\psi_1, \ldots, \psi_\ell)$, with $\psi_j \in \overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$. The *rational Krylov subspace* with poles Ψ generated by A and v is defined as

$$\begin{aligned}\mathcal{RK}_\ell(A, v, \Psi) = q_\ell(A)^{-1}\mathcal{K}_\ell(A, v) = \text{span}\{q_\ell(A)^{-1}v, q_\ell(A)^{-1}Av,\\ \ldots, q_\ell(A)^{-1}A^\ell v\},\end{aligned}$$

where $q_\ell(z) = \prod_{j=1}^{\ell}(z - \psi_j)$ and if $\psi_j = \infty$, then we omit the factor $(z - \psi_j)$ from $q_\ell(z)$.

Note that, a Krylov subspace is a particular rational Krylov subspace where all poles are chosen equal to ∞: $\mathcal{RK}_\ell(A, v, (\infty, \ldots, \infty)) = \mathcal{K}_\ell(A, v)$. A common strategy of pole selection consists in alternating 0 and ∞. The resulting vector space is known in the literature as the *extended Krylov subspace* [13].

We denote by $\mathcal{P}_\ell$ the set of polynomials of degree at most ℓ, and by $\mathcal{R}_{\ell,\ell}$ the set of rational functions $g(z)/l(z)$ with $g(z), l(z) \in \mathcal{P}_\ell$. Given $\Psi = \{\psi_1, \ldots, \psi_\ell\} \subset \overline{\mathbb{C}}$, we indicate with $\frac{\mathcal{P}_\ell}{\Psi}$ the set of rational functions of the form $g(z)/l(z)$, with $g(z) \in \mathcal{P}_\ell$ and $l(z) := \prod_{\psi_j\in\Psi\setminus\{\infty\}}(z - \psi_j)$.

It is well-known that Krylov subspaces contain the action of related rational matrix functions of A on the (block) vector v, if the poles of the rational functions are a subset of the poles used to construct the approximation space.

Lemma 1 (Exactness property) *Let A be a $n \times n$ matrix, $v \in \mathbb{C}^{n\times s}$ be a (block) vector and $\Psi = \{\psi_1, \ldots, \psi_\ell\} \subset \overline{\mathbb{C}}$. If $U_\mathcal{P}, U_\mathcal{R}$ are orthonormal bases of $\mathcal{K}_\ell(A, v)$ and $\mathcal{RK}_\ell(A, v, \Psi)$, respectively, then:*

1. $f(z) \in \mathcal{P}_\ell \implies f(A)v = U_\mathcal{P} f(A_\ell)(U_\mathcal{P}^* v) \in \mathcal{K}_\ell(A, v), \quad A_\ell = U_\mathcal{P}^* A U_\mathcal{P},$
2. $f(z) \in \frac{\mathcal{P}_\ell}{\Psi} \implies f(A)v = U_\mathcal{R} f(A_\ell)(U_\mathcal{R}^* v) \in \mathcal{RK}_\ell(A, v, \Psi), \quad A_\ell = U_\mathcal{R}^* A U_\mathcal{R},$

Lemma 1 enables to prove the quasi-optimality of the Galerkin projection described in Sect. 3. Indeed, if $\mathcal{W} := \mathcal{RK}(A, v, \Psi)$, then [18]

$$\|x_{\mathcal{W}} - x\|_2 \leq 2 \cdot \|v\|_2 \cdot \min_{r(z) \in \frac{\mathcal{P}_\ell}{\Psi}} \max_{z \in [a,b]} |f(z) - r(z)|. \tag{6}$$

The optimal choice of poles for generating the rational Krylov subspaces is problem dependent and linked to the rational approximation of the function $f(z)$ on $[a, b]$. We investigate how to perform this task when f is either a Laplace–Stieltjes or Cauchy–Stieltjes function.

3.2 Simultaneous approximation of resolvents and matrix exponentials

The integral expression (1) reads as

$$f(A)v = \int_0^\infty g(t, A)\mu(t)\, dt, \qquad g(t, A) \in \{e^{-tA}, (tI + A)^{-1}\}$$

when evaluated at a matrix argument. Since the projection is a linear operation we have

$$x_{\mathcal{W}} = W f(A_{\mathcal{W}}) v_{\mathcal{W}} = \int_0^\infty W g(t, A_{\mathcal{W}}) v_{\mathcal{W}}\, \mu(t) dt.$$

This suggests to look for a space approximating uniformly well, in the parameter t, matrix exponentials and resolvents, respectively. A result concerning the approximation error in the L^2 norm for $t \in \mathbf{i}\mathbb{R}$ is given in [14, Lemma 4.1]. The proof is obtained exploiting some results on the skeleton approximation of $\frac{1}{t+\lambda}$ [26]. We provide a pointwise error bound, which can be obtained by following the same steps of the proof of [14, Lemma 4.1]. We include the proof for completeness.

Theorem 1 *Let A be Hermitian positive definite with spectrum contained in $[a, b]$ and U be an orthonormal basis of $\mathcal{U}_{\mathcal{R}} = \mathcal{RK}_\ell(A, v, \Psi)$. Then, $\forall t \in [0, \infty)$, we have the following inequality:*

$$\|(tI + A)^{-1}v - U(tI + A_\ell)^{-1}v_\ell\|_2 \leq \frac{2}{t + a}\|v\|_2 \min_{r(z) \in \frac{\mathcal{P}_\ell}{\Psi}} \frac{\max_{z \in [a,b]} |r(z)|}{\min_{z \in (-\infty, 0]} |r(z)|} \tag{7}$$

where $A_\ell = U^* A U$ and $v_\ell = U^* v$.

Proof Following the construction in [26], we consider the function $f_{\text{skel}}(\lambda, t)$ defined by

$$f_{\text{skel}}(t, \lambda) := \begin{bmatrix} \frac{1}{t_1+\lambda} & \cdots & \frac{1}{t_\ell+\lambda} \end{bmatrix} M^{-1} \begin{bmatrix} \frac{1}{t+\lambda_1} \\ \vdots \\ \frac{1}{t+\lambda_\ell} \end{bmatrix}, \qquad M_{ij} = \frac{1}{t_j + \lambda_i} \in \mathbb{C}^{\ell \times \ell},$$

where M_{ij} are the entries of M and (t_j, λ_i) is an $\ell \times \ell$ grid of interpolation nodes. The function $f_{\text{skel}}(t, \lambda)$ is usually called Skeleton approximation and it is practical

for approximating $\frac{1}{t+\lambda}$; indeed its relative error takes the explicit form: $1-(t+\lambda)f_{\text{skel}}(t,\lambda)=\frac{r(\lambda)}{r(-t)}$ with $r(z)=\prod_{j=1}^{\ell}\frac{z-\lambda_j}{t_j+z}$. If this ratio of rational functions is small, then $f_{\text{skel}}(t,\lambda)$ is a good approximation of $\frac{1}{t+\lambda}$ and—consequently—$f_{\text{skel}}(t,A)$ is a good approximation of $(tI+A)^{-1}$. Note that, for every fixed t, $f_{\text{skel}}(t,\lambda)$ is a rational function in λ with poles $-t_1,\dots,-t_\ell$. Therefore, using the poles $\psi_j=-t_j$, $j=1,\dots,\ell$ for the projection we may write, thanks to (6):

$$\|(tI+A)^{-1}v-U(tI+A)^{-1}v_\ell\|_2\le\frac{2}{t+a}\|v\|_2\frac{\max_{z\in[a,b]}|r(z)|}{\min_{z\in(-\infty,0]}|r(z)|}.$$

Taking the minimum over the possible choices of the parameters λ_j we get (7). □

Concerning the rational approximation of the (parameter dependent) exponential, the idea is to rely on its Laplace transform that involves the resolvent:

$$e^{-tA}=\frac{1}{2\pi\mathbf{i}}\lim_{T\to\infty}\int_{-\mathbf{i}T}^{\mathbf{i}T}e^{st}(sI+A)^{-1}\,ds. \tag{8}$$

In this formulation, it is possible to exploit the Skeleton approximation of $\frac{1}{s+\lambda}$ in order to find a good choice of poles, independently on the parameter t. For proving the main result we need the following technical lemma whose proof is given in the "Appendix 2".

Lemma 2 *Let $\mathcal{L}^{-1}[\widehat{r}(s)]$ be the inverse Laplace transform of $\widehat{r}(s)=\frac{1}{s}\frac{p(s)}{p(-s)}$, where $p(s)$ is a polynomial of degree ℓ with positive real zeros contained in $[a,b]$. Then,*

$$\|\mathcal{L}^{-1}[\widehat{r}(s)]\|_{L^\infty(\mathbb{R}_+)}\le\gamma_{\ell,\kappa},\qquad\gamma_{\ell,\kappa}:=2.23+\frac{2}{\pi}\log\left(4\ell\cdot\sqrt{\frac{\kappa}{\pi}}\right),$$

where $\kappa=\frac{b}{a}$.

Theorem 2 *Let A be Hermitian positive definite with spectrum contained in $[a,b]$ and U be an orthonormal basis of $\mathcal{U}_{\mathcal{R}}=\mathcal{RK}_\ell(A,v,\Psi)$, where $\Psi=\{\psi_1,\dots,\psi_\ell\}\subseteq[-b,-a]$. Then, $\forall t\in[0,\infty)$, we have the following inequality:*

$$\|e^{-tA}v-Ue^{-tA_\ell}v_\ell\|_2\le4\gamma_{\ell,\kappa}\|v\|_2\max_{z\in[a,b]}|r_\Psi(z)|, \tag{9}$$

where $A_\ell=U^*AU$, $v_\ell=U^*v$, $\kappa:=\frac{b}{a}$, $r_\Psi(z)\in\mathcal{R}_{\ell,\ell}$ is the rational function defined by $r_\Psi(z):=\prod_{j=1}^{\ell}\frac{z+\psi_j}{z-\psi_j}$ and $\gamma_{\ell,\kappa}$ is the constant defined in Lemma 2.

Proof We consider the Skeleton approximation of $\frac{1}{s+\lambda}$ by restricting the choice of poles in both variables to Ψ

$$f_{\text{skel}}(s,\lambda):=\left[\frac{1}{\lambda-\psi_1}\ \cdots\ \frac{1}{\lambda-\psi_\ell}\right]M^{-1}\begin{bmatrix}\frac{1}{s-\psi_1}\\ \vdots\\ \frac{1}{s-\psi_\ell}\end{bmatrix},\qquad M_{ij}=-\frac{1}{\psi_i+\psi_j},$$

where M_{ij} denote the entries of M. Then, by using (8) for A and A_ℓ we get

$$e^{-tA}v - Ue^{-tA_\ell}v_\ell = \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} \int_{-\mathbf{i}T}^{\mathbf{i}T} e^{st}(sI + A)^{-1}v - e^{st}U(sI + A_\ell)^{-1}v_\ell \, ds.$$

Adding and removing the term $e^{st} f_{\mathrm{skel}}(s, A)v = e^{st}U f_{\mathrm{skel}}(s, A_\ell)U^*v$ inside the integral (the equality holds thanks to Lemma 1) we obtain the error expression

$$\begin{aligned}
e^{-tA}v - Ue^{-tA_\ell}v_\ell &= \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} \int_{-\mathbf{i}T}^{\mathbf{i}T} e^{st}\left[(sI + A)^{-1}v - U(sI + A_\ell)^{-1}v_\ell\right] ds \\
&= \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} \int_{-\mathbf{i}T}^{\mathbf{i}T} e^{st}(sI + A)^{-1}\left[I - (sI + A) f_{\mathrm{skel}}(s, A)\right] v \, ds \\
&\quad - \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} U \int_{-\mathbf{i}T}^{\mathbf{i}T} e^{st}(sI + A_\ell)^{-1}\left[I - (sI + A_\ell) f_{\mathrm{skel}}(s, A_\ell)\right] v_\ell \, ds \\
&= \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} \int_{-\mathbf{i}T}^{\mathbf{i}T} e^{st}(sI + A)^{-1} r_\Psi(A) r_\Psi(-s)^{-1} v \, ds \\
&\quad - \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} U \int_{-\mathbf{i}T}^{\mathbf{i}T} e^{st}(sI + A_\ell)^{-1} r_\Psi(A_\ell) r_\Psi(-s)^{-1} v_\ell \, ds.
\end{aligned}$$

Since A and A_ℓ are normal, the above integrals can be controlled by the maximum of the corresponding scalar functions on the spectrum of A (and A_ℓ), which yields the bound

$$\begin{aligned}
\|e^{-tA}v - Ue^{-tA_\ell}v_\ell\|_2 &\le 2 \max_{\lambda\in[a,b]} |h(t, \lambda)|, \\
h(t, \lambda) &:= \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} \int_{-iT}^{iT} e^{st} \frac{1}{s + \lambda} \frac{r_\Psi(\lambda)}{r_\Psi(-s)} \, ds.
\end{aligned}$$

We note that $r_\Psi(\lambda)$ can be pulled out of the integral, since it does not depend on s, and thus the above can be rewritten as

$$\begin{aligned}
h(t, \lambda) &= r_\Psi(\lambda) \cdot \mathcal{L}^{-1}\left[\frac{1}{\lambda + s} \frac{p(s)}{p(-s)}\right](t) \\
&= r_\Psi(\lambda) \cdot \mathcal{L}^{-1}\left[\frac{s}{s + \lambda}\right] \star \mathcal{L}^{-1}\left[\frac{1}{s} \frac{p(s)}{p(-s)}\right](t) \\
&= r_\Psi(\lambda) \cdot (\delta(t) - \lambda e^{-\lambda t}) \star \mathcal{L}^{-1}\left[\frac{1}{s} \frac{p(s)}{p(-s)}\right](t),
\end{aligned}$$

where $p(s)$ is as in Lemma 2 and $\delta(t)$ indicates the Dirac delta function. Since the 1-norm of $\delta(t) - \lambda e^{-t\lambda}$ is equal to 2, using Young's inequality we can bound $\|h(t, \lambda)\|_\infty \le 2\|\mathcal{L}^{-1}\left[\frac{1}{s} \frac{p(s)}{p(-s)}\right]\|_\infty$. Therefore, we need to estimate the infinity norm of $\mathcal{L}^{-1}\left[\frac{1}{s} \frac{p(s)}{p(-s)}\right](t)$. Such inverse Laplace transform can be uniformly bounded in t by using Lemma 2 with a constant that only depends on ℓ and b/a:

$$|h(\lambda, t)| \leq 2\gamma_{\ell,\kappa}|r_\Psi(\lambda)|.$$

This completes the proof. □

Remark 1 The constant provided by Lemma 2 is likely not optimal. Indeed, experimentally it seems to hold that $\gamma_{\ell,\kappa} = 1$ for any choice of poles in the negative real axis—not necessarily contained in $[-b, -a]$—and this has been verified in many examples. If this is proved, then the statement of Theorem 1 can be made sharper by removing the factor $\gamma_{\ell,\kappa}$.

3.3 Bounds for the rational approximation problems

Theorems 1 and 2 show the connection between the error norm and certain rational approximation problems. In this section we discuss the optimal values of such problems in the cases of interests.

Definition 5 Let $\Psi \subset \overline{\mathbb{C}}$ be a finite set, and I_1, I_2 closed subsets of $\overline{\mathbb{C}}$. Then, we define[1]

$$\theta_\ell(I_1, I_2, \Psi) := \min_{r(z) \in \frac{\mathcal{P}_\ell}{\Psi}} \frac{\max_{I_1} |r(z)|}{\min_{I_2} |r(z)|}.$$

The θ_ℓ functions enjoy some invariance and inclusion properties, which we report here, and will be extensively used in the rest of the paper.

Lemma 3 *Let I_1, I_2 be subsets of the complex plane, and $\Psi \subset \overline{\mathbb{C}}$. Then, the map θ_ℓ satisfies the following properties:*

(i) (shift invariance) For any $t \in \mathbb{C}$, it holds $\theta_\ell(I_1 + t, I_2 + t, \Psi + t) = \theta(I_1, I_2, \Psi)$.

(ii) (monotonicity) $\theta_\ell(I_1, I_2, \Psi)$ is monotonic with respect to the inclusion on the parameters I_1 and I_2:

$$I_1 \subseteq I_1', I_2 \subseteq I_2' \implies \theta_\ell(I_1, I_2, \Psi) \leq \theta_\ell(I_1', I_2', \Psi).$$

(iii) (Möbius invariance) If $M(z)$ is a Möbius transform, that is a rational function $M(z) = (\alpha z + \beta)/(\gamma z + \delta)$ with $\alpha\delta \neq \beta\gamma$, then

$$\theta_\ell(I_1, I_2, \Psi) = \theta_\ell(M(I_1), M(I_2), M(\Psi)).$$

Proof Property (i) follows by (iii) because applying a shift is a particular Möbius transformation. Note that, generically, when we compose a rational function $r(z) = \frac{p(z)}{h(z)} \in \frac{\mathcal{P}_\ell}{\Psi}$ with $M^{-1}(z)$ we obtain another rational function of (at most) the same degree and with poles $M(\Psi)$. Hence, we obtain

[1] We allow the slight abuse of notation of writing $|r(\infty)|$ as the limit of $|r(z)|$ as $|z| \to \infty$, in the case either I_1 or I_2 contains the point at infinity.

$$\theta_\ell(I_1, I_2, \Psi) = \min_{r(z)\in\frac{\mathcal{P}_\ell}{\Psi}} \frac{\max_{I_1}|r(z)|}{\min_{I_2}|r(z)|} = \min_{r(z)\in\frac{\mathcal{P}_\ell}{\Psi}} \frac{\max_{M(I_1)}|r(M^{-1}(z))|}{\min_{M(I_2)}|r(M^{-1}(z))|}$$
$$= \min_{r(z)\in\frac{\mathcal{P}_\ell}{M(\Psi)}} \frac{\max_{M(I_1)}|r(z)|}{\min_{M(I_2)}|r(z)|} = \theta_\ell(M(I_1), M(I_2), M(\Psi)).$$

Property (ii) follows easily from the fact that the maximum taken on a larger set is larger, and the minimum taken on a larger set is smaller. □

Now, we consider the related optimization problem, obtained by allowing Ψ to vary:

$$\min_{\Psi\subset\overline{\mathbb{C}},|\Psi|=\ell} \theta_\ell(I_1, I_2, \Psi) = \min_{r(z)\in\mathcal{R}_{\ell,\ell}} \frac{\max_{z\in I_1}|r(z)|}{\min_{z\in I_2}|r(z)|}. \tag{10}$$

The latter was posed and studied by Zolotarev in 1877 [33], and it is commonly known as the *third Zolotarev problem*. We refer to [3] for a modern reference where the theory is used to recover bounds on the convergence of rational Krylov methods and ADI iterations for solving Sylvester equations.

In the case $I_1 = -I_2 = [a, b]$ (10) simplifies to

$$\min_{r(z)\in\mathcal{R}_{\ell,\ell}} \frac{\max_{z\in[a,b]}|r(z)|}{\min_{z\in[a,b]}|r(-z)|}$$

which admits the following explicit estimate.

Theorem 3 (Zolotarev) *Let $I = [a, b]$, with $0 < a < b$. Then*

$$\min_{\Psi\subset\overline{\mathbb{C}},\ |\Psi|=\ell} \theta_\ell(I, -I, \Psi) \le 4\rho_{[a,b]}^{\ell}, \qquad \rho_{[a,b]} := \exp\left(-\frac{\pi^2}{\log(4\kappa)}\right), \qquad \kappa = \frac{b}{a}.$$

In addition, the optimal rational function $r_\ell^{[a,b]}(z)$ that realizes the minimum has the form

$$r_\ell^{[a,b]}(z) := \frac{p_\ell^{[a,b]}(z)}{p_\ell^{[a,b]}(-z)}, \qquad p_\ell^{[a,b]}(z) := \prod_{j=1}^{\ell}(z + \psi_{j,\ell}^{[a,b]}), \qquad \psi_{j,\ell}^{[a,b]} \in -I.$$

We denote by $\Psi_\ell^{[a,b]} := \{\psi_{1,\ell}^{[a,b]}, \dots, \psi_{\ell,\ell}^{[a,b]}\}$ the set of poles of $r_\ell^{[a,b]}(z)$.

Explicit expression for the elements of $\Psi_\ell^{[a,b]}$ are available in terms of elliptic functions, see [14, Theorem 4.2].

Remark 2 The original version of Zolotarev's result involves $\exp(-\frac{\pi^2}{\mu(\kappa^{-1})})$ in place of $\rho_{[a,b]}$, where $\mu(\cdot)$ is the Grötzsch ring function. For simplicity, in this paper we prefer the slightly suboptimal form involving the logarithm. We remark that for large κ (which is usually the case when considering rational Krylov methods) the difference is negligible [1, Section 17.3].

We use Theorem 3 and the Möbius invariance property as building blocks for bounding (7). The idea is to map the set $[-\infty, 0] \cup [a, b]$ into $[-1, -\widehat{a}] \cup [\widehat{a}, 1]$—for some $\widehat{a} \in (0, 1)$—with a Möbius transformation; then make use of Theorem 3 and Lemma 3(iii) to provide a convenient choice of Ψ for the original problem.

Lemma 4 *The Möbius transformation*

$$T_C(z) := \frac{\Delta + z - b}{\Delta - z + b}, \qquad \Delta := \sqrt{b^2 - ab},$$

maps $[-\infty, 0] \cup [a, b]$ *into* $[-1, -\widehat{a}] \cup [\widehat{a}, 1]$, *with* $\widehat{a} := \frac{\Delta+a-b}{\Delta-a+b} = \frac{b-\Delta}{\Delta+b}$. *The inverse map* $T_C(z)^{-1}$ *is given by:*

$$T_C^{-1}(z) := \frac{(b + \Delta)z + b - \Delta}{1 + z}.$$

Moreover, for any $0 < a < b$ *it holds* $\widehat{a}^{-1} \leq \frac{4b}{a}$, *and therefore* $\rho_{[\widehat{a},1]} \leq \rho_{[a,4b]}$.

Proof By direct substitution, we have $T_C(-\infty) = -1$, and $T_C(b) = 1$; moreover, again by direct computation one verifies that $T_C(0) + T_C(a) = 0$, which implies that $T_C([-\infty, 0]) = [-1, -\widehat{a}]$ and $T_C([a, b]) = [\widehat{a}, 1]$. Then, we have

$$\widehat{a}^{-1} = \frac{\Delta + b}{b - \Delta}, \qquad \Delta = b\sqrt{1 - a/b}.$$

Using the relation $\sqrt{1 - t} \leq 1 - \frac{t}{2}$ for any $0 \leq t \leq 1$, we obtain that $\widehat{a}^{-1} \leq \frac{2b - \frac{a}{2}}{\frac{a}{2}} \leq 4\frac{b}{a}$, which concludes the proof. □

Remark 3 We note that the estimate $\rho_{[\widehat{a},1]} \leq \rho_{[a,4b]}$ is asymptotically tight, that is the limit of $\rho_{[\widehat{a},1]}/\rho_{[a,4b]} \to 1$ as $b/a \to \infty$. For instance, if $b/a = 10$ then the relative error between the two quantities is about $2 \cdot 10^{-2}$, and for $b/a = 1000$ about $5 \cdot 10^{-5}$. Since the interest for this approach is in dealing with matrices that are not well-conditioned, we consider the error negligible in practice.

In light of Theorem 3 and Lemma 4, we consider the choice

$$\Psi_{C,\ell}^{[a,b]} := T_C^{-1}(\Psi_\ell^{[\widehat{a},1]}) \tag{11}$$

in Theorem 1. This yields the following estimate.

Corollary 1 *Let* A *be Hermitian positive definite with spectrum contained in* $[a, b]$ *and* U *be an orthonormal basis of* $\mathcal{U}_\mathcal{R} = \mathcal{RK}_\ell(A, v, \Psi_{C,\ell}^{[a,b]})$. *Then,* $\forall t \in [0, \infty)$

$$\|(tI + A)^{-1}v - U(tI + A_\ell)^{-1}v_\ell\|_2 \leq \frac{8}{t + a}\|v\|_2 \rho_{[a,4b]}^\ell, \tag{12}$$

where $A_\ell = U^*AU$ *and* $v_\ell = U^*v$.

When considering Laplace–Stieltjes functions, we may choose as poles $\Psi_\ell^{[a,b]}$ which are the optimal Zolotarev poles on the interval $[a, b]$. This enables to prove the following result, which builds on Theorem 2.

Corollary 2 *Let A be Hermitian positive definite with spectrum contained in $[a, b]$ and U be an orthonormal basis of $\mathcal{U}_\mathcal{R} = \mathcal{RK}_\ell(A, v, \Psi_\ell^{[a,b]})$. Then, $\forall t \in [0, \infty)$*

$$\|e^{-tA}v - Ue^{-tA_\ell}v_\ell\|_2 \leq 8\gamma_{\ell,\kappa}\|v\|_2\rho_{[a,b]}^{\frac{\ell}{2}}, \tag{13}$$

where $A_\ell = U^*AU$ and $v_\ell = U^*v$.

Proof The proof relies on the fact that the optimal Zolotarev function evaluated on the interval $[a, b]$ can be bounded by $2\rho_{[a,b]}^{\frac{\ell}{2}}$ [5, Theorem 3.3]. Since its zeros and poles are symmetric with respect to the imaginary axis and real, we can apply Theorem 2 to obtain (13). □

3.4 Convergence bounds for Stieltjes functions

Let us consider $f(z)$ a Stieltjes function of the general form (1). Then the error of the rational Krylov method for approximating $f(A)v$ is given by

$$\begin{aligned}\|f(A)v - Uf(A_\ell)v_\ell\|_2 &= \left\| \int_0^\infty [g(t, A)v - Ug(t, A_\ell)v_\ell]\,\mu(t)\,dt \right\|_2 \\ &\leq \int_0^\infty \|g(t, A)v - Ug(t, A_\ell)v_\ell\|_2\,\mu(t)\,dt\end{aligned}$$

where $g(t, A)$ is either a parameter dependent exponential or resolvent function. Therefore Corollary 1 and Corollary 2 provide all the ingredients to study the error of the rational Krylov projection, when the suggested pole selection strategy is adopted.

Corollary 3 *Let $f(z)$ be a Laplace–Stieltjes function, A be Hermitian positive definite with spectrum contained in $[a, b]$, U be an orthonormal basis of $\mathcal{U}_\mathcal{R} = \mathcal{RK}_\ell(A, v, \Psi_\ell^{[a,b]})$ and $x_\ell = Uf(A_\ell)v_\ell$ with $A_\ell = U^*AU$ and $v_\ell = U^*v$. Then*

$$\|f(A)v - x_\ell\|_2 \leq 8\gamma_{\ell,\kappa} f(0^+)\|v\|_2\rho_{[a,b]}^{\frac{\ell}{2}}, \tag{14}$$

where $\gamma_{\ell,\kappa}$ is defined as in Theorem 2, and $f(0^+) := \lim_{z\to 0^+} f(z)$.

Proof Since $f(z)$ is a Laplace–Stieltjes function, we can express the error as follows:

$$\begin{aligned}\|f(A)v - x_\ell\|_2 &\leq \int_0^\infty \left\|e^{-tA}v - Ue^{-tA_\ell}U^*v\right\|_2 \mu(t)\,dt \\ &\leq 8\gamma_{\ell,\kappa}\int_0^\infty \mu(t)\,dt\|v\|_2\rho_{[a,b]}^{\frac{\ell}{2}} \\ &= 8\gamma_{\ell,\kappa} f(0^+)\|v\|_2\rho_{[a,b]}^{\frac{\ell}{2}},\end{aligned}$$

where we applied (6) and Corollary 2 to obtain the second inequality. □

Remark 4 In order to be meaningful, Corollary 3 requires the function $f(z)$ to be finite over $[0, \infty)$, which might not be the case in general (consider for instance $x^{-\alpha}$, which is both Cauchy and Laplace–Stieltjes). Nevertheless, the result can be applied to $f(z+\eta)$, which is always completely monotonic for a positive η, by taking $0 < \eta < a$. A value of η closer to a gives a slower decay rate, but a smaller constant $f(0^+)$. Similarly, if $f(z)$ happens to be completely monotonic on an interval larger than $[0, \infty)$, then bounds with a faster asymptotic convergence rate (but a larger constant) can be obtained considering $\eta < 0$.

Corollary 1 allows to state the corresponding bound for Cauchy–Stieltjes functions. The proof is analogous to the one of Corollary 3.

Corollary 4 *Let $f(z)$ be a Cauchy–Stieltjes function, A be Hermitian positive definite with spectrum contained in $[a, b]$, U be an orthonormal basis of $\mathcal{U}_{\mathcal{R}} = \mathcal{RK}_\ell(A, v, \Psi_{C,\ell}^{[a,b]})$ with $\Psi_{C,\ell}^{[a,b]}$ as in (11) and $x_\ell = Uf(A_\ell)v_\ell$ with $A_\ell = U^*AU$ and $v_\ell = U^*v$. Then*

$$\|f(A)v - x_G\|_2 \leq 8f(a)\|v\|_2 \rho_{[a,4b]}^{\ell}. \tag{15}$$

3.5 Nested sequences of poles

From the computational perspective, it is more convenient to have a nested sequence of subspaces $\mathcal{W}_1 \subseteq \ldots \mathcal{W}_j \subseteq \mathcal{W}_{j+1} \subseteq \ldots$, so that ℓ can be chosen adaptively. For example, in [19] the authors propose a greedy algorithm for the selection of the poles taylored to the evaluation of Cauchy–Stieltjes matrix functions. See [15,16] for greedy pole selection strategies to be applied in different—although closely related—contexts.

The choices of poles proposed in the previous sections require to a priori determine the degree ℓ of the approximant x_ℓ. Given a target accuracy, one can use the convergence bounds in Corollary 3–4 to determine ℓ. Unfortunately, this is likely to overestimate the minimum value of ℓ that provides the sought accuracy.

An option, that allows to overcome this limitation, is to rely on the method of *equidistributed sequences* (EDS), as described in [14, Section 4]. The latter exploits the limit—as $\ell \to \infty$—of the measures generated by the set of points $\Psi_\ell^{[a,b]}$, $\Psi_{C,\ell}^{[a,b]}$ to return infinite sequences of poles that are guaranteed to provide the same asymptotic rate of convergence. More specifically, the EDS $\{\widetilde{\sigma}_j\}_{j\in\mathbb{N}}$ associated with $\Psi_\ell^{[a,1]}$ is obtained with the following steps:

(i) Select $\zeta \in \mathbb{R}^+ \setminus \mathbb{Q}$ and generate the sequence $\{s_j\}_{j\in\mathbb{N}} := \{0,\ \zeta - \lfloor\zeta\rfloor,\ 2\zeta - \lfloor 2\zeta\rfloor,\ 3\zeta - \lfloor 3\zeta\rfloor,\ \ldots\}$, where $\lfloor\cdot\rfloor$ indicates the greatest integer less than or equal to the argument; this sequence has as asymptotic distribution (in the sense of EDS) the Lebesgue measure on $[0, 1]$.

(ii) Compute the sequence $\{t_j\}_{j\in\mathbb{N}}$ such that $g(t_j) = s_j$ where

$$g(t) := \frac{1}{2M}\int_{a^2}^{t} \frac{dy}{\sqrt{(y-a^2)y(1-y)}}, \qquad M := \int_0^1 \frac{dy}{\sqrt{(1-y^2)(1-(1-a^2)y^2)}},$$

(iii) Define $\widetilde{\sigma}_j := \sqrt{t_j}$.

More generally, the EDS associated with $\Psi_\ell^{[a,b]}$, $\Psi_{C,\ell}^{[a,b]}$ are obtained by applying either a scaling or the Möbius transformation (11) to the EDS for $\Psi_\ell^{[a,1]}$.

In our implementation, only the finite portion $\{\widetilde{\sigma}_j\}_{j=0,\dots,\ell-1}$ is—incrementally—generated for computing x_ℓ. As starting irrational number we select $\zeta = \frac{1}{\sqrt{2}}$ and each quantity t_j is approximated by applying the Newton's method to the equation $g(t_j) - s_j = 0$. The initialization of the Newton iteration is done by approximating $\hat{t} \mapsto g(e^{\hat{t}}) - s_j$ with a linear function on the domain of interest, and then using the exponential of its only root as starting point. This is done beforehand selecting $t = a^2$ and $t = a$ as interpolation points; in our experience, with such starting point Newton's method converges in a few steps.

3.6 Some numerical tests

3.6.1 Laplace–Stieltjes functions

Let us consider the 1D diffusion problem over $[0, 1]$ with zero Dirichlet boundary conditions

$$\frac{\partial u}{\partial t} = \epsilon \frac{\partial^2 u}{\partial x^2} + f(x), \qquad u(x, 0) \equiv 0, \qquad \epsilon = 10^{-2},$$

discretized using central finite differences in space with 50,000 points, and integrated by means of the exponential Euler method with time step $\Delta t = 0.1$. This requires to evaluate the action of the Laplace–Stieltjes matrix function $\varphi_1(\frac{\epsilon}{h^2}\Delta t A)v$, where A is the tridiagonal matrix $A = \text{tridiag}(-1, 2, -1)$. We test the convergence rates of various choices of poles by measuring the absolute error when using a random vector v. Figure 1 (left) reports the results associated with: the poles from Corollary 2, the corresponding EDS computed as described in Sect. 3.5 and the extended Krylov method. It is visible that the three approximations have the same convergence rate, although the choice of poles from Corollary 2 and the EDS performs slightly better. We mention that, since A is ill-conditioned, polynomial Krylov performs poorly on this example.

We keep the same settings and we test the convergence rates for the Laplace–Stieltjes function $z^{-\frac{3}{2}}W(z)$ where $W(z)$ is the Lambert W function [22]. The plot in Fig. 1(right) shows that after about 10 iterations the convergence rate of the extended Krylov method deteriorates, while the poles from Corollary 2 and the EDS provide the best convergence rate.

3.6.2 Inverse square root

Let us test the pole selection strategies for Cauchy–Stieltjes functions, by considering the evaluation of $f(z) = z^{-\frac{1}{2}}$ on the $n \times n$ matrix $\text{tridiag}(-1, 2, -1)$, for $n = 10^4, 5 \cdot 10^4, 10^5$. The list of methods that we consider includes: the poles $\Psi_{C,\ell}^{[a,b]}$ from Corollary 1, the associated EDS, the extended Krylov method and the adaptive strategy proposed in [19] for Cauchy–Stieltjes functions. The latter is implemented in the `markovfunmv` package available at http://guettel.com/markovfunmv/ which we

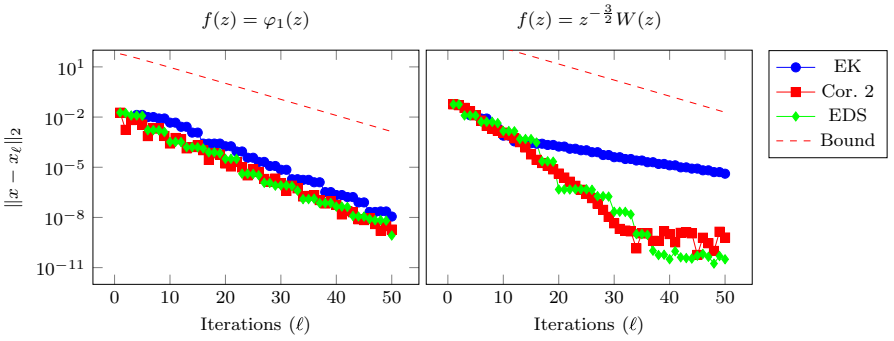


Fig. 1 Convergence history of the different projection spaces for the evaluation of $\varphi_1(A)v$ and $A^{-\frac{3}{2}}W(A)v$ for a matrix argument of size $50{,}000 \times 50{,}000$. The methods tested are extended Krylov (EK), rational Krylov with the poles from Corollary 2 and rational Krylov with nested poles obtained as in Sect. 3.5. The bound in the left figure is obtained directly from Corollary 2. The bound in the right figure has been obtained as in Remark 4

used for producing the results reported in Fig. 2. The poles from Corollary 1 and the extended Krylov method yield the best and the worst convergence history, respectively, for all values of n. The EDS and `markovfunm` perform similarly for $n = 10^4$, but as n increases the decay rate of `markovfunm` deteriorates significantly.

We consider a second numerical experiment which keeps the same settings apart from the size of the matrix argument which is fixed to $n = 10^5$. Then, we measure the number of iterations and the computational time needed by the methods using nested sequences of poles, i.e. EK, EDS, `markovfunm`, to reach different target values for the relative error $\frac{\|x - x_\ell\|_2}{\|x\|_2}$. The EK method has the cheapest iteration cost because it exploits the pre-computation of the Cholesky factor of the matrix A for the computation of the orthogonal basis. However, as testified by the results in Table 2, the high number of iterations makes EK competitive only for the high relative error 10^{-1}. The iteration costs of EDS and `markovfunm` is essentially the same since they only differ in the computation of the poles, which requires a negligible portion of the computational time. Therefore, the comparison between EDS and `markovfunm` goes along with the number of iterations which makes the former more efficient.[2] We remark that in the situation where precomputing the Cholesky gives a larger computational benefit, and memory is not an issue, EK may be competitive again.

We conclude the numerical experiments on the inverse square root by considering matrix arguments with different distributions of the eigenvalues. More precisely, we set A as the diagonal matrix of dimension $n = 5 \cdot 10^4$ with the following spectrum configurations:

(i) Equispaced values in the interval $[\frac{1}{n}, 1]$,
(ii) Eigenvalues of trid$(-1, 2 + 10^{-3}, -1)$ (shifted Laplacian),
(iii) 20 Chebyshev points in $[10^{-3}, 10^{-1}]$ and $n - 20$ Chebyshev points in $[10, 10^3]$.

[2] To make a fair comparison between the methods, for this test we relied on the rational Arnoldi implementation found in `markovfunm` for the implementation of Algorithm 1 using EDS poles.

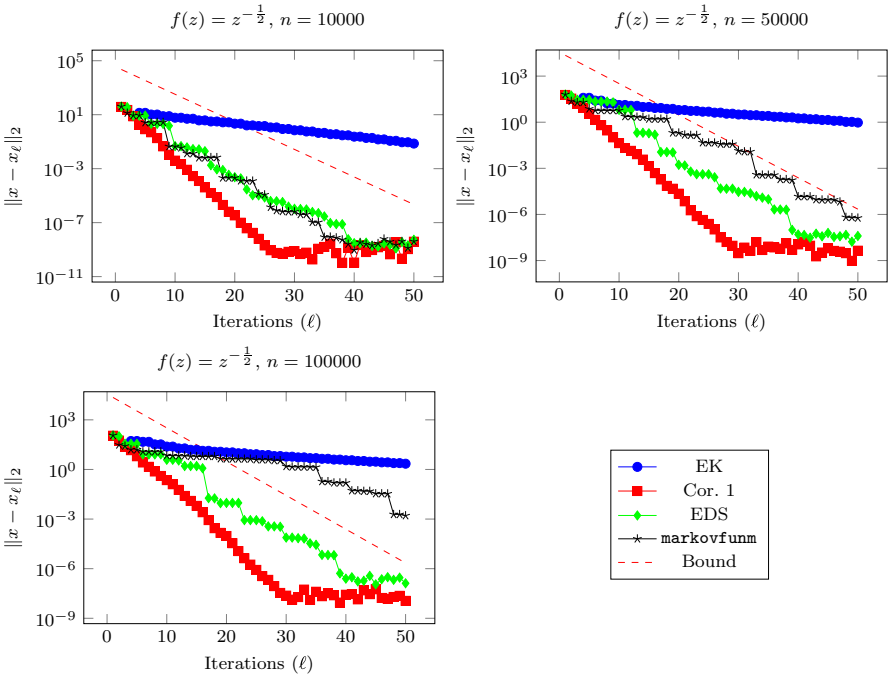


Fig. 2 Convergence history of the different projection spaces for the evaluation of $A^{-\frac{1}{2}}v$, with $A = \text{trid}(-1, 2, -1)$, for different sizes n of the matrix argument. The methods tested are extended Krylov (EK), rational Krylov with the poles from Corollary 1, rational Krylov with nested poles obtained as in Sect. 3.5 (EDS) and rational Krylov with the poles of `markovfunm`. The bound is obtained from Corollary 1

Table 2 Comparison of the time and number of iterations required for computing $A^{-\frac{1}{2}}v$ with different relative tolerances using `markovfunm`, EDS, and extended Krylov

Relative error	`markovfunm`		EDS		Extended Krylov	
	Time (s)	Its	Time (s)	Its	Time (s)	Its
10^{-1}	0.37	18	0.10	7	0.32	20
10^{-2}	0.76	29	0.26	14	2.17	64
10^{-3}	1.17	38	0.37	18	4.79	106
10^{-4}	1.64	47	0.43	20	8.16	144
10^{-5}	2.01	53	0.58	24	12.32	180
10^{-6}	2.56	61	0.82	31	16.72	212

The argument A is the $100{,}000 \times 100{,}000$ matrix $\text{trid}(-1, 2, -1)$

The convergence histories of the different projection spaces are reported in Fig. 3. For all the eigenvalues configurations, EDS and `markovfunm` provide comparable performances. The poles from Corollary 1 performs as EDS and `markovfunm` on (ii) and slightly better on (i) and (iii). Once again, EK is the one providing the slowest convergence rate on all examples.

3.6.3 Other Cauchy–Stieltjes functions

Finally, we test the convergence rate of the different pole selection strategies for the Cauchy–Stieltjes functions $\frac{1-e^{-\sqrt{z}}}{z}$, $z^{-0.2}$, $z^{-0.8}$ and the matrix argument $A = \mathrm{trid}(-1, 2, -1)$.

The results reported in Fig. 4 show that in all cases the poles from Corollary 1 and the extended Krylov method provide the best and the worst convergence rates, respectively. The EDS converges faster than `markovfunm` apart from the case of $z^{-0.2}$ where the two strategies perform similarly.

4 Evaluating Stieltjes functions of matrices with Kronecker structure

We consider the task of computing $f(\mathcal{M})v$ where $\mathcal{M} = I \otimes A - B^T \otimes I$. This problem often stems from the discretizations of 2D differential equations, such as the matrix transfer method used for fractional diffusion equations [32].

We assume that $v = \mathrm{vec}(F)$, where $F = U_F V_F^T$ where U_F and V_F are tall and skinny matrices. For instance, when $f(z) = z^{-1}$, this is equivalent to solving the matrix equation $AX - XB = F$. It is well-known that, if the spectra of A and B are separated, then the low-rank property is numerically inherited by X [5]. For more general functions than z^{-1}, a projection scheme that preserves the Kronecker structure has been proposed in [8] using polynomial Krylov methods. We briefly review it in Sect. 4.1. The method proposed in [8] uses tensorized polynomial Krylov subspaces, so it is not well-suited when A and B are ill-conditioned, as it often happens discretizing differential operators. Therefore, we propose to replace the latter with a tensor product of rational Krylov subspaces and we provide a strategy for the pole selection. This enables a faster convergence and an effective scheme for the approximation of the action of such matrix function in a low-rank format.

The case of Laplace–Stieltjes functions, described in Sect. 4.2, follows easily by the analysis performed for the pole selection with a generic matrix A. The error analysis for Cauchy–Stieltjes functions, presented in Sect. 4.3, requires more care and builds on the theory for the solution of Sylvester equations.

4.1 Projection methods that preserve Kronecker structure

If A, B are $n \times n$ matrices, applying the projection scheme described in Sect. 3 requires to build an orthonormal basis W for a (low-dimensional) subspace $\mathcal{W} \subseteq \mathbb{C}^{n^2}$, together with the projections of $W^*\mathcal{M}W = H$ and $v_{\mathcal{W}} = W^*v$. Then the action of $f(\mathcal{M})$ on v is approximated by:

$$f(\mathcal{M})v \approx W f(H) v_{\mathcal{W}}.$$

The trick at the core of the projection scheme proposed in [8] consists in choosing a tensorized subspace of the form $\mathcal{W} := \mathcal{U} \otimes \mathcal{V}$, spanned by an orthonormal basis of the form $W = U \otimes V$, where U and V are orthonormal bases of $\mathcal{U} \subseteq \mathbb{C}^n$ and

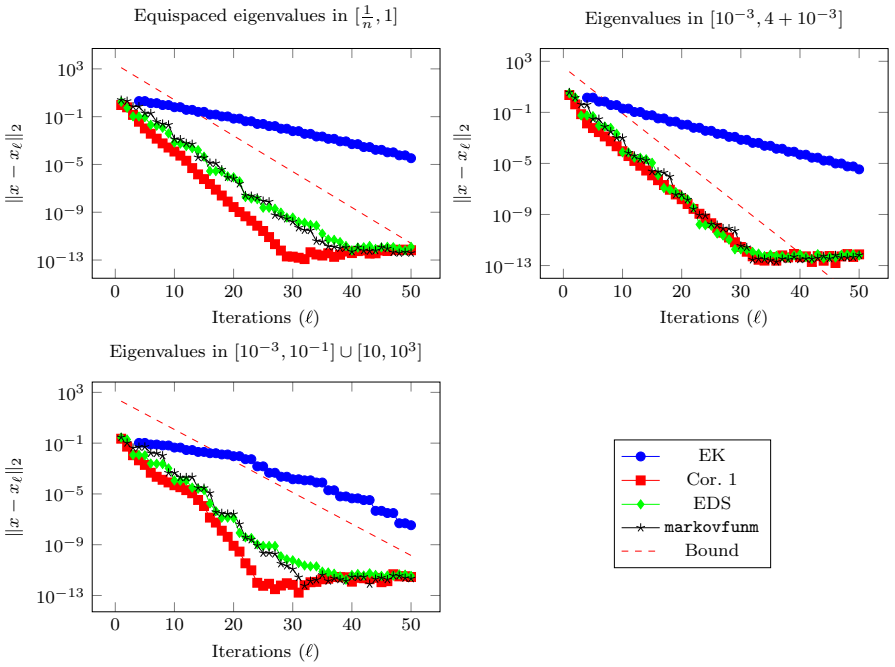


Fig. 3 Convergence history of the different projection spaces for the evaluation of $A^{-\frac{1}{2}}v$ for a diagonal matrix argument of size $50{,}000 \times 50{,}000$ with different eigenvalue distributions. The methods tested are extended Krylov (EK), rational Krylov with the poles from Corollary 1, rational Krylov with nested poles obtained as in Sect. 3.5 (EDS) and rational Krylov with the poles of `markovfunm`. The bound is obtained from Corollary 1

$\mathcal{V} \subseteq \mathbb{C}^n$, respectively. With this choice, the projection of $\mathcal{M}$ onto $\mathcal{U} \otimes \mathcal{V}$ retains the same structure, that is

$$(U \otimes V)^* \mathcal{M}(U \otimes V) = I \otimes A_U - B_V^T \otimes I,$$

where $A_\mathcal{U} = U^* A U$ and $B_\mathcal{V} = V^* B V$.

Since in our case $v = \mathrm{vec}(F)$ and $F = U_F V_F^T$, this enables to exploit the low-rank structure as well. Indeed, the projection of F onto $\mathcal{U} \otimes \mathcal{V}$ can be written as $v_\mathcal{W} = \mathrm{vec}(F_\mathcal{W}) = \mathrm{vec}((U^* U_F)(V_F^T V))$. The high-level structure of the procedure is sketched in Algorithm 1.

At the core of Algorithm 1 is the evaluation of the matrix function on the projected matrix $I \otimes A_\mathcal{U} - B_\mathcal{V}^T \otimes I$. Even when U, V have a low dimension $k \ll n$, this matrix is $k^2 \times k^2$, so it is undesirable to build it explicitly and then evaluate $f(\cdot)$ on it.

When $f(z) = z^{-1}$, it is well-known that such evaluation can be performed in k^3 flops by the Bartels-Stewart algorithm [2], in contrast to the k^6 complexity that would be required by a generic dense solver for the system defined by $I \otimes A_\mathcal{U} - B_\mathcal{V}^T \otimes I$. For a more general function, we can still design a $\mathcal{O}(k^3)$ procedure for the evaluation of $f(\cdot)$ in our case. Indeed, since $A_\mathcal{U}$ and $B_\mathcal{V}$ are Hermitian, we may diagonalize them using a unitary transformation as follows:

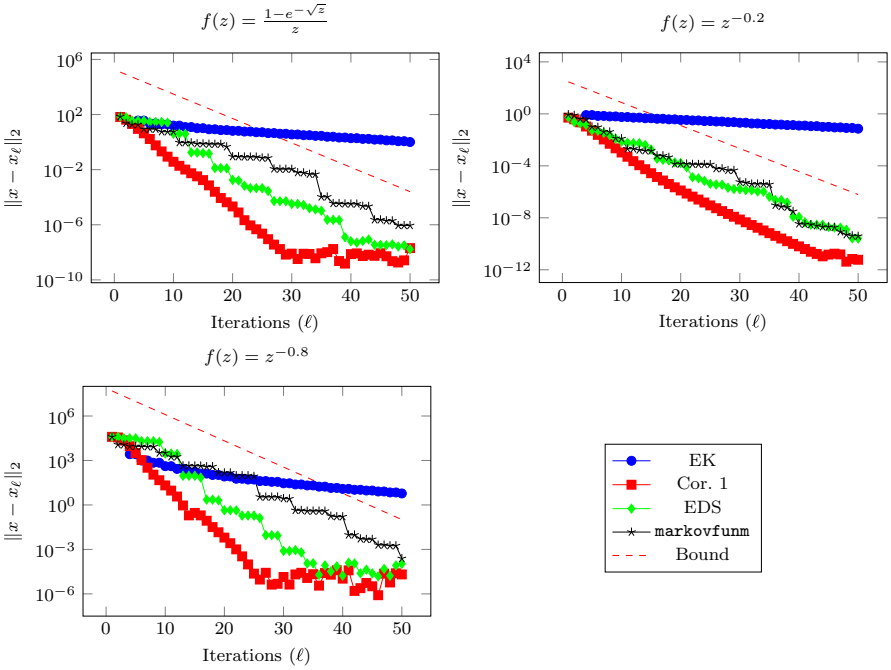


Fig. 4 Convergence history of the different projection spaces for the evaluation of $f(A)v$ for different Cauchy–Stieltjes functions $f(z)$ and the matrix argument$A = \text{trid}(-1, 2, -1)$ of size $50{,}000 \times 50{,}000$. The methods tested are extended Krylov (EK), rational Krylov with the poles from Corollary 1, rational Krylov with nested poles obtained as in Sect. 3.5 (EDS) and rational Krylov with the poles of `markovfunm`. The bound is obtained from Corollary 1

Algorithm 1 Approximate $\text{vec}^{-1}(f(\mathcal{M})\text{vec}(F))$

procedure KroneckerFun(f, A, B, U_F, V_F) ▷ Compute $f(\mathcal{M})\text{vec}(F)$
1: $U, V \leftarrow$ orthonormal bases for the selected subspaces.
2: $A_\mathcal{U} \leftarrow U^* A U$
3: $B_\mathcal{V} \leftarrow V^* B V$
4: $F_\mathcal{W} \leftarrow U^* U_F (V_F^T V)$
5: $Y \leftarrow \text{vec}^{-1}(f(I \otimes A_\mathcal{U} - B_\mathcal{V}^T \otimes I)\text{vec}(F_\mathcal{W}))$.
6: **return** UYV^*
end procedure

$$Q_A^* A_\mathcal{U} Q_A = D_A, \qquad Q_B^* B_\mathcal{V} Q_B = D_B.$$

Then, the evaluation of the matrix function $f(z)$ with argument $I \otimes A_\mathcal{U} - B_\mathcal{V}^T \otimes I$ can be recast to a scalar problem by setting

$$f(I \otimes A_\mathcal{U} - B_\mathcal{V}^T \otimes I)\text{vec}(U^* F V) = \left(\overline{Q}_B \otimes Q_A\right) f(\mathcal{D}) \left(Q_B^T \otimes Q_A^*\right) \text{vec}(U^* F V),$$

where $\mathcal{D} := I \otimes D_A - D_B \otimes I$. If we denote by $X = \mathrm{vec}^{-1}(f(\mathcal{M})\mathrm{vec}(F))$ and with D the matrix defined by $D_{ij} = (D_A)_{ii} - (D_B)_{jj}$, then

$$X = Q_A \left[f^\circ(D) \circ (Q_A^* U^* F V Q_B) \right] Q_B^*,$$

where $\circ$ denotes the Hadamard product and $f^\circ(\cdot)$ the function $f(\cdot)$ applied *componentwise* to the entries of D: $[f^\circ(D)]_{ij} = f(D_{ij})$.

Assuming that the matrices Q_A, Q_B and the corresponding diagonal matrices D_A, D_B, are available, this step requires k^2 scalar function evaluation, plus 4 matrix-matrix multiplications, for a total computational cost bounded by $\mathcal{O}(c_f \cdot k^2 + k^3)$, where c_f denotes the cost of a single function evaluation. The procedure is described in Algorithm 2.

Algorithm 2 Evaluation of $f(I \otimes A_{\mathcal{U}} - B_{\mathcal{V}}^T \otimes I)\mathrm{vec}(U^* F V)$ for normal $k \times k$ matrices $A_{\mathcal{U}}$, $B_{\mathcal{V}}$

```
procedure FUNM_DIAG(f, A_U, B_V, U*FV)
    (Q_A, D_A) ← EIG(A_U)
    (Q_B, D_B) ← EIG(B_V)
    F_W ← Q_A^* U^* F V Q_B
    for i, j = 1, ..., n do
        X_ij ← f((D_A)_ii + (D_B)_jj) · (F_W)_ij
    end for
    return vec(Q_A X Q_B^*)
end procedure
```

4.2 Convergence bounds for Laplace–Stieltjes functions of matrices with Kronecker structure

The study of approximation methods for Laplace–Stieltjes functions is made easier by the following property of the matrix exponential: whenever M, N commute, then $e^{M+N} = e^M e^N$. Since the matrices $B^T \otimes I$ and $I \otimes A$ commute, we have

$$x = \mathrm{vec}(X) = f(\mathcal{M})v = \int_0^\infty e^{-t\mathcal{M}} v \mu(t)\, dt = \mathrm{vec}\left(\int_0^\infty e^{-tA} U_F V_F^T e^{tB} \mu(t)\, dt \right).$$

Consider projecting the matrix $\mathcal{M}$ onto a tensorized subspace spanned by the Kronecker products of unitary matrices $U \otimes V$. This, combined with Algorithm 1, yields an approximation whose accuracy is closely connected with the one of approximating e^{-tA} by projecting using U, and e^{tB} using V. As discussed in Sect. 3, there exists a choice of poles that approximates uniformly well the matrix exponential, and this can be leveraged here as well.

Corollary 5 *Let $f(z)$ be a Laplace–Stieltjes function, $A, -B$ be Hermitian positive definite with spectrum contained in $[a, b]$ and X_ℓ be the approximation of $X =$*

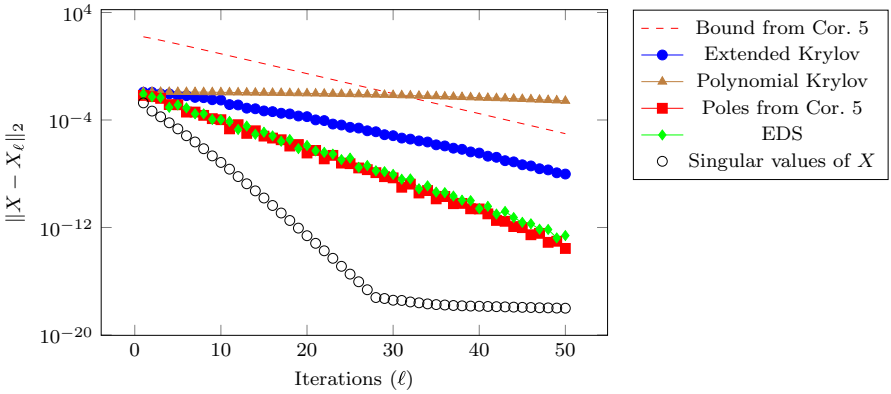


Fig. 5 Convergence history of the different projection spaces for the evaluation of $\varphi_1(\mathcal{M})v$ with the Kronecker structured matrix $\mathcal{M} = I \otimes A + A \otimes I$, where A is of size 1000×1000 and has condition number about $5 \cdot 10^5$. The singular values of the true solution X are reported as well

$\mathrm{vec}^{-1}(f(\mathcal{M})\mathrm{vec}(F))$ *obtained using Algorithm 1 with* $U \otimes V$ *orthonormal basis of* $\mathcal{U}_\mathcal{R} \otimes \mathcal{V}_\mathcal{R} = \mathcal{RK}_\ell(A, U_F, \Psi_\ell^{[a,b]}) \otimes \mathcal{RK}_\ell(B^T, V_F, \Psi_\ell^{[a,b]})$. *Then,*

$$\|X - X_\ell\|_2 \le 16\gamma_{\ell,\kappa} f(0^+)\rho_{[a,b]}^{\frac{\ell}{2}}\|F\|_2.$$

Proof If $f(z)$ is a Laplace–Stieltjes function, we may express the error matrix $X - X_\ell$ as follows:

$$X - X_\ell = \int_0^\infty \left[e^{-tA}Fe^{tB} - Ue^{-tA_\ell}U^*FVe^{tB_\ell}V^*\right]\mu(t)\,dt,$$

where $A_\ell = U^*AU$ and $B_\ell = V^*BV$. Adding and subtracting the quantity $Ue^{-tA_\ell}U^*Fe^{tB}$ yields the following inequalities:

$$\begin{aligned}\|X - X_\ell\|_2 &\le \int_0^\infty \|e^{-tA}F - Ue^{-tA_\ell}(U^*F)\|_2\|e^{tB}\|_2\mu(t)\,dt \\ &\quad + \int_0^\infty \|e^{tB^T}F^T - Ve^{tB_\ell^T}(V^*F^T)\|_2\|e^{-tA_\ell}\|_2\mu(t)\,dt \\ &\le 16\gamma_{\ell,\kappa}\int_0^\infty \mu(t)\,dt \cdot \rho_{[a,b]}^{\frac{\ell}{2}}\|F\|_2\end{aligned}$$

where in the last step we used Corollary 2 for both addends. □

Example 1 To test the proposed projection spaces we consider the same matrix A of Example 3.6.1, and we evaluate the function φ_1 to $\mathcal{M} = I \otimes A + A \otimes I$, applied to a vector $v = \mathrm{vec}(F)$, where F is a random rank 1 matrix, generated by taking the outer product of two unit vectors with normally distributed entries. The results are reported in Fig. 5.

4.3 Convergence bounds for Cauchy–Stieltjes functions of matrices with Kronecker structure

As already pointed out in Sect. 3, evaluating a Cauchy–Stieltjes function requires a space which approximates uniformly well the shifted inverses of the matrix argument under consideration. When considering a matrix $\mathcal{M} = I \otimes A - B^T \otimes I$ which is Kronecker structured, this acquires a particular meaning.

In fact, relying on the integral representation (3) of $f(z)$ we obtain:

$$f(\mathcal{M})v = \int_0^\infty \mu(t)(tI + \mathcal{M})^{-1}\mathrm{vec}(F)\, dt = \int_0^\infty \mu(t)X_t\, dt,$$

where $X_t := \mathrm{vec}^{-1}((tI + \mathcal{M})^{-1}\mathrm{vec}(F))$ solves the matrix equation

$$(tI + A)\, X_j - X_j B = F. \tag{16}$$

Therefore, to determine a good projection space for the function evaluation, we should aim at determining a projection space where these parameter dependent Sylvester equations can be solved uniformly accurate. We note that, unlike in the Laplace–Stieltjes case, the evaluation of the resolvent does not split into the evaluation of the shifted inverses of the factors, and this does not allow to apply Theorem 1 for the factors A and B.

A possible strategy to determine an approximation space is using polynomial Krylov subspaces $\mathcal{K}_m(tI + A, U_F) \otimes \mathcal{K}_m(B^T, V_F)$ for solving (16) at a certain point t. Thanks to the shift invariance of polynomial Krylov subspaces, all these subspaces coincide with $\mathcal{U}_P \otimes \mathcal{V}_P = \mathcal{K}_m(A, U_F) \otimes \mathcal{K}_m(B^T, V_F)$. This observation is at the core of the strategy proposed in [8], which makes use of $\mathcal{U}_P \otimes \mathcal{V}_P$ in Algorithm 1. This allows to use the convergence theory for linear matrix equations to provide error bounds in the Cauchy–Stieltjes case, see [8, Section 6.2].

Since rational Krylov subspaces are usually more effective in the solution of Sylvester equations, it is natural to consider their use in place $\mathcal{U}_P \otimes \mathcal{V}_P$. However, they are not shift invariant, and this makes the analysis not straightforward. Throughout this section, we denote by $U \otimes V$ the orthonormal basis of the tensorized rational Krylov subspace

$$\mathcal{U}_R \otimes \mathcal{V}_R := \mathcal{RK}_\ell(A, U_F, \Psi) \otimes \mathcal{RK}_\ell(B^T, V_F, \Xi) \tag{17}$$

where $\Psi := \{\psi_1, \ldots, \psi_\ell\}$ and $\Xi := \{\xi_1, \ldots, \xi_\ell\}$ are the prescribed poles. We define the following polynomials of degree (at most) ℓ:

$$p(z) := \prod_{j=1, \psi_j \neq \infty}^{\ell} (z - \psi_j), \qquad q(z) := \prod_{j=1, \xi_j \neq \infty}^{\ell} (z - \xi_j) \tag{18}$$

and we denote by $A_\ell = U^* A U$, $B_\ell = V^* B V$ the projected $(\ell k \times \ell k)$-matrices, where k is the number of columns of U_F and V_F.

In Sect. 4.3.1, we recall and slightly extend some results about rational Krylov methods for Sylvester equations i.e., the case $f(z) = z^{-1}$. This will be the building block for the convergence analysis of the approximation of Cauchy–Stieltjes functions in Sect. 4.3.2.

4.3.1 Convergence results for Sylvester equations

Algorithm 1 applied to $f(z) = z^{-1}$ coincides with the Galerkin projection method for Sylvester equations [28], whose error analysis can be found in [3]; the results in that paper relate the Frobenius norm of the residual to a rational approximation problem. We state a slightly modified version of Theorem 2.1 in [3], that enables to bound the residual in the Euclidean norm. The proof is reported in the "Appendix 1".

Theorem 4 *Let $A, -B$ be Hermitian positive definite with spectrum contained in $[a, b]$ and X_ℓ be the approximate solution returned by Algorithm 1 using $f(z) = z^{-1}$ and the orthonormal basis $U \otimes V$ of $\mathcal{U}_R \otimes \mathcal{V}_R = \mathcal{RK}_\ell(A, U_F, \Psi) \otimes \mathcal{RK}_\ell(B^T, V_F, \Xi)$, then*

$$\|AX_\ell - X_\ell B - F\|_2 \leq (1+\kappa) \max\{\theta_\ell(I_A, I_B, \Psi),\ \theta_\ell(I_B, I_A, \Xi)\}\|F\|_2.$$

Remark 5 Using the mixed norm inequality $\|AB\|_F \leq \|A\|_F \|B\|_2$, one can state the bound in the Frobenius norm as well:

$$\|AX_G^{(\ell)} - X_G^{(\ell)} B - F\|_F \leq (1+\kappa)\sqrt{\theta_\ell^2(I_A, I_B, \Psi) + \theta_\ell^2(I_B, I_A, \Xi)} \cdot \|F\|_F,$$

which is tighter than the one in [3].

For our analysis, it is more natural to bound the approximation error of the exact solution X, instead of the norm of the residual. Since the residual is closely related with the backward error of the underlying linear system, bounding the forward error $\|X - X_\ell\|_2$ causes the appearances of an additional condition number.

Corollary 6 *If X_ℓ is the approximate solution of the linear matrix equation $AX - XB = F$ returned by Algorithm 1 as in Theorem 4, then*

$$\|X_\ell - X\|_2 \leq \frac{a+b}{2a^2} \max\{\theta_\ell(I_A, I_B, \Psi),\ \theta_\ell(I_B, I_A, \Xi)\}\|F\|_2$$

Proof We note that $X_\ell - X$ solves the Sylvester equation $A(X_\ell - X) - (X_\ell - X)B = R$, where $R := AX_\ell - X_\ell B - F$ verifies $\|R\|_2 \leq (1+\kappa)$ $\max\{\theta_\ell(I_A, I_B, \Psi),\ \theta_\ell(I_B, I_A, \Xi)\}\|F\|_2$, thanks to Theorem 4. In view of [21, Theorem 2.1] $\|X_\ell - X\|_2$ is bounded by $\frac{1}{2a}\|R\|_2$. □

4.3.2 Error analysis for Cauchy–Stieltjes functions

In view of Eq. 16, the evaluation of Cauchy–Stieltjes function is closely related to solving (in a uniformly accurate way) parameter-dependent Sylvester equations. This connection is clarified by the following result.

Theorem 5 *Let $f(z)$ be a Cauchy–Stieltjes function, $A, -B$ be Hermitian positive definite with spectrum contained in $[a, b]$ and X_ℓ be the approximate evaluation of $f(z)$ returned by Algorithm 1 using the orthonormal basis $U \otimes V$ of the subspace $\mathcal{U}_R \otimes \mathcal{V}_R = \mathcal{RK}_\ell(A, U_F, \Psi) \otimes \mathcal{RK}_\ell(B^T, V_F, \Xi)$. Then,*

$$\|X - X_\ell\|_2 \leq f(2a) \cdot (1+\kappa) \cdot \|F\|_2 \cdot \max_{t \geq 0} \left[\max \left\{ \theta_\ell(I_A, I_B - t, \Psi),\ \theta_\ell(I_B, I_A + t, \Xi) \right\} \right], \tag{19}$$

where $\kappa = \frac{b}{a}$ and $\theta_\ell(\cdot, \cdot, \cdot)$ is as in Definition 5.

Proof Applying the definition of $f(\mathcal{M})$ we have $f(\mathcal{M})\mathrm{vec}(F) = \int_0^\infty (tI + \mathcal{M})^{-1}\mathrm{vec}(F)\mu(t)\, dt$. We note that, for any $t \geq 0$, the vector $\mathrm{vec}(X_t) := (tI + \mathcal{M})^{-1}\mathrm{vec}(F)$ is such that X_t solves the Sylvester equation $(tI + A)X_t - X_t B = F$. Then, we can write X as $X = \int_0^\infty X_t \mu(t)\, dt$.

Let us consider the approximation UY_tV^* to X_t obtained by solving the projected Sylvester equation $(tI + U^*AU)Y_t - Y_t(V^*BV) = U^*FV$, and $Y = \int_0^\infty Y_t \mu(t)\, dt$. We remark that $\mathcal{RK}_\ell(A, U_F, \Psi) = \mathcal{RK}_\ell(tI + A, U_F, \Psi + t)$.

Then, relying on Corollary 6, we can bound the error $R_t := \|X_t - UY_tV^*\|_2$ with

$$R_t \leq C(t) \cdot \max \left\{ \theta_\ell\left(I_A + t, I_B, \Psi + t\right), \theta_\ell\left(I_B, I_A + t, \Xi\right) \right\} \|F\|_2,$$

where $C(t) := \frac{2(t+a+b)}{(t+2a)^2}$. Making use of Lemma 3(i) we get:

$$R_t \leq C(t) \cdot \underbrace{\max \left\{ \theta_\ell\left(I_A, I_B - t, \Psi\right), \theta_\ell\left(I_B, I_A + t, \Xi\right) \right\}}_{:= \Theta_\ell(t)} \|F\|_2.$$

An estimate for the error on X is obtained by integrating R_t:

$$\begin{aligned} \|X - X_\ell\|_2 &\leq \int_0^\infty \mu(t) \frac{2(t + a + b)\|F\|_2}{(t + 2a)^2} \Theta_\ell(t) dt \\ &\leq (1 + \kappa)\|F\|_2 \int_0^\infty \frac{\mu(t)}{t + 2a} \Theta_\ell(t) dt \\ &\leq f(2a) \cdot (1 + \kappa) \cdot \|F\|_2 \cdot \max_{t \geq 0} \Theta_\ell(t), \end{aligned}$$

where we used that the function $\frac{2(t+a+b)}{t+2a}$ is maximum over $[0, \infty]$ at $t = 0$. □

Inspired by Theorem 4, we look at the construction of rational functions that make the quantities $\theta_\ell(I_A, I_B - t, \Psi)$ and $\theta_\ell(I_B, I_A + t, \Xi)$ small. If we choose $\Xi = -\Psi$ then (19) simplifies to

$$\|X - X_\ell\|_2 \leq f(2a) \cdot (1 + \kappa) \cdot \|F\|_2 \cdot \max_{t \geq 0} \theta_\ell(I_A, I_B - t, \Psi), \tag{20}$$

because $\theta_\ell(I_A, I_B - t, \Psi) = \theta_\ell(I_A, -I_A - t, \Psi) = \theta_\ell(-I_A, I_A + t, -\Psi) = \theta_\ell(I_B, I_A + t, -\Psi)$, in view of Lemma 3(iii).

Similarly to the analysis done for Cauchy–Stieltjes function for a generic matrix A, we may consider a Möbius transform that maps the Zolotarev problem involving the point at infinity in a more familiar form. More precisely, we aim at mapping the set $[-\infty, -a] \cup [a, b]$ into $[-1, -\widetilde{a}] \cup [\widetilde{a}, 1]$—for some $\widetilde{a} \in (0, 1)$. Then, we make use of Theorem 3 and Lemma 3(iii) to provide a choice of Ψ that makes the quantity $\theta_\ell(I_A, I_B - t, \Psi)$ small, independently of t.

Lemma 5 *The Möbius transformation*

$$T(z) := \frac{\Delta + z - b}{\Delta - z + b}, \qquad \Delta := \sqrt{b^2 - a^2},$$

maps $[-\infty, -a] \cup [a, b]$ *into* $[-1, -\widetilde{a}] \cup [\widetilde{a}, 1]$, *with* $\widetilde{a} := \frac{\Delta+a-b}{\Delta-a+b}$. *The inverse map* $T(z)^{-1}$ *is:*

$$T^{-1}(z) := \frac{(b + \Delta)z + b - \Delta}{1 + z}.$$

In addition, we have $\widetilde{a}^{-1} \leq 2b/a$, *and therefore* $\rho_{[\widetilde{a},1]} \leq \rho_{[a,2b]}$.

Proof The proof can be easily obtained following the same steps of Lemma 4. As in that case, the overestimate introduced by the inequality $\rho_{[\widetilde{a},1]} \leq \rho_{[a,2b]}$ is negligible in practice (see Remark 3). □

In light of the previous result, we consider Theorem 5 with the choice of poles

$$\Psi = \Psi_{C_2,\ell}^{[a,b]} := T^{-1}(\Psi_\ell^{[\widetilde{a},1]}), \qquad \Xi = -\Psi_{C_2,\ell}^{[a,b]}, \tag{21}$$

where $\Psi_\ell^{[\widetilde{a},1]}$ indicates the set of optimal poles and zeros—provided by Theorem 3—for the domain $[-1, \widetilde{a}] \cup [\widetilde{a}, 1]$. This yields the following.

Corollary 7 *Let* $f(z)$ *be a Cauchy–Stieltjes function with density* $\mu(t)$, $A, -B$ *be Hermitian positive definite with spectrum contained in* $[a, b]$ *and* X_ℓ *the approximate evaluation of* $f(z)$ *returned by Algorithm 1 using the orthonormal basis* $U \otimes V$ *of the subspace* $\mathcal{RK}_\ell(A, U_F, \Psi_{C_2,\ell}^{[a,b]})) \otimes \mathcal{RK}_\ell(B^T, V_F, -\Psi_{C_2,\ell}^{[a,b]}))$, *where* $\Psi_{C_2,\ell}^{[a,b]}$ *is as in (21). Then,*

$$\|X - X_\ell\|_2 \leq 4 \cdot f(2a) \cdot (1 + \kappa) \cdot \|F\|_2 \cdot \rho_{[a,2b]}^\ell, \qquad \rho_{[a,2b]} := \exp\left(-\frac{\pi^2}{\log\left(\frac{8b}{a}\right)}\right).$$

Proof By setting $I_A = I$, $I_B = -I$ in the statement of Theorem 5 we get (20), so that we just need the bound

$$\begin{aligned}&\theta_\ell(I_A, I_B - t, T^{-1}(\Psi_\ell^{[\widetilde{a},1]}))\\ &\quad = \theta_\ell(I_A, -I_A - t, T^{-1}(\Psi_\ell^{[\widetilde{a},1]})) \leq \theta_\ell(I_A, [-\infty, -a], T^{-1}(\Psi_\ell^{[\widetilde{a},1]}))\end{aligned}$$

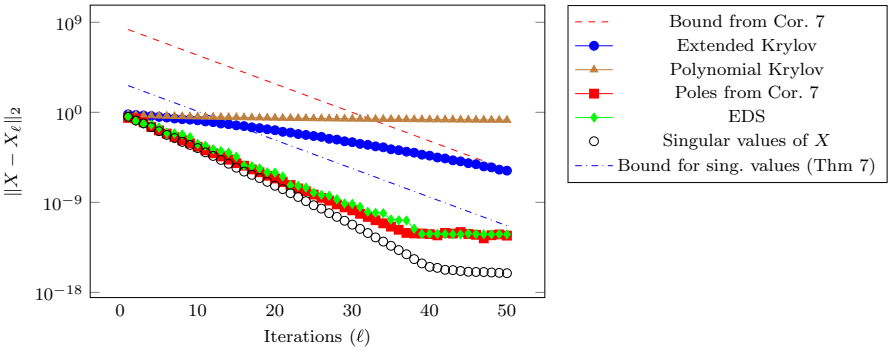


Fig. 6 Convergence history of the different projection spaces for the evaluation of $\mathcal{M}^{-\frac{1}{2}}v$ with the Kronecker structured matrix $\mathcal{M} = I \otimes A + A \otimes I$, where A is of size 1000×1000 and has condition number about $5 \cdot 10^5$. The singular values of the true solution X and the bound given in Theorem 7 are reported as well

$$= \theta_\ell([\widetilde{a}, 1], [-1, -\widetilde{a}], \Psi_\ell^{[\widetilde{a},1]}) \leq 4\rho_{[\widetilde{a},1]}^{\ell},$$

where the first inequality follows from Lemma 3(ii) and the last equality from Lemma 3(iii) applied with the map $T(z)$. The claim follows combining this inequality $\rho_{[\widetilde{a},1]} \leq \rho_{[a,2b]}$ from Lemma 5. □

Example 2 We consider the same matrix A of Example 3.6.2, and we evaluate the inverse square root of $\mathcal{M} = I \otimes A + A \otimes I$, applied to a vector $v = \mathrm{vec}(F)$, where F is a random rank 1 matrix, generated by taking the outer product of two unit vectors with normally distributed entries.

We note that, in Fig. 6, the bound from Corollary 7 accurately predicts the asymptotic convergence rate, even though it is off by a constant; we believe that this is due to the artificial introduction of $(1+\kappa)$ in the Galerkin projection bound, which is usually very pessimistic in practice [3].

4.4 Low-rank approximability of *X*

The Kronecker-structured rational Krylov method that we have discussed provides a practical way to compute the evaluation of the matrix function under consideration. However, it can be used also theoretically to predict the decay in the singular values of the computed matrix X, and therefore to describe its approximability properties in a low-rank format.

4.4.1 Laplace–Stieltjes functions

In the Laplace–Stieltjes case, we may employ Corollary 5 directly to provide an estimate for the decay in the singular values.

Theorem 6 *Let $f(z)$ be a Laplace–Stieltjes function and $\mathcal{M} = I \otimes A - B^T \otimes I$ where $A, -B$ are Hermitian positive definite with spectra contained in $[a, b]$. Then,*

if $\mathrm{vec}(X) = f(\mathcal{M})\mathrm{vec}(F)$, *with* $F = U_F V_F^T$ *of rank* k, *we have*

$$\sigma_{1+\ell k}(X) \le 16\gamma_{\ell,\kappa} f(0^+)\rho_{[a,b]}^{\frac{\ell}{2}}\|F\|_2.$$

Proof We note that the approximation X_ℓ obtained using the rational Krylov method with the poles given by Corollary 5 has rank (at most) ℓk, and $\|X - X_\ell\|_2 \le 16\gamma_{\ell,\kappa} f(0^+)\rho_{[a,b]}^{\frac{\ell}{2}}$. The claim follows by applying the Eckart–Young theorem. □

4.4.2 Cauchy–Stieltjes functions

In the case of Cauchy–Stieltjes function, the error estimate in Corollary 7 would provides a result completely analogue to Theorem 6. However, the bound obtained this way involves the multiplicative factor $1 + \kappa$; this can be avoided relying on an alternative strategy.

The idea is to consider the close connection between the rational problem (10) and the approximate solution returned by the *factored Alternating Direction Implicit method* (fADI) [3,5,6]. More specifically, for $t \ge 0$ let us denote with X_t, the solution of the shifted Sylvester equation

$$(tI + A)X_t - X_t B = U_F V_F^*. \tag{22}$$

In view of (16), X_t is such that $X = \int_0^\infty X_t \mu(t)dt$. Running fADI for ℓ iterations, with shift parameters $T^{-1}(\Psi_\ell^{[\widetilde{a},1]}) = \{\alpha_1, \ldots, \alpha_\ell\}$ and $T^{-1}(-\Psi_\ell^{[\widetilde{a},1]}) = \{\beta_1, \ldots, \beta_\ell\}$, provides an approximate solution $X_\ell^{ADI}(t)$ of (22) such that its column and row span belong to the spaces

$$\mathcal{U}_\ell^{\mathrm{ADI}}(t) = \mathcal{RK}(A, U_F, \{\alpha_1 - t, \ldots, \alpha_\ell - t\}), \quad \mathcal{V}_\ell^{\mathrm{ADI}} = \mathcal{RK}(B^T, V_F, \{\beta_1, \ldots, \beta_\ell\}).$$

Note that the space V_ℓ^{ADI} does not depend on t because the right coefficient of (22) does not depend on t. If we denote by $U_\ell^{\mathrm{ADI}}(t)$ and V_ℓ^{ADI} orthonormal bases for these spaces, we have $X_\ell^{\mathrm{ADI}}(t) = U_\ell^{\mathrm{ADI}}(t)Y_\ell^{\mathrm{ADI}}(t)(V_\ell^{\mathrm{ADI}})^*$, and using the ADI error representation [3,5] we obtain $\|X_t - X_\ell^{\mathrm{ADI}}(t)\|_2 \le \|X_t\|_2 \rho_{[a,2b]}^{\ell}$.

In particular, $X_\ell^{ADI}(t)$ is a uniformly good approximation of X_t having rank (at most) ℓk and its low-rank factorization has the same right factor $\forall t \ge 0$.

Theorem 7 *Let* $f(z)$ *be a Cauchy–Stieltjes function and* $X = \mathrm{vec}^{-1}(f(\mathcal{M})\mathrm{vec}(F))$, *with* $\mathcal{M} := I \otimes A - B^T \otimes I$, *where* $A, -B$ *are Hermitian positive definite with spectra contained in* $[a, b]$. *Then the singular values* $\sigma_j(X)$ *of the matrix* X *verifies:*

$$\sigma_{1+\ell k}(X) \le 4 f(2a)\rho_{[a,2b]}^{\ell}\|F\|_2.$$

Proof Let us define $\widehat{X}_\ell := \int_0^\infty X_\ell^{\mathrm{ADI}}(t)\mu(t)dt = \int_0^\infty U_\ell^{\mathrm{ADI}}(t)\mu(t)dt \cdot Y_\ell^{\mathrm{ADI}}(V_\ell^{\mathrm{ADI}})^T$. Since V_ℓ^{ADI} does not depend on t we can take it out from the integral, and therefore $\widehat{X}_\ell$ has rank bounded by ℓk. Then, applying the Eckart–Young theorem we have

the inequality

$$\sigma_{1+\ell s}(X) \leq \|X - \widehat{X}_\ell\|_2 \leq \int_0^\infty \|X_t - X_\ell^{\mathrm{ADI}}(t)\|_2 \mu(t) dt \leq 4 \int_0^\infty \rho_{[a,2b]}^\ell \|X_t\|_2 \mu(t) dt$$
$$\leq 4 \int_0^\infty \frac{\mu(t)}{(t+2a)} dt \; \rho_{[a,2b]}^\ell \|F\|_2 = 4 f(2a) \rho_{[a,2b]}^\ell \|F\|_2.$$

□

5 Conclusions, possible extensions and open problems

We have presented a pole selection strategy for the rational Krylov methods when approximating the action of Laplace–Stieltjes and Cauchy–Stieltjes matrix functions on a vector. The poles have been shown to provide a fast convergence rate and explicit error bounds have been established. The theory of equidistributed sequences has been used to obtained a nested sequence of poles with the same asymptotic convergence rate. Then, the approach presented in [8] that addresses the case of a matrix argument with a Kronecker sum structure has been extended to use rational Krylov subspaces. We have proposed a pole selection strategy that ensures a good exponential rate of convergence of the error norm. From the theoretical perspective we established decay bounds for the singular values of $\mathrm{vec}^{-1}(f(I \otimes A - B^T \otimes I)\mathrm{vec}(F))$ when F is low-rank. This generalizes the well known low-rank approximability property of the solutions of Sylvester equations with low-rank right hand side. Also in the Kronecker structured case, it has been shown that relying on equidistributed sequences is an effective practical choice.

There are some research lines that naturally stem from this work. For instance, we have assumed for simplicity to be working with Hermitian positive definite matrices. This assumption might be relaxed, by considering non-normal matrices with field of values included in the positive half plane. Designing an optimal pole selection for such problems would require the solution of Zolotarev problems on more general domains, and deserves further study. In addition, since the projected problem is also non-normal, the fast diagonalization approach for the evaluation proposed in Sect. 4.1 might not be applicable or stable, and therefore an alternative approach would need to be investigated.

Acknowledgements The author wish to thank Paul Van Dooren and André Ran for fruitful discussions about Lemma 2.

Proof of Theorem 4

According to [3, Theorem 2.1], the residue $R := AX_\ell - X_\ell B - F$ can be written[3] as $\rho = \rho_{12} + \rho_{21}$, with

$$\rho_{12} = U \cdot r_B^G(A_\ell)^{-1} \cdot F \cdot r_B^G(B) \qquad \rho_{21} = r_A^G(A) \cdot F \cdot r_A^G(B_\ell)^{-1} V^*,$$

where $r_A^G(z) := \det(zI - A_\ell)/p(z)$, and $r_B^G(z) = \det(zI - B_\ell)/q(z)$, with $p(z), q(z)$ defined as in (18). In addition, it is shown that $\rho_{12} = UU^*\rho(I - VV^*)$ and $\rho_{21} = (I - UU^*)\rho VV^*$.

Moreover, the proof of [3, Theorem 2.1] shows that, for any choice of (ℓ, ℓ)-rational function $r_B(x)$ with poles $z_1, \ldots, z_\ell$, we can further decompose ρ_{12} as $\rho_{12} = U(J_1 - J_2)$, where

$$J_1 = \frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (zI - A_\ell)^{-1} U^* F \cdot \frac{r_B(B)}{r_B(z)} V^* dz,$$
$$J_2 = \mathcal{S}_{A_\ell, B}\left(-\frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (zI - A_\ell)^{-1} U^* F V (zI - B_\ell)^{-1} \frac{r_B(B_\ell)}{r_B(z)} V^* dz \right),$$

with $\mathcal{S}_{A,B}(X) := AX - XB$ and Γ_A a path encircling once the interval I_A but not I_B.

With a direct integration we get

$$J_1 = r_B(A_\ell)^{-1} U^* F \cdot r_B(B) V^*,$$

which yields $\|J_1\|_2 \le \|F\|_2 \cdot \|r_B(A_\ell)^{-1}\|_2 \|r_B(B_\ell)\|_2$. Let $\widetilde{B} := VB_\ell V^* - c(I - VV^*)$. Then,[4]

$$\begin{aligned}
&\mathcal{S}_{A_\ell, \widetilde{B}}(\mathcal{S}_{A_\ell, B}^{-1}(J_2)) \\
&= \mathcal{S}_{A_\ell, \widetilde{B}}\left(-\frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (zI - A_\ell)^{-1} U^* F V (zI - B_\ell)^{-1} \frac{r_B(B_\ell)}{r_B(z)} V^* dz \right) \\
&= -\frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (A_\ell - zI)(zI - A_\ell)^{-1} U^* F V (zI - B_\ell)^{-1} \frac{r_B(B_\ell)}{r_B(z)} V^* dz \\
&\quad - \frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (zI - A_\ell)^{-1} U^* F V (zI - B_\ell)^{-1} (zI - B_\ell) \frac{r_B(B_\ell)}{r_B(z)} V^* dz \\
&= \frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} U^* F V (zI - B_\ell)^{-1} \frac{r_B(B_\ell)}{r_B(z)} V^* dz \\
&\quad - \frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (zI - A_\ell)^{-1} U^* F V \frac{r_B(B_\ell)}{r_B(z)} V^* dz
\end{aligned}$$

[3] In the original statement of [3, Theorem 2.1] the residual is decomposed in three parts; the missing term is equal to zero whenever the projection subspace contains the right hand side, which is indeed our case.

[4] The matrix $\widetilde{B}$ is not used in the original proof of [3], which contains a minor typo. There, the operator $\mathcal{S}_{A_\ell, \widetilde{B}}$ is replaced by $\mathcal{S}_{A_\ell, B_\ell}$ which does not have compatible dimensions.

$$= -\frac{1}{2\pi \mathbf{i}} \int_{\Gamma_A} (zI - A_\ell)^{-1} U^* F V \frac{r_B(B_\ell)}{r_B(z)} V^* dz = -r_B(A_\ell)^{-1} U^* F V r_B(B_\ell) V^*,$$

where we used that $V^* \widetilde{B} = B_\ell V^*$ and that the integral on the path Γ_A of $(zI - B_\ell)^{-1}/r_B(z)$ vanishes. Notice that $\|\mathcal{S}_{A,B}(X)\|_2 \leq (\|A\|_2 + \|B\|_2)\|X\|_2$ and $\|\mathcal{S}_{A,B}^{-1}(X)\|_2 \leq \|X\|_2 / \min_{i,j} |\lambda_i(A) - \lambda_j(B)|$ [21, Theorem 2.1]. We get $\|J_2\|_2 \leq \kappa \|r_B(A_\ell)^{-1}\|_2 \|r_B(B_\ell)\|_2 \|F\|_2$ and consequently

$$\|\rho_{12}\| \leq \|J_1\|_2 + \|J_2\|_2 \leq (1 + \kappa) \|r_B(A_\ell)^{-1}\|_2 \|r_B(B_\ell)\|_2 \|F\|_2.$$

Taking the minimum over all (ℓ, ℓ)-rational functions with poles Ξ provides $\|\rho_{12}\|_2 \leq (1+\kappa)\theta_\ell(I_B, I_A, \Xi)\|F\|_2$. Analogously one obtains the similar estimate for ρ_{21} swapping the role of A and B. Since ρ_{12} and ρ_{21} have orthonormal rows and columns, we have $\|\rho_{12} + \rho_{21}\|_2 = \max\{\|\rho_{12}\|_2, \|\rho_{21}\|_2\}$, which concludes the proof.

Bounding an inverse Laplace transform

The proof of Theorem 1 requires to bound the infinity norm of an inverse Laplace transform of a particular rational function, given in Lemma 2. The purpose of this appendix is to provide the details of its proof, that uses elementary arguments even though it is quite long.

Let us consider the following functions, usually called *sine integral* functions, that will be useful in the following proofs:

$$\mathrm{Si}(x) := \int_0^x \frac{\sin(t)}{t}\, dt, \qquad \mathrm{si}(x) := \int_x^\infty \frac{\sin(t)}{t}\, dt.$$

It is known that $\mathrm{si}(x) + \mathrm{Si}(x) = \frac{\pi}{2}$, and that $0 \leq \mathrm{Si}(x) \leq 1.852$ (see [1, Section 6.16]), and therefore $|\mathrm{si}(x)| \leq \frac{\pi}{2}$. We will need the following result, which involved integral of the sinc function by some particular measure.

Lemma 6 *Let $g(t)$ be a decreasing and positive $\mathcal{C}^1$ function over an interval $[0, \gamma]$. Then, the following inequality holds:*

$$\left| \int_0^\gamma \frac{\sin(s) g(s)}{s}\, ds \right| \leq 1.852 \cdot g(0).$$

Proof Integrating by parts yields $I = \mathrm{Si}(s)g(s)\Big|_0^\gamma - \int_0^\gamma \mathrm{Si}(s)g'(s)\, ds$. The first term is equal to $\mathrm{Si}(\gamma)g(\gamma)$, which can be bounded by $1.852 \cdot g(\gamma)$. The second part can be bounded in modulus with

$$\left| \int_0^\gamma \mathrm{Si}(s)g'(s)\, ds \right| \leq -\max_{[0,\gamma]} |\mathrm{Si}(s)| \cdot \int_0^\gamma g'(s)\, ds = (g(0) - g(\nu)) \max_{[0,\nu]} |\mathrm{Si}(s)|,$$

where we have used that $g'(s)$ is negative, so $|g'(s)| = -g'(s)$. Combining the two inequalities we have

$$|I| \leq 1.852 \cdot g(\gamma) + 1.852 \cdot (g(0) - g(\gamma)) = 1.852 \cdot g(0).$$

□

Given a set of positive real points γ_j enclosed in a interval $[a, b]$ with $a > 0$, we define the rational function

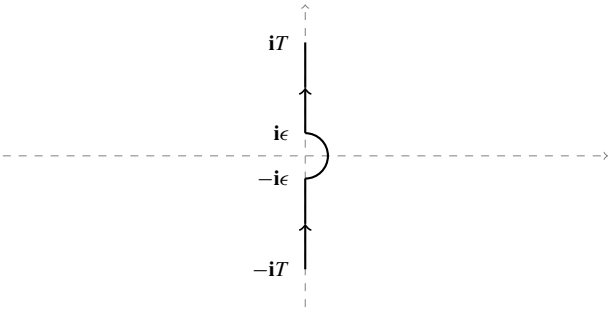
$$\widehat{r}(s) := \frac{1}{s} \frac{p(s)}{p(-s)}. \tag{23}$$

Note that $\widehat{r}(s)$ has poles enclosed in the negative half-plane which ensures that $\lim_{t\to\infty} \mathcal{L}^{-1}[\widehat{r}(s)] = 1$. In particular $\mathcal{L}^{-1}[\widehat{r}(s)]$ is bounded on $\mathbb{R}_+$. We are now ready to prove Lemma 2.

Proof of Lemma 2 We write the inverse Laplace transform as follows:

$$f(t) = \frac{1}{2\pi \mathbf{i}} \lim_{T\to\infty} \int_{-\mathbf{i}T}^{\mathbf{i}T} \widehat{r}(s) e^{st} ds.$$

The integration path needs to be chosen to keep all the poles on its left, including zero. Therefore, we choose the path γ_ϵ that goes from $-\mathbf{i}T$ to $-\mathbf{i}\epsilon$, follows a semicircular path around 0 on the right, and then connects $\mathbf{i}\epsilon$ to $\mathbf{i}T$. It is sketched in the following figure:



Splitting the integral in the three segments we obtain the formula:

$$f(t) = \frac{1}{2\pi \mathbf{i}} \int_{\partial B(0,\epsilon)\cap\mathbb{C}_+} \widehat{r}(s) e^{st}\, ds + \lim_{T\to\infty} \left(\frac{1}{2\pi \mathbf{i}} \int_{-\mathbf{i}T}^{\mathbf{i}\epsilon} \widehat{r}(s) e^{st} ds + \frac{1}{2\pi \mathbf{i}} \int_{\mathbf{i}\epsilon}^{\mathbf{i}T} \widehat{r}(s) e^{st} ds \right). \tag{24}$$

Concerning the first term, it is immediate to see that the integrand uniformly converges to the $1/s$ for $\epsilon \to 0$, and therefore the first terms goes to $\frac{1}{2}$ for ϵ small. We now focus on the second part.

We can rephrase the ratio of polynomials defining $r(s)$ as follows:

$$\frac{p(s)}{p(-s)} = \prod_{j=1}^{\ell} \frac{\gamma_j - s}{\gamma_j + s}, \qquad \gamma_j \in [a, b], \quad 0 < a < b.$$

Then, we note that the above ratio restricted to the points of the form $\mathbf{i}s$ yields a complex number of modulus one, that must have the form $\frac{p(\mathbf{i}s)}{p(-\mathbf{i}s)} = e^{\mathbf{i}\theta(s)}$, where

$$\theta(s) := \arg\left(\frac{p(\mathbf{i}s)}{p(-\mathbf{i}s)}\right) = \sum_{j=1}^{\ell} \arg(\gamma_j - \mathbf{i}s) - \arg(\gamma_j + \mathbf{i}s) = -2\sum_{j=1}^{\ell} \operatorname{atan}\left(\frac{s}{\gamma_j}\right) \in [-\ell\pi, \ell\pi].$$

In particular, $\lim_{s\to\infty} \theta(s) = -\ell\pi$ and for $s > 0$ it holds

$$\ell\pi + \theta(s) = \sum_{j=1}^{\ell} 2\left(\frac{\pi}{2} - \operatorname{atan}\left(\frac{s}{\gamma_j}\right)\right) = \sum_{j=1}^{\ell} 2\left(\int_0^{\infty} \frac{1}{1+x^2}dx - \int_0^{\frac{s}{\gamma_j}} \frac{1}{1+x^2}dx\right) \tag{25}$$

$$= 2\sum_{j=1}^{\ell} \int_{\frac{s}{\gamma_j}}^{\infty} \frac{1}{1+x^2}dx \le 2\sum_{j=1}^{\ell} \int_{\frac{s}{\gamma_j}}^{\infty} \frac{1}{x^2}dx = 2\frac{\sum_{j=1}^{\ell}\gamma_j}{s} \le \frac{2\ell b}{s}. \tag{26}$$

This allows to rephrase the integrals of (24) in the more convenient form

$$\frac{1}{2\pi\mathbf{i}} \int_{\mathbf{i}\epsilon}^{\mathbf{i}T} \widehat{r}(s)e^{st}\, ds = \frac{1}{2\pi\mathbf{i}} \int_{\epsilon}^{T} \mathbf{i}\cdot\widehat{r}(\mathbf{i}s)e^{\mathbf{i}st}\, ds = \frac{(-1)^{\ell}}{2\pi\mathbf{i}} \int_{\epsilon}^{T} \frac{e^{\mathbf{i}(st+\theta(s))}}{s}\, ds.$$

Since we are summing the integral between $[\epsilon, \infty]$ and $[-\infty, \epsilon]$ we can drop the odd part of the integrand, and rewrite their sum as follows:

$$\frac{(-1)^{\ell}}{2\pi\mathbf{i}} \int_{\epsilon}^{T} \frac{e^{\mathbf{i}(st+\theta(s))}}{s}\, ds + \frac{(-1)^{\ell}}{2\pi\mathbf{i}} \int_{-T}^{\epsilon} \frac{e^{\mathbf{i}(st+\theta(s))}}{s}\, ds = \frac{(-1)^{\ell}}{\pi} \int_{\epsilon}^{T} \frac{\sin(st+\theta(s))}{s}\, ds.$$

The above integral is well-defined even if we let $\epsilon \to 0$, we can can take the limit in (24) which yields exactly the value $\frac{1}{2}$ for the first term, and we have reduced the problem to estimate $f(t) = \frac{1}{2} + \frac{(-1)^{\ell}}{\pi}\int_0^{\infty} \frac{\sin(st+\theta(s))}{s}\, ds$. To bound the integral, we split the integration domain in three parts:

$$\frac{1}{\pi}\int_0^{\infty} \frac{\sin(st+\theta(s))}{s} ds = \underbrace{\frac{1}{\pi}\int_0^{\nu} \frac{\sin(st+\theta(s))}{s} ds}_{I_1} + \underbrace{\frac{1}{\pi}\int_{\nu}^{\xi} \frac{\sin(st+\theta(s))}{s} ds}_{I_2}$$
$$+ \underbrace{\frac{1}{\pi}\int_{\xi}^{\infty} \frac{\sin(st+\theta(s))}{s} ds}_{I_3},$$

where we choose $\nu = a\tan(\frac{\pi}{4\ell})$ and $\xi = 4\ell b$. For later use, we note that $\frac{a\pi}{4\ell} \le \nu \le \frac{a}{\ell}$. Concerning I_1, we can further split the integral as $I_1 = \frac{1}{\pi}\int_0^{\nu} \frac{\sin(st)\cos(\theta(s))}{s} ds + \frac{1}{\pi}\int_0^{\nu} \frac{\cos(st)\sin(\theta(s))}{s} ds$. Note that $|\theta(s)| \le 2s\sum_{j=1}^{\ell} \gamma_j^{-1}$, which can be obtained making use of the inequality $|\operatorname{atan}(t)| \le |t|$. We can bound the second integral term as

follows:

$$\frac{1}{\pi}\left|\int_0^{\nu}\frac{\cos(st)\sin(\theta(s))}{s}ds\right| \le \frac{1}{\pi}\int_0^{\nu}\frac{\cos(st)|\theta(s)|}{s}ds \le \nu\frac{1}{\pi}\sum_{j=1}^{\ell}\gamma_j^{-1} \le \frac{1}{\pi},$$

where we have used $\nu \le \frac{a}{\ell}$ and $\sum_j^{\ell}\gamma_j^{-1} \le \frac{\ell}{a}$. The first part can be bounded making use of Lemma 6, by introducing the change of variable $y = st$, which yields

$$\frac{1}{\pi}\int_0^{\nu}\frac{\sin(st)\cos(\theta(s))}{s}ds = \frac{1}{\pi}\int_0^{t\nu}\frac{\sin(y)\cos(\theta(y/t))}{y}dy.$$

Note that on $[0, t\nu]$ the function $\cos(\theta(y/t))$ is indeed decreasing, thanks to our choice of ν, and therefore the above can be bounded in modulus by $\frac{1}{\pi}\left|\int_0^{t\nu}\frac{\sin(y)\cos(\theta(y/t))}{y}dy\right| \le \frac{1.852}{\pi}$, where we have used that $\cos(\theta(0)) = 1$, and applied Lemma 6.

Concerning I_2 we have

$$|I_2| = \left|\frac{1}{\pi}\int_{\nu}^{\xi}\frac{\sin(st+\theta(s))}{s}ds\right| \le \frac{1}{\pi}\int_{\nu}^{\xi}\frac{1}{s}ds = \frac{1}{\pi}\log\left(\frac{\xi}{\nu}\right) \le \frac{1}{\pi}\log\left(\frac{16\ell^2 b}{a\pi}\right)$$

Concerning I_3, we perform the same splitting for a since of a sum that we had for I_1, yielding

$$I_3 = \underbrace{\frac{1}{\pi}\int_{\xi}^{\infty}\frac{\sin(st)\cos(\theta(s))}{s}ds}_{I_4} + \underbrace{\frac{1}{\pi}\int_{\xi}^{\infty}\frac{\cos(st)\sin(\theta(s))}{s}ds}_{I_5}.$$

By using (26) we have that $\forall s \in [\xi, \infty)$:

$$\cos(\theta(s)) = \cos(-\ell\pi + \varphi(s)) = (-1)^{\ell}\cos(\varphi(s)), \qquad 0 \le \varphi(s) \le \frac{2\ell b}{s}.$$

Using the Lagrange expression for the residual of the Taylor expansion we get $\cos(\varphi(s)) \underbrace{=}_{\psi\in[0,\varphi(s)]} 1 - \sin(\psi(s))\varphi(s)$. This enables bounding I_4 as follows:

$$\begin{aligned}
|I_4| &= \frac{1}{\pi}\left|\int_{\xi}^{\infty}\frac{\sin(st)\cos(\theta(s))}{s}ds\right| \le \frac{1}{\pi}\left|\int_{\xi}^{\infty}\frac{\sin(st)}{s}ds\right| \\
&\quad + \frac{1}{\pi}\int_{\xi}^{\infty}\left|\frac{\sin(st)\sin(\psi(s))\varphi(s)}{s}\right|ds \\
&\le \frac{1}{\pi}\left|\int_{\xi}^{\infty}\frac{\sin(st)}{s}ds\right| + \frac{1}{\pi}\int_{\xi}^{\infty}\frac{2\ell b}{s^2}ds = \frac{1}{\pi}\left(|\mathrm{si}(\xi)| + \frac{2kb}{\xi}\right) \le \frac{1}{2} \\
&\quad + \frac{2\ell b}{\xi\pi} \le \frac{1}{2} + \frac{1}{2\pi}.
\end{aligned}$$

Analogously, for bounding I_5, we remark that by using (26) we have that $\forall s \in [\xi, \infty)$:

$$\sin(\theta(s)) = \sin(-\ell\pi + \varphi(s)) = (-1)^{\ell} \sin(\varphi(s)), \qquad 0 \le \varphi(s) \le \frac{2\ell b}{s}.$$

Hence,

$$\begin{aligned} |I_5| &\le \frac{1}{\pi} \left| \int_{\xi}^{\infty} \frac{\cos(st)\sin(\theta(s))}{s}\, ds \right| \le \frac{1}{\pi} \int_{\xi}^{\infty} \frac{|\sin(\varphi(s))|}{s}\, ds \\ &\le \frac{1}{\pi} \int_{\xi}^{\infty} \frac{|\varphi(s)|}{s}\, ds \le \frac{2kb}{\xi\pi} \le \frac{1}{2\pi}. \end{aligned}$$

Combining all these inequalities, we have

$$\begin{aligned} \|f(t)\|_{L^\infty(\mathbb{R}_+)} &\le \frac{1}{2} + \frac{1}{\pi} + \frac{1.852}{\pi} + \frac{1}{\pi} \log\left(16\ell^2 \frac{b}{a\pi}\right) \\ &\quad + \frac{1}{2} + \frac{1}{\pi} \le 2.23 + \frac{2}{\pi} \log\left(4\ell \cdot \sqrt{\frac{b}{\pi a}}\right). \end{aligned}$$

□

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables, vol. 55. Courier Corporation, Chelmsford (1965)
2. Bartels, R.H., Stewart, G.W.: Algorithm 432: solution of the matrix equation $AX + XB = C$. Commun. ACM **15**, 820–826 (1972)
3. Beckermann, B.: An error analysis for rational Galerkin projection applied to the Sylvester equation. SIAM J. Numer. Anal. **49**(6), 2430–2450 (2011). https://doi.org/10.1137/110824590
4. Beckermann, B., Reichel, L.: Error estimates and evaluation of matrix functions via the Faber transform. SIAM J. Numer. Anal. **47**(5), 3849–3883 (2009). https://doi.org/10.1137/080741744
5. Beckermann, B., Townsend, A.: Bounds on the singular values of matrices with displacement structure. SIAM Rev. **61**(2), 319–344 (2019). https://doi.org/10.1137/19M1244433. Revised reprint of "On the singular values of matrices with displacement structure" [MR3717820]
6. Benner, P., Li, R.C., Truhar, N.: On the ADI method for Sylvester equations. J. Comput. Appl. Math. **233**(4), 1035–1045 (2009). https://doi.org/10.1016/j.cam.2009.08.108
7. Benzi, M., Klymko, C.: Total communicability as a centrality measure. J. Complex Netw. **1**(2), 124–149 (2013)
8. Benzi, M., Simoncini, V.: Approximation of functions of large matrices with Kronecker structure. Numer. Math. **135**(1), 1–26 (2017). https://doi.org/10.1007/s00211-016-0799-9
9. Berg, C.: Stieltjes–Pick–Bernstein–Schoenberg and their connection to complete monotonicity. In: Positive Definite Functions: From Schoenberg to Space-Time Challenges, pp. 15–45 (2008)
10. Bernstein, S.: Sur les fonctions absolument monotones. Acta Math. **52**(1), 1–66 (1929). https://doi.org/10.1007/BF02547400
11. Braess, D.: Nonlinear Approximation Theory, vol. 7. Springer, Berlin (2012)
12. Breiten, T., Simoncini, V., Stoll, M.: Low-rank solvers for fractional differential equations. Electron. Trans. Numer. Anal. **45**, 107–132 (2016)
13. Druskin, V., Knizhnerman, L.: Extended Krylov subspaces: approximation of the matrix square root and related functions. SIAM J. Matrix Anal. Appl. **19**(3), 755–771 (1998)

14. Druskin, V., Knizhnerman, L., Zaslavsky, M.: Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts. SIAM J. Sci. Comput. **31**(5), 3760–3780 (2009). https://doi.org/10.1137/080742403
15. Druskin, V., Lieberman, C., Zaslavsky, M.: On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. SIAM J. Sci. Comput. **32**(5), 2485–2496 (2010)
16. Druskin, V., Simoncini, V.: Adaptive rational Krylov subspaces for large-scale dynamical systems. Syst. Control Lett. **60**(8), 546–560 (2011). https://doi.org/10.1016/j.sysconle.2011.04.013
17. Frommer, A., Güttel, S., Schweitzer, M.: Efficient and stable Arnoldi restarts for matrix functions based on quadrature. SIAM J. Matrix Anal. Appl. **35**(2), 661–683 (2014). https://doi.org/10.1137/13093491X
18. Güttel, S.: Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. GAMM-Mitt. **36**(1), 8–31 (2013). https://doi.org/10.1002/gamm.201310002
19. Güttel, S., Knizhnerman, L.: A black-box rational Arnoldi variant for Cauchy–Stieltjes matrix functions. BIT **53**(3), 595–616 (2013). https://doi.org/10.1007/s10543-013-0420-x
20. Hochbruck, M., Ostermann, A.: Exponential integrators. Acta Numer. **19**, 209–286 (2010). https://doi.org/10.1017/S0962492910000048
21. Horn, R.A., Kittaneh, F.: Two applications of a bound on the Hadamard product with a Cauchy matrix. Electron. J. Linear Algebra **3**, 4–12 (1998). https://doi.org/10.13001/1081-3810.1010. Dedicated to Hans Schneider on the occasion of his 70th birthdayDedicated to Hans Schneider on the occasion of his 70th birthday
22. Kalugin, G.A., Jeffrey, D.J., Corless, R.M., Borwein, P.B.: Stieltjes and other integral representations for functions of Lambert W. Integral Transf. Spec. Funct. **23**(8), 581–593 (2012)
23. Kressner, D., Massei, S., Robol, L.: Low-rank updates and a divide-and-conquer method for linear matrix equations. SIAM J. Sci. Comput. **41**(2), A848–A876 (2019). https://doi.org/10.1137/17M1161038
24. Massei, S., Mazza, M., Robol, L.: Fast solvers for two-dimensional fractional diffusion equations using rank structured matrices. SIAM J. Sci. Comput. **41**(4), A2627–A2656 (2019). https://doi.org/10.1137/18M1180803
25. Massei, S., Palitta, D., Robol, L.: Solving rank-structured Sylvester and Lyapunov equations. SIAM J. Matrix Anal. Appl. **39**(4), 1564–1590 (2018). https://doi.org/10.1137/17M1157155
26. Oseledets, I.V.: Lower bounds for separable approximations of the Hilbert kernel. Sb. Math. **198**(3), 137–144 (2007). https://doi.org/10.1070/SM2007v198n03ABEH003842
27. Schweitzer, M.: Restarting and error estimation in polynomial and extended Krylov subspace methods for the approximation of matrix functions. Ph.D. thesis, Universitätsbibliothek Wuppertal (2016)
28. Simoncini, V.: Computational methods for linear matrix equations. SIAM Rev. **58**(3), 377–441 (2016). https://doi.org/10.1137/130912839
29. Susnjara, A., Perraudin, N., Kressner, D., Vandergheynst, P.: Accelerated filtering on graphs using Lanczos method. arXiv preprint arXiv:1509.04537 (2015)
30. Townsend, A., Olver, S.: The automatic solution of partial differential equations using a global spectral method. J. Comput. Phys. **299**, 106–123 (2015). https://doi.org/10.1016/j.jcp.2015.06.031
31. Widder, D.V.: The Laplace Transform. Princeton Mathematical Series, vol. 6. Princeton University Press, Princeton (1941)
32. Yang, Q., Turner, I., Liu, F., Ilić, M.: Novel numerical methods for solving the time-space fractional diffusion equation in two dimensions. SIAM J. Sci. Comput. **33**(3), 1159–1180 (2011). https://doi.org/10.1137/100800634
33. Zolotarev, E.: Application of elliptic functions to questions of functions deviating least and most from zero. Zap. Imp. Akad. Nauk. St. Petersb. **30**(5), 1–59 (1877)