Received: 7 December 2023 | Revised: 9 May 2024 | Accepted: 27 May 2024

DOI: 10.1002/nla.2571

RESEARCH ARTICLE

WILEY

Error bounds for the approximation of matrix functions with rational Krylov methods

Igor Simunec

Faculty of Sciences, Scuola Normale Superiore, Pisa, Italy

Correspondence
Igor Simunec, Faculty of Sciences, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa (PI), Italy.
Email: igor.simunec@sns.it

Funding information
INDAM (Italian Institute of High Mathematics)

Abstract

We obtain an expression for the error in the approximation of $f(A)\boldsymbol{b}$ and $\boldsymbol{b}^T f(A)\boldsymbol{b}$ with rational Krylov methods, where A is a symmetric matrix, $\boldsymbol{b}$ is a vector and the function f admits an integral representation. The error expression is obtained by linking the matrix function error with the error in the approximate solution of shifted linear systems using the same rational Krylov subspace, and it can be exploited to derive both a priori and a posteriori error bounds. The error bounds are a generalization of the ones given in Chen et al. for the Lanczos method for matrix functions. A technique that we employ in the rational Krylov context can also be applied to refine the bounds for the Lanczos case.

KEYWORDS

error bound, matrix function, rational Krylov method

1 | INTRODUCTION

An important problem in numerical linear algebra is the computation of the action of a matrix function $f(A)$[1] on a vector $\boldsymbol{b}$, where the matrix A is usually large and sparse. The matrix-vector product $f(A)\boldsymbol{b}$ can be efficiently approximated by means of polynomial[2–4] or rational Krylov methods,[5–10] without resorting to the expensive computation of the whole matrix function $f(A)$. Each iteration of a polynomial or rational Krylov method requires, respectively, the computation of a matrix-vector product or the solution of a shifted linear system with the matrix A.

In applications it is often essential to have bounds or estimates on the error of the Krylov approximation, in order to either predict the number of iterations needed to reach a given accuracy or to have a reliable stopping criterion. In the literature there have been several papers devoted to developing a priori and a posteriori error bounds and estimates, both in the polynomial[11–13] and in the rational Krylov setting.[14–16] In particular, the authors of Reference 13 analyze the error of the Lanczos method for functions of Hermitian matrices, and by means of the Cauchy integral formula they derive an expression for the error that can be used to obtain both a priori and a posteriori error bounds.

In this work we follow a similar approach to analyze the error in the approximation of $f(A)\boldsymbol{b}$ by means of a rational Krylov method, and we derive both a priori and a posteriori bounds in the case of a symmetric matrix A, for functions that have a certain integral representation. The results that we obtain can be interpreted as a generalization to the rational Krylov case of the ones presented in Reference 13.

The rest of the paper is organized as follows. In Section 1.1 we introduce basic definitions regarding Krylov subspaces and some notation that is used throughout the paper. In Section 2 we establish properties of the residuals of shifted linear systems solved with a rational Krylov method. These properties are used in Section 3 to derive integral expressions of the matrix function error in terms of errors or residuals of shifted linear systems, which are in turn used to obtain a priori

and a posteriori error bounds. These bounds are discussed in detail, respectively, in Sections 3.2 and 3.3. In Section 4 we include an bound on the residual of a linear system solved with a rational Krylov method, which is required for the a priori bounds on the matrix function error. In Section 5 we generalize our approach to quadratic forms $\boldsymbol{b}^T f(A)\boldsymbol{b}$, and in Section 6 we illustrate the bounds with some numerical experiments. Section 7 contains some concluding remarks.

1.1 | Basic definitions

Given a matrix $A \in \mathbb{R}^{n\times n}$ and a vector $\boldsymbol{b} \in \mathbb{R}^n$, a polynomial Krylov subspace is defined by

$$\mathcal{K}_m(A,\boldsymbol{b}) = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, \dots, A^{m-1}\boldsymbol{b}\} = \{p(A)\boldsymbol{b} : p \in \Pi_{m-1}\},$$

where Π_{m-1} denotes the set of polynomials of degree $\leq m-1$. Given a sequence of poles $\{\xi_j\}_{j\geq 1} \subset (\mathbb{C}\cup\{\infty\})\setminus\sigma(A)$, where $\sigma(A)$ denotes the spectrum of A, we can define the denominator polynomial $q_{m-1}(z) = \prod_{j=1,\xi_j\neq\infty}^{m-1}(z-\xi_j)$ and the associated rational Krylov subspace

$$\mathcal{Q}_m(A,\boldsymbol{b}) = q_{m-1}(A)^{-1}\mathcal{K}_m(A,\boldsymbol{b}) = \left\{ r(A)\boldsymbol{b} : r(z) = \frac{p_{m-1}(z)}{q_{m-1}(z)}, \text{with } p_{m-1}\in\Pi_{m-1}\right\}.$$

Note that $\mathcal{K}_m(A,\boldsymbol{b})$ is a special case of a rational Krylov subspace, obtained by taking all poles $\xi_j = \infty$.

The subspaces $\mathcal{Q}_j(A,\boldsymbol{b})$ for $j = 1, \dots, m$ form a nested sequence of strictly increasing dimension, provided that m is smaller than the invariance index M, that is, the smallest integer such that $\mathcal{K}_M(A,\boldsymbol{b}) = \mathcal{K}_{M+1}(A,\boldsymbol{b})$. For simplicity, we are going to assume that this is always the case. An orthonormal basis $V_m = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_m]$ of the subspace $\mathcal{Q}_m(A,\boldsymbol{b})$ can be constructed iteratively using the rational Arnoldi algorithm, introduced by Ruhe .[5]

Using the notation $A_m := V_m^T A V_m$, the matrix-vector product $f(A)\boldsymbol{b}$ can be approximated with

$$\mathtt{f}_m := V_m f(A_m) V_m^T \boldsymbol{b} = V_m f(A_m)\boldsymbol{e}_1 \|\boldsymbol{b}\|_2. \tag{1}$$

Each iteration of the rational Arnoldi algorithm requires the solution of a shifted linear system with the matrix $A-\xi_j I$ and one of the poles ξ_j. One iteration can thus be significantly more expensive than a polynomial Krylov iteration, but the convergence of (1) to $f(A)\boldsymbol{b}$ can be much faster provided that the poles are chosen well. When spectral information of A is available, the poles for the rational Krylov subspace can be chosen by exploiting rational approximations of f on the spectrum or field of values of A .[9,14] Greedy adaptive strategies have also been proposed in the literature, see for instance Reference 17. In the sections that follow we are going to derive a priori and a posteriori bounds for the error $\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2$.

The rational Arnoldi algorithm also outputs two upper Hessenberg matrices $\underline{H_m}$ and $\underline{K_m}$ of size $(m+1)\times m$ that satisfy the rational Arnoldi relation

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m}.$$

We will denote by H_m and K_m the leading principal $m\times m$ blocks of $\underline{H_m}$ and $\underline{K_m}$. If $\xi_m = \infty$, then the last row of $\underline{K_m}$ is zero and K_m is invertible in Sec. 3.1 of Reference 9, so we can rewrite the rational Arnoldi relation as

$$AV_m K_m = V_{m+1}\underline{H_m},$$

and A_m can be computed via the identity

$$A_m = H_m K_m^{-1}.$$

We refer the reader to References 9 and 18 for a more thorough discussion of rational Krylov methods for the computation of matrix functions, and to Reference 19 for an analysis of rational Krylov decompositions.

We consider the class of functions $f : \mathcal{S} \to \mathbb{C}$ that admit the integral expression

$$f(x) = \int_\Gamma (x - z)^{-1} \, \mathrm{d}\mu(z), \tag{2}$$

for a suitable contour $\Gamma \subset \mathbb{C}$ and measure $\mathrm{d}\mu(z)$. The representation (2) reduces to the Cauchy integral formula when f is analytic, $\mathrm{d}\mu(z) = -\frac{1}{2\pi i} f(z)\,\mathrm{d}z$, Γ is a simple closed curve and $\mathcal{S}$ is the interior of Γ, and to the Cauchy–Stieltjes integral representation when $\Gamma = (-\infty, 0]$, $\mathcal{S} = \mathbb{C} \setminus \Gamma$ and $\mathrm{d}\mu(z)$ is a positive measure (see, e.g., Def. 2 of Reference 16). The following are some examples of Cauchy–Stieltjes functions in Ex. 1.3 and 1.4 of Reference 14:

$$\begin{aligned} x^{-\alpha} &= \frac{\sin(\alpha\pi)}{\pi} \int_{-\infty}^{0} \frac{(-z)^{-\alpha}}{x - z}\,\mathrm{d}z, \qquad \alpha \in (0,1), \quad x \in \mathbb{C} \setminus (-\infty, 0], \\ \frac{\log(1+x)}{x} &= -\int_{-\infty}^{-1} \frac{z^{-1}}{x - z}\,\mathrm{d}z, \qquad x \in \mathbb{C} \setminus (-\infty, -1]. \end{aligned} \tag{3}$$

Rational approximations of Cauchy–Stieltjes (and more generally, Markov) functions have been investigated for instance in References 14 and 20.

The integral expression (2) for f still holds when the scalar variable x is replaced by a matrix A with eigenvalues contained in $\mathcal{S}$. This is easy to see when A is diagonalizable; for a general matrix A, we refer to Thm. 6.2.28 of Reference 21 for a proof in the case of the Cauchy integral formula. For Cauchy–Stieltjes functions, one way to prove the matrix identity is to show first that it holds for Jordan blocks by using an integral representation of the derivatives of f (see, e.g., Sec. 3 of Reference 22). Using the matrix version of the integral representation (2), we can write the error in the rational Krylov approximation of $f(A)\boldsymbol{b}$ in the form

$$f(A)\boldsymbol{b} - \mathtt{f}_m = \int_\Gamma \left((A - zI)^{-1}\boldsymbol{b} - V_m (A_m - zI)^{-1} V_m^T \boldsymbol{b} \right) \mathrm{d}\mu(z) = \int_\Gamma \mathtt{err}_m(z)\,\mathrm{d}\mu(z), \tag{4}$$

where $\mathtt{err}_m(z)$ denotes the error in the solution of the shifted linear system $(A - zI)\boldsymbol{x} = \boldsymbol{b}$ using the rational Krylov subspace $\mathcal{Q}_m(A, \boldsymbol{b})$. Motivated by (4), in Section 2 we analyze the solution of shifted linear systems with rational Krylov methods, starting by recalling the polynomial Krylov case.

2 | RESULTS ON RATIONAL KRYLOV METHODS FOR SHIFTED LINEAR SYSTEMS

2.1 | Background: solution of linear systems with FOM

Let us start by recalling some basic facts regarding the solution of shifted linear systems using the full orthogonalization method (FOM),[23] which will simplify the presentation of the next part.

Recall that after m iterations FOM constructs a basis V_m of the polynomial Krylov subspace $\mathcal{K}_m(A, \boldsymbol{b})$ and approximates the solution to the linear system $A\boldsymbol{x} = \boldsymbol{b}$ with

$$\boldsymbol{x}_m = V_m A_m^{-1} \boldsymbol{e}_1 \|\boldsymbol{b}\|_2, \qquad \text{where } A_m = V_m^T A V_m.$$

This is equivalent to imposing that the residual $\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{x}_m$ is orthogonal to the basis V_m.

The residual $\boldsymbol{r}_m$ can be elegantly expressed in terms of the characteristic polynomial of A_m. If we denote by $\chi_m(z)$ the characteristic polynomial of A_m, that is $\chi_m(z) = \det(zI - A_m)$, by eq. (3.8) of Reference 24 we have

$$\boldsymbol{r}_m = \frac{1}{\chi_m(0)} \chi_m(A)\boldsymbol{b}. \tag{5}$$

Moreover, since $\boldsymbol{r}_m \in \mathcal{K}_{m+1}(A, \boldsymbol{b})$ and $\boldsymbol{r}_m \perp \mathcal{K}_m(A, \boldsymbol{b})$, the residual $\boldsymbol{r}_m$ is proportional to the next basis vector $\boldsymbol{v}_{m+1}$ (see, e.g. Prop. 6.7 of Reference 23).

A similar argument can be used for the solution of a shifted linear system $(A - tI)\boldsymbol{x}(t) = \boldsymbol{b}$, for any $t \in \mathbb{R}$. Due to the shift invariance of polynomial Krylov subspaces, that is, $\mathcal{K}_m(A - tI, \boldsymbol{b}) = \mathcal{K}_m(A, \boldsymbol{b})$, after m iterations of FOM applied to the shifted system we have constructed the same basis V_m as in the case of the linear system $A\boldsymbol{x} = \boldsymbol{b}$, and therefore the approximate solution is given by

$$\boldsymbol{x}_m(t) = V_m(A_m^t)^{-1}\boldsymbol{e}_1\|\boldsymbol{b}\|_2, \qquad \text{where } A_m^t := V_m^T(A - tI)V_m = A_m - tI.$$

The residual can be expressed using (5) with A_m^t in place of A_m, yielding

$$\boldsymbol{r}_m(t) = \boldsymbol{b} - (A - tI)\boldsymbol{x}_m(t) = \frac{1}{\chi_m^t(0)}\chi_m^t(A - tI)\boldsymbol{b},$$

where

$$\chi_m^t(z) := \det(zI - A_m^t) = \chi_m(z + t).$$

With simple algebraic manipulations, the above expression can be rewritten as

$$\boldsymbol{r}_m(t) = \frac{1}{\chi_m(t)}\chi_m(A)\boldsymbol{b} = \frac{\chi_m(0)}{\chi_m(t)}\boldsymbol{r}_m(0). \tag{6}$$

In other words, the residuals of shifted linear systems are all collinear, and they are proportional to $\boldsymbol{v}_{m+1}$, the next vector in the Krylov basis. The constant in (6) can be written explicitly in terms of the eigenvalues of the projected matrix A_m. These facts are well known in the literature, see for instance Prop. 2.1 of Reference 25 and Lemma 5 of Reference 3.

2.2 | Solution of shifted linear systems with rational Krylov methods

The results that we derive in this section have strong similarities with the ones presented in Section 2.1 for FOM. Consider the family of shifted linear system

$$(A - zI)\boldsymbol{x} = \boldsymbol{b}, \qquad z \in \mathbb{C}.$$

In analogy to FOM and to (1), we can extract from the rational Krylov subspace $\mathcal{Q}_m(A, \boldsymbol{b})$ the approximate solution

$$\mathtt{x}_m(z) = V_m(A_m - zI)^{-1}\boldsymbol{e}_1\|\boldsymbol{b}\|_2. \tag{7}$$

Let us introduce the notation

$$\begin{aligned}\mathtt{res}_m(z) &= \boldsymbol{b} - (A - zI)\mathtt{x}_m(z)\\ \mathtt{err}_m(z) &= (A - zI)^{-1}\mathtt{res}_m(z),\end{aligned}$$

for the linear system residual and error, respectively.

Lemma 1. *Assuming that K_m is nonsingular, the shifted linear system residual $\mathtt{res}_m(z)$ can be written as*

$$\mathtt{res}_m(z) = -(I - V_mV_m^T)(h_{m+1,m}I - k_{m+1,m}A)\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1}(A_m - zI)^{-1}\boldsymbol{e}_1\|\boldsymbol{b}\|_2, \tag{8}$$

where $h_{m+1,m}$ and $k_{m+1,m}$ are the last subdiagonal elements of $\underline{H_m}$ and $\underline{K_m}$, respectively.

Proof. Let us consider the rational Arnoldi decomposition

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m},$$

obtained after m iterations of the rational Arnoldi algorithm. We can write this relation more explicitly as

$$AV_mK_m + A\boldsymbol{v}_{m+1}k_{m+1,m}\boldsymbol{e}_m^T = V_mH_m + \boldsymbol{v}_{m+1}h_{m+1,m}\boldsymbol{e}_m^T. \tag{9}$$

The ratios of corresponding subdiagonal elements of $\underline{H_m}$ and $\underline{K_m}$ are the poles $\xi_1, \ldots, \xi_m$ of the rational Krylov subspace (see Sec. 5.1 of Reference 18 or Reference 19). In particular, $h_{m+1,m}k_{m+1,m}^{-1} = \xi_m$, so $k_{m+1,m} = 0$ when $\xi_m = \infty$. For all $z \in \mathbb{C}$ we have

$$(A - zI)V_m = V_m(H_mK_m^{-1} - zI) + (h_{m+1,m}I - k_{m+1,m}A)\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1}. \tag{10}$$

Recalling that $A_m = V_m^TAV_m$, we see from (9) that

$$A_m = H_mK_m^{-1} - k_{m+1,m}V_m^TA\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1},$$

and combining with (10) we get for A_m the identity

$$\begin{aligned}(A - zI)V_m &= V_m(A_m - zI) + \big(h_{m+1,m}I - k_{m+1,m}(I - V_mV_m^T)A\big)\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1}\\ &= V_m(A_m - zI) + (I - V_mV_m^T)(h_{m+1,m}I - k_{m+1,m}A)\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1},\end{aligned} \tag{11}$$

where we used the fact that $\boldsymbol{v}_{m+1} \perp \operatorname{span}(V_m)$ for the last equality. Using (11), the residual $\mathtt{res}_m(z)$ can be written as

$$\mathtt{res}_m(z) = -(I - V_mV_m^T)(h_{m+1,m}I - k_{m+1,m}A)\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1}(A_m - zI)^{-1}\boldsymbol{e}_1\|\boldsymbol{b}\|_2,$$

concluding the proof. ∎

This shows that the residuals are collinear for all $z \in \mathbb{C}$. In particular, when $\xi_m = \infty$ we have $k_{m+1,m} = 0$ and (8) simplifies to

$$\mathtt{res}_m(z) = -h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^TK_m^{-1}(A_m - zI)^{-1}\boldsymbol{e}_1\|\boldsymbol{b}\|_2, \tag{12}$$

that is, the residuals are collinear to the next basis vector $\boldsymbol{v}_{m+1}$, as in the polynomial case (see Section 2.1). The identity (12) has been used in past literature, see for instance Sec. 6.6.2 of Reference 18 and Sect. 4.3 of Reference 26.

For simplicity, let us focus first on the case $z = 0$, and let us use the notation $\mathtt{res}_m(0) = \mathtt{res}_m$ and similarly $\mathtt{x}_m(0) = \mathtt{x}_m$. The residual $\mathtt{res}_m$ is orthogonal to $\mathcal{Q}_m(A, \boldsymbol{b})$ because of the factor $(I - V_mV_m^T)$ in (8); recalling that $\xi_m = h_{m+1,m}/k_{m+1,m}$, we also see that $\mathtt{res}_m$ belongs to the subspace

$$(I - \xi_m^{-1}A)\mathcal{Q}_{m+1}(A, \boldsymbol{b}) = q_{m-1}(A)^{-1}\mathcal{K}_{m+1}(A, \boldsymbol{b}) \supset \mathcal{Q}_m(A, \boldsymbol{b}).$$

If we let χ_m denote the characteristic polynomial of A_m, that is $\chi_m(z) = \det(zI - A_m)$, by Lemma 4.5 of Reference 18 we have

$$\chi_m(A)q_{m-1}(A)^{-1}\boldsymbol{b} \perp \mathcal{Q}_m(A, \boldsymbol{b}),$$

and moreover $\chi_m(A)q_{m-1}(A)^{-1}\boldsymbol{b} \in q_{m-1}(A)^{-1}\mathcal{K}_{m+1}(A, \boldsymbol{b})$. The vectors $\chi_m(A)q_{m-1}(A)^{-1}\boldsymbol{b}$ and $\boldsymbol{r}_m$ belong to the same subspace of dimension $m + 1$ which contains $\mathcal{Q}_m(A, \boldsymbol{b})$ and they are both orthogonal to $\mathcal{Q}_m(A, \boldsymbol{b})$, so they must be collinear, that is, there exists a constant $\alpha \in \mathbb{R}$ such that

$$\mathtt{res}_m = \alpha\chi_m(A)q_{m-1}(A)^{-1}\boldsymbol{b}.$$

The value of the constant α can be determined by looking at the constant terms in the identity

$$q_{m-1}(A)\boldsymbol{b} - q_{m-1}(A)A\mathtt{x}_m = \alpha\chi_m(A)\boldsymbol{b},$$

which gives us

$$q_{m-1}(0) = \alpha \chi_m(0).$$

As a consequence, we have the following elegant expression for the residual with $z = 0$,

$$\mathtt{res}_m(0) = \frac{q_{m-1}(0)}{\chi_m(0)} \chi_m(A) q_{m-1}(A)^{-1} \boldsymbol{b}. \tag{13}$$

This result can now be easily generalized to all $z \in \mathbb{C}$. Consider the shifted linear system $(A - zI)\boldsymbol{x} = \boldsymbol{b}$, which has the approximate solution

$$\mathtt{x}_m(z) = V_m (A_m^z)^{-1} \boldsymbol{e}_1 \|\boldsymbol{b}\|_2, \qquad \text{where } A_m^z = V_m^T (A - zI) V_m = A_m - zI.$$

Observe that the subspace $\mathcal{Q}_m(A, \boldsymbol{b})$ with poles $\{\xi_1, \ldots, \xi_m\}$ coincides with the subspace $\mathcal{Q}_m(A - zI, \boldsymbol{b})$ defined using the shifted poles $\{\xi_1 - z, \ldots, \xi_m - z\}$. The denominator polynomial associated to this rational Krylov subspace is therefore

$$q_{m-1}^z(x) = \prod_{j=1}^{m-1} (x - (\xi_j - z)) = q_{m-1}(x + z),$$

and the characteristic polynomial χ_m^z of the projected matrix A_m^z is

$$\chi_m^z(x) = \det(xI - A_m^z) = \chi_m(x + z).$$

If we write the residual $\mathtt{res}_m(z) = \boldsymbol{b} - (A - zI)\mathtt{x}_m(z)$ using (13) with A replaced by $A - zI$, we get

$$\begin{aligned} \mathtt{res}_m(z) &= \frac{q_{m-1}^z(0)}{\chi_m^z(0)} \chi_m^z(A - zI) q_{m-1}^z(A - zI)^{-1} \boldsymbol{b} \\ &= \frac{q_{m-1}(z)}{\chi_m(z)} \chi_m(A) q_{m-1}(A)^{-1} \boldsymbol{b} \\ &= \frac{q_{m-1}(z)}{q_{m-1}(0)} \cdot \frac{\chi_m(0)}{\chi_m(z)} \, \mathtt{res}_m(0). \end{aligned} \tag{14}$$

If we let $\theta_j^{(m)}, j = 1, \ldots, m$, denote the eigenvalues of A_m, we obtain the following more explicit expression for the shifted residual:

$$\mathtt{res}_m(z) = \varphi_m(z) \mathtt{res}_m(0), \qquad \text{where } \varphi_m(z) = \prod_{j=1}^{m-1} \frac{\xi_j - z}{\xi_j} \prod_{j=1}^{m} \frac{\theta_j^{(m)}}{\theta_j^{(m)} - z}.$$

Given $w \in \mathbb{C}$, it is easy to derive from (14) the identity

$$\mathtt{res}_m(z) = \frac{q_{m-1}(z)}{q_{m-1}(w)} \cdot \frac{\chi_m(w)}{\chi_m(z)} \, \mathtt{res}_m(w). \tag{15}$$

Furthermore, comparing (15) with (8), we also obtain

$$\frac{q_{m-1}(z)}{q_{m-1}(w)} \cdot \frac{\chi_m(w)}{\chi_m(z)} = \frac{\boldsymbol{e}_m^T K_m^{-1} (A_m - zI)^{-1} \boldsymbol{e}_1}{\boldsymbol{e}_m^T K_m^{-1} (A_m - wI)^{-1} \boldsymbol{e}_1}.$$

We are going to use the previous expressions in Section 3 to obtain bounds for the error in the approximation of $f(A)\boldsymbol{b}$.

3 | BOUNDS FOR THE MATRIX FUNCTION ERROR

Recalling (4), the error in the approximation of $f(A)\boldsymbol{b}$ can be written as

$$f(A)\boldsymbol{b} - \mathtt{f}_m = \int_\Gamma \mathtt{err}_m(z)\,\mathrm{d}\mu(z) = \int_\Gamma (A - zI)^{-1}\mathtt{res}_m(z)\,\mathrm{d}\mu(z).$$

Using (15) we can write $\mathtt{res}_m(z)$ in terms of $\mathtt{res}_m(w)$ for any $z, w \in \mathbb{C}$. A similar relation also holds for $\mathtt{err}_m(z)$ and $\mathtt{err}_m(w)$, indeed we have

$$\begin{aligned}\mathtt{err}_m(z) &= (A - zI)^{-1}\mathtt{res}_m(z)\\ &= \frac{q_{m-1}(z)}{q_{m-1}(w)} \cdot \frac{\chi_m(w)}{\chi_m(z)}(A - zI)^{-1}\mathtt{res}_m(w)\\ &= \frac{q_{m-1}(z)}{q_{m-1}(w)} \det\big(h_{w,z}(A_m)\big)\; h_{w,z}(A)\;\mathtt{err}_m(w),\end{aligned} \tag{16}$$

where we used the notation $h_{w,z}(t) := \dfrac{t-w}{t-z}$, borrowed from Reference 13. We are also going to use the notation $k_z(t) := (t-z)^{-1}$.

The identities (15) and (16) can be used to prove the following theorem, which gives us an expression for the error in the approximation of $f(A)\boldsymbol{b}$ in terms of the error or residual of shifted linear systems.

Theorem 1. *Assume that f has the integral representation (2), and let $\mathtt{f}_m$ be the approximation (1) to $f(A)\boldsymbol{b}$ after m iterations of a rational Krylov method with denominator polynomial q_{m-1}. Let $\mathcal{D} = \{\xi_1, \dots, \xi_{m-1}\} \cup \sigma(A) \cup \sigma(A_m)$ and let w be a measurable function $w : \Gamma \to \mathbb{C} \setminus \mathcal{D}$. We have*

$$\begin{aligned}f(A)\boldsymbol{b} - \mathtt{f}_m &= \int_\Gamma \frac{q_{m-1}(z)}{q_{m-1}(w(z))} \det\big(h_{w(z),z}(A_m)\big)\; h_{w(z),z}(A)\;\mathtt{err}_m(w(z))\,\mathrm{d}\mu(z)\\ &= \int_\Gamma \frac{q_{m-1}(z)}{q_{m-1}(w(z))} \det\big(h_{w(z),z}(A_m)\big)\; k_z(A)\;\mathtt{res}_m(w(z))\,\mathrm{d}\mu(z),\end{aligned} \tag{17}$$

Proof. In (4), for each $z \in \Gamma$ replace $\mathtt{err}_m(z)$ with the expressions in (16) using $w = w(z)$ to obtain the first identity in (17). The second identity is then easily obtained by recalling the definition of k_z. ∎

Remark 1. An identity similar to the one in Theorem 1 has been given in Corol. 2.5 of Reference 13 for the error of the approximation of $f(A)\boldsymbol{b}$ with the Lanczos method, using the Cauchy integral representation of f. Indeed, Theorem 1 reduces to Corol. 2.5 of Reference 13 by taking $q_{m-1}(z) \equiv 1$, $\mathrm{d}\mu(z) = -\dfrac{1}{2\pi i} f(z)\,\mathrm{d}z$ and a constant $w(z) = z$. Apart from the factors involving the denominator q_{m-1} of the rational Krylov subspace, an important difference from the result in Reference 13 is the use of a function $w(z)$ instead of a constant parameter w. This modification, which might appear minor at a quick glance, turns out to be a crucial element for dealing with the rational Krylov case. We are going to thoroughly discuss this point in Section 3.1.

Although it is not practically feasible to evaluate the error expression in (17) directly since it depends explicitly on the matrix A, it can be employed to derive bounds for the error that are easily computable. The following corollary is a simple consequence of Theorem 1, and it is the analogue of Thm. 2.6 of Reference13 for rational Krylov instead of Lanczos. We use the notation $\|g\|_X = \max_{x\in X} |g|$ for the supremum norm of the function g on the compact set $X \subset \mathbb{C}$, and $\|A\|_2$ for the spectral norm of the matrix A. We denote the eigenvalues of A_m by $\lambda_i(A_m)$, $i = 1, \dots, m$.

Corollary 1. *With the same hypothesis of Theorem 1, assume further that A is symmetric and that for some compact sets $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_m \subset \mathbb{R} \setminus \Gamma$ we have $\sigma(A) \subset \mathcal{S}_0$, and $\lambda_i(A_m) \in \mathcal{S}_i$ for all $i = 1, \dots, m$. Then we have the inequalities*

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \prod_{j=0}^{m} \|h_{w(z),z}\|_{\mathcal{S}_j} \cdot \|\mathtt{err}_m(w(z))\|_2\;|\mathrm{d}\mu(z)|,$$

and

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \prod_{j=1}^{m} \|h_{w(z),z}\|_{S_j} \cdot \|k_z\|_{S_0} \cdot \|\mathtt{res}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|.$$

Proof. Since the eigenvalues of A are contained in S_0, we have

$$\|h_{w(z),z}(A)\|_2 = \max_{\lambda\in\sigma(A)} |h_{w(z),z}(\lambda)| \le \|h_{w(z),z}\|_{S_0}.$$

Similarly, we have

$$|\det\big(h_{w(z),z}(A_m)\big)| = \prod_{j=1}^{m} |h_{w(z),z}(\lambda_j(A_m))| \le \prod_{j=1}^{m} \|h_{w(z),z}\|_{S_j}.$$

By applying the above inequalities to the first identity in Theorem 1, we obtain

$$\begin{aligned}\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 &\le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot |\det(h_{w(z),z}(A_m))| \cdot \|h_{w(z),z}(A)\|_2 \|\mathtt{err}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)| \\ &\le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot \|h_{w(z),z}\|_{S_0} \cdot \prod_{j=1}^{m} \|h_{w(z),z}\|_{S_j} \cdot \|\mathtt{err}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|.\end{aligned}$$

The second inequality follows with the same argument, using the error expression in terms of $\mathtt{res}_m(w(z))$ from Theorem 1. ∎

Observe that if we use a constant function $w(z) \equiv w$ in Corollary 1 the bounds are simplified, since the terms $\|\mathtt{err}_m(w)\|_2$ and $\|\mathtt{res}_m(w)\|_2$ no longer depend on z and thus can be bounded independently from the integral term.

Corollary 1 can be interpreted as either an a priori or an a posteriori bound, depending on the choices for the sets S_j. For example, we can easily obtain an a priori bound by taking $S_j = [a, b] \supset \sigma(A)$ for $j = 0, 1, \dots, m$. On the other hand, once A_m has been computed and its eigenvalues are known, we can obtain an a posteriori bound by taking $S_0 = [a, b]$ and $S_j = \{\lambda_j(A_m)\}$, for $j = 1, \dots, m$. The two following corollaries are restatements of Corollary 1 in the a priori and a posteriori setting, respectively, using the choices described above for the sets $S_j, j = 0, \dots, m$.

Corollary 2. *Under the assumptions of Theorem 1, if A is symmetric with $\sigma(A) \subset [a, b]$ contained in the interior of Γ, we have the a priori bounds*

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot \|h_{w(z),z}\|_{[a,b]}^{m+1} \cdot \|\mathtt{err}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|,$$

and

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot \|h_{w(z),z}\|_{[a,b]}^{m} \cdot \|k_z\|_{[a,b]} \cdot \|\mathtt{res}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|.$$

Corollary 3. *Under the assumptions of Theorem 1, if A is symmetric with $\sigma(A) \subset [a, b]$ contained in the interior of Γ, we have the a posteriori bounds*

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))} \cdot \frac{\chi_m(w(z))}{\chi_m(z)}\right| \cdot \|h_{w(z),z}\|_{[a,b]} \cdot \|\mathtt{err}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|,$$

and

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))} \cdot \frac{\chi_m(w(z))}{\chi_m(z)}\right| \cdot \|k_z\|_{[a,b]} \cdot \|\mathtt{res}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|.$$

If additional a priori information on the spectrum of A is available, it can be incorporated in the sets S_j in order to get more precise bounds. For example, assume that in addition to the spectral interval $[a, b]$ we know that all except for 10 eigenvalues of A are contained in the smaller interval $[a, c]$, with $a < c < b$. Then, by the Cauchy interlacing theorem Thm. 8.1.7 of Reference 27 we can conclude that at most 10 eigenvalues of A_m are not inside the interval $[a, c]$, so we can take $S_j = [a, b]$ for $j = 1, \dots, 10$ and $S_j = [a, c]$ for $j = 11, \dots, m$. This kind of choice for the sets S_j exploits more spectral information of the matrix compared to the bounds in Corollary 2, so it would result in more refined bounds that are able to better capture the convergence behavior of rational Krylov methods.

3.1 | Bound discussion

The error bounds in Corollaries 2 and 3 depend on several parameters, such as the curve Γ and the function $w(z)$, and on the choice between $\|\mathtt{err}_m(w(z))\|_2$ and $\|\mathtt{res}_m(w(z))\|_2$.

Since the factors $\|h_{w,z}\|_{[a,b]}$ and $\|k_z\|_{[a,b]}$ diverge when z approaches the interval $[a, b]$, when using the Cauchy integral formula it is a good idea to take Γ as far as possible from the spectrum of A. For example, for a function such as $f(z) = \sqrt{z}$ with a growth of $o(|z|)$ for $|z| \to \infty$, we can choose Γ as a large circular arc centered in the origin joined with a keyhole contour around the negative real line to avoid the branch cut of f. Since $|f(z)/z| = o(1)$ as $|z| \to \infty$, the integral on the large circular arc in (17) vanishes as the radius goes to infinity, so by taking the limit we can consider the contour $\Gamma = (-\infty, 0]$. See Sec. 3.1 of Reference 13 for further discussion on the choice of Γ.

Regarding the choice between the bound in terms of $\|\mathtt{err}_m(w(z))\|_2$ and the one in terms of $\|\mathtt{res}_m(w(z))\|_2$, it is clear that for an a posteriori bound it makes more sense to use the residual norm, since it can be computed quite cheaply once the Krylov basis is available, while the formulation with the error norm requires the exact solution of the linear system, which is significantly more expensive to obtain and therefore impractical.

On the other hand, for an a priori bound neither $\mathtt{res}_m(w(z))$ nor $\mathtt{err}_m(w(z))$ are available because the Krylov basis has not been constructed yet, so in a realistic scenario they have to be bounded using an a priori bound for the linear system error or residual. Which formulation gives the better bound would depend on the quality of the bound for the linear system error or residual, but numerical experiments seem to indicate that the formulation based on the error is usually more accurate.

The same can be observed for a posteriori bounds as well: the bound in Corollary 3 with the exact error is sharper than the bound with the exact residual. A possible explanation for this fact is that the inequality

$$\|h_{w,z}(A)\mathtt{err}_m(w)\|_2 \le \|h_{w,z}\|_{[a,b]} \cdot \|\mathtt{err}_m(w)\|_2,$$

is tighter than

$$\|(A - zI)^{-1}\mathtt{res}_m(w)\|_2 \le \|k_z\|_{[a,b]} \cdot \|\mathtt{res}_m(w)\|_2,$$

especially when w can be freely chosen (note that the left-hand sides in the two equations above are equal). However, this observation is only meaningful for a priori bounds, because in practice one would always use the residual formulation for an a posteriori bound, since the exact residual norm is cheap to compute when the rational Krylov basis is available. See also Section 3.3 for more details on the computation of the a posteriori bounds.

A crucial factor for the effective use of the bounds in Corollaries 2 and 3 is the choice of the function $w(z)$. The simplest option (and the cheapest for evaluation) is to use a constant function $w(z) \equiv w$, similar to Reference 13, since with this choice we only have to compute or bound the residual (or error) of a single shifted linear system. This turns out to be the most convenient choice for a posteriori bounds, but it often gives underwhelming results when used to obtain a priori bounds, and it may even cause the bounds to diverge (see Section 6.1 for an example); in this setting, it is much more effective to use a general function $w(z)$.

To have a better understanding of this phenomenon, let us focus on the first bound in Corollary 2, and let us look at each factor in the integrand separately:

- the factor $|q_{m-1}(z)/q_{m-1}(w)| = \prod_{j=1}^{m-1} |(z - \xi_j)/(w - \xi_j)|$ is large when w is closer to the set of poles $\{\xi_j\}_j$ compared to z, and small when w is farther than z from the poles;
- the factor $\|h_{w,z}\|_{[a,b]}^{m+1}$ is small when w is closer than z to the spectral interval $[a, b]$, and large otherwise;

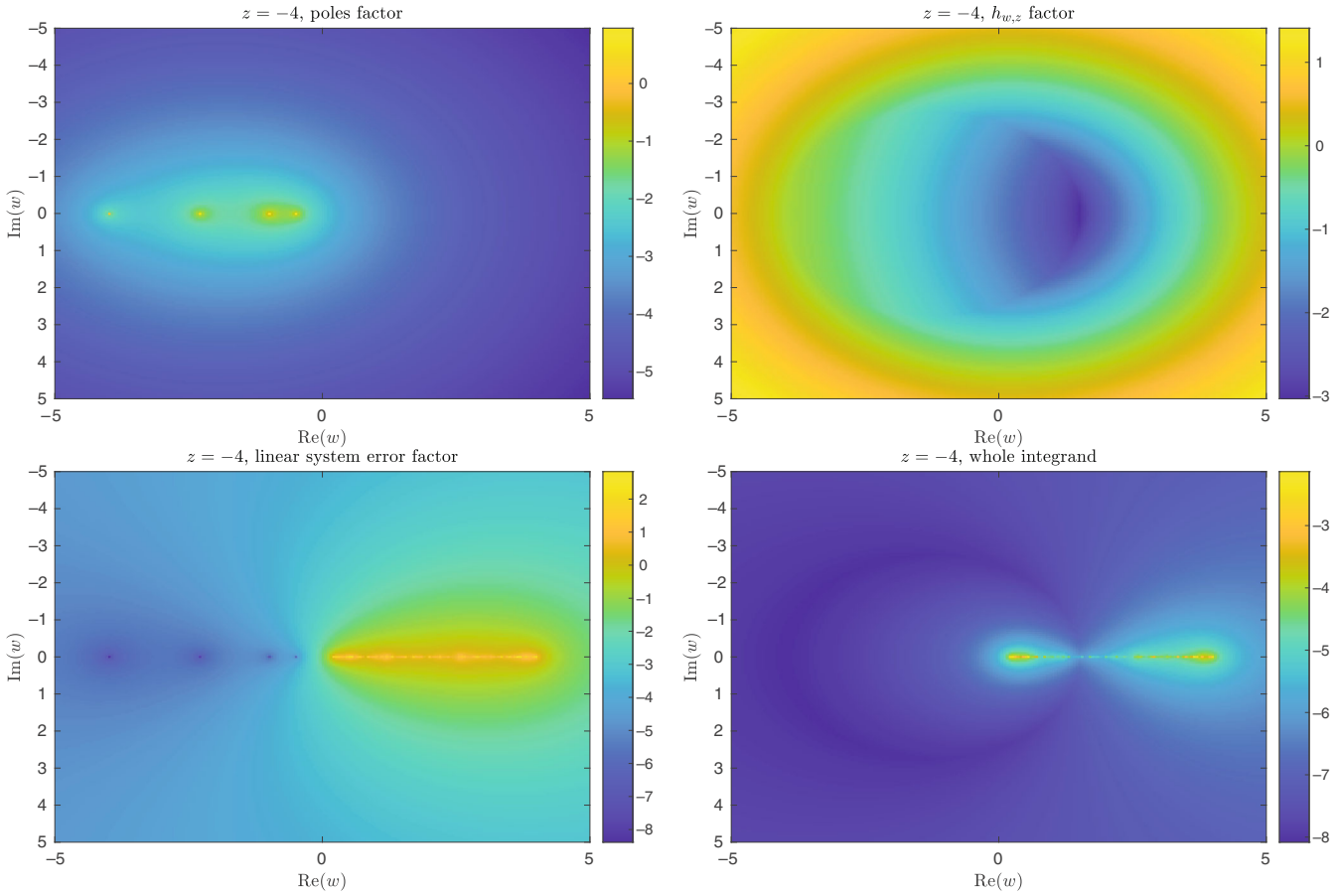

FIGURE 1 Behavior of the different factors in the first a priori bound in Corollary 2 for $w \in [-5, 5] \times [-5i, 5i]$, for $z = -4$.

- the factor $\|\mathtt{err}_m(w)\|_2$ is large when w is close to $[a, b]$, and small when w is close to the poles or when w has a large modulus.

A consequence of the different behavior of the factors and their dependence on z is that the value of the parameter w that minimizes the integrand can be significantly different for different values of z, so it is difficult to find a single w that makes the integrand small for all $z \in \Gamma$. This is illustrated in Figures 1 and 2 with a simple example where the spectrum of the matrix A is contained in $[0.2, 4]$ and the rational Krylov subspace has four poles, for $z = -4$ and $z = -0.3$ and w that varies in the complex square $[-5, 5] \times [-5i, 5i]$. Note that the poles factors in Figures 1 and 2 only differ by a constant multiplicative factor, and that the linear system error factors are exactly the same. However, due to the different behavior of the $h_{w,z}$ factor in the two cases, the full integrands (on the bottom right in Figures 1 and 2) look significantly different.

The values of $\|h_{w,z}\|_{[a,b]}$ in Figures 1 and 2 are computed using Lemma 2, which generalizes Lemma 3.1 of Reference 13 to the case of complex w and real z. We mention that this result can also be generalized to the case of both w and z complex with a similar proof, but the statement is less elegant in that case.

Lemma 2. *Let* $[a, b] \subset \mathbb{R}$, $z \in \mathbb{R} \setminus [a, b]$ *and* $w \in \mathbb{C}$. *Define*

$$\lambda^\star = \frac{|w|^2 - z\,Re(w)}{Re(w) - z},$$

and let

$$h^\star = \begin{cases} |h_{w,z}(\lambda^\star)| = \left|\dfrac{Im(w)}{w - z}\right| & \text{if } \lambda^\star \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

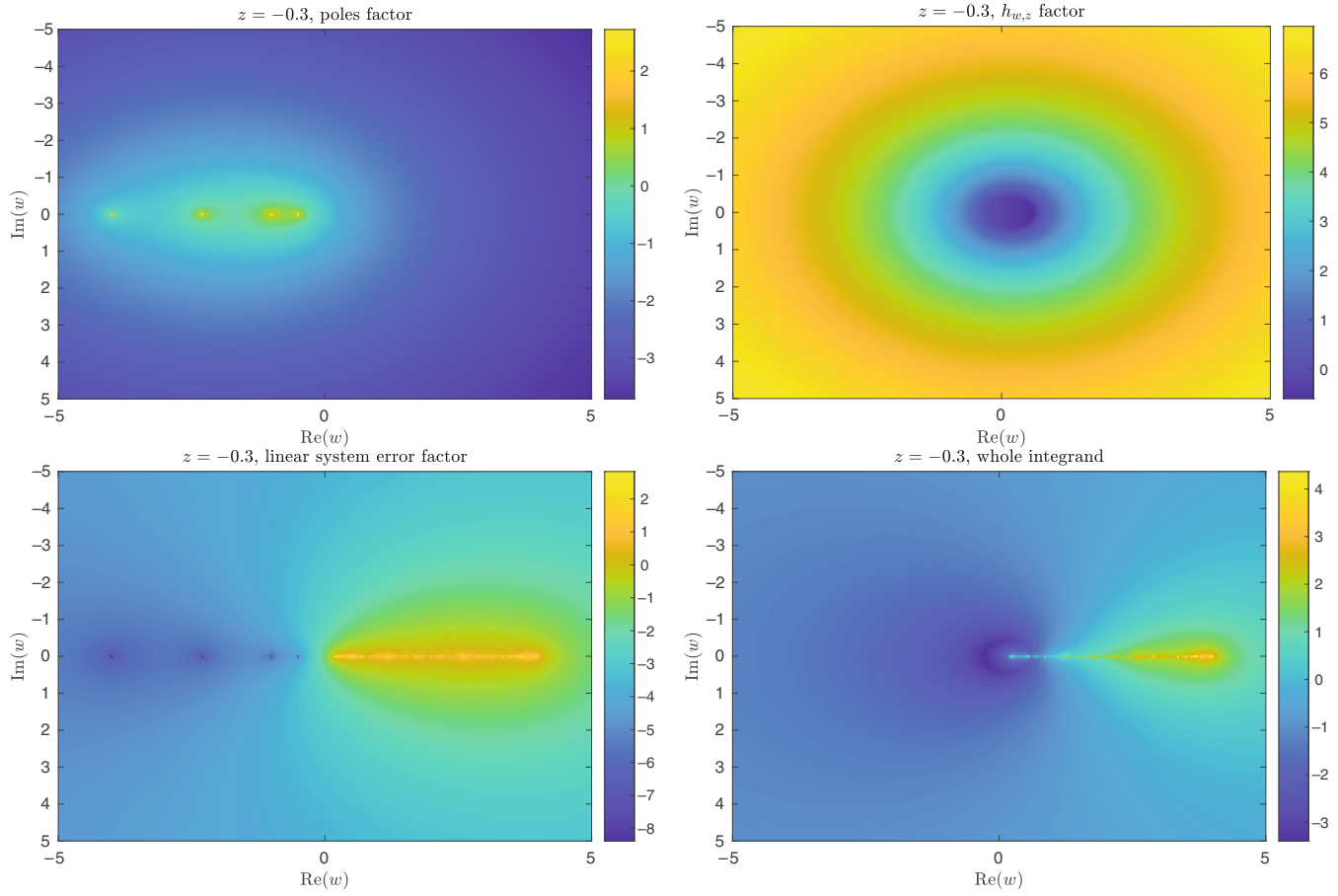

FIGURE 2 Behavior of the different factors in the first a priori bound in Corollary 2 for $w \in [-5, 5] \times [-5i, 5i]$, for $z = -0.3$.

We have

$$\|h_{w,z}\|_{[a,b]} = \max\left\{\left|\frac{a-w}{a-z}\right|, \left|\frac{b-w}{b-z}\right|, h^\star\right\}.$$

Proof. The proof is similar to the proof of Lemma 3.1 of Reference 13. For any $\lambda \in \mathbb{R}$, we have

$$H(\lambda) := |h_{w,z}(\lambda)|^2 = \left|\frac{\lambda - w}{\lambda - z}\right|^2 = \frac{(\lambda - \mathrm{Re}(w))^2 + \mathrm{Im}(w)^2}{(\lambda - z)^2},$$

and

$$\frac{\mathrm{d}H}{\mathrm{d}\lambda} = \frac{2(\lambda - \mathrm{Re}(w))(\lambda - z)^2 - 2(\lambda - z)((\lambda - \mathrm{Re}(w))^2 + \mathrm{Im}(w)^2)}{(\lambda - z)^4}.$$

It follows that $\frac{\mathrm{d}H}{\mathrm{d}\lambda} = 0$ only for $\lambda = \lambda^\star$, so the only possible local maxima of $h_{w,z}(\lambda)$ in $[a, b]$ are for $\lambda = a$, $\lambda = b$ and $\lambda = \lambda^\star$, if $\lambda^\star \in [a, b]$. Finally, with some simple algebraic manipulations it can be shown that $|h_{w,z}(\lambda^\star)| = |\mathrm{Im}(w)/(w - z)|$. ∎

We can see from Figures 1 and 2 that the values of w for which the integrand is smallest are close to z. Indeed, it turns out that the bounds in Corollary 1 are minimized precisely when $w(z) \equiv z$.

Proposition 1. *Under the same assumptions of Corollary 1, the bounds in Corollary 1 are minimized for* $w(z) \equiv z$.

Proof. Let us consider the first bound in Corollary 1. For $w(z) \equiv z$, we have $h_{z,z}(t) = 1$ for all $t \neq z$, so the bound becomes simply

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \leq \int_\Gamma \|\mathtt{err}_m(z)\|_2 \, |\mathrm{d}\mu(z)|.$$

Now, consider any function $w(z)$ that satisfies the assumptions in Theorem 1 and recall (16), which implies that

$$\|\mathtt{err}_m(z)\|_2 = \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot |\det(h_{w(z),z}(A_m))| \cdot \|h_{w(z),z}(A)\ \mathtt{err}_m(w(z))\|_2.$$

Using the assumptions of Corollary 1, we can bound

$$|\det(h_{w(z),z}(A_m))| \leq \prod_{j=1}^{m} \|h_{w(z),z}\|_{S_j},$$

and

$$\|h_{w(z),z}(A)\mathtt{err}_m(w(z))\|_2 \leq \|h_{w(z),z}\|_{S_0} \cdot \|\mathtt{err}_m(w(z))\|_2.$$

If we plug these inequalities in the expression for $\|\mathtt{err}_m(z)\|_2$, we obtain

$$\begin{aligned}\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 &\leq \int_\Gamma \|\mathtt{err}_m(z)\|_2 \, |\mathrm{d}\mu(z)| \\ &\leq \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot \prod_{j=1}^{m} \|h_{w(z),z}\|_{S_j} \cdot \|h_{w(z),z}\|_{S_0} \cdot \|\mathtt{err}_m(w(z))\|_2 \, |\mathrm{d}\mu(z)|,\end{aligned}$$

which coincides with the first bound in Corollary 1. Since this holds for any admissible function $w(z)$, the function $w(z) = z$ must be a minimizer for the right-hand side of the bound.

The same can be proved for the second bound in Corollary 1, by observing that for $w(z) = z$ the bound simplifies to

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \leq \int_\Gamma \|k_z\|_{S_0} \cdot \|\mathtt{res}(z)\|_2 \, |\mathrm{d}\mu(z)|,$$

and then using the same strategy, recalling the identity

$$\|\mathtt{res}(z)\|_2 = \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right| \cdot |\det(h_{w(z),z}(A_m))| \cdot \|\mathtt{res}(w(z))\|_2.$$

■

Although using $w(z) \equiv z$ is not very practical because it would require the computation of the residual or error norm of a different shifted linear system for each point used in the discretization of the integral, the result of Proposition 1 provides some theoretical insight regarding the choice of the function $w(z)$. In the following subsections we discuss practical ways to choose the function $w(z)$ in order to approximately minimize the bounds, both in the a priori and a posteriori setting.

3.2 | A priori bounds

Let us consider the first a priori bound in Corollary 2. For any $z \in \Gamma$, the ideal choice for $w(z)$ would be one that minimizes the integrand in the bound, that is,

$$w(z) = \operatorname*{argmin}_{w \in \mathbb{C}\setminus D} \left(\frac{1}{|q_{m-1}(w)|} \cdot \|h_{w,z}\|_{[a,b]}^{m+1} \cdot \|\mathtt{err}_m(w)\|_2 \right),$$

Algorithm 1. A priori error bound for $f(A)\boldsymbol{b}$

Input: symmetric matrix $A \in \mathbb{R}^{n\times n}$, vector $\boldsymbol{b} \in \mathbb{R}^n$, real interval $[a,b]$ s.t. $[a,b] \supset \sigma(A)$, function f with integral representation (2), poles $\{\xi_1, \dots, \xi_{m-1}\}$.

Output: a priori bound $\varepsilon_m \in \mathbb{R}$ such that $\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \varepsilon_m$, with $\mathtt{f}_m$ defined in (1) using poles $\{\xi_1, \dots \xi_{m-1}\}$.

1: Choose a discrete set $\mathcal{W} \subset \mathbb{C} \setminus [a,b] \cup \{\xi_1, \dots, \xi_{m-1}\}$ (e.g., take $\mathcal{W}$ to be a discretization of Γ).

2: For all $w \in \mathcal{W}$, compute $\varphi_m(w)$ such that $\|\mathtt{err}_m(w)\| \le \varphi_m(w)$ (e.g., using Theorem 2).

3: Choose a quadrature formula for the integral over Γ, with quadrature nodes $\mathcal{Z} \subset \Gamma$.

4: For all $z \in \mathcal{Z}$, compute $w(z) = \operatorname{argmin}_{w\in\mathcal{W}} \left(\frac{1}{|q_{m-1}(w)|} \cdot \|h_{w,z}\|_{[a,b]}^{m+1} \cdot \varphi_m(w) \right)$, using [Lemma 3.1 of Reference 13] or Lemma 2 to evaluate $\|h_{w,z}\|_{[a,b]}$.

5: Using the chosen quadrature formula, compute $\varepsilon_m \approx \int_\Gamma \left| \frac{q_{m-1}(z)}{q_{m-1}(w(z))} \right| \cdot \|h_{w(z),z}\|_{[a,b]}^{m+1} \cdot \varphi_m(w(z)) \, |\mathrm{d}\mu(z)|$.

where we have dropped the factors that do not depend on w, and the set D is the one defined in Theorem 1. We have seen in Proposition 1 that the bound is minimized by taking $w(z) = z$. However, in practice we do not have access to $\|\mathtt{err}_m(w)\|_2$ a priori, so we must instead rely on an a priori bound for the linear system error of the form $\|\mathtt{err}_m(w)\|_2 \le \varphi_m(w)$, and then select

$$w(z) = \operatorname*{argmin}_{w\in\mathbb{C}\setminus D} \left(\frac{1}{|q_{m-1}(w)|} \cdot \|h_{w,z}\|_{[a,b]}^{m+1} \cdot \varphi_m(w) \right). \tag{18}$$

In general, the choice $w(z) = z$ is not going to be the minimizer of (18), but we may heuristically expect that choosing $w(z)$ near z gets us close to the optimum. The optimization problem can be approximately solved numerically by replacing $\mathbb{C} \setminus D$ with a finite set $\mathcal{W}$ of candidates for w, and by finding for each z the value of $w \in \mathcal{W}$ that minimizes the function on the right-hand side. The a priori bound is summed up in Algorithm 1.

Note that in a practical implementation of the bounds, the integral will be approximately evaluated using a quadrature formula, so the number of values of $z \in \Gamma$ for which we have to find the best $w \in \mathcal{W}$ will be finite. For each quadrature node z, the minimization over the set $\mathcal{W}$ requires the computation of $\|h_{w,z}\|_{[a,b]}$ for all $w \in \mathcal{W}$, which can be done efficiently using Lemma 3.1 of Reference 13 if w is real, or with Lemma 2 if z is real. Additionally, for each $w \in \mathcal{W}$ we have to compute $q_{m-1}(w)$ and $\varphi_m(w)$. Although these computations may significantly increase the cost of quadrature if the cardinality of $\mathcal{W}$ is large, it is important to note that their cost is independent of the matrix dimension n, and therefore it eventually becomes negligible compared to the cost of constructing the Krylov basis as the size of A increases.

A possible choice for the set $\mathcal{W}$ is a discretization of the integration contour Γ, which ensures that for each $z \in \Gamma$ there exists a $w \in \mathcal{W}$ that is quite close. Alternatively, one could decide to use a different set $\mathcal{W}_z$ for each $z \in \Gamma$, for instance by taking a small number of points in the neighborhood of z, or even a single point, such as the $w \in \mathcal{W}$ that is closest to z. However, this kind of choice heavily relies on the heuristic that the optimum of (18) is attained for w close to z, which is not always true. For example, this heuristic may fail when the upper bound $\|\mathtt{err}_m(w)\|_2 \le \varphi_m(w)$ significantly overestimates the error norm only for certain values of w.

3.3 | A posteriori bounds

Let us consider now the a posteriori bounds. We focus on the second bound in Corollary 3, since the formulation with the linear system residual is the most likely to be useful in practice. In this case, the best choice for the function $w(z)$ would be one that minimizes the integrand in the bound, that is,

$$w(z) = \operatorname*{argmin}_{w\in\mathbb{C}\setminus D} \frac{|\chi_m(w)|}{|q_{m-1}(w)|} \cdot \|\mathtt{res}_m(w)\|_2,$$

Algorithm 2. A posteriori error bound for $f(A)\boldsymbol{b}$

Input: symmetric matrix $A \in \mathbb{R}^{n\times n}$, vector $\boldsymbol{b} \in \mathbb{R}^n$, real interval $[a,b]$ s.t. $[a,b] \supset \sigma(A)$, function f with integral representation (2), V_m orthonormal basis of $\mathcal{Q}_m(A,\boldsymbol{b})$ with poles $\{\xi_1,\dots,\xi_{m-1}\}$, $A_m = V_m^T A V_m$.

Output: a posteriori bound $\epsilon_m \in \mathbb{R}$ such that $\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \le \epsilon_m$, with $\mathtt{f}_m$ defined in (1) using poles $\{\xi_1,\dots\xi_{m-1}\}$.

1: Choose any $w \in \mathbb{C} \setminus [a,b] \cup \{\xi_1,\dots,\xi_{m-1}\}$.

2: Compute $\|\mathtt{res}_m(w)\|_2 = (A - wI)V_m(A_m - wI)^{-1}e_1\|\boldsymbol{b}\|_2$.

3: Choose a quadrature formula for the integral over Γ, and use it compute

$$\epsilon_m \approx \int_\Gamma \left| \frac{q_{m-1}(z)}{q_{m-1}(w)} \cdot \frac{\chi_m(w)}{\chi_m(z)} \right| \cdot \|k_z\|_{[a,b]} \cdot \|\mathtt{res}_m(w)\|_2 \, |\mathrm{d}\mu(z)|.$$

where again we have only included the terms that depend on w. However, it follows from (15) that the quantity $\mathtt{res}_m(w) \cdot \chi_m(w)\, q_{m-1}(w)^{-1}$ is constant in w, so any value of w can be used to evaluate the bound and would give the same result obtained by taking $w(z) \equiv z$, which minimizes the bound by Proposition 1. The most efficient choice is then to use a single w for all $z \in \Gamma$. Note that with the choice $w(z) = z$, the bound can be rewritten in the simple form

$$\|f(A)\boldsymbol{b} - \mathtt{f}_m\| \le \int_\Gamma \|(A - zI)^{-1}\|_2 \cdot \|\mathtt{res}_m(z)\|_2 \, |\,\mathrm{d}\mu(z)|,$$

which can be easily obtained directly from (4). The advantage of the formulation with $w(z) \equiv w$ is that the corresponding bound can be evaluated by computing the residual of a single linear system.

The residual-based a posteriori bound is summarized in Algorithm 2. The evaluation of the a posteriori bound after m Krylov iterations requires the computation of the residual

$$\mathtt{res}_m(w) = (A - wI)\mathtt{x}_m(w) = (A - wI)V_m(A_m - wI)^{-1}\boldsymbol{e}_1\|\boldsymbol{b}\|_2.$$

Note that it is likely that an eigenvalue decomposition of A_m has been already computed for the evaluation of $f(A_m)$, so the $m \times m$ linear system $(A_m - wI)^{-1}\boldsymbol{e}_1$ can be solved with just $O(m^2)$ cost, with no need to compute a factorization of the matrix $A_m - wI$. So we can compute $\mathtt{x}_m(w)$ for $O(mn)$ cost, and we can obtain $\mathtt{res}_m(w)$ with one additional matrix-vector multiplication with A. Alternatively, if we can keep in memory the matrix AV_m (whose columns are often computed anyway in the rational Arnoldi algorithm), we can entirely avoid additional matrix-vector multiplications with A and cheaply compute $\mathtt{res}_m(w)$ for a cost of $O(mn)$. In contrast, the first bound in Corollary 3 formulated in terms of linear system errors is significantly more expensive to compute, since in that case taking w constant is not equivalent to taking $w(z) = z$ and hence optimization of $w(z)$ is beneficial, and the exact solution of several shifted linear systems is required to compute the corresponding error norms.

4 | AN A PRIORI BOUND FOR THE LINEAR SYSTEM RESIDUAL

In this section we present an a priori bound on the residual of a shifted linear system, which can be used to bound $\|\mathtt{res}_m(w(z))\|_2$ and $\|\mathtt{err}_m(w(z))\|_2$ in Corollary 2.

Consider the shifted linear system

$$(A - wI)\boldsymbol{x} = \boldsymbol{b}, \qquad w \in \mathbb{C}.$$

This linear system coincides with a Sylvester equation in which one of the coefficient matrices is 1×1,

$$AX - XB = \boldsymbol{b}\boldsymbol{c}^T,$$

where $X = \boldsymbol{x} \in \mathbb{R}^{n\times 1}$, $B = w \in \mathbb{R}^{1\times 1}$ and $\boldsymbol{c} = 1 \in \mathbb{R}$. As a consequence, the residual norm $\|\mathtt{res}_m(w)\|_2$ of the shifted linear system can be bounded by using the bounds developed in Reference 28 for the residual of Sylvester equations

solved using the rational Krylov subspaces $\mathcal{Q}_m(A, \boldsymbol{b})$ and $\mathcal{Q}_1(B^T, \boldsymbol{c})$. Since B and $\boldsymbol{c}$ are scalars, a Krylov subspace of dimension 1 generated with B^T and $\boldsymbol{c}$ coincides with the whole space, and the approximate solution of the Sylvester equation obtained using the method from Reference 28 coincides with the rational Krylov approximation to the solution of a shifted linear system described in Section 2.2. The theorem that follows is obtained by specializing Thm. 2.3 of Reference 28 to the case of a shifted linear system with a symmetric positive definite matrix A. The notation in the statement of the theorem is changed from Reference 28 in order to match the notation used in the rest of the paper.

Theorem 2 (Thm. 2.3 of Reference 28). *Assume that A is symmetric with spectrum contained in $[\lambda_{min}, \lambda_{max}]$, and let the set of poles $\{\xi_1, \ldots, \xi_{m-1}\}$ be closed under complex conjugation. For any $w \in \mathbb{R} \setminus [\lambda_{min}, \lambda_{max}] \cup \{\xi_1, \ldots, \xi_{m-1}\}$, we have*

$$\|\mathtt{res}_m(w)\|_2 \leq \|\boldsymbol{b}\|_2 \, (4 + c) \, \gamma_m,$$

where

$$c = 2\sqrt{2}\sqrt{\kappa_2(A - wI)},$$

and

$$\gamma_m = \left|\frac{\varphi(w) - 1}{\varphi(w) + 1}\right| \cdot \prod_{j=1}^{m-1} \left|\frac{\varphi(w)/\varphi(\xi_j) - 1}{\varphi(w)/\varphi(\overline{\xi_j}) + 1}\right|, \qquad \varphi(z) = \sqrt{\frac{z - \lambda_{max}}{z - \lambda_{min}}}.$$

Proof. We show that this theorem follows by applying the second part of Thm. 2.3 of Reference 28 to the linear system $(A - wI)\boldsymbol{x} = \boldsymbol{b}$, interpreted as a Sylvester equation. In this proof we use some notation from.[28]

Note that in our setting we have $B = w \in \mathbb{R}$, so the size of the Krylov subspace associated to B is $n = 1$, and the corresponding pole is $z_{B,1} = \infty$. On the other hand, the poles $\{z_{A,1}, \ldots, z_{A,m}\}$ used for the rational Krylov subspace associated to A are given by $\{\infty, \xi_1, \ldots, \xi_{m-1}\}$, where the pole at infinity appears because of the different notation used in the definition of rational Krylov subspaces in Reference 28.

The numerical range of B is $W(B) = \{w\}$, so we have $u_{B,n} \equiv 0$ and

$$\gamma_{B,A} = \max_{z \in [\lambda_{\min}, \lambda_{\max}]} u_{B,n}(z) = 0.$$

We have

$$c_3 = 2\sqrt{2}\sqrt{\frac{\max\{|\lambda_{\min} - w|, |\lambda_{\max} - w|\}}{\min\{|\lambda_{\min} - w|, |\lambda_{\max} - w|\}}} = 2\sqrt{2}\sqrt{\kappa_2(A - wI)} =: c,$$

and

$$\gamma_{A,B} = u_{A,m}(w) = \left|\frac{\varphi(w) - 1}{\varphi(w) + 1}\right| \cdot \prod_{j=1}^{m-1} \left|\frac{\varphi(w)/\varphi(\xi_j) - 1}{\varphi(w)/\varphi(\overline{\xi_j}) + 1}\right| =: \gamma_m, \qquad \varphi(z) := \sqrt{\frac{z - \lambda_{\max}}{z - \lambda_{\min}}},$$

where the first factor in γ_m corresponds to the pole at infinity.

Observing that the Sylvester equation residual $\|S_{A,B}(X - X^G_{m,n})\|_F$ from Reference 28 corresponds to the shifted linear system residual $\|\mathtt{res}_m(w)\|_2$ in our setting, by Thm. 2.3 of Reference 28 we can conclude that

$$\begin{aligned}\|\mathtt{res}_m(w)\|_2 &\leq \|\boldsymbol{b}\|_2 \, (4 \ \max\{\gamma_{A,B} + \gamma_{B,A}\} + c_3 \, (\gamma_{A,B} + \gamma_{B,A})) \\ &= \|\boldsymbol{b}\|_2 \, (4 + c_3) \, \gamma_m.\end{aligned}$$

■

Remark 2. The original statement of Thm. 2.3 of Reference 28 holds also for nonsymmetric matrices, by using the numerical ranges of A and B and the associated Green functions. The main assumption on the matrices

A and B is that $W(A) \cap W(B) = \emptyset$, which in our setting corresponds to $w \notin W(A)$. Here we prefer to consider only the symmetric case, since it is difficult to compute the numerical range of a general matrix in practice.

Remark 3. The bound on the residual norm $\|\mathtt{res}_m(w)\|_2$ also provides a bound for the error norm $\|\mathtt{err}_m(w)\|_2$ via the inequality

$$\begin{aligned}\|\mathtt{err}_m(w)\|_2 &\le \|(A - wI)^{-1}\|_2 \cdot \|\mathtt{res}_m(w)\|_2 \\ &\le \max\{|\lambda_{\min} - w|^{-1}, |\lambda_{\max} - w|^{-1}\} \cdot \|\mathtt{res}_m(w)\|_2.\end{aligned} \tag{19}$$

The right-hand side of the inequality (19) can be used in (18) as the function $\varphi_m(w)$.

5 | ERROR BOUNDS FOR QUADRATIC FORMS

In this section we adapt the statement of Theorem 1 to the case of quadratic forms with $f(A)$, and we show that it leads to improved bounds by exploiting the symmetry of A. Let us denote by $\mathtt{q}_m = \boldsymbol{b}^T \mathtt{f}_m$ the approximation to $\boldsymbol{b}^T f(A)\boldsymbol{b}$ obtained after m steps of a rational Krylov method. From (4) we immediately have the identity

$$\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathtt{q}_m = \int_\Gamma \boldsymbol{b}^T \mathtt{err}_m(z)\, \mathrm{d}\mu(z).$$

We can write $\boldsymbol{b} = (A - zI)\mathtt{x}_m(z) + \mathtt{res}_m(z)$, and using the symmetry of A we have

$$\begin{aligned}\boldsymbol{b}^T \mathtt{err}_m(z) &= \mathtt{x}_m(z)^T (A - zI)\mathtt{err}_m(z) + \mathtt{res}_m(z)^T \mathtt{err}_m(z) \\ &= \mathtt{x}_m(z)^T \mathtt{res}_m(z) + \mathtt{res}_m(z)^T \mathtt{err}_m(z) \\ &= \mathtt{res}_m(z)^T \mathtt{err}_m(z),\end{aligned}$$

where in the last equality we used the fact that $\mathtt{res}_m(z) \perp \mathcal{Q}_m(A, \boldsymbol{b})$, so $\mathtt{x}_m(z)^T \mathtt{res}_m(z) = 0$; see Lemma 1. The error for the approximation of $\boldsymbol{b}^T f(A)\boldsymbol{b}$ is therefore given by

$$\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathtt{q}_m = \int_\Gamma \mathtt{res}_m(z)^T \mathtt{err}_m(z)\, \mathrm{d}\mu(z). \tag{20}$$

Combining this error expression with (15) leads to the following theorem, which is the equivalent of Theorem 1 for quadratic forms.

Theorem 3. *Assume that A is symmetric and that f has the integral representation (2), and denote by $\mathtt{q}_m$ the approximation to $\boldsymbol{b}^T f(A)\boldsymbol{b}$ after m iterations of a rational Krylov method with denominator polynomial q_{m-1}. Let $\mathcal{D} = \{\xi_1, \dots, \xi_{m-1}\} \cup \sigma(A) \cup \sigma(A_m)$ and let w be a function $w : \Gamma \to \mathbb{C} \setminus \mathcal{D}$. We have*

$$\begin{aligned}\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathtt{q}_m &= \int_\Gamma \left(\frac{q_{m-1}(z)}{q_{m-1}(w(z))} \det(h_{w(z),z}(A_m))\right)^2 \mathtt{res}_m(w(z))^T k_z(A) \mathtt{res}_m(w(z))\, \mathrm{d}\mu(z) \\ &= \int_\Gamma \left(\frac{q_{m-1}(z)}{q_{m-1}(w(z))} \det(h_{w(z),z}(A_m))\right)^2 \mathtt{res}_m(w(z))^T h_{w(z),z}(A) \mathtt{err}_m(w(z))\, \mathrm{d}\mu(z).\end{aligned}$$

Proof. In (20), use $\mathtt{err}_m(z) = (A - zI)^{-1}\mathtt{res}_m(z)$ and use (15) to replace $\mathtt{res}_m(z)$ with $\mathtt{res}_m(w)$. ∎

The following corollary can be derived from the two error expressions in Theorem 3 with a proof similar to the one of Corollary 1.

Corollary 4. *With the same hypothesis as in Theorem 3, assume that for some $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_m \subset \mathbb{C}$ we have $\sigma(A) \subset \mathcal{S}_0$, and $\lambda_i(A_m) \in \mathcal{S}_i$ for all $i = 1, \dots, m$. Then we have*

$$\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathtt{q}_m\|_2 \le \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right|^2 \prod_{j=1}^m \|h_{w(z),z}\|_{\mathcal{S}_j}^2 \cdot \|k_z\|_{\mathcal{S}_0} \cdot \|\mathtt{res}_m(w(z))\|_2^2\, |\mathrm{d}\mu(z)|,$$

and similarly

$$\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathfrak{q}_m\|_2 \leq \int_\Gamma G_m(z)\ |\mathrm{d}\mu(z)|,$$

where

$$G_m(z) = \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right|^2 \prod_{j=1}^{m} \|h_{w(z),z}\|_{S_j}^2 \cdot \|h_{w(z),z}\|_{S_0} \cdot \|\mathtt{res}_m(w(z))\|_2 \cdot \|\mathtt{err}_m(w(z))\|_2.$$

Note that the first bound in Corollary 4 has the same structure as the second one in Corollary 1, with the difference that several of the factors are squared. This is in line with the fact that the error $\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathfrak{q}_m\|_2$ is related to the error in the approximation of f with rational functions of type $(2m-1, 2m-2)$, instead of type $(m-1, m-1)$, so we can expect to have roughly twice the convergence rate compared to the error $\|f(A)\boldsymbol{b} - \mathfrak{f}_m\|_2$ (see e.g. Remark 3.2 of Reference 9 or Prop. 4.4 of Reference 26). A similar result has been obtained in Sec. 6 of Reference 13 for the Lanczos method.

Corollary 4 can be specialized by choosing different sets S_j to obtain a priori and a posteriori bounds. This is formalized in the following two corollaries.

Corollary 5. *Under the assumptions of Theorem 3, we have the a priori bounds*

$$\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathfrak{q}_m\|_2 \leq \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right|^2 \|h_{w(z),z}\|_{[a,b]}^{2m} \cdot \|k_z\|_{[a,b]} \cdot \|\mathtt{res}_m(w(z))\|_2^2\ |\mathrm{d}\mu(z)|,$$

and

$$\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathfrak{q}_m\|_2 \leq \int_\Gamma H_m(z)\ |\mathrm{d}\mu(z)|,$$

where

$$H_m(z) = \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))}\right|^2 \prod_{j=1}^{m} \|h_{w(z),z}\|^{2m+1} \cdot \|\mathtt{res}_m(w(z))\|_2 \cdot \|\mathtt{err}_m(w(z))\|_2.$$

Corollary 6. *Under the assumptions of Theorem 3, we have the a posteriori bound*

$$\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathfrak{q}_m\|_2 \leq \int_\Gamma \left|\frac{q_{m-1}(z)}{q_{m-1}(w(z))} \cdot \frac{\chi_m(w(z))}{\chi_m(z)}\right|^2 \cdot \|k_z\|_{[a,b]} \cdot \|\mathtt{res}_m(w(z))\|_2^2\ |\mathrm{d}\mu(z)|.$$

Remark 4. With the same arguments used in the proof of Proposition 1, it can be shown that the bounds in Corollary 4 are minimized by taking $w(z) = z$, and with this choice the bounds have much simpler expressions that can be derived directly from (20). From these expressions, it is easy to see that for $w(z) = z$ the second bound in Corollary 5 is always sharper than the first one, because of the inequality $\|\mathtt{err}_m(z)\|_2 \leq \|k_z\|_{[a,b]}\|\mathtt{res}_m(z)\|_2$. In the experiments of Section 6 we observe this also when $w(z)$ is an approximate minimizer of the integrand, assuming that the exact linear system error and residual norms are known. On the other hand, when using a priori bounds on the linear system error and residual norms, which bound is better depends on the linear system bounds that we use. In particular, if we have an a priori bound on the linear system residual and we use it to obtain a bound on the error via (19), then the first bound in Corollary 5 is always sharper than the second one, because of the inequality $\|k_z\|_{[a,b]} \leq \|h_{w,z}\|_{[a,b]}\|k_w\|_{[a,b]}$.

Let us briefly discuss how to select the function $w(z)$. In the case of a priori bounds, similarly to the case of $f(A)\boldsymbol{b}$, we should select $w(z)$ as a minimizer of the integrand evaluated at z, where in practice $\|\mathtt{res}_m(w)\|_2$ is replaced by an a priori bound of the form $\psi_m(w) \geq \|\mathtt{res}_m(w)\|_2$, and $\|\mathtt{err}_m(w)\|_2$ is replaced by $\varphi_m(w) \geq \|\mathtt{err}_m(w)\|_2$. In the end, we obtain

$$w(z) = \operatorname*{argmin}_{w \in \mathbb{C}\setminus D} \left(\frac{1}{|q_{m-1}(w)|^2} \cdot \|h_{w,z}\|_{[a,b]}^{2m} \cdot \psi_m(w)^2 \right),$$

for the first bound in Corollary 5, and

$$w(z) = \underset{w\in\mathbb{C}\setminus D}{\operatorname{argmin}}\left(\frac{1}{|q_{m-1}(w)|^2}\cdot\|h_{w,z}\|_{[a,b]}^{2m+1}\cdot\varphi_m(w)\psi_m(w)\right),$$

for the second bound. In practice, the minimization problem can be solved approximately by taking w in a finite set $\mathcal{W}$, which can be chosen for example as a discretization of Γ, similarly to the case of bounds for $f(A)\boldsymbol{b}$.

When considering a posteriori bounds the situation is simpler, since by (15) for any $w \in \mathbb{C}\setminus D$ we have

$$\left(\frac{q_{m-1}(z)}{q_{m-1}(w)}\cdot\frac{\chi_m(w)}{\chi_m(z)}\right)\mathtt{res}_m(w) = \mathtt{res}_m(z),$$

so we can choose $w(z) \equiv w$ for any fixed w and obtain the same bound.

6 | NUMERICAL EXPERIMENTS

In this section we include some experiments that demonstrate the performance of the error bounds for the approximation of $f(A)\boldsymbol{b}$ and $\boldsymbol{b}^T f(A)\boldsymbol{b}$ when the function f is either Cauchy–Stieltjes or analytic in a neighborhood of the spectrum of A. We use the functions $f(z) = z^{-1/2}$ and $f(z) = \log(1+z)/z$ with the Cauchy–Stieltjes formulations in (3), and the function $f(z) = z^{1/2}$ with the Cauchy integral formula on the contour $\Gamma = (-\infty, 0]$. In all experiments, the poles used to construct the rational Krylov subspace are the asymptotically optimal poles for Cauchy–Stieltjes functions proposed in Reference 16. We assume that an interval $[a, b]$ that contains $\sigma(A)$ is known. This is required to compute the poles, as well as for the evaluation of the a priori and a posteriori bounds. Note that if no spectral information on A is available, it is still possible to compute estimates of the extremal eigenvalues of A by using the matrix A_m computed while running the rational Krylov method, and use them to estimate the a posteriori bounds. The integrals required for the evaluation of the bounds are computed numerically using the MATLAB `integral` function.

6.1 | Comparison between bounds with and without optimization

We first conduct a simple experiment to show that the optimization of the parameter w discussed in Section 3.2 is fundamental for the applicability of a priori bounds. We consider the computation of $A^{-1/2}\boldsymbol{b}$, for a 1000×1000 symmetric matrix A with logspaced eigenvalues in the interval $[10^{-1}, 10^{1}]$, and we compare the error with the residual-based a posteriori bound from Corollary 3 with $w = 0$ (`a posteriori bound`, $w = 0$), and the error-based a priori bounds from Corollary 2 for some fixed values of w (`a priori bound`, $w = 0, -1, 20$) and with the optimization described in Algorithm 1, where $\mathcal{W}$ contains 50 logspaced points in $[-10^4, -10^{-4}]$ (`a priori bound, optimized` w). The linear system error norms in the a priori bounds are bounded using (19) and Theorem 2. The results are depicted in Figure 3. The dotted lines are a priori bounds obtained using 500 different values of w in $\mathbb{R}\setminus[10^{-1}, 10^{1}]$. It turns out that none of the a priori bounds without optimization converge in this case, and they give no useful information, with several of them quickly diverging. On the other hand, the bound obtained by optimizing w for each $z \in \Gamma$ used to evaluate the integral correctly captures the convergence behavior, although it is off by a couple of order of magnitude. The a posteriori bound requires no optimization, confirming the discussion in Section 3.3.

In the next experiment we compare the impact of the cardinality of the set $\mathcal{W}$ on the quality of the optimized bounds. We consider the computation of $f(A)\boldsymbol{b}$, where $f(z) = \log(1+z)/z$ or $f(z) = \sqrt{z}$, A is a 1000×1000 symmetric matrix with logspaced eigenvalues in $[10^{-2}, 10^{2}]$ and $\boldsymbol{b}$ is a random vector. We compare the a priori bound from Algorithm 1 with different cardinalities of the set $\mathcal{W}$, both with $\varphi_m(w)$ given by the linear system error bound from Theorem 2 (`a priori bound`, $|W| = 25, 50, 100$, `bounded error`) and with the exact linear system errors (`a priori bound`, $|W| = 25, 50, 100$, `exact error`). In all cases, the set $\mathcal{W}$ is chosen as a discretization of $[-10^5, -10^{-5}]$ with logspaced points. We also make a comparison with the first a posteriori bound form Corollary 3 using $w(z) = z$ and the exact linear system error norms (`a posteriori bound`, $w(z) = z$, `exact error`), which by Proposition 1 is the best bound that can be obtained using the approach that we consider in this work, but it requires to compute or bound the error norm of

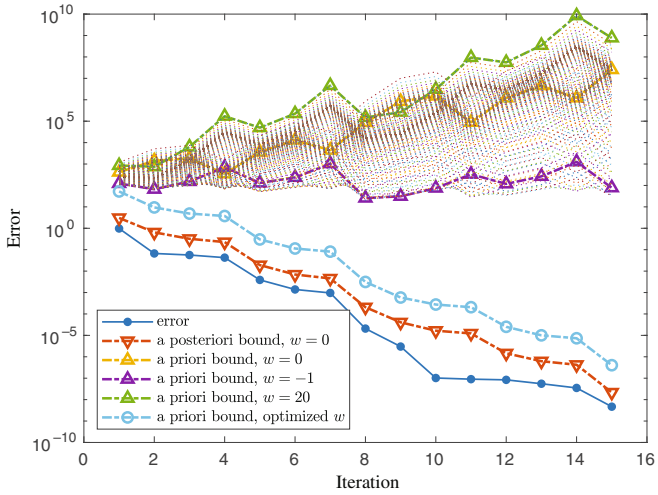

FIGURE 3 A priori and a posteriori error bounds for $A^{-1/2}\boldsymbol{b}$, where A is a symmetric 1000×1000 matrix with logspaced eigenvalues in $[10^{-1}, 10^{1}]$.

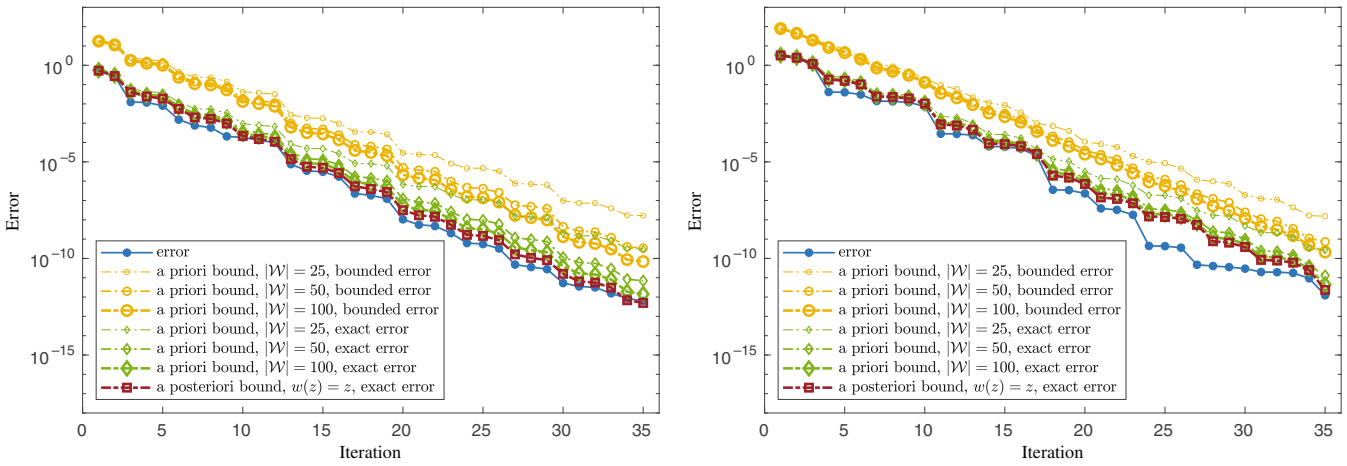

FIGURE 4 Comparison of optimized a priori bounds with different sets $\mathcal{W}$ for $f(A)\boldsymbol{b}$, where A is a 1000×1000 symmetric matrix with logspaced eigenvalues in $[10^{-2}, 10^{2}]$. Left: $f(z) = \log(1+z)/z$. Right: $f(z) = \sqrt{z}$.

a linear system for each point used in the discretization of the integral. The results are shown in Figure 4. For both functions, we see that the a priori bound with $|\mathcal{W}| = 100$ and exact error norms is very close to the a posteriori bound with $w(z) = z$. Both bounds are also quite close to the exact error. The bounds with $|\mathcal{W}| = 50$ are also only slightly worse than the ones with $|\mathcal{W}| = 100$.

6.2 | Bounds for matrix-vector products

We compare the following bounds for the error in the computation of $f(A)\boldsymbol{b}$:

- the residual-based a posteriori bound from Algorithm 2 with $w = 0$ (`a posteriori bound, fixed` w, `exact residual`),
- the error-based a priori bound from Algorithm 1 with optimized w and linear system error bounded with (19) and Theorem 2 (`a priori bound, optimized` w, `bounded error`),

- the error-based a posteriori bound from Corollary 3 with optimized w and exact linear system error (`a posteriori bound, optimized` w, `exact error`),
- the error-based a priori bound from Algorithm 1 with optimized w and exact linear system error (`a priori bound, optimized` w, `exact error`),
- the a priori bound from Corol. 4 of Reference 16 for Cauchy-Stieltjes functions (`a priori bound slope`, Corol. 4 of Reference 16),
- a simple error estimate based on the difference of consecutive approximations (`consecutive differences`), i.e., approximate $\|f(A)\boldsymbol{b} - \mathtt{f}_m\|_2 \approx \|\mathtt{f}_m - \mathtt{f}_{m+1}\|_2$.

We mention that the poles used in Corol. 4 of Reference 16 are not nested, so they are not convenient for the construction of a rational Krylov subspace. Instead, we use a nested sequence of poles based on an equidistributed sequence, as shown in Sec. 3.5 of Reference 16; this pole sequence is guaranteed to have the same asymptotic rate of convergence as predicted by Corol. 4 of Reference 16, but the actual bound on the error may not be satisfied. The error estimate based on consecutive differences requires running the method for an additional iteration to be computed, and although it is usually quite reliable it may underestimate the true error when convergence is slow or stagnating; this behavior might be undesirable in situations in which a guarantee on the error bound is required. Note that the third and especially the fourth bound cannot be used in a practical setting, but they can provide interesting theoretical insight. Recall that the residual-based a posteriori bound is independent of the choice of w, so we can simply take $w = 0$ (see Section 3.3), but the error-based a posteriori bound does not have this property and hence benefits from the optimization of w, similarly to the a priori bounds. For all optimized bounds, we use as $\mathcal{W}$ a discretization of $[-10^3\lambda_{\max}(A), -10^{-3}\lambda_{\min}(A)]$ with 100 logspaced points.

The results for the two functions $f(z) = \log(1+z)/z$ and $f(z) = \sqrt{z}$ are shown in Figure 5, using as A the symmetric positive definite matrix `Nasa/nasa2146` of size 2146×2146 from the SuiteSparse Matrix Collection[29] and as $\boldsymbol{b}$ a random vector. We notice that the a priori and a posteriori bounds that use the exact error are slightly better than the residual-based a posteriori bound, and they almost overlap with each other. By comparing the a priori bounds that use the exact linear system error with the ones that use the bound in Theorem 2, we can conclude that most of the overestimation in the a priori error bounds for $f(A)\boldsymbol{b}$ comes from the overestimation of the linear system error when combining (19) with Theorem 2. Indeed, if we had access to the exact linear system errors in advance, we would be able to obtain a priori the fourth bound in Figure 5, which captures the convergence behavior very well. Although this scenario is clearly not realistic, an interesting consequence of this observation is that any improvement to a priori bounds for the linear system error would automatically lead to an equal improvement in the a priori bounds for $f(A)\boldsymbol{b}$ of Corollary 2, with the limit case of an exact linear system error yielding a bound for $f(A)\boldsymbol{b}$ that is quite close to the true error. The a priori bound (Corol. 4 of Reference 16) correctly predicts the convergence rate of the rational Krylov approximation for $f(z) = \log(1+z)/z$ and

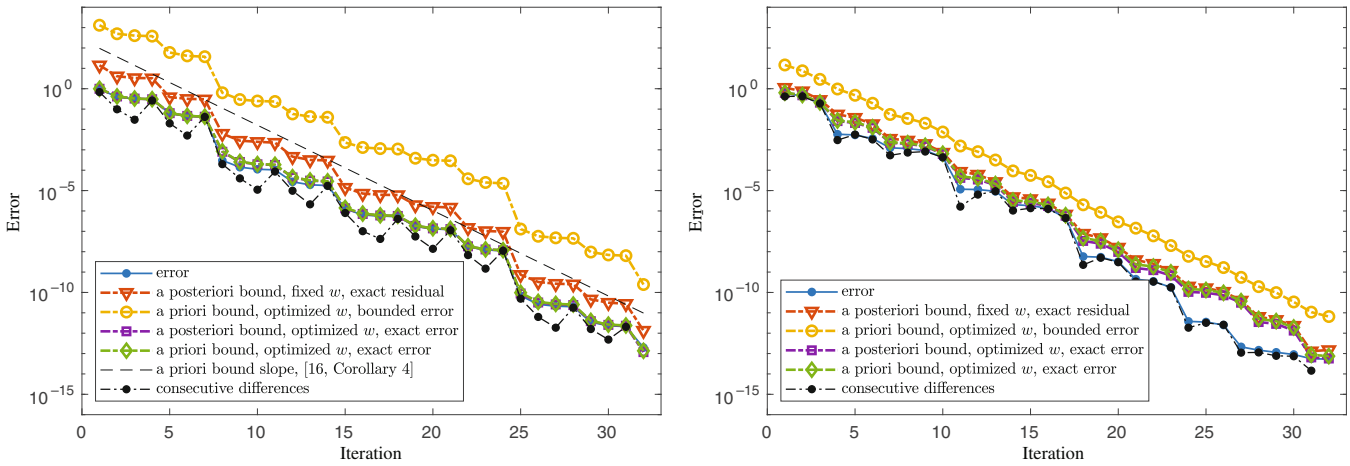

FIGURE 5 A priori and a posteriori bounds for $f(A)\boldsymbol{b}$, where A is the 2146×2146 symmetric matrix `Nasa/nasa2146`. Left: $f(z) = \log(1+z)/z$. Right: $f(z) = \sqrt{z}$.

it is more accurate than the a priori bound with bounded linear system error from Algorithm 1, although as we discussed above it might not be reliable as an upper bound because we are using the nested pole sequence from Sec. 3.5 of Reference 16. Note that this bound does not hold for the function $f(z) = \sqrt{z}$ since it is not Cauchy–Stieltjes, but it is still able to predict the convergence rate quite well, due to the fact that $f(z) = z \cdot z^{-1/2}$ and $g(z) = z^{-1/2}$ is Cauchy–Stieltjes. The error estimates based on consecutive differences are usually quite accurate, but they sometimes underestimate the error; this is especially noticeable for the function $f(z) = \log(1+z)/z$.

6.3 | Bounds for quadratic forms

We compare the following bounds for the error in the computation of the quadratic form $\boldsymbol{b}^T f(A)\boldsymbol{b}$:

- the a posteriori bound from Corollary 6 with $w = 0$, using the exact linear system residual (`a posteriori bound, fixed` w, `exact residual`),
- the first a priori bound from Corollary 5 with optimized w and linear system residual bounded using Theorem 2 (`a priori bound, optimized` w, `bounded residual`),
- the first a priori bound from Corollary 5 with optimized w and exact linear system residual (`a priori bound, optimized` w, `exact residual`),
- the second a priori bound from Corollary 5 with optimized w and exact linear system error and residual (`a priori bound, optimized` w, `exact residual + error`),
- a simple error estimate based on the difference of consecutive approximations (`consecutive differences`), that is, approximate $\|\boldsymbol{b}^T f(A)\boldsymbol{b} - \mathfrak{q}_m\|_2 \approx \|\mathfrak{q}_m - \mathfrak{q}_{m+1}\|_2$.

As in the previous experiment, for all optimized bounds we use as the set $\mathcal{W}$ a discretization of $[-10^3\lambda_{\max}(A), -10^{-3}\lambda_{\min}(A)]$ with 100 logspaced points. Recall that the second bound in Corollary 5 with $\|\mathtt{res}_m(w)\|_2$ bounded with Theorem 2 and $\|\mathtt{err}_m(w)\|_2$ bounded using (19) is always less accurate than the first one (see Remark 4), so we omit it from our experiments.

The results for the two functions $f(z) = \log(1+z)/z$ and $f(z) = \sqrt{z}$ are shown in Figure 6, where we also include the error for $f(A)\boldsymbol{b}$ for comparison. We can notice that the convergence for the quadratic form $\boldsymbol{b}^T f(A)\boldsymbol{b}$ is roughly twice as fast as the convergence for the matrix-vector product with $f(A)$. Similarly to the case of the bounds for $f(A)\boldsymbol{b}$, the a priori bound with exact linear system residual is significantly better than the bound that uses Theorem 2. Moreover, the a priori bound that uses both the exact residual and the exact error is even more accurate; this is in agreement with the observations in Remark 4, where we show that for the best possible choice of w, that is, $w(z) = z$, the bound in Corollary 5 that uses both the error and the residual is more accurate than the bound that only uses $\|\mathtt{res}_m(z)\|_2$. The error estimate based on consecutive differences is very accurate for $f(z) = \sqrt{z}$, but it occasionally underestimates the error for $f(z) = \log(1+z)/z$.

6.4 | Comparison with Lanczos setting

In this section we replicate the experiment in Fig. 4b of Reference 13 to show that the optimization of w can also lead to improvements for the error bounds in the polynomial Krylov (Lanczos) setting.

We consider the computation of $A^{1/2}\boldsymbol{b}$, where A is a 1000×1000 symmetric matrix with uniformly spaced eigenvalues in $[10^{-2}, 10^2]$, using a polynomial Krylov method. The a priori and a posteriori bounds used in Reference 13 can be obtained as a special case of the error-based bounds in Corollaries 2 and 3, by taking $q_{m-1}(z) \equiv 1$ and fixing $w = 0$. They are respectively denoted by `a priori bound, fixed` w, `exact error` and `a posteriori bound, fixed` w, `exact error` in Figure 7. We compare these bounds with the ones obtained by optimizing w as described in Section 3.2, which we label as `a priori bound, optimized` w, `exact error` and `a posteriori bound, optimized` w, `exact error`. In particular, the optimized a priori bound coincides with the one given in Algorithm 1 when all the poles $\xi_j = \infty$. Note that since we are using $\|\mathtt{err}_m(w)\|_2$ instead of $\|\mathtt{res}_m(w)\|_2$ in the a posteriori bound, the discussion in Section 3.3 does not apply and the bound can be improved by optimizing the choice of w. Similar to Reference 13, we assume that the error norms $\|\mathtt{err}_m(w)\|_2$ are known exactly. This assumption allows us to theoretically investigate the

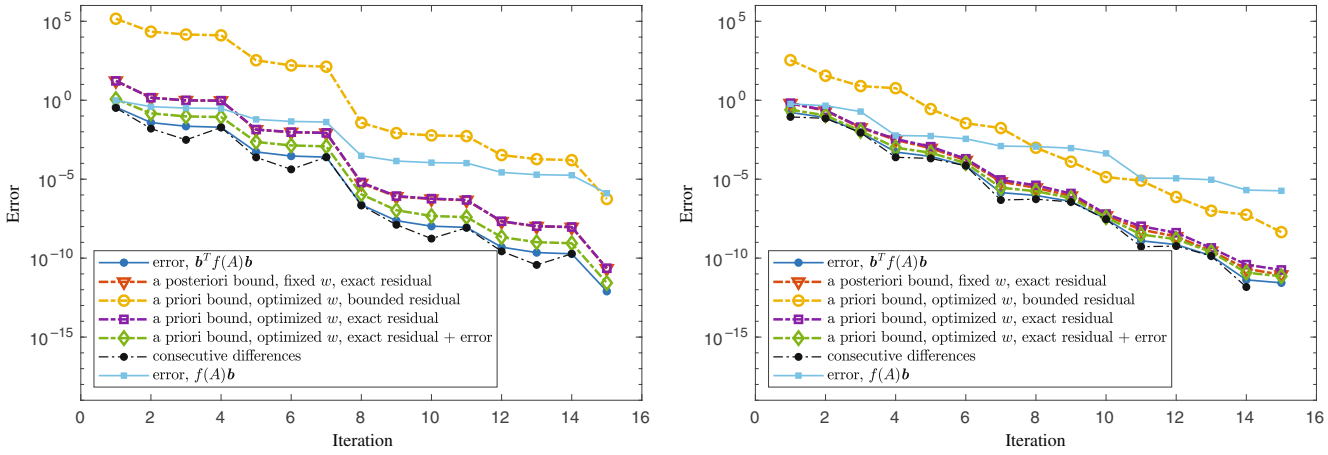

FIGURE 6 A priori and a posteriori bounds for $\boldsymbol{b}^T f(A)\boldsymbol{b}$, where A is the 2146×2146 symmetric matrix `Nasa/nasa2146`. Left: $f(z) = \log(1+z)/z$. Right: $f(z) = \sqrt{z}$.

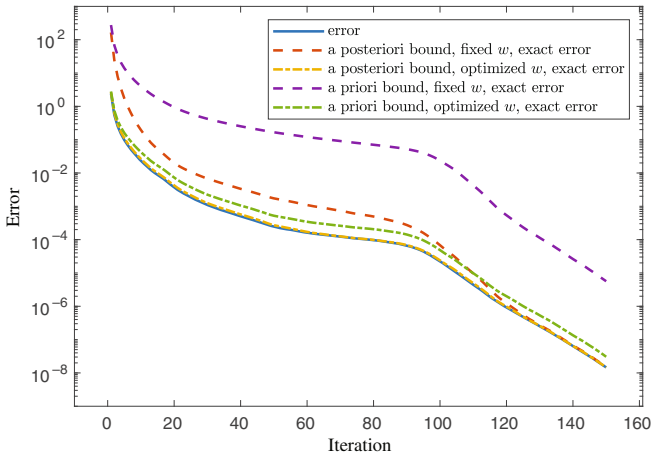

FIGURE 7 Comparison of bounds for $A^{1/2}\boldsymbol{b}$, where A is a 1000×1000 symmetric matrix with linearly spaced eigenvalues in $[10^{-2}, 10^{2}]$, with and without optimization of w.

error bounds by eliminating any dependence on possibly inaccurate bounds on the linear system error norms, although it is not realistic in a practical scenario.

The results of the comparison are shown in Figure 7, where the optimized bounds are computed using as set $\mathcal{W}$ a discretization of $[-10^5, -10^{-5}]$ with 25 logspaced points. We can see that optimizing w gives significant improvements, especially for the a priori bound; the optimized a posteriori bound is practically overlapping with the error curve. If the number of points in $\mathcal{W}$ is increased to 100, the accuracy of the optimized a priori bound increases, and it becomes essentially indistinguishable from the optimized a posteriori bound. It should be mentioned that the plots in Fig. 4 of Reference 13 consider the A-norm of the error, while here we consider the 2-norm, so the error curves shown in Figure 7 are slightly different from the ones in Reference 13.

7 | CONCLUSIONS

We have derived error bounds for the approximation of $f(A)\boldsymbol{b}$ and $\boldsymbol{b}^T f(A)\boldsymbol{b}$ with rational Krylov methods, by exploiting properties of rational Arnoldi decomposition and an integral representation of the function f. The bounds that we have obtained generalize some bounds for Lanczos-based matrix function approximation given in.[13] In the rational Krylov setting, the more complicated expression for the matrix function error poses an additional challenge compared to the

polynomial Krylov case; we were able to overcome this obstacle by numerically optimizing the parameter w that appears in the bounds. The same strategy can be also applied to the Lanczos setting, and it can lead to improvements in those bounds as well.

ACKNOWLEDGMENTS

The author is thankful to the two anonymous reviewers for their insightful comments. Open access publishing facilitated by Scuola Normale Superiore, as part of the Wiley - CRUI-CARE agreement.

CONFLICT OF INTEREST STATEMENT

The author declares no potential conflict of interests.

FUNDING INFORMATION

The author acknowledges financial support from INDAM (Italian Institute of High Mathematics) via the INDAM-GNCS project *Metodi basati su matrici e tensori strutturati per problemi di algebra lineare di grandi dimensioni*.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Igor Simunec https://orcid.org/0000-0002-6266-924X

REFERENCES

1. Higham NJ. Functions of matrices: theory and computation. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2008. https://doi.org/10.1137/1.9780898717778
2. Saad Y. Analysis of some Krylov subspace approximations to the matrix exponential operator. SIAM J Numer Anal. 1992;29(1):209–28. https://doi.org/10.1137/0729014
3. van den Eshof J, Frommer A, Lippert T, Schilling K, van der Vorst HA. Numerical methods for the QCDd overlap operator. I. Sign-function and error bounds. Comput Phys Commun. 2002;146(2):203–24.
4. Frommer A, Güttel S, Schweitzer M. Efficient and stable Arnoldi restarts for matrix functions based on quadrature. SIAM J Matrix Anal Appl. 2014;35(2):661–83. https://doi.org/10.1137/13093491X
5. Ruhe A. Rational Krylov algorithms for nonsymmetric eigenvalue problems. Recent advances in iterative methods. The IMA volumes in mathematics and its applications. Volume 60. New York: Springer; 1994. p. 149–64.
6. Druskin V, Knizhnerman L. Extended Krylov subspaces: approximation of the matrix square root and related functions. SIAM J Matrix Anal Appl. 1998;19(3):755–71. https://doi.org/10.1137/S0895479895292400
7. Moret I, Novati P. RD-rational approximations of the matrix exponential. BIT. 2004;44(3):595–615.
8. Güttel S, Knizhnerman L. A black-box rational Arnoldi variant for Cauchy-Stieltjes matrix functions. BIT. 2013;53(3):595–616. https://doi.org/10.1007/s10543-013-0420-x
9. Güttel S. Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. GAMM-Mitt. 2013;36(1):8–31.
10. Benzi M, Simunec I. Rational Krylov methods for fractional diffusion problems on graphs. BIT. 2022;62(2):357–85. https://doi.org/10.1007/s10543-021-00881-0
11. Frommer A, Simoncini V. Stopping criteria for rational matrix functions of Hermitian and symmetric matrices. SIAM. J Sci Comput. 2008;30(3):1387–412. https://doi.org/10.1137/070684598
12. Frommer A, Schweitzer M. Error bounds and estimates for Krylov subspace approximations of Stieltjes matrix functions. BIT. 2016;56(3):865–92. https://doi.org/10.1007/s10543-015-0596-3
13. Chen T, Greenbaum A, Musco C, Musco C. Error bounds for Lanczos-based matrix function approximation. SIAM J Matrix Anal Appl. 2022;43(2):787–811. https://doi.org/10.1137/21M1427784
14. Beckermann B, Reichel L. Error estimates and evaluation of matrix functions via the Faber transform. SIAM J Numer Anal. 2009;47(5):3849–83. https://doi.org/10.1137/080741744
15. Moret I, Novati P. Krylov subspace methods for functions of fractional differential operators. Math Comp. 2019;88(315):293–312. https://doi.org/10.1090/mcom/3332
16. Massei S, Robol L. Rational Krylov for Stieltjes matrix functions: convergence and pole selection. BIT. 2021;61(1):237–73. https://doi.org/10.1007/s10543-020-00826-z
17. Druskin V, Lieberman C, Zaslavsky M. On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. SIAM J Sci Comput. 2010;32(5):2485–96. https://doi.org/10.1137/090774082

18. Güttel S. Rational Krylov methods for operator functions, Dissertation. Freiberg: Technische Universität Bergakademie Freiberg. 2010. available as MIMS Eprint 2017.39. Available from http://eprints.ma.man.ac.uk/2586/
19. Berljafa M, Güttel S. Generalized rational Krylov decompositions with an application to rational approximation. SIAM J Matrix Anal Appl. 2015;36(2):894–916. https://doi.org/10.1137/140998081
20. Beckermann B, Bisch J, Luce R. On the rational approximation of Markov functions, with applications to the computation of Markov functions of Toeplitz matrices. arXiv preprint arXiv:2106.05098 2022 https://arxiv.org/pdf/2106.05098.pdf
21. Horn RA, Johnson CR. Topics in matrix analysis. Cambridge: Cambridge University Press; 1994 Corrected reprint of the 1991 original.
22. Berg C. Mateu J, Porcu E, editors. Stieltjes-pick-Bernstein-Schoenberg and their connection to complete monotonicity; 2008. p. 15–45.
23. Saad Y. Iterative methods for sparse linear systems. 2nd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2003. https://doi.org/10.1137/1.9780898718003
24. Paige CC, Parlett BN, van der Vorst HA. Approximate solutions and eigenvalue bounds from Krylov subspaces. Numer Linear Algebra Appl. 1995;2(2):115–33. https://doi.org/10.1002/nla.1680020205
25. Simoncini V. Restarted full orthogonalization method for shifted linear systems. BIT. 2003;43(2):459–66. https://doi.org/10.1023/A:1026000105893
26. Benzi M, Rinelli M, Simunec I. Computation of the von Neumann entropy of large matrices via trace estimators and rational Krylov methods. Numer Math. 2023;155(3-4):377–414. https://doi.org/10.1007/s00211-023-01368-6
27. Golub GH, Van Loan CF. Matrix computations. Johns Hopkins studies in the mathematical sciences. 4th ed. Baltimore, MD: Johns Hopkins University Press; 2013.
28. Beckermann B. An error analysis for rational Galerkin projection applied to the Sylvester equation. SIAM J Numer Anal. 2011;49(6):2430–50. https://doi.org/10.1137/110824590
29. Davis TA, Hu Y. The University of Florida sparse matrix collection. ACM Trans Math Softw. 2011;38(1):1–25. https://doi.org/10.1145/2049662.2049663

How to cite this article: Simunec I. Error bounds for the approximation of matrix functions with rational Krylov methods. Numer Linear Algebra Appl. 2024;31(5):e2571. https://doi.org/10.1002/nla.2571