The work of Angelo A. Casulli was partially supported by the GNCS/INdAM project “Metodi low-rank per problemi di algebra lineare con struttura data-sparse”.

1. Introduction

Generalized matrix functions (GMFs) are an extension of the notion of matrix functions based on the singular value decomposition (SVD) instead of the spectral decomposition. They were introduced for the first time in [15], with the purpose of extending the definition of matrix functions to rectangular matrices. Although the introduction of GMFs dates to the Seventies [15], they have become a more active area of research only in recent years. For instance, theoretical aspects of generalized matrix functions have been investigated in [1, 5, 21], while efficient numerical methods have been developed in [3, 4]. For applications of generalized matrix functions, we direct the reader to [1, 3, 4] and the references therein.

In many applications that involve standard matrix functions, it is only required to compute matrix-vector products of the form \( f(A)b \), where the matrix \( A \) is usually large and sparse. In this case, the (expensive) computation of the whole matrix \( f(A) \) can be bypassed by using methods that directly approximate the product \( f(A)b \), such as Krylov methods. These methods only require matrix-vector products and possibly the solution of shifted linear systems with the matrix \( A \).

A similar situation arises when GMFs are involved: indeed, it is often required to compute the action of a generalized matrix function on a vector \( f \) [3, 4], and hence it is preferable to use a method that avoids the computation of the whole GMF by means of an SVD.

This problem was recently investigated in [3, 4], using methods based on the Golub-Kahan bidiagonalization in [3], and Chebyshev polynomial interpolation in [4].
In this paper, we propose a generalization of the method proposed in [3], using the interpretation of the Golub-Kahan bidiagonalization in terms of Krylov subspaces. Performing $k$ steps of the Golub-Kahan bidiagonalization of a matrix $A$ with starting vector $b$ is equivalent to the simultaneous computation of orthonormal bases of the polynomial Krylov subspaces $P_k(A^T A, b)$ and $P_k(A A^T, A b)$. By replacing the polynomial Krylov subspaces with their rational counterparts, we obtain a rational Krylov method for the computation of the action of a GMF on a vector.

As can be expected by analogy with standard matrix functions, in the case of non-analytic functions and functions of low regularity these rational methods have a faster convergence than the method based on the Golub-Kahan bidiagonalization. However, their increased effectiveness comes at the cost of having to solve a linear system at each iteration, while the methods discussed in [3, 4] only require matrix-vector products.

The Golub-Kahan bidiagonalization of a matrix can be computed with a short recurrence, which relies on the fact that the projected matrix is bidiagonal. This structure is unfortunately not preserved in the rational case that we consider here. However, we are still able to construct a short recurrence to update the rational Krylov bases and the projected matrix, using the fact that the projected matrix is a quasiseparable matrix.

We also prove error bounds that link the error of the method from [3] based on the Golub-Kahan bidiagonalization and the rational methods introduced in this paper with, respectively, the error of uniform polynomial and rational approximation of the function $f$. These bounds are a direct generalization of the bounds for standard matrix functions, and they can be proved with the same techniques. Although the connection of GMFs to standard matrix functions is well-known, to the best of our knowledge these error bounds for the approximation of GMFs have never appeared in previous literature.

The paper is organized as follows. In Section 2 we introduce the notation used throughout the paper. In Section 3 we recall the definition of standard matrix functions and GMFs and we present some of their properties. In Section 4 we briefly introduce the class of rational Krylov methods for standard matrix functions. The use of polynomial and rational Krylov methods in the context of generalized matrix functions is discussed in Section 5. Section 6 is dedicated to the proof of the error bounds and related discussion. Some numerical experiments to compare the different methods and illustrate the error bounds are presented in Section 7, and Section 8 contains concluding remarks.

2. Notation

We denote by $\mathbb{R}^{m \times n}$ the space of $m \times n$ real matrices. We use bold letters for vectors, e.g. $v \in \mathbb{R}^n$. The entries of vector $v$ are given by $v_1, \ldots, v_n$, and the entries of matrix $A \in \mathbb{R}^{m \times n}$ are $a_{ij}$. We also use a MATLAB-like notation: $\text{diag}(d_1, \ldots, d_n)$ represents an $n \times n$ diagonal matrix with entries $d_1, \ldots, d_n$ on the diagonal; for $i \leq j$ and $h \leq k$, we denote by $A(i : j, h : k)$ the submatrix of $A$ corresponding to row indices from $i$ to $j$ and column indices from $h$ to $k$.

We denote by $\text{triu}(A)$ the upper triangular part of matrix $A$, and more generally by $\text{triu}(A, k)$ the matrix with all zeroes below the $k$-th diagonal whose other entries coincide with those of $A$. Diagonals above the main diagonal are represented with
a positive index, so that $\text{triu}(A, 1)$ indicates the strictly upper triangular part of $A$. Similarly, $\text{tril}(A)$ and $\text{tril}(A, -1)$ denote the lower triangular and strictly lower triangular part of $A$, respectively. We use the same notation also for rectangular matrices. We denote by $A^+$ the Moore-Penrose pseudoinverse of a matrix $A$.

3. Matrix functions

The goal of this section is to define generalized matrix functions (GMFs) and introduce their main properties. We begin by recalling some basic concepts about standard matrix functions, and then we introduce GMFs and some of their properties.

3.1. Standard matrix functions. The concept of matrix function is a natural way to generalize the evaluation of a function on a square matrix. For simplicity, we treat only the case of diagonalizable matrices. The general definitions and a thorough description of matrix functions can be found in the monograph [16].

Let $A$ be an $n \times n$ matrix. Assume that $A$ is diagonalizable, i.e. $A = VDV^{-1}$, where $D = \text{diag}(d_1, \ldots, d_n)$. Given a function $f$ defined on the set $\{d_1, \ldots, d_n\}$ the matrix function of $f$ applied on $A$ is defined as

$$f(A) = Vf(D)V^{-1},$$

where $f(D) = \text{diag}(f(d_1), \ldots, f(d_n))$.

Equivalently, if $p(x) = \sum_{i=0}^{n-1} p_i x^i$ is a polynomial that interpolates $f$ in $d_1, \ldots, d_n$, the matrix function can be defined as

$$f(A) = p(A) = \sum_{i=0}^{n-1} p_i A^i.$$

The computation of a matrix function using the first of the two definitions requires knowledge of the eigenvalues of $A$. However, if $n$ is large, finding the eigenvalues of $A$ can be unfeasible.

A widely used technique to obtain a good approximation of $f(A)$ is to find a low-degree polynomial $q$ that approximates the interpolant polynomial $p$, and to approximate $f(A)$ with $q(A)$. Polynomial Krylov methods, presented in Section 4, use a similar strategy to approximate $f(A)b$, by using $A$ and $b$ to construct the polynomial $q$.

3.2. Generalized matrix functions. Generalized matrix function were first introduced in [15], with the purpose of extending the definition of matrix functions to rectangular matrices. They are defined in a similar way with respect to the standard matrix functions, but the singular value decomposition is used instead of the diagonalization.

Let $A \in \mathbb{R}^{m \times n}$ and let $A = U\Sigma V^T$ be its SVD, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m \times n}$ is defined as

$$\Sigma_{i,j} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 & \text{otherwise}, \end{cases}$$

where $r \leq \min\{m, n\}$ is the rank of $A$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the nonzero singular values of $A$. 
Given a function $f$ defined on the set $\{\sigma_1, \ldots, \sigma_r\}$ the generalized matrix function of $f$ applied on $A$ is defined as
\[ f^o(A) = U f^o(\Sigma)V^T, \]
where
\[ f^o(\Sigma)_{i,j} = \begin{cases} f(\sigma_i) & \text{if } i = j \leq r \\ 0 & \text{otherwise}. \end{cases} \]

Observe that a GMF can be expressed in terms of the compact SVD of the matrix $A$, that is $A = U_s \Sigma V_s^T$, where $U_s \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ have orthonormal columns, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$. In such case, we have
\[ f^o(A) = U_s f(\Sigma_r)V_s^T. \]

Since the definition of a GMF only depends on the values of $f$ on the nonzero singular values of $A$, we can always assume that $f$ is an odd function, and in particular that $f(0) = 0$.

**Remark 3.1.** [4, Theorem 2.1] If $p$ is a polynomial that interpolates $f$ in $\sigma_1, \ldots, \sigma_r$ it holds that $f^o(A) = p^o(A)$. Moreover, since $\sigma_i > 0$ for $i = 1, \ldots, r$, we can always take $p$ as an odd polynomial, i.e. $p(z) = q(z^2)z$ for some polynomial $q$.

Next, we list some properties of GMFs that will be required in the following sections. A discussion of additional properties of generalized matrix functions can be found in [4].

**Lemma 3.2.** Let $S \in \mathbb{R}^{n \times n}$ be a symmetric matrix and let $f$ be defined on the singular values of $S$. If $S$ is positive definite, or $S$ is positive semidefinite and $f(0) = 0$, it holds
\[ f^o(S) = f(S). \]

**Proposition 3.3.** Let $p$ be an odd polynomial, i.e. we can write $p(z) = q(z^2)z$ for some polynomial $q$. Then, for any matrix $A \in \mathbb{R}^{m \times n}$ it holds
\[ p^o(A) = q(AA^T)A = Aq(A^T A). \]

For the proof of Proposition 3.3 we refer to [4, Section 2.2].

**Corollary 3.4.** Let $r$ be a rational function with an odd numerator and an even denominator, i.e., $r(z) = \frac{p(z^2)}{q(z^2)}z$ for some polynomials $p$ and $q$, such that the singular values of $A \in \mathbb{R}^{m \times n}$ and 0 are not roots of $q$. Then it holds
\[ r^o(A) = q(AA^T)^{-1}p(AA^T)A = Aq(A^T A)^{-1}p(A^T A). \]

**Proof.** The proof is practically the same as the proof of Proposition 3.3. We report it for completeness. Let $A = U\Sigma V^T$ be an SVD of $A$. Then we have $AA^T = U\Sigma^2 U^T$, and more generally $s(AA^T) = Us(\Sigma^2)U^T$ for any polynomial $s$.

Since $0$ is not a root of $q$, we have $r(0) = 0$ and hence by Lemma 3.2 we have
\[ r^o(A) = Ur^o(\Sigma)V^T = Ur(\Sigma)V^T = Uq(\Sigma^2)^{-1}p(\Sigma^2)\Sigma V^T. \]

Note that $q(\Sigma^2)^{-1}$ is well defined because the singular values of $A$ are not roots of $q$. Now, multiplying by $U^TU = I$ we obtain
\[ r^o(A) = Uq(\Sigma^2)^{-1}U^T U p(\Sigma^2)U^T U \Sigma V^T = q(AA^T)^{-1}p(AA^T)A. \]

The second identity can be obtained in a similar way. □
The following proposition gives a formulation of the above results for general functions. See also [3, Theorem 10] for an alternative formulation of the same identity.

**Proposition 3.5.** Let $A \in \mathbb{R}^{m \times n}$ and let $f$ be a function defined on the nonzero singular values of $A$. Defining $g(z) = \frac{f(\sqrt{z})}{\sqrt{z}}$, for $z \neq 0$, it holds

$$f^\circ(A) = g^\circ(AA^T)A = Ag^\circ(A^TA).$$

Moreover, if $\lim_{z \to 0} \frac{f(z)}{z} = 0$, we can define $g(0) = 0$. Then $g(AA^T)$ and $g(A^TA)$ are well defined, and we have

$$f^\circ(A) = g(AA^T)A = Ag(A^TA).$$

**Proof.** Let $p(z) = q(z^2)z$ be an odd polynomial that interpolates $f$ in the nonzero singular values of $A$. From Proposition 3.3 and Remark 3.1 we have

$$f^\circ(A) = p^\circ(A) = q(AA^T)A = Ag(A^TA).$$

Since $p$ interpolates $f$ in the nonzero singular values of $A$, $q$ interpolates $g$ in the squares of the nonzero singular values of $A$, which are the nonzero singular values of $AA^T$ and $A^TA$. Hence $q^\circ(AA^T) = g^\circ(AA^T)$ and $q^\circ(A^TA) = g^\circ(A^TA)$.

If $g(0) = 0$, by Lemma 3.2 we also have $g^\circ(A^TA) = g(A^TA)$ and $g^\circ(AA^T) = g(AA^T)$. \hfill $\square$

**Remark 3.6.** If $AA^T$ is positive definite, by Lemma 3.2 it holds $f^\circ(A) = g(AA^T)A$ without the assumption $\lim_{z \to 0} \frac{f(z)}{z}$, and similarly if $A^TA$ is positive definite we have $f^\circ(A) = Ag(A^TA)$. Note that we could also artificially define $g(0) = 0$ without any assumptions on $f$, since the definition of a GMF only depends on the nonzero singular values of the matrix. However, this would cause $g$ to be discontinuous at 0 and it would not be very useful in practice.

The following proposition links $f^\circ(A)$ with $f^\circ(A^T)$, which will be useful when $A$ is rectangular. We recall that $A^+$ denotes the Moore-Penrose pseudoinverse of $A$, which can be defined as $A^+ = h^\circ(A^T) = h^\circ(A)^T$ for $h(z) = z^{-1}$. The more general statement in Proposition 3.7 can be seen as a generalization of [3, Proposition 7(iv)] and [15, Theorem 4(d)].

**Proposition 3.7.** Let $A \in \mathbb{R}^{m \times n}$ and let $f$ be a function defined on the nonzero singular values of $A$. Then it holds

$$f^\circ(A) = (A^+)^T f^\circ(A^T)A.$$

More generally, assume that $f(z) = g(z)h(z)k(z)$, where $g, h, k$ are functions defined on the nonzero singular values of $A$. Then it holds

$$f^\circ(A) = g^\circ(A)h^\circ(A^T)k^\circ(A).$$

**Proof.** We directly prove the generalized version, since the first statement simply follows by taking $g(z) = z^{-1}$, $h(z) = f(z)$ and $k(z) = z$.

Let $A$ have the singular value decomposition $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m \times n}$ has the singular values of $A$ on the main
diagonal. We have:
\[
g^\circ(A) h^\circ(A^T) k^\circ(A) = U g^\circ(\Sigma) V^T V h^\circ(\Sigma^T) U^T U k^\circ(\Sigma) V^T
\]
\[
= U g^\circ(\Sigma) h^\circ(\Sigma^T) k^\circ(\Sigma) V^T
\]
\[
= U f^\circ(\Sigma) V^T = f^\circ(A),
\]
where we used that \(g^\circ(\Sigma) h^\circ(\Sigma^T) k^\circ(\Sigma) = f^\circ(\Sigma)\), which can be verified directly. \(\square\)

Remark 3.8. The same proof of Proposition \(\text{[57]}\) can be used to show that, if \(f(z)g(z) = h(z)k(z)\), then it holds
\[
f^\circ(A^T)g^\circ(A) = h^\circ(A^T)k^\circ(A).
\]
In particular we have \(A^T f^\circ(A) = f^\circ(A^T)A\).

4. Rational Krylov methods

The class of Krylov methods provides an efficient way to compute approximations to expressions of the form \(f(A)b\). The main idea behind these methods is to construct a low dimensional subspace \(S_k \subset \mathbb{R}^n\) for some integer \(k \ll n\) using information from \(A\) and \(b\), and then to approximate \(f(A)b\) with an appropriate vector from \(S_k\).

A popular choice for the approximation subspace \(S_k\) is the polynomial Krylov subspace
\[
P_k(A, b) = \text{span}\{b, Ab, \ldots, A^{k-1}b\} = \{p_{k-1}(A)b : p \in \Pi_{k-1}\},
\]
where \(\Pi_{k-1}\) denotes the set of polynomials of degree \(\leq k - 1\).

More generally, using a sequence of poles \(\{\xi_k\}_{k \geq 1} \subseteq (\mathbb{C} \cup \{\infty\}) \setminus \sigma(A)\), one can define the rational Krylov subspace
\[
Q_k(A, b) = q_{k-1}(A)^{-1} P_k(A, b) = \left\{ r_{k-1}(A)b : r_{k-1}(z) = \frac{p_{k-1}(z)}{q_{k-1}(z)}, \text{with } p_{k-1} \in \Pi_{k-1} \right\},
\]
where \(q_{k-1}(z) = \prod_{j=1}^{k-1} (1 - z/\xi_j)\). In the case when all poles are located at \(\infty\), we have \(q_{k-1}(z) \equiv 1\) and hence we recover the polynomial Krylov subspace \(P_k(A, b)\).

It is easy to verify that the Krylov subspaces \(Q_k(A, b)\) form a nested sequence, and that \(Q_k(A, b) = k\) as long as \(k\) is smaller than the invariance index \(K\) of the sequence, i.e. the smallest integer such that \(Q_K(A, b) = Q_{K+1}(A, b)\) (or, equivalently, \(P_K(A, b) = P_{K+1}(A, b)\)).

For \(k \leq K\), an orthonormal basis \(\{v_1, \ldots, v_k\}\) of \(Q_k(A, b)\) can be computed with the rational Arnoldi algorithm, introduced by Ruhe in \([24]\). In the basic algorithm, the first basis vector is chosen as \(v_1 = b/\|b\|_2\). Then, given a set of vectors \(\{v_1, \ldots, v_j\}\) which form an orthonormal basis of \(Q_j(A, b)\), the next basis vector \(v_{j+1}\) is computed by orthonormalizing the vector \((I - \frac{1}{\xi_{j+1}} A)^{-1} A v_j\) against the previously computed basis vectors. To prevent the algorithm from failing, it is required that \((I - \frac{1}{\xi_{j+1}} A)^{-1} A v_j \in Q_{k+1}(A, b) \setminus Q_k(A, b)\); this property is almost always satisfied in practice, however there are no theoretical guarantees that it holds. An approach for finding a vector \(w_j\) that guarantees that \((I - A)^{-1} w_j \in Q_{k+1}(A, b) \setminus Q_k(A, b)\) and is (near-)optimal in some sense was recently discussed in \([8]\), using the notion of continuation pairs \((\eta_j, \rho_j, t_j)\): in general such a vector \(w_j\) is of the form \((\rho_m A - \eta_m) V_j t_j\), where \(V_j = [v_1 \ldots v_j] \in \mathbb{C}^{n \times j}\).
Using the matrix with orthonormal columns \( V_k = [v_1 \ldots v_k] \in \mathbb{C}^{n \times k} \), we can compute the following approximation to \( f(A)b \) from the subspace \( Q_k(A,b) \):

\[
\bar{y}_k = V_k f(V_k^* A V_k) V_k^* b = V_k f(A_k) e_1,
\]

where \( A_k = V_k^* A V_k \) is the projection of \( A \) on the subspace \( Q_k(A,b) \), and \( e_1 \) denotes the first vector of the canonical basis of \( \mathbb{C}^n \). The accuracy of the approximation \( \bar{y}_k \) is largely dependent on the pole sequence \( \{\xi_k\}_{k \geq 1} \). The problem of choosing a sequence of poles that is effective for a particular function \( f \) and a given set containing the spectrum of \( A \) has often been discussed in the literature: we refer, for instance, to [14] and the references therein.

Some specific sequences of poles lead to special cases of the rational Krylov subspaces \( Q_k(A,b) \): if all the poles are equal to \( \xi \in \mathbb{C} \setminus \sigma(A) \), then \( Q_k(A,b) \) is a Shift-and-Invert Krylov subspace,

\[
Q_k(A,b) = \mathcal{P}_k \left( \left( I - \frac{1}{\xi} A \right)^{-1}, b \right),
\]

that was first investigated for the computation of matrix functions in [15, 28]. If the poles alternate between 0 and \( \infty \), we obtain the extended Krylov subspace, introduced in [10], which is of the form

\[
Q_{2k}(A,b) = A^{-k} \mathcal{P}_{2k}(A,b) = \mathcal{P}_{2k}(A, A^{-k} b).
\]

We refer to [13] for an extensive discussion on rational Krylov methods for the computation of matrix functions.

5. Krylov methods for GMFs

The computation of \( f^\diamond(A)b \) by using an SVD of \( A \) can be unfeasible if the size of \( A \) is large. A possible way to approximate the product of a generalized matrix function times a vector is to use two rectangular matrices with orthonormal columns to project the matrix \( A \) onto a smaller space and then to compute the generalized matrix function of the projected matrix: let \( k \ll n \) and let \( U_k, V_k \in \mathbb{R}^{n \times k} \) with orthonormal columns, \( B_k \in \mathbb{R}^{k \times k} \) such that \( A \approx U_k B_k V_k^T \), then

\[
f^\diamond(A)b \approx U_k f^\diamond(B_k) V_k^T b,
\]

where the matrix \( f^\diamond(B_k) \) can be computed by means of the singular value decomposition.

In this section we describe how to compute such a projected matrix using the Golub-Kahan bidiagonalization, which is equivalent to using a polynomial Krylov method on the matrices \( A^T A \) and \( AA^T \). This strategy corresponds to the “third approach” discussed in [3, Section 5.4]; the numerical results of [3] indicate that (5.1) is often more effective than the other approaches they propose, which are based on Gauss and Gauss-Radau quadrature formulas.

In Sections 5.2 and 5.3 we generalize this approach to the rational Krylov case and we show that a short recurrence like the one of the Golub-Kahan bidiagonalization can be obtained in the rational case too.

5.1. Golub-Kahan bidiagonalization. The first method we describe for the computation of a truncated SVD is the Golub-Kahan bidiagonalization introduced for the first time in 1965.
\textbf{Theorem 5.1.} Let $A \in \mathbb{R}^{m \times n}$, with $m > n$. There exist orthogonal matrices $P \in \mathbb{R}^{m \times m}, Q \in \mathbb{R}^{n \times n}$ such that

\begin{equation}
\begin{bmatrix}
\alpha_1 & \beta_1 & \ldots & \ldots & 0 \\
0 & \alpha_2 & \beta_2 & \ldots & \\
\vdots & \vdots & \ddots & \ddots & \\
0 & \ldots & \ldots & 0 & \alpha_n \\
0 & \ldots & \ldots & \ldots & 0 \\
\end{bmatrix}
= \begin{bmatrix}
P^T A Q
\end{bmatrix}
\end{equation}

Moreover the first column of $Q$ can be chosen arbitrarily.

The proof of Theorem 5.1 is constructive and it is usually called Householder bidiagonalization process. It can be found in [12, Section 5.4].

In the case of large matrices the full bidiagonalization is too expensive. The goal of the Golub-Kahan bidiagonalization is to extract good approximations of singular values and singular vectors before the full bidiagonalization is completed.

Let

\begin{equation}
P = [p_1 \ldots p_m], \quad Q = [q_1 \ldots q_n]
\end{equation}

be a column partitioning of $P$ and $Q$. From Theorem 5.1 it follows that $AQ = PB$ and $A^T P = QB^T$; using these relations we have

\begin{align}
AQ_k &= \alpha_k p_k + \beta_{k-1} p_{k-1}, \\
A^T p_k &= \alpha_k q_k + \beta_k q_{k+1},
\end{align}

for $k = 1, \ldots, n$, with the convention that $\beta_0 p_0 = 0$ and $\beta_n q_{n+1} = 0$.

Let $r_k = AQ_k - \beta_{k-1} p_{k-1}$. Using the orthogonality of the columns of $P$, we have

\begin{equation}
\alpha_k = \pm \|r_k\|_2, \quad p_k = \frac{r_k}{\alpha_k} \quad \text{(if } \alpha_k \neq 0)\end{equation}

Similarly defining $s_k = A^T p_k - \alpha_k q_k$ we have

\begin{equation}
\beta_k = \pm \|s_k\|_2, \quad q_{k+1} = \frac{s_k}{\beta_k} \quad \text{(if } \beta_k \neq 0)\end{equation}

Hence, given $p_{k-1}, q_k, \beta_{k-1}$, we can compute $p_k, q_{k+1}, \beta_k$ and $\alpha_k$.

Defining $P_k = [p_1 \ldots p_k], Q_k = [q_1 \ldots q_k]$ and

\begin{equation}
B_k = \begin{bmatrix}
\alpha_1 & \beta_1 & 0 & \ldots & 0 \\
0 & \alpha_2 & \beta_2 & \ldots & \\
\vdots & \vdots & \ddots & \ddots & \\
0 & \ldots & \ldots & 0 & \alpha_n \\
0 & \ldots & \ldots & \ldots & 0 \\
\end{bmatrix},
\end{equation}

after the $k$-th step of (5.4) and (5.5) we have

\begin{align}
AQ_k &= P_k B_k, \\
A^T P_k &= Q_k B_k^T + \alpha_k e_k^T,
\end{align}
assuming $\alpha_k > 0$. It can be shown that
\begin{align}
\text{span}\{q_1, \ldots, q_k\} &= \mathcal{P}_k(A^T A, q_1), \\
\text{span}\{p_1, \ldots, p_k\} &= \mathcal{P}_k(A A^T, A q_1),
\end{align}
(5.7)

thus the convergence of the Golub-Kahan bidiagonalization follows from the convergence of the Lanczos method applied on $A^T A$ and $AA^T$.

For a given $k$, we can approximate $A$ with $P_k B_k Q_k^T$, and hence we can compute an approximation of the SVD of $A$ by computing the SVD of $B_k$. For further information on the Golub-Kahan bidiagonalization we refer to [12, Chapter 10].

The vector $y = f^\circ(A) b$ can be approximated with the expression
\begin{equation}
\tilde{y}_k = P_k f^\circ(B_k) Q_k^T b = \|b\|_2 P_k f^\circ(B_k) e_1 \in \mathcal{P}_k(A A^T, Ab).
\end{equation}
(5.8)

We refer to this approximation as a polynomial Krylov method for GMFs.

5.2. **Rational Krylov methods for GMFs.** As we saw in the previous section, the Golub-Kahan bidiagonalization computes orthonormal bases for the polynomial Krylov subspaces $\mathcal{P}_k(A^T A, b)$ and $\mathcal{P}_k(A A^T, Ab)$. By analogy with that approach, in this section we propose to compute an approximation to $f^\circ(A) b$ using the rational Krylov subspaces
\begin{equation}
Q_k(A^T A, b) \quad \text{and} \quad Q_k(A A^T, Ab),
\end{equation}
(5.9)

where $q_{k-1}(z) = \prod_{j=1}^{k-1} (1 - z/\xi_j)$, for a given pole sequence $\{\xi_j\}_{j \geq 1}$.

Assume that we have constructed two matrices with orthonormal columns $P_k$ and $Q_k$, such that $\text{span}(P_k) = Q_k(A A^T, Ab)$ and $\text{span}(Q_k) = Q_k(A^T A, b)$. Then, defining $B_k = P_k^T A Q_k$, by analogy with the polynomial Krylov approach we can introduce the vector
\begin{equation}
\tilde{y}_k = P_k f^\circ(B_k) Q_k^T b = P_k f^\circ(B_k) e_1,
\end{equation}
(5.10)

which is an approximation to $f^\circ(A) b$ from the subspace $Q_k(A A^T, Ab)$.

First of all, notice that it is sufficient to compute only one of the two rational Krylov subspaces: indeed, since it holds $A Q_k(A^T A, b) = Q_k(A A^T, Ab)$, we can compute an orthonormal basis of the subspace $\mathcal{Q}_k(A A^T, ab)$ simply by orthonormalizing the columns of $A Q_k$. This is equivalent to computing a QR decomposition $A Q_k = W_k R_k$, where $W_k$ has orthonormal columns and $R_k$ is upper triangular, so we can set $P_k = W_k$. Moreover, notice that we also have
\begin{equation*}
B_k = P_k^T A Q_k = W_k^T W_k R_k = R_k,
\end{equation*}
i.e. with the QR decomposition we also recover the matrix $B_k$, without the need to project $A$ explicitly. The basis $Q_k$ of the subspace $Q_k(A^T A, b)$ can be computed by applying the rational Arnoldi algorithm to the matrix $A^T A$ with initial vector $b$.

5.3. **Short recurrence for rational Golub-Kahan algorithm.** In the Golub-Kahan bidiagonalization (without reorthogonalization), we can compute the last column of $P_k$, $Q_k$ and $B_k$ just by knowing a few previous columns of $P_k$ and $Q_k$, by means of the equations (5.3). This short recurrence is possible because of the bidiagonal structure of the matrix $B_k$, that unfortunately is not preserved when we perform a rational Krylov method.
In this section we show that, if a rational Krylov method is used, the matrix $B_k$ is a quasiseparable matrix (see Definition 5.3). This structure extends the bidiagonal form that is obtained during the Golub-Kahan bidiagonalization. Using such structure we build a short recurrence that allows us to update the matrix $P_k$ and the matrix $B_k$ avoiding the full orthogonalization related to the computation of the QR factorization of the matrix $AQ_k$.

**Definition 5.2.** A matrix $S \in \mathbb{R}^{n \times n}$ is called a semiseparable matrix if all submatrices taken out of the lower and upper triangular part of the matrix have rank at most 1, that is
\[
\text{rank } S(i : n, 1 : i) \leq 1 \quad \text{and} \quad \text{rank } S(1 : i, i : n) \leq 1,
\]
for every $i = 1, \ldots, n$. Moreover, $S$ is called a generator representable semiseparable matrix if the lower and upper triangular parts of the matrix are derived from a rank 1 matrix, that is
\[
\text{tril}(S) = \text{tril}(uv^T) \quad \text{and} \quad \text{triu}(S) = \text{triu}(pq^T),
\]
for $u, v, p, q \in \mathbb{R}^n$.

**Definition 5.3.** A matrix $S$ is called a quasiseparable matrix if all the subblocks taken out of the strictly lower triangular part of the matrix (respectively, the strictly upper triangular part) are of rank at most 1, that is
\[
\text{rank } S(i + 1 : n, 1 : i) \leq 1 \quad \text{and} \quad \text{rank } S(1 : i, i + 1 : n) \leq 1,
\]
for every $i = 1, \ldots, n$.

The following theorem from [29, Section 1.5.2] gives us a complete characterization of invertible semiseparable matrices.

**Theorem 5.4.** The inverse of an invertible tridiagonal matrix is a semiseparable matrix, and vice versa. Moreover, the inverse of an invertible irreducible tridiagonal matrix is a generator representable semiseparable matrix and vice versa.

As it has been proved in [13, Section 5.2], if we perform a rational Krylov algorithm on a symmetric matrix we obtain a particular equivalence called rational Arnoldi decomposition. See also [7, Section 2] and [23, eq. (2.2)].

**Theorem 5.5.** Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$, let $Q_{k+1}(A, b)$ be the rational Krylov space with nonzero poles $\{\xi_1, \ldots, \xi_k\}$ and assume that $k$ is less than the invariance index of the Krylov subspace. Let $Q_{k+1} \in \mathbb{R}^{n \times (k+1)}$ be the matrix with orthonormal columns generated by the Arnoldi algorithm, such that $\text{span}(Q_{k+1}) = Q_{k+1}(A, b)$. Then the following relation holds:
\[
AQ_{k+1}K_k = Q_{k+1}H_k, \quad \text{with} \quad K_k = L_k + D_kH_k,
\]
where $H_k \in \mathbb{R}^{(k+1) \times k}$ is a full rank tridiagonal irreducible symmetric matrix, $L_k$ is the $(k + 1) \times k$ identity matrix and $D_k = \text{diag}(0, \frac{1}{\xi_1}, \ldots, \frac{1}{\xi_{k-1}})$, where $\frac{1}{\infty} = 0$.

Starting from (5.11), we are going to prove that the projection of the symmetric matrix $A$ on the Krylov subspace (i.e., $Q_{k+1}^T AQ_{k+1}$) is the sum of a diagonal matrix and a semiseparable matrix.
**Theorem 5.6.** Let \( A \in \mathbb{R}^{n \times n} \) be a symmetric matrix and let \( Q_{k+1} \in \mathbb{R}^{n \times (k+1)} \) be the matrix with orthonormal columns generated by the rational Arnoldi algorithm using poles \( \{\xi_1, \ldots, \xi_k\} \) different from zero and infinity. Assuming that \( k+1 \) is less than the invariance index, we have

\[
J_{k+1} := Q_{k+1}^T AQ_{k+1} = S + \hat{D}_k,
\]

where \( \hat{D}_k = \text{diag}(0, \xi_1, \ldots, \xi_k) \) and \( S \) is a symmetric generator representable semiseparable matrix.

**Proof.** Using \( \xi_{k+1} = \infty \), from Theorem 5.6, we obtain the relation

\[
AQ_{k+1}K_{k+1} = Q_{k+2}H_{k+1},
\]

where \( K_{k+1} \) and \( H_{k+1} \) are tridiagonal, \( e_{k+2}^T K_{k+1} = 0^T \) and \( e_{k+2}^T H_{k+1} \) is a multiple of \( e_{k+1}^T \). Hence, multiplying (5.12) on the left by \( Q_{k+2}^T \) and taking the first \( k+1 \) columns and rows, we have

\[
J_{k+1} = H_{k+1}K_{k+1}^{-1}.
\]

Let us define \( \hat{D}_k = \text{diag}(\gamma, \xi_1, \ldots, \xi_k) \), for \( \gamma \in \mathbb{R} \). Notice that, since the first column of \( Q_{k+1} \) is not an eigenvector of \( A \), the entry in position (1,2) of \( J_{k+1} \) has to be different from 0. From this it can be noticed that \( J_{k+1} - \hat{D}_k \) is a symmetric generator representable semiseparable matrix if and only if \( J_{k+1} - \hat{D}_k \) is invertible. Moreover, taking \( \gamma \neq h_{1,1} \) and \( \gamma \neq 0 \) we have that \( J_{k+1} - \hat{D}_k \) is invertible and its inverse can be computed as follows:

\[
\left( H_{k+1}K_{k+1}^{-1} - \hat{D}_k \right)^{-1} = \left( -\hat{D}_k(K_{k+1} - \hat{D}_k^{-1}H_{k+1})K_{k+1}^{-1} \right)^{-1} = \left( -K_{k+1}(K_{k+1} - \hat{D}_k^{-1}H_{k+1})^{-1}\hat{D}_k^{-1} \right).
\]

Since \( K_{k+1} = I_{k+1} + D_kH_{k+1} \) where \( D_k = \text{diag}(0, \frac{1}{\xi_1}, \ldots, \frac{1}{\xi_k}) \), we have

\[
\left( H_{k+1}K_{k+1}^{-1} - \hat{D}_k \right)^{-1} = -K_{k+1}(I_{k+1} + (D_k - \hat{D}_k^{-1})H_{k+1})^{-1}\hat{D}_k^{-1} = -K_{k+1}^{-1}(I_{k+1} + (D_k - \hat{D}_k^{-1})H_{k+1})^{-1}\hat{D}_k^{-1} = -K_{k+1}^{-1}(I_{k+1} + \frac{1}{\gamma - h_{1,1}} e_1e_1^T H_{k+1})^{-1}\hat{D}_k^{-1} = -(K_{k+1}^{-1} + \frac{1}{\gamma - h_{1,1}} K_{k+1} e_1 e_1^T H_{k+1})\hat{D}_k^{-1}.
\]

The third equality follows from the Sherman-Morrison formula, using the fact that \( \gamma \neq h_{1,1} \). This also shows that the matrix \( J_{k+1} - \hat{D}_k \) is indeed invertible.

The obtained matrix is an irreducible tridiagonal matrix since \( K_{k+1} \) and \( H_{k+1} \) have such structure and \( \gamma \neq 0 \). Hence, using Theorem 5.4, we have that \( J_{k+1} - \hat{D}_k \) is a generator representable semiseparable matrix, and therefore also \( J_{k+1} - \hat{D}_k \) is.

The following corollary generalizes the statement of Theorem 5.6 to the case with poles at \( \infty \).

**Corollary 5.7.** If there exists a pole \( \xi_i = \infty \), the projected matrix \( J_{k+1} = Q_{k+1}^T AQ_{k+1} \) is still a quasiseparable matrix.
Proof. Let \( \{\xi^{(j)}\}_{j \in \mathbb{N}} \) be a sequence of real numbers outside of the convex hull of \( \sigma(A) \) such that
\[
\lim_{j \to \infty} \xi^{(j)} = \infty,
\]
and let \( J_{k+1}^{(j)} \) be the projected matrix obtained by using poles equal to
\[
\{\xi_1, \ldots, \xi_{i-1}, \xi^{(j)}, \xi_{i+1}, \ldots, \xi_k\}.
\]
Since the basis computed by the rational Arnoldi process depends continuously on the poles, we have that
\[
\lim_{j \to \infty} J_{k+1}^{(j)} = J_{k+1}.
\]
From Theorem 5.6 we have that the matrices \( J_{k+1}^{(j)} \) are quasiseparable for all \( j \). Since the quasiseparable matrices are a closed set \([29, \text{Section 1.4.1}]\), we have the thesis. \( \square \)

Notice that, if the matrix \( A \) is symmetric positive semidefinite and \( k+1 \) is less than the invariance index, the projected matrix \( J_k = Q_k^T AQ_k \) has to be positive definite. Indeed, if there exists a vector \( x \neq 0 \) such that \( J_k x = 0 \), we have that \( AQ_k x = 0 \). In particular, since \( Q_k x \in q_{k-1}(A)^{-1} \mathcal{P}_k(A, b) \), where \( q_{k-1} \) as defined in Section 4, there exist \( \alpha_0, \ldots, \alpha_j, j \leq k-1 \), with \( \alpha_j \neq 0 \) such that
\[
Q_k x = q_{k-1}(A)^{-1} \sum_{i=0}^{j} \alpha_i A^i b,
\]
and so
\[
0 = AQ_k x = q_{k-1}(A)^{-1} \sum_{i=0}^{j} \alpha_i A^{i+1} b.
\]
This implies that \( A^{i+1} b \in \mathcal{P}_j(A, b) \), but this is impossible because \( k+1 \) is less than the invariance index. In particular, this implies the existence of the Cholesky factorization of the matrix \( J_k \) if the matrix \( A \) is symmetric positive semidefinite.

The matrix \( B_k = P_k^T AQ_k \) we are interested in when using rational Krylov methods for GMFs is exactly the transpose of the Cholesky factor of the matrix \( J_k = Q_k^T (A^T A) Q_k \). Indeed \( B_k \) is upper triangular and
\[
J_k = (AQ_k)^T AQ_k = (P_k B_k)^T (P_k B_k) = B_k^T B_k.
\]
The following proposition gives us that the matrix \( B_k \) is the upper triangular part of a rank one plus diagonal matrix.

Proposition 5.8. Let \( A \in \mathbb{R}^{n \times n} \) be symmetric positive definite and let \( L \in \mathbb{R}^{n \times n} \) be its Cholesky factor (i.e., \( L \) is lower triangular and \( A = LL^T \)). If there exist \( u, v \in \mathbb{R}^n \) such that \( \text{tril}(A, -1) = \text{tril}(vu^T, -1) \), then \( \text{tril}(L, -1) = \text{tril}(vux^T, -1) \) for \( x \in \mathbb{R}^n \).

Proof. It can be easily proved that the last row of \( \text{tril}(L, -1) \) is equal to
\[
v_n \begin{bmatrix} u_1 \cdots u_{n-1} \end{bmatrix} \begin{bmatrix} L_{n-1}^T \\ 0 \end{bmatrix},
\]
where \( L_{n-1} \) is the leading principal submatrix of \( L \) of size \( n-1 \). Using this fact the thesis can be proved recursively. \( \square \)
Remark 5.9. Notice that if \( \text{triu}(J_k, 1) = \text{triu}(uw^T, 1) \), the vector \( v \) cannot have zero entries: indeed if there exists \( s \leq k \) such that \( v_s = 0 \), then, as a consequence of the proof of Theorem 5.6, the matrix \( J_s - \text{diag}(\gamma, \xi_1, \ldots, \xi_{s-1}) \) has the last column equal to zero for all \( \gamma \in \mathbb{R} \), but this is impossible since for an appropriate choice of \( \gamma \) this matrix has to be invertible.

Exploiting the quasiseparable structure of the matrix \( B_k \), we can compute the matrices \( B_k \) and \( P_k \) by only performing a few scalar products. Indeed, if we let

\[
B_k = \begin{bmatrix}
d_1 & \beta_1 & \gamma_1 & \ast \\
d_2 & \beta_2 & \gamma_2 & \ast \\
& \ddots & \ddots & \ddots \\
0 & d_{k-2} & \beta_{k-2} & \gamma_{k-2} \\
& & d_{k-1} & \beta_{k-1} \\
& & & d_k
\end{bmatrix},
\]

and we define \( x_k = [P_{k-1} \ 0]B_k e_k \), we have that

\[
A q_k = A Q_k e_k = P_k B_k e_k = d_k p_k + x_k.
\]

(5.13)

Using the fact that the submatrix of \( B_k \) that involves the last two columns and all except for the last two rows has rank at most 1, we can compute \( x_k \) with the recursive relation

\[
x_k = \frac{\gamma_{k-2}}{\beta_{k-2}} x_{k-1} + \frac{\gamma_{k-2}}{\beta_{k-2}} p_{k-1}.
\]

This allows us to compute \( d_k, \beta_{k-1}, \gamma_{k-2} \) and \( p_k \) with only two scalar products.

The \( k \)-th step of the procedure is summarized in Algorithm 1.

**Algorithm 1:** \( k \)-th step of rational Golub-Kahan algorithm

| Input: \( A, q_k, p_{k-1}, p_{k-2}, x_{k-1} \) |
| Output: \( p_k, d_k, \beta_{k-1}, \gamma_{k-2}, x_k \) |
| 1 \( w \leftarrow A q_k \) |
| 2 \( \beta_{k-1} \leftarrow w^T p_{k-1} \) |
| 3 \( \gamma_{k-2} \leftarrow w^T p_{k-2} \) |
| 4 \( x_k \leftarrow \gamma_{k-2} x_{k-1} + \beta_{k-1} p_{k-1} \) |
| 5 \( w \leftarrow w - x_k \) |
| 6 \( d_k \leftarrow \| w \|_2 \) |
| 7 \( p_k = \frac{w}{d_k} \) |

Notice that during the \( k \)-th step of the procedure, we do not require the first \( k - 1 \) columns of \( Q_k \). Moreover, for the computation of the projected solution \( \bar{y}_k \) defined in (5.10), we do not need the matrix \( Q_k \). For this reason, the computation of the matrix \( Q_k \) can be performed by using a short recurrence rational Lanczos algorithm, as the one presented in [13, Section 5.2], and we can keep in memory only the last two columns of \( Q_k \). After the \( k \)-th step of the algorithm, the \( k \)-th column of the matrix \( B_k \) can be computed using the newly computed quantities and the previous column, by exploiting the rank structure of the matrix \( B_k \).

Algorithm 1 reduces to the standard Golub-Kahan bidiagonalization if all the poles are chosen equal to infinity.
Remark 5.10. In the algorithm it is implicitly assumed that $\beta_i \neq 0$ for each $i$. In practice this hypothesis is always satisfied, however, as observed in Remark 5.9 if $k$ is less than the invariance index there is at least one nonzero off-diagonal entry in the $k$-th column of $B_k$. Hence we could modify the algorithm to avoid the issue of $\beta_i = 0$.

Remark 5.11. The algorithm presented in this section also works if some of the poles are equal to zero. However, the proof of this fact requires slightly different tools, and hence we omitted it for brevity.

6. Error bounds

In this section we prove some error bounds for the approximation of $f^*(A)b$ using the polynomial and rational Krylov methods described above. These bounds link the approximation error with the error of polynomial and rational approximation of $f$ on an interval containing the singular values of $A$. Our results are the analogue of the ones that hold for standard matrix functions, and they can be proved in a similar way.

We first find an upper and lower bound for the singular values of $B_k$. For convenience, given a matrix $A \in \mathbb{R}^{m \times n}$, throughout this section we are going to use an extended notation for singular values, defining $\sigma_j := 0$ for all $j$ such that $\min\{m, n\} < j \leq \max\{m, n\}$.

Lemma 6.1. Let $\sigma_1$ and $\sigma_n$ be the first and $n$-th singular value of $A \in \mathbb{R}^{m \times n}$, respectively. Then the singular values of $B_k$ belong to the interval $[\sigma_n, \sigma_1]$.

Proof. Let $\eta_1$ and $\eta_k$ be the largest and smallest singular value of $B_k$, respectively. Using the variational characterization of eigenvalues, we have

$$\eta_k^2 = \min_{v \in \mathbb{R}^k} v^T B_k^T B_k v = \min_{v \in \mathbb{R}^k} v^T Q_k^T A^T P_k P_k^T A Q_k v.$$  

Since the columns of $Q_k$ are an orthonormal basis of the subspace $Q_k(A^T A, b)$, we get

$$\eta_k^2 = \min_{w \in Q_k(A^T A, b)} w^T A^T P_k P_k^T A w = \min_{w \in Q_k(A^T A, b)} w^T A^T A w,$$  

where for the last equality we used the fact that $Aw \in Q_k(AA^T, Ab)$ and therefore $P_k P_k^T Aw = Aw$. Finally, we obtain

$$\eta_k^2 = \min_{w \in Q_k(A^T A, b)} w^T A^T A w \geq \min_{w \in \mathbb{R}^n} w^T A^T A w = \lambda_{\min}(A^T A) = \sigma_n^2.$$  

With the same procedure, we also get

$$\eta_1^2 = \max_{w \in Q_k(A^T A, b)} w^T A^T A w \leq \max_{w \in \mathbb{R}^n} w^T A^T A w = \lambda_{\max}(A^T A) = \sigma_1^2.$$  

Observe that if $A \in \mathbb{R}^{m \times n}$ is rectangular with $n > m$, we always have $\sigma_n = 0$, and hence $B_k$ may have singular values arbitrarily close to 0 even if $\sigma_{\min\{m, n\}}(A) > 0$. This fact is going to affect the error bounds in Theorem 6.3 and Theorem 6.7.
As an example, consider the $1 \times 2$ matrix $A = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and the vector $b = \begin{bmatrix} \epsilon \\ 1 \end{bmatrix}$ for small $\epsilon > 0$. For $k = 1$, we have $Q_1 = b/\|b\|_2 = \frac{1}{\sqrt{1+\epsilon^2}} b$, and $P_1 = Ab/\|Ab\|_2 = 1$. So we have $B_1 = P_1^T AQ_1 = \epsilon \in \mathbb{R}^{1 \times 1}$, and hence $B_1$ can have an arbitrarily small singular value even if $\sigma_1(A) = 1$.

6.1. Polynomial error bounds. We first prove the error bounds in the polynomial case. Recall that a polynomial Krylov method computes an approximation to $y = f^\circ(A)b$ from the subspace $\mathcal{P}_k(AA^T, Ab)$ as

$$y_k = P_k f^\circ(B_k)Q_k^T b = \|b\|_2 P_k f^\circ(B_k)e_1,$$

where $B_k = P_k^T AQ_k$, and $P_k$ and $Q_k$ are the matrices computed in the Golub-Kahan bidiagonalization of $A$, satisfying $\text{span}(P_k) = \mathcal{P}_k(AA^T, Ab)$ and $\text{span}(Q_k) = \mathcal{P}_k(AA^T, Ab)$.

A key observation for proving the bounds is the exactness of the approximation (6.1) when $f$ is an odd polynomial, stated in the following lemma.

**Lemma 6.2.** Assume that $f = p_{2k-1}$ is an odd polynomial of degree $\leq 2k - 1$. Then the approximation $\bar{y}_k$ given by (6.1) is exact, i.e., it holds $y = \bar{y}_k$.

**Proof.** It is sufficient to prove this for $f(z) = z^{2\ell-1}$, $\ell = 1, \ldots, k$. For $\ell = 1$, we have

$$y = Ab \quad \text{and} \quad \bar{y}_k = P_k (P_k^T AQ_k)Q_k^T b = Ab,$$

since $Q_k Q_k^T b = b$, and $P_k P_k^T Ab = Ab$ for all $k \geq 1$.

For $\ell > 1$, we have by Proposition 6.3 that

$$y = f^\circ(A) = (AA^T)^{\ell-1} Ab$$

and

$$\bar{y}_k = P_k f^\circ(B_k)e_1 = P_k (B_k B_k^T)^{\ell-1} B_k Q_k^T b.$$

Recalling the definition of $B_k$, we get

$$\bar{y}_k = P_k (B_k B_k^T)^{\ell-2} (P_k^T AQ_k Q_k^T A^T P_k) (P_k^T AQ_k) Q_k^T b.$$

We have $Q_k Q_k^T b = b$ since $b \in \mathcal{P}_1(A^T A, b)$, and similarly $P_k P_k^T Ab = Ab$ since $Ab \in \mathcal{P}_1(AA^T, Ab)$. Hence, letting $b_2 := A^T Ab \in \mathcal{P}_2(A^T A, b)$, we obtain

$$\bar{y}_k = P_k (B_k B_k^T)^{\ell-2} (P_k^T AQ_k) Q_k^T b_2,$$

which is the same situation as before, with $\ell$ replaced by $\ell - 1$, and $b \in \mathcal{P}_1(A^T A, b)$ replaced by $b_2 \in \mathcal{P}_2(A^T A, b)$.

Since $k \geq \ell$, we can repeat the above procedure until we are left with

$$\bar{y}_k = P_k (P_k^T AQ_k) Q_k^T b_\ell, \quad \text{where} \quad b_\ell = (A^T A)^{\ell-1} b \in \mathcal{P}_\ell(A^T A, b).$$

Then, because $k \geq \ell$, we have that $Q_k Q_k^T b_\ell = b_\ell$, and likewise $P_k P_k^T Ab_\ell = Ab_\ell$, since $Ab_\ell \in \mathcal{P}_\ell(AA^T, Ab)$, so we obtain $\bar{y}_k = A(A^T A)^{\ell-1} b$. By comparing with (6.2), we finally get $y = \bar{y}_k$. □

Using Lemma 6.2, we can prove the following theorem.
**Theorem 6.3.** Let \( A \in \mathbb{R}^{m \times n} \), and let \( \sigma_1, \sigma_n \) and \( \sigma_m \) be the first, \( n \)-th and \( m \)-th singular value of \( A \), respectively. Let \( \tilde{y}_k \) be the approximation to \( y = f^\circ(A)b \) given by (6.1). Then the following inequality holds:

\[
(6.4) \quad \| f^\circ(A)b - \tilde{y}_k \|_2 \leq 2\|b\|_2 \min_{p \in \Pi_{k-1}} \| f(z) - p(z^2)z \|_\infty, |\sigma_n, \sigma_1|.
\]

Moreover, if \( A \) is square with \( \sigma_m = \sigma_n > 0 \), or if \( \lim_{z \to \infty} \frac{f(z)}{z} = 0 \), it also holds that

\[
(6.5) \quad \| f^\circ(A)b - \tilde{y}_k \|_2 \leq 2\|Ab\|_2 \min_{p \in \Pi_{k-1}} \| f(z)/z - p(z^2)z \|_\infty, |\sigma_{\max(m,n)}, \sigma_1|.
\]

**Proof.** Let \( p \) be a polynomial of degree \( \leq k - 1 \). Then \( p_{2k-1}(z) = p(z^2)z \) is an odd polynomial of degree \( \leq 2k - 1 \), and by Lemma 6.2 we have

\[
(6.6) \quad p_{2k-1}(A)b = P_kp_{2k-1}(B_k)Q_k^T b.
\]

By adding and subtracting the quantity in (6.6) to \( f^\circ(A)b - \tilde{y}_k \), we get

\[
(6.7) \quad f^\circ(A)b - \tilde{y}_k = [f^\circ(A) - p_{2k-1}(A)]b - P_k[f^\circ(B_k) - p_{2k-1}(B_k)]Q_k^T b.
\]

By invariance of the 2-norm under unitary transformations, we have

\[
\| f^\circ(A) - p_{2k-1}(A) \|_2 = \| f - p_{2k-1} \|_{\infty, \sigma_{\text{sing}}(A)} \leq \| f - p_{2k-1} \|_{\infty, |\sigma_{\max(m,n)}, \sigma_1|},
\]

and similarly, by Lemma 6.1 it holds

\[
\| f^\circ(B_k) - p_{2k-1}(B_k) \|_2 \leq \| f - p_{2k-1} \|_{\infty, |\sigma_n, \sigma_1|}.
\]

Combining the above inequalities with (6.7), we get

\[
\| f^\circ(A)b - \tilde{y}_k \|_2 \leq 2\|b\|_2 \| f - p_{2k-1} \|_{\infty, |\sigma_n, \sigma_1|},
\]

and by taking the minimum over \( p \in \Pi_{k-1} \) we obtain (6.4).

To prove (6.5), recall that if \( \sigma_m > 0 \) or \( \lim_{z \to \infty} \frac{f(z)}{z} = 0 \), by Proposition 3.5 we have \( f^\circ(A) = g(AA^T)A \), where \( g(z) = f(\sqrt{z})/\sqrt{z} \), and similarly, if \( \sigma_n > 0 \) or \( \lim_{z \to \infty} \frac{f(z)}{z} = 0 \) it holds \( f^\circ(B_k) = g(B_kB_k^T)B_k \). Therefore, by also using Proposition 3.3 we can rewrite (6.7) in the form

\[
f^\circ(A)b - \tilde{y}_k = [g(AA^T) - p(AA^T)]Ab - P_k[g(B_kB_k^T) - p(B_kB_k^T)]B_kQ_k^T b.
\]

Given that the eigenvalues of \( B_kB_k^T \) are the squares of the singular values of \( B_k \), with a similar argument as before we obtain

\[
\| f^\circ(A) - \tilde{y}_k \|_2 \leq \| Ab \|_2 \left( \| g - p \|_{\infty, |\sigma_{\max(m,n)}, \sigma_1|} + \| g - p \|_{\infty, |\sigma_n^2, \sigma_m^2|} \right)
\]

\[
\leq 2\|Ab\|_2 \| f(z)/z - p(z^2)z \|_\infty, |\sigma_{\max(m,n)}, \sigma_1|.
\]

As before, (6.5) follows by taking the minimum over \( p \in \Pi_{k-1} \). \( \square \)

**Remark 6.4.** Observe that if the matrix \( A \) is not square, we have \( \sigma_{\max(m,n)} = 0 \), and hence the bound (6.5) always involves a polynomial approximation over the whole interval \([0, \sigma_1]\), even when \( \sigma_{\min(m,n)} > 0 \). If \( A \in \mathbb{R}^{m \times n} \) is rectangular with \( m < n \), then the bound (6.4) also involves the whole interval \([0, \sigma_1]\).

A possible strategy to overcome this issue is to use Proposition 3.7 and write

\[
y = f^\circ(A)b = (A^+)^T f^\circ(A^T)Ab.
\]
The vector \( \mathbf{w} = f^\circ(A^T)\mathbf{A} \) can be approximated using a Krylov method on \( A^T \), and then \( \mathbf{y} \) can be recovered by solving the least squares problem

\[
(6.8) \quad \mathbf{y} = (A^+)^T \mathbf{w} = \arg \min_{\mathbf{y}} \| A^T \mathbf{y} - \mathbf{w} \|_2.
\]

By rewriting the problem in this form, if \( m < n \) and \( \sigma_m > 0 \) we get a bound involving approximation on the smaller interval \( [\sigma_m, \sigma_1] \) for the approximation of \( \mathbf{w} \), which translates to a bound for the approximation of \( \mathbf{y} \).

The bound (6.4) can be manipulated to obtain a more explicit bound. Assume that \( \sigma_n > 0 \), and let \( I = [\sigma_n, \sigma_1] \). The polynomial \( p(z^2)z \) is odd, and we can assume that \( f \) is also odd, so we have

\[
\min_{p \in \Pi_{k-1}} \| f(z) - p(z^2)z \|_{\infty, I} = \min_{p \in \Pi_{k-1}} \| f(z) - p(z^2)z \|_{\infty, (-I) \cup I}
\]

\[
= \min_{q \in \Pi_{k-1}} \| f(z) - q(z) \|_{\infty, (-I) \cup I},
\]

where we used the fact that the polynomial of best approximation on \( (-I) \cup I \) for an odd function is itself odd.

Bounds on the asymptotic rate of convergence for the polynomial approximation of a function on the union of disjoint intervals \( (-I) \cup I \) have been developed in [9]. Using [27, Theorem 8.2] in the proof of [9, Theorem 1] to get explicit inequalities instead of asymptotic bounds, we obtain the following quantitative version of [9, Theorem 1].

We denote by \( E_\rho \) the ellipse with foci at \( \pm 1 \) and vertices at \( \pm \frac{1}{2}(\rho + \frac{1}{\rho}) \), and by \( \tilde{E}_\rho \) its image under the linear function that maps \( [-1, 1] \) to the interval \([a^2, b^2]\). The ellipse \( \tilde{E}_\rho \) has foci at \( a^2 \) and \( b^2 \), and vertices at \( \frac{1}{2}(a^2 + b^2) \pm \frac{1}{2}(\rho + \frac{1}{\rho})(b^2 - a^2) \).

**Theorem 6.5.** Let \( 0 < a < b \) and let \( I = [a, b] \). Assume that \( f|_{-I} \) and \( f|_{I} \) are, respectively, the restrictions of a function \( f_1 \) analytic in the left half-plane \( \text{Re}(z) < 0 \), and of a function \( f_2 \) analytic in the right half-plane \( \text{Re}(z) > 0 \). Then, for all \( k > 0 \) and \( 1 < \rho \leq \frac{b + a}{b - a} \) it holds

\[
(6.9) \quad \min_{p \in \Pi_{2k-1}} \| f(z) - p(z) \|_{\infty, (-I) \cup I} \leq C \frac{\rho}{\rho - 1} \rho^{-k},
\]

where \( C = C(\rho) = M_1 + M_2 + \frac{1}{n}(N_1 + N_2) \), with

\[
M_1 = \max_{z \in \tilde{E}_\rho} |f_1(\sqrt{z})|, \quad N_1 = \max_{z \in \tilde{E}_\rho} |f_1(\sqrt{z})/\sqrt{z}|,
\]

\[
M_2 = \max_{z \in \tilde{E}_\rho} |f_2(\sqrt{z})|, \quad N_2 = \max_{z \in \tilde{E}_\rho} |f_2(\sqrt{z})/\sqrt{z}|,
\]

where \( \tilde{E}_\rho \) denotes the ellipse defined above.

**Remark 6.6.** If we take \( \rho = \frac{b + a}{b - a} \) the ellipse \( \tilde{E}_\rho \) has a vertex at 0. In this case, depending on the function \( f \), we may have \( N_i = +\infty \) and hence \( C = +\infty \). In such a situation the bound only makes sense for \( \rho < \frac{b + a}{b - a} \).

Assuming that \( \sigma_n > 0 \), by plugging (6.9) in the bound (6.4), we get

\[
(6.10) \quad \| f^\circ(A)\mathbf{b} - \bar{y}_k \|_2 \leq 2C\| \mathbf{b} \|_2 \frac{\rho}{\rho - 1} \rho^{-k}, \quad \text{for } 1 < \rho \leq \frac{\sigma_1 + \sigma_n}{\sigma_1 - \sigma_n},
\]

where \( C \) is as defined in Theorem 6.5.
6.2. Rational error bounds. Next, we prove similar error bounds for the rational approximation (5.10). The strategy of the proof is the same as in the polynomial case, even though the proof itself is a bit more technical. We start by stating the result analogous to Theorem 6.3. Recall that the denominator in the rational Krylov space $Q_k(AA^T, Ab)$ is given by the polynomial $q_{k-1}(z) = \prod_{j=1}^{k-1} (1 - z/\xi_j)$, where $\{\xi_j\}_{j \geq 1}$ is a sequence of poles in $(\mathbb{C} \cup \{\infty\}) \setminus \sigma(AA^T)$. For convenience, in this section we use as denominator polynomial for $Q_k(AA^T, b)$ the polynomial $\tilde{q}_{k-1}(z) = \prod_{j=1}^{k-1} (z - \xi_j)$. This polynomial differs from $q_{k-1}$ only by a multiplicative constant, and hence it identifies the same rational Krylov subspace. The results of this section, such as Theorem 6.7, can be equivalently stated in terms of $q_{k-1}$ or $\tilde{q}_{k-1}$.

**Theorem 6.7.** Let $A \in \mathbb{R}^{m \times n}$, and let $\sigma_1$, $\sigma_n$, and $\sigma_m$ be the first, $n$-th and $m$-th singular value of $A$, respectively. Let $\tilde{y}_k$ be the approximation to $y = f^*(A)b$ from $Q_k(AA^T, Ab)$ given by (5.10). Then the following inequality holds:

$$
\|f^*(A)b - \tilde{y}_k\|_2 \leq 2\|b\|_2 \min_{p \in \Pi_{k-1}} \|f(z) - q_{k-1}(z^2)^{-1}p(z^2)z\|_{\infty,[\sigma_n,\sigma_1]}.
$$

Moreover, if $A$ is square with $\sigma_m = \sigma_n > 0$, or if $\lim_{z \to 0} \frac{f(z)}{z} = 0$, it also holds that

$$
\|f^*(A)b - \tilde{y}_k\|_2 \leq 2\|Ab\|_2 \min_{p \in \Pi_{k-1}} \|f(z)/z - q_{k-1}(z^2)^{-1}p(z^2)\|_{\infty,[\sigma_{\max\{m,n\}},\sigma_1]}.
$$

Note that the same issues discussed after Theorem 6.3 in the case of rectangular matrices also arise in the rational case, and the same approach proposed in Remark 6.4 can be used to address them.

Similarly to the polynomial case, the bound (6.11) can be rewritten by exploiting the fact that $f$ can be assumed to be odd and hence that the best approximant on a symmetric interval is odd, yielding

$$
\|f^*(A)b - \tilde{y}_k\|_2 \leq 2\|b\|_2 \min_{p \in \Pi_{k-1}} \|f(z) - q_{k-1}(z^2)^{-1}p(z)\|_{\infty,(I)\cup I},
$$

where again $I = [\sigma_n,\sigma_1]$. However, rational approximation on disjoint intervals is a complicated problem, and hence this formulation might be less useful in practice.

A more practical way to rewrite the bound (6.11) is the following:

$$
\|f^*(A)b - \tilde{y}_k\|_2 \leq 2\|b\|_2 \min_{p \in \Pi_{k-1}} \|f(z) - q_{k-1}(z^2)^{-1}p(z^2)z\|_{\infty,[\sigma_n,\sigma_1]} = 2\|b\|_2 \min_{p \in \Pi_{k-1}} \|\sqrt{z}(f(\sqrt{z})/\sqrt{z} - q_{k-1}(z)^{-1}p(z))\|_{\infty,[\sigma_n^2,\sigma_1^2]} \leq 2\|b\|_2 \min_{p \in \Pi_{k-1}} \|f(\sqrt{z})/\sqrt{z} - q_{k-1}(z)^{-1}p(z)\|_{\infty,[\sigma_n^2,\sigma_1^2]}
$$

Although we get an additional factor $\sigma_1$, this bound relates the error with a uniform rational approximation problem on a real interval. This approximation problem is well-studied in the literature, and it is the same that appears when computing standard matrix functions with rational Krylov methods, so it can be a viable tool for selecting good poles.

Before we prove Theorem 6.7, we state and prove a few auxiliary lemmas. As in the polynomial case, the key point for the proof is the exactness of the rational
Lemma 6.8. Let $k \geq j + 1$, and let $v \in Q_j(A^T A, b)$, with $\bar{q}_{k-1}(z) = \prod_{i=1}^{k-1} (z - \xi_i)$. Then:

\begin{align}
(B_k^T B_k - \xi_j I)^{-1}Q_k^T v &= Q_k^T (A^T A - \xi_j I)^{-1}v \\
B_k^T B_k(B_k^T B_k - \xi_j I)^{-1}Q_k^T v &= Q_k^T A^T A(A^T A - \xi_j I)^{-1}v
\end{align}

Proof. To prove (6.15), it is enough to show that

$$(B_k^T B_k - \xi_j I)Q_k^T (A^T A - \xi_j I)^{-1}v = Q_k^T v.$$ 

Letting $x = (A^T A - \xi_j I)^{-1}v \in Q_{j+1}(A^T A, b)$ and observing that $Q_k Q_k^T x = x$, we have

$$(B_k^T B_k - \xi_j I)Q_k^T (A^T A - \xi_j I)^{-1}v = Q_k^T (A^T P_k P_k^T A - \xi_j I)Q_k Q_k^T x$$

$$= Q_k^T (A^T P_k P_k^T A x - \xi_j x)$$

$$= Q_k^T (A^T A - \xi_j I) x = Q_k^T x,$$

where we also used that $P_k P_k^T A x = A x$, since $A x \in Q_{j+1}(A A^T, Ab)$. This proves (6.15).

To prove (6.16), by using (6.15) we have

$$B_k^T B_k (B_k^T B_k - \xi_j I)^{-1}Q_k^T v = B_k^T B_k Q_k^T x$$

$$= Q_k^T A P_k P_k^T A Q_k Q_k^T x.$$ 

Now, observe that $x \in Q_{j+1}(A^T A, b)$ and that $A x \in Q_{j+1}(A A^T, A x)$, and hence it holds $Q_k Q_k^T x = x$ and $P_k P_k^T A x = A x$. With these facts, we get

$$B_k^T B_k (B_k^T B_k - \xi_j I)^{-1}Q_k^T v = Q_k^T A A^T x,$$

which is equivalent to (6.16). $\Box$

Lemma 6.9. Assume that $f$ is of the form $f(z) = q_{k-1}(z^2)^{-1} p(z^2)$, where $p \in \Pi_{k-1}$. Then the approximation $\bar{y}_k$ from the rational Krylov subspace $Q_k(A A^T, Ab)$ given by (5.10) is exact, i.e., it holds $y = \bar{y}_k$.

Proof. It is sufficient to show that $y = \bar{y}_k$ for $f(z) = \frac{z^{2\ell-1}}{q_{k-1}(z^2)}$, for $\ell = 1, \ldots, k$.

By Corollary 3.3, we have

$$y = f^\circ(A)b = A(A^T A)^{\ell-1} \bar{q}_{k-1}(A^T A)^{-1} b.$$ 

Similarly, we have

$$\bar{y}_k = P_k f^\circ(B_k)Q_k^T b = P_k B_k (B_k^T B_k)^{\ell-1} \bar{q}_{k-1}(B_k^T B_k)^{-1} Q_k^T b$$

$$= P_k B_k \prod_{j=\ell}^{k-1} (B_k^T B_k - \xi_j I)^{-1} \prod_{j=1}^{\ell-1} [(B_k^T B_k)(B_k^T B_k - \xi_j I)^{-1}] Q_k^T b.$$
Now, define the vectors $t_m \in \mathbb{C}^k$ as
\[
    t_m = \begin{cases}
        \prod_{j=1}^{m-1} \left[ (B_k^T B_k)^{-1} - \xi_j I \right] Q_k^T b, & \text{if } 1 \leq m \leq \ell, \\
        \prod_{j=1}^{\ell-1} \left[ (B_k^T B_k)^{-1} - \xi_j I \right] Q_k^T b, & \text{if } \ell < m \leq k.
    \end{cases}
\]
It is straightforward to see that $t_{m+1} = B_k^T B_k (B_k^T B_k - \xi_m I)^{-1} t_m$ for $1 \leq m < \ell$, and that $t_{m+1} = (B_k^T B_k - \xi_m I)^{-1} t_m$ for $\ell \leq m < k$. By Lemma 6.8 we have
\[
    t_2 = B_k^T B_k (B_k^T B_k - \xi_1 I)^{-1} Q_k^T b = Q_k^T A (A^T A - \xi_1 I)^{-1} b,
\]
and hence $t_2 = Q_k^T b_2$, with $b_2 \in Q_2(A^T A, b)$.

By induction, assume that $t_m = Q_k^T b_m$, where $b_m \in Q_m(A^T A, b)$. Then, if $m < \ell$, we have
\[
    t_{m+1} = B_k^T B_k (B_k^T B_k - \xi_m I)^{-1} Q_k^T b_m = Q_k^T A (A^T A - \xi_m I)^{-1} b_m,
\]
where we used (6.16). Defining $b_{m+1} = A^T A (A^T A - \xi_m I)^{-1} b_m$, we get that $t_{m+1} = Q_k^T b_{m+1}$, where $b_{m+1} \in Q_{m+1}(A^T A, b)$. The case $\ell \leq m < k$ is similar: by using (6.15), we find that $t_{m+1} = Q_k^T b_{m+1}$, with $b_{m+1} = (A^T A - \xi_m I)^{-1} b_m \in Q_{m+1}(A^T A, b)$.

Considering $m = k$, by the above discussion we have obtained that
\[
    t_k = Q_k^T b_k = Q_k^T (A^T A)^{-1} \prod_{j=1}^{\ell-1} (A^T A - \xi_j I)^{-1} b
    = Q_k^T (A^T A)^{-1} \tilde{q}_{k-1} (A^T A)^{-1} b,
\]
with $b_k = (A^T A)^{-1} \tilde{q}_{k-1} (A^T A)^{-1} b \in Q_k(A^T A, b)$. Therefore we have
\[
    \bar{y}_k = P_k B_k t_k = P_k P_k^T A Q_k Q_k^T b_k = Ab_k,
\]
where we used $Q_k Q_k^T b_k = b_k$ and $P_k P_k^T Ab_k = Ab_k$. Hence we conclude that $y = \bar{y}_k$.

We now have all the elements to prove Theorem 6.7. The proof follows the same strategy as the proof of Theorem 6.3.

Proof of Theorem 6.7. Let $p$ be a polynomial of degree $\leq k - 1$. Then the rational function $r(z) = \tilde{q}_{k-1}(z^2)^{-1} p(z^2) z$ satisfies the assumptions of Lemma 6.9 and hence we have
\[
    r^\circ(A) b = P_k r^\circ(B_k) Q_k^T b.
\]
By adding and subtracting the quantity in (6.17) to $f^\circ(A) b - \bar{y}_k$, we get
\[
    f^\circ(A) b - \bar{y}_k = [f^\circ(A) - r^\circ(A)] b - P_k [f^\circ(B_k) - r^\circ(B_k)] Q_k^T b.
\]
Since by Lemma 6.1 the nonzero singular values of $B_k$ are contained in the interval $[\sigma_n, \sigma_1]$, by invariance of the 2-norm under unitary transformations, we have
\[
    \| f^\circ(A) - r^\circ(A) \|_2 \leq \| f - r \|_{\infty, [\sigma_{\min(m, n)}, \sigma_1]},
\]
\[
    \| f^\circ(B_k) - r^\circ(B_k) \|_2 \leq \| f - r \|_{\infty, [\sigma_n, \sigma_1]}.
\]
Combining the above inequalities with (6.18), we get
\[
    \| f^\circ(A) b - \bar{y}_k \|_2 \leq 2 \| b \|_2 \| f - r \|_{\infty, [\sigma_n, \sigma_1]}.
\]
and by taking the minimum over \( p \in \Pi_{k-1} \) we obtain (6.11).

To prove (6.12), if \( \sigma_n = \sigma_m > 0 \) or \( \lim_{z \to \infty} \frac{f(z)}{z} = 0 \), by Proposition 3.3 we can write \( f^\circ(A) = g(ABA^T)A \) and \( f^\circ(B_k) = g(B_kB_k^TB_k)B_k \), where \( g(z) = f(\sqrt{z})/\sqrt{z} \). Thus, using also Corollary 3.4 we can rewrite (6.13) in the form

\[
f^\circ(A)b - \bar{y}_k = h(ABA^T)Ab - P_kh(B_kB_k^TB_k)B_kQ_k^Tb,
\]

where \( h(z) = g(z) - \tilde{q}_{k-1}(z)^{-1}p(z) \).

Given that the eigenvalues of \( B_kB_k^TB_k \) are the squares of the singular values of \( B_k \), using Lemma 6.1 and proceeding as above we obtain

\[
\|f^\circ(A) - \bar{y}_k\|_2 \leq \|Ab\|_2 \left( \|h\|_{\infty, [\sigma_m^2, \sigma_n^2]} + \|h\|_{\infty, [\sigma_n^2, \sigma_1^2]} \right) = 2\|Ab\|_2 \left( f(z)/z - \tilde{q}_{k-1}(z)^{-1}p(z) \right)_{\infty, [\sigma_{\max}^{2\sigma_{\max}}]}.
\]

As before, (6.12) follows by taking the minimum over \( p \in \Pi_{k-1} \).

We mention that the results of this section can also be obtained by exploiting the link between GMFs of \( A \) and standard functions of the matrix \( A = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \).

Indeed, it was observed in [3] that for an odd function \( f \) it holds

\[
f(A) = \begin{bmatrix} 0 & f^\circ(A) \\ f^\circ(A^T) & 0 \end{bmatrix}.
\]

If we define the orthogonal matrix \( U_{2k} = \begin{bmatrix} P_k & 0 \\ 0 & Q_k \end{bmatrix} \) and the vector \( c = \begin{bmatrix} 0 \\ b \end{bmatrix} \in \mathbb{R}^{m+n} \), we have that

\[
U_{2k}^T A U_{2k} = \begin{bmatrix} 0 & B_k \\ B_k^T & 0 \end{bmatrix}
\]

and

\[
f(A)c = \begin{bmatrix} f^\circ(A)b \\ 0 \end{bmatrix},
\]

\[
U_{2k} f(U_{2k}^T A U_{2k}) U_{2k}^T c = \begin{bmatrix} P_k f^\circ(B_k)Q_k^Tb \\ 0 \end{bmatrix}.
\]

Moreover, it can be proved that the columns of \( U_{2k} \) are an orthonormal basis for a rational Krylov subspace \( Q_{2k}(A, c) \), the poles of which consist of a single pole at \( \infty \), and \( \pm \theta_j, j = 1, \ldots, k-1 \), where \( \theta_j = \xi_j \) for each \( j \). This fact is straightforward to prove in the polynomial case, where all \( \theta_j \) are equal to \( \infty \).

An alternative derivation of the error bounds in Theorem 6.3 and Theorem 6.7 could then be obtained by combining the above fact with Lemma 6.1 and error bounds concerning rational Krylov approximation of standard matrix functions (see, e.g., [11] Corollary 3.4).

6.3. An optimal pole for the Shift-and-Invert method. In this section we use the error bounds in Theorem 6.7 combined with a known result from approximation theory to find a pole that optimizes the bounds in the case of a single repeated pole (Shift-and-Invert method) located on the negative real line.

We consider the case of a nonsingular square matrix \( A \in \mathbb{R}^{n \times n} \), with singular values contained in the interval \([\sigma_{\min}, \sigma_{\max}]\), with \( \sigma_{\min} > 0 \).

Note that with a change of variables the bound (6.12) can be rewritten in the form

\[
\|f^\circ(A)b - \bar{y}_k\|_2 \leq 2 \|Ab\|_2 \min_{p \in \Pi_{k-1}} \|g(z) - \tilde{q}_{k-1}(z)^{-1}p(z)\|_{\infty, [\sigma_{\min}^2, \sigma_{\max}^2]}.
\]
where the function $g$ is defined as $g(z) = f(\sqrt{z})/\sqrt{z}$.

In the case of a single repeated pole $\xi < 0$, we have $q_{k-1}(z) = (z - \xi)^{k-1}$ and
\[
\left\{ (z - \xi)^{-k+1} p(z) : p \in \Pi_{k-1} \right\} = \left\{ p((z - \xi)^{-1}) : p \in \Pi_{k-1} \right\}.
\]

By defining $h(z) = g(z^{-1} + \xi)$, so that $g(z) = h((z - \xi)^{-1})$, we get
\[
\min_{p \in \Pi_{k-1}} \| g(z) - (z - \xi)^{-k+1} p(z) \|_{\infty,[\sigma^2_{\min},\sigma^2_{\max}]} = \min_{p \in \Pi_{k-1}} \| h(z) - p(z) \|_{\infty,[\mu_{\min},\mu_{\max}]},
\]
where $\mu_{\min} := (\sigma^2_{\max} - \xi)^{-1}$ and $\mu_{\max} := (\sigma^2_{\min} - \xi)^{-1}$. Notice that for $\xi < 0$ it holds indeed $0 < \mu_{\min} \leq \mu_{\max} \leq (-\xi)^{-1}$.

The minimum in (6.19) can be bounded using the following result in approximation theory, adapted from [19, Proposition 3.1]. Its proof relies on classical bounds for Faber series, see [11, Corollary 2.2].

**Proposition 6.10.** Let $\xi < 0$, and assume that $h(z) = g(z^{-1} + \xi)$ is analytic in the strip $0 < \text{Re} z < (-\xi)^{-1}$ and continuous in $[0,(-\xi)^{-1}]$. Then, for any integer $k \geq 1$ it holds
\[
\min_{p \in \Pi_{k-1}} \| h(z) - p(z) \|_{\infty,[\mu_{\min},\mu_{\max}]} \leq 2M \frac{\rho^k}{1 - \rho},
\]
where $M = \| h(z) \|_{\infty,[0,(-\xi)^{-1}]}$ and
\[
\rho = \max \left\{ \frac{\sqrt{\sigma^2_{\max} - \xi} - \sqrt{\sigma^2_{\min} - \xi}}{\sqrt{\sigma^2_{\max} - \xi} + \sqrt{\sigma^2_{\min} - \xi}}, \frac{\sigma_{\max} \sqrt{\sigma^2_{\min} - \xi} - \sigma_{\min} \sqrt{\sigma^2_{\max} - \xi}}{\sigma_{\max} \sqrt{\sigma^2_{\min} - \xi} + \sigma_{\min} \sqrt{\sigma^2_{\max} - \xi}} \right\}.
\]

It follows from the analysis after [19, Proposition 3.1] that the bound (6.20) is optimized by choosing $\xi = -\sigma_{\min} \sigma_{\max}$. This choice leads to the following bound for the Shift-and-Invert iterates:
\[
\| y - \tilde{y}_k \|_2 \leq 2\| b \|_2 M \sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}} \exp \left( -2k \frac{\sqrt{\sigma_{\min}}}{\sigma_{\max}} \right).
\]

We remark that the original result (see [19, equation (3.4)]) exhibited an error like $\exp \left( -2k \frac{\sqrt{\sigma_{\min}}}{\sqrt{\sigma_{\max}}} \right)$ for a symmetric matrix $A$ with spectrum in $[a,b] \subset (0, +\infty)$, when using the Shift-and-Invert method with the optimal pole $\xi = -\sqrt{ab}$. The fact that in (6.21) we have $\sqrt{\sigma_{\min}}$ instead of $\sqrt{\sigma_{\max}}$ is not surprising, since we are essentially applying the result from [19] to the matrix $A^T A$, whose spectrum is contained in $[\sigma^2_{\min}, \sigma^2_{\max}]$.

### 7. Numerical results

In this section we present some numerical experiments with the purpose of illustrating the error bounds and comparing the different methods proposed in the previous sections. We test the methods on randomly generated matrices with a prescribed distribution of singular values, obtained by taking two random orthogonal matrices $U, V \in \mathbb{R}^{n \times n}$ and constructing $A = U \Sigma V^T$, where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n \times n}$. The random orthogonal matrices are obtained by taking a matrix $B \in \mathbb{R}^{n \times n}$ with entries from the normal distribution $\mathcal{N}(0,1)$ and computing the QR factorization $B = QR$. If the diagonal entries of $R$ are nonnegative, then $Q$ is a random
orthogonal matrix from the Haar distribution, a natural uniform probability distribution on the manifold of \( n \times n \) orthogonal matrices [26].

The experiments were done with MATLAB, using the \texttt{ratkrylov} function from the Rational Krylov Toolbox [6] for the implementation of the rational Arnoldi algorithm. The plots display the relative 2-norm error \( \| f^\circ(A)b - \bar{y}_k \|_2 \), where \( \bar{y}_k \) is the approximation defined in (5.8) or (5.10), depending on the Krylov method that was used.

7.1. Error bounds. We start by illustrating in Figure 1 the sharpness of the bound (6.10) for the polynomial Krylov method. Under the assumptions of Theorem 6.5 the rate of convergence in the bound only depends on the interval \([\sigma_n, \sigma_1]\), and hence we can expect it to be pessimistic for most functions. Indeed, for entire functions such as \( \sinh(z) \) and \( \sin(z) \) (see Figure 1(b)) the convergence is much faster than the bound (6.10); however, the bound can capture the asymptotic rate of convergence for certain functions with lower regularity such as \( \sqrt{z} \), for suitable singular value distributions (see Figure 1(a)). This implies that, under the same assumptions of Theorem 6.5, it is only possible to improve the multiplicative constant in the bound (6.10). Note that in Figure 1 the multiplicative constant in the bound (6.10) was ignored for better visualization.

In Figure 2 we compare the convergence of the rational Krylov methods and we test the sharpness of the bounds (6.13) and (6.21). We use the Shift-and-Invert method with the pole \( \xi = -\sigma_{\min}\sigma_{\max} \), the extended Krylov method [10], that alternates poles at \( \infty \) with poles at 0, and a general Krylov method with an asymptotically optimal pole sequence for Laplace-Stieltjes functions, developed in [17]. The poles were selected using the interval \([\sigma_n^2, \sigma_1^2]\), with reference to the
The function $f(z) = \sqrt{z} \log(1 + \sqrt{z})$ is such that the function
\[
\frac{f(\sqrt{z})}{\sqrt{z}} = \frac{\log(1 + \sqrt{z})}{\sqrt{z}}
\]
is Laplace-Stieltjes, or equivalently, completely monotonic [25 Definition 1.3]. This follows from [25 Theorem 3.7] and the fact that $\log(1 + z)/z$ is completely monotonic. An approximation from above to the bound (6.13) was evaluated using a quasi-optimal polynomial $p$ computed by replacing the uniform norm with the 2-norm on a discrete set of points in $I \cup (-I)$. We can see in Figure 2 that the convergence of the rational Krylov method with asymptotically optimal poles closely follows the bound (6.13), and that the convergence rate of the Shift-and-Invert method is correctly predicted by the bound (6.21). The convergence speed of the extended Krylov method is comparable to the one of the Shift-and-Invert method. Note that, as in the polynomial case, the bound (6.21) displayed in Figure 2 does not include the multiplicative constant.

7.2. Rectangular case. Next, we investigate the performance of the methods in the case of a rectangular matrix $A \in \mathbb{R}^{m \times n}$, and the effectiveness of the strategy proposed in Remark 6.4 when $m < n$ to reduce the computation of $f^\odot(A)b$ to the computation of $f^\odot(A^T)A^Tb$. We report in Figure 3 the convergence plots of the rational Krylov method with asymptotically optimal poles, for the functions $f(z) = \sqrt{z}$ and $f(z) = z \log(z)$. We can observe that the convergence is similar for the function $z \log(z)$ (Figure 3(b)), while there is a large benefit in using the alternative expression (6.8) in the case of the function $f(z) = \sqrt{z}$. This is likely due to the fact that $\sqrt{z}$ has a large derivative close to zero, and hence roundoff errors in the smallest computed singular values of the matrix $B_k$ are extremely amplified when applying the function $f$. Indeed, we can see in Figure 3(a) that it...
Figure 3. Convergence of the rational Krylov methods with asymptotically optimal poles for the approximation of $f^\varpi(A)b$, where $A$ is a rectangular 1000 × 1500 matrix whose singular values are Chebyshev points of the second kind for the interval $[10^{-2}, 10]$. The red line shows the convergence of the method described in Remark 6.4, which computes $f^\varpi(A)b$ by first computing $f^\varpi(A^T)Ab$ and then solving a least squares problem.

is not possible to get below a relative accuracy of $10^{-8}$ if we directly approximate $f^\varpi(A)b$, while we can reach a relative error of about $10^{-13}$ if we use the connection with $f^\varpi(A^T)Ab$, since in this case the projected matrix $B_k$ has no singular values close to zero.

7.3. Finite precision issues. In finite precision, one of the main practical problems of the Krylov methods based on a short recurrence (such as, for instance, the Lanczos method) is the loss of orthogonality in the computed basis vector. This phenomenon has been studied for the polynomial Lanczos case in [22]. A brief study of the problem for the rational Lanczos case can be found in [23].

As can be expected, the algorithm presented in Section 6.3 also suffers from this numerical instability. However, our experiments show that this loss of orthogonality deteriorates only slightly the accuracy of the algorithm: if the poles are chosen to guarantee a moderate number of iterations for convergence, it appears that the error produced by comparing the short recurrence algorithm with the one that uses full orthogonalization remains rather small, and it stops growing after a few iterations (see Figure 4). This effect has been already studied in [20] for the approximation of the product between a standard matrix function and a vector by means of the polynomial Lanczos algorithm.

8. Conclusions

In this paper we have proposed the use of rational Krylov methods in the computation of the action of a generalized matrix function on a vector. We have developed an extension of the Golub-Kahan bidiagonalization to the rational case, that uses
a short recurrence to compute the basis vectors of the rational Krylov subspace. We have proved error bounds for the computation of GMFs with polynomial and rational Krylov methods, that relate the error of approximating $f^\circ(A)b$ with the best uniform polynomial or rational approximation of the function $f$ on a real interval containing the singular values of $A$, and we have conducted experiments to investigate the sharpness of such bounds. The experiments we performed also show that rational Krylov methods are particularly effective compared to polynomial Krylov methods when the function $f$ or its derivatives have singularities close to the singular values of $A$.

Acknowledgements

The authors would like to thank Michele Benzi for his support and advice.

References

1. Fredrik Andersson, Marcus Carlsson, and Karl-Mikael Perfekt, Operator-Lipschitz estimates for the singular value functional calculus, Proc. Amer. Math. Soc. 144 (2016), no. 5, 1867–1875. MR 3460149
2. Francesca Arrigo and Michele Benzi, Edge modification criteria for enhancing the communicability of digraphs, SIAM J. Matrix Anal. Appl. 37 (2016), no. 1, 443–468. MR 3479697
3. Francesca Arrigo, Michele Benzi, and Caterina Fenu, Computation of generalized matrix functions, SIAM J. Matrix Anal. Appl. 37 (2016), no. 3, 836–860. MR 3521545
4. Jared L. Aurentz, Anthony P. Austin, Michele Benzi, and Vassilis Kalantzis, *Stable computation of generalized matrix functions via polynomial interpolation*, SIAM J. Matrix Anal. Appl. **40** (2019), no. 1, 210–234. MR 3910862

5. Michele Benzi and Ru Huang, *Some matrix properties preserved by generalized matrix functions*, Spec. Matrices **7** (2019), 27–37. MR 3898254

6. Mario Berljafa, Steven Elsworth, and Stefan Güttel, *A rational Krylov toolbox for MATLAB*, MIMS EPrint 2014.56, Manchester Institute for Mathematical Sciences, University of Manchester, Manchester, UK (2014).

7. Mario Berljafa and Stefan Güttel, *Generalized rational Krylov decompositions with an application to rational approximation*, SIAM J. Matrix Anal. Appl. **36** (2015), no. 2, 894–916. MR 3361437

8. *Parallelization of the rational Arnoldi algorithm*, SIAM J. Sci. Comput. **39** (2017), no. 5, S197–S221. MR 3716554

9. Charles K. Chui and Maurice Hasson, *Degree of uniform approximation on disjoint intervals*, Pacific J. Math. **105** (1983), no. 2, 291–297. MR 691605

10. Vladimir Druskin and Leonid Knizhnerman, *Extended Krylov subspaces: approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl. **19** (1998), no. 3, 755–771. MR 1616584

11. S. W. Ellacott, *Computation of Faber series with application to numerical polynomial approximation in the complex plane*, Math. Comp. **40** (1983), no. 162, 575–587. MR 689474

12. Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, fourth ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 2013. MR 3024913

13. Stefan Güttel, *Rational Krylov Methods for Operator Functions*, Ph.D. thesis, Technische Universität Bergakademie Freiberg, Germany, 2010, Dissertation available as MIMS Eprint 2017.39.

14. *Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection*, GAMM-Mitt. **36** (2013), no. 1, 8–31. MR 3095912

15. J. B. Hawkins and A. Ben-Israel, *On generalized matrix functions*, Linear and Multilinear Algebra **1** (1973), no. 2, 163–171. MR 330185

16. Nicholas J. Higham, *Functions of Matrices. Theory and Computation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. MR 2396439

17. Stefano Massei and Leonardo Robol, *Rational Krylov for Stieltjes matrix functions: convergence and pole selection*, BIT **61** (2021), no. 1, 237–273. MR 4235307

18. Igor Moret and Paolo Novati, *RD-rational approximations of the matrix exponential*, BIT **44** (2004), no. 3, 595–615. MR 2106019

19. *Krylov subspace methods for functions of fractional differential operators*, Math. Comp. **88** (2019), no. 315, 293–312. MR 3854059

20. Cameron Musco, Christopher Musco, and Aaron Sidford, *Stability of the Lanczos method for matrix function approximation*, Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, PA, 2018, pp. 1605–1624. MR 3775893

21. Vanni Noferini, *A formula for the Fréchet derivative of a generalized matrix function*, SIAM J. Matrix Anal. Appl. **38** (2017), no. 2, 434–457. MR 3656502

22. C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl. **18** (1976), no. 3, 341–349. MR 501834

23. Davide Palitta, Stefano Pozza, and Valeria Simoncini, *The short-term rational Lanczos method and applications*, 2021, arXiv:2103.04054.

24. Axel Ruhe, *Rational Krylov algorithms for nonsymmetric eigenvalue problems*, Recent Advances in Iterative Methods, IMA Vol. Math. Appl., vol. 60, Springer, New York, 1994, pp. 149–164. MR 1332748

25. René L. Schilling, Renming Song, and Zoran Vondraček, *Bernstein Functions. Theory and Applications*, second ed., De Gruyter Studies in Mathematics, vol. 37, Walter de Gruyter & Co., Berlin, 2012. MR 2978140

26. G. W. Stewart, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal. **17** (1980), no. 3, 403–409. MR 581487

27. Lloyd N. Trefethen, *Approximation Theory and Approximation Practice*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013. MR 3012510
28. Jasper van den Eshof and Marlis Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput. **27** (2006), no. 4, 1438–1457. MR 2199756

29. Raf Vandebril, Marc Van Barel, and Nicola Mastronardi, *Matrix Computations and Semiseparable Matrices: Linear Systems. Vol. 1*, Johns Hopkins University Press, Baltimore, MD, 2008. MR 2378139

Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy
Email address: angelo.casulli@sns.it

Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy
Email address: igor.simunec@sns.it