Computation of generalized matrix functions with rational Krylov methods

Angelo A. Casulli†  Igor Simunec†

†Scuola Normale Superiore di Pisa

Due Giorni di Algebra Lineare Numerica e Applicazioni

Napoli, 14-15 febbraio 2022
1. Background
   - Generalized matrix functions
   - Rational Krylov subspaces

2. Computation of GMFs
   - Golub-Kahan bidiagonalization
   - Rational Krylov methods
   - Short term recurrence
   - Error bounds

3. Numerical results
1. **Background**
   - Generalized matrix functions
   - Rational Krylov subspaces

2. **Computation of GMFs**
   - Golub-Kahan bidiagonalization
   - Rational Krylov methods
   - Short term recurrence
   - Error bounds

3. Numerical results
Let $A$ be a diagonalizable $n \times n$ matrix, with spectral decomposition

$$A = V \Lambda V^{-1}, \quad \text{where} \ \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

If $f : \mathbb{C} \to \mathbb{C}$ is a function defined on the spectrum of $A$, the **matrix function** $f(A)$ is defined as

$$f(A) = V f(\Lambda) V^{-1}, \quad \text{where} \ f(\Lambda) = \text{diag}(f(\lambda_1), \ldots, f(\lambda_n)).$$

For a general matrix $A$, a matrix function can be defined via the Jordan canonical form.
Generalized matrix functions (GMFs) are an extension of standard matrix functions to the rectangular case, defined using the SVD instead of a spectral decomposition [Hawkins–Ben-Israel, 1973].

Given an \( m \times n \) matrix \( A \) with SVD \( A = U \Sigma V^T \) and a function \( f : \mathbb{R}^+ \to \mathbb{R} \), a generalized matrix function of \( A \), denoted by \( f^\diamond(A) \), is defined as

\[
f^\diamond(A) = U f^\diamond(\Sigma) V^T \in \mathbb{R}^{m \times n},
\]

where \( f^\diamond(\Sigma) \) is diagonal with entries

\[
f^\diamond(\Sigma)_{ii} = \begin{cases} 
  f(\sigma_i) & \text{if } \sigma_i > 0, \\
  0 & \text{otherwise}.
\end{cases}
\]

This definition does not depend on the value of \( f \) at \( z = 0 \), so we can always assume that \( f(0) = 0 \) and \( f \) is odd.
Examples

- For $f(z) = z$, $f^\diamond (A) = A$.
- For $f(z) = z^3$, $f^\diamond (A) = AA^T A$.
- For $f(z) = z^{-1}$, $f^\diamond (A)^T$ is the Moore-Penrose pseudoinverse $A^+$.
- If $f$ is odd and we define $A = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$, then

\[
  f(A) = \begin{bmatrix} 0 & f^\diamond (A) \\ f^\diamond (A^T) & 0 \end{bmatrix}.
\]
Rational Krylov subspaces

Expressions of the form $f(A)b$ can be efficiently computed using polynomial and rational Krylov subspaces.

Given a sequence of poles $\{\xi_j\}_{j \geq 1} \subset (\mathbb{C} \cup \infty) \setminus \Lambda(A)$, the associated rational Krylov subspace is

$$Q_k(A, b) = q_{k-1}(A)^{-1}K_k(A, b),$$

where $q_{k-1}(z) = \prod_{j=1}^{k-1} (z - \xi_j)$ and $K_k(A, b)$ is the Krylov subspace

$$K_k(A, b) = \text{span}\{b, Ab, \ldots, A^{k-1}b\}.$$  

An orthonormal basis $V_k$ of $Q_k(A, b)$ can be computed with the rational Arnoldi algorithm, which requires the solution of $k - 1$ shifted linear systems with the matrix $A$. 
1. Background
   - Generalized matrix functions
   - Rational Krylov subspaces

2. Computation of GMFs
   - Golub-Kahan bidiagonalization
   - Rational Krylov methods
   - Short term recurrence
   - Error bounds

3. Numerical results
Computation of $f^\Diamond(A)b$ via GK bidiagonalization

In [Arrigo-Benzi-Fenu, 2016] the following method is proposed for the efficient computation of $f^\Diamond(A)b$.

- **Perform** $k$ steps of Golub-Kahan bidiagonalization on $A$. This produces a $k \times k$ bidiagonal matrix $B_k$ and matrices $P_k$, $Q_k$ with orthonormal columns, such that $B_k = P_k^T A Q_k$.

- **Approximate** $f^\Diamond(A)b$ with
  \[ y_k = P_k f^\Diamond(B_k) Q_k^T b = P_k f^\Diamond(B_k) e_1 \| b \|_2. \]
In [Arrigo-Benzi-Fenu, 2016] the following method is proposed for the efficient computation of $f^{\diamond}(A)b$.

- Perform $k$ steps of Golub-Kahan bidiagonalization on $A$. This produces a $k \times k$ bidiagonal matrix $B_k$ and matrices $P_k$, $Q_k$ with orthonormal columns, such that $B_k = P_k^T A Q_k$.
- Approximate $f^{\diamond}(A)b$ with
  \[
  y_k = P_k f^{\diamond}(B_k) Q_k^T b = P_k f^{\diamond}(B_k) e_1 \|b\|_2.
  \]

Another efficient method based on Chebyshev interpolation is proposed in [Aurentz-Austin-Benzi-Kalantzis, 2019].

Both methods have very good performance for analytic functions such as $\sin(z)$, $\sinh(z)$. 
The matrices $P_k$ and $Q_k$ constructed in the Golub-Kahan bidiagonalization process are orthonormal bases of the polynomial Krylov subspaces

$$\text{span } Q_k = \mathcal{K}_k(A^T A, b) \quad \text{and} \quad \text{span } P_k = \mathcal{K}_k(AA^T, Ab).$$

In this talk we present a new class of methods, obtained by replacing $\mathcal{K}_k(A^T A, b)$ and $\mathcal{K}_k(AA^T, Ab)$ with the corresponding rational Krylov subspaces.
The matrices $P_k$ and $Q_k$ constructed in the Golub-Kahan bidiagonalization process are orthonormal bases of the polynomial Krylov subspaces

$$\text{span } Q_k = \mathcal{K}_k(A^T A, b) \quad \text{and} \quad \text{span } P_k = \mathcal{K}_k(A A^T, A b).$$

In this talk we present a new class of methods, obtained by replacing $\mathcal{K}_k(A^T A, b)$ and $\mathcal{K}_k(A A^T, A b)$ with the corresponding rational Krylov subspaces.

Let $Q_k$ and $P_k$ be matrices with orthonormal columns such that

$$\text{span } Q_k = Q_k(A^T A, b) \quad \text{and} \quad \text{span } P_k = Q_k(A A^T, A b).$$

Defining $B_k = P_k^T A Q_k$, we can approximate $f^\diamondsuit(A) b$ with

$$y_k = P_k f^\diamondsuit(B_k) Q_k^T b = P_k f^\diamondsuit(B_k) e_1 \|b\|_2.$$
Computational remarks

- A basis $Q_k$ of $Q_k(A^T A, b)$ can be computed by solving $k - 1$ shifted linear systems with $A^T A$.

- Since $Q_k(AA^T, Ab) = AQ_k(A^T A, b)$, we can obtain $P_k$ and $B_k$ with a thin QR decomposition of $AQ_k$. 

In the Golub-Kahan bidiagonalization, $B_k$ is bidiagonal, so the next columns of $P_k$, $Q_k$ and $B_k$ can be computed with a short term recurrence.

In the rational case, $B_k$ is upper triangular but not bidiagonal. However, we can still compute the columns of $P_k$ and $B_k$ with a short recurrence by exploiting the quasiseparable structure of $B_k$, i.e. that all the blocks in its strictly upper triangular part have rank 1.
Computational remarks

- A basis $Q_k$ of $Q_k(A^T A, b)$ can be computed by solving $k - 1$ shifted linear systems with $A^T A$.

- Since $Q_k(AA^T, Ab) = AQ_k(A^T A, b)$, we can obtain $P_k$ and $B_k$ with a thin QR decomposition of $AQ_k$.

- In the Golub-Kahan bidiagonalization, $B_k$ is **bidiagonal**, so the next columns of $P_k$, $Q_k$ and $B_k$ can be computed with a short term recurrence.
Computational remarks

- A basis $Q_k$ of $Q_k(A^T A, b)$ can be computed by solving $k - 1$ shifted linear systems with $A^T A$.

- Since $Q_k(AA^T, Ab) = AQ_k(A^T A, b)$, we can obtain $P_k$ and $B_k$ with a thin QR decomposition of $AQ_k$.

- In the Golub-Kahan bidiagonalization, $B_k$ is bidiagonal, so the next columns of $P_k$, $Q_k$ and $B_k$ can be computed with a short term recurrence.

- In the rational case, $B_k$ is upper triangular but not bidiagonal.

- However, we can still compute the columns of $P_k$ and $B_k$ with a short recurrence by exploiting the quasiseparable structure of $B_k$, i.e. that all the blocks in its strictly upper triangular part have rank 1.
Short term recurrence

Let

\[
B_k = \begin{bmatrix}
  d_1 & \beta_1 & \gamma_1 & \cdots & * \\
  \vdots & \ddots & \ddots & \ddots & \vdots \\
  d_k & 0 & \beta_{k-2} & \gamma_{k-2} \\
  d_{k-1} & 0 & \beta_{k-1} \\
  d_k & 0 & & & \\
\end{bmatrix}.
\]

We have

\[
Aq_k = AQ_k e_k = P_k B_k e_k = d_k p_k + x_k, \quad \text{where } x_k = [P_{k-1} \ 0] B_k e_k.
\]
Short term recurrence

Let

\[
B_k = \begin{bmatrix}
d_1 & \beta_1 & \gamma_1 & \cdots & * \\
& \ddots & \ddots & \ddots & \ddots \\
& & d_{k-2} & \beta_{k-2} & \gamma_{k-2} \\
& & & d_{k-1} & \beta_{k-1} \\
& & & & d_k
\end{bmatrix}.
\]

We have

\[
Aq_k = AQ_k e_k = P_k B_k e_k = d_k p_k + x_k,
\]

where \(x_k = [P_{k-1} \ 0] B_k e_k\).

Using the fact that the submatrices of \(B_k\) in the strictly upper triangular part have rank at most 1, we can compute \(x_k\) with the recursive relation

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

Algorithm 1: Short recurrence rational Krylov approximation of $f^\circ(A)b$

**Input:** $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $f$, $\{\xi_1, \ldots, \xi_{k-1}\}$

**Output:** $y_k \in Q_k(AA^T, Ab)$ s.t. $y_k \approx f^\circ(A)b$

1. $q_1 = \frac{b}{\|b\|_2}$
2. $w_1 = (I - A^T A / \xi_1)^{-1} A^T A q_1$  // can use other choices
3. Compute $q_2$ by orthogonalizing $w_1$ against $q_1$
4. Compute the QR decomposition $[p_1, p_2] \begin{bmatrix} d_1 & \beta_1 \\ 0 & d_2 \end{bmatrix} = [q_1, q_2]$
5. Define $B_2 = \begin{bmatrix} d_1 & \beta_1 \\ 0 & d_2 \end{bmatrix}$ and $x_2 = \beta_1 p_1$
6. for $j = 2, \ldots, k - 1$ do
7. \hspace{1em} $w_j = (I - A^T A / \xi_j)^{-1} A^T A q_j$  // can use other choices
8. \hspace{1em} Compute $q_{j+1}$ by orthogonalizing $w_j$ against $[q_1, \ldots, q_j]$
9. \hspace{1em} Compute $p_{j+1}, d_{j+1}, \beta_j, \gamma_{j-1}$ with short recurrence
10. \hspace{1em} $B_{j+1} = \begin{bmatrix} B_j & s_{j+1} \\ 0 & d_{j+1} \end{bmatrix}$, where $s_{j+1} = \begin{bmatrix} \gamma_{j-1} (B_j)_{1:j-1,j} \\ \beta_j \end{bmatrix}$
11. $P_k = [p_1, \ldots, p_k]$
12. Compute $f^\circ(B_k)$, e.g. via an SVD of $B_k$
13. $y_k = P_k f^\circ(B_k) e_1 \|b\|_2$
Let $A \in \mathbb{R}^{m \times n}$, and let $\sigma_1, \sigma_n$ be the first and $n$-th singular values of $A$, where we use the notation $\sigma_n := 0$ when $n > m$.

Theorem

Let $y_k \in Q_k(AA^T, Ab)$ be the approximation to $f^{\diamond}(A)b$ obtained after $k$ steps of a rational Krylov method, with $q_{k-1}(z) = \prod_{j=1}^{k-1}(z - \xi_j)$ as denominator polynomial. Then

$$
\|f^{\diamond}(A)b - y_k\|_2 \leq 2\|b\|_2 \min_{p \in \mathcal{P}_{k-1}} \|f(z) - q_{k-1}(z^2)^{-1}p(z^2)z\|_\infty, [\sigma_n, \sigma_1].
$$

The bound can be also reformulated as

$$
\|f^{\diamond}(A)b - y_k\|_2 \leq 2\sigma_1\|b\|_2 \min_{p \in \mathcal{P}_{k-1}} \left\|\frac{f(\sqrt{z})}{\sqrt{z}} - \frac{p(z)}{q_{k-1}(z)}\right\|_\infty, [\sigma_n^2, \sigma_1^2].
$$
How to deal with $\sigma_n = 0$

When $A \in \mathbb{R}^{m \times n}$ with $n > m$, we have $\sigma_n = 0$ and the matrix $B_k$ can have arbitrarily small singular values even if $\sigma_m > 0$.

**Example.** Consider $A = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} \varepsilon & 1 \end{bmatrix}^T$, for small $\varepsilon > 0$. Then we have $Q_1 = b/\|b\|_2 = \frac{1}{\sqrt{1+\varepsilon^2}}b$ and $P_1 = Ab/\|Ab\|_2 = 1$. So we get $B_1 = P_1^T AQ_1 = \frac{\varepsilon}{\sqrt{1+\varepsilon^2}}$, which can be arbitrarily close to zero.
How to deal with $\sigma_n = 0$

When $A \in \mathbb{R}^{m \times n}$ with $n > m$, we have $\sigma_n = 0$ and the matrix $B_k$ can have arbitrarily small singular values even if $\sigma_m > 0$.

**Example.** Consider $A = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} \epsilon & 1 \end{bmatrix}^T$, for small $\epsilon > 0$. Then we have $Q_1 = b/\|b\|_2 = \frac{1}{\sqrt{1+\epsilon^2}}b$ and $P_1 = Ab/\|Ab\|_2 = 1$. So we get $B_1 = P_1^T AQ_1 = \frac{\epsilon}{\sqrt{1+\epsilon^2}}$, which can be arbitrarily close to zero.

We can overcome this difficulty with the identity

$$f^{\diamond}(A)b = (A^+)^T f^{\diamond}(A^T)Ab,$$

by first computing $w = f^{\diamond}(A^T)Ab$ with a rational Krylov method on $A^T$ and then recovering $f^{\diamond}(A)b$ as the solution of the least squares problem

$$f^{\diamond}(A)b = (A^+)^T w = \arg \min_y \|A^T y - w\|_2.$$

The projected matrix now has singular values in the interval $[\sigma_m, \sigma_1]$. 
1. Background
   - Generalized matrix functions
   - Rational Krylov subspaces

2. Computation of GMFs
   - Golub-Kahan bidiagonalization
   - Rational Krylov methods
   - Short term recurrence
   - Error bounds

3. Numerical results
Numerical results

We performed experiments on random matrices with prescribed singular values to investigate the sharpness of the error bounds.

We compare the following methods:

- **polynomial** Krylov method;
- **extended** Krylov method, with alternating poles at 0 and $\infty$;
- **Shift-and-Invert** Krylov method, with repeated pole $\xi = -\sigma_{\text{min}}\sigma_{\text{max}}$;
- rational Krylov method with **asymptotically optimal poles** for Laplace-Stieltjes functions, from [Massei-Robol, 2020].

The poles for the fourth method were chosen according to the bound

$$\|f^\diamond(A)b - y_k\|_2 \leq 2\sigma_1\|b\|_2 \min_{p \in P_{k-1}} \left\| \frac{f(\sqrt{z})}{\sqrt{z}} - \frac{p(z)}{q_{k-1}(z)} \right\|_\infty, [\sigma_n^2, \sigma_1^2].$$
Convergence of the polynomial Krylov method, for a $2000 \times 2000$ matrix whose singular values are the Chebyshev points of the second kind for the interval $[10^{-1}, 10]$. 

Convergence – Polynomial
Convergence of rational Krylov methods for \( f^{\circ}(A)b \), where \( A \) is a 2000 \( \times \) 2000 matrix with logspaced singular values in the interval \([1, 10]\) (left) or \([10^{-1}, 10]\) (right), and \( f(z) = \sqrt{z} \log(1 + \sqrt{z}) \).
Convergence of the asymptotically optimal rational Krylov method for $f^\circ(A)b$, where $A$ is a rectangular $1000 \times 1500$ matrix whose singular values are Chebyshev points of the second kind in the interval $[10^{-2}, 10]$. 

Igor Simunec (SNS)  
Computation of GMFs with rational Krylov methods  
17 / 20
Loss of orthogonality

Effects of the loss of orthogonality in the rational Golub-Kahan algorithm for the approximation of \( f^\diamond (A) b \), where \( f(z) = \sqrt{z} \) and \( A \) is a 2000 × 2000 matrix with logspaced singular values in the interval \([10^{-1}, 10^2]\).
Conclusions

- We have proposed a class of rational Krylov methods for the computation of $f^{\Diamond}(A)b$. 

Conclusions

- We have proposed a class of rational Krylov methods for the computation of $f^\diamond(A)b$.
- The projected matrix $B_k$ and the Krylov basis $P_k$ can be computed with a short recurrence by exploiting the quasiseparable structure of $B_k$. 
Conclusions

- We have proposed a class of rational Krylov methods for the computation of $f^\diamond (A) b$.

- The projected matrix $B_k$ and the Krylov basis $P_k$ can be computed with a short recurrence by exploiting the quasiseparable structure of $B_k$.

- We have proved error bounds that relate the convergence rate of these methods to rational approximation of $f$ on $[\sigma_n, \sigma_1]$.
Conclusions

- We have proposed a class of rational Krylov methods for the computation of $f^\diamondsuit(A)b$.
- The projected matrix $B_k$ and the Krylov basis $P_k$ can be computed with a short recurrence by exploiting the quasiseparable structure of $B_k$.
- We have proved error bounds that relate the convergence rate of these methods to rational approximation of $f$ on $[\sigma_n, \sigma_1]$.
- Our numerical experiments indicate that the bounds can accurately predict convergence, and show that the rational Krylov methods converge faster than the polynomial ones when $f$ has low regularity.
Conclusions

- We have proposed a class of rational Krylov methods for the computation of $f^{\diamond}(A)b$.

- The projected matrix $B_k$ and the Krylov basis $P_k$ can be computed with a short recurrence by exploiting the quasiseparable structure of $B_k$.

- We have proved error bounds that relate the convergence rate of these methods to rational approximation of $f$ on $[\sigma_n, \sigma_1]$.

- Our numerical experiments indicate that the bounds can accurately predict convergence, and show that the rational Krylov methods converge faster than the polynomial ones when $f$ has low regularity.
Thank you for your attention!

Our preprint on arXiv:


References:


Backup slides
Practical polynomial error bound

Assume $\sigma_1 > 0$, let $1 < \rho \leq \frac{\sigma_1 + \sigma_n}{\sigma_1 - \sigma_n}$, and denote by $E_{\rho}$ the ellipse with vertices at $\frac{1}{2} (\sigma_n^2 + \sigma_1^2) \pm \frac{1}{4} (\rho + \frac{1}{\rho})(\sigma_n^2 - \sigma_1^2)$ and foci at $\sigma_n^2$ and $\sigma_1^2$.

**Theorem (Bernstein)**

Let the function $g$ be analytic in the interior of the ellipse $E_{\rho}$, and assume that $\max_{z \in E_{\rho}} |g(z)| \leq M$. Then

$$\min_{p \in \mathcal{P}_k} \| g(z) - p(z) \|_{\infty, [\sigma_n^2, \sigma_1^2]} \leq \frac{2M}{\rho - 1} \rho^{-k}.$$

If the function $f(\sqrt{z}/\sqrt{z})$ is analytic in the interior of $E_{\rho}$, we get

$$\| f^\circ(A) b - y_k \|_2 \leq 4M\sigma_1 \| b \|_2 \frac{\rho}{\rho - 1} \rho^{-k},$$

where $M = \max_{z \in E_{\rho}} |f(\sqrt{z}/\sqrt{z})|$ and $1 < \rho \leq \frac{\sigma_1 + \sigma_n}{\sigma_1 - \sigma_n}$. 
Let $y_k$ be the approximation to $f^\diamond(A)b$ from the Shift-and-Invert Krylov method with the single pole $\xi = -\sigma_{\min} \sigma_{\max}$. Then

$$\|f^\diamond(A)b - y_k\|_2 \leq 2\|b\|_2 M \sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}} \exp \left(-2k \sqrt{\frac{\sigma_{\min}}{\sigma_{\max}}} \right),$$

where

$$M = \|h(z)\|_{\infty,[0,-\xi^{-1}]}, \quad h(z) = \frac{f(\sqrt{z^{-1}} + \xi)}{\sqrt{z^{-1}} + \xi}.$$ 

The function $h$ has the property that

$$\frac{f(\sqrt{z})}{\sqrt{z}} = h((z - \xi)^{-1}).$$
Proof sketch of polynomial error bound

If \( f = p_{2k-1} \) is an odd polynomial of degree \( \leq 2k - 1 \), then it is not hard to see that \( y_k = f^\diamond (A)b \).

For a general \( f \), take an approximating odd polynomial \( p_{2k-1} \in \mathcal{P}_{2k-1} \). Then, defining \( h_k = f - p_{2k-1} \), we have

\[
\| f^\diamond (A)b - y_k \|_2 \leq \| h_k^\diamond (A)b \|_2 + \| P_k h_k^\diamond (B_k) Q_k^T b \|_2.
\]

Since we have

\[
\| h_k^\diamond (A) \|_2 \leq \| h_k \|_\infty, [\sigma_{\min\{m,n\}}, \sigma_1],
\]

\[
\| h_k^\diamond (B_k) \|_2 \leq \| h_k \|_\infty, [\sigma_n, \sigma_1],
\]

we obtain

\[
\| f^\diamond (A)b - y_k \|_2 \leq 2 \| b \|_2 \| h_k \|_\infty, [\sigma_n, \sigma_1].
\]

The statement follows by minimizing over the polynomial \( p_{2k-1} \).
Comparison between polynomial and rational Krylov for $f^{\diamond}(A)b$, where $A$ is the $8490 \times 8490$ adjacency matrix of the directed graph p2p-Gnutella30 and $b$ is the vector of all ones. Left: $f_1(z) = \sinh(z)$. Right: $f_2(z) = z^{1/3}$.
### Polynomial vs Rational – Execution time

<table>
<thead>
<tr>
<th>function</th>
<th>polynomial</th>
<th>rational</th>
<th>rational</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>$t_k$</td>
<td>$E_k$</td>
</tr>
<tr>
<td>sinh($z$)</td>
<td>11</td>
<td>0.0103</td>
<td>1.54e-10</td>
</tr>
<tr>
<td>$z^{1/3}$</td>
<td>2000</td>
<td>41.9933</td>
<td>2.48e-03</td>
</tr>
</tbody>
</table>

Number of iterations $k$, execution time $t_k$ in seconds required to achieve tolerance $\text{tol} = 10^{-9}$, and actual error $E_k$ at iteration $k$. The execution times are for the short recurrence implementations, obtained as an average over 10 runs.