Mini-project - Computational Linear Algebra MATH-453

Decay of the entries of matrix functions

Author: Francesco Sala
Professor: Daniel Kressner
Supervisor: Haoze He

Date: June 2nd, 2023

Abstract

Polynomial approximations of matrix functions are widely used in numerical analysis when assembling $f(A)$ is too expensive. If a matrix A is banded, it is possible to show that, under suitable conditions, the entries of the matrix function $f(A)$ exhibit an exponential decay when moving away from the diagonal, thus making the polynomial approximation an effective approximation of $f(A)$. This project focuses on the results presented in [1]. The exponential and inverse matrix functions were studied, and empirical results were compared with theoretical ones. It was possible to show a superlinear convergence of the approximation of the $\exp(A)$ by means of the Chebyshev polynomials of degree k $p_k(A)$. Consequently, since $p_k(A)$ is mk-banded for A m-banded, this result can be considered an empirical verification of the decay of the entries of $f(A)$, provided that f is sufficiently smooth.

Contents

1 Introduction

Let us consider a matrix $A \in \mathbb{C}^{n\times n}$. A major result from linear algebra states that A can be expressed in the canonical Jordan form $Z^{-1}AZ = J = diag(J_1, ..., J_p)$, where:

$$J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix} \tag{1}$$

is the k-th Jordan block, and the sum of the dimensions of all the Jordan blocks equals n. Note that the Jordan matrix J is unique (up to reordering of the blocks J_k), but the transformation matrix Z is not. Let the index of λ_i be the size of the largest Jordan block associated with λ_i, and denote it by $\mathrm{ind}_{\lambda_i}(A)$. The matrix function associated with a scalar function f is defined as:

$$f(A) := p(A) \tag{2}$$

where $p(z)$ is the unique Hermite interpolating polynomial of degree smaller than $\sum_{i=1}^{s} \mathrm{ind}_{\lambda_i}(A)$ such that:

$$\frac{\partial^g}{\partial z^g} p(\lambda_i) = \frac{\partial^g}{\partial z^g} f(\lambda_i), \quad g \in 0, 1, \ldots, \mathrm{ind}_{\lambda_i}(A) - 1, \quad i \in 1, 2, \ldots, s \tag{3}$$

provided that these derivatives exist. Alternatively, the matrix function f can also be defined through the Jordan canonical form as:

$$f(A) := Z f(J) Z^{-1} = Z \,\mathrm{diag}(f(J_k))\, Z^{-1} \tag{4}$$

where:

$$f(J_k) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \frac{1}{2}f^{(2)}(\lambda_k) & \cdots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & f'(\lambda_k) & \cdots & \frac{f^{(m_k-2)}(\lambda_k)}{(m_k-2)!} \\ & & f(\lambda_k) & \ddots & \vdots \\ & & & \ddots & f'(\lambda_k) \\ & & & & f(\lambda_k) \end{bmatrix} \tag{5}$$

We can now dive into the following sections[1].

[1]The theoretical background is adapted from [2] and [3].

2 Bounds on the entries of a matrix function

Let F be an analytic function on a simply connected open region of the complex plane containing the interval $[-1, 1]$: there exist ellipses with foci in -1 and 1 such that F analytic in their interiors. If $\alpha > 1$ and $\beta > 0$ are the half axes of the ellipse, we can define $\chi = \alpha + \beta$, which is the number that completely defines the ellipse $\mathcal{E}_\chi$. Additionally, let $M(\chi) = \max_{z \in \mathcal{E}} |F(z)|$ be the maximum absolute value of F over $\mathcal{E}_\chi$.

Theorem 2.1 (Bernestein's theorem [1])**.** *Let F be an analytic function in the interior of the ellipse $\mathcal{E}_\chi$, $\chi > 1$, and continuous on $\mathcal{E}_\chi$. Let $F(z)$ be real for $z \in \mathbb{R}$. Then:*

$$E_k(F) \leq K_0 q^{k+1} \tag{6}$$

where

$$K_0 = \frac{2\chi M(\chi)}{\chi - 1},\ q = \frac{1}{\chi} \tag{7}$$

and $E_k(F) = \inf\{||F - p||_\infty : p \in P_k\}$.

Note that the decay expressed by the theorem becomes slower for $\chi \to 1^+$, while the "larger" the ellipse over which F is analytical, the faster the decay. Note additionally that, by its definition, $E_k(F)$ is the *best* approximation error. Before proceeding any further, let us show that the product of two banded matrices is banded. Given A and B a-banded and b-banded respectively, i.e. such that $A_{ij} = 0$ for $|i - j| > \frac{a}{2}$ and $B_{ij} = 0$ for $|i - j| > \frac{b}{2}$, their product AB is $(a + b)$-banded. Indeed we have:

$$(AB_{ij}) = \sum_{k=1}^{n} A_{ik} B_{kj} = \sum_{\substack{|k-i| \leq \frac{a}{2} \\ |k-j| \leq \frac{b}{2}}} A_{ik} B_{kj}$$

If $|i-j| > \frac{a}{2} + \frac{b}{2}$ there is no k that can satisfy both conditions in the summation at the same time, and thus AB_{ij} will be $(a+b)$-banded. In particular, this results shows, by induction, that A^k, with A m-banded, will be km-banded.

Theorem 2.2. *Let A be symmetric, tridiagonal (2-banded) and such that $[-1, 1]$ is the smallest interval containing $\sigma(A)$ (its spectrum). Let*

$$K = \max\{K_0, ||F(A)||_2\} \tag{8}$$

with F and K_0 as above. Then we have:

$$|(F(A))_{ij}| \leq K q^{|i-j|} \tag{9}$$

Proof. First, we have shown that A^k will be $2k$-banded, and hence any matrix polynomial $p_k(A)$ will be $2k$-banded. From Eq. 6 (Bernstein's theorem), we have:

$$||F(A) - p_k(A)||_2 = \max_{x \in \sigma(A)} |F(x) - p_k(x)| \leq ||F - p_k||_\infty \leq K_0 q^{k+1} \tag{10}$$

If $i \neq j$, we select k such that $|i-j| = k+1$ and since $p_k(A)_{ij} = 0$ for these indices (i, j), we have:

$$|(F(A))_{ij}| = |(F(A))_{ij} - (p_k(A))_{ij}| \leq ||F(A) - p_k(A)||_2 \leq K_0 q^{k+1} \leq K_0 q^{|i-j|} \leq K q^{|i-j|} \tag{11}$$

Conversely, if $i = j$ (diagonal entries), we have:

$$|(F(A))_{ii}| \leq ||F(A)||_2 \tag{12}$$

and by definition of K, $||F(A)||_2 \leq K$, and thus the theorem holds true even in this case.

□

It is now interesting to study what happens if this theorem is applied to a m-banded matrix with eigenvalues contained in a generical interval $I = [a, b]$. We will make use of affine functions as in [1] to derive a meaningful result. In particular, let A be a symmetric matix, and let $a = \lambda_{min}(A)$ and $b = \lambda_{max}(A)$. We can introduce the affine function ψ so that:

$$\psi : \mathbb{C} \to \mathbb{C}, \quad \psi(\lambda) = \frac{2\lambda - (a+b)}{b-a} \tag{13}$$

This function will map the eigenvalues of A from the interval $[a, b]$ to the interval $[-1, 1]$, as required by Theorem 2.2. We can then define a new matrix $B = \psi(A)$:

$$B = \psi(A) = \frac{2}{b-a} A - \frac{a+b}{b-a} I \tag{14}$$

whose eigenvalues are in $[-1, 1]$. If we want to study the behavior of $f(A)$, with f analytic on a simply connected region containing [a,b], and such that $f(\lambda) \in \mathbb{R}$ if $\lambda \in \mathbb{R}$, we can study $F = f \circ \psi^{-1}$, which will satisfy the hypothesis of the Theorem 2.2. By definition of F, we have:

$$F(B) = f \circ \psi^{-1}(B) = f \circ \psi^{-1}(\psi(A)) = f(A) \tag{15}$$

Let $\alpha > 1$ and $\beta > 0$ be the semi-axes of the ellipse $\mathcal{E}_\chi$ over which F is analytic; let γ and δ be the semi-axes of the ellipse $\mathcal{N}_\zeta$ over which f is analytic,

with $\zeta = \delta + \gamma$, and let $w = \frac{b-a}{2}$. Note that the semi-axes are simply stretched by the trasformation ψ^{-1}, according to:

$$\gamma = \frac{b-a}{2}\alpha$$
$$\delta = \frac{b-a}{2}\beta$$

and thus:

$$\alpha = \frac{2}{b-a}\gamma$$
$$\beta = \frac{2}{b-a}\delta$$

The theoretical bound in Eq. 9 can thereby be applied, with the following result:

$$|F(B)_{ij}| = |f(A)_{ij}| \leq Kq^{|i-j|} \tag{16}$$

where $K = \max\{K_0, ||f(A)||_2\}$, and:

$$\begin{aligned} K_0 &= \frac{2\chi \max_{x\in\mathcal{E}_\chi}\{|F(x)|\}}{\chi - 1} \\ &= \frac{2(\alpha+\beta)\max_{x\in\mathcal{E}_\chi}\{|F(x)|\}}{\alpha+\beta-1} \\ &= \frac{\left(\frac{4}{b-a}\right)(\gamma+\delta)\max_{\hat{x}\in\mathcal{N}_\zeta}\{|f(\hat{x})|\}}{\left(\frac{2}{b-a}\right)(\gamma+\delta)-1} \\ &= \frac{2\zeta \max_{\hat{x}\in\mathcal{N}_\zeta}\{|f(\hat{x})|\}}{\zeta - w} \end{aligned}$$

while:

$$q = \frac{1}{\chi} = \frac{1}{\alpha+\beta} = \frac{b-a}{2(\gamma+\delta)} \tag{17}$$

It is therefore possible to bound the entries of a more generical $f(A)$ once the ellipse $\mathcal{N}_\zeta$ is known.

Finally, note that the bound can be rearranged using the Frobenius norm. Indeed, for $A \in \mathbb{C}^{n\times n}$:

$$||A||_F \leq \sqrt{n}||A||_2 \tag{18}$$

since

$$||A||_F^2 = \sum_{i=1}^{n} ||a_i||_2^2 = \sum_{i=1}^{n} ||Ae_i||_2^2 \leq \sum_{i_1}^{N} ||A||_2^2 ||e_i||_2^2 = n||A||_2^2 \tag{19}$$

where a_i, $i = 1, \ldots, n$ denotes the columns of A. Therefore, we can make use of this result to reexpress Eq. 6 and the first inequality in the proof of Theorem 2.2 to get to:

$$||f(A) - p_k(A)||_F \leq \sqrt{n} K q^{k+1} \tag{20}$$

where all terms are as defined above. Recall that this bound holds for a sequence of polynomials p_k that realize the *best approximation error*.

3 Numerical study

Let us consider the matrix A:

$$A = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{100 \times 100} \tag{21}$$

which is a tridiagonal 2-banded matrix. Besides, A is a Toepliz matrix; for such a class of tridiagonal matrices, the eigenvalues can be computed in a closed form according to the following formula:

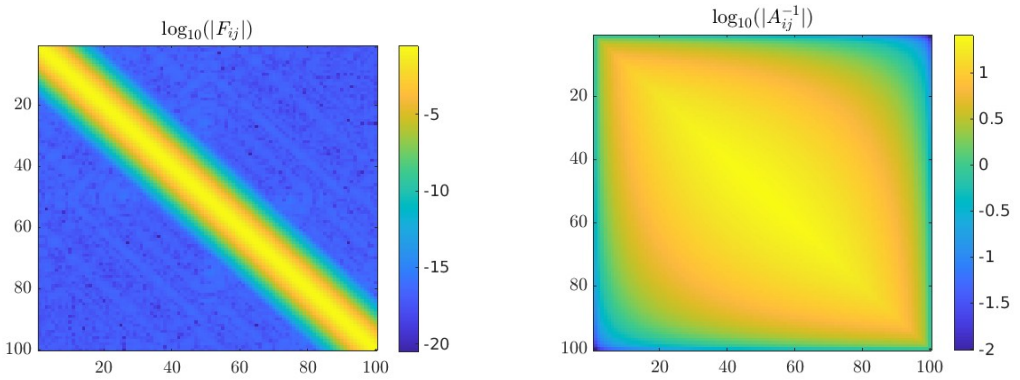
$$a_1 + 2\sqrt{a_2 a_3} \cos\left(\frac{k\pi}{n+1}\right), \quad k = 1, \ldots, n \tag{22}$$

where a_1, a_2, and a_3 are the constant entries of the three non-zero diagonals. For the given A, the eigenvalues are hence:

$$-2 + 2\cos\left(\frac{k\pi}{101}\right), \quad k = 1, \ldots, 100 \tag{23}$$

and are all contained in the interval $(-4, 0)$; the larger the matrix, the closer $\lambda_{min}(A)$ and $\lambda_{max}(A)$ to the extrema of the interval.

The computation of $f(A) = \exp(A)$ by means of the MATLAB command `expm` shows an exponential decay of the magnitude of the entries of $f(A)$ as displayed in Fig. 1a: for $|i - j| > 16$ the values of the entries of $\exp(A)$ are comparable to the ε-machine. Additionally, it can be noted that $f(A)$ is characterized by a banded structure.



(a) Logarithm of the entries of $f(A) = exp(A)$. *(b) Logarithm of the entries of $f(A) = A^{-1}$.*

Figure 1: Logarithm of the entries of $f(A)$.

On the contrary, if $f(A) = A^{-1}$, the resulting plot differs significantly from the previous one, as depicted in Fig. 1b, obtained using the MATLAB command `inv`. In this case, the decay is significantly slower and despite the bandedness of A, A^{-1} is a full matrix with almost no negligible entry. In particular, the magnitude of each entry does not go below 10^{-2} over the entire matrix.

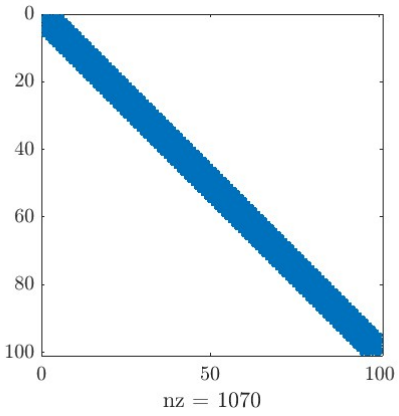
3.1 Decay of the entries and approximation of $f(A)$

The decay of the entries of a matrix function was studied by first looking at the approximation error in the Frobenius norm and then by exploiting the theoretical results presented in the previous sections. First of all, in order to approximate $f(A)$ using a polynomial of degree k such that:

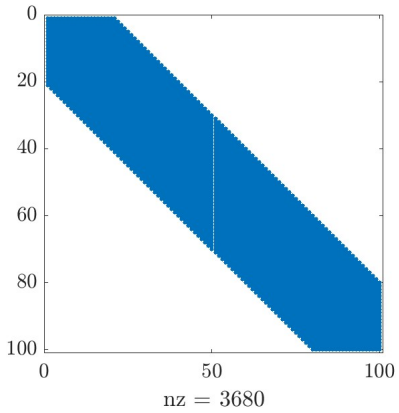
$$f(A) \approx p_k(A) \tag{24}$$

a MATLAB function `fm.m` was implemented. Such a function makes use of the MATLAB package Chebfun to interpolate f at Chebyshev nodes in the interval $[\lambda_{min}, \lambda_{max}]$. In particular, it defines the inverse of the function ψ defined in Eq. 13 to approximate the function $F = f \circ \psi^{-1}$ over the interval $[-1, 1]$. Now, in the proof of Theorem 2.2, we stated that $p_k(A)$ is expected to be $2k$-banded for A tridiagonal. This fact can be numerically verified: Fig. 2 shows the sparsity pattern of both $p_5(A)$ and $p_{20}(A)$ are plotted using the MATLAB command `spy`: both plots visually confirm that, given A 2-banded, $p_k(A)$ will be $2k$-banded. In light of this result and that of Fig. 1, we expect that $p_k(A)$ be a good approximation of $\exp(A)$ and not of A^{-1}.

Therefore, it is now interesting to numerically study what happens to the entries of $f(A)$ when approximating it through $p_k(A)$ for different values of k.



(a) Sparsity pattern of $p_5(A)$.



(b) Sparsity pattern of $p_{20}(A)$.

Figure 2

In particular, the Frobenius norm was used to estimate the distance between the function $f(A)$ and its approximation $p_k(A)$. Fig. 3 shows how the approximation error behaves when changing the value of k both for the exponential and inverse functions, using a plot with logarithmic y-axis. Along with the approximation error, the bound given by Eq. 20 is plotted as well as a reference. To compute the bound, it is necessary to compute the terms that appear in Eq. 20. For this purpose, let $a = \lambda_{min}(A)$, $b = \lambda_{max}(A)$ be the two foci of the ellipse $\mathcal{N}_\zeta$; if we define a value of the vertical semi-axis δ, then the other semi-axis γ is given by:

$$\gamma = \sqrt{w^2 + \delta^2}$$

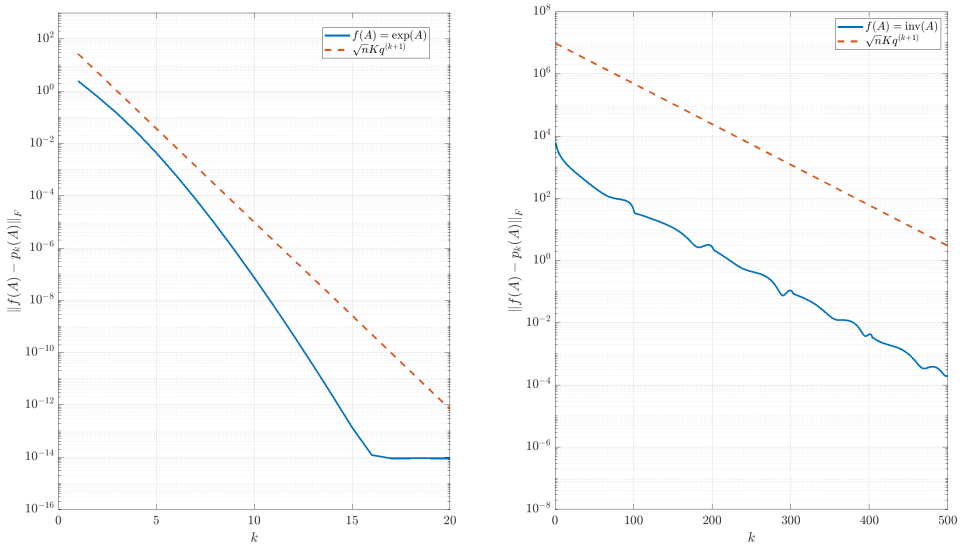
with $w = \frac{b-a}{2}$. Besides, the maximum in the absolute value of both the $\exp(x)$ and x^{-1} over the ellipse $\mathcal{N}_\zeta$ is reached for the point on the x-axis with the largest coordinate, i.e. for $x = \frac{b+a}{2} + \gamma$. Furthermore, while $f(x) = \exp(x)$ is analytic and continuous over the entire complex plane, allowing a free choice of δ, for $f(x) = x^{-1}$ the ellipse must not touch the y-axis to satisfy the hypotheses of Bernstein's theorem, i.e. $\gamma < \frac{|b+a|}{2}$. This implies:

$$0 < \delta < \sqrt{\left(\frac{|b+a|}{2}\right)^2 - w^2} \tag{25}$$

and for the given matrix A, we get the bound $\delta < 0.0622$.

By looking at Fig. 3, it is possible to note that the approximation of the exponential converges significantly faster than that of the inverse, reaching a Frobenius norm $\sim 1 \times 10^{-14}$ for $k = 20$, while the error on the inverse function is above 10

for $k = 100$, and, despite decaying, the convergence is decidedly slower. Additionally, although the theoretical bound in Eq. 20 refers to the case of the best approximation error, the use of Chebyshev polynomials attains an approximation error below the theoretical bound. This may lead to conclude that the Chebyshev approximation may not be far from the best attainable, although a more formal proof is required to prove this argument.



(a) $f(A) = exp(A)$, $\delta = 5$. *(b)* $f(A) = A^{-1}$, $\delta = 0.06$.

Figure 3: Convergence analysis in the Frobenius norm.

Moreover, the ellipse $\mathcal{N}_\zeta$ over which the exponential meets the hypothesis of Theorem 2.1 may extend to the entire complex plane, and hence the sum of the semi-axes ζ may be increased to show a faster convergence: for every value of k, an optimal value of ζ can be computed. This explains the superlinear convergence obtained analytically. On the contrary, the ellipse over which the inverse function is analytic and continuous is constrained by the singularity in 0. Therefore, we cannot prove a convergence much faster than the one plotted even if we were to use the sequence of polynomials that realize the best approximation error[2]. Finally, it is now possible to link these results about the approximation of f using a polynomial with the decay of a matrix function. Since $p_k(A)$ is $2k$-banded, a small value of $||f(A) - p_k(A)||_F$ ensures that $f(A)$ has indeed small-magnitude entries far from the diagonal. This allows approximating $f(A)$ using a polynomial with a trade-off:

[2]Initially the function `fm.m` computed the approximation of f using Chebyshev polynomials, and then the evaluation of $p(A)$ was performed using `polyvalm`, i.e. moving to the canonical basis. This led to numerical instabilities, and a different approach has therefore been adopted in order not to move from the Chebyshev basis to the canonical one.

the larger the degree of the polynomial k, the better the approximation, but we will have to deal with a less sparse matrix. Moreover, the smallest interval that contains the eigenvalues of A must be known: the presented approach works well for a Toepliz matrix, whose eigenvalues are known in closed form.

4 Summary

The decay of the entries of a matrix function was studied, using the functions $\exp(x)$ and x^{-1} as benchmarks. It was possible to show that, under the hypothesis of $f(x)$ analytic and continuous over an ellipse in the complex plane, the function $f(A)$ displays a decay of its entries for a banded matrix A, consistently with the definition of analytic functions and the consideration that $p_k(A)$ is mk-banded for a m-banded matrix A. Additionally, the convergence of the approximation of $f(A)$ using Chebyshev polynomials was shown to be significantly fast for $\exp(A)$, suggesting that these polynomials may not be far from the best reachable approximation. Further work could focus on studying these polynomials and their relationship with the best approximation error. In conclusion, the implemented method for approximating $f(A)$ can be satisfactorily used as long as f satisfies the hypotheses of Bernstein's theorem, and the eigenvalues of A are known or can be efficiently computed.

References

[1] M. Benzi and G. Golub, "Bounds for the entries of matrix functions with applications to preconditioning," *BIT Numerical Mathematics*, vol. 39, Mar. 1998. DOI: 10.1023/A:1022362401426.

[2] N. Higham, *Functions of Matrices: Theory and Computation* (Other Titles in Applied Mathematics). Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2008, ISBN: 9780898717778. [Online]. Available: https://books.google.ch/books?id=S6gpNn1JmbgC.

[3] D. Kressner, "Computational linear algebra - lecture notes," Feb. 2023.