

Matrix functions

Nathan Rousselot
Mathematical Engineering Department
KU Leuven

August 14, 2023

1 Introduction

In this document we introduce the notion of matrix functions. Say we have a function $f : \mathbb{C} \rightarrow \mathbb{C}$, then we can define the function f on a matrix \mathbf{A} as follows: $f : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$. In the following, we will first provide some theoretical background on matrix function. Then, we will describe the numerical issues coming with manipulating matrix functions. Finally, we will provide algorithm for efficient computation of matrix functions.

2 Theoretical background

2.1 Natural Definition

Polynomial functions

Let $p : \mathbb{C} \rightarrow \mathbb{C}$ be a polynomial function of degree d :

$$p(t) = \sum_{k=0}^d c_k t^k \quad (2.1)$$

Then, considering a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, and posing $\mathbf{A}^0 = I_n$, we can define the polynomial function $p : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ on a matrix \mathbf{A} as follows:

$$p(\mathbf{A}) = \sum_{k=0}^d c_k \mathbf{A}^k \quad (2.2)$$

Rational functions

Let $f : \mathbb{C} \rightarrow \mathbb{C}$ be a rational function of the form:

$$f(t) := \frac{p(t)}{q(t)} \quad (2.3)$$

It is not immediate how one would approach this function with a matrix. We want to define

$$f(\mathbf{A}) := q(\mathbf{A})^{-1}p(\mathbf{A}) \quad (2.4)$$

However, this is not well defined if $q(\mathbf{A})$ is singular. In other words, we need to make sure that $q(\mathbf{A})$ is invertible. This is the case if and only if $q(\lambda) \neq 0$ for all eigenvalues λ of \mathbf{A} . This is a very strong condition, and it is not always possible to find a rational function f such that $q(\lambda) \neq 0$ for all eigenvalues λ of \mathbf{A} . We note that a choice in notation has been made in equation 2.4, the other notation is still valid.

Lemma 1. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$, and let $p : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ and $q : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$, two matrix polynomials. Then,*

$$q(\mathbf{A})p(\mathbf{A}) = p(\mathbf{A})q(\mathbf{A}) \quad (2.5)$$

Proof. Let $f(\mathbf{A}) := q(\mathbf{A})p(\mathbf{A})$, and $g(\mathbf{A}) = p(\mathbf{A})q(\mathbf{A})$. Then

$$\begin{aligned} f(\mathbf{A}) &= \left(\sum_{k=0}^d c_k \mathbf{A}^k \right) \left(\sum_{j=0}^m b_j \mathbf{A}^j \right) \\ &= \sum_{k=0}^d \sum_{j=0}^m c_k b_j \mathbf{A}^{j+k} \\ &= \sum_{j=0}^m \sum_{k=0}^d b_j c_k \mathbf{A}^{k+j} \\ &= \left(\sum_{j=0}^m b_j \mathbf{A}^j \right) \left(\sum_{k=0}^d c_k \mathbf{A}^k \right) = g(\mathbf{A}) \end{aligned}$$

□

Theorem 1. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$, and let $p : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ and $q : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$, two matrix polynomials such that $q(\mathbf{A})$ is non-singular. Then,*

$$q(\mathbf{A})^{-1}p(\mathbf{A}) = p(\mathbf{A})q(\mathbf{A})^{-1} \quad (2.6)$$

Proof. Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a matrix such that $q(\mathbf{A})$ is non-singular

$$q(\mathbf{A})^{-1}p(\mathbf{A}) = q(\mathbf{A})^{-1}p(\mathbf{A})q(\mathbf{A})q(\mathbf{A})^{-1}$$

Using Lemma 1

$$\begin{aligned} q(\mathbf{A})^{-1}p(\mathbf{A})q(\mathbf{A})q(\mathbf{A})^{-1} &= q(\mathbf{A})^{-1}q(\mathbf{A})p(\mathbf{A})q(\mathbf{A})^{-1} \\ &= p(\mathbf{A})q(\mathbf{A})^{-1} \end{aligned}$$

□

Power Series

Let $f : \mathbb{C} \rightarrow \mathbb{C}$ be a function that can be expressed as a power series:

$$f(t) = \sum_{k=0}^{\infty} c_k t^k \quad (2.7)$$

Then, considering a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, and posing $\mathbf{A}^0 = I_n$, we can define the power series function $f : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ on a matrix \mathbf{A} as follows:

$$f(\mathbf{A}) = \sum_{k=0}^{\infty} c_k \mathbf{A}^k \quad (2.8)$$

In the scalar case, we know that the power series converges if $|t| < r$, where r is the radius of convergence. Obviously this translates in the matrix power series

Theorem 2. *Let $f : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ be a matrix power series. Then, the series converges if and only if $\rho(\mathbf{A}) < r$, where $\rho(\mathbf{A})$ is the spectral radius of \mathbf{A} , and r is the radius of convergence of the scalar power series.*

Proof is provided in Frommer and Simoncini 2008. In the case of a finite-order Laurent Series, *i.e.*:

$$f(t) = \sum_{k=-d}^d c_k t^k \quad (2.9)$$

For the matrix case, we need to ensure convergence (similarly to power series), but also ensure existence of the inverse, as Laurent series do have negative powers. If both of those conditions are satisfied, we can write the Laurent series as a matrix function:

$$f(\mathbf{A}) = \sum_{k=-d}^d c_k \mathbf{A}^k \quad (2.10)$$

2.2 Spectrum-Based Definition

Diagonalizable Matrices

Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a diagonalizable matrix. That means that there exists a matrix $\mathbf{V} \in \mathbb{C}^{n \times n}$ such that \mathbf{V} is invertible, and $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of \mathbf{A} :

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \quad (2.11)$$

Theorem 3. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a diagonalizable matrix. Then we can define the function $f(\mathbf{A})$ as:*

$$f(\mathbf{A}) := \mathbf{V}f(\mathbf{\Lambda})\mathbf{V}^{-1} \quad (2.12)$$

with

$$f(\mathbf{\Lambda}) = \begin{bmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_n) \end{bmatrix} \quad (2.13)$$

This property is very handy, as it allows us to compute matrix functions by simply applying the function to the eigenvalues of the matrix. Computationally, this avoids inverting matrices, and lots of matrix products. However, this property is only valid for diagonalizable matrices. This property puts constraints on $f(\mathbf{A})$, as its eigenvectors must form a basis \mathbf{F}^n . Another more practical constraint, that is sufficient but not necessary, is if \mathbf{A} is a full rank matrix, then it is diagonalizable.

Defective Matrices

In some cases, the matrix \mathbf{A} is not diagonalizable, that means the sum of the dimensions of the eigenspaces is less than n , we call that a *Defective Matrix*. In that case, we can generalize the principle of diagonalization using the Jordan canonical form of \mathbf{A} :

$$\mathbf{A} = \mathbf{V}\mathbf{J}\mathbf{V}^{-1} \quad (2.14)$$

where \mathbf{J} is a Jordan matrix, and \mathbf{V} is a matrix containing the generalized eigenvectors of \mathbf{A} . The Jordan matrix is a block diagonal matrix, where each block is a Jordan block. A Jordan block is a matrix of the form:

$$\mathbf{J}_k(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \quad (2.15)$$

Theorem 4. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a defective matrix. Then we can define the function $f(\mathbf{A})$ as:*

$$f(\mathbf{A}) := \mathbf{V}f(\mathbf{J})\mathbf{V}^{-1} \quad (2.16)$$

with

$$f(\mathbf{J}) = \begin{bmatrix} f(\mathbf{J}_1(\lambda_1)) & 0 & \cdots & 0 \\ 0 & f(\mathbf{J}_2(\lambda_2)) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\mathbf{J}_k(\lambda_k)) \end{bmatrix} \quad (2.17)$$

and

$$f(\mathbf{J}_i(\lambda_i)) = \begin{bmatrix} f(\lambda_i) & f'(\lambda_i) & \frac{f''(\lambda_i)}{2!} & \cdots & \frac{f^{(k-1)}(\lambda_i)}{(k-1)!} \\ 0 & f(\lambda_i) & f'(\lambda_i) & \cdots & \frac{f^{(k-2)}(\lambda_i)}{(k-2)!} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_i) & f'(\lambda_i) \\ 0 & 0 & \cdots & 0 & f(\lambda_i) \end{bmatrix} \quad (2.18)$$

Obviously, both definitions of matrix functions, based on diagonalization and Jordan canonical form, presuppose that the spectral radius of \mathbf{A} , $\rho(\mathbf{A})$, is less than r , the radius of convergence.

2.3 Interpolation-based definition

Interestingly, in this section we will show that for any $\mathbf{A} \in \mathbb{C}^{n \times n}$ and any sufficiently differentiable f , we can find a polynomial p such that $f(\mathbf{A}) = p(\mathbf{A})$. First, let us observe from previous sections that only the eigenvalues of \mathbf{A} are actually important for matrix polynomials. Also recall that every

matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ with spectrum $\{\lambda_1, \dots, \lambda_n\}$ has a minimal polynomial $\phi_{\mathbf{A}}$ given by

$$\phi_{\mathbf{A}}(t) := \prod_{i=1}^k (t - \lambda_k)^{n_k} \quad (2.19)$$

which is the unique monic minimal degree ($\deg(\phi_{\mathbf{A}}) = n_1 + \dots + n_k \leq n$) polynomial such that $\phi_{\mathbf{A}}(\mathbf{A}) = 0$.

Theorem 5. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a matrix with eigenvalues $\{\lambda_1, \dots, \lambda_k\}$, and minimal polynomial given by equation 2.19. Then for any two polynomials p_1, p_2 we have that $p_1(\mathbf{A}) = p_2(\mathbf{A})$ if and only if*

$$\forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\} : p_1^{(i)}(\lambda_j) = p_2^{(i)}(\lambda_j).$$

Proof. Let $\mathbf{A} \in \mathbb{C}^{n \times n}$, with spectrum $\{\lambda_1, \dots, \lambda_k\}$, and two polynomials p_1 and p_2 such that $p_1(\mathbf{A}) = p_2(\mathbf{A})$. Let us define q such that

$$q := p_1 - p_2$$

Then, $q(\mathbf{A}) = 0$, and is thus divisible by $\phi_{\mathbf{A}}$, meaning that

$$\forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\}, q(\lambda_j) = 0 \Rightarrow p_1^{(i)}(\lambda_j) = p_2^{(i)}(\lambda_j)$$

Similarly, consider two polynomials p_1 and p_2 such that

$$\forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\}, p_1^{(i)}(\lambda_j) = p_2^{(i)}(\lambda_j)$$

For $i = 0$, $q := p_1 - p_2 = 0$ on the spectrum of \mathbf{A} , and is then divisible by $\phi_{\mathbf{A}}$. Then

$$q = K\phi_{\mathbf{A}}$$

with K a polynomial. Then, $q(\mathbf{A}) = K(\mathbf{A})\phi_{\mathbf{A}}(\mathbf{A}) = 0$ since by definition, $\phi_{\mathbf{A}}(\mathbf{A}) = 0$. And thus, $p_1(\mathbf{A}) = p_2(\mathbf{A})$.

From this reasoning, we conclude that

$$\begin{aligned} & p_1(\mathbf{A}) = p_2(\mathbf{A}) \\ \Leftrightarrow & \forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\} : p_1^{(i)}(\lambda_j) = p_2^{(i)}(\lambda_j) \end{aligned}$$

□

In theorem 5, the conditions involve the evaluation of the polynomials p_1 and p_2 and their derivatives up to order $n_k - 1$ at the eigenvalues of A . However, when the spectrum of A is simple, each eigenvalue λ_j is of multiplicity $n_j = 1$. This means that there are no higher order terms corresponding to these eigenvalues in the minimal polynomial, or in other words, there are no repeated roots. Consequently, there is no need to consider the derivatives of the polynomials p_1 and p_2 because there are no repeated roots for the polynomials to “match up” with. Therefore, in this simpler case, we only need to check that the polynomials p_1 and p_2 agree at the eigenvalues of A . In formal terms, the condition becomes:

Corollary 5.1. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a matrix with eigenvalues $\{\lambda_1, \dots, \lambda_k\}$, and minimal polynomial given by equation 2.19. If \mathbf{A} has simple spectrum, then for any two polynomials p_1, p_2 we have that $p_1(\mathbf{A}) = p_2(\mathbf{A})$ if and only if*

$$\forall j \in \{1, \dots, k\} : p_1(\lambda_j) = p_2(\lambda_j).$$

From this corollary, and from theorem 5, we can confirm our earlier statement : only the spectrum of \mathbf{A} is important for matrix polynomials. More importantly, we observe that $p(\mathbf{A})$ is uniquely defined by its values on the spectrum of \mathbf{A} . It seems then natural to extend this definition to any function f .

Definition 1. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a matrix with minimal polynomial given as in equation 2.19, and let f be a function that is at least $\max_k \{n_k - 1\}$ times differentiable. Say p is its (n_1, \dots, n_k) -Hermite interpolant i.e. the polynomial satisfying*

$$\forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\} : p^{(i)}(\lambda_j) = f^{(i)}(\lambda_j)$$

of minimal degree. Then we define $f(\mathbf{A}) = p(\mathbf{A})$.

3 The matrix-vector product $f(\mathbf{A})\mathbf{b}$

3.1 Introduction

To motivate the need for a matrix-vector product, we will consider a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$. Let us consider the matrix exponential, such as :

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} \tag{3.20}$$

Not only does this computation is very heavy (see section 5.1 for computation strategies), but it also affects the structure of the matrix. Say for example, we have the following laplacian matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & 0 & \dots & 0 & 0 & -1 & 2 \end{bmatrix}$$

Obviously, \mathbf{A} is a sparse matrix, and also a rank-structured one : it is a tridiagonal matrix. However, by computing $e^{\mathbf{A}}$, one will get a dense matrix. Storing a dense matrix often becomes challenging as the dimension of the problem grows. Besides storing, computing such a dense matrix is also an issue. However, the matrix-vector product $f(\mathbf{A})\mathbf{b}$ can be stored and computed efficiently.

3.2 The Method

As motivated by 3.1, when encountering specific structure in \mathbf{A} , such as sparsity, there is a lot to gain if we can store $f(\mathbf{A})\mathbf{b}$. In this section we will work towards a way to evaluate $f(\mathbf{A})\mathbf{b}$ in an intuitive way, thanks to Arnoldi method.

3.2.1 Formal Definitions

Firstly, we will define the notion of $\phi_{\mathbf{A},\mathbf{b}}$, i.e. the minimal polynomial of \mathbf{A} with respect to the vector \mathbf{b} . This is simply the polynomial

$$\phi_{\mathbf{A},\mathbf{b}}(t) := \prod_{i=1}^k (t - \lambda_i)^{m_k} \quad (3.21)$$

of minimal degree such that $\phi_{\mathbf{A},\mathbf{b}}(\mathbf{A})\mathbf{b} = \mathbf{0}$. Here $\lambda_1, \dots, \lambda_k$ are again the eigenvalues of \mathbf{A} .

Lemma 2. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a matrix with eigenvalues $\{\lambda_1, \dots, \lambda_k\}$, and minimal polynomial given by equation 2.19. Then for any two polynomials p_1, p_2 we have that $p_1(\mathbf{A})\mathbf{b} = p_2(\mathbf{A})\mathbf{b}$ if and only if*

$$\forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\} : p_1^{(i)}(\lambda_j) = p_2^{(i)}(\lambda_j).$$

Proof. From theorem 5, we know that

$$\begin{aligned} p_1(\mathbf{A}) &= p_2(\mathbf{A}) \\ \Leftrightarrow \forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\} : p_1^{(i)}(\lambda_j) &= p_2^{(i)}(\lambda_j) \end{aligned}$$

We also know that

$$p_1(\mathbf{A}) = p_2(\mathbf{A}) \Leftrightarrow p_1(\mathbf{A})\mathbf{b} = p_2(\mathbf{A})\mathbf{b}$$

Thus, we conclude that

$$\begin{aligned} p_1(\mathbf{A})\mathbf{b} &= p_2(\mathbf{A})\mathbf{b} \\ \Leftrightarrow \forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, n_k - 1\} : p_1^{(i)}(\lambda_j) &= p_2^{(i)}(\lambda_j) \end{aligned}$$

□

Theorem 6. *Let f be a sufficiently differentiable function that has no singularities on the spectrum of a given matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$. Then, with p the unique Hermite interpolating polynomial of \mathbf{A} w.r.t. \mathbf{b} i.e.*

$$\forall j \in \{1, \dots, k\} : \forall i \in \{0, \dots, m_k - 1\} : p^{(i)}(\lambda_j) = f^{(i)}(\lambda_j)$$

we have that $f(\mathbf{A})\mathbf{b} = p(\mathbf{A})\mathbf{b}$.

Proof. Let f be a function that is at least $\max_k\{m_k - 1\}$ times differentiable, and let p_1 be its (m_1, \dots, m_k) -Hermite interpolant. Then, by definition 1, we have that

$$f(\mathbf{A}) = p_1(\mathbf{A})$$

Then, by lemma 2, let us define p_2 such that

$$p_2(\mathbf{A})\mathbf{b} = p_1(\mathbf{A})\mathbf{b}$$

Then, we have that

$$p_1(\mathbf{A})\mathbf{b} = p_2(\mathbf{A})\mathbf{b} = f(\mathbf{A})\mathbf{b}$$

□

3.2.2 The Arnoldi Method

The Arnoldi method is a reduction method that allows low-rank approximation of a given matrix \mathbf{A} . It is based on the Hessenberg reduction of a matrix \mathbf{A} . More formally

Definition 2. Let $\mathbf{A} \in \mathbb{C}^{n \times n}$, the Arnoldi method approximates its Hessenberg reduction given by

$$\mathbf{Q}^* \mathbf{A} \mathbf{Q} = \mathbf{H} \quad (3.22)$$

where $\mathbf{Q} \in \mathbb{C}^{n \times n}$ is unitary, and $\mathbf{H} \in \mathbb{C}^{n \times n}$ is upper Hessenberg. As it is a (low-rank) approximation, Arnoldi will compute the following decomposition

$$\mathbf{Q}_k^* \mathbf{A} \mathbf{Q}_k = \mathbf{H}_k \quad (3.23)$$

with $\mathbf{Q}_k \in \mathbb{C}^{n \times k}$ unitary, and $\mathbf{H}_k \in \mathbb{C}^{k \times k}$ upper Hessenberg. That means that \mathbf{H}_k is the orthogonal projection of \mathbf{A} onto the Krylov subspace $\mathcal{K}_k(\mathbf{A}, \mathbf{q}_1)$, with \mathbf{q}_1 the first column of \mathbf{Q} .

4 Algorithms

In this section, we will implement basic algorithms. From the previous section, we will distinguish two cases :

- the dense case, where $f(\mathbf{A})$ is to be computed
- and the case where only the matrix-vector product $f(\mathbf{A})\mathbf{b}$ is to be computed

4.1 Dense case

5 Applications

5.1 Matrix Exponential

Consider the simple system of ODEs

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x} \quad (5.24)$$

with initial condition $\mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n$. Then we know the solution to be given by $\mathbf{x}(t) = e^{At}\mathbf{x}_0$. However, for all but the stablest systems, this is not a good method, due to issues such as stability and stiffness.

References

Frommer, Andreas and Valeria Simoncini (2008). “Matrix functions”. In: *Model order reduction: theory, research aspects and applications*. Springer, pp. 275–303.