# Numerical linear algebra and functions of matrices

Parts of a lecture given in the "Master 2 de Mathématiques Appliquées, Lille–Littoral–Valenciennes"
by Bernhard Beckermann, http://math.univ-lille1.fr/∼bbecker
March 2013

## Table of contents



## 1 Introduction and organization of this text

Matrix functions are useful tools not only in applied mathematics and scientific computing, but also in various other fields like control theory, elecromagnetism or the research on complex networks like social networks. Over the last decade one can observe a very important research activity in the field of matrix functions $f(A)$ or $f(A)b$ with $A \in \mathbb{C}^{n\times n}$ with spectrum $\sigma(A)$, and $b \in \mathbb{C}^n$, using tools both from numerical linear algebra and approximation theory. As an example, for $f(z) = 1/z$ we solve a system of linear equations, but also $f(z) = \exp(z)$, $f(z) = \log(z)$ and other functions play an important role in applications, such as the solution of a system of ODEs obtained for instance through a discretization in the spacial variable of a partial differential equation.

### 1.1 Organization of this text

One fascinating aspect of matrix functions is that it combines several fields of research

- Basic definitions and applications : §2, §3
- Links between norms of matrix functions and norms of functions : §4
- Polynomial and rational approximation of a function of a real or complex variable : §5, §6
- Algorithmic aspects from numerical linear algebra : §5, §7, §8.

This text has been extracted from some lecture notes of a lecture at the University of Lille. Those who want to have a quick look should scan through the basic definitions and properties

of §2, and maybe have a quick look at the applications mentioned in §3. The rest of the text is written for those readers who desire a more detailed introduction to matrix functions before coming to the spring school, and only §2 is required to read any of the other above-mentioned three parts. The interested readers might want to try to solve some of the suggested exercises, have a look at proofs (which are included only if they help for understanding), and/or use one of the references below for further reading.

Here is a more detailed summary: In §2 we give a proper definition of matrix functions based on the Jordan normal form and some other formulations trough matrix polynomials and the Cauchy formula, together with elementary properties of matrix functions. A small number of applications for particular matrix functions are enumerated in §3.
One method for computing approximately $f(A)$ is to compute $g(A)$ for a "simpler" function $g$ "close" to $f$. We give in §4 some general tools to relate $\|f(A)-g(A)\|$ to the supremum of $f-g$ on some subset of the complex plane. This enables us to construct in §5 "good" polynomial and rational approximants of $f$ and $f(A)$. In §6 we give (in french) some tools for best polynomial or rational approximation of analytic functions in the complex plane. This section on complex approximation is probably far more advanced compared to the rest of the text, and can be omitted without any harm.
In what follows we shortly expose two different linear algebra techniques for computing matrix functions. The direct Schur-Parlett method for computing matrix functions (exactly up to rounding errors) is exposed in §7. Finally, we give in §8 some basics on (rational) Krylov techniques for approaching $f(A)b$, which will be further discussed in several lectures of the spring school.

### 1.2 Further reading

These lecture notes follow closely the reference [6]. An interested reader can also find some complements in the books [12], [13] et [15] for general aspects on numerical linear algebra, [4] on Krylov methods, [8] on rational approximation, [1] for aspects on linear control, and finally [14] for the pseudo spectrum.

## 2 Three ways of defining matrix functions

Whereas it is easy to define the polynomial of a matrix or an entire function as the exponential, things become more tricky for functions defined only on subsets of the complex plane. Following [6, Section 1] we will here go through the Jordan normal form.

**2.1. The Jordan normal form:** *Each $A \in \mathbb{C}^{n\times n}$ is similar to a matrix in Jordan normal form : $\exists Z \in \mathbb{C}^{n\times n}$ invertible such that $Z^{-1}AZ = J = \mathrm{diag}(J_1, ..., J_p)$ block diagonal, with*

$$J_k = J_{m_k}(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & 0 & \cdots & 0 \\ 0 & \lambda_k & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \lambda_k & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k\times m_k}$$

*called* **Jordan block** *and $m_1 + .... + m_p = n$. Thus $\sigma(A) = \{\lambda_1, ..., \lambda_p\}$, where the $\lambda_j$ are not necessarily distinct. We call* **index** *$m(\lambda)$ of an eigenvalue $\lambda$ the size of the largest Jordan block associated to $\lambda$.*
**Special case:** *if the index of all eigenvalues is equal to one (or $m_1 = ... = m_p = 1$) we say that $A$ is* **diagonalizable**. *Here the columns of $Z$ are the corresponding eigenvectors.*

**Special case:** *if $A$ is hermitian ($A = A^*$) or normal ($AA^* = A^*A$) then it is diagonalizable, and one can chose a unitary $Z$ ($Z^*Z = I_n$, in other words, an orthonormal basis of eigenvectors).*

Notice that the number of blocks and their size as well as the index of each eigenvalue are invariants of $A$, but by no means the Jordan normal form $A = ZJZ^{-1}$ is unique.

**2.2. Definition of a function of a matrix:** *Let $f$ be a function defined on the spectrum of $A$, meaning that $\forall \lambda \in \sigma(A)$ of index $m(\lambda)$ we know $f^{(j)}(\lambda)$ for $j = 0, 1, ..., m(\lambda) - 1$. Then*

$$f(A) = Z \operatorname{diag}(f(J_1), ..., f(J_p))Z^{-1}, \quad f(J_m(\lambda)) = \begin{bmatrix} \frac{f(\lambda)}{0!} & \frac{f'(\lambda)}{1!} & \cdots & \frac{f^{(m-1)}(\lambda)}{(m-1)!} \\ 0 & \frac{f(\lambda)}{0!} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{f'(\lambda)}{1!} \\ 0 & \cdots & 0 & \frac{f(\lambda)}{0!} \end{bmatrix}.$$

In particular, if $A$ is diagonalizable, $Z^{-1}AZ = \operatorname{diag}(\lambda_1, ..., \lambda_n)$, then $f(A) = Z \operatorname{diag}(f(\lambda_1), ..., f(\lambda_n))Z^{-1}$. Clearly, the drawback of Definition 2.2 is that one could believe that the value of a function of a matrix depends on the particular choice of the Jordan normal form. It will follow implicitly from Corollary 2.5 below that this is not true. But, until then, let us always take the same normal form for a given matrix $A$.

**2.3. Lemma on elementary operations for matrix functions for the same matrix:**

*(a)* $(f + g)(A) = f(A) + g(A)$ *(matrix sum).*

*(b)* $(f \cdot g)(A) = f(A) \cdot g(A) = g(A) \cdot f(A)$ *(matrix product).*

*(c) For* $f(z) = \alpha \in \mathbb{C}$ *we have* $f(A) = \alpha I_n$ *(matrix identity).*

*(d) If* $f(z) = \frac{1}{\alpha - z}$ *for some* $\alpha \in \mathbb{C} \setminus \sigma(A)$ *then* $f(z) = (\alpha I_n - A)^{-1}$, *the* **resolvent** *in* $\alpha$ *(matrix inversion).*

We get from 2.3 that polynomials of matrices are evaluated as expected

$$A^0 = I_n, \quad \text{and for an integer } \ell > 0: \quad A^\ell = A \cdot A^{\ell-1} = \underbrace{A \cdot A \cdot ... \cdot A}_{\ell \text{ times.}}$$

Thus for a rational function with poles $\notin \sigma(A)$

$$r(z) = \frac{a_0 + a_1 z + ... + a_j z^j}{b_0 + b_1 z + ... + b_k z^k} = c \frac{(z - x_1)...(z - x_j)}{(z - y_1)...(z - y_k)}$$

we get

$$\begin{aligned} r(A) &= (a_0 I_n + a_1 A + ... + a_j A^j)(b_0 I_n + b_1 A + ... + b_k A^k)^{-1} \\ &= c(A - x_1 I_n)...(A - x_j I_n)(A - y_1 I_n)^{-1}...(A - y_k I_n)^{-1} \end{aligned}$$

where we notice that any two factors permute. Thus the value of a polynomial or a rational function of a matrix does not depend on the particular choice of the Jordan normal form.

**2.4. Exercice :**

*(a) For the characteristic polynomial* $\chi(\lambda) = \det(\lambda I_n - A)$, *show that* $\chi(A) = 0 (\in \mathbb{C}^{n \times n})$.

*(b) Show that there exists a unique monic polynomial* $\psi$ *of minimal degree such that* $\psi(A) = 0$ *(called* **minimal polynomial***). Verify the formula* $\psi(z) = \prod_{j=1}^{s} (z - \lambda_{k_j})^{m(\lambda_{k_j})}$ *with* $\lambda_{k_1}, ..., \lambda_{k_s}$ *the distinct eigenvalues of* $A$.

The following results shows that, for computing matrix functions, at least in theory it is sufficient to know to evaluate a polynomial of a matrix.

**2.5. Corollary on the representation through polynomials :** *If $p$ interpolates $f$ on $\sigma(A)$ (in the Hermite sense)*

$$\forall \lambda \in \sigma(A) \quad \forall j = 0, 1, ..., m(\lambda) - 1 : \quad f^{(j)}(\lambda) = p^{(j)}(\lambda)$$

*then $f(A) = p(A)$. There exists a unique such $p$ (called the interpolation polynomial of $(f, A)$) with $\deg p < \deg \psi$ and $\psi$ the minimal polynomial of $A$.*

**2.6. Exercise :**

*(a) Show that $f(A^*) = (f(A))^*$ if $f(\overline{z}) = \overline{f(z)}$.*

*(b) Show that $f(XAX^{-1}) = Xf(A))X^{-1}$.*

*(c) Show that if $X$ permutes with $A$ then also with $f(A)$.*

*(d) Show that $f(\operatorname{diag}(A, B)) = \operatorname{diag}(f(A), f(B))$.*

*(e) Let $A, B \in \mathbb{C}^{n\times n}$, $A$ invertible. Show that $AB$ and $BA$ have the same Jordan blocks. Conclude that $Af(BA) = f(AB)A$* (first for a polynomial $f$).

*(f)* Let $g$ be defined on the spectrum of $A$ et suppose that $f^{(j)}(g(\lambda))$ exists for all $\lambda \in \sigma(A)$ et $j = 0, 1, ..., m(\lambda) - 1$. Show that $(f \circ g)(A) = f(g(A))$ (*in comparing the size of the Jordan blocks of $A$ and $g(A)$, replace $f$ by an appropriate polynomial.*).*

**2.7. Exercise :** *The DFT matrix of order $n$ is defined by $F_n = \frac{1}{\sqrt{n}}(\exp(-2\pi i \frac{jk}{n}))_{j,k=0,1,...,n-1}$. Show that $F_n$ is unitary, complex symmetric, and that $F_n^4 = I_n$. Deduce explicitly* $\exp(\pi F_n)$.

**2.8. Exercise :** *Show that an upper block triangular matrix is factorizable as follows*

$$M = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} = \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$$

*if and only if $X$ is solution of the Sylvester equation $AX - XB = C$. In this case, show that*

$$f(M) = \begin{bmatrix} f(A) & f(A)X - Xf(B) \\ 0 & f(B) \end{bmatrix}.$$

**2.9. Exercise :** *Consider the following non degenerate bidiagonal matrix*

$$M = \begin{bmatrix} \lambda_1 & d_1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & d_{n-1} \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

*with $d_j \neq 0$, and $D = \operatorname{diag}(1, d_1, d_1d_2, ..., d_1...d_{n-1})$. Show that*

$$f(M) = D^{-1} \begin{bmatrix} f[\lambda_1] & f[\lambda_1, \lambda_2] & \cdots & f[\lambda_1, ..., \lambda_n] \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & f[\lambda_{n-1}] & f[\lambda_{n-1}, \lambda_n] \\ 0 & \cdots & 0 & f[\lambda_n] \end{bmatrix} D$$

*with $f[\lambda_j, ..., \lambda_k]$ a divided difference.*

**2.10. Exercise :** *For $X \in \mathbb{C}^{n\times r}$, $Y \in \mathbb{C}^{r\times n}$, $YX$ of rank $r$, $\alpha \in \mathbb{C}$, use the identity*

$$\begin{bmatrix} \alpha I_n & X \\ 0 & \alpha I_r + YX \end{bmatrix} \begin{bmatrix} I & 0 \\ Y & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ Y & I \end{bmatrix} \begin{bmatrix} \alpha I_n + XY & X \\ 0 & \alpha I_r \end{bmatrix}$$

*for showing that*

$$f(\alpha I_n + XY) = f(\alpha) I_n + X(YX)^{-1}[f(\alpha I_r + YX) - f(\alpha) I_r]Y,$$

*and* $(I_n + XY)^{-1} = I_n - X(I_r + YX)^{-1}Y$.

There is a last formula for computing matrix functions based on the Cauchy integral formula which is particularly useful for error estimates, which can be generalized for functions of Hilbert space operators.

**2.11. Theorem on Cauchy formula :** *Let $f$ be analytic on some open $\Omega \subset \mathbb{C}$, and $\Gamma \subset \Omega$ a system of Jordan curves encircling each $\lambda \in \sigma(A)$ exactly one time, with mathematically positive orientation, then*

$$f(A) = \frac{1}{2\pi i} \int_\Gamma f(\zeta)(\zeta I_n - A)^{-1}\, d\zeta.$$

*Proof.* According to Lemma 2.3, it is sufficient to show this formula for a Jordan block. We observe that

$$(\zeta I_n - J_m(\lambda))^{-1} = \begin{bmatrix} \frac{1}{\zeta-\lambda} & \frac{1}{(\zeta-\lambda)^2} & \cdots & \frac{1}{(\zeta-\lambda)^m} \\ 0 & \frac{1}{\zeta-\lambda} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{(\zeta-\lambda)^2} \\ 0 & \cdots & 0 & \frac{1}{\zeta-\lambda} \end{bmatrix},$$

and the claimed formula follows by comparing element by element using Cauchy's integral formula for a function and its derivatives. □

By comparing both integral representations, we also immediately obtain the following.

**2.12. Corollary, series expansion :** *If $f(z) = \sum_{j=0}^{\infty} a_j z^j$ admits a convergence radius $R > \rho(A) := \max\{|\lambda| : \lambda \in \sigma(A)\}$ (the spectral radius of $A$), then $f(A) = \sum_{j=0}^{\infty} a_j A^j$* (convergence in norm, that is, $\|f(A) - \sum_{j=0}^{k} a_j A^j\| \to 0$ pour $k \to \infty$).

By replacing $A$ by $A - z_0 I_n$, on may also obtain similar results for expansions around any $z_0 \neq 0$.

We should warn the reader that

| no method exposed in §2 should be used like a black box procedure |
|---|

for computing matrix functions for the following reasons:

- the size of a Jordan block is not stable under perturbations (unless the matrix is diagonalizable), thus Definition 2.2 does not have a counterpart in finite precision arithmetic,
- the approach 2.5 of evaluating $P(A)$ for $P$ the interpolation polynomial of $(f, A)$ suffers from the same drawback even for clustering eigenvalues. In addition, one should know how to efficiently evaluate $P(A)$ in a numerically stable manner, see §5. To fix ideas, for diagonalizable $A$ with distinct eigenvalues, $A = Z\,\mathrm{diag}(\lambda_1, ..., \lambda_n)Z^{-1}$, $Z = (y_1, ..., y_n)$, $Z^{-*} = (\tilde{y}_1, ..., \tilde{y}_n)$, we have that

$$P(z) = \sum_{j=1}^{n} \ell_j(z) f(\lambda_j), \quad \ell_j(z) = \prod_{k\neq j} \frac{z - \lambda_k}{\lambda_j - \lambda_k},$$

and thus $\ell_j(A) = Z\,\mathrm{diag}(0, ..., 0, 1, 0, ..., 0)Z^{-1} = y_j \tilde{y}_j^*$, which gives exactly the formula $P(A) = \sum_j P(\lambda_j) y_j \tilde{y}_j^*$ which we have seen already in 2.2,

- The approach 2.12 seems to give us a "simple" method of computing $\exp(A), \cos(A), \cos(\sqrt{A})$, but it could be quite sensitive to finite precision arithmetic and in particular cancellations, see [10] and and the discussion at the end of [6, Section 4.2 and Section 4.3],
- If one wants to define $\log(A)$, $\sqrt{A}$ and similar multi-valued functions, one has first to choose correctly in 2.11 the set $\Omega$. For this type of functions, in general $\Omega = \mathbb{C} \setminus (-\infty, 0]$, and thus one has to exclude matrices with eigenvalues $< 0$. Also, one should know how to select the contour $\Gamma$... Subsequently, additional errors are introduced by applying quadrature formulas, see, e.g., [5].

# 3 Motivation: some particular matrix functions and applications

A nice exposition about various applications of matrix functions for different tasks of scientific computing can be found in [6, Chapter 2], other examples will be given in the lectures of the Lille spring school. The aim of this section is just to summarize some few basic ones.

**3.1. Example: fractional powers**
*For $\gamma \in \mathbb{R} \setminus \mathbb{Z}$, one usually defines (the principal branch of) $f(z) = z^\gamma$ on $\Omega = \mathbb{C} \setminus (-\infty, 0]$ via polar coordinates: if $z = re^{i\phi}$ with $r > 0, \phi \in (-\pi, \pi)$ then $f(z) = r^\gamma e^{i\gamma\phi} = \exp(\gamma(\log(r) + i\phi))$, implying that $f$ is (single-valued and) analytic in $\Omega$.*
*For $-1 < \gamma < 0$, the function $f(z) = z^\gamma$ admits a representation like a* **Markov function**

$$f(z) = \int_a^b \frac{d\mu(x)}{z - x}, \quad -\infty \le a < b \le \infty, \quad \mu \text{ a positive measure on } [a, b],$$

*here $z^\gamma = \frac{\sin(|\gamma|)}{\pi} \int_{-\infty}^0 \frac{|x|^\gamma}{z-x} dx$. E.g., $z^{-1/2} = \int_{-\infty}^0 \frac{1}{\pi\sqrt{|x|}} \frac{dx}{z-x}$, and $z^{1/2}$ could be considered as $z$ times the Markov function $z^{-1/2}$.*
*Beside splitting techniques, another striking example for the usefulness of matrix square roots $A^{1/2}$ is to form a "geometric mean" between two hermitian positive definite (hpd) $A, B$ of order $n$: We define $X = A\#B$ to be the unique solution of the matrix equation $XA^{-1}X = B$, and get*[1] *the formula $A\#B = B^{1/2}(B^{-1/2}AB^{-1/2})^{1/2}B^{1/2}$. The geometric matrix mean is attractive for its many (partly non-trivial) properties such as commutativity, or for instance the equivalence being true for any hermitian $X$*

$$\begin{bmatrix} A & X \\ X & B \end{bmatrix} \text{ hpd} \iff A\#B - X \text{ hpd.}$$

*See [6, Chapter 2.2] and the references therein for more details about matrix means, and [6, Chapters 6-7] for commputational aspects for fractional powers of matrices.*

**3.2. Example: the logarithm**
*$f(z) = \log(z)$ is also defined (and analytic) on $\Omega = \mathbb{C} \setminus (-\infty, 0]$ via polar coordinates: $\log(re^{i\phi}) = \log(r) + i\phi$, $r > 0, \phi \in (-\pi, \pi)$, and again we obtain a modification of a Markov function*

$$\frac{\log(1+z)}{z} = \int_{-\infty}^{-1} \frac{1}{|x|} \frac{1}{z-x}.$$

*The logarithm can be helpful to compute determinants since*

$$\det(A) = \prod_{j=1}^n \lambda_j = \exp(\sum_{j=1}^n \log(\lambda_j)) = \exp(\mathit{trace}(\log(A))),$$

[1] This formula can be verified by setting $X = B^{1/2} Z B^{1/2}$.

*but it does also occur in the study of certain (stationary) Markov chains of order n [6, Chapter 2.3]: the entry $(i,j)$ of a* **transition matrix** *$P(t)$ does indicate the probability $\in [0,1]$ that an object in state $i$ at time $s$ passes to the state $j$ at time $t+s$ (independent of $s$). As a consequence, $P(t)$ is stochastic, meaning that it contains only elements $\in [0,1]$, $1 \in \sigma(P(t))$ is the largest eigenvalue, and the corresponding right eigenvector is $(1,1,...,1)^T$. Such a transition matrix has the semi-group property $P(t)P(s) = P(s+t)$ for all $s,t \geq 0$, implying that $P(t) = \exp(Qt)$, with[2] the generator $Q = \log(P(1))$. For computational aspects for the matrix logarithm see [6, Chapter 11].*

**3.3. Example: The sign function**

*The sign function is defined on $\Omega = \mathbb{C} \setminus (i\mathbb{R})$ by*

$$sign(z) = \begin{cases} +1 & \text{if } Re(z) > 0, \\ -1 & \text{if } Re(z) < 0, \end{cases}$$

*and thus $sign(z) = z(z^2)^{-1/2}$. If $A$ is diagonalizable, with no imaginary eigenvalues, then $f(A)$ with $f(z) = (1 + sign(z))/2$ can be easily seen to be a projector onto the set generated by the eigenvectors with eigenvalues $Re(\lambda) > 0$.*

*In QCD (lattice quantum chromodynamics) one requires to solve large sparse systems $Ax = b$ (via Krylov methods requiring an efficient black box matrix-vector multiplication with $A$) where $A = \mathrm{diag}(\pm 1) - sign(H)$ and $H$ a large sparse hermitian matrix, see [6, Chapter 2.7] and the references therein. This a very prominent example where it is not feasible to compute or even to store $sign(H)$ in order to implement such a matrix-vector product.*

*The sign function is also useful for writing the solution of certain Syvester matrix equations $AX - XB = C$ in the unknown $X$: suppose that the eigenvalues of $A$ (and of $B$) have a strictly positive real part (and strictly negative, respectively), such that $sign(A) = I$, $sign(B) = -I$. Then according to 2.8*

$$sign\left(\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}\right) = \begin{bmatrix} I & 2X \\ 0 & -I \end{bmatrix}.$$

*We should mention that the above assuptions on $A, B$ are common in linear control theory, in particular in the the context $B = -A^*$ of Lyapunov equations. One may find in [6, Section 2.4] a similar approch for the algebraic Riccati equation, and in [6, Chapter 5] more about computational aspects for the matrix sign function.*

**3.4. Exercise:**

*Let $A, B \in \mathbb{C}^{n \times n}$, with $\sigma(AB)$ having an empty intersection with $(-\infty, 0]$. Then with $C = A(BA)^{-1/2} = (AB)^{1/2}B^{-1}$*

$$sign\left(\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & C \\ C^{-1} & 0 \end{bmatrix}.$$

**3.5. Example: Systems of ordinary differential equations and the exponential function**

*Starting from a parabolic PDE (such as the heat equation)*

$$\frac{\partial}{\partial t} w(x,t) + Lw(x,t) = f(x,t), \quad x \in \Omega$$

*plus (say, homogeneous) boundary conditions for all $t$ and $x \in \partial\Omega$ plus initial conditions for $t = 0$ and $x \in \Omega$, with $L$ a space differential operator, we can obtain a system of ordinary differential equations via a space discretization: given a finite element linear space $V_h$ of dimension $n$ with basis $v_{1,h}(x), v_{2,h}(x), ..., v_{n,h}(x)$, we look for $w_h(x,t) = \sum_{j=1}^n y_j(t) v_{j,h}(x)$ such that*

$$\forall v \in V_h: \quad \langle \frac{\partial}{\partial t} w_h + Lw_h - f, v \rangle = 0, \quad \langle w_h(x,0) - w(x,0), v \rangle = 0.$$

[2] For knowing the week transition it could be more efficient to compute $P(1/52) = P^{1/52}$, see 3.1.

*Introducing the mass and stiffness matrices*[3]

$$M = (\langle v_{j,h}, v_{k,h}\rangle)_{j,k}, \quad K = (\langle Lv_{j,h}, v_{k,h}\rangle)_{j,k}, \quad F_f(t) = (\langle f, v_{k,h}\rangle)_k, \quad y(t) = (y_j(t))_j,$$

*we are left with the system* $M\dot{y}(t) + Ky(t) = F_f(t)$, $My(0) = F_{w(x,0)}$. *By definition, the stiffness matrix is hpd, and introducing the Cholesky decomposition* $M = CC^T$ *and the variable* $x(t) = C^T y(t)$, $A = -C^{-1}KC^{-T}$ *we are left with*

$$\dot{x}(t) = Ax(t) + g(t), \quad x(0) \in \mathbb{R}^n \quad \textit{given.}$$

*For* $g(t) = 0$ *(homogeneous first order differential equation with constant coefficients) we find the solution* $y(t) = \exp(At)y(0)$, *and for general* $g$ *by variation of constants*

$$x(t) = \exp(At)x(0) + \int_0^t \exp(A(t-s))g(s)\,ds.$$

*In particular, if* $g(t) = g(0)$ *does not depend on* $t$, *then (by expansion in series)*

$$x(t) = \exp(At)x(0) + t\varphi_1(At)g(0), \quad \varphi_1(z) := \frac{\exp(z) - 1}{z}.$$

*For systems of ODE of second order we can either rewrite our problem as a first order system of double size, or write the solution in terms of trigonometric functions.*

**3.6. Example: Link with linear control theory**
*A continuous stationary linear dynamic system with* $m$ *entries and* $p$ *outputs (short MIMO) consists of finding the output* $y \in \mathcal{C}([0,+\infty);\mathbb{C}^p)$ *corresponding to the input* $u \in \mathcal{C}([0,+\infty);\mathbb{C}^m)$ *via*

$$y(t) = Cx(t), \quad \dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = 0,$$

*with the state variable* $x \in \mathcal{C}([0,+\infty);\mathbb{C}^n)$, *and thus* $C \in \mathbb{C}^{p\times n}$, $B \in \mathbb{C}^{n\times m}$, $A \in \mathbb{C}^{n\times n}$. *Compared with the heat equation in 3.5, one would for instance like to know (approximately) the mean of the temperature as a function of* $t$ *depending on some part of the initial conditions.*
*Introducing the one-sided Laplace transform (any function defined on* $[0,\infty)$ *(the futur) is continued by* $0$ *on* $(-\infty, 0]$ *(the past))*

$$\mathcal{L}(x)(s) = \int_0^\infty \exp(-st)x(t)\,ds,$$

*one observes by integration by parts that* $\mathcal{L}(\dot{x})(s) = s\mathcal{L}(x)(s)$, *and thus* $\mathcal{L}(x)(s) = (sI - A)^{-1}B\mathcal{L}(u)(s)$ *or* $\mathcal{L}(y)(s) = C(sI - A)^{-1}B\mathcal{L}(u)(s)$, *with the* $p \times m$*–valued* **transfer function** $R_n(s) = C(sI - A)^{-1}B$. *For instance, in realistic simulations of integrated circuits via RLC circuits one easily finds that* $n \geq 10^6$, *and thus one of the aims of linear control, called model reduction, is to replace*

$$\begin{bmatrix} A & B \\ C & 0^{p\times m} \end{bmatrix} \in \mathbb{C}^{(n+p)\times(n+m)} \quad by \quad \begin{bmatrix} \widetilde{A} & \widetilde{B} \\ \widetilde{C} & 0^{p\times m} \end{bmatrix} \in \mathbb{C}^{(\widetilde{n}+p)\times(\widetilde{n}+m)}$$

*such that* $\widetilde{n} \ll n$, *and the output* $\widetilde{y}(t)$ *of the new system for the input* $u(t)$ *is not too far from the original output* $y(t)$. *Applying the Plancherel equality we find that* $\|\mathcal{L}(y)\|_{L^2(i\mathbb{R})} = \|y\|_{L^2([0,+\infty))}$, *and thus require that the difference of the two transfer functions* $R_n$ *and* $\widetilde{R}_{\widetilde{n}}$ *should be small on the imaginary axis (and, for physical reasons, with the eigenvalues of* $A$, *also all eigenvalues of* $\widetilde{A}$ *should be in the left half plane).*

[3] If as in the heat equation $L = -\Delta$ is the Laplacien then the stiffness matrix $K$ is hpd.

*We give one more ingredient, see the lecture of Paul Van Dooren: there exist numbers $\sigma_1 \geq ... \geq \sigma_n$ called Hankel singular values*[4] *such that, for any choice of $\widetilde{A}, \widetilde{B}, \widetilde{C}$ we have*

$$\sup_{\zeta \in i\mathbb{R}} |R_n(\zeta) - \widetilde{R}_{\widetilde{n}}(\zeta)| \geq \sigma_{\widetilde{n}},$$

*and Glover (1984) showed that there is some construction for $\widetilde{A}, \widetilde{B}, \widetilde{C}$ called balanced truncation such that*

$$\sup_{\zeta \in i\mathbb{R}} |R_n(\zeta) - \widetilde{R}_{\widetilde{n}}(\zeta)| \leq 2\sigma_{\widetilde{n}} + ... + 2\sigma_n.$$

*The drawback of this approach is that balanced truncation is quite costly and only feasible if $n$ is sufficiently small, and other approaches through rational Krylov spaces are much more interesting, see again the lecture of Paul Van Dooren. In any case, the link with matrix functions is very fruitful, for expressing for instance the energy of a linear dynamical system.*

**3.7. Example: the exponential and the Sylvester equation**
*Let $\sigma(A), \sigma(-B) \subset \{Re(z) < 0\}$, then the matrix-valued differential equation $\dot{Y}(t) = AY(t) - Y(t)B$, $Y(0) = C$ has the solution $Y(t) = \exp(At)C\exp(-Bt)$. Notice that $\lim_{t\to\infty} Y(t) = 0$. Thus setting $X = -\int_0^\infty \exp(At)C\exp(-Bt)\,dt$ we find that*

$$AX - XB = -\int_0^\infty \dot{Y}(t)\,dt = Y(0) - Y(\infty) = C,$$

*in other words, we have got an integral formula for the solution $X$ of our Sylvester equation $AX - XB = C$.*
*Notice also that*

$$\exp(sA)X - X\exp(sB) = \int_0^\infty \exp((s+t)A)C\exp(-tB)\,dt + \int_0^\infty \exp(tA)C\exp((s-t)B)\,dt$$
$$= \int_0^s \exp(tA)\exp((s-t)B)\,dt$$

*which together with 2.8 shows that*

$$\exp\left(s\begin{bmatrix} A & C \\ 0 & C \end{bmatrix}\right) = \begin{bmatrix} \exp(sA) & \int_0^s \exp(tA)\exp((s-t)B)\,dt \\ 0 & \exp(B) \end{bmatrix}.$$

# 4 The concept of K-spectral sets (Von Neumann, numerical range, pseudo-spectrum)

For being able to approach $f(A)$ by $g(A)$ for a (rational or polynomial) function $g$ "close" to $f$, the following notion is helpful.[5]

**4.1. Definition of $K$-spectral sets**
*A closed $\Omega \subset \mathbb{C}$ is called $K$-spectral for a matrix $A$ if there exists a numerical constant $K$ such that for each function $f$ analytic in a neighborhood of $\Omega$ we have*

$$\|f(A)\| \leq K\,\|f\|_\Omega, \quad \textit{with} \quad \|f\|_\Omega := \sup_{z\in\omega} |f(z)|.$$

[4] They are known to be the singular values of $X$ the solution of the Sylvester equation $AX + XA = BC$.
[5] See also "C. Badea, B. Beckermann, Spectral sets, in: L. Hogben, Handbook of Linear Algebra, second edition (2013)." and the references therein.

Notice that necessarily $\sigma(A) \subset \Omega$ for any $K$-spectral set (since otherwise consider $f_a(z) = \frac{1}{z-a}$ for $a \to \lambda \in \sigma(A) \setminus \Omega$).

**4.2. Lemma: the spectrum is $K$-spectral:**
*Let $A$ be diagonalizable, $Z^{-1}AZ$ being diagonal, then $\sigma(A)$ is $K$-spectral for $A$ with $K = cond(Z) = \|Z\| \, \|Z^{-1}\|$.*

From the example

$$f(z) = z, \quad A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \|f(A)\| = 1, \quad \|f\|_{\sigma(A)} = 0$$

we see that having a diagonalizable $A$ is essential in 4.2. Here we could only give an upper bound in terms of the derivatives of $f$, or otherwise one has to make $\Omega$ larger, which will be the object of the subsequent considerations.
Before going further, notice that if $\phi$ is analytic in a neighborhood of $\Omega$ then $\Omega$ being $K$-spectral for $A$ implies that $\phi(\Omega)$ is $K$-spectral for $\phi(A)$. Thus the following result allows to deal with half planes or complements of open disks being images of a closed disk under a homographic transformation.

**4.3. Von Neumann's Theorem**
*The disk $\Omega = \{z \in \mathbb{C} : |z - \omega| \leq R\}$ with $R \geq \|A - \omega I\|$ is 1-spectral for $A$ (e.g., the closed unit disk $\mathbb{D}$ is 1-spectral for any matrix with $\|A\| \leq 1$).*

*Proof.* We follow [11, §154] and consider only the case $R > \|A - \omega I\|$ for sufficiently small $R$, the general case following by passing to a limit. Let

$$U = \frac{A - \omega I}{R}, \quad g(z) = f(\omega + Rz)$$

such that $\|U\| < 1$, $g$ being analytic in $\mathbb{D}$ and $f(A) = g(U)$. Then following 2.11

$$\begin{aligned} f(A) &= \frac{1}{2\pi i} \int_{|w|=1} g(w)(wI - U)^{-1} dw = \frac{1}{2\pi} \int_{|w|=1} g(w)(wI - U)^{-1} w \, |dw| \\ &= \frac{1}{2\pi} \int_{|w|=1} g(w) \Big( \frac{1}{2} I + \frac{1}{2}(wI + U)(wI - U)^{-1} \Big) \, |dw| \\ &= \frac{g(0)}{2} + \frac{1}{2\pi} \int_{|w|=1} g(w) \Big( \frac{1}{2}(wI + U)(wI - U)^{-1} \Big) \, |dw|. \end{aligned}$$

Similarly, by the residual theorem (or Neumann series)

$$\begin{aligned} &\frac{1}{2\pi} \int_{|w|=1} g(w) \Big( \frac{1}{2}(wI + U)(wI - U)^{-1} \Big)^* \, |dw| \\ &\frac{1}{2\pi i} \int_{|w|=1} g(w) \Big( \frac{1}{2}(I + wU^*)(I - wU^*)^{-1} \Big) \, \frac{dw}{w} = \frac{g(0)}{2}. \end{aligned}$$

Thus

$$f(A) = \frac{1}{2\pi} \int_{|w|=1} g(w) M(w) \, |dw|, \quad M(w) = \mathrm{Re}\Big( (I + U/w)(I - U/w)^{-1} \Big),$$

with $M(w)$ hemitian positive definite since with $y = (I - U/w)\widetilde{y}$

$$y^* M(w) y = \Re(\widetilde{y}^* (I - U/w)^* (I + U/w)\widetilde{y}) = \widetilde{y}^* \widetilde{y} - \widetilde{y}^* U^* U \widetilde{y} \geq \widetilde{y}^* \widetilde{y}(1 - \|U\|^2) > 0.$$

As a consequence, for any vectors $x, y$ of norm 1 using twice Cauchy-Schwarz:

$$|x^* f(A) y| \leq \|g\|_{\mathbb{D}} \frac{1}{2\pi} \int_{|w|=1} \sqrt{x^* M(w) x} \sqrt{y^* M(w) y}\, |dw|$$
$$\leq \|g\|_{\mathbb{D}} \sqrt{\frac{1}{2\pi} \int_{|w|=1} x^* M(w) x\, |dw| \frac{1}{2\pi} \int_{|w|=1} y^* M(w) y\, |dw|} = \|f\|_\Omega \|x\| \|y\| \leq \|g\|_{\mathbb{D}},$$

as claimed above. □

Using $\phi(z) = \frac{1-z}{1+z}$, Theorem 4.3 implies the following.

**4.4. Corollary for half planes:**
*If $A$ is semi-definite positive (i.e., $Re(y^* A y) \geq 0$ for all vectors $y$), then $\Omega = \{Re(z) \geq 0\}$ is 1-spectral for $A$, in particular $\|\exp(-A)\| \leq 1$.*

Let us mention a recent generalization obtained by Michel Crouzeix in 2007.

**4.5. Definition of numerical range:**
*The numerical range (or field of values) [7] of a matrix $A \in \mathbb{C}^{n \times n}$ is given by*

$$W(A) = \Big\{ \frac{y^* A y}{y^* y} : y \in \mathbb{C}^n \setminus \{0\} \Big\}.$$

It is not difficult to verify that $W(A)$ is a compact subset of $\{|z| \leq \|A\|\}$ containing $\sigma(A)$, and that for diagonal (or normal) $A$ we obtain for $W(A)$ the convex hull of $\sigma(A)$. Concerning our preceding example,

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad W(A) = \frac{\mathbb{D}}{2}$$

being much smaller than $\mathbb{D} = \{|z| \leq \|A\|\}$.

**4.6. Lemma of convexity [9, Theorem V.3.1] :**
*$W(A)$ is convex.*

**4.7. Crouzeix's Theorem :**
*The numerical range $W(A)$ is $K$-spectral for $A$ with $K \leq 11.08$.*
*In particular, the disk centered at 0 with radius $\max\{|z| : z \in W(A)\}$ (called numerical radius) is 2-spectral for $A$.*

The above example

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad f(A) = \begin{bmatrix} 0 & f'(0) \\ 0 & 0 \end{bmatrix}, \quad W(A) = \frac{\mathbb{D}}{2}$$

shows that the numerical range constant satisfies $K \geq 2$, and it is conjectured by Crouzeix that $K = 2$. This conjecture has been shown to be true for $2 \times 2$ matrices in [2] where the numerical range has the shape of a filled ellipse.
Recall from Lemma 4.3 that for any normal matrix $A$ we have that $\|f(A)\| \leq \|f\|_{\sigma(A)} \leq \|f\|_{conv(\sigma(A))} = \|f\|_{W(A)}$. For non normal matrices, it is also possible to consider:

**4.8. Definition of the pseudo-spectrum**
*For $\epsilon > 0$, let $\sigma_\epsilon(A) = \{z \in \mathbb{C} : \|(zI - A)^{-1}\| \geq 1/\epsilon$.*

Clearly, $\sigma_\epsilon(A)$ increases with $\epsilon$. One shows that $\sigma_\epsilon(A)$ is compact, but not necessarily convex, and even not necessarily connected, since $\bigcap_{\epsilon > 0} \sigma_\epsilon(A) = \sigma(A)$.

**4.9. Theorem on equivalent representations of the pseudo-spectrum [14] :**
*The following statements are equivalent:*

$$\begin{array}{ll} (i) & z \in \sigma_\epsilon(A), \\ (ii) & \exists y \in \mathbb{C}^n, \|y\| = 1, \quad and \quad \|(A - zI)y\| \leq \epsilon, \\ (iii) & \exists E \in \mathbb{C}^{n\times n}, \|E\| \leq \epsilon, \quad and \quad z \in \sigma(A + E). \end{array}$$

**4.10. Exercise :**
*By establishing*

$$\|(xI - A)^{-1} - (yI - A)^{-1}\| \leq \|(xI - A)^{-1}\|^2/(1 - |x - y|\,\|(xI - A)^{-1}\|)$$

*if* $|x - y|\,\|(xI - A)^{-1}\| < 1$, *show that* $z \mapsto \|(zI - A)^{-1}\|$ *is continuous in* $\mathbb{C} \setminus \sigma(A)$. *Deduce that* $\sigma_\epsilon(A)$ *is closed, with boundary* $\partial\sigma_\epsilon(A) = \{z \in \mathbb{C} : \|(zI - A)^{-1}\| = 1/\epsilon\}$.

Starting from the Cauchy formula 2.11 for matrix functions, the following result is immediate.

**4.11. Corollary on the pseudo-spectrum [14] :**
*Any compact* $\Omega$ *containing the pseudo-spectrum* $\sigma_\epsilon(A)$ *is* $K$*-spectral for* $A$ *with* $K = length(\partial\Omega)/(2\pi\epsilon)$.

**4.12. Exercise :**
*For* $\epsilon > 0$, *show the following perturbation result*

$$\bigcup_{\|E\|\leq\epsilon} W(A + E) = \{z \in \mathbb{C} : dist(z, W(A)) \leq \epsilon\}.$$

# 5 Polynomial and rational approximation

This section discusses some basic methods for computing $f(A)$ for general (sufficiently smooth) functions. We should warn the reader that for classical functions like the exponential or fractional powers, particular methods have been developed (based for instance on a scaling and squaring principle for the exponential) which quite often are more efficient [6].
Considering the findings of §4, given a $K$-spectral set $\Omega$ for $A$ and $f$ analytic in $\Omega$, one could imagine to approach $f(A)$ by $p(A)$ with a polynomial $p$ and $\|f(A) - p(A)\| \leq K\,\|f - p\|_\Omega$. Two questions become however immediate:

| | |
|---|---|
| Q1) | how to evaluate $p(A)$, and |
| Q2) | how to find a suitable $p$ such that the right-hand side $\|f - p\|_\Omega$ is small. |

The answer to Q1) depends of course on the basis used to write down $p$, we could use for instance the basis of monomials $p(A) = \sum_{j=0}^m p_j A^j$. Writing $M(n)$ for the complexity of multiplying two matrices of order $n$ (which of course also depends on the structure of these matrices), a naive implementation would compute successively all powers of $A$ through $A^{j+1} = AA^j$. Since the powers could be computed on the fly, we arrive at a complexity of $mM(n)$ and a storage requirement of $\mathcal{O}(n^2)$, and the same is true for a Horner type implementation.

**5.1. Algorithm of Paterson and Stockmeyer**
*By adding coefficients* 0, *suppose that* $m = rs - 1$ *for two natural integers* $r, s$. *Writing*

$$p(A) = \sum_{j=0}^{r-1} B_j (A^s)^j, \quad B_j = \sum_{k=0}^{s-1} p_{sj+k} A^k,$$

*the first one evaluated by a Horner-type scheme, we arrive at a total complexity of* $(r + s)M(m)$ *(minimal for* $r \approx s \approx \sqrt{m}$*), but we need to store also the* $s + 1$ *matrices* $A^1, ..., A^s$.

One may show that there exists a constant $c$ (depending not on $n, p, A$) such that, for the output $X_m$ of Algorithm 5.1 in finite precision arithmetic

$$\|p(A) - X_m\| \leq \frac{c\epsilon n^2}{1 - c\epsilon n} \| \sum_j |p_j| |A|^j \|,$$

where $\epsilon$ is the machine precision, and where the absolute value in $|A|$ is taken elementwise.
The simple example $n = 1$, $p$ the $m$th partial sum of $f(z) = \exp(z)$ and $A = (-5)$ shows that the upper bound might be much larger as $\|f(A)\|$ because of cancellations. Of course, for scalar arguments, one may overcome this cancellation by computing inverses of partial sums of $(\exp(-A))$, but such a simple trick is no longer true for $2 \times 2$ matrices.
The following exercise gives us an idea of the error in exact arithmetic for partial sums of Taylor series.

**5.2. Exercise**
*Let $f$ be analytic in $\{|z| \leq \rho(A)\}$, then*

$$E_m := f(A) - \sum_{j=0}^{m} \frac{f^{(j)}(0)}{j!} A^j = \int_0^1 f^{(m+1)}(tA) \frac{(1-t)^m}{m!} A^{m+1} \, dt.$$

*Deduce that*

$$\|E_m\| \leq \frac{\|A^{m+1}\|}{(m+1)!} \max_{t \in [0,1]} \|f^{(m+1)}(tA)\|.$$

Applying 5.1 and 5.2 with $f(z) = \cos(z)$ and odd $m$, it remains to give an upper bound for $\max_{t \in [0,1]} \|\cos(tA)\|$, and $\| \sum_{j=0}^{(m-1)/2} \frac{|A|^{2j}}{2j!} \|$, for instance in both cases $\|\cosh(|A|)\| \approx \cosh(\|A\|)$. Thus for obtaining a small relative error $\|\cos(A) - X_m\| / \|\cos(A)\|$, we would desire that $\cosh(\|A\|)/\|\cos(A)\|$ is of moderate size, which is true as long as $\|A\| \leq 1$, but not in general. This clearly shows that it is preferable to apply several times the reduction formula $\cos(A) = 2\cos^2(A/2) - I$ before applying Taylor sums. Similar scaling arguments are true for the exponential function, the logarithm, and fractional powers.

Instead of working with Taylor sums, we could also compute polynomial interpolants $p$ of $f$ in some Newton basis. Here the ordering of the interpolation points $x_1, x_2, ...$ is essential (Leja ordering increases numerical stability). For evaluating $p(A)$, we could use either a Horner-type implementation or (in case of cyclically repeated interpolation points) a variant of 5.1. As we will see later, an appropriate choice of the interpolation points depending on $\Omega$ will enable us to give also an answer to question Q2).

For evaluating $R(A)$ for a rational function $R = P/Q$, $\deg P, \deg Q \leq m$, one has at least three possibilities:

- evaluate separately $P(A)$ and $Q(A)$ by the techniques in the beginning of the section, requiring one inversion of complexity $I(n)$, and about $2mM(n)$;
- evaluate at $z = A$ the partial fraction decomposition which generically takes the form

  $$R(z) = c_0 + \sum_{j=1}^{m} \frac{c_j}{z - z_j},$$

  here we have have a complexity of $mI(n)$;
- expand $R$ in a finite continued fraction following the techniques below.

**5.3. Continued fractions**
*As for sums and products, studying an infinite continued fraction*

$$C = \beta_0 + \frac{\alpha_1\rfloor}{\lceil \beta_1} + \frac{\alpha_2\rfloor}{\lceil \beta_2} + \frac{\alpha_3\rfloor}{\lceil \beta_3} + ...$$

*means that one has to study the convergence of the sequence of convergents*

$$C_m = \beta_0 + \frac{\alpha_1\rfloor}{\lceil \beta_1} + \frac{\alpha_2\rfloor}{\lceil \beta_2} + ... + \frac{\alpha_m\rfloor}{\lceil \beta_m} = \frac{P_m}{Q_m}$$

*which are computed by the initializations* $P_{-1} = 1, P_0 = \beta_0, Q_{-1} = 0, Q_0 = 1$, *and the two forward recurrencies*

$$\left\{ \begin{array}{c} P_{m+1} \\ Q_{m+1} \end{array} \right\} = \beta_{m+1} \left\{ \begin{array}{c} P_m \\ Q_m \end{array} \right\} + \alpha_{m+1} \left\{ \begin{array}{c} P_{m-1} \\ Q_{m-1} \end{array} \right\}.$$

*With one of the choices*

$$(i) \qquad \alpha_{m+1} = -\frac{C_{m+1} - C_m}{C_m - C_{m-1}}, \quad \beta_{m+1} = 1 - \alpha_{m+1},$$

$$(ii) \qquad \alpha_{m+1} = \frac{P_m Q_{m+1} - Q_m P_{m+1}}{P_m Q_{m-1} - Q_m P_{m-1}}, \quad \beta_{m+1} = \frac{P_{m+1} Q_{m-1} - Q_{m+1} P_{m-1}}{P_m Q_{m-1} - Q_m P_{m-1}}$$

*one may construct a continued fraction with convergents* $C_m = P_m/Q_m$. *However, a numerically stable evaluation of a continued fraction is done backwards: we have* $C_m = C_0^{(m)}$, *with*

$$C_m^{(m)} = \beta_m, \quad \text{and for } k = m-1, m-2, ..., 0\text{:} \quad C_k^{(m)} = \beta_k + \frac{\alpha_{k+1}}{C_{k+1}^{(m)}}.$$

*Supposing that all* $\alpha_k, \beta_k$ *are polynomials of degree at most* $s$, *we can therefore compute the* $m$*th convergent evaluated at* $A$ *with a complexity of* $mI(n) + sM(n)$ *(provided that we store some powers of* $A$*).*

**5.4. Rational interpolants with prescribed poles:**
*For a polynomial* $Q$ *and* $x_1, ..., x_m$ *with* $Q(x_j) \neq 0$, *we are looking for a polynomial* $P$ *of degree* $< m$ *such that* $R_{m,Q} = P/Q$ *interpolates* $f$ *at the interpolation points* $x_1, ..., x_m$ *(counting multiplicities).*
*The following properties are true:*

**Lagrange formula:** *if* $x_1, ..., x_m$ *are distinct then*

$$R_{m,Q}(z) = B(z) \sum_{j=1}^{m} \frac{f(x_j)}{(z - x_j) B'(x_j)}, \quad B(z) = \frac{\prod_{j=1}^{m} (z - x_j)}{Q(z)}.$$

**Link with polynomial interpolants and Cauchy formula:** $P = QR_{m,Q}$ *is interpolation polynomial of* $Qf$ *at* $x_1, ..., x_m$ *(existence and uniqueness of* $R_{m,Q}$*), in particular*

$$f(z) - R_{m,Q}(z) = B(z)\, [x_1, ..., x_m, z](Qf).$$

**Hermite formula:** *If* $f$ *is analytic in* $\Omega_0 \supset \Omega$ *and* $x_1, ..., x_m \in Int(\Omega_0)$ *then*

$$f(z) - R_{m,Q}(z) = B(z)\, \frac{1}{2\pi i} \int_{\partial\Omega_0} \frac{f(\zeta)}{B(\zeta)} \frac{d\zeta}{\zeta - z}$$

**Choice of interpolation points:** *following the Hermite formula, we desire to find for fixed $Q$ a numerator of $B$ such that $\|B\|_\Omega \|1/B\|_{\partial\Omega_0}$ is as small as possible. If we also optimize the denominator, we are left with the third Zolotarev problem on the two sets $\Omega$ and $\partial\Omega_0$. Thus, typically, the interpolation points lie in $\Omega$ whereas the poles of $B$ should be close to $\partial\Omega_0$, i.e., the roots of $Q$ simulate singularities of $f$.*

**Link with best rational approximation:** *Using the triangular inequality in the Lagrange formula, we get using the classical Céa Lemma trick with the Lebesgue constant $L$ that:*

$$\|f - R_{m,Q}\|_\Omega \leq (1+L) \min_{p\in\mathcal{P}_{m-1}} \|f - \frac{p}{Q}\|_\Omega, \quad L = \|\sum_{j=1}^{m} \Big| \frac{B(z)}{(z-x_j)B'(x_j)} \Big| \|_\Omega.$$

*Thus we could alternatively choose $x_1, ..., x_m$ minimizing $L$ which means that these interpolation points do represent "well" the continuous set $\Omega$ (and the denominator $Q$).*

**5.5. Definition: Rational interpolants with free poles**
*Here for given $x_1, ..., x_{2m} \in \mathbb{C}$ one tries to find $R_m = P_m/Q_m$ with $P_m \in \mathcal{P}_{m-1}, Q_m \in \mathbb{P}_m \setminus \{0\}$, such that $fQ_m - P_m$ vanishes at $x_1, ..., x_{2m}$ counting multiplicities. In general, $Q_m(x_j) \neq 0$, and thus the rational function $R_m$ interpolates $f$ at $x_1, ..., x_{2m}$.*
*The following properties are true:*

**Existence and uniqueness:** *Polynomials $P_m, Q_m$ as above exist, the fraction $P_m/Q_m$ is unique (but we may have unattainable points $x_j$ such that $R_m(x_j) \neq f(x_j)$).*

**Link with rational interpolants with fixed poles:** *for any $Q \in \mathcal{P}_m$ we have $R_m = R_{2m,QQ_m}$.*

**Link with Padé approximants:** *if $x_1 = x_2 = ... = x_{2m} = 0$ then the $R_m$ are Padé approximants[6] approximants of $f$ of type $[m-1|m]$.*

**Link with continued fraction 5.3(ii):** *the $\beta_j$ are polynomials of degree 1, and the $\alpha_k$ of degree 2 (with roots $x_{2k-1}, x_{2k}$).*

The philosophy behind rational interpolants with free poles is that the poles find themselves the singularities of the function. The most classical example is given for $f(z) = \log(z+1)$ where the Padé approximants $R_m$ converge uniformly on any compact subset of $\mathbb{C} \setminus (-\infty, -1]$, the interval $(-\infty, -1]$ containing all the poles [8]. This has to be compared with Taylor sums which only converge on disks.

For a Markov function $f(z) = \int_a^b \frac{d\mu(x)}{z-x}$ (see §3) one can be much more precise: if $\{x_1, ..., x_{2m}\} \subset \mathbb{C} \setminus [a, b]$ is symmetric with respect to the real axis, then the denominators $Q_m$ fulfill an orthogonality relation with varying weights, implying that all poles are simple and in $(a, b)$, and all residuals positive. One may deduce the so-called Markov theorem saying that if the interpolation points remain bounded away from $[a, b]$ then $R_m \to f$ uniformly on any compact subset of $\mathbb{C} \setminus [\alpha, \beta]$. Moreover, in case $x_1 = x_2 = ... \in (\beta, \infty)$ one gets the a posteriori estimate that maximum of $f - R_m$ in a disk $|z - x_1| \leq r < |\beta - z_1|$ is attained in $x_1 - r$.

# 6 Best approximants, Faber polynomials and the Faber transform

Soit $\mathbb{E} \subset \mathbb{C}$ un convexe compact et $f$ analytique dans un voisinage de $\mathbb{E}$. Dans ce chapitre on cherche à minimiser $\|f - p\|_\mathbb{E}$ pour un polynôme de degré $\leq n$ (ou une fonction rationnelle à numerateur de degré $\leq n$ et à pôles fixes).

[6] For many special functions, continued fractions with explicit coefficients are known with convergents being Padé approximants, see [8].

Dans ce chapitre nous allons introduire en 6.3 une transformée dite de Faber associée à $\mathbb{E}$ qui renvoie $w^j$ sur un polynôme $F_j$ de degré $j$ dit polynôme de Faber. Dans un premier temps nous allons montrer en 6.6 que les sommes partielles de la serie de Faber sont assez proche des meilleurs approximants polynomiaux de $f$ sur $\mathbb{E}$ par rapport à la norme du sup sur $\mathbb{E}$. Dans le cas particulier du disque $\mathbb{E} = \mathbb{D}$, on retrouve alors le lien connu entre meilleure approximation polynomiale et sommes de Taylor. En 6.7 nous explicitons ce résultat pour la classe des fonctions de Markov. Ensuite nous allons établir en 6.8 un résultat similaire pour des meilleurs approximants rationnels à pôles fixes. En complément du §4, nous concluons en 6.9 qu'il est possible de relier $\|(f-p)(A)\|$ à la norme du max de $\mathcal{F}^{-1}(f-p)$ sur $\mathbb{D}$.

On note par $\phi : \mathbb{C} \setminus \mathbb{E} \mapsto \mathbb{C} \setminus \mathbb{D}$ l'application de Riemann (l'unique bijection analytique conforme vérifiant $\phi(\infty) = \infty, \phi'(\infty) > 0$ et $\forall\, z : \phi'(z) \neq 0$), et $\psi = \phi^{-1}$. L'ensemble de niveau $\mathbb{E}_R$ pour $R > 1$ est défini par son complément $\mathbb{E}_R^c = \{z \notin \mathbb{E} : |\phi(z)| > R\}$.

**6.1. Définition :**
*On définit $F_j(z)$ pour $z \in int(\mathbb{E})$, $|w| \geq 1$ (ou $z \in \mathbb{E}$, $|w| > 1$) par la fonction génératrice*

$$\frac{w\psi'(w)}{\psi(w) - z} = \sum_{j=0}^{\infty} \frac{F_j(z)}{w^j}.$$

Pour l'exemple $\mathbb{E} = \mathbb{D}$, nous avons $\psi(w) = w$, et $F_j(z) = z^j$. Pour l'exemple $\mathbb{E} = [-1, 1]$, $\psi(w) = \frac{1}{2}(w + \frac{1}{w})$, et $F_0(z) = 1$ et pour $j \geq 1$ : $F_j(\psi(w)) = w^j + \frac{1}{w^j} = 2T_j(\psi(w))$, avec $T_j$ le $j$ième polynôme de Chebyshev.

**6.2. Lemme :**
*$F_j$ est un polynôme de degré $j$, $F_0(z) = 1$, et pour $j \geq 1$ : $F_j(\psi(w)) - w^j$ est analytique dans $|w| > 1$ inclus $\infty$ et s'annule en $\infty$.*

*Proof.* La série génératrice étant absolument convergente pour $|w| = 1 + \epsilon > 1$, on obtient pour $k \in \mathbb{Z}$, $z \in \mathbb{E}$

$$(*) \qquad \begin{aligned} &\frac{1}{2\pi i} \int_{|w|=1} w^k \frac{w\psi'(w)}{\psi(w) - z} \frac{dw}{w} \\ &= \sum_{j=0}^{\infty} F_j(z) \frac{1}{2\pi i} \int_{|w|=1+\epsilon} w^{k-j} \frac{dw}{w} = \begin{cases} F_k(z) & k \geq 0, \\ 0 & k < 0. \end{cases} \end{aligned}$$

en particulier en écrivant $\phi(\zeta)^j - P(\zeta)$ analytique en $\mathbb{C}\setminus\mathbb{E}$ et s'annulant en $\infty$ avec $P$ un polynôme de degré $j$

$$F_j(z) - P(z) = \frac{1}{2\pi i} \int_{\partial\mathbb{E}} (\phi(\zeta)^j - P(\zeta)) \frac{d\zeta}{\zeta - z} = 0$$

d'après le théorème de Cauchy. □

Voici un résultat utilisant la convexité de $\mathbb{E}$.

**6.3. Définition et Théorème :**
*Pour $P$ polynôme et $z \in int(\mathbb{E})$, soit*

$$\mathcal{F}(P)(z) = \frac{1}{2\pi i} \int_{|w|=1} P(w)\, 2Re\Big(\frac{w\psi'(w)}{\psi(w) - z}\Big) \frac{dw}{w}.$$

**(a)** *$\mathcal{F}(1)(z) = 2$, et pour $j \geq 1$ : $\mathcal{F}(w^j)(z) = F_j(z)$.*
**(b)** *$\|\mathcal{F}(P)\|_{\mathbb{E}} \leq 2\,\|P\|_{\mathbb{D}}$, en particulier $\|F_j\|_{\mathbb{E}} \leq 2$.*
**(c)** *$\mathcal{F}(P)(\psi(w)) - P(w)$ est analytique dans $|w| > 1$ inclus $\infty$.*

*Proof.* Pour $j \geq 0$

$$\mathcal{F}(w^j)(z) = \frac{1}{2\pi}\int_{|w|=1} w^j \frac{w\psi'(w)}{\psi(w)-z}\frac{dw}{iw} + \overline{\frac{1}{2\pi}\int_{|w|=1} w^{-j}\frac{w\psi'(w)}{\psi(w)-z}\frac{dw}{iw}}$$

ce qui d'après $(*)$ vaut 2 si $j=0$, et $F_j(z)$ pour $j>0$, ce qui démontre (a). Pour une preuve de la partie $(b)$, notons d'abord que pour $w=e^{it}$ nous avons $\frac{dw}{iw} = dt > 0$. Aussi, on montre que $w\psi'(w)/|w\psi'(w)|$ nous donne la normale extérieure de $\mathbb{E}$ au point $z=\psi(w)$. Donc par convexité pour $z \in \mathrm{int}(\mathbb{E})$

$$\mathrm{Re}\Big(\frac{w\psi'(w)}{\psi(w)-z}\Big) > 0,$$

ce qui permet d'estimer

$$|\mathcal{F}(P)(z)| \leq \frac{\|P\|_{\mathbb{D}}}{2\pi}\int_{|w|=1}\Big|2\mathrm{Re}\Big(\frac{w\psi'(w)}{\psi(w)-z}\Big)\frac{dw}{iw}\Big| = \|P\|_{\mathbb{D}}\mathcal{F}(1) = 2\,\|P\|_{\mathbb{D}}.$$

□

**6.4. Exercice :**
*En suivant le raisonnement de la preuve du théorème 4.3 de Neumann, montrer que $W(A) \subset \mathbb{E}$ et $p = \mathcal{F}(P)$ pour un polynôme $P$ implique que $\|p(A)\| \leq 2\,\|P\|_{\mathbb{D}}$, et en particulier $\|F_j(A)\| \leq 2$.*

**6.5. Exercice :**
*Soit $f$ analytique dans un voisinage de $\mathbb{E}_R$ pour $R>1$, alors avec*

$$f_j := \frac{1}{2\pi i}\int_{|w|=1}\frac{f(\psi(w))}{w^j}\frac{dw}{w}$$

*dits coefficients de Faber montrer que $f_j = \mathcal{O}(R^{-j})_{j\to\infty}$, et que les sommes partielles de la somme de Faber $\sum_{j=0}^{\infty} f_j F_j(z)$ convergent vers $f$ uniformément dans $\mathbb{E}$.*

L'exo 6.5 nous permet d'étendre la définition de $\mathcal{F}$ à tout $P$ analytique dans un voisinage de $\mathbb{D}$, tout en gardant les propriétés 6.3(b),(c), et

$$\mathcal{F}\Big(\frac{f_0}{2} + \sum_{j=1}^{\infty} f_j w^j\Big)(z) = \sum_{j=0}^{\infty} f_j F_j(z).$$

On a le résultat suivant concernant la meilleure approximation polynômiale de $f$ sur $\mathbb{E}$.

**6.6. Corollaire :**
*Soit $f$ analytique dans un voisinage de $\mathbb{E}$, alors*

$$|f_{m+1}| \leq \sqrt{\sum_{j=m+1}^{\infty}|f_j|^2} \leq \min_{P\in\mathcal{P}_m}\|f-P\|_{\mathbb{E}} \leq \|f - \sum_{j=0}^{m} f_j F_j\|_{\mathbb{E}} \leq 2\sum_{j=m+1}^{\infty}|f_j|.$$

Ce corollaire 6.6 nous donne un encadrement précis et un "bon" approximant explicite si les $f_j$ décroissent rapidement, voir l'exemple suivant. Notons que la première et troisième inégalité sont évidentes, et la quatrième découle de l'estimation de $\|F_j\|_{\mathbb{E}}$ donnée dans 6.3(b). Une preuve de la deuxième inégalité va nous demander un peu d'effort, elle découlera comme cas particulier du théorème 6.8 ci-dessous.

**6.7. Corollaire :**
*Soit $\mathbb{E}$ symétrique par rapport à l'axe réelle, et $[a,b] \subset \mathbb{R}$, à gauche de $\mathbb{E}$ (c'est-à-dire, $b < \min\{Re(z) : z \in \mathbb{E}\}$), alors pour la fonction de Markov*

$$f(z) = \int_a^b \frac{d\mu(x)}{z-x},$$

*pour $j \geq m+1$ nous avons*

$$|f_j| \leq |\phi(b)|^{m+1-j}|\,|f_{m+1}|, \quad \sum_{j=m+1}^{\infty} |f_j| \leq \frac{2}{|\phi(b)|^{m+1}}\,\|f\|_{\mathbb{E}}$$

*($\Longrightarrow$ l'estimation du 1.6.6 est précise à un facteur $2/(1-|\phi(b)|^{-1})$ près).*

*Proof.* D'après le théorème de Fubini nous avons

$$f_j \;=\; \frac{1}{2\pi i}\int_{|w|=1}\int_a^b \frac{d\mu(x)}{\psi(w)-x}\frac{dw}{w^{j+1}} = \int_a^b\int_a^b d\mu(x)\frac{1}{2\pi i}\int_{|w|=1}\frac{1}{\psi(w)-x}\frac{dw}{w^{j+1}}.$$

L'intégrant de l'intégrale en $w$ admet une seule singularité dans $\mathbb{D}^c \cup \{\infty\}$, au point $w = \phi(x)$. Donc, par le théorème des résidus en analyse complexe (ou tout simplement par le théorème de Cauchy après changement de variables $\zeta = \psi(w)$ et changement d'orientation de la courbe d'intégration),

$$\frac{1}{2\pi i}\int_{|w|=1}\frac{1}{\psi(w)-x}\frac{dw}{w^{j+1}} = -\frac{1}{\psi'(\phi(x))}\frac{1}{\phi(x)^{j+1}} = -\frac{\phi'(x)}{\phi(x)^{j+1}}.$$

Par unicité de l'application de Riemann et symétrie de $\mathbb{E}$, nous avons $\phi(\overline{z}) = \overline{\phi(z)}$ pour tout $z \notin \mathbb{E}$, en particulier, $\phi(x)$ et $\phi'(x) \neq 0$ sont réels pour $x \in \mathbb{R} \setminus \mathbb{E} \supset [a,b]$. Comme de plus $\phi'(\infty) > 0$, $\phi(\infty) = \infty$, nous déduisons que $\phi' > 0$ dans $\mathbb{R} \setminus [a,b]$, et donc $\phi$ est croissant et négatif sur $[a,b]$, et $1/|\phi|$ croit sur $[a,b]$. Donc

$$|f_j| \;=\; \Big|\int_a^b \frac{\phi'(x)}{\phi(x)^{j+1}}d\mu(x)\Big| = \int_a^b \frac{|\phi'(x)|}{|\phi(x)|^{j+1}}d\mu(x) \leq |\phi(b)|^{m+1-j}|\,|f_{m+1}|.$$

Du théorème 3.1 de l'article [K. C. Toh and L. N. Trefethen, The Kreiss matrix theorem on a general complex domain, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 145–165] on sait que, pour tout domaine $\mathbb{E}$ simplement connexe pas forcement convexe,

$$\forall z \notin \mathbb{E}: \quad \operatorname{dist}(z,\mathbb{E})\frac{|\phi'(z)|}{|\phi(z)|-1} \in [\frac{1}{2}, 2].$$

Par conséquent,

$$\sum_{j=m+1}^{\infty}|f_j| = \int_a^b \frac{|\phi'(x)|\,d\mu(x)}{(1-|\phi(x)|^{-1})|\phi(x)|^{m+2}} \leq \frac{2}{|\phi(b)|^{m+1}}\int_a^b \frac{d\mu(x)}{\operatorname{dist}(x,\mathbb{E})} = \frac{2}{|\phi(b)|^{m+1}}\,\|f\|_{\mathbb{E}}.$$

□

Nous allons maintenant démontrer un résultat similaire à 1.6.6 pour les fonctions rationnelles à pôles prescrits. Pour $w_1, ..., w_m \in \mathbb{C}\setminus\mathbb{D}$ soit $Q(w) = \prod_{j=1}^m (1-w/w_j)$, et $q(z) = \prod_{j=1}^m (z-\psi(w_j))$. On va supposer dans la suite que les $w_j$ (et donc les $z_j = \psi(w)$) soient distincts. Néanmoins, les idées de preuve restent valables après des passages à la limite, par exemple $w_1 \to w_2$, mais aussi

$w_1 \to \infty$ (et donc $1 - w/w_1 \to 1$, ce qui veut dire que $Q$ sera de degré $< m$). En particulier, on aura la situation du théorème 1.6.6 en faisant tendre tous les $w_j$ vers $\infty$.
Rappelons quelques petites éléments de la théorie des espaces de Hardy: on note

$$\|F\|_2 := \sqrt{\frac{1}{2\pi}\int_{|w|=1} |F(w)|^2\,|dw|}$$

pour une fonction $F$ de carré intégrable sur le cercle d'unité. L'identité $\frac{1}{2\pi}\int_{|w|=1} w^{j-k}|dw| = \delta_{j,k}$ plus la théorie des espaces $H^2$ montre que

$$\|F\|_2 = \sqrt{\sum_{j=-\infty}^{+\infty} |F_j|^2}$$

si $F$ est analytique dans la couronne $1-\epsilon < |w| < 1+\epsilon$ et y admet alors un développement de Laurent $F(w) = \sum_{j=-\infty}^{+\infty} F_j w^j$.

**6.8. Théorème :**
*Soit $f$ analytique dans un voisinage de $\mathbb{E}$. Notons $R_m = P_m/Q$ l'interpolant de $F(w) = f_0/2 + \sum_{j=1}^{\infty} f_j w^j$ aux points $0$ et $1/\overline{w}_1, ..., 1/\overline{w}_m$, et*

$$B(w) := w\prod_{j=1}^{m}\frac{w - 1/\overline{w}_j}{1 - w/w_j}, \quad \frac{p_m}{q} := \mathcal{F}(\frac{P_m}{Q}), \quad b_j := \frac{1}{2\pi i}\int_{|u|=1}\frac{f(\psi(w))}{B(u)}\frac{du}{u^j}.$$

*Alors*

$$|b_1| \le \sqrt{\sum_{j=1}^{\infty}|b_j|^2} \le \min_{p\in\mathcal{P}_m}\|f - \frac{p}{q}\|_{\mathbb{E}} \le \|f - \frac{p_m}{q}\|_{\mathbb{E}} \le 2\sum_{j=1}^{\infty}|b_j|.$$

Avant de se lancer dans la preuve, notons que pour $w_1, ..., w_m \to \infty$, $B$ devient $w^{m+1}$ et donc $b_j = f_{j+m}$. Aussi, $P_m/Q = P_m$ devient la somme partielle de $F$ d'ordre $m$, donc le corollaire 6.6 est en effet un cas limite du théorème 6.8.

*Proof.* Dans un premier temps, montrons que $p_m \in \mathcal{P}_m$, c'est-à-dire, $p_m/q$ est effectivement un candidat pour notre problème de minimisation. En écrivant la décomposition en termes simples et en utilisant la fonction génératrice des polynômes de Faber nous obtenons

$$\begin{aligned}
\mathcal{F}\Big(\frac{P_m(w)}{Q(w)}\Big)(z) &= \mathcal{F}\Big(c_0 + \sum_{j=1}^{m}\frac{c_j}{w - w_j}\Big)(z) = \mathcal{F}\Big(c_0 - \sum_{j=1}^{m}\frac{c_j}{w_j}\sum_{k=0}^{\infty}\frac{w^k}{w_j^k}\Big)(z)\\
&= c_0\mathcal{F}(1)(z) - \sum_{j=1}^{m}\frac{c_j}{w_j}\sum_{k=0}^{\infty}\frac{\mathcal{F}(w^k)(z)}{w_j^k}\\
&= c_0 + \frac{P_m}{Q}(0) - \sum_{j=1}^{m}\frac{c_j}{w_j}\frac{w_j\psi'(w_j)}{\psi(w_j) - z} = c_0 + \frac{P_m}{Q}(0) - \sum_{j=1}^{m}\frac{c_j\psi'(w_j)}{z - \psi(w_j)}
\end{aligned}$$

étant clairement un élément de $\mathcal{P}_m/q$. D'ailleurs, cette formule très explicite permet de construire sur ordinateur $p_m/q$ sachant la décomposition en termes simples de $P_m/Q$.
On passe maintenant à une preuve de la troisième inégalité sachant que la deuxième est triviale. Observons d'abord que $F - P_m/Q$ est analytique dans un voisinage de $|w| \le 1+\epsilon$ pour un $\epsilon > 0$. D'après la formule d'Hermite 5.4, nous obtenons pour $|w| = 1$ sachant que $|B(w)| = 1$

$$\begin{aligned}
\Big|F(w) - \frac{P_m}{Q}(w)\Big| &= |B(w)|\Big|\frac{1}{2\pi i}\int_{|u|=1+\epsilon}\frac{F(u)}{B(u)}\frac{du}{u-w}\Big|\\
&= \Big|\frac{1}{2\pi i}\int_{|u|=1+\epsilon}\frac{f(\psi(u))}{B(u)}\frac{du}{u-w}\Big| = \Big|\sum_{j=1}^{\infty} b_j w^{j-1}\Big| \le \sum_{j=1}^{\infty}|b_j|,
\end{aligned}$$

où dans la deuxième égalité on a utilisé le fait que

$$u \mapsto \frac{F(u) - f(\psi(u))}{B(u)} \frac{1}{u-w}$$

est analytique dans $|u| \geq 1+\epsilon$ inclus $\infty$, voire 6.3(c), avec un double zéro en $\infty$. En utilisant 6.3(b), on en déduit que

$$\|f - \frac{p_m}{q}\|_{\mathbb{E}} = \|\mathcal{F}(F - \frac{P_m}{Q})\|_{\mathbb{E}} \leq 2\,\|F - \frac{P_m}{Q}\|_{\mathbb{D}} \leq 2\sum_{j=1}^{\infty} |b_j|,$$

c'est-à-dire, il reste seulement la première inégalité à établir.
Pour tout $p \in \mathcal{P}_m$ nous pouvons écrire

$$(f - \frac{p}{q})(\psi(w)) = w(\widetilde{F}(w) - \frac{P}{Q}(w)) + H(w), \quad \widetilde{F}(w) = \sum_{j=1}^{\infty} F_j w^j = \frac{F(w) - F(0)}{w},$$

avec $P \in \mathcal{P}_{m-1}$, et $H$ analytique dans $|u| > 1$ d'après 6.3(c) (développer $f - p/q$ en série de Faber). Comme le terme à gauche du second membre est analytique dans un voisinage du disque, et s'annule en 0, nous obtenons alors

$$\|f - \frac{p}{q}\|_{\mathbb{E}}^2 = \|(f - \frac{p}{q}) \circ \psi\|_{\partial\mathbb{D}}^2 \geq \|(f - \frac{p}{q}) \circ \psi\|_2^2 = \|\widetilde{F} - \frac{P}{Q}\|_2^2 + \|H\|_2^2.$$

Notons qu'il existe un polynôme $\widetilde{P} \in \mathcal{P}_{m-1}$ de sorte que

$$\frac{P_m}{Q} - F(0) = \frac{P_m}{Q} - \frac{P_m}{Q}(0) = w\,\frac{\widetilde{P}}{Q} \quad \text{et alors} \quad \|\widetilde{F} - \frac{\widetilde{P}}{Q}\|_2^2 = \|F - \frac{P_m}{Q}\|_2^2 = \sum_{j=1}^{\infty} |b_j|^2,$$

la dernière égalité découlant de la représentation intégrale de $|F - P/Q|$ donnée ci-dessus. En effet, le lecteur vérifie aisément que $\widetilde{P}/Q$ n'est rien que l'interpolant $\in \mathcal{P}_{m-1}/Q$ de $\widetilde{F}$ aux points $1/\overline{w}_1, ..., 1/\overline{w}_m$. En combinant ces deux chaînes d'inégalités, il est suffisant de démontrer que

$$\|\widetilde{F} - \frac{\widetilde{P}}{Q}\|_2 = \min_{P \in \mathcal{P}_{m-1}} \|\widetilde{F} - \frac{P}{Q}\|_2,$$

autrement dit, on connaît le meilleur approximant par rapport à la norme $\|\cdot\|_2$ induite par un produit scalaire

$$\langle G, H\rangle = \frac{1}{2\pi}\int_{|w|=1} F(w)\overline{G(w)}|dw| = \frac{1}{2\pi i}\int_{|w|=1} F(w)\overline{G(w)}\frac{dw}{w},$$

défini, disons, sur l'espace vectoriel des fonctions analytiques dans un voisinage fixe de $\mathbb{D}$ (c'est en effet le produit scalaire de l'espace plus grand $H^2$ de Hardy). On sait minimiser au sens des moindres carrés:

$$\frac{\widetilde{P}}{Q}(w) = \sum_{j=1}^{m} \frac{e_j}{w - w_j}$$

est meilleur approximant par rapport à $\|\cdot\|_2$ de $\widetilde{F}$ si et seulement si l'erreur $\widetilde{F} - \frac{\widetilde{P}}{Q}$ est orthogonal à toute fonction dans $\mathcal{P}_{m-1}/Q$, avec base $1/(w - w_\ell)$, $\ell = 1, ..., m$. Il faut et il suffit alors que

$$\left[\ \langle \widetilde{F}, \tfrac{1}{w-w_\ell}\rangle\ \right]_{\ell=1,...,m} = \left[\ \langle \tfrac{1}{w-w_j}, \tfrac{1}{w-w_\ell}\rangle\ \right]_{\ell,j=1,...,m} \left[\ c_j\ \right]_{j=1,...,m}.$$

Un petit calcul de résidus montre que

$$\langle \widetilde{F}, \frac{1}{w-w_\ell}\rangle = -F(1/\overline{w}_\ell)/\overline{w}_\ell, \qquad \langle \frac{1}{w-w_j}, \frac{1}{w-w_\ell}\rangle = -\frac{1}{1/\overline{w}_\ell - w_j}/\overline{w}_\ell$$

et donc notre système est équivalent au fait que $\widetilde{P}/Q$ interpole $\widetilde{F}$ aux points $1/\overline{w}_1, ..., 1/\overline{w}_m$, comme désiré ci-dessus. □

**6.9. Exercise :**
*Avec les notations de 1.6.8, si $W(A) \subset \mathbb{E}$ alors*

$$\|f(A) - p_m(A)q(A)^{-1}\| \leq \|F - P_m/Q\|_{\mathbb{D}} \leq \sum_{j=1}^{\infty} |b_j|.$$

# 7 Direct computation : the Parlett-Schur approach

The Parlett method for computing $f(A)$, implemented as `funm` under Matlab, can be summarized as follows

1. compute the Schur normal form $A = UTU^*$ with unitary $U$ and upper triangular $T$. We then have $f(A) = Uf(T)U^*$ by 2.6(b);

2. introduce a block partition

$$T = \begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,\ell} \\ 0 & T_{2,2} & \cdots & T_{2,\ell} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & T_{\ell,\ell} \end{bmatrix}$$

   with square $T_{j,j}$;

3. compute $F_{j,j} = F(T_{j,j})$ by some direct method (see §5, Taylor expansion, polynomial interpolants, rational approximation of $f$);

4. then

$$f(T) = \begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & F_{1,\ell} \\ 0 & F_{2,2} & \cdots & F_{2,\ell} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & F_{\ell,\ell} \end{bmatrix}$$

   with the same block partition as $T$, where the blocks $F_{j,k}$ for increasing $j-k \geq 1$ are obtained through the Sylvester equation in the unknown rectangular matrix $F_{j,k}$

$$T_{j,j}F_{j,k} - F_{j,k}T_{k,k} = \sum_{i=j}^{k-1} F_{j,i}T_{i,k} - \sum_{i=j+1}^{k} T_{j,i}F_{i,k} \tag{1}$$

   which itself follows from the identity $Tf(T) = f(T)T$, see 2.6(c).

The aim of this section is to study more closely each of the above steps, see also [6, Chapter 4.6] and [6, Chapter 9].

## 7.1 Defining and computing the Schur normal form

As a general reference we refer the reader to [3].

**7.1. Lemma on existence of a Schur normal form**
*Any matrix $A \in \mathbb{C}^{n\times n}$ can be factorized as $A = UTU^{-1}$ with $U, T \in \mathbb{C}^{n\times n}$, $U^{-1} = U^*$ (unitary) and $T$ upper triangular.*

*Proof.* By recurrence on $n$: we multiply $A$ on the right with $Q$ and on the left with $Q^{-1}$ where the unitary $Q$ contains in its first column an eigenvector of $A$. □

We notice that the set of diagonal elements of $T$ concides with $\sigma(T) = \sigma(A)$. Also, with $A$ hermitian/normal, also $T$ is hermitian/normal, implying that $T$ must be diagonal ($\longrightarrow$ for normal $A$ we get the Jordan normal form).
In general, such a Schur decomposition of $A$ is not unique. Also, the technique employed in the proof for constructing $U$ should not be implemented on a computer as it is, since it requires many eigenvector computations. Indeed, we can directly use the $QR$ method (originally designed for computing eigenvalues).

**7.2. The $QR$ method:**
*Starting from $A_0 = A$ and a sequence of $\mu_0, \mu_1, ... \in \mathbb{C}$, we compute successively for $k = 0, 1, 2, ...$*

$$A_k - \mu_k I = Q_k R_k \quad \textit{(QR decomposition, } Q_k \textit{ unitary, } R_k \textit{ upper triangular),}$$
$$A_{k+1} = R_k Q_k + \mu_k I.$$

Notice that $A_k = Q_{k-1}^* A_{k-1} Q_{k-1} = ... = U_k^* A U_k$ with unitary $U_k = Q_{k-1}...Q_1 Q_0$, that is, $A$ and $A_k$ are similar. It is possible to show (at least for symmetric $A$ and for a large choice of shift parameters $\mu_k$) that $A_k \to T$ upper triangular, and that $U_k \to U$ unitary, and thus we obtain the Schur decomposition $T = U^* A U$. In practice it is observed that the convergence is quite fast, such that after $\mathcal{O}(n)$ iterations one obtains an acceptable precision.
A basic (naive) implementation of 7.2 requires $\mathcal{O}(n^3)$ arithmetic operations for each iteration, which can be reduced to $\mathcal{O}(n^2)$ provided $A_0$ is already in upper Hessenberg form (i.e., the $(j,k)$ entry is zero for $j > k+1$), as seen as follows. In the sequel of §7 we adapt matlab notation for submatrices.
Recall that for all $x \in \mathbb{C}^n$ there exists $w \in \mathbb{C}^n$ of norm 1 such that $H(w) = 1 - 2ww^*$ is unitary (called Householder transformation) and that $H(w)x = \|x\|e_1$, with $e_1$ the first canonical vector. Recall also that the matrix mutiplication $H(w)B = B - w(w^*B)$ can be implemented with a complexity $O(n^2)$ by adding a multiple of $w$ to each column of $B$ (and similarly for $BH(w)$).

**7.3. Reduction to Hessenberg form:**
*We just describe the first reduction, and apply the same procedure to all lower right submatrices: let $H_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$ be a Householder transformation with $H_{n-1}A(2:n,1) = \alpha e_1$, then*

$$\left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & H_{n-1} & \\ 0 & & & \end{array}\right] A \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & H_{n-1} & \\ 0 & & & \end{array}\right]^* = \left[\begin{array}{c|ccc} * & * & \cdots & * \\ \hline \alpha & & & \\ 0 & & & \\ \vdots & & * & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & H_{n-1} & \\ 0 & & & \end{array}\right] = \left[\begin{array}{c|ccc} * & * & \cdots & * \\ \hline \alpha & & & \\ 0 & & & \\ \vdots & & * & \\ 0 & & & \end{array}\right].$$

*Thus we get in complexity $O(n^3)$ an upper Hessenberg $A_0 = Q^* A Q$ with unitary*

$$Q = \left[\begin{array}{cc} I_1 & 0 \\ 0 & H_{n-1} \end{array}\right] \left[\begin{array}{cc} I_2 & 0 \\ 0 & H_{n-2} \end{array}\right] ... \left[\begin{array}{cc} I_{n-2} & 0 \\ 0 & H_2 \end{array}\right].$$

**7.4. Details for implementing the QR method:**
*Let us show by recurrence that if $A_k$ is lower Hessenberg then also $A_{k+1}$, which will allow us to deduce an implementation of the QR method with complexity $\mathcal{O}(n^2)$ for each iteration (giving an overall complexity of $O(n^3)$). At step $j \in 1, ..., n-1$ one constructs a Givens rotation $G_j = \mathrm{diag}(I_{j-1}, H_j, I_{n-1-j})$ with unitary $H_j \in \mathbb{C}^{2\times 2}$ such that*

$$H_j(G_{j-1}...G_1(A_k - \mu_k I))(j : j+1, j) = (\alpha_j, 0)^T$$

*for some $\alpha_j > 0$, and thus $G_{j-1}...G_1(A_k - \mu_k I)$ is upper Hessenberg, the first $j$ elements on the first lower diagonal being equal to 0, but not yet the other ones. Thus setting*

$$Q_k^* = G_{n-1}...G_1, \quad R_k = Q_k^*(A_k - \mu_k I)$$

*we have obtained our QR decomposition, with the updating formulas $U_{k+1} = Q_k G_1...G_{n-1}$ and $A_{k+1} = \mu_k I + R_k Q_k = \mu_k I + R_k G_{n-1}^*...G_1^*$. Since any multiplication with $G_j$ on the left (or $G_j^*$ on the right) consists of replacing only rows (columns) $j$ and $j+1$ by a linear combination of both ones, the resulting matrix $A_{k+1}$ indeed has a Hessenberg structure, and one iteration has the complexity $\mathcal{O}(n^2)$.*
*In practice, one observes that there is an index $\gamma_k$ (decreasing in $k$) such that*

$$A_k(j+1, j) \quad \textit{are "small" for } j = \gamma_k, \gamma_k + 1, ..., n-1 \textit{ but not for } j = \gamma_k - 1$$

*(where "small" could mean that $|A_k(j+1,j)| \leq \tau(|A_k(j,j)| + |A_k(j+1,j+1)|)$ with the tolerance $\tau$ of order of a multiple of the machine precision). In this case one decides to make a "deflation" and continues to apply the QR decomposition only for the upper left submatrix of order $\gamma_k$, or, equivalently (at least in exact arithmetic), does no longer use Gives factors $G_j$ for $j \geq \gamma_k$.*
*It remains to have a practical idea of how to choose the shift parameters $\mu_k$ (which strongly influences the convergence behavior): common choices are the Rayleigh shift $\mu_k = A_k(\gamma_k, \gamma_k)$, the Wilkinson shift where $\mu_k$ is the eigenvalue of $A_k(\gamma_k - 1 : \gamma_k, \gamma_k - 1 : \gamma_k)$ the closest to $A_k(\gamma_k, \gamma_k)$, and finally the "double shift implicit" where one uses both eigenvalues (here the implementation is more complicated [3], but the advantage is that one can use real arithmetic for real data, with the price to pay that there might be nontriangular $2 \times 2$ blocks on the diagonal of $T$).*

As one expects from manipulating unitary matrices, it can be shown that the computation of the Schur factorization through the QR method is backward stable.

## 7.2 Computing matrix functions through the Sylvester equation

We still have to discuss an efficient and numerically stable way of solving (1). By simplifying notation, consider the Sylvester equation $AX - XB = C$ in the unknown $X \in \mathbb{C}^{p\times q}$, with square upper triangular matrices $A, B$ of order $p$ and $q$, respectively. In our setting, $p, q$ are small compared to $n$, but $p \neq q$ is possible.

**7.5. Lemma on the solvability of the Sylvester equation:**
*The Sylvester equation $AX - XB = C$ has a unique solution for all right-hand sides $C$ if and only if $\sigma(A) \cap \sigma(B)$ is empty.*

*Proof.* If $\lambda \in \sigma(A) \cap \sigma(B)$ then by choosing as $x$ a corresponding right-hand eigenvector of $A$ and as $y^*$ a corresponding left-hand eigenvector of $B$, we get for $C = 0$ the two solutions $X \in \{0, xy^*\}$.
Conversely, we may assume by using if necessary the Schur decomposition that $B$ is upper triangular. Then for the $j$th column of $AX - XB = C$ we get that

$$(A - B(j,j)I)X(:,j) = C(:,j) - \sum_{k=0}^{j-1} X(:,k)B(k,j).$$

Since $B(j,j) \in \sigma(B)$ and $A - B(j,j)I$ is invertible by assumption on the spectra, this formula allows to compute successively all columns of $X$. $\square$

In Parlett's approach (1), the data and in particular the matrix $C$ in general are only available up to some precision, since they are themselves results of floating point operations. In order to monitor the accumulation of rounding errors, one wishes that a small perturbation $\Delta C$ of the right-hand side does not imply a large perturbation $\Delta X$ of the solution, i.e., with $A(X+\Delta X) - (X+\Delta X)B = C + \Delta C$ one wants to keep

$$\frac{\|\Delta X\|}{\|\Delta C\|} = \frac{\|\Delta X\|}{\|A\Delta X - \Delta X B\|}$$

of moderate size. This motivates heuristically the desire to have a quantity $\text{spread}(A,B) := \min_Y \|AY - YB\|/\|Y\|$ as large as possible, the criterion for our partitionning of $T$, see §7.3. Unfortunately, we only dispose of upper bounds which we will be able to maximize. Indeed, $\text{spread}(A,B) \leq \text{dist}(\sigma(A), \sigma(B))$, which can be easily seen by considering $Y = xy^*$, with $x$ right eigenvectors of $A$, and $y^*$ left eigenvectors of $B$.

**7.6. Exercice:**
*For normal $A, B$, show that spread$(A,B) \leq$ dist$(\sigma(A), \sigma(B)) \leq \sqrt{\max(p,q)}$spread$(A,B)$.*

### 7.3 How to partition?

As we have seen in §7.2, the stability of the Parlett-Schur method depends on the fact whether $\text{spread}(T_{j,j}, T_{k,k})$ is sufficiently large for all $j \neq k$. Let us stay here with the complex QR method where $T$ is upper triangular. In this case, we try to partition $\sigma(A)$, the diagonal of $T$, in such a way that $\text{dist}(\sigma(T_{j,j}), \sigma(T_{k,k}))$, the distance between the diagonal elements of $T_{j,j}$ and those of $T_{k,k}$ becomes as large as possible for all $j \neq k$. In particular, two identical or "close" eigenvalues should be in the same block.
It is easy to find a partition in classes $F_1, ..., F_\ell$ of the discrete set $\sigma(A)$ such that any two members of two different classes have a distance of at least $\delta$, the threshold $\delta > 0$ being fixed by the users. For instance, one may find the connected components in an undirected graph where two vertices = elements of $\sigma(A)$ are connected by an edge if their distance is $< \delta$.
By making the link between the QR method and the (block) power method, one may show that, in general, $|t_{1,1}| \leq |t_{2,2}| \leq ... \leq |t_{n,n}|$ for the diagonal elements of $T$, which makes it likely that members of the same class are successive elements on the diagonal of $T$, simplifying the partitioning of $T$.
However, it might be necessary to permute some diagonal elements of $T$ (through simultaneous permutations of rows and columns of $T$) destroying the structure of $T$. It is sufficient to consider the case of a permutation of the elements $t_{j,j}$ and $t_{j+1,j+1}$: here one constructs a unitary Schur factor $H_j$ of order 2 (as in the proof of Lemma 7.1) such that

$$H_j \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} T(j:j+1, j:j+1) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^* H_j^* = H_j \begin{bmatrix} t_{j+1,j+1} & 0 \\ t_{j,j+1} & t_{j,j} \end{bmatrix} H_j^* = \begin{bmatrix} t_{j+1,j+1} & * \\ 0 & t_{j,j} \end{bmatrix}.$$

Thus a similarity transformation with $\text{diag}(I_{j-1}, H_j, I_{n-j-1})$ will reestablish the correct shape in complexity $\mathcal{O}(n)$, with an overall complexity of $\mathcal{O}(n^3)$.

## 8 The Arnoldi (or Rayleigh-Ritz) approximation of $f(A)b$

In many applications $A$ is large but sparse which makes it impossible to compute $f(A)$ by some direct method as that of §7. For approaching $f(A)b$ we therefore require particular methods

only based on a matrix-vector product for $A$, or perhaps on solving systems with some shifted counterpart of $A$.
For a projection method, we dispose of an orthonormal basis $v_1, ..., v_m$ of a linear subspace $\mathcal{K}_m \subset C^n$, where we suppose for convenience that $v_1 = b/\|b\|$ (in what follows we will suppose without loss of generality that $\|b\| = 1$). The orthonormal basis will be arranged in a matrix $V_m = (v_1, ..., v_m) \in \mathbb{C}^{n\times m}$. The Arnoldi (or Rayleigh-Ritz) approximant of $f(A)b$ is given by the expression

$$x_m = V_m f(A_m) V_m^* b, \quad \text{with } A_m := V_m^* A V_m \in \mathbb{C}^{m\times m}.$$

Notice that $V_m^* b = e_1$ the first canonical vector in $\mathbb{C}^m$. Also, using 2.5 one easily checks that $x_m = V_m f(A_m) V_m^* V_m V_m^* b = f(V_m V_m^* A V_m V_m^*) V_m V_m^* b = f(V_m V_m^* A) b$ with the orthogonal projector $V_m V_m^*$ onto $\mathcal{K}_m$, but, in general, $x_m$ is different[7] from the projection of $f(A)b$ onto $\mathcal{K}_m$. The computational cost for $x_m$ becomes affordable as long as $m$ is of moderate size: here one computes (the first column of) $f(A_m)$ by some direct method, and needs to store $V_m$. However, we need to insure that $f$ is defined on $\sigma(A_m)$, for which it will be convenient to suppose in what follows that $f$ is defined in some neighborhood of $W(A)$ (since $W(A_m) \subset W(A)$).

**8.1. Polynomial Krylov spaces:**
*Consider $\mathcal{K}_m = \mathcal{K}_m(A; b) = span\{b, Ab, A^2 b, ..., A^{m-1} b\}$, here supposed to be of dimension $m$. Given an orthonormal basis $v_1, ..., v_j$ of $\mathcal{K}_j \subset \mathcal{K}_{j+1}$, it will be therefore sufficient to find $v_{j+1} \in \mathcal{K}_{j+1}$ of unit norm which is orthogonal to $v_1, ..., v_j$. This is accomplished by the Arnoldi method [13], by making $Av_j$ orthogonal to $v_1, ..., v_j$ and by normalizing. In other words,*

$$h_{j+1,j} v_{j+1} = A v_j - h_{1,j} v_1 - ... - h_{j,j} v_j,$$

*which can be written in matrix form as $AV_m = V_{m+1} \overline{H_m}$ with $\overline{H_m} \in \mathbb{C}^{(m+1)\times m}$ having upper Hessenberg structure. In particular,*[8]

$$A_m = V_m^* A V_m = [I_m, 0] \overline{H_m} =: H_m.$$

**8.2. Lemma: exactness property for polynomial Krylov**
*For all $f \in \mathcal{P}_{m-1}$ we have that the Arnoldi approximation $x_m = V_m f(A_m) V_m^* b$ coincides with $f(A)b$.*

*Proof.* It is sufficient to show this statement for $f_k(z) = z^k$, $k = 0, 1, ..., m-1$, which will be done by reccurence on $k$: since $b \in \mathcal{K}_m$, we have that $V_m f_0(A_m) V_m^* b = V_m V_m^* b = f_0(A)b$.
By construction of $\mathcal{K}_m$, there holds $f_k(A) \in \mathcal{K}_m$, and thus for $k \geq 1$

$$f_k(A)b = V_m V_m^* A f_{k-1}(A) b = V_m V_m^* A V_m f_{k-1}(A_m) V_m^* b = V_m f_k(A_m) V_m^* b.$$

□

A combination of 8.2 with 2.5 shows that the Arnoldi error $f(A)b - x_m$ is indeed an interpolation error evaluated at $A$:

**8.3. Corollary: Arnoldi error is interpolation error**
*Let $p_{m-1} \in \mathcal{P}_{m-1}$ be an interpolant for $(f, A_m)$, then $V_m f(A_m) V_m^* b = p_{m-1}(A)b$.*

[7] Taking $f(z) = 1/z$ leads to $x_m = V_m A_m^{-1} V_n^* b$ or, equivalently, to the requirement that $x_m \in \mathcal{K}_m$ is such that its residual $b - Ax_m$ is orthogonal to $\mathcal{K}_m$, whereas $V_m V_m^* A^{-1} b$ is the element of $K_m$ being closest to $A^{-1}b$ (minimal error).

[8] We see from this formula that, for symmetric $A$, also $H_m$ is symmetric, in other words, the above formula becomes the Lanczos three term recurrence relation, since $h_{1,j} = ... = h_{j-2,j} = 0$, and $h_{j-1,j} = h_{j,j-1}$ has been already computed in the iteration before. Because of potential rounding errors, it might be however suitable to keep the above full orthogonalization procedure.

At least for symmetric $A$ (and thus $A_m$) we can conclude that the Arnoldi error will be small if the eigenvalues of $A_m$ (also called Ritz values) represent "well" the eigenvalues of $A$. Of course in general there are not enough Ritz values to approach each eigenvalue of $A$, but for a small Arnoldi error it could be sufficient to insure that the Lebesgue function of the Ritz values is not too large on the eigenvalues, see below.
Let us consider a generalization.

**8.4. Rational Krylov spaces:**
*Given some parameters $z_1, z_2, z_3, ... \in \mathbb{C} \cup \{\infty\}$ with $z_j \neq 0$ (otherwise translation), one defines*

$$\widetilde{\mathcal{K}}_m = \mathcal{K}_m(A, q_{m-1}(A)^{-1}b) = \Big\{ \frac{p}{q_{m-1}}(A)b : p \in \mathcal{P}_{m-1} \Big\}, \quad q_m(z) = \prod_{j=1}^{m}(1 - \frac{z}{z_j}),$$

*provided that $q_{m-1}(A)$ is invertible (which is for instance true if $z_j \notin W(A)$). Notice that $\widetilde{\mathcal{K}}_m = \mathcal{K}_m(A, b)$ for the parameters $z_1 = ... = z_{m-1} = \infty$. We suppose as before that $\widetilde{K}_m$ is of dimension $m$, and denote by $v_1, ..., v_m$ an orthogonal basis of $\widetilde{\mathcal{K}}_m$.*
*In practice, one does not compute such a basis directly via 8.1 in forming the vector $q_{m-1}(A)^{-1}b$. Following the original work of Ruhe, for computing $v_{j+1}$ from the orthogonal basis of $\widetilde{\mathcal{K}}_j$, one does merely choose a continuation vector $v \in \mathcal{K}_j$, orthogonalizes $A(I - A/z_j)^{-1}v$ (obtained by solving a system of linear equations) against $v_1, ..., v_j$, and normalizes. Because of the occurrence of the resolvent, the link between the matrix of recurrence coefficients and the matrix $A_m = V_m^* A V_m$ is more complicated, we omit details.*

**8.5. Lemma: exactness property for rational Krylov**
*For all $f \in \mathcal{P}_{m-1}/q_{m-1}$ we have that the rational Arnoldi approximation $x_m = V_m f(A_m) V_m^* b$ coincides with $f(A)b$.*

*Proof.* Write $c = q_{m-1}(A)^{-1}b$, and let us denote by $\widetilde{V}_m$ the matrix having as columns an orthogonal basis of $\mathcal{K}_m(A, c)$, and $\widetilde{A}_m = \widetilde{V}_m^* A \widetilde{V}_m$. Since we have two orthogonal bases of the same space, there exists an unitary matrix $U$ of order $m$ such that $V_m = \widetilde{V}_m U$, showing that $x_m = \widetilde{V}_m U f(U^* \widetilde{A}_m U) U^* \widetilde{V}_m^* b = \widetilde{V}_m f(\widetilde{A}_m) \widetilde{V}_m^* b$. Write more explicitly $f = p/q_{m-1}$ with $p \in \mathcal{P}_{m-1}$, we get applying 8.2 that $b = q_{m-1}(A)c = \widetilde{V}_m q_{m-1}(\widetilde{A}) \widetilde{V}_m^* c$ and thus $\widetilde{V}_m^* b = q_{m-1}(\widetilde{A}) \widetilde{V}_m^* c$, implying that

$$r(A)b = p(A)c = \widetilde{V}_m p(\widetilde{A}_m) \widetilde{V}_m^* c = \widetilde{V}_m f(\widetilde{A}_m) q_{m-1}(\widetilde{A}) \widetilde{V}_m^* c = x_m.$$

□

**8.6. Corollary: rational Arnoldi error is interpolation error**
*Let $p_{m-1}/q_{m-1} \in \mathcal{P}_{m-1}/q_{m-1}$ be an interpolant for $(f, A_m)$, then $V_m f(A_m) V_m^* b = \frac{p_{m-1}}{q_{m-1}}(A)b$.*

**8.7. Remark: link with orthogonal rational functions**
*Consider the scalar product for rational functions*[9]

$$\langle P, Q \rangle = \Big( Q(A)b \Big)^* \Big( P(A)b \Big).$$

*By construction,*

$$\exists\, \varphi_j \in \mathcal{P}_j/q_j, \quad v_{j+1} = \varphi_j(A)b, \quad \langle \varphi_j, \varphi_k \rangle = \delta_{j,k},$$

*and thus these rational functions $\varphi_0, ..., \varphi_{m-1}$ are orthogonal rational functions (ORF). According to 8.5 we know that $e_{j+1} = V_m^* v_{j+1} = V_m^* V_m \varphi(A_m) V_m^* b = \varphi_j(A_m) e_1$.*

[9] By the present assumptions, this "scalar product" is only positive definite on $\mathcal{P}_{m-1}/q_{m-1}$, but let us consider a sufficiently small $m$.

*Let us introduce the auxiliary vectors* $\widetilde{v}_{m+1} = \widetilde{\varphi}_m(A)b$, $\widetilde{\varphi}_m \in \mathcal{P}_m/q_{m-1}$ *by supposing that*

$$v_1, ..., v_m, \widetilde{v}_{m+1} \text{ is an orthonormal basis for the poles } z_1, z_2, ..., z_{m-1}, \widetilde{z}_m = \infty.$$

*We show below that the numerator* $\widetilde{\varphi}_m q_{m-1}$ *is a non-trivial multiple of the characteristic polynomial* $\chi$ *of* $A_m$ *which coincides with the minimal polynomial of* $A_m$*. Hence the rational interpolant in 8.6 is unique, and in addition we may use the theory of zeros of orthogonal rational functions to learn more about*[10] *how (rational) Ritz values approach eigenvalues of* $A$*.*

*Proof.* Since $\chi/q_{m-1} \in \mathcal{P}_m/q_{m-1}$ with orthonormal basis $\varphi_0, ..., \varphi_{m-1}, \widetilde{\varphi}_m$, there exist $c, c_0, ..., c_{m-1} \in \mathbb{C}$ such that

$$\frac{\chi}{q_{m-1}} = c\widetilde{\varphi}_m + \sum_{j=0}^{m-1} c_j \varphi_j.$$

Factorizing $\chi(z) = (z-a)\widetilde{\chi}(z)$ with $\widetilde{\chi} \in \mathcal{P}_{m-1}$, we may apply 8.5 and obtain for $j \in \{0, ..., m-1\}$

$$\begin{aligned} c_j &= \langle \frac{\chi}{q_{m-1}}, \varphi_j \rangle = v_{j+1}^*(A - aI)\frac{\widetilde{\chi}}{q_{m-1}}(A)b \\ &= e_{j+1}^* V_m^*(A - aI)V_m \frac{\widetilde{\chi}}{q_{m-1}}(A_m)e_1 = e_{j+1}^* \frac{\chi}{q_{m-1}}(A_m)e_1 \end{aligned}$$

and thus $c_j = 0$ since $\chi(A_m) = 0$. Thus $\chi/q_{m-1}$ is a (non-trivial) multiple of $\widetilde{\varphi}_m$, as claimed above. If the minimal polynomial of $A_m$ would be of degree $< m$, we could expand it as in the first part of the proof, but now with $c = 0$, and the same argument shows that such a polynomial must be trivial, a contradiction. □

# References


[1] Claude Brezinski. *Computational Aspects of Linear Control.* Kluwer, Dordrecht, 2002.

[2] Michel Crouzeix. Bounds for analytical functions of matrices. *Integral Equations and Operator Theory*, 48:461–477, 2004.

[3] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[4] Anne Greenbaum. *Iterative Methods for Solving Linear Systems.* SIAM, Philadelphia, Baltimore, Maryland, USA, 1997.

[5] Nicholas Hale, Nicholas J. Higham, and Lloyd N. Trefethen. Computing $A^\alpha$, $\log(A)$ and related matrix functions by contour integrals. MIMS EPrint 2007.103, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, August 2007. To appear in SIAM J. Numer. Anal.

[6] Nicholas J. Higham. *Functions of Matrices: Theory and Computation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[7] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.


[10] This task might be tricky for general matrices, but at least for $A$ normal we may write our scalar product as $\langle P, Q \rangle = \int P(z)\overline{Q(z)}\, d\mu(z)$ with a positive discrete measure $\mu$ supported on (a subset of) $\sigma(A)$. Thus one remains with the question how roots of ORF do approach the discrete support of orthogonality.

[8] George A. Baker Jr. and Peter Graves-Morris. *Padé Approximants*, volume 59 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, second edition, 1996.

[9] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, second edition, 1976.

[10] Cleve B. Moler and Charles F. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49, 2003.

[11] Frigyes Riesz and Béla Sz.-Nagy. *Functional Analysis*. Blackie & Son, London and Glasgow, second edition, 1956.

[12] Youcef Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, Manchester, and Halsted Press, New York, 1992.

[13] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2003.

[14] Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, NJ, USA, 2005.

[15] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.