Convergence Estimates of Krylov Subspace Methods for the Approximation of Matrix Functions Using Tools from Potential Theory

Güttel, Stefan

2006

MIMS EPrint: **2015.34**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Diploma Thesis

Convergence Estimates of Krylov Subspace Methods for the Approximation of Matrix Functions Using Tools from Potential Theory

Stefan Güttel

June 2006

Convergence Estimates of Krylov Subspace Methods for the Approximation of Matrix Functions Using Tools from Potential Theory

Contents

Preface

This work is about the numerical evaluation of the expression

$$f(A)\boldsymbol{b},$$

where $A \in \mathbb{C}^{N\times N}$ is an arbitrary square matrix, $\boldsymbol{b} \in \mathbb{C}^N$ is a vector and f is a suitable matrix function. This task is of very high importance in all applied sciences since it is a generalization of the following problems, to name just a few:

- **Solve the linear system of equations $A\boldsymbol{x} = \boldsymbol{b}$.**
  The solution is $\boldsymbol{x} = f(A)\boldsymbol{b}$, where $f(z) = 1/z$.
- **Solve an ordinary differential equation $\boldsymbol{y}'(t) = A\boldsymbol{y}(t)$ with given initial value $\boldsymbol{y}(0) = \boldsymbol{b}$.** The solution is $\boldsymbol{y}(t) = f(tA)\boldsymbol{b}$, where $f(z) = \exp(z)$.
- **Solve identification problems in stochastic semigroups.** Here one needs to compute $f(A)\boldsymbol{b}$ with $f(z) = \log(z)$ (see Singer, Spilermann [29]).
- **Simulate Brownian motion of molecules.** Here one needs to determine $f(A)\boldsymbol{b}$ with $f(z) = \sqrt{z}$ (see Ericsson [9]).

In the first chapter we will define the term $f(A)$. There are different equivalent approaches. A constructive one is to involve the *Jordan canonical form* of the matrix A. Later we shall see that $f(A) = p_{f,A}(A)$, where $p_{f,A}$ is a polynomial of degree $\leq N-1$ that interpolates f at the eigenvalues of A. In practical applications N is very large and the spectrum of A is not known. Therefore we will determine an f-interpolating polynomial $p_{f,m}$ of low degree $m - 1 \ll N$ and hope that

$$p_{f,m}(A)\boldsymbol{b} \approx f(A)\boldsymbol{b}.$$

The resulting methods are called *Krylov subspace methods* or *polynomial methods* and they are considered in Chapter 2.

The choice of the interpolation nodes for $p_{f,m}$ is an important issue. If the interpolation nodes are *uniformly distributed* on a compact subset of $\mathbb{C}$, we may analyze the *asymptotic* convergence behavior of the arising methods using theory of *interpolation* and *best approximation*. This is done in Chapter 3.

Another very popular choice of interpolation nodes are *Ritz values*. The resulting *Arnoldi approximations* converge in many cases very fast to $f(A)\boldsymbol{b}$. To explain this, it is necessary to describe the behavior of Ritz values. In Chapter 4 we will present a theory on the *convergence of Ritz values*, which was mainly developed by Beckermann and Kuijlaars (see [1, 18, 19]). This theory involves tools from *potential theory*.

Files

All computations in this work have been carried out using MATHWORKS MATLAB, VERSION 6.5 R13. The operating system was MICROSOFT WINDOWS XP PROFESSIONAL, SP 2. The necessary files can be found on the attached CD-ROM. All figures may be reproduced by the reader and are marked by a symbol and an *identifier* at the right margin of the page. A graphical user interface for the easy access to the corresponding `.m`-files is provided. It is executed from the command line of MATLAB by typing


Identifier

```
>> cd X:        % change to CD-ROM drive X
>> cd mat
>> guirun
```

The `.m`-files may also be accessed directly from the subfolder `FILES/Identifier` by running `rundemo.m`. Note that some of the examples require the *Schwarz-Christoffel-toolbox* written by Driscoll [3], which should be added to the *search patch* of MATLAB. Moreover, access to the MAPLE kernel is necessary for some files.

Additionally, the CD-ROM contains this document and two presentations about the subject of this work as `.pdf`- and LaTeX-files, as well as all the figures shown here. For a detailed *File List* we refer to page 97.

Some Words about Notation

Throughout this work, matrices are uppercase letters and vectors are bold lowercase letters. The null matrix is denoted by O and the identity matrix by I. $\boldsymbol{\xi}_m$ denotes the m-th unit coordinate vector, whereas $\boldsymbol{e}_m$ denotes the error and $\boldsymbol{u}_m$ is a column of an unitary matrix. Mainly in the first chapter we will use the space-saving toep-operator. It takes a vector argument and constructs a Toeplitz matrix (i.e., a matrix with constant diagonal entries) by using the vector entries as diagonal values, where the main diagonal value is underlined. If the size of the constructed matrix is not clear from the context, it follows the toep-operator:

$$\operatorname{toep}(1,\underline{2},3,4) = \begin{bmatrix} 2 & 3 & 4 & \\ 1 & 2 & 3 & 4 \\ & 1 & 2 & 3 \\ & & 1 & 2 \end{bmatrix} \in \mathbb{C}^{4\times 4}, \quad \operatorname{toep}(1,\underline{2},3,4) = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix} \in \mathbb{C}^{2\times 2}.$$

The operator diag arranges (block-)diagonal matrices by taking a list of matrices $J_1, J_2, \ldots, J_k$:

$$\operatorname{diag}(J_1, J_2, \ldots, J_k) = \begin{bmatrix} \boxed{J_1} & & & \\ & \boxed{J_2} & & \\ & & \ddots & \\ & & & \boxed{J_k} \end{bmatrix}.$$

On page 98 we give a detailed overview about the symbols used here.

Acknowledgements

This Diploma Thesis was written during my studies at the *Technische Universität Bergakademie Freiberg* and the *University of Cyprus*. I owe a particular dept of gratitude to *Prof. Dr. Michael Eiermann* and *Prof. Dr. Nikos Stylianopoulos* for introducing me to Krylov subspaces and potential theory, and for their invaluable support, not only in mathematical matters. I thank *PD Dr. Oliver Ernst* for the patient proof-reading of the manuscript. I am grateful to *Anna* for her encouragements and private language courses, although still it is all Greek to me. And finally, I wish to say a big thank-you to *my parents* for their patience and care.

1 Matrix Functions

The aim of this chapter is to give a meaning to the term $f(A)$, where $A \in \mathbb{C}^{N\times N}$ is a given square matrix and $f(z)$ is a complex-valued function of a complex variable $z \in \mathbb{C}$. We explain which requirements f has to fulfill in order that $f(A)$ is defined and how it is defined. Since there are different definitions, we have to clarify in which sense they are compatible to each other: some of them hold only for polynomials, others hold only if f is analytic in a domain that contains the eigenvalues of A, etc. The most common and constructive approach involves the Jordan canonical form of a matrix. Another very important viewpoint, especially for the following chapters, is to consider f as an interpolation polynomial. At the end of this chapter we list some properties of matrix functions.

1.1 Polynomial Matrix Functions

Let $p(z)$ be a polynomial of degree m with complex coefficients α_j, i.e., $p(z) = \alpha_m z^m + \alpha_{m-1} z^{m-1} + \cdots + \alpha_0$. This will be denoted by $p(z) \in \mathcal{P}_m(z)$. Since the powers $I, A, A^2, \ldots$ exist, we may insert A in p and the following definition is justified.

Definition 1.1. *$p(A)$ is defined as*

$$\boxed{p(A) := \alpha_m A^m + \alpha_{m-1} A^{m-1} + \cdots + \alpha_0 I \in \mathbb{C}^{N\times N}.} \tag{D1}$$

We say p is a polynomial matrix function.

We no longer have to distinguish between $\mathcal{P}_m(z)$ and the set of polynomials in A of degree $\leq m$. We simply write $\mathcal{P}_m$. In the following lemma we summarize some important properties of polynomial matrix functions.

Lemma 1.2. *Let $p \in \mathcal{P}_m$ be a polynomial, $A \in \mathbb{C}^{N\times N}$ and $A = TJT^{-1}$, where $J = \operatorname{diag}(J_1, J_2, \ldots, J_k)$ is block-diagonal. Then*

(i) $p(A) = Tp(J)T^{-1}$,

(ii) $p(J) = \operatorname{diag}\left(p(J_1), p(J_2), \ldots, p(J_k)\right)$,

(iii) *If* $A\boldsymbol{v} = \lambda\boldsymbol{v}$ *then* $p(A)\boldsymbol{v} = p(\lambda)\boldsymbol{v} \qquad (\boldsymbol{v} \in \mathbb{C}^N)$,

(iv) *Given another polynomial* $\tilde{p} \in \mathcal{P}_{\tilde{m}}$ *then* $p(A)\tilde{p}(A) = \tilde{p}(A)p(A)$.

Proof. **(i)** By (D1) we have

$$\begin{aligned} p(A) &= p(TJT^{-1}) \\ &= \alpha_m \left(TJT^{-1}\right)^m + \alpha_{m-1}\left(TJT^{-1}\right)^{m-1} + \cdots + \alpha_0 I \\ &= T\left(\alpha_m J^m + \alpha_{m-1}J^{m-1} + \cdots + \alpha_0 I\right)T^{-1} \\ &= Tp(J)T^{-1}. \end{aligned}$$

(ii) Powers of block-matrices do not alter the block-structure.

(iii) Using (D1) we obtain

$$\begin{aligned} p(A)\boldsymbol{v} &= \alpha_m A^m\boldsymbol{v} + \alpha_{m-1}A^{m-1}\boldsymbol{v} + \cdots + \alpha_0\boldsymbol{v} \\ &= \alpha_m \lambda^m\boldsymbol{v} + \alpha_{m-1}\lambda^{m-1}\boldsymbol{v} + \cdots + \alpha_0\boldsymbol{v} \\ &= p(\lambda)\boldsymbol{v}. \end{aligned}$$

(iv) holds because powers of A commute: $A^\mu A^\nu = A^\nu A^\mu$. □

1.2 The Jordan Canonical Form

A factorization $A = TJT^{-1}$ with $J = \operatorname{diag}(J_1, J_2, \ldots, J_k)$ can always be found. Every square matrix A is similar to a block-diagonal *Jordan matrix* J, where each *Jordan block* $J_j = J_j(\lambda_j) \in \mathbb{C}^{n_j\times n_j}$ has entries λ_j on the main diagonal and 'ones' on the first upper diagonal ($j = 1, 2, \ldots, k$):

$$J_j(\lambda_j) := \operatorname{toep}(\underline{\lambda_j}, 1) = \begin{bmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_j & 1 \\ & & & & \lambda_j \end{bmatrix}.$$

We say $J = T^{-1}AT$ is a *Jordan canonical form* (JCF) of A. The numbers λ_j are the eigenvalues of A and the columns of T are the corresponding *generalized eigenvectors*. In general the computation of a JCF is very expensive and unstable. Nevertheless it will be useful to extend our definition of polynomial matrix functions to wider function classes.

Assume that the JCF of A consists of one single Jordan block, i.e.,

$$J := \mathrm{toep}(\underline{\lambda}, 1) \in \mathbb{C}^{n\times n},\ n = N.$$

Let $p_m(z) := z^m$ be the *monomial of degree* m. Then $p_m(J)$ is an upper triangular Toeplitz matrix and its i-th diagonal[1] contains the values $\binom{m}{i}\lambda^{m-i}$. In other words,

$$p_m(J) = \mathrm{toep}\left(\underline{\binom{m}{0}\lambda^m}, \ldots, \binom{m}{i}\lambda^{m-i}, \ldots, \binom{m}{m}\lambda^0\right) \in \mathbb{C}^{n\times n}. \tag{1.1}$$

To explain this we write $J = \lambda I + E$ with $E := \mathrm{toep}(\underline{0}, 1) \in \mathbb{C}^{n\times n}$ and note that $E^0 = I$, $E^2 = \mathrm{toep}(\underline{0}, 0, 1)$, $E^3 = \mathrm{toep}(\underline{0}, 0, 0, 1), \ldots$ and $E^m = O$ for $m \geq n$. Because I and E commute we may apply the Binomial Theorem, resulting in

$$p_m(J) = (\lambda I + E)^m = \sum_{i=0}^{m} \binom{m}{i}\lambda^{m-i}E^i,$$

from which the assertion (1.1) follows.

We observe that

$$p_m^{(i)}(\lambda) = \frac{m!}{(m-i)!}\lambda^{m-i} = i!\binom{m}{i}\lambda^{m-i}.$$

Here $p_m^{(i)}$ is the i-th derivative of the function p_m. Note that $p_m^{(i)} \equiv 0$ if $i > m$. Consequently, (1.1) can be rewritten as

$$p_m(J) = \mathrm{toep}\left(\underline{p_m(\lambda)}, \ldots, \frac{p_m^{(i)}(\lambda)}{i!}, \ldots, \frac{p_m^{(n-1)}(\lambda)}{(n-1)!}\right).$$

Finally, we replace p_m by a function $f : \mathbb{C} \supseteq D \to \mathbb{C}$ and find that $f(J)$ is well defined if $f(\lambda), f'(\lambda), \ldots, f^{(n-1)}(\lambda)$ exist. We are led to give the following definition of $f(A)$ (see Lancaster, Tismenetsky [20]).

[1] The i-th diagonal of a matrix $M = [m_{\mu,\nu} : 1 \leq \mu, \nu \leq n]$ contains all entries $m_{\mu,\nu}$ that satisfy $\nu - \mu = i$. This is MATLAB-enumeration of diagonals.

Definition 1.3. *Given $A \in \mathbb{C}^{N\times N}$ with a Jordan canonical form $J = T^{-1}AT$, where $J = \operatorname{diag}(J_1, J_2, \ldots, J_k)$ and $J_j = J_j(\lambda_j) \in \mathbb{C}^{n_j\times n_j}$ $(j = 1, 2, \ldots, k)$. Let U be an open subset of $\mathbb{C}$ such that $\{\lambda_1, \lambda_2, \ldots, \lambda_k\} \subset U$. Let f be a function $f : U \subseteq D \to \mathbb{C}$.*

We say that f is defined for *A if $f(\lambda_j), f'(\lambda_j), \ldots, f^{(d_{\lambda_j}-1)}(\lambda_j)$ exist, where d_{λ_j} is the size of the largest Jordan block associated with the eigenvalue λ_j.*

If f is defined for A we set

$$\boxed{f(A) := T \operatorname{diag}\left(f(J_1), f(J_2), \ldots, f(J_k)\right) T^{-1},} \tag{D2}$$

where

$$\boxed{f(J_j) := \operatorname{toep}\left(\underline{f(\lambda_j)}, \ldots, \frac{f^{(i)}(\lambda_j)}{i!}, \ldots, \frac{f^{(n_j-1)}(\lambda_j)}{(n_j-1)!}\right).} \tag{D2'}$$

Remarks 1.4.

(i) *$f(A)$ is uniquely determined by* (D2).

Proof. The JCF is unique up to a permutation of the Jordan blocks (see Meyer [22]). Given another JCF $\tilde{J} = \tilde{T}^{-1}A\tilde{T}$. Then there exists a *permutation matrix* $P \in \{0,1\}^{N\times N}$, $P^TP = I$, such that $\tilde{J} = PJP^T$ and $\tilde{T} = TP^T$. Therefore

$$\tilde{T}f(\tilde{J})\tilde{T}^{-1} = f(\tilde{T}\tilde{J}\tilde{T}^{-1}) = f(TP^TPJP^TPT^{-1}) = f(TJT^{-1}) = f(A).$$

□

(ii) *By $\Lambda(A)$ we denote the* spectrum of A. *The* minimal polynomial of A *is defined as*

$$\psi_A(z) := \prod_{\lambda\in\Lambda(A)} (z-\lambda)^{d_\lambda}.$$

ψ_A is the monic polynomial of smallest degree $d := \sum_{\lambda\in\Lambda(A)} d_\lambda$ that annihilates A (i.e., $\psi_A(A) = O$)[2]. ψ_A is uniquely determined.

Proof. T is invertible. Therefore $\psi_A(A) = T\psi_A(J)T^{-1} = O$ if and only if $\psi_A(J) = O$. Let $J_\star(\lambda)$ be a largest Jordan block to an eigenvalue λ. Then all other Jordan blocks J_j to the same eigenvalue are leading submatrices

[2]Indeed, this is an equivalent definition of the minimal polynomial.

of $J_\star$ and all matrices $\psi_A(J_j)$ are leading submatrices of $\psi_A(J_\star)$. Hence, it is sufficient to prove that $\psi_A(J_\star) = O$. But this is obvious since ψ_A has a root λ of multiplicity d_λ:

$$\begin{aligned}\psi_A(J_\star(\lambda)) &= \operatorname{toep}\left(\underline{\psi_A(\lambda)}, \ldots, \frac{\psi_A^{(i)}(\lambda)}{i!}, \ldots, \frac{\psi_A^{(d_\lambda-1)}(\lambda)}{(d_\lambda-1)!}\right) \\ &= \operatorname{toep}(\underline{0}, \ldots, 0, \ldots, 0).\end{aligned}$$

Conversely, we assume that λ is only a root of multiplicity $\nu \leq d_\lambda - 1$. Then $\psi_A^{(\nu)}(\lambda) \neq 0$ and $\psi_A(J_\star(\lambda)) \neq O$. Therefore ψ_A is not annihilating A and thus not a minimal polynomial of A, which is a contradiction.
Now we prove the uniqueness of ψ_A. Assume that $\tilde{\psi}_A$ is another minimal polynomial of A. Then by definition $\deg(\tilde{\psi}_A) = d$. Consequently, $\beta(\tilde{\psi}_A - \psi_A)$ is a monic polynomial (for some scaling constant $0 \neq \beta \in \mathbb{C}$) of degree $< d$ that annihilates A. This is a contradiction to ψ_A and $\tilde{\psi}_A$ having the minimal degree d. Thus, uniqueness is proven. □

(iii) *By χ_A we denote the* characteristic polynomial of A,

$$\chi_A(z) := \det(zI - A) = \prod_{j=1}^{k}(z - \lambda_j)^{n_j}.$$

Obviously, ψ_A is a divisor of χ_A, i.e.,

$$\psi_A \mid \chi_A\,. \tag{1.2}$$

(iv) *If all the λ_j are pairwise distinct (i.e., each eigenvalue occurs in exactly one Jordan block) then*

$$\psi_A(z) = \prod_{j=1}^{k}(z - \lambda_j)^{n_j} = \chi_A(z).$$

Such matrices are called nonderogatory.

(v) *By construction it is clear that for monomials $p_m(z) = f(z)$*

$$(\mathrm{D1}) = (\mathrm{D2}).$$

But the equivalence of both definitions also persists for all $p \in \mathcal{P}_m$. Given $p(z) = \sum_{j=0}^{m} \alpha_j z^j$ then by Lemma 1.2 we obtain

$$p(A) = \sum_{j=0}^{m} \alpha_j p_j(A) = T\left(\sum_{j=0}^{m} \alpha_j p_j(J)\right) T^{-1} = Tp(J)T^{-1}.$$

1.3 Polynomial Interpolation I

The following theorem clarifies the connection between matrix functions and interpolation polynomials.

Theorem 1.5.

(i) *There holds*

$$f(A) = p(A)$$

if and only if

$$\boxed{f^{(i)}(\lambda) = p^{(i)}(\lambda) \qquad \text{for all } \lambda \in \Lambda(A),\ i = 0, 1, \ldots, d_\lambda - 1.} \tag{HIP}$$

These are $d := \deg(\psi_A)$ *interpolation conditions on* p.

(ii) *There exists a uniquely determined polynomial* $p_{f,A} \in \mathcal{P}_{d-1}$ *that satisfies* (HIP). *We say* $p_{f,A}$ *is the* Hermite interpolating polynomial satisfying (HIP).

(iii) *Every polynomial* p *that satisfies* (HIP) *can be represented in the form*

$$p(z) = p_{f,A}(z) + \psi_A(z)h(z)$$

for some polynomial $h(z)$.

Proof. **(i)** Let $J = T^{-1}AT$ be a JCF of A, $J = \operatorname{diag}(J_1, J_2, \ldots, J_k)$. Clearly, $f(A) = p(A)$ if and only if $f(J) = p(J)$. By definition,

$$\begin{aligned} f(J_j) &= \operatorname{toep}\left(\underline{f(\lambda_j)}, \ldots, \frac{f^{(i)}(\lambda_j)}{i!}, \ldots, \frac{f^{(n_j-1)}(\lambda_j)}{(n_j-1)!}\right) \\ &\overset{\text{(HIP)}}{=} \operatorname{toep}\left(\underline{p(\lambda_j)}, \ldots, \frac{p^{(i)}(\lambda_j)}{i!}, \ldots, \frac{p^{(n_j-1)}(\lambda_j)}{(n_j-1)!}\right) \\ &= p(J_j), \end{aligned}$$

where the second equality holds for all $j = 1, 2, \ldots, k$ if and only if the interpolation conditions (HIP) are satisfied.

(ii) Given another polynomial $\hat{p}_{f,A}$ that satisfies (HIP). Then $\hat{p}_{f,A} - p_{f,A}$ has d roots (counted by multiplicities) and thus $\hat{p}_{f,A} - p_{f,A} \equiv 0$. Hence $p_{f,A}$ is unique. Now existence follows from uniqueness: With

$$p_{f,A}(z) = \alpha_{d-1}z^{d-1} + \alpha_{d-2}z^{d-2} + \cdots + \alpha_0,$$

(HIP) is a system of d linear equations for d unknowns α_i and can be written as $M\boldsymbol{\alpha} = \boldsymbol{f}$, where $M \in \mathbb{C}^{d\times d}$, $\boldsymbol{f} \in \mathbb{C}^d$ and $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_{d-1}]^T$. M is invertible because of the already proven uniqueness of $p_{f,A}$. Hence $\boldsymbol{\alpha} = M^{-1}\boldsymbol{f}$ exists.

(iii) Set $r(z) := p(z) - p_{f,A}(z)$. Since $p(A) = p_{f,A}(A)$ due to (i), we have $r(A) = O$. Thus, r must contain a factor ψ_A, i.e., $r(z) = \psi_A(z)h(z)$. Conversely, if $p(z) = p_{f,A}(z) + \psi_A(z)h(z)$ then $p(A) = p_{f,A}(A) + \psi_A(A)h(A) = p_{f,A}(A) + O = p_{f,A}(A)$. □

Example 1.6. *Let $A = [\alpha]$ for some constant $\alpha \in \mathbb{C}$. Then $\psi_A(z) = z - \alpha$ and $\deg(\psi_A) = 1$. Therefore $f(A) = p_{f,A}(A)$ with $\deg(p_{f,A}) = 0$, namely $p_{f,A}(z) = f(\alpha)I$. This is a degenerate case.*

Example 1.7. *Find a polynomial p such that $p(A) = \exp(A)$, where*

$$A = \begin{bmatrix} 1 & 6 & 4 & 0 & -8 \\ 0 & 7 & 4 & 0 & -8 \\ 2 & 0 & -1 & -1 & -2 \\ 2 & -4 & 0 & 0 & 2 \\ 2 & 6 & 3 & -1 & -9 \end{bmatrix}.$$

A Jordan canonical form of A is $J = T^{-1}AT$, where

$$T = \begin{bmatrix} 1 & -4 & 2 & -4 & 1 \\ 0 & -4 & 2 & -4 & 1 \\ 0 & -2 & -1 & -1 & 0 \\ 2 & 2 & 2 & 1 & 0 \\ 0 & -5 & 2 & -4 & 1 \end{bmatrix} \text{ and } J = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

We read off: J has $k = 4$ Jordan blocks J_j to the associated eigenvalues $\lambda_1 = 1, \lambda_2 = -1, \lambda_3 = 0, \lambda_4 = -1$. The sizes n_j of the blocks are $n_1 = n_3 = n_4 = 1$ and $n_2 = 2$. Thus, $d_1 = d_0 = 1$ and $d_{-1} = 2$. The minimal polynomial of A is

$$\psi_A(z) = (z-1)(z+1)^2 z\,.$$

We have to determine a polynomial $p(z)$ that satisfies the following $d = 4$ Hermite interpolation conditions*:*

$$\begin{aligned} p(\lambda_1) = p(1) &\overset{!}{=} \exp(1) = \mathrm{e}, \\ p(\lambda_2) = p(-1) &\overset{!}{=} \exp(-1) = 1/\mathrm{e}, \\ p'(\lambda_2) = p'(-1) &\overset{!}{=} \exp(-1) = 1/\mathrm{e}, \\ p(\lambda_3) = p(0) &\overset{!}{=} \exp(0) = 1. \end{aligned}$$

A solution is

$$p(z) = \frac{\mathrm{e}^2 - 4\mathrm{e} + 5}{4\mathrm{e}} z^3 + \frac{(\mathrm{e}-1)^2}{2\mathrm{e}} z^2 + \frac{\mathrm{e}^2 + 4\mathrm{e} - 7}{4\mathrm{e}} z + 1$$

and there holds

$$p(A) = \exp(A).$$

Because $p \in \mathcal{P}_{d-1}$ we already found the unique Hermite interpolating polynomial.

Remarks 1.8.

(i) *Every matrix function $f(\cdot)$ can be represented* pointwise *(i.e., for a fixed A) as a polynomial $p_{f,A}(A) \in \mathcal{P}_{d-1}$, $d = \deg(\psi_A)$.*

(ii) *$f(A)$ depends only on the values of $f, f', \ldots$ on $\Lambda(A)$. Thus, $f(A)$ and $f(B)$ have the same polynomial representation if A and B have the same minimal polynomial (e.g., if A and B are similar matrices).*

(iii) *If all Jordan blocks have size 1×1 and thus J is a diagonal matrix (this happens if and only if A has n linearly independent eigenvectors, e.g., if A is normal) then* (HIP) *reduces to a* Lagrange interpolation problem*:*

$$\boxed{f(\lambda) = p(\lambda) \qquad \text{for all } \lambda \in \Lambda(A).} \tag{LIP}$$

1.4 The Components of a Matrix

We want to derive a (more or less) explicit formula for the Hermite interpolating polynomial $p_{f,A} \in \mathcal{P}_{d-1}$ that fulfills (HIP) for a function f (see Theorem 1.5). By the way, this will lead us to another definition of $f(A)$ generalizing the *Cauchy integral formula*. The following derivation has been adapted from Gantmacher [11].

Let

$$\psi_A(z) = \prod_{\lambda \in \Lambda(A)} (z - \lambda)^{d_\lambda}$$

denote the minimal polynomial of A. We represent the rational function $p_{f,A}(z)/\psi_A(z)$, where $\deg(p_{f,A}) < \deg(\psi_A) = d$, as a sum of partial fractions:

$$\frac{p_{f,A}(z)}{\psi_A(z)} = \sum_{\lambda \in \Lambda(A)} \left(\frac{\alpha_{\lambda,0}}{(z-\lambda)^{d_\lambda}} + \frac{\alpha_{\lambda,1}}{(z-\lambda)^{d_\lambda - 1}} + \cdots + \frac{\alpha_{\lambda,d_\lambda - 1}}{(z-\lambda)} \right), \tag{1.3}$$

where $\alpha_{\lambda,i}$ are certain constants we want to determine for $\lambda \in \Lambda(A)$ and $i = 0, 1, \ldots, d_\lambda - 1$. Therefor we multiply both sides of (1.3) by $(z-\lambda)^{d_\lambda}$ and set $\psi_{A,\lambda}(z) := \psi_A(z)/(z-\lambda)^{d_\lambda}$. We obtain

$$\frac{p_{f,A}(z)}{\psi_{A,\lambda}(z)} = \alpha_{\lambda,0} + \alpha_{\lambda,1}(z-\lambda) + \cdots + \alpha_{\lambda,d_\lambda - 1}(z-\lambda)^{d_\lambda - 1} + (z-\lambda)^{d_\lambda} R_\lambda(z),$$

where $R_\lambda(z)$ is a rational function with $R_\lambda(\lambda) \neq \infty$.

From the last equation the following can be easily verified:

$$\alpha_{\lambda,i} = \frac{1}{i!} \left[\frac{p_{f,A}(z)}{\psi_{A,\lambda}(z)} \right]^{(i)}_{z=\lambda}. \tag{1.4}$$

By (HIP) we know that $p^{(i)}_{f,A}(\lambda) = f^{(i)}(\lambda)$ $(\lambda \in \Lambda(A);\ i = 0, 1, \ldots, d_\lambda - 1)$. Furthermore no higher derivatives of $p_{f,A}$ occur in (1.4). Therefore we may replace $p_{f,A}$ by f:

$$\alpha_{\lambda,i} = \frac{1}{i!} \left[\frac{f(z)}{\psi_{A,\lambda}(z)} \right]^{(i)}_{z=\lambda}. \tag{1.5}$$

Hence all the $\alpha_{\lambda,i}$ can be obtained and we may determine $p_{f,A}(z)$ by multiplying (1.3) by $\psi_A(z)$:

$$p_{f,A}(z) = \sum_{\lambda \in \Lambda(A)} \left(\alpha_{\lambda,0} + \alpha_{\lambda,1}(z-\lambda) + \cdots + \alpha_{\lambda,d_\lambda - 1}(z-\lambda)^{d_\lambda - 1} \right) \psi_{A,\lambda}(z). \tag{1.6}$$

By substituting in (1.6) the expressions (1.5) for the coefficients $\alpha_{\lambda,i}$ and gathering the terms that contain the same factor $f^{(i)}(\lambda)$, we may represent $p_{f,A}(z)$ in the form

$$p_{f,A}(z) = \sum_{\lambda \in \Lambda(A)} \left(f(\lambda)\varphi_{\lambda,0}(z) + f'(\lambda)\varphi_{\lambda,1}(z) + \cdots + f^{(d_\lambda - 1)}(\lambda)\varphi_{\lambda,d_\lambda - 1}(z) \right), \tag{1.7}$$

where $\varphi_{\lambda,i} \in \mathcal{P}_{d-1}$ $(\lambda \in \Lambda(A);\ i = 1, 2, \ldots, d_\lambda - 1)$ are d polynomials that are completely determined when $\psi_A(z)$ is given and do not depend on the function f. Choosing functions $f_{\lambda,i}(z)$ such that

$$f_{\lambda,i}^{(\nu)}(z) = \begin{cases} 1, & z = \lambda,\ i = \nu; \\ 0, & \text{otherwise}, \end{cases}$$

for all $z \in \Lambda(A)$, then the associated Hermite interpolating polynomials $p_{\lambda,i}$ fulfill (HIP) by definition, i.e., $p_{\lambda,i}^{(\nu)}(\lambda) = f_{\lambda,i}^{(\nu)}(\lambda)$. Therefore (1.7) yields

$$\varphi_{\lambda,i}^{(\nu)}(z) = \begin{cases} 1, & z = \lambda,\ i = \nu; \\ 0, & \text{otherwise}, \end{cases} \tag{1.8}$$

for all $z \in \Lambda(A)$.

Hence all the $\varphi_{\lambda,i}(z)$ are linearly independent $(f \equiv 0 \Rightarrow p_{f,A} \equiv 0 \Rightarrow f^{(i)}(\lambda) = 0)$. Thus, $\{\varphi_{\lambda,i}(\lambda) : \lambda \in \Lambda(A);\ i = 0, 1, \ldots, d_\lambda - 1\}$ is a basis of $\mathcal{P}_{d-1}$, the *Hermite basis*.

Definition 1.9. *With the polynomials $\varphi_{\lambda,i}$ from above the matrices $C_{\lambda,i} := \varphi_{\lambda,i}(A)$ define the* components of A.

We summarize some properties of the components of a matrix:

Theorem 1.10. *Let $C_{\lambda,i} \in \mathbb{C}^{N\times N}$ $(\lambda \in \Lambda(A);\ i = 0, 1, \ldots, d_\lambda - 1)$ be the components of $A \in \mathbb{C}^{N\times N}$ and let $J = T^{-1}AT = \operatorname{diag}(J_1, J_2, \ldots, J_k)$ be a JCF of A, where $J_j = J_j(\lambda_j) \in \mathbb{C}^{n_j \times n_j}$ for $j = 1, 2, \ldots, k$. Then there holds*

(i) $\{ C_{\lambda,i} : \lambda \in \Lambda(A);\ i = 0, 1, \ldots, d_\lambda - 1 \} \subset \mathbb{C}^{N\times N}$ *is a set of linearly independent matrices,*

(ii)

$$f(A) = \sum_{\lambda\in\Lambda(A)} \sum_{i=0}^{d_\lambda - 1} f^{(i)}(\lambda) C_{\lambda,i}\ , \tag{1.9}$$

(iii) $\sum_{\lambda\in\Lambda(A)} C_{\lambda,0} = I$ *and* $\sum_{\lambda\in\Lambda(A)} \lambda C_{\lambda,0} + C_{\lambda,1} = A$,

(iv) $C_{\lambda,i} C_{\mu,j} = C_{\mu,j} C_{\lambda,i}$,

(v) $C_{\lambda,i} = T\,\mathrm{diag}(D_1, D_2, \ldots, D_k)T^{-1}$, *where* $D_j \in \mathbb{C}^{n_j,n_j}$ $(j = 1, 2, \ldots, k)$ *and*

$$D_j = \begin{cases} I, & \lambda_j = \lambda,\ i = 0; \\ \mathrm{toep}(\underbrace{0, \ldots, 0}_{i\text{-times}}, 1/i!), & \lambda_j = \lambda,\ 0 < i \leq n_j - 1; \\ O, & \textit{otherwise.} \end{cases}$$

Proof. **(i)** $\sum_{\lambda,i} c_{\lambda,i}C_{\lambda,i} = 0 \Rightarrow O = \sum_{\lambda,i} c_{\lambda,i}\varphi_{\lambda,i}(A) \in \mathcal{P}_{d-1} \Rightarrow c_{\lambda,i} = 0$, otherwise we would have found a minimal polynomial of degree $d-1$, which is a contradiction. **(ii)** results from (1.7) using the Definition 1.9. **(iii)** follows from (ii) setting $f(z) = 1$ or $f(z) = z$, respectively. The components are polynomials in A. Thus, **(iv)** is an implication of Lemma 1.2, (iv). **(v)** Definition 1.3 yields

$$C_{\lambda,i} = \varphi_{\lambda,i}(A) = T\varphi_{\lambda,i}(J)T^{-1} = T\,\mathrm{diag}\left(\varphi_{\lambda,i}(J_1), \varphi_{\lambda,i}(J_2), \ldots, \varphi_{\lambda,i}(J_k)\right)T^{-1},$$

where

$$\varphi_{\lambda,i}(J_j) = \mathrm{toep}\left(\underline{\varphi_{\lambda,i}(\lambda_j)}, \ldots, \frac{\varphi_{\lambda,i}^{(\nu)}(\lambda_j)}{\nu!}, \ldots, \frac{\varphi_{\lambda,i}^{(n_j-1)}(\lambda_j)}{(n_j-1)!}\right).$$

The values $\varphi_{\lambda,i}^{(\nu)}(\lambda_j)$ are given by (1.8): $\varphi_{\lambda,i}^{(\nu)}(\lambda_j) = \delta_{i,\nu}$, which yields $\varphi_{\lambda,i}(J_j) = D_j$ and therefore the assertion. □

Remark 1.11. *Equation* (1.9) *is often referred to as the* spectral resolution of A for f.

Example 1.12. *The components of the matrix A from Example 1.7 are*

$$C_{-1,0} = T\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}T^{-1},\quad C_{-1,1} = T\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}T^{-1},$$

$$C_{0,0} = T\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}T^{-1},\quad C_{1,0} = T\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}T^{-1}.$$

There holds $\exp(A) = \mathrm{e}^{-1}C_{-1,0} + \mathrm{e}^{-1}C_{-1,1} + \mathrm{e}^{0}C_{0,0} + \mathrm{e}^{1}C_{1,0}$.

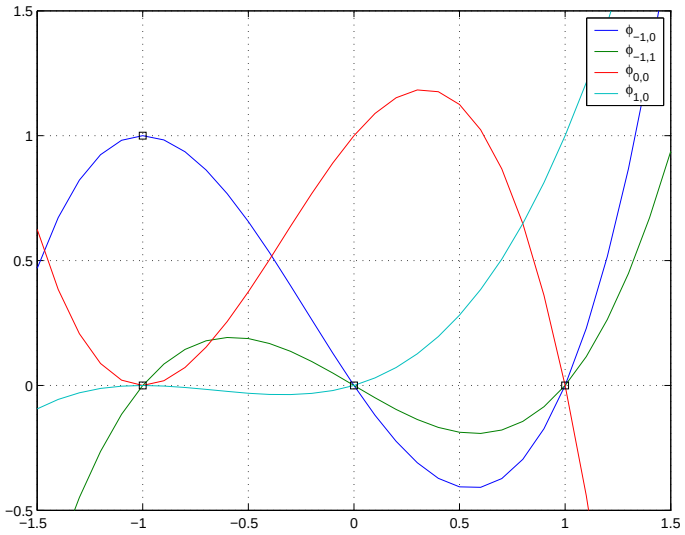


Figure 1.1: The Hermite basis polynomials $\varphi_{\lambda,i}$ of the matrix A from Example 1.7.

1.5 Cauchy Integral Formula

We begin with some fundamental definitions from complex analysis as they are needed for what follows. For further studies the reader may consult Walsh [32] and Henrici [16].

Definitions 1.13. *A* path γ *is a continuous function* $\gamma : [a, b] \to \mathbb{C}$. *A* closed path γ *is a path that satisfies* $\gamma(a) = \gamma(b)$. *A* simple path γ *is a path that satisfies* $\gamma(s) = \gamma(t) \Rightarrow s = a$ *and* $t = b$ *for all* $a \leq s < t \leq b$. *For a closed path* γ *we define the* winding number around $z \in \mathbb{C}$ *as*

$$\mathrm{wind}_z(\gamma) := [\arg(\gamma(b) - z) - \arg(\gamma(a) - z)]/2\pi,$$

where arg *has to be chosen continuous along* γ. *The* interior of a closed path γ *is defined as*

$$\mathrm{int}(\gamma) := \{z \in \mathbb{C} \ : \ \mathrm{wind}_z(\gamma) \neq 0\}.$$

The exterior of a closed path γ *is*

$$\mathrm{ext}(\gamma) := \{z \in \mathbb{C} \ : \ \mathrm{wind}_z(\gamma) = 0\}.$$

An open set Ω *is* connected *if any two points of it can be joined by a path that is contained in* Ω. Ω *is a* domain *if it is nonempty, open and connected.*

The image of a path γ is a curve Γ*, i.e.,*

$$\Gamma = \gamma([a,b]) := \{\gamma(t) \in \mathbb{C} \ : \ t \in [a,b]\}.$$

A curve is closed (simple) *if it is the image of a closed (simple) path. A simple closed curve is called* Jordan curve*. The* winding number of a curve Γ around $z \in \mathbb{C}$ *is defined as the winding number of γ around z, where γ is a path whose image is Γ. The* interior (exterior) of a curve Γ *is defined as the interior (exterior) of a path whose image is Γ. By the* Jordan Curve Theorem *it is known that the exterior of a Jordan curve is an unbounded domain (i.e., nonempty, connected and open) and its interior is a simply connected bounded domain. The latter is often called* Jordan domain.

Theorem 1.14 (Cauchy). *Let $f(z)$ be a function that is analytic within the interior of a Jordan curve Γ and extends continuously to it. Then the* Cauchy integral formula

$$\boxed{f^{(i)}(z) = \frac{i!}{2\pi \mathbb{i}} \int_\Gamma \frac{f(\zeta)}{(\zeta - z)^{i+1}}\, d\zeta} \tag{CIF}$$

holds for $i = 0, 1, \ldots$ and any $z \in \operatorname{int}(\Gamma)$.

Proof. See, for example, Henrici [16, p. 211]. □

Lemma 1.15. *Let $A \in \mathbb{C}^{N\times N}$, $\zeta \notin \Lambda(A)$ and $C_{\lambda,i}$ be the components of A. There holds*

$$R_\zeta(A) := (\zeta I - A)^{-1} = \sum_{\lambda \in \Lambda(A)} \sum_{i=0}^{d_\lambda - 1} \frac{i!}{(\zeta - \lambda)^{i+1}} C_{\lambda,i}. \tag{1.10}$$

$R_\zeta(A)$ is the resolvent of A to ζ.

Proof. For $\zeta \notin \Lambda(A)$, $(\zeta I - A)$ is invertible because $\mathcal{N}(\zeta I - A) = \{\boldsymbol{0}\}$. The spectral resolution (1.9) of A for $f_\zeta(\lambda) = 1/(\zeta - \lambda)$ (defined for all $\lambda \neq \zeta$) yields the desired equivalence in (1.10). □

Theorem 1.16. *Let $A \in \mathbb{C}^{N\times N}$ and Γ be a Jordan curve such that $\Lambda(A) \subset \operatorname{int}(\Gamma)$. Let $f(z)$ be analytic in $\operatorname{int}(\Gamma)$ and continuous on Γ, then*

$$\boxed{f(A) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma f(\zeta)(\zeta I - A)^{-1}\, d\zeta = \frac{1}{2\pi \mathbb{i}} \int_\Gamma f(\zeta) R_\zeta(A)\, d\zeta.} \tag{D3}$$

Proof. By multiplying both sides of (1.10) by $f(\zeta)/(2\pi\mathbb{i})$ and integrating along Γ we obtain

$$\begin{aligned}\int_\Gamma \frac{f(\zeta)}{2\pi\mathbb{i}}(\zeta I - A)^{-1}\,d\zeta &= \int_\Gamma \frac{f(\zeta)}{2\pi\mathbb{i}} \sum_{\lambda\in\Lambda(A)} \sum_{i=0}^{d_\lambda-1} \frac{i!}{(\zeta-\lambda)^{i+1}} C_{\lambda,i}\,d\zeta \\ &= \sum_{\lambda\in\Lambda(A)} \sum_{i=0}^{d_\lambda-1} \left(\frac{i!}{2\pi\mathbb{i}} \int_\Gamma \frac{f(\zeta)}{(\zeta-\lambda)^{i+1}}\,d\zeta\right) C_{\lambda,i} \\ &\overset{\text{(CIF)}}{=} \sum_{\lambda\in\Lambda(A)} \sum_{i=0}^{d_\lambda-1} f^{(i)}(\lambda) C_{\lambda,i} \\ &\overset{(1.9)}{=} f(A).\end{aligned}$$

□

1.6 Polynomial Interpolation II

We have a look at more general interpolation polynomials. Let

$$\omega(z) = (z-\mu_1)^{n_1}(z-\mu_2)^{n_2}\cdots(z-\mu_k)^{n_k}$$

be an arbitrary monic polynomial, μ_j pairwise distinct.

Let $p_{f,\omega}$ be a polynomial such that

$$f^{(i)}(\mu_j) = p_{f,\omega}^{(i)}(\mu_j) \qquad \text{for } j = 1,2,\ldots,k;\ i = 0,1,\ldots,n_j-1. \tag{gHIP}$$

These are $d := \deg(\omega)$ interpolation conditions to $p_{f,\omega}$. We say $p_{f,\omega} \in \mathcal{P}_{d-1}$ *is the Hermite interpolating polynomial to* f *at the nodes* ω. ω is often referred to as the *nodal polynomial* to the *nodes* $\mu_1,\mu_2,\ldots,\mu_k$.

Remark 1.17. *Since we can always construct a matrix* A *such that* ω *is the minimal polynomial of* A, (gHIP) *is nothing but a generalization of* (HIP), *page 11.*

The following theorem provides an analytic representation of $p_{f,\omega}$.

Theorem 1.18 (Hermite Formula). *Let* Γ *be a Jordan curve such that* $\{\mu_1,\mu_2,\ldots,\mu_k\} \subset \operatorname{int}(\Gamma)$ *and let* f *be analytic in* $\operatorname{int}(\Gamma)$ *and extend continuously to* Γ. *There holds*

$$p_{f,\omega}(z) = \frac{1}{2\pi\mathbb{i}} \int_\Gamma \frac{\omega(\zeta)-\omega(z)}{\zeta - z} \frac{f(\zeta)}{\omega(\zeta)}\,d\zeta.$$

Proof. $\omega(\zeta) - \omega(z)$ is a polynomial in z of degree d with a root in ζ. Hence $\zeta - z$ is a divisor of it. For this reason we may write

$$p_{f,\omega}(z) = \frac{1}{2\pi\mathbb{i}} \int_\Gamma \left(\sum_{j=0}^{d-1} \alpha_j(\zeta) z^j \right) \frac{f(\zeta)}{\omega(\zeta)}\, d\zeta = \sum_{j=0}^{d-1} \left(\frac{1}{2\pi\mathbb{i}} \int_\Gamma \alpha_j(\zeta) \frac{f(\zeta)}{\omega(\zeta)}\, d\zeta \right) z^j,$$

which is obviously a polynomial of degree $d-1$.

$p_{f,\omega}(z)$ also fulfills the interpolation conditions (gHIP):

$$p_{f,\omega}^{(i)}(\mu_j) = \frac{i!}{2\pi\mathbb{i}} \int_\Gamma \frac{\omega(\zeta) - \omega(\mu_j)}{(\zeta - \mu_j)^i} \frac{f(\zeta)}{\omega(\zeta)}\, d\zeta = f^{(i)}(\mu_j),$$

where we used (CIF) and the fact that $\omega(\mu_j) = 0$.

By Theorem 1.5 it follows that $p_{f,\omega}$ is the Hermite interpolating polynomial satisfying (gHIP). □

From the last formula we can immediately derive the interpolation error formula due to Hermite:

Lemma 1.19 (Interpolation error). *Let Γ be a Jordan curve such that $\{\mu_1, \mu_2, \ldots, \mu_k\} \subset \operatorname{int}(\Gamma)$ and let f be analytic in $\operatorname{int}(\Gamma)$ and extend continuously to Γ. There holds*

$$f(z) - p_{f,\omega}(z) = \frac{1}{2\pi\mathbb{i}} \int_\Gamma \frac{f(\zeta)}{(\zeta - z)} \frac{\omega(z)}{\omega(\zeta)}\, d\zeta.$$

Proof. Represent $f(z)$ using (CIF). □

1.7 Power Series

We examined polynomials in $A \in \mathbb{C}^{N\times N}$ of degree $m < +\infty$. We may also consider matrix functions f that are defined by power series:

$$f(A) := \sum_{j=0}^{+\infty} \alpha_j A^j = \lim_{m\to+\infty} \sum_{j=0}^{m} \alpha_j A^j. \tag{1.11}$$

We have to take care about the convergence behavior of this expression. Let $|||\cdot|||$ denote an arbitrary matrix norm on $\mathbb{C}^{N\times N}$ that satisfies $\varrho(A) \le |||A|||$. The *Cauchy convergence criteria* is

$$\sum_{j=0}^{+\infty} \alpha_j A^j \text{ converges} \quad \Leftrightarrow \quad \forall \varepsilon > 0\ \exists n_\varepsilon \in \mathbb{N}_0 : \left|\left|\left| \sum_{j=n_\varepsilon}^{+\infty} \alpha_j A^j \right|\right|\right| < \varepsilon.$$

We assume that f is analytic around 0 and has convergence radius τ (i.e., $|f(z)| < +\infty$ for $|z| < \tau$). Then

$$\left|\!\left|\!\left| \sum_{j=n_\varepsilon}^{+\infty} \alpha_j A^j \right|\!\right|\!\right| \leq \sum_{j=n_\varepsilon}^{+\infty} |\alpha_j| \, |||A|||^j ,$$

thus, $\varrho(A) \leq |||A||| < \tau$ would be a sufficient criteria for the convergence of (1.11) because Taylor series converge absolutely. Here $\varrho(A) := \max\{|\lambda| : \lambda \in \Lambda(A)\}$ denotes the *spectral radius of A*.

Theorem 1.20. *Let f be analytic in an open set $U \ni 0$ and let $f(z) = \sum_{j=0}^{+\infty} \alpha_j z^j$ be the Taylor expansion of f in 0 with convergence radius $\tau \in (0, +\infty]$. Then f is defined for every matrix A with $\varrho(A) < \tau$ and*

$$\boxed{f(A) = \sum_{j=0}^{+\infty} \alpha_j A^j = \lim_{m \to +\infty} \sum_{j=0}^{m} \alpha_j A^j.} \tag{D4}$$

Example 1.21. *Let $f(z) = \exp(z)$. f has convergence radius $\tau = +\infty$. Thus, f is defined for every $A \in \mathbb{C}^{N \times N}$ and there holds*

$$f(A) = \exp(A) = \sum_{j=0}^{+\infty} \frac{A^j}{j!}.$$

Remark 1.22. **(i)** *If f is of the form*

$$f(z) = \sum_{j=0}^{+\infty} \alpha_j (z - z_0)^j$$

and for all eigenvalues $\lambda \in \Lambda(A)$ there holds $|f(\lambda)| < +\infty$ then

$$f(A) := \sum_{j=0}^{+\infty} \alpha_j (A - z_0 I)^j$$

is well defined.

(ii) *If $J = \operatorname{toep}(\underline{z_0}, 1) \in \mathbb{C}^{N \times N}$ then by definition of a matrix function*

$$\begin{aligned} f(J) &= \operatorname{toep}\left(\underline{f(z_0)}, \ldots, \frac{f^{(i)}(z_0)}{i!}, \ldots, \frac{f^{(N-1)}(z_0)}{(N-1)!} \right) \\ &= \operatorname{toep}\left(\underline{\alpha_0}, \ldots, \alpha_i, \ldots, \alpha_{N-1} \right), \end{aligned}$$

i.e., we can read off the coefficients of the truncated power series of f in the first row of $f(J)$.

1.8 Properties of Matrix Functions

Firstly, we may extend Lemma 1.2 to general matrix functions:

Lemma 1.23. *Let $A = TJT^{-1} \in \mathbb{C}^{N\times N}$, where $J = \operatorname{diag}(J_1, J_2, \ldots, J_k)$ is block-diagonal and let f be defined for A. There holds*

(i) $f(A) = Tf(J)T^{-1}$,

(ii) $f(J) = \operatorname{diag}\left(f(J_1), f(J_2), \ldots, f(J_k)\right)$,

(iii) *If $A\boldsymbol{v} = \lambda\boldsymbol{v}$ then $f(A)\boldsymbol{v} = f(\lambda)\boldsymbol{v}$,*

(iv) *Given another function $\tilde{f}$ that is defined for A. Then $f(A)\tilde{f}(A) = \tilde{f}(A)f(A)$.*

Proof. Apply Lemma 1.2 to the polynomial representation of $f(A)$ and $\tilde{f}(A)$. □

Given two scalar functions f and g that are defined for A. By $p_{f,A}$ and $p_{g,A}$ we denote the corresponding Hermite interpolating polynomials that satisfy $p_{f,A}(A) = f(A)$ and $p_{g,A}(A) = g(A)$. Clearly, this polynomials fulfill

$$\begin{aligned} p_{\alpha f,A} &= \alpha\, p_{f,A} \qquad (\alpha \in \mathbb{C}), \\ p_{f+g,A} &= p_{f,A} + p_{g,A}, \\ p_{fg,A} &= p_{f,A}\, p_{g,A}, \end{aligned}$$

where the polynomials $p_{\alpha f,A}$, $p_{f+g,A}$ and $p_{fg,A}$ Hermite-interpolate the functions αf, $f+g$ and fg at the roots of ψ_A. These three identities imply that any scalar rational identity will be fulfilled by the corresponding matrix functions.

Example 1.24. *The following equations hold, provided that all the involved terms are defined:*

$$\begin{aligned} \sin^2(A) + \cos^2(A) &= I, \\ \sin(A)\left(\cos(A)\right)^{-1} &= \tan(A), \\ \exp(\mathbb{i}A) &= \cos(A) + \mathbb{i}\sin(A), \\ \log(\alpha A) &= \log(\alpha)I + \log(A), \\ (I - A)^{-1} &= I + A + A^2 + \cdots \qquad (\textit{if } \varrho(A) < 1), \\ A &= \Re(A) + \mathbb{i}\,\Im(A), \\ &\vdots \end{aligned}$$

2 Krylov Subspace Methods

Given a matrix $A \in \mathbb{C}^{N\times N}$ and a vector $\boldsymbol{b} \in \mathbb{C}^N$, our task is the computation of

$$f(A)\boldsymbol{b}$$

for a given matrix function f that is defined for A. This should be accomplished in an elegant and efficient way, with regard to computation speed as well as memory requirements.

In most cases, N is very large and A is sparse. In general, $f(A)$ is not sparse and thus it would not be reasonable to first determine $f(A)$ and then multiply the result by $\boldsymbol{b}$.

Definition 2.1. *The* m-th Krylov subspace of A and $\boldsymbol{b}$ *is defined as*

$$\mathcal{K}_m(A, \boldsymbol{b}) := \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, A^2\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}\} \qquad (m \geq 1),$$

where $\operatorname{span}\{\ldots\}$ *is the set of all linear combinations of the vectors in braces. For ease of notation we abbreviate* $\mathcal{K}_m(A, \boldsymbol{b})$ *by* $\mathcal{K}_m$.

In Chapter 1 we proved that

$$f(A) = p_{f,A}(A),$$

where $p_{f,A}$ is a polynomial of degree $d-1$ that interpolates the function f in the Hermite sense at the roots of ψ_A. Hence

$$f(A)\boldsymbol{b} = p_{f,A}(A)\boldsymbol{b} \in \mathcal{K}_d(A, \boldsymbol{b}).$$

Krylov subspace methods for the approximation of matrix functions are iterative methods that choose their iterates $\boldsymbol{x}_m$ from Krylov spaces $\mathcal{K}_1, \mathcal{K}_2, \ldots$ In other words,

$$\boldsymbol{x}_m = p_m(A)\boldsymbol{b} \in \mathcal{K}_m(A, \boldsymbol{b})$$

for some polynomial p_m of degree $m-1$. For obvious reasons such methods are also known as *polynomial methods*. The element $\boldsymbol{x}_m$ is called *Krylov approximation of order* m. How the polynomial p_m is chosen, depends on the concrete method at hand.

A 'good' Krylov subspace method should return the exact result $\boldsymbol{x}_m = f(A)\boldsymbol{b}$ if the Krylov subspace $\mathcal{K}_m$ contains it. One might claim that such Krylov subspace methods are not iterative methods since they terminate after a finite number of steps. This is true provided that we ignore rounding errors. In practical applications we will start the iteration and run until some *stopping condition* is fulfilled. Note that finding such a stopping condition is not always a trivial task for general matrix functions: a residual or error norm may not be available. This is one of our main motivations to seek convergence estimates of Krylov subspace methods.

In the following sections we will list important properties of Krylov subspaces, introduce the *Arnoldi process* and some Krylov subspace methods, where we concentrate on the *Arnoldi method* and a *generalized interpolation method*.

2.1 Properties of Krylov Subspaces

Lemma 2.2. *By ψ_A we denote the minimal polynomial of A. There exists an index $L = L(A, \boldsymbol{b}) \leq \deg(\psi_A)$ such that*

$$\mathcal{K}_1(A, \boldsymbol{b}) \subsetneqq \mathcal{K}_2(A, \boldsymbol{b}) \subsetneqq \cdots \subsetneqq \mathcal{K}_L(A, \boldsymbol{b}) = \mathcal{K}_{L+1}(A, \boldsymbol{b}) = \cdots$$

$\mathcal{K}_L(A, \boldsymbol{b})$ is the first of the nested Krylov subspaces that is invariant to A.

Proof. It is obvious with Definition 2.1 that the Krylov subspaces are nested subspaces of $\mathbb{C}^N$, i.e.,

$$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots \subseteq \mathcal{K}_m \subseteq \cdots \subseteq \mathbb{C}^N.$$

This chain must become stationary because of the finite dimension of $\mathbb{C}^N$. Thus, there exists a minimal index $L = L(A, \boldsymbol{b})$ with $\mathcal{K}_L = \mathcal{K}_{L+1} = \cdots$

Now assume that $\mathcal{K}_m = \mathcal{K}_{m+1}$ for some index m. This means that $A^m\boldsymbol{b} \in \mathcal{K}_m$, or equivalently

$$A^m\boldsymbol{b} = \alpha_0\boldsymbol{b} + \alpha_1 A\boldsymbol{b} + \cdots + \alpha_{m-1}A^{m-1}\boldsymbol{b}$$

for some coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{m-1} \in \mathbb{C}$. We multiply this equation by A and obtain

$$A^{m+1}\boldsymbol{b} = \alpha_0 A\boldsymbol{b} + \alpha_1 A^2\boldsymbol{b} + \cdots + \alpha_{m-1}A^m\boldsymbol{b}.$$

Therefore $A^{m+1}\boldsymbol{b} \in \mathcal{K}_{m+2}$ is a linear combination of elements from $\mathcal{K}_{m+1}$. Thus, $\mathcal{K}_{m+1} = \mathcal{K}_{m+2}$. Continuing by induction on $m \to m+1$ yields the assertion.

From $\psi_A(A)\boldsymbol{b} = \boldsymbol{0}$ it follows that $L \leq \deg(\psi_A)$. □

Corollary 2.3. *There holds*

$$\dim(\mathcal{K}_m) = \min\{m, L\}.$$

Definition 2.4. *By $\psi_{A,\boldsymbol{b}}(z)$ we denote the monic polynomial of smallest degree for which $\psi_{A,\boldsymbol{b}}(A)\boldsymbol{b} = \boldsymbol{0}$. We say $\psi_{A,\boldsymbol{b}}$ is the* minimal polynomial of $\boldsymbol{b}$ with respect to A.

Lemma 2.5. *$\psi_{A,\boldsymbol{b}}$ is uniquely determined and of the form*

$$\psi_{A,\boldsymbol{b}}(z) = \prod_{\lambda \in \Lambda(A)} (z-\lambda)^{c_\lambda}. \tag{2.1}$$

Proof. First we prove the uniqueness of $\psi_{A,\boldsymbol{b}}$. Assume that $\tilde{\psi}_{A,\boldsymbol{b}}$ is another minimal polynomial of $\boldsymbol{b}$ with respect to A. Then by definition $\deg(\tilde{\psi}_{A,\boldsymbol{b}}) = \deg(\psi_{A,\boldsymbol{b}})$. Consequently, $p := \beta(\tilde{\psi}_{A,\boldsymbol{b}} - \psi_{A,\boldsymbol{b}})$ is a monic polynomial (for some scaling constant $0 \neq \beta \in \mathbb{C}$) that satisfies $p(A)\boldsymbol{b} = \boldsymbol{0}$ and is of lower degree than $\psi_{A,\boldsymbol{b}}$. This is a contradiction.

We turn to (2.1). Assume that $\psi_{A,\boldsymbol{b}}$ contains a factor $(z-\tilde{\lambda})$, where $\tilde{\lambda} \notin \Lambda(A)$. Then $(A - \tilde{\lambda}I)$ is invertible and by definition of $\psi_{A,\boldsymbol{b}}$ we have

$$(A-\tilde{\lambda}I)^{-1}\psi_{A,\boldsymbol{b}}(A)\boldsymbol{b} = \boldsymbol{0}.$$

Hence $(z-\tilde{\lambda})^{-1}\psi_{A,\boldsymbol{b}}(z)$ is a minimal polynomial of $\boldsymbol{b}$ with respect to A and its degree is lower than $\deg(\psi_{A,\boldsymbol{b}})$. This is a contradiction. □

Lemma 2.6. *There holds $L = \deg(\psi_{A,\boldsymbol{b}})$.*

Proof. By Lemma 2.2, $L = L(A, \boldsymbol{b})$ is the smallest integer for which $A^L\boldsymbol{b}$ is linearly dependent on $\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{L-1}\boldsymbol{b}$. Therefore we will find uniquely determined coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{L-1} \in \mathbb{C}$ such that

$$A^L\boldsymbol{b} = \alpha_0\boldsymbol{b} + \alpha_1 A\boldsymbol{b} + \cdots + \alpha_{L-1}A^{L-1}\boldsymbol{b}.$$

Thus,

$$\psi_{A,\boldsymbol{b}}(z) := z^L - \alpha_{L-1}z^{L-1} - \cdots - \alpha_0$$

is the minimal polynomial of $\boldsymbol{b}$ with respect to A and $\deg(\psi_{A,\boldsymbol{b}}) = L$. □

Theorem 2.7. *Let f be defined for A and let*

$$\psi_{A,\boldsymbol{b}}(z) = \prod_{\lambda\in\Lambda(A)} (z-\lambda)^{c_\lambda}$$

be the minimal polynomial of $\boldsymbol{b}$ with respect to A. By $p_{f,A,\boldsymbol{b}} \in \mathcal{P}_{L-1}$ we denote the unique Hermite interpolating polynomial satisfying

$$p^{(i)}_{f,A,\boldsymbol{b}}(\lambda) = f^{(i)}(\lambda) \qquad \text{for all } \lambda \in \Lambda(A),\ i = 0, 1, \ldots, c_\lambda - 1.$$

Then

$$f(A)\boldsymbol{b} = p_{f,A,\boldsymbol{b}}(A)\boldsymbol{b}.$$

Proof. Let $J = T^{-1}AT = \operatorname{diag}(J_1, J_2, \ldots, J_k)$ be a JCF of A. We note that $\psi_{A,\boldsymbol{b}}(A)\boldsymbol{b} = \boldsymbol{0}$ if and only if

$$\psi_{A,\boldsymbol{b}}\left(\begin{bmatrix} \boxed{J_1} & & & \\ & \boxed{J_2} & & \\ & & \ddots & \\ & & & \boxed{J_k} \end{bmatrix}\right)\boldsymbol{a} = \boldsymbol{0}, \text{ where } \boldsymbol{a} := T^{-1}\boldsymbol{b} = \begin{bmatrix} \boxed{\boldsymbol{a}_1} \\ \boxed{\boldsymbol{a}_2} \\ \vdots \\ \boxed{\boldsymbol{a}_k} \end{bmatrix}.$$

The length of each of the $\boldsymbol{a}_j$ corresponds to the size of the Jordan block J_j ($j = 1, 2, \ldots, k$). By reading the above equation block-wise we obtain

$$\psi_{A,\boldsymbol{b}}(J_j)\boldsymbol{a}_j = \boldsymbol{0}.$$

Let $J_j \in \mathbb{C}^{n\times n}$ be a fixed Jordan block associated with the eigenvalue λ. $\psi_{A,\boldsymbol{b}}$ has a root of multiplicity c_λ in λ. With the definition of a matrix function it follows that

$$\psi_{A,\boldsymbol{b}}(J_j) = \operatorname{toep}\left(\underbrace{0, \ldots, 0}_{c_\lambda\text{-times}}, *, \ldots, *\right),$$

where all the $*$ are nonzero entries. This implies

$$\boldsymbol{a}_j = [\underbrace{*, \ldots, *}_{c_\lambda\text{-times}}, 0, \ldots, 0]^T.$$

Hence, for $f(J_j)\boldsymbol{a}_j = p_{f,A,\boldsymbol{b}}(J_j)\boldsymbol{a}_j$ to hold, it is sufficient that in

$$f(J_j) = \operatorname{toep}\left(\underline{f(\lambda)}, \ldots, \frac{f^{(i)}(\lambda)}{i!}, \ldots, \frac{f^{(n-1)}(\lambda)}{(n-1)!}\right)$$

only $f(\lambda), f'(\lambda), \ldots, f^{(c_\lambda - 1)}(\lambda)$ are interpolated by $p_{f,A,\boldsymbol{b}}$. □

Remark 2.8. *From the last proof it is easy to see that for all $\lambda \in \Lambda(A)$ there holds $c_\lambda \leq d_\lambda$, where d_λ is the multiplicity of the root λ in ψ_A. Together with* (2.1) *and* (1.2) *this yields*

$$\psi_{A,\boldsymbol{b}} \mid \psi_A \mid \chi_A\,. \tag{2.2}$$

2.2 Nonderogatory Matrices

For the fast convergence of Krylov subspace methods it would be advantageous if the subspaces $\mathcal{K}_m(A, \boldsymbol{b})$ were to become stationary very early. In other words: we hope that L is small. Unfortunately, this is not always the case. At least we have a complete characterization of the worst-case, namely $L = N$.

It is clear that

$$L = N \iff \deg(\psi_{A,\boldsymbol{b}}) = N \overset{(2.2)}{\Longrightarrow} \psi_A \equiv \chi_A.$$

Definition 2.9. *A matrix $A \in \mathbb{C}^{N\times N}$ for which $\psi_A \equiv \chi_A$ is said to be* nonderogatory. *A vector $\boldsymbol{b} \in \mathbb{C}^N$ for which $L(A, \boldsymbol{b}) = N$ is said to be* cyclic for A.

Remark 2.10. *A matrix A is nonderogatory if and only if its Jordan canonical form contains one and only one Jordan block to each eigenvalue $\lambda \in \Lambda(A)$. This is equivalent to the following assertions:*

- **(i)** *All eigenvectors of A associated with the same eigenvalue are linearly dependent.*
- **(ii)** *The JCF of A is (up to a permutation of the Jordan blocks) uniquely determined by the characteristic (= minimal) polynomial of A.*

Moreover, two nonderogatory matrices are similar if and only if their characteristic polynomials agree.

Lemma 2.11. *Let*

$$\chi_A(z) = z^N + \alpha_{N-1}z^{N-1} + \cdots + \alpha_0$$

be the characteristic polynomial of $A \in \mathbb{C}^{N\times N}$*. Then* A *is nonderogatory if and only if it is similar to the* companion matrix $C_{\boldsymbol{\alpha}}$ *of its characteristic polynomial,*

$$C_{\boldsymbol{\alpha}} := \begin{bmatrix} 0 & & & & -\alpha_0 \\ 1 & 0 & & & -\alpha_1 \\ & 1 & 0 & & -\alpha_2 \\ & & \ddots & \ddots & \vdots \\ & & & \ddots\ 0 & -\alpha_{N-2} \\ & & & 1 & -\alpha_{N-1} \end{bmatrix} \in \mathbb{C}^{N\times N}.$$

Proof. [⇐] First it has to be shown that χ_A is the characteristic polynomial of $C_{\boldsymbol{\alpha}}$. This can be done by expanding $\det(zI - C_{\boldsymbol{\alpha}})$ along the first column and proceeding by induction on the dimension of $C_{\boldsymbol{\alpha}}$. For a detailed proof see Meyer [22, p. 648].

Secondly, we show that $C_{\boldsymbol{\alpha}}$ is nonderogatory: Assumed there is a monic polynomial $\psi(z) = z^d + \beta_{d-1}z^{d-1} + \cdots + \beta_0$ of degree $d < N$ that annihilates $C_{\boldsymbol{\alpha}}$. Then

$$\boldsymbol{0} = \psi(C_{\boldsymbol{\alpha}})\boldsymbol{\xi}_1 = C_{\boldsymbol{\alpha}}^d\boldsymbol{\xi}_1 + \beta_{d-1}C_{\boldsymbol{\alpha}}^{d-1}\boldsymbol{\xi}_1 + \cdots + \beta_0\boldsymbol{\xi}_1 = \boldsymbol{\xi}_{d+1} + \beta_{d-1}\boldsymbol{\xi}_d + \cdots + \beta_0\boldsymbol{\xi}_1, \quad (2.3)$$

i.e., $\boldsymbol{\xi}_{d+1}$ is linearly dependent on $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_d$, which is impossible and therefore a contradiction. Consequently, the minimal and the characteristic polynomial of $C_{\boldsymbol{\alpha}}$ coincide and $C_{\boldsymbol{\alpha}}$ is nonderogatory.

Being nonderogatory is invariant under a similarity transformation. Since A is similar to $C_{\boldsymbol{\alpha}}$, it is nonderogatory.

[⇒] A and $C_{\boldsymbol{\alpha}}$ are nonderogatory matrices with the same characteristic polynomial. Therefore they are similar. □

Lemma 2.12. *There exists a vector* $\boldsymbol{b} \in \mathbb{C}^N$ *that is cyclic for* $A \in \mathbb{C}^{N\times N}$ *if and only if* A *is nonderogatory.*

Proof. [⇐] By Lemma 2.11 there exists an invertible matrix $T \in \mathbb{C}^{N\times N}$ such that $C_{\boldsymbol{\alpha}} = T^{-1}AT$, where $C_{\boldsymbol{\alpha}}$ is the companion matrix to the characteristic polynomial of A. $\boldsymbol{\xi}_1$ is cyclic for $C_{\boldsymbol{\alpha}}$ because of (2.3). Equivalently, $\boldsymbol{b} := T\boldsymbol{\xi}_1$ is cyclic for $T^{-1}A$. There holds $N = \dim(T\mathcal{K}_N(T^{-1}A, \boldsymbol{b})) = \dim(\mathcal{K}_N(A, \boldsymbol{b}))$. Thus, $\boldsymbol{b}$ is cyclic for A.

[⇒] Assume that A is not nonderogatory, i.e., $\deg(\psi_A) < N$. Because of (2.2) this implies $L = \deg(\psi_{A,\boldsymbol{b}}) < N$ and thus $\boldsymbol{b}$ is not cyclic for A. □

2.3 The Arnoldi Process

Let $m \leq L$. We will construct an orthonormal basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\} \subset \mathbb{C}^N$ of the m-th Krylov subspace $\mathcal{K}_m(A, \boldsymbol{b})$ with $\operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_j\} = \mathcal{K}_j(A, \boldsymbol{b})$ for any $j \leq m$. This is done by a Gram-Schmidt procedure, which for Krylov matrices is known as the *Arnoldi process*. $\|\cdot\|$ denotes always the 2-norm of a vector or a matrix.

Definition 2.13. *A matrix* $H_m = [h_{i,j} : 1 \leq i, j \leq m]$ *is said to be an* upper Hessenberg matrix *if* $j + 1 < i \Rightarrow h_{i,j} = 0$, *i.e.,*

$$H_m = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,m-1} & h_{1,m} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,m-1} & h_{2,m} \\ & h_{3,2} & \cdots & h_{3,m-1} & h_{3,m} \\ & & \ddots & \vdots & \vdots \\ & & & h_{m,m-1} & h_{m,m} \end{bmatrix}.$$

If $h_{j+1,j} \neq 0$ $(j = 1, 2, \ldots, m-1)$ *then* H_m *is said to be* unreduced.

Algorithm 2.14: Arnoldi process

Input: $A \in \mathbb{C}^{N\times N}$, $\boldsymbol{b} \in \mathbb{C}^N$, $m \leq L$.

Output: $V := [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$, $\boldsymbol{v}_{m+1}$, $H_m := [h_{i,j} : 1 \leq i, j \leq m]$, $h_{m+1,m}$.

```
v_1 := b/‖b‖
for j = 1, 2, ..., m do
    w := A v_j
    for i = 1, 2, ..., j do                    /* orthogonalize w */
        h_{i,j} := v_i^H w
        w := w − h_{i,j} v_i
    h_{j+1,j} := ‖w‖                           /* h_{j+1,j} = 0 iff j = L */
    if h_{j+1,j} > 0 then
        v_{j+1} := w/h_{j+1,j}
    else
        v_{j+1} := 0
```

Theorem 2.15 (Arnoldi decomposition). *Given $A \in \mathbb{C}^{N\times N}$, $b \in \mathbb{C}^N$ and $m < L$. There exists a matrix $V_m \in \mathbb{C}^{N\times m}$ with orthonormal columns, a vector $\boldsymbol{v}_{m+1} \in \mathbb{C}^N$ satisfying $V_m^H \boldsymbol{v}_{m+1} = \boldsymbol{0}$, an unreduced upper Hessenberg matrix $H_m \in \mathbb{C}^{m\times m}$ and $h_{m+1,m} \geq 0$ such that*

$$AV_m = V_m H_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{\xi}_m^T\,. \tag{2.4}$$

For $m = L$ this reduces to

$$AV_L = V_L H_L\,. \tag{2.5}$$

Proof. Let $j \in \{1, 2, \ldots, m\}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_j$ be an orthonormal basis of $\mathcal{K}_j(A, \boldsymbol{b})$. Lines 3–6 of Algorithm 2.14 read as

$$\boldsymbol{w} := A\boldsymbol{v}_j - \sum_{i=1}^{j} h_{i,j}\boldsymbol{v}_i\,, \tag{2.6}$$

where

$$h_{i,j} := \boldsymbol{v}_i^H \boldsymbol{w},$$

i.e., $\boldsymbol{w}$ is orthogonal to $\mathcal{K}_j$. Because of Corollary 2.3 we have $A\boldsymbol{v}_j \in \mathcal{K}_j$ (and thus $\boldsymbol{w} = \boldsymbol{0}$) only if $j = L$.

By Lines 7–11 we set

$$h_{j+1,j} := \|\boldsymbol{w}\| \tag{2.7}$$

and define $\boldsymbol{v}_{j+1}$, which is orthogonal to $\mathcal{K}_j$ and satisfies

$$h_{j+1,j}\boldsymbol{v}_{j+1} = \boldsymbol{w}. \tag{2.8}$$

Only if $j = L$ we have $h_{j+1,j} = 0$ and $\boldsymbol{v}_{j+1} = \boldsymbol{0}$.

(2.6) and (2.8) yield

$$A\boldsymbol{v}_j = \sum_{i=1}^{j+1} h_{i,j}\boldsymbol{v}_i\,,$$

which for $j = 1, 2, \ldots, m$ can be written in matrix form

$$A[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m] = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m, \boldsymbol{v}_{m+1}]\tilde{H}_{m+1,m} \tag{2.9}$$

with

$$\tilde{H}_{m+1,m} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,m-1} & h_{1,m} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,m-1} & h_{2,m} \\ & h_{3,2} & \cdots & h_{3,m-1} & h_{3,m} \\ & & \ddots & \vdots & \vdots \\ & & & h_{m,m-1} & h_{m,m} \\ & & & & h_{m+1,m} \end{bmatrix} .$$

We define $H_m := [h_{i,j} : 1 \leq i, j \leq m]$ by removing the last row of $\tilde{H}_{m+1,m}$. Then H_m is an unreduced upper Hessenberg matrix and (2.9) can be rewritten as

$$A[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m] = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]H_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{\xi}_m^T .$$

By setting $V := [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$ the assertion is obtained. □

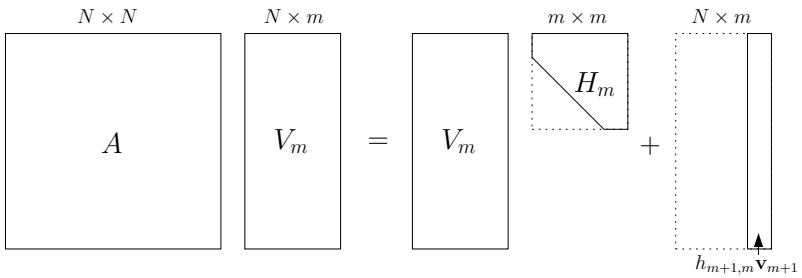


Figure 2.1: Scheme of the Arnoldi decomposition

Corollary 2.16. *For $1 \leq m \leq L$ there holds*

$$H_m = V_m^H A V_m .$$

Remark 2.17. *We say H_m is the* compression of A onto $\mathcal{K}_m(A, \boldsymbol{b})$. *$H_m$ represents the operation of A on $\mathcal{K}_m$ with respect to the* Arnoldi basis $\{\boldsymbol{v}_j : j = 1, 2, \ldots, m\}$. *We shall see that this representation is free of redundance in the sense that H_m is nonderogatory.*

Lemma 2.18. *H_m is a nonderogatory matrix ($1 \leq m \leq L$).*

Proof. Otherwise there would exist two linearly independent eigenvectors $\boldsymbol{x}, \boldsymbol{y}$ to the same eigenvalue λ. Define $\boldsymbol{z} := \alpha\boldsymbol{x} + \beta\boldsymbol{y}$, $\boldsymbol{z} = [z_1, z_2, \ldots, z_m]^T \neq \boldsymbol{0}$ such that $z_m \neq 0$. $H_m\boldsymbol{z} = \lambda\boldsymbol{z}$ and H_m unreduced yield $z_{m-1} = 0$, $z_{m-2} = 0, \ldots$ inductively. Therefore $\boldsymbol{z} = \boldsymbol{0}$, which is a contradiction. □

Lemma 2.19. *There hold*

(i) *H_m has strictly positive entries on its lower subdiagonal, i.e.,*

$$h_{j+1,j} > 0 \qquad \text{for } j = 1, 2, \ldots, m-1.$$

Furthermore, $h_{m+1,m} > 0$ for $m < L$ and $h_{L+1,L} = 0$.

(ii)

$$\boldsymbol{\xi}_m^T H_m^{k-1} \boldsymbol{\xi}_1 = \begin{cases} 0, & k < m; \\ \prod_{j=1}^{m-1} h_{j+1,j}, & k = m. \end{cases}$$

Proof. **(i)** This is an immediate consequence of the definition of $h_{j+1,j}$ given in (2.7). **(ii)** Direct multiplication $H_m H_m^{k-1}$ shows that $\boldsymbol{\xi}_k^T H_m^{k-1} \boldsymbol{\xi}_1 = h_{j+1,j} \boldsymbol{\xi}_{k-1}^T H_m^{k-2} \boldsymbol{\xi}_1$ for $k = 2, 3, \ldots, m$. By recursively applying this equation to itself the assertion is obtained. □

2.4 Arnoldi Approximation to $f(A)\boldsymbol{b}$

Lemma 2.20. *Let $m < L$ be fixed and $p(z) = \alpha_m z^m + \alpha_{m-1} z^{m-1} + \cdots + \alpha_0 \in \mathcal{P}_m$. With the notation from Theorem 2.15 there holds*

$$p(A)\boldsymbol{b} = \|\boldsymbol{b}\| V_m p(H_m) \boldsymbol{\xi}_1 + \|\boldsymbol{b}\| \alpha_m \gamma_m \boldsymbol{v}_{m+1}, \tag{2.10}$$

where $\gamma_m = \prod_{j=1}^{m} h_{j+1,j}$. In particular, for any $p \in \mathcal{P}_{m-1}$ this reduces to

$$p(A)\boldsymbol{b} = \|\boldsymbol{b}\| V_m p(H_m) \boldsymbol{\xi}_1. \tag{2.11}$$

Proof. It is sufficient to prove

$$A^j \boldsymbol{b} = \|\boldsymbol{b}\| V_m H_m^j \boldsymbol{\xi}_1 \qquad \text{for } j < m \tag{2.12}$$

$$\text{and} \qquad A^m \boldsymbol{b} = \|\boldsymbol{b}\| V_m H_m^m \boldsymbol{\xi}_1 + \|\boldsymbol{b}\| \gamma_m \boldsymbol{v}_{m+1}. \tag{2.13}$$

By construction of the Algorithm 2.14 there holds $\boldsymbol{b} = \|\boldsymbol{b}\| V_m \boldsymbol{\xi}_1$. Therefore the assertion (2.12) is true for $j = 0$:

$$A^0 \boldsymbol{b} = \boldsymbol{b} = \|\boldsymbol{b}\| V_m \boldsymbol{\xi}_1 = \|\boldsymbol{b}\| V_m H_m^0 \boldsymbol{\xi}_1.$$

Let (2.12) hold for $j = 0, 1, \ldots, k-1$. Then by induction,

$$\begin{aligned} A^k \boldsymbol{b} &= A(A^{k-1}\boldsymbol{b}) \\ &\overset{(2.12)}{=} A(\|\boldsymbol{b}\| V_m H_m^{k-1} \boldsymbol{\xi}_1) \\ &= \|\boldsymbol{b}\| (AV_m) H_m^{k-1} \boldsymbol{\xi}_1 \\ &\overset{(2.4)}{=} \|\boldsymbol{b}\| (V_m H_m + h_{m+1,m} \boldsymbol{v}_{m+1} \boldsymbol{\xi}_m^T) H_m^{k-1} \boldsymbol{\xi}_1 \\ &= \|\boldsymbol{b}\| V_m H_m^k \boldsymbol{\xi}_1 + \|\boldsymbol{b}\| h_{m+1,m} (\boldsymbol{\xi}_m^T H_m^{k-1} \boldsymbol{\xi}_1) \boldsymbol{v}_{m+1}. \end{aligned}$$

By setting $\gamma_m := h_{m+1,m} \boldsymbol{\xi}_m^T H_m^{k-1} \boldsymbol{\xi}_1$ and applying Lemma 2.19, (ii), we obtain the assertion (2.12) if $k < m$ or (2.13) if $k = m$, respectively. □

Definition 2.21. *Let f be defined for H_m. The* Arnoldi approximation of order m to $f(A)\boldsymbol{b}$ *is defined as*

$$\boxed{\boldsymbol{f}_m := \|\boldsymbol{b}\| V_m f(H_m) \boldsymbol{\xi}_1.}$$

The algorithm to obtain the Arnoldi approximation $\boldsymbol{f}_m$ can be summarized as follows.

Algorithm 2.22: Arnoldi method I

Input: $A \in \mathbb{C}^{N \times N}$, $\boldsymbol{b} \in \mathbb{C}^N$, $m \leq L$.

Output: Arnoldi approximation $\boldsymbol{f}_m$.

1 Determine V_m, H_m using the Arnoldi process 2.14.

2 Set $\boldsymbol{f}_m := \|\boldsymbol{b}\| V_m f(H_m) \boldsymbol{\xi}_1$.

Remark 2.23. **(i)** *We are still left with the problem of evaluating $f(H_m)$ (actually, we need only the first column $f(H_m)\boldsymbol{\xi}_1$). The dimension of this problem has been reduced from N to m.*

(ii) *Even though we assume that f is defined for A, H_m may have eigenvalues in points where f is not defined. In this case, $f(H_m)$ is not defined and therefore the Arnoldi approximation of order m remains undefined too. For the solution of linear systems of equations (i.e., $f(z) = 1/z$), this problem is known as* Galerkin breakdown.

2.5 Ritz Values

Let $1 \leq m \leq L$. The eigenvalues of the matrix H_m are called *Ritz(m) values of* A. With $\chi_m(z)$ we denote the characteristic polynomial of H_m. χ_m is also called *Ritz(m) polynomial of* A.[1] Since H_m is nonderogatory by Lemma 2.18, we have that $\chi_m(z)$ is the minimal polynomial of H_m.

Lemma 2.24. *There holds*

$$\chi_L = \psi_{A,\boldsymbol{b}}.$$

This implies

$$\boldsymbol{f}_L = f(A)\boldsymbol{b}.$$

Proof. We know that χ_L and $\psi_{A,b}$ are monic polynomials of degree L. Set $p := \psi_{A,\boldsymbol{b}} - \chi_L \in \mathcal{P}_{L-1}$ and assume $p \not\equiv 0$. Note that $p(H_L) = \psi_{A,\boldsymbol{b}}(H_L)$. Since H_L is nonderogatory, p has to fulfill L interpolation conditions and therefore $p \equiv \psi_{A,\boldsymbol{b}}$. But this is a contradiction to $\deg(\psi_{A,\boldsymbol{b}}) = L$.

Theorem 2.7 asserts that $f(A)\boldsymbol{b} = p_{f,A,\boldsymbol{b}}(A)\boldsymbol{b}$, where $p_{f,A,\boldsymbol{b}} \in \mathcal{P}_{L-1}$ interpolates f at the roots of $\psi_{A,\boldsymbol{b}} = \chi_L$. Moreover we know from Theorem 1.5 that $f(H_L) \in \mathcal{P}_{L-1}$ is a polynomial $p_{f,H_L}(H_L)$, where p_{f,H_L} interpolates f at the eigenvalues of H_L. Therefore $p_{f,H_L} = p_{f,A,\boldsymbol{b}}$ and the assertion is obtained. □

Lemma 2.25. *There holds*

$$V_m^H \chi_m(A)\boldsymbol{b} = \boldsymbol{0}.$$

Proof. Let $m < L$. Multiply (2.10) by V_m^H from the left setting $p := \chi_m$. For $m = L$ the assertion follows from Lemma 2.24, because $\chi_L(A)\boldsymbol{b} = \psi_{A,\boldsymbol{b}}(A)\boldsymbol{b} = \boldsymbol{0}$. □

Remark 2.26. *Let $m < L$. Then $\chi_m(A)\boldsymbol{b}$ is an element of $\mathcal{K}_{m+1}(A, \boldsymbol{b})$ that satisfies*

$$\chi_m(A)\boldsymbol{b} \perp \mathcal{K}_m\,.$$

In what follows, we denote the set of monic polynomials of degree m by $\mathcal{P}_m^\infty$.

[1] Both H_m and χ_m depend also on the vector $\boldsymbol{b}$. For ease of notation, this will not be mentioned explicitly in the sequel.

Lemma 2.27. *χ_m minimizes the norm $\|p(A)\boldsymbol{b}\|$ among all $p \in \mathcal{P}_m^\infty$.*

Proof. Lemma 2.25 yields $V_m V_m^H \chi_m(A)\boldsymbol{b} = \boldsymbol{0}$. Note that $V_m V_m^H$ is the orthogonal projector onto $\mathcal{K}_m(A, \boldsymbol{b})$. From this it follows

$$(V_m V_m^H \chi_m(A)\boldsymbol{b}, \boldsymbol{x}) = 0 \qquad \text{for all } \boldsymbol{x} \in \mathcal{K}_m,$$

or equivalently,

$$(\chi_m(A)\boldsymbol{b}, V_m V_m^H \boldsymbol{x}) = 0 \qquad \text{for all } \boldsymbol{x} \in \mathcal{K}_m.$$

Writing $\chi_m(z) = z^m - q(z)$, where $q \in \mathcal{P}_{m-1}$, we obtain

$$(A^m\boldsymbol{b} - q(A)\boldsymbol{b}, A^j\boldsymbol{b}) = 0 \qquad \text{for } j = 1, 2, \ldots, m-1.$$

These are the *normal equations* for minimizing the 2-norm of $A^m\boldsymbol{b} - q(A)\boldsymbol{b}$ among all $q \in \mathcal{P}_{m-1}$. □

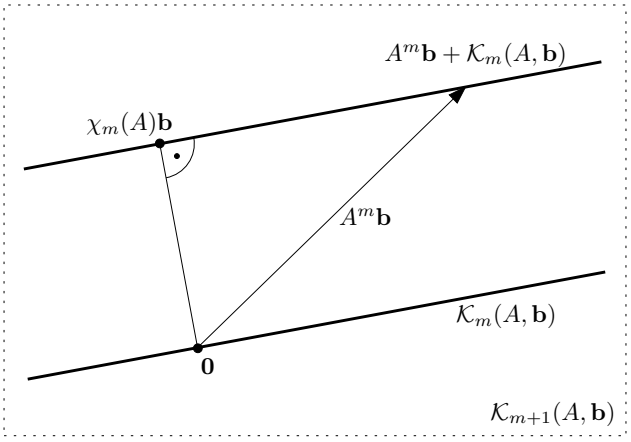


Figure 2.2:
$\mathcal{K}_m(A, \boldsymbol{b})$ is a hyperplane of $\mathcal{K}_{m+1}(A, \boldsymbol{b})$ and $\chi_m(A)\boldsymbol{b}$ can be interpreted as the best approximation to the origin $\boldsymbol{0}$ out of the linear manifold $A^m\boldsymbol{b} + \mathcal{K}_m(A, \boldsymbol{b})$.

With the notion of Ritz values we may reformulate Algorithm 2.22. Instead of evaluating $f(H_m)$, we determine the polynomial $p_{f,m}$ that interpolates f at the Ritz(m) values of A. By Theorem 1.5 there holds

$$f(H_m) = p_{f,m}(H_m).$$

Since $p_{f,m}$ is a polynomial of degree $m-1$, equation (2.11) yields

$$\boldsymbol{f}_m = p_{f,m}(A)\boldsymbol{b}.$$

Algorithm 2.28: Arnoldi method II

Input: $A \in \mathbb{C}^{N\times N}$, $\boldsymbol{b} \in \mathbb{C}^N$, $m \leq L$.

Output: Arnoldi approximation $\boldsymbol{f}_m$.

1 Determine V_m, H_m using the Arnoldi process 2.14.

2 Determine the Ritz(m) values of A, i.e., determine the eigenvalues of H_m.

3 Determine $p_{f,m} \in \mathcal{P}_{m-1}$ that interpolates f (in the Hermite sense) at the Ritz(m) values.

4 Set $\boldsymbol{f}_m := \|\boldsymbol{b}\| V_m p_{f,m}(H_m)\boldsymbol{\xi}_1 \quad (= \|\boldsymbol{b}\| V_m f(H_m)\boldsymbol{\xi}_1 = p_{f,m}(A)\boldsymbol{b})$.

To summarize, it has become apparent that the Arnoldi method is simply an interpolation process, where the nodes for the f-interpolating polynomial $p_{f,m}$ are the Ritz(m) values of A. This interpolation nodes are implicitly chosen by the method *independently* of the function f. Although χ_m fulfills the minimizing property from Lemma 2.27, there is no guarantee that the Ritz(m) values of A are a good choice to achieve a fast decrease of the approximation error $\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\|$.

We recall the formula for the interpolation error from Lemma 1.19 and assume that f is sufficiently smooth. In our context the error formula takes the form

$$f(z) - p_{f,m}(z) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma \frac{f(\zeta)}{(\zeta - z)} \frac{\chi_m(z)}{\chi_m(\zeta)}\, d\zeta,$$

where Γ is a Jordan curve such that all Ritz(m) values are contained in its interior. We obtain

$$f(A)\boldsymbol{b} - p_{f,m}(A)\boldsymbol{b} = \left(\frac{1}{2\pi \mathbb{i}} \int_\Gamma f(\zeta)(\zeta I - A)^{-1} \frac{\chi_m(A)}{\chi_m(\zeta)}\, d\zeta \right) \boldsymbol{b},$$

or equivalently

$$f(A)\boldsymbol{b} - \boldsymbol{f}_m = \left(\frac{1}{2\pi \mathbb{i}} \int_\Gamma f(\zeta)(\zeta I - A)^{-1} \frac{1}{\chi_m(\zeta)}\, d\zeta \right) \chi_m(A)\boldsymbol{b}.$$

Therefore we get for the error of the Arnoldi approximation of order m,

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq \frac{1}{2\pi} \left\| \int_\Gamma f(\zeta)(\zeta I - A)^{-1} \frac{1}{\chi_m(\zeta)}\, d\zeta \right\| \|\chi_m(A)\boldsymbol{b}\|.$$

Lemma 2.27 asserts that χ_m is the minimizer of $\|p(A)\boldsymbol{b}\|$ among all $p \in \mathcal{P}_m^\infty$. Hence $\|\chi_m(A)\boldsymbol{b}\|$ is as small as possible and in this sense, choosing the Ritz(m) values of A as interpolation points seems reasonable. On the other hand, we do not

know how the integral term behaves. Clearly, we would like $|\chi_m|$ to be large on Γ, especially in regions where $|f|$ is large. But this is beyond our influence since χ_m is only determined by A and $\boldsymbol{b}$ and not by f.

Example 2.29. *This example demonstrates that the Ritz values may be 'blind' until the end of the Arnoldi process, i.e., none of the eigenvalues of A is approximated by a Ritz(m) value as long as $m < N$.*

We approximate $f(A)\boldsymbol{b}$ for $f(z) = 1/(\beta - z)$, where $A = C_{\boldsymbol{\alpha}}$ is the nonderogatory companion matrix introduced in Lemma 2.11. We set $\boldsymbol{b} := \boldsymbol{\xi}_1$ and note that $A\boldsymbol{\xi}_j = \boldsymbol{\xi}_{j+1}$ for $j = 1, 2, \ldots, N-1$. Therefore the matrix V_m produced by the Arnoldi process is simply $V_m = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m]$ and by Corollary 2.16 we have

$$H_m = \begin{cases} \operatorname{toep}(1, \underline{0}), & m < N; \\ A, & m = N. \end{cases}$$

It is obvious what happens while $m < N$: all the Ritz(m) values are equal to 0. Let $\beta = 0 \notin \Lambda(A)$. Then f is defined for A but none of the functions $f(H_m)$ is defined. The Arnoldi method will break down.

Let $\beta \neq 0$ and $\beta \notin \Lambda(A)$. For all $m < L$ the functions $f(H_m)$ are defined and for the Arnoldi approximation $\boldsymbol{f}_m$ there holds $\boldsymbol{f}_m = p_{f,m}(A)\boldsymbol{b}$, where $p_{f,m} \in \mathcal{P}_{m-1}$ interpolates $f, f', \ldots, f^{(m-1)}$ at the point 0. On the other hand, the eigenvalues of A are the roots of $\chi_A(z) = z^N + \alpha_{N-1}z^{N-1} + \cdots + \alpha_0$ which may be chosen arbitrarily. Therefore we cannot expect $\boldsymbol{f}_m$ to be a good approximation to $f(A)\boldsymbol{b}$.

2.6 The Lanczos Process

Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian matrix. By Corollary 2.16 we have $H_m = V_m^H A V_m$. Therefore H_m is Hermitian and moreover symmetric since it has only real values on its lower subdiagonal (Lemma 2.19). Thus, we may write

$$H_m = \begin{bmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_m \\ & & \beta_m & \alpha_m \end{bmatrix},$$

where α_j, β_j are real numbers. If we let $V_L = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L]$ we have by (2.5)

$$A\boldsymbol{v}_j = \beta_j \boldsymbol{v}_{j-1} + \alpha_j \boldsymbol{v}_j + \beta_{j+1} \boldsymbol{v}_{j+1}$$

for $j = 2, 3, \ldots, L-1$. This three-term recurrence for $\boldsymbol{v}_{j+1}$ is used by the *Lanczos process* to construct the orthonormal basis vectors of $\mathcal{K}_L$. The Lanczos process is mathematically equivalent to the Arnoldi process (Algorithm 2.14) applied to a Hermitian matrix (see Saad [26, p. 185–187]).

Algorithm 2.30: Lanczos process

Input: $A \in \mathbb{C}^{N \times N}$ Hermitian, $\boldsymbol{b} \in \mathbb{C}^N$, $m \leq L$.

Output: $V := [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$, $\boldsymbol{v}_{m+1}$, H_m, β_{m+1}.

```
1  v_0 := 0
2  β_1 := 0
3  v_1 := b/‖b‖
4  for j = 1, 2, ..., m do
5      w := A v_j
6      w := w − β_j v_{j−1}                    /* orthogonalize w */
7      α_j := v_j^H w
8      w := w − α_j v_j
9      β_{j+1} := ‖w‖                          /* β_{j+1} = 0 iff j = L */
10     if β_{j+1} > 0 then
11         v_{j+1} := w/β_{j+1}
12     else
13         v_{j+1} := 0
```

Now let A have the eigenvalues

$$\lambda_{\min} := \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N =: \lambda_{\max}\,.$$

By $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_m$ we denote the Ritz(m) values of A, i.e., the real eigenvalues of the symmetric matrix H_m. The following theorem is often referred to as the *interlacing property* of Ritz values of Hermitian matrices.

Theorem 2.31. *For $m < L$ there holds*

$$\lambda_{\min} \leq \theta_1 < \theta_2 < \cdots < \theta_m \leq \lambda_{\max},$$

and each of the intervals

$$(-\infty, \theta_1], [\theta_1, \theta_2], \ldots, [\theta_{m-1}, \theta_m], [\theta_m, +\infty)$$

contains at least one eigenvalue of A.

Proof. See Golub, Van Loan [12, Chapter 9]. □

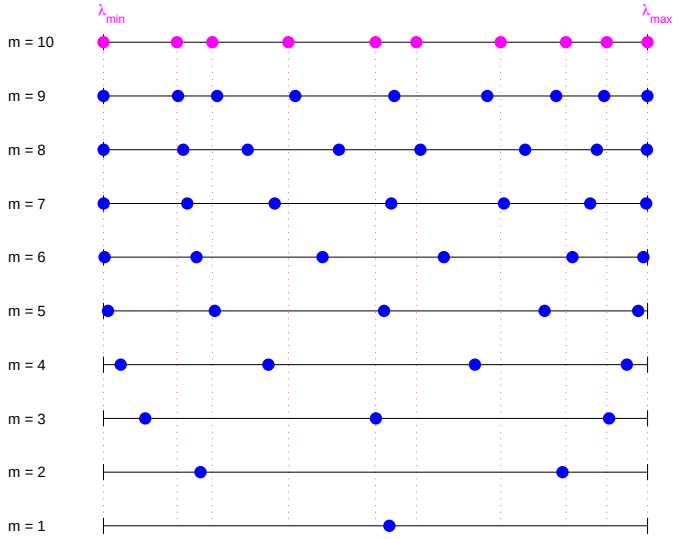


Figure 2.3:
The blue dots show the Ritz(m) values of a Hermitian matrix A of size 10×10 for $m = 1, 2, \ldots, 9$. The magenta dots and the vertical dotted lines indicate the eigenvalues of A that coincide with the Ritz(10) values. Note that two Ritz(m) values are separated by at least one vertical dotted line.


2Lanczos

An immediate consequence of the last theorem is

Corollary 2.32. *In any interval $(-\infty, x]$ ($x \in \mathbb{R}$) the number of Ritz(m) values does not exceed the number of eigenvalues.*

Remark 2.33. *If a function f is defined on the interval $K := [\lambda_{\min}, \lambda_{\max}]$ then $f(H_m)$ is defined for every $m \leq L$ since all the Ritz(m) values are contained in K. Hence it is assured that no (Galerkin) breakdown will occur in the Arnoldi method, Algorithm 2.28.*

2.7 Residual and Error Minimizing Methods

In all the previous algorithms for the approximation of $f(A)\boldsymbol{b}$ we are still faced with one of the tasks

- determine $f(H_m)\boldsymbol{\xi}_1$,
- determine the eigenvalues of H_m and an f-interpolating polynomial.

In case of the function $f(z) := 1/z$ we can avoid this problems, since $f = f^{-1}$ is self-inverse. This can be exploited to construct a 'control-equation' for the Krylov approximations. We consider the problem

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{2.14}$$

where A is an invertible matrix. Hence $f(A)$ is defined and $\boldsymbol{x} = f(A)\boldsymbol{b}$ solves the linear system (2.14).

For an arbitrary vector $\boldsymbol{x}_m$ we define the *residual* $\boldsymbol{r}_m$ by

$$\boldsymbol{r}_m := \boldsymbol{b} - A\boldsymbol{x}_m.$$

We assume that $\boldsymbol{x}_m \in \mathcal{K}_m(A, \boldsymbol{b})$, since we dare to construct a Krylov subspace method. Then $\boldsymbol{x}_m$ has the representation

$$\boldsymbol{x}_m = p_m(A)\boldsymbol{b},$$

where p_m is a polynomial of degree $m-1$. Moreover, the last two equations yield

$$\begin{aligned} \boldsymbol{r}_m &= \boldsymbol{b} - Ap_m(A)\boldsymbol{b} \\ &= (I - Ap_m(A))\boldsymbol{b}. \end{aligned}$$

By defining $\tilde{p}_m(z) := 1 - zp_m(z)$, we may write

$$\boldsymbol{r}_m = \tilde{p}_m(A)\boldsymbol{b}. \tag{2.15}$$

Obviously, $\tilde{p}_m$ is a polynomial of degree m that satisfies $\tilde{p}_m(0) = 1$. We say that $\tilde{p}_m$ is a *residual polynomial of degree* m and denote this by $\tilde{p}_m \in \mathcal{P}_m^0$. There is an immediate connection between the polynomials p_m and $\tilde{p}_m$. Because of

$$p_m(z) = \frac{1 - \tilde{p}_m(z)}{z}$$

it is easy to see that p_m interpolates the function $1/z$ in the Hermite sense at the roots of the associated residual polynomial $\tilde{p}_m$, and this holds for every Krylov subspace method.

A *residual minimizing method* is characterized by a minimizing property for the residual, i.e., $\tilde{p}_m$ is chosen such that $|||\boldsymbol{r}_m|||$ is minimized, where $|||\cdot|||$ denotes a fixed vector norm of $\mathbb{C}^N$. In other words,

$$|||\boldsymbol{r}_m||| = |||\tilde{p}_m(A)\boldsymbol{b}||| = \min_{p\in\mathcal{P}_m^0} |||p(A)\boldsymbol{b}||| .$$

Note that $|||\boldsymbol{r}_m||| = 0$ if and only if the associated Krylov approximation $\boldsymbol{x}_m$ solves (2.14). Since $A^{-1}\boldsymbol{b} \in \mathcal{K}_L$, the above minimizing property assures that $\boldsymbol{x}_L$ solves (2.14) in exact arithmetic. Practical implementations of minimal residual methods make use of the Arnoldi decomposition (2.4): since $\boldsymbol{x}_m \in \mathcal{K}_m$ can be represented in the form $V_m\boldsymbol{y}_m$ for some $\boldsymbol{y}_m \in \mathbb{C}^m$, we have

$$\boldsymbol{r}_m = \boldsymbol{b} - AV_m\boldsymbol{y}_m = \boldsymbol{b} - V_{m+1}\tilde{H}_{m+1,m}\boldsymbol{y}_m .$$

Now let $\|\cdot\|$ denote the 2-norm. The resulting method is then called *generalized minimal residual method* (GMRES). Note that $\boldsymbol{b} = \|\boldsymbol{b}\|V_{m+1}\boldsymbol{\xi}_1$ and therefore

$$\|\boldsymbol{r}_m\| = \left\|\|\boldsymbol{b}\|\boldsymbol{\xi}_1 - \tilde{H}_{m+1,m}\boldsymbol{y}_m\right\| = \min_{\boldsymbol{y}\in\mathbb{C}^m} \left\|\|\boldsymbol{b}\|\boldsymbol{\xi}_1 - \tilde{H}_{m+1,m}\boldsymbol{y}\right\| .$$

This is a least squares problem. Once we obtained the minimizer $\boldsymbol{y}_m$, we set $\boldsymbol{x}_m := V_m\boldsymbol{y}_m$.

If A is Hermitian, we can make further simplifications by using the Lanczos process (Algorithm 2.30) instead of the Arnoldi process. The resulting method is called *minimal residual method* (MINRES).

Now we turn to *error minimizing methods*. For an arbitrary vector $\boldsymbol{x}_m$ we define the *error* $\boldsymbol{e}_m$ by

$$\boldsymbol{e}_m := A^{-1}\boldsymbol{b} - \boldsymbol{x}_m.$$

Note that $A\boldsymbol{e}_m = \boldsymbol{r}_m$. By assuming $\boldsymbol{x}_m \in \mathcal{K}_m(A, \boldsymbol{b})$, equation (2.15) yields

$$\boldsymbol{e}_m = \tilde{p}_m(A)A^{-1}\boldsymbol{b}$$

for some $\tilde{p}_m \in \mathcal{P}_m^0$. An *error minimizing method* is characterized by a minimizing property for the error, i.e., $\tilde{p}_m$ is chosen such that $|||\boldsymbol{e}_m|||$ is minimized. In other

words,

$$|||\, \boldsymbol{e}_m \,||| = \left|||\, \tilde{p}_m(A)A^{-1}\boldsymbol{b} \,\right||| = \min_{p\in\mathcal{P}_m^0} \left|||\, p(A)A^{-1}\boldsymbol{b} \,\right|||\,.$$

For the same reasoning as above, it is assured that $\boldsymbol{x}_L$ solves (2.14).

Let A be a symmetric positive definite matrix. We may minimize the A-norm $\|\boldsymbol{e}_m\|_A := (\boldsymbol{e}_m^H A \boldsymbol{e}_m)^{1/2}$ of the error, resulting in the widely used *CG method* (see Stiefel [17]). The iterates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ of the CG method are characterized by

$$\|A^{-1}\boldsymbol{b} - \boldsymbol{x}_m\|_A = \|\boldsymbol{e}_m\|_A = \min_{p\in\mathcal{P}_m^0} \|p(A)A^{-1}\boldsymbol{b}\|_A.$$

For implementations of the mentioned algorithms we refer to the books of Greenbaum [13] and Saad [26].

2.8 A Generalized Interpolation Method

Given are a *nodal matrix*

$$M = \begin{bmatrix} \mu_{1,1} & & & \\ \mu_{2,1} & \mu_{2,2} & & \\ \mu_{3,1} & \mu_{3,2} & \mu_{3,3} & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

and the corresponding *nodal polynomials*

$$\omega_m(z) = \prod_{j=1}^{m} (z - \mu_{m,j}) \qquad (m = 1, 2, \ldots).$$

We assume that f (and $f', f'', \ldots$ if necessary) is defined on all nodes.

Algorithm 2.34: Generalized interpolation method

Input: $A \in \mathbb{C}^{N\times N}$, $\boldsymbol{b} \in \mathbb{C}^N$, nodes $\mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,m}$.

Output: Krylov approximation $\boldsymbol{g}_m$.

1 Determine $q_{f,m} \in \mathcal{P}_{m-1}$ that Hermite-interpolates f at $\mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,m}$.

2 Set $\boldsymbol{g}_m := q_{f,m}(A)\boldsymbol{b}$.

Remarks 2.35. **(i)** *If the nodes $\mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,m}$ are the Ritz(m) values of A, this algorithm coincides with the Arnoldi method (Algorithm 2.28).*

(ii) *We have to clarify for which nodes we can expect that the $\boldsymbol{g}_m$ approximate $f(A)\boldsymbol{b}$ well and how fast the approximation error decreases with m.*

(iii) *This method has the advantage that, once $q_{f,m}$ is determined, we can evaluate it for different A and $\boldsymbol{b}$. Then of course, the quality of the Krylov approximations $\boldsymbol{g}_m$ will vary.*

(iv) *This algorithm allows to choose the nodes explicitly, i.e., we may adapt the interpolation nodes to the function f.*

3 Polynomial Interpolation and Best Approximation

The class of Krylov subspace methods is much too large to discuss each known algorithm for the approximation of $f(A)\boldsymbol{b}$. Such algorithms will differ for special f and A. For example, if A is Hermitian and we use Arnoldi approximations

$$\boldsymbol{f}_m := \|\boldsymbol{b}\| V_m f(H_m)\boldsymbol{\xi}_1,$$

we will find that H_m is a symmetric tridiagonal matrix. This can be exploited to improve calculation speed and memory storage need as well.

It seems reasonable to examine Krylov subspace methods as what they are: *polynomial interpolation methods*. From now on we will assume that $A \in \mathbb{C}^{N\times N}$ is a *normal* matrix, i.e., $A^H A = AA^H$. Normal matrices can be written in the form $A = UDU^H$, where D is a diagonal matrix with the eigenvalues of A as diagonal entries and U is an unitary matrix. We say: normal matrices are *unitarily diagonalizable*. For a matrix function f that is defined for a normal matrix A, the properties of the 2-norm $\|\cdot\|$ yield

$$\|f(A)\| = \|f(D)\| = \max\{|f(\lambda)| : \lambda \in \Lambda(A)\}.$$

More generally, if A is only required to be *diagonalizable*, i.e., $A = XDX^{-1}$ with an invertible matrix X, then

$$\|f(A)\| \leq \|X\|\|X^{-1}\| \max\{|f(\lambda)| : \lambda \in \Lambda(A)\}.$$

Many of the inequalities developed here hold for normal matrices but can be easily extended to diagonalizable matrices by involving the term $\|X\|\|X^{-1}\|$.

3.1 Some Approximation Theory

Let Ω be a compact (i.e., closed and bounded) subset of $\mathbb{C}$. By $C(\Omega)$ we denote the set of continuous functions $f : \Omega \to \mathbb{C}$. For all $f \in C(\Omega)$ there holds $|f| \in C(\Omega)$ and, by *Weierstrass' Theorem*, $|f|$ attains a maximum on Ω. We define

$$\|f\|_\Omega := \max_{z\in\Omega} |f(z)|.$$

$\|\cdot\|_\Omega$ is called the *uniform norm on* Ω. A sequence $(f_m)_{m\geq 1} \subset C(\Omega)$ *converges uniformly to* f if $\|f - f_m\|_\Omega \to 0$ for $m \to +\infty$. We denote this by $f_m \Rightarrow f$.

Definition 3.1. *Given a linear space $\mathcal{V}$ with norm $|||\cdot|||$ and $v \in \mathcal{V}$. Let U be an arbitrary subset of $\mathcal{V}$. We say $u^* \in U$ is an* element of best approximation from U to v *if*

$$|||v - u^*||| = \min_{u\in U} |||v - u|||.$$

Theorem 3.2 (Existence of best approximations). *Let $\mathcal{U}$ be a finite-dimensional normed linear subspace of $\mathcal{V}$. Then for every $v \in \mathcal{V}$ there exists an element $u^* \in \mathcal{U}$ of best approximation to v.*

Proof. We define $U_0 := \{u \in \mathcal{U} : |||v - u||| \leq |||v|||\}$, which is a closed and bounded subset of a finite-dimensional space, thus compact. Set $d := \inf_{u\in U_0} |||v - u|||$ and let $(u_i)_{m\geq 1} \subset U_0$ be a minimizing sequence, i.e., $|||v - u_i||| \to d$ as $i \to +\infty$. By the compactness of U_0, this sequence has at least one accumulation point $u^* \in U_0$ and we can assume that $|||u_i - u^*||| \to 0$. Hence,

$$|||v - u^*||| \leq |||v - u_i||| + |||u_i - u^*||| \to d = \inf_{u\in U_0} |||v - u|||.$$

Because of

$$\inf_{u\in U_0} |||v - u||| = \inf_{u\in\mathcal{U}} |||v - u|||$$

and $u^* \in \mathcal{U}$, the minimum is attained and u^* is the element of best approximation to v. □

For our purposes we will identify

$$\begin{aligned} |||\cdot||| &= \|\cdot\|_\Omega, \\ \mathcal{V} &= C(\Omega), \\ v &= f(\zeta), \\ \mathcal{U} &= \mathcal{P}_{m-1} \end{aligned}$$

and conclude that the problem

$$\text{for } f \in C(\Omega) \text{ find } p^* \in \mathcal{P}_{m-1} \text{ such that } \|f - p^*\|_\Omega = \min_{p \in \mathcal{P}_{m-1}} \|f - p\|_\Omega$$

has a solution. The following theorem provides us the uniqueness of p^*, which we refer to as the *polynomial uniform best approximation of degree $m-1$ to f on Ω*.

Theorem 3.3 (Tonelli). *Let $\Omega \subset \mathbb{C}$ be compact and contain more than m points. For a given $f \in C(\Omega)$ we set*

$$M^* := \min_{p \in \mathcal{P}_{m-1}} \|f - p\|_\Omega$$

and let p^* be such a minimizing polynomial. Then

(i) *there are at least $m+1$ distinct points $z \in \Omega$ at which $|f(z) - p^*(z)| = M^*$,*

(ii) *p^* is unique.*

Proof. See Davis [2, pp. 143–145]. □

If Ω is a real compact set and f is real-valued, then **(i)** from the above theorem is also a sufficient condition for that p^* is the best approximation to f, provided $f - p^*$ takes on its extreme value with alternating sign on Ω.

Theorem 3.4 (Oscillating property). *Let $\Omega \subset \mathbb{R}$ be compact and $f \in C(\Omega)$ real-valued. For $p \in \mathcal{P}_{m-1}$ we set*

$$M := \|f - p\|_\Omega.$$

Then p is the polynomial uniform best approximation to f on Ω if and only if there are $m+1$ distinct points $x_1 < x_2 < \ldots < x_{m+1}$, $x_i \in \Omega$, such that

$$f(x_i) - p(x_i) = \pm M \qquad \text{for} \quad i = 1, 2, \ldots, m+1$$

with alternating sign (i.e., $f(x_i) - p(x_i) = p(x_{i+1}) - f(x_{i+1})$).

Proof. Let $p^* \in \mathcal{P}_{m-1}$ be the uniform best approximation to f on Ω. By definition we have $\|f - p^*\|_\Omega = M^* \leq M$. Now assume $M^* < M$. We set $m_i := f(x_i) - p^*(x_i)$ and note that $|m_i| < M$ for $i = 1, 2, \ldots, m+1$. There holds

$$p^*(x_i) - p(x_i) = (f(x_i) - m_i) - (f(x_i) - \pm M) = \pm M - m_i.$$

The polynomial $p^* - p \in \mathcal{P}_{m-1}$ has $m+1$ points of alternating sign in $x_1, x_2, \ldots, x_{m+1}$. Hence it has m roots and therefore $p^* - p \equiv 0$. □

3.2 Chebyshev Polynomials

We will give a brief survey to the classical Chebyshev polynomials. The interested reader will find more details and proofs on this topic in Davis [2, pp. 60–64].

Chebyshev Polynomials on $[-1, 1]$

Definition 3.5. *The* Chebyshev polynomial of degree m *is defined as*

$$T_m(x) := \cos(m \arccos x) \qquad (x \in [-1, 1];\ m = 0, 1, \ldots).$$

We have to prove that T_m is indeed a polynomial, but first we will give the following recurrence relation.

Lemma 3.6. *There holds*

$$T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x) \qquad \text{for } m = 1, 2, \ldots \tag{3.1}$$

Proof. By adding the equations

$$\begin{aligned} \cos(m+1)\theta &= \cos m\theta \cos\theta - \sin m\theta \sin\theta \\ \cos(m-1)\theta &= \cos m\theta \cos\theta + \sin m\theta \sin\theta \end{aligned}$$

we get

$$\cos(m+1)\theta = 2\cos m\theta \cos\theta - \cos(m-1)\theta.$$

By setting $\cos\theta = x$ and $\cos m\theta = T_m(x)$, the assertion is obtained. □

Since $T_0(x) = 1$ and $T_1(x) = x$, the recurrence relation (3.1) yields

Corollary 3.7. T_m *is a polynomial of degree* m *and of the form*

$$T_m(x) = 2^{m-1}x^m + \textit{terms of lower degree.}$$

The following well known result is easily verified and thus given without proof.

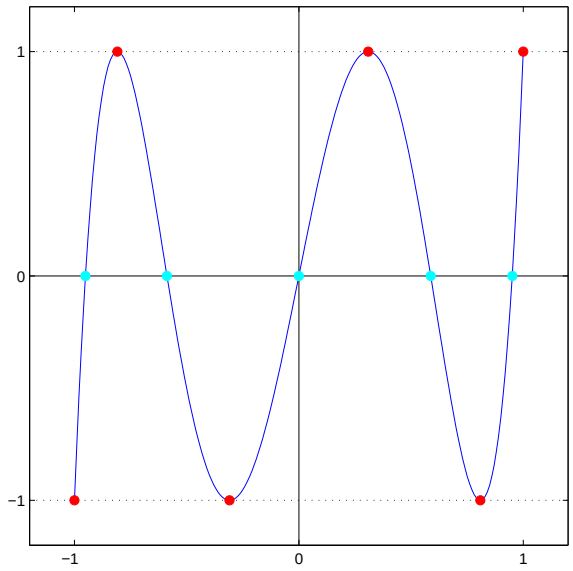
Lemma 3.8. T_m *has* m *simple roots* x_k *in* $(-1, 1)$*, where*

$$x_k = \cos\frac{2k-1}{2m}\pi \qquad (k = 1, 2, \ldots, m).$$

There holds $|T_m(x)| \le 1$ *for all* $x \in [-1, 1]$*. Equality holds for the* $m + 1$ *points*

$$x'_k = \cos\frac{2k}{2m}\pi \qquad (k = 0, 1, \ldots, m),$$

where the value $T_m(x'_k) = \pm 1$ *is taken with alternating sign.*




3T-5

Figure 3.1: The graph of $T_5(x)$ on $[-1, 1]$ with its roots and extreme points.

Definition 3.9. *The* normalized Chebyshev polynomial of degree m *is defined as*

$$\widetilde{T}_m(x) := \frac{1}{2^{m-1}} T_m(x) \qquad (m = 0, 1, \ldots).$$

Since $\widetilde{T}_m$ is a monic polynomial of degree m, we can write $\widetilde{T}_m(x) = x^m - p^*(x)$ for some $p^* \in \mathcal{P}_{m-1}$. Furthermore, $|\widetilde{T}_m|$ takes on its extreme value $M := \frac{1}{2^{m-1}}$ at $m+1$ distinct points in $\Omega := [-1, 1]$ with alternating sign. Theorem 3.4 yields that p^* is the unique best approximating polynomial to $f(x) = x^m$ on Ω. We state this result as a theorem.

Theorem 3.10 (Chebyshev). *There holds*

$$\|\widetilde{T}_m\|_\Omega = \min_{p \in \mathcal{P}_m^\infty} \| p \|_\Omega = \frac{1}{2^{m-1}},$$

where $\mathcal{P}_m^\infty$ *denotes the set of all monic polynomials of degree* m *and* $\Omega = [-1, 1]$.

Chebyshev Polynomials in $\mathbb{C}$

Now we let $z \in \mathbb{C}$ be fixed and set $T_0(z) = 1, T_1(z) = z$. The recurrence relation (3.1) is well defined for complex arguments, i.e.,

$$T_{m+1}(z) = 2zT_m(z) - T_{m-1}(z) \qquad \text{for } m = 1, 2, \ldots \tag{3.2}$$

In order to study the behavior of the m-th Chebyshev polynomial $T_m = T_m(z)$ in the complex plane, we observe that (3.2) is a *difference equation of order* 2 with characteristic polynomial

$$\chi(\zeta) = \zeta^{m+1} - 2z\zeta^m + \zeta^{m-1},$$

which has a root of multiplicity $m-1$ at 0 and the two non-trivial roots

$$w := z + \sqrt{z^2-1} \quad \text{and} \quad z - \sqrt{z^2-1} = w^{-1}.$$

Note that $w + w^{-1} = 2z$. From the theory of difference equations it is known that $T_m(z)$ is a linear combination of w^m and w^{-m} that satisfies the initial conditions $T_0(z) = 1$ and $T_1(z) = z$. Therefore we have

$$T_m(z) = \frac{1}{2}\left(w^m + w^{-m}\right), \qquad \text{where } z := \frac{1}{2}\left(w + w^{-1}\right). \tag{3.3}$$

The mapping

$$\Psi : w \mapsto \frac{1}{2}\left(w + w^{-1}\right) = z$$

is the well known *Joukowski transformation*. It is the conformal bijection from $\overline{\mathbb{C}}\setminus\overline{\mathbb{D}}$ onto $\overline{\mathbb{C}}\setminus[-1,1]$ with $\Psi(\infty) = \infty$ and $\Psi'(\infty) = \frac{1}{2}$. For every $R > 1$, Ψ maps the circle $\{w : |w| = R\} =: \mathbb{T}_R$ to an ellipse E_R with semiaxes $\frac{1}{2}(R + R^{-1})$ and $\frac{1}{2}(R - R^{-1})$. From (3.3) we obtain

$$|T_m(z)| = \frac{1}{2}|w^m|\left|1 + w^{-2m}\right| \qquad \text{for } z \in E_R,$$

and since $|w^m| = R^m$ for $w \in \mathbb{T}_R$ we have

$$|T_m(z)| = \frac{R^m}{2}\left|1 + w^{-2m}\right| \qquad \text{for } z \in E_R. \tag{3.4}$$

Remark 3.11. *For $\widetilde{T}_m$ the equation* (3.4) *is of the form*

$$\left|\widetilde{T}_m(z)\right| = \frac{R^m}{2^m}\left|1 + w^{-2m}\right| \qquad \textit{for } z \in E_R. \tag{3.5}$$

Shifted Chebyshev Polynomials

Finally, we transform the Chebyshev polynomials to an arbitrary positive real interval

$$K := [\lambda_{\min}, \lambda_{\max}] \qquad (0 < \lambda_{\min} < \lambda_{\max})$$

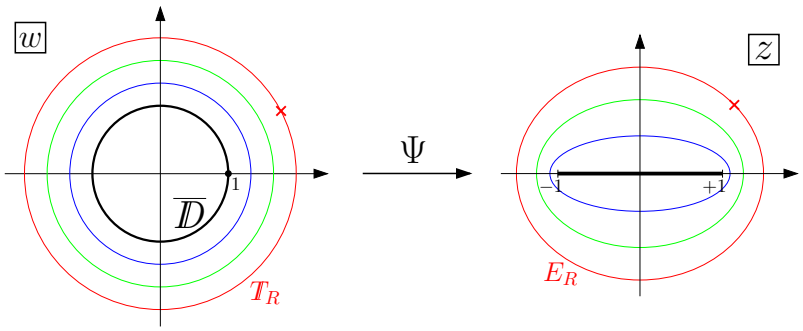


Figure 3.2: The Joukowski transformation Ψ.

and normalize them at the origin. The linear mapping

$$z \mapsto \frac{2z - \lambda_{\max} - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}$$

furnishes the transformation from K onto $[-1, 1]$. We define the *(shifted) Chebyshev polynomial of degree m on the interval K* as

$$T_m^K(z) := \frac{T_m\left(\frac{2z-\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)}{T_m\left(\frac{-\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)}.$$

Clearly, this polynomial satisfies $T_m^K(0) = 1$. Later on we shall make use of the following assertion.

Lemma 3.12. *There holds*

$$\max_{z\in K} |T_m^K(z)| = \min_{p\in\mathcal{P}_m^0} \max_{z\in K} |p(z)| = 2\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{-m}\right)^{-1},$$

where $\kappa := \lambda_{\max}/\lambda_{\min}$.

Proof. T_m^K is a polynomial of degree m, say, $T_m^K(z) = \alpha_m z^m + \alpha^{m-1} z^{m-1} + \cdots + \alpha_1 z + 1$. It takes on its extreme value

$$M := \frac{1}{\left|T_m\left(\frac{-\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)\right|}$$

in $m+1$ points of K with alternating sign, because it is just a shifted version of T_m that takes on the alternating extreme values ± 1 in $[-1, 1]$ (cf. Lemma 3.8). By Theorem 3.4, we have that $(\alpha^{m-1} z^{m-1} + \cdots + \alpha_1 z + 1) \in \mathcal{P}_{m-1}^0$ is the uniform best

approximation to $-\alpha_m z^m$ on K. Thus, $\max_{z\in K}|T_m^K(z)| = M$ is minimal. At last we determine M. We fix $z := \left(\frac{-\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)$ and note that

$$w := -\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \quad \text{satisfies} \quad z = \frac{1}{2}(w + w^{-1}),$$

where $\kappa = \lambda_{\max}/\lambda_{\min} > 1$, hence $w < 0$. From Lemma (3.3) we obtain

$$M = \frac{1}{|T_m(z)|} = \frac{1}{\left|\frac{1}{2}(w^m + w^{-m})\right|} = \frac{2}{(-w)^m + (-w)^{-m}},$$

which is the assertion. □

3.3 A Generalized Approximation Method

To complete our survey, we mention that the polynomials $(q_{f,m})_{m\geq 1}$ we used to construct Krylov subspace methods for the approximation of $f(A)\boldsymbol{b}$ may not only arise from interpolation processes. One may think of methods that determine a polynomial $q_{f,m}^* \in \mathcal{P}_{m-1}$ of best approximation to f on a compact set Ω and define the Krylov approximations by

$$\boldsymbol{a}_m := q_{f,m}^*(A)\boldsymbol{b} \in \mathcal{K}_m.$$

Indeed, some *semi-iterative methods* may be put into this framework, e.g. the *Chebyshev method*. A prototype of such approximation methods looks like this:

Algorithm 3.13: Generalized approximation method

Input: $A \in \mathbb{C}^{N\times N}$, $\boldsymbol{b} \in \mathbb{C}^N$, $m \geq 1$, Ω compact, $f \in C(\Omega)$.

Output: Krylov approximation $\boldsymbol{a}_m$.

1 Determine a polynomial best approximation $q_{f,m}^* \in \mathcal{P}_{m-1}$ to f on Ω.

2 Set $\boldsymbol{a}_m := q_{f,m}^*(A)\boldsymbol{b}$.

Remark 3.14. *If Ω consists of m points at most, this algorithm reduces to an interpolation problem. If Ω consists of exactly m points, $q_{f,m}^*$ will be the unique Lagrange interpolation polynomial of degree $m-1$. If Ω includes more than m points, then $q_{f,m}^*$ is still unique due to Theorem 3.3.*

3.4 Error of Polynomial Methods

Lemma 3.15. *Given $q \in \mathcal{P}_{m-1}$, $A \in \mathbb{C}^{N\times N}$ normal and $\boldsymbol{b} \in \mathbb{C}^N$. For every function f that is defined for A there holds*

$$\|f(A)\boldsymbol{b} - q(A)\boldsymbol{b}\| \leq \|\boldsymbol{b}\| \max_{\lambda\in\Lambda(A)} |f(\lambda) - q(\lambda)|.$$

Proof. We write $A = UDU^H$, where U is unitary and D is a diagonal matrix. From the properties of matrix functions and the 2-norm it follows

$$\begin{aligned} \|f(A)\boldsymbol{b} - q(A)\boldsymbol{b}\| &\leq \|f(A) - q(A)\|\|\boldsymbol{b}\| \\ &= \|\boldsymbol{b}\|\|U(f(D) - q(D))U^H\| \\ &= \|\boldsymbol{b}\| \max_{\lambda\in\Lambda(A)} |f(\lambda) - q(\lambda)|. \end{aligned}$$

□

Remark 3.16. *This result is as important as it is simple. It asserts that, in order to obtain results about the error of a polynomial method, we may study*

$$\max_{\lambda\in\Lambda(A)} |f(\lambda) - q(\lambda)|$$

for some polynomial $q \in \mathcal{P}_{m-1}$. More generally, we may consider

$$\max_{\lambda\in\Omega} |f(\lambda) - q(\lambda)| = \|f - q\|_\Omega,$$

where Ω is a compact set containing $\Lambda(A)$. For non-normal matrices it is not valid that the error of polynomial methods is primarily determined by the spectrum. This is why the investigations of the convergence of Krylov subspace methods for arbitrary matrices is much more complicated.

One may ask what is the 'best possible' Krylov approximation $\boldsymbol{g}_m^* \in \mathcal{K}_m(A, \boldsymbol{b})$ to $f(A)\boldsymbol{b}$ that we can expect from a polynomial method in general. Again we decompose $A = UDU^H$, $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N]$ and note that

$$\|f(A)\boldsymbol{b} - q(A)\boldsymbol{b}\|^2 = \|(f(D) - q(D))U^H\boldsymbol{b}\|^2 = \sum_{i=1}^{N} |\boldsymbol{u}_i^H\boldsymbol{b}(f(\lambda_i) - q(\lambda_i))|^2. \quad (3.6)$$

Thus, the *best approximation* $\boldsymbol{g}_m^*$ *to* $f(A)\boldsymbol{b}$ *out of* $\mathcal{K}_m$ *with respect to the 2-norm* can be obtained by minimizing (3.6) among all polynomials of degree $m-1$. This is a *weighted least squares problem*. Once we obtained the minimizing polynomial q^* we set $\boldsymbol{g}_m^* := q^*(A)\boldsymbol{b}$.

The Generalized Interpolation Method

Let us now consider f-interpolating polynomials $q_{f,m}$ that arise from the generalized interpolation method, Algorithm 2.34. With the formula for the Hermite interpolation error (Lemma 1.19) we can present some results (cf. Gaier, [10, pp. 59–61]).


3Taylor

Example 3.17. *Given a nodal polynomial* $\omega_m(z) := z^m$*, i.e., zero is the only interpolation node and it has multiplicity* m*. Let* f *be analytic in* $\mathbb{D}_R := \{z \ : \ |z| < R\}$ *and continuous on* $\overline{\mathbb{D}}_R := \{z \ : \ |z| \le R\}$ *for* $R > 0$*. The resulting* f*-interpolating polynomial* $q_{f,m}$ *is*

$$\begin{aligned} q_{f,m}(z) &= \frac{1}{2\pi \mathbb{i}} \int_{|\zeta|=R} \frac{\zeta^m - z^m}{\zeta - z} \frac{f(\zeta)}{\zeta^m}\, d\zeta \\ &= \frac{1}{2\pi \mathbb{i}} \int_{|\zeta|=R} \frac{1-(z/\zeta)^m}{1-(z/\zeta)} \frac{f(\zeta)}{\zeta}\, d\zeta \\ &= \sum_{i=0}^{m-1} \frac{1}{2\pi \mathbb{i}} \int_{|\zeta|=R} (z/\zeta)^i \frac{f(\zeta)}{\zeta}\, d\zeta \\ &= \sum_{i=0}^{m-1} \frac{f^{(i)}(0)}{i!} z^i \end{aligned} \tag{3.7}$$

for all $z \in \mathbb{D}_R$*. This is the truncated Taylor expansion of* f *at* 0*. We used the formula for the partial sum of a geometric sequence and the Cauchy integral formula. Since Taylor expansions converge uniformly in their convergence disk, we have* $q_{f,m}(z) \rightrightarrows f(z)$ *on* $\overline{\mathbb{D}}_R$ *for* $m \to +\infty$*. Lemma 3.15 implies that* $\|f(A)\boldsymbol{b} - q_{f,m}(A)\boldsymbol{b}\| \to 0$ *if we can choose* R *such that* $\varrho(A) \le R$*.*

Consider the function $f_\alpha(z) := (\alpha - z)^{-1}$*,* $0 \neq \alpha \in \mathbb{C}$*.* f_α *is analytic in* $\mathbb{D}_R$ *and continuous on* $\overline{\mathbb{D}}_R$ *for* $R < |\alpha|$*. Since* $f_\alpha^{(i)}(0) = i!\alpha^{-i-1}$*, equation (3.7) yields*

$$q_{f_\alpha,m}(z) = \sum_{i=0}^{m-1} \alpha^{-i-1} z^i = \frac{1-(z/\alpha)^m}{\alpha - z}.$$

The error satisfies

$$f_\alpha(z) - q_{f_\alpha,m}(z) = \frac{(z/\alpha)^m}{\alpha - z} \to 0 \tag{3.8}$$

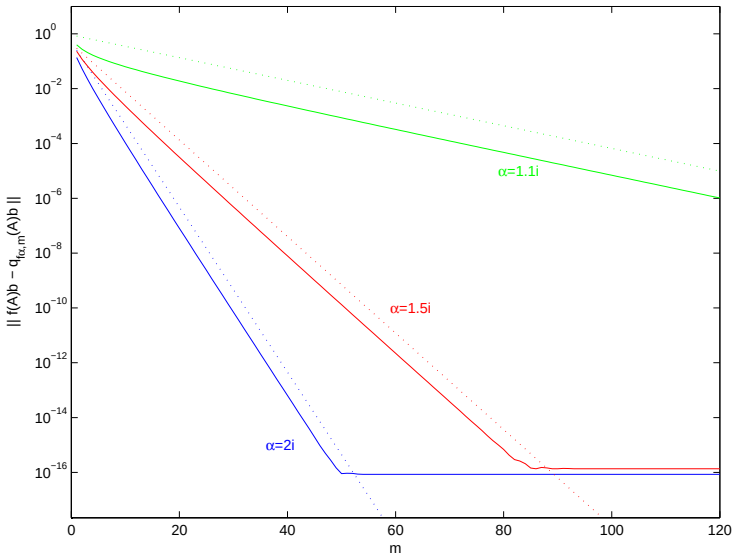


Figure 3.3:
The solid lines show $\|f_\alpha(A)\boldsymbol{b} - q_{f_\alpha,m}(A)\boldsymbol{b}\|$ for different values of α, corresponding to Example 3.17. In all cases $A \in \mathbb{C}^{101\times 101}$ is a diagonal matrix with equidistant eigenvalues in $[-1,1]$ and $\boldsymbol{b} = [1,1,\ldots,1]/\sqrt{101}$. The dotted line is the error bound (3.9).

for $m \to +\infty$ and all $|z| \le R$. For an arbitrary set $S \subseteq \mathbb{C}$ and $z \in \mathbb{C}$ we define

$$\operatorname{dist}(z,S) := \inf_{s\in S} |s - z|.$$

Using Lemma 3.15 and (3.8) *we obtain*

$$\|f_\alpha(A)\boldsymbol{b} - q_{f_\alpha,m}(A)\boldsymbol{b}\| \le \|\boldsymbol{b}\|_2 \frac{\varrho(A)^m}{|\alpha|^m \operatorname{dist}(\alpha,\Lambda(A))}, \tag{3.9}$$

i.e., we can expect the resulting polynomial method to converge fast if the singularity α is far from the origin and the eigenvalues of A. The latter should be centered tightly around zero such that $\varrho(A)$ is small.

Example 3.18. *Let $\omega_m(z) := z^m - 1$ and f be analytic in $\mathbb{D}_R$ and continuous on $\overline{\mathbb{D}}_R$, $R > 1$. The resulting interpolation polynomials are*

$$\hat{q}_{f,m}(z) = \frac{1}{2\pi \mathrm{i}} \int_{|\zeta|=R} \frac{\zeta^m - z^m}{\zeta - z} \frac{f(\zeta)}{\zeta^m - 1}\, d\zeta.$$

Because $\hat{q}_{f,m}$ has a similar form to $q_{f,m}$ from the previous example, we consider

$$\begin{aligned} \hat{q}_{f,m}(z) - q_{f,m}(z) &= \frac{1}{2\pi \mathrm{i}} \int_{|\zeta|=R} \frac{\zeta^m - z^m}{\zeta - z} \left(\frac{f(\zeta)}{\zeta^m - 1} - \frac{f(\zeta)}{\zeta^m} \right) d\zeta \\ &= \frac{1}{2\pi \mathrm{i}} \int_{|\zeta|=R} \frac{\zeta^m - z^m}{\zeta - z} \frac{f(\zeta)}{\zeta^{2m} - \zeta^m}\, d\zeta. \end{aligned}$$

Let $|z| = \rho > R$. *Then*

$$\begin{aligned} |\hat{q}_{f,m}(z) - q_{f,m}(z)| &\leq \frac{2\pi R}{2\pi} \max_{|z|=\rho,|\zeta|=R} \frac{|z^m\,((\zeta/z)^m - 1)\,f(\zeta)|}{|\zeta^{2m}(\zeta - z)(1 - 1/\zeta^m)|} \\ &= \frac{\rho^m}{R^{2m-1}} O(1) \qquad \text{for } m \to +\infty. \end{aligned} \tag{3.10}$$

This tends to zero if $\rho < R^2$. *The maximum principle yields that the maximum in the above formula is attained for* z *satisfying* $|z| = \rho$, *even if we allow* $z \in \overline{\mathbb{D}}_\rho$. *Hence* (3.10) *holds for all* $z \in \overline{\mathbb{D}}_\rho$. *With*

$$|f(z) - \hat{q}_{f,m}(z)| \leq |f(z) - q_{f,m}(z)| + |q_{f,m}(z) - \hat{q}_{f,m}(z)|$$

it is assured that $\|f(A)\boldsymbol{b} - \hat{q}_{f,m}(A)\boldsymbol{b}\| \to 0$ *if we can choose* R *sufficiently large that* $\varrho(A) < R$.

Consider again the function $f_\alpha(z) = (\alpha - z)^{-1}$. *It is easily verified that the interpolating polynomial* $\hat{q}_{f_\alpha,m}$ *is*

$$\hat{q}_{f_\alpha,m}(z) = \frac{1 - \omega_m(z)/\omega_m(\alpha)}{\alpha - z},$$

since it is a polynomial of degree $m - 1$ *that fulfills the interpolation conditions. Thus,*

$$\begin{aligned} \hat{q}_{f_\alpha,m}(z) &= \frac{1 - (z^m - 1)/(\alpha^m - 1)}{\alpha - z} \\ &= \frac{1 - (z/\alpha)^m}{(\alpha - \alpha^{-m+1})(1 - z/\alpha)} \\ &= \frac{1}{\alpha - \alpha^{-m+1}} \sum_{i=0}^{m-1} \frac{z^i}{\alpha^i} \end{aligned}$$

and

$$f_\alpha(z) - \hat{q}_{f_\alpha,m}(z) = \frac{z^m - 1}{(\alpha^m - 1)(\alpha - z)}.$$

Together with Lemma 3.15 we obtain

$$\|f_\alpha(A)\boldsymbol{b} - \hat{q}_{f_\alpha,m}(A)\boldsymbol{b}\| \leq \|\boldsymbol{b}\| \frac{\varrho(A)^m + 1}{|\alpha^m - 1|\,\operatorname{dist}(\alpha, \Lambda(A))}. \tag{3.11}$$

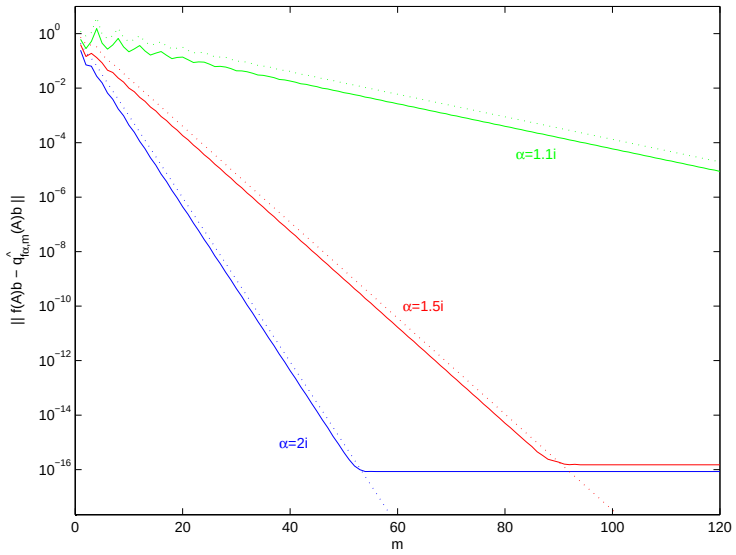


Figure 3.4:
The solid lines show $\|f_\alpha(A)\boldsymbol{b} - \hat{q}_{f_\alpha,m}(A)\boldsymbol{b}\|$ (Example 3.18). Here we interpolate at the m-th roots of unity. A and $\boldsymbol{b}$ are the same as in the previous example. The dotted line is the error bound (3.11).


3Cheby

Example 3.19. *We distribute m interpolation nodes on the interval $[-1,1]$ according to the roots of the Chebyshev polynomial $\widetilde{T}_m$, i.e., $\omega_m(z) := \widetilde{T}_m(z)$. By Lemma 1.19 the interpolation error is*

$$f(z) - \tilde{q}_{f,m}(z) = \frac{1}{2\pi\mathbb{i}} \int_{E_R} \frac{\widetilde{T}_m(z)}{\widetilde{T}_m(\zeta)} \frac{f(\zeta)}{\zeta - z}\, d\zeta,$$

where the curve E_R is the ellipse with semiaxes $\frac{1}{2}(R + R^{-1})$ and $\frac{1}{2}(R - R^{-1})$, $R > 1$. Furthermore, it is assumed that f is analytic in the interior of E_R and extends continuously to it. Recall that

$$\max_{z\in[-1,1]} \left|\widetilde{T}_m(z)\right| = \frac{1}{2^{m-1}}$$

due to Theorem 3.10. For $\zeta \in E_R$ we get from equation (3.5)

$$\left|\widetilde{T}_m(\zeta)\right| = \frac{R^m}{2^m} \left|1 + w^{-2m}\right|,$$

where $|w| = R$. It is easily verified that for any w with $|w| = R > 1$ there holds

$$0 < 1 - R^{-2} \le \left|1 + w^{-2m}\right|.$$

This yields

$$\begin{aligned} |f(z) - \tilde{q}_{f,m}(z)| &= \frac{1}{2\pi}\frac{1}{2^{m-1}}\frac{2^m}{R^m(1-R^{-2})}O(1) \\ &= \frac{1}{R^m}O(1) \end{aligned}$$

for all $z \in [-1,1]$. *Let* $\Lambda(A) \subset [-1,1]$. *By Lemma 3.15 there holds* $\|f(A)\boldsymbol{b} - \hat{q}_{f,m}(A)\boldsymbol{b}\| = \frac{1}{R^m}O(1)$ *for each function* f *that is analytic in the interior of* E_R *and extends continuously to it.*

The interpolation error for the function $f_\alpha(z) = 1/(\alpha - z)$ *is*

$$f_\alpha(z) - \tilde{q}_{f_\alpha,m}(z) = \frac{\widetilde{T}_m(z)}{\widetilde{T}_m(\alpha)(\alpha - z)}.$$

Choose $R > 1$ *such that* $\alpha \in E_R$. *Then*

$$\max_{z\in[-1,1]} |f_\alpha(z) - \tilde{q}_{f_\alpha,m}(z)| \leq \frac{2}{R^m(1-R^{-2})\operatorname{dist}(\alpha,[-1,1])}.$$

If all the eigenvalues of A *are contained in* $[-1,1]$ *we obtain with Lemma 3.15*

$$\|f_\alpha(A)\boldsymbol{b} - \tilde{q}_{f_\alpha,m}(A)\boldsymbol{b}\| \leq \|\boldsymbol{b}\|\frac{2}{R^m(1-R^{-2})\operatorname{dist}(\alpha,[-1,1])}. \tag{3.12}$$

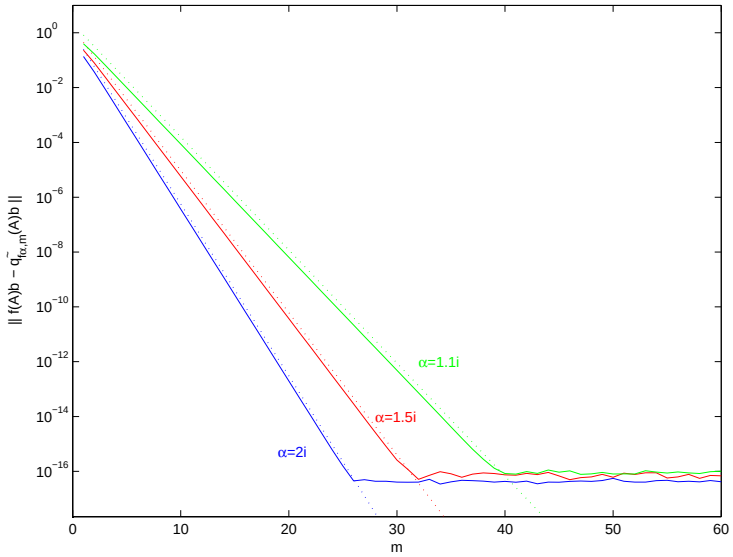


Figure 3.5:
The solid lines show $\|f_\alpha(A)\boldsymbol{b} - \tilde{q}_{f_\alpha,m}(A)\boldsymbol{b}\|$ (Example 3.19). Here we interpolate at the roots of $\widetilde{T}_m$. A and $\boldsymbol{b}$ are the same as in the previous examples. The dotted lines show the error bound (3.12). Note the different scaling of the m-axis.

Error Estimate for Arnoldi Approximations

When choosing the Ritz(m) values of A as interpolation nodes, we can show a stronger assertion than Lemma 3.15 provides. Roughly speaking, the following Lemma asserts that the Arnoldi approximations $\boldsymbol{f}_m$ (cf. Algorithms 2.22 and 2.28) are not worse than twice the best we can get from any polynomial method for a slightly larger matrix $\tilde{A}$ that satisfies $\Lambda(\tilde{A}) = \Lambda(A) \cup \Lambda(H_m)$.

Lemma 3.20. *Let A be normal and $\Lambda(A) \cup \Lambda(H_m) \subseteq \Omega$, Ω compact. Then the Arnoldi approximations $\boldsymbol{f}_m$ fulfill*

$$\|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| \leq 2\|\boldsymbol{b}\| \min_{p \in \mathcal{P}_{m-1}} \max_{\lambda \in \Omega} |f(\lambda) - p(\lambda)|.$$

Proof. Let $p \in \mathcal{P}_{m-1}$. Then $p(A)\boldsymbol{b} = V_m p(H_m) V_m^H \boldsymbol{b}$, as asserted in Lemma 2.20. With the definition of $\boldsymbol{f}_m$ we get

$$\begin{aligned} \|f(A)\boldsymbol{b} - \boldsymbol{f}_m\| &= \|f(A)\boldsymbol{b} - V_m f(H_m) V_m^H \boldsymbol{b} + V_m p(H_m) V_m^H \boldsymbol{b} - p(A)\boldsymbol{b}\| \\ &\leq \|\boldsymbol{b}\| \left(\|f(A) - p(A)\| + \|f(H_m) - p(H_m)\|\right) \\ &\leq 2\|\boldsymbol{b}\| \max_{\lambda \in \Omega} |f(\lambda) - p(\lambda)|. \end{aligned}$$

We take the infimum among all $p \in \mathcal{P}_{m-1}$ over this inequality and note that this infimum is attained because of Theorem 3.2. The proof is complete. □

3.5 Interpolation in Uniformly Distributed Points

Let $\Omega \subset \mathbb{C}$ be a compact set such that its complement $\Omega^C = \mathbb{C} \setminus \Omega$ is a simply connected domain. Then by *Riemann Mapping Theorem* there exists a conformal map

$$z = \Psi(w) = cw + c_0 + c_1 w^{-1} + \cdots$$

from $\overline{\mathbb{D}}^C$ onto Ω^C with $\Psi(\infty) = \infty$ and $\Psi'(\infty) = c > 0$. c is called the *capacity* of $\partial\Omega$. By

$$\Phi(z) = c^{-1} z + \cdots$$

we denote the inverse function of Ψ. Note that Φ is a conformal map from Ω^C onto $\overline{\mathbb{D}}^C$. For every $R > 1$ we define the *level curve*

$$L_R := \{z \in \mathbb{C} \ : \ |\Phi(z)| = R\} \subset \Omega^C.$$

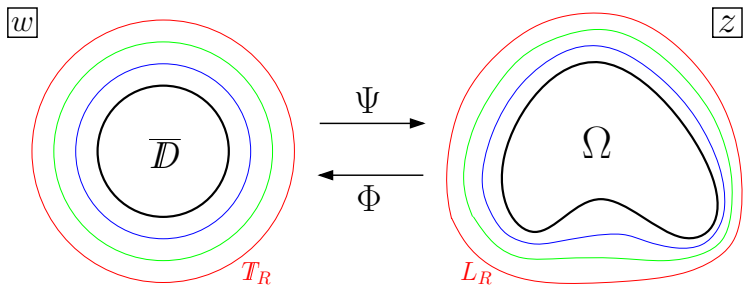


Figure 3.6: The conformal maps Ψ, Φ and the level curves L_R.

All the L_R are disjoint Jordan curves because they are the image of $\mathbb{T}_R$ under the bijective analytic transformation Ψ.

Let $(\omega_m)_{m\geq 1}$ be a sequence of *nodal polynomials for* Ω, i.e., each ω_m is a monic polynomial of degree m and all of its roots $\mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,m}$ are contained in Ω. We define the numbers

$$M_m := \|\omega_m\|_\Omega = \max_{z\in\Omega} |\omega_m(z)|.$$

By the *maximum principle*, the maximum M_m is attained on $\partial\Omega = \partial(\Omega^C)$ and there holds

$$M_m \geq c^m \qquad \text{for } m = 1, 2, \ldots \tag{3.13}$$

To prove this, we consider the function

$$H_m(z) := \frac{\omega_m(z)}{(c\,\Phi(z))^m},$$

which is analytic in Ω^C. Moreover, $H_m(z) \to 1$ as $z \to \infty$. The *maximum principle* implies

$$\max_{z\in L_R} |H_m(z)| \geq 1 \qquad \text{for all } R > 1,$$

and, since $\Phi(z) = R$ for $z \in L_R$, this yields

$$\max_{z\in L_R} |\omega_m(z)| \geq (cR)^m \qquad \text{for all } R > 1.$$

By taking $R \to 1$, the assertion (3.13) is obtained. The following definition is now justified.

Definition 3.21. *The nodes associated with the sequence $(\omega_m)_{m\geq 1}$ of nodal polynomials for Ω are* uniformly distributed on Ω *if*

$$\sqrt[m]{M_m} \to c \qquad \textit{for } m \to +\infty. \tag{3.14}$$

Example 3.22. *The roots of the (normalized) Chebyshev polynomials* $\widetilde{T}_m(z)$ *are uniformly distributed on* $\Omega = [-1, 1]$:

By Theorem 3.10 we have $M_m = 1/2^{m-1}$. *The conformal map from* $\overline{\mathbb{D}}^C$ *onto* $[-1, 1]^C$ *is the Joukowski transformation* $z = \Psi(w) = \frac{1}{2}(w + w^{-1})$. *Hence* $c = \frac{1}{2}$ *and the condition* (3.14) *is satisfied.*

Example 3.23. *The roots of the (shifted) Chebyshev polynomials* $T_m^K(z)$ *are uniformly distributed on* K.

Now let f be *analytic on* Ω (i.e., analytic in an open subset of $\mathbb{C}$ that contains Ω). By $q_{f,m}(z)$ we denote the Hermite interpolating polynomial of degree $m-1$ that interpolates f at the roots of ω_m. The following theorem gives the connection between the uniform distribution of the nodes and the convergence of the corresponding interpolation process.

Theorem 3.24 (Kalmár-Walsh). *The convergence*

$$q_{f,m}(z) \Rightarrow f(z) \qquad (z \in \Omega,\ m \to +\infty)$$

takes place for each function f *analytic on* Ω *if and only if the interpolation nodes are uniformly distributed on* Ω.

Proof. See Gaier [10, pp. 65–66]. □

The following theorem gives an assertion about the rate of convergence.

Theorem 3.25. *Assume that* $R > 1$ *is the largest number such that* f *is analytic inside* L_R. *The interpolating polynomials* $q_{f,m}$ *with uniformly distributed nodes on* Ω *then satisfy the condition*

$$\limsup_{m \to +\infty} \sqrt[m]{\|f - q_{f,m}\|_\Omega} = \frac{1}{R} =: k(\Omega, f). \tag{3.15}$$

Proof. See Gaier [10, pp. 66–67]. □

Definition 3.26. *A sequence of* f*-interpolating polynomials* $(q_{f,m})_{m \geq 1}$ converges maximally to f on Ω *if the condition* (3.15) *is satisfied. The number* $k(\Omega, f)$ *is called the* asymptotic convergence factor *of the sequence* $\left(\sqrt[m]{\|f - q_{f,m}\|_\Omega}\right)_{m \geq 1}$.

The term 'maximal convergence' was introduced by Walsh, see [32, p. 79]. It is justified by the fact that $1/R$ is the best possible (i.e., smallest) asymptotic convergence factor that holds for *all* functions which are analytic inside L_R. If f is an *entire* function (i.e., analytic in the whole complex plane $\mathbb{C}$), the constant R may be chosen arbitrarily large. In this case we expect *superlinear convergence*.

Interpolation in Fejér Points

Let Ω have a sufficiently smooth boundary $\partial\Omega$, e.g. a Jordan curve. By the *Theorem of Caratheodory-Osgood* there exists a bijective continuous extension $\widetilde{\Psi} : \mathbb{D}^C \to \Omega^C \cup \partial\Omega$ of Ψ to the boundary.

Definition 3.27. *The* Fejér points $\{\mu_{m,j} : j = 1, 2, \ldots, m\}$ of order m on Ω *are the images under* $\widetilde{\Psi}$ *of the* m*-th roots of unity, i.e.,*

$$\mu_{m,j} := \widetilde{\Psi}\left(\exp\left(2\pi\mathbb{i}(j-1)/m\right)\right) \qquad \text{for } j = 1, 2, \ldots, m.$$

Theorem 3.28 (Fejér). *The Fejér points are uniformly distributed on* Ω*.*

Proof. See Gaier [10, p. 67–69]. □

Application to the Generalized Interpolation Method

We turn back to the generalized interpolation method, Algorithm 2.34. Recall that A is required to be normal. Choose a compact set Ω such that $\Lambda(A) \subset \Omega$ and Ω^C is a simply connected domain. Let f be analytic on Ω and R chosen according to Theorem 3.25. By $\{q_{f,m}\}_{m\geq 1}$ we denote a sequence of f-interpolating polynomials of degree $m-1$ to f with uniformly distributed nodes on Ω. By Lemma 3.15 and Theorem 3.25 we have immediately that

$$\limsup_{m\to+\infty} \left(\frac{\|f(A)\boldsymbol{b} - q_{f,m}(A)\boldsymbol{b}\|}{\|\boldsymbol{b}\|} \right)^{1/m} \leq \frac{1}{R}.$$

Now it seems reasonable to use uniformly distributed interpolation points for the generalized interpolation method. In this case we know that the error should behave asymptotically like R^{-m}.


3Fejer

Example 3.29. *Let* $f(z) := 1/z$. *Let* $A \in \mathbb{C}^{100\times 100}$ *be a normal matrix with randomly and evenly distributed eigenvalues inside a L-shaped polygon* Ω, *see Figure 3.7(a). We compute the map* $\widetilde{\Psi}$ *using the* Schwarz-Christoffel-toolbox *for* MATLAB *(see Driscoll [3]). With the help of the function* `evalinv` *we determine*

$$R \approx 1.6421,$$

which is the value for which the origin 0 *lies on the level curve* L_R. *See also Figure 3.7(b). For a fixed* m *we determine the Fejér points of order* m *on* Ω *and evaluate the interpolating polynomial* $q_{f,m}$ *that interpolates* f *at the Fejér points. For this task we use the* variable precision arithmetic *of* MAPLE *in order to avoid stability problems. A more practical implementation makes use of recurrence schemes to compute the coefficients of* $q_{f,m}$ *in* Newton form, *see Novati [24]. Finally, the Krylov approximation to* $f(A)\boldsymbol{b}$ *is* $\boldsymbol{g}_m := q_{f,m}(A)\boldsymbol{b}$. *In Figure 3.7(d) we plot the logarithmic error* $\log(|f(z)-q_{f,m}(z)|+\varepsilon)$ *for* $m = 16$. *(The small positive constant* ε *is added to avoid* $\log(0)$, *that would be attained at least in the interpolation points.) Note how well* f *is approximated on* Ω, *indicated by the dark blue color.*

The greatest advantage of generalized interpolation methods with uniformly distributed interpolation points *(GIMUD) is, that the interpolating polynomials* $q_{f,m}$ *can be applied to every matrix whose spectrum is contained in* Ω *without worsening the asymptotic convergence factor* $1/R$. *Once* $q_{f,m}$ *is determined,* $q_{f,m}(A)\boldsymbol{b}$ *is easily evaluated for different* A *and* $\boldsymbol{b}$, *e.g., using the* Horner scheme. *Such problems arise very often, for example, if we solve a partial differential equation with the* method of lines. *In this case we will have to solve a* set *of ordinary differential equations and this involves the evaluation of* $\exp(tA)\boldsymbol{b}$, *where* $t > 0$ *and* $\boldsymbol{b}$ *varies for each evaluation.*

One of the drawbacks of (GIMUD) is that we first have to know at least the outlying eigenvalues of A *in order to determine* Ω. *Therefor we may determine some Ritz(m) values of* A. *They often have the property to approximate the eigenvalues of* A *at the edge of the spectrum, even if* m *is small (cf. Chapter 4). Because we first have to run another Krylov subspace method in order to use (GIMUD), it is often called a* hybrid method.

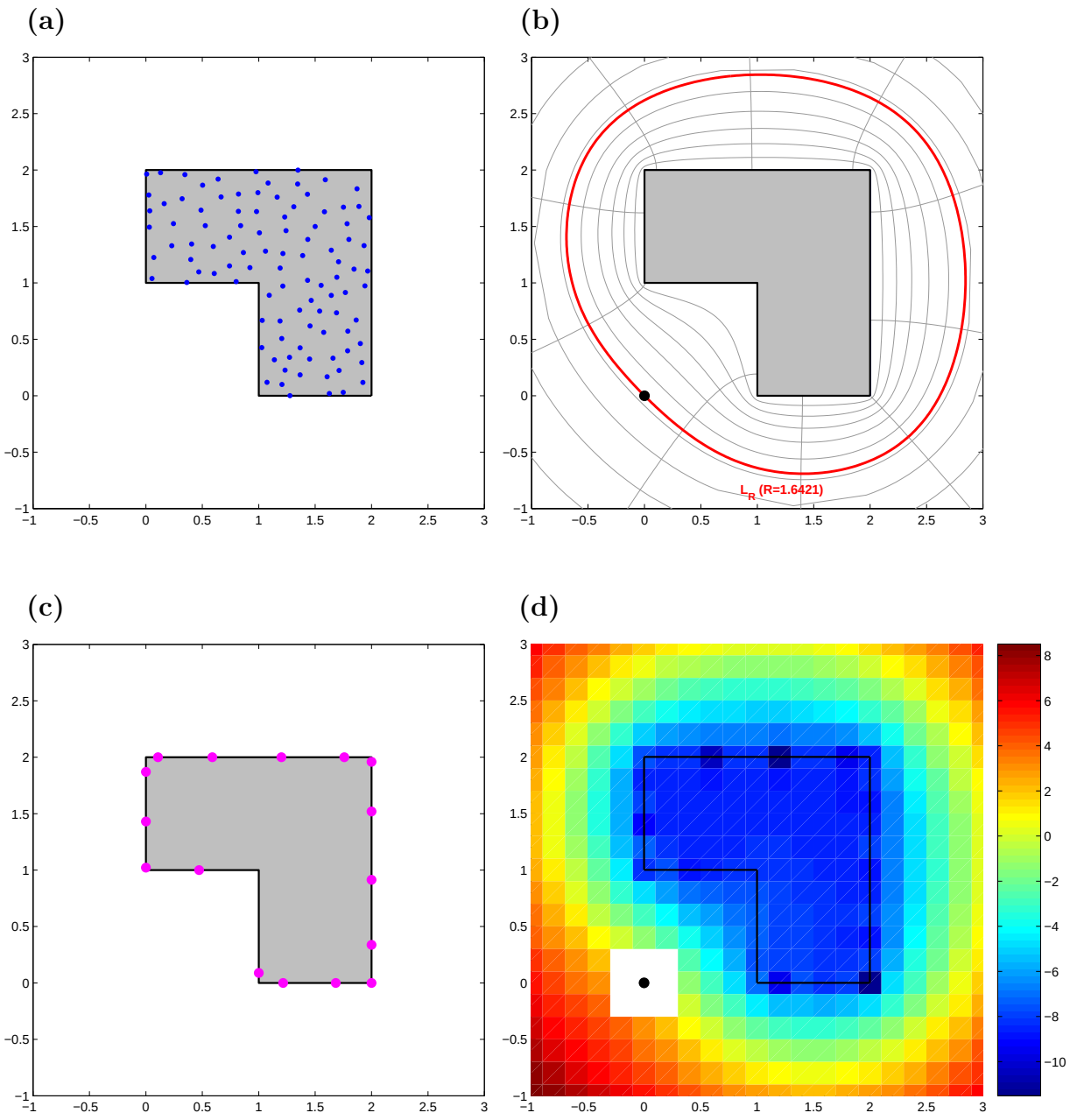


Figure 3.7:
Illustration to Example 3.29. (a) L-shaped polygon Ω (grey filled) with the 100 random eigenvalues (blue dots). (b) The image of an orthogonal grid under the map Ψ (grey lines). The critical level curve L_R is in red. The black dot in the origin indicates the singularity of f. (c) Fejér points of order 16 on Ω. (d) The colors indicate the value of $\log(|f(z) - q_{f,m}(z)| + \varepsilon)$.

In Figure 3.8 we plot the error curves of the interpolation method using

- *Fejér points on Ω (magenta),*
- *equidistant points on the boundary of the polygon (blue),*
- *Ritz values (green)*

as interpolation nodes. The erratic behavior of the blue error curve is caused by a strong oscillation of the corresponding interpolation polynomials inside Ω if the degree is high.

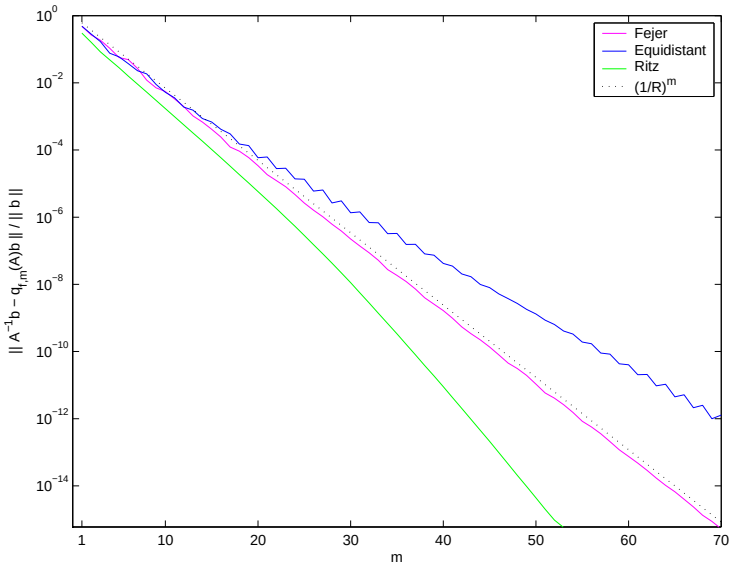


Figure 3.8:
Error of the interpolation method using different interpolation nodes. The dotted line indicates the asymptote R^{-m}.

3.6 Convergence of the CG Method

Recall from Section 2.7 that the iterates of the CG method $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ minimize the A-norm of the error $\boldsymbol{e}_m = A^{-1}\boldsymbol{b} - \boldsymbol{x}_m$:

$$\|\boldsymbol{e}_m\|_A = \min_{p \in \mathcal{P}_m^0} \|p(A)A^{-1}\boldsymbol{b}\|_A.$$

A is symmetric positive definite and therefore normal. Thus,

$$\begin{aligned} \|\boldsymbol{e}_m\|_A &= \min_{p\in\mathcal{P}_m^0} \|A^{1/2}p(A)A^{-1}\boldsymbol{b}\| \\ &= \min_{p\in\mathcal{P}_m^0} \|Up(D)U^H A^{1/2}A^{-1}\boldsymbol{b}\| \\ &\leq \min_{p\in\mathcal{P}_m^0} \|p(D)\|\|A^{1/2}A^{-1}\boldsymbol{b}\| \\ &= \|A^{-1}\boldsymbol{b}\|_A \min_{p\in\mathcal{P}_m^0} \max_{\lambda\in\Lambda(A)} |p(\lambda)|. \end{aligned} \tag{3.16}$$

(Although $A^{1/2}$ is not uniquely determined, $\|A^{1/2}\boldsymbol{v}\|$ is unique for every vector $\boldsymbol{v}$.)

The problem

$$M := \min_{p\in\mathcal{P}_m^0} \max_{\lambda\in\Lambda(A)} |p(\lambda)|$$

is a polynomial uniform best approximation problem on the discrete set $\Lambda(A)$. Let $\lambda_{\min}$ ($\lambda_{\max}$) denote the smallest (largest) eigenvalue of A and set $\kappa := \lambda_{\max}/\lambda_{\min}$. We replace $\Lambda(A)$ by the interval $K := [\lambda_{\min}, \lambda_{\max}]$. There holds

$$M \leq \tilde{M} := \min_{p\in\mathcal{P}_m^0} \max_{\lambda\in K} |p(\lambda)|,$$

because a polynomial $\tilde{p} \in \mathcal{P}_m^0$ for which the minimum $\tilde{M}$ is attained fulfills $|\tilde{p}(\lambda)| \leq \tilde{M}$ for all $\lambda \in \Lambda(A)$. The minimizer $\tilde{p}$ is the shifted Chebyshev polynomial on K of degree m. Thus, Lemma 3.12 yields

$$\tilde{M} = 2\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{-m}\right)^{-1}.$$

Using (3.16) we obtain the following error bound for the CG method

$$\frac{\|\boldsymbol{e}_m\|_A}{\|A^{-1}\boldsymbol{b}\|_A} \leq 2\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{-m}\right)^{-1} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m. \tag{3.17}$$

Example 3.30. *We consider four symmetric positive definite matrices* A_i :

- A_1 *has* 100 *eigenvalues at the roots of* $T_{100}^{[1,100]}$, *i.e., the eigenvalues are uniformly distributed on the interval* $[1, 100]$,
- A_2 *has* 100 *equidistant eigenvalues in the interval* $[1, 100]$,
- A_3 *has* 98 *equidistant eigenvalues in the interval* $[20, 80]$ *and two separated eigenvalues* $\{1, 100\}$,
- A_4 *has* 100 *equidistant eigenvalues in the interval* $[20, 80]$.

In Figure 3.9 we plot the error norms

$$\frac{\|\boldsymbol{e}_m\|_A}{\|A^{-1}\boldsymbol{b}\|_A}$$

of the CG iterates $\boldsymbol{x}_m$. *Note that the error bound* (3.17) *is the same for the matrices* A_1, A_2 *and* A_3, *since* $\kappa = 100$ *for all of them. Thus, we expect a geometric decrease of the error with rate*

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = \frac{9}{11}.$$

For the matrix A_4 *we have* $\kappa = 4$ *and therefore expect a geometric decrease of the error with rate* $1/3$.

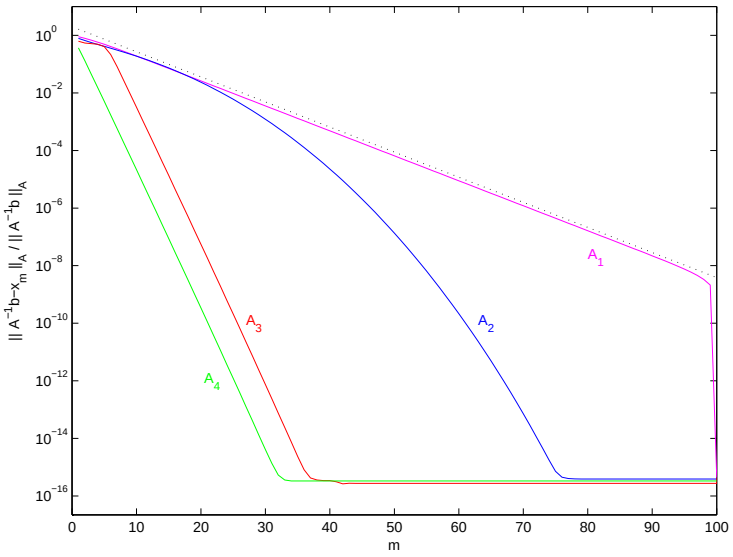


Figure 3.9:
Error of the CG iterates for the matrices A_i. The dotted line is the error bound (3.17) for $\kappa = 100$.

We make the following observations:

(i) *The error curve for the matrix A_1 (magenta) behaves as predicted by the dotted error bound* (3.17)*. This suggests that uniformly distributed eigenvalues are the worst case for the convergence behavior of the CG method. This observation can also be made for polynomial interpolation methods in general.*

(ii) *The error curves for A_2 (blue) and A_3 (red) decrease much faster than the error bound suggests, so that we have a much too pessimistic prediction of the error decrease. The reason is that the error bound* (3.17) *does not take into account the* fine structure *of the spectrum, i.e., the* distribution *of the eigenvalues in the interior of the interval* $[1, 100]$*.*

(iii) *After a few initial iterates, the error curve for A_3 behaves like the error curve for A_4 (green), although the predicted convergence rates differ by a factor ≈ 2.5. After the separated eigenvalues $\{1, 100\}$ of A_3 'have been found' by the underlying interpolation process, the interpolation actually takes place on a smaller interval* $[20, 80]$*.*

Several attempts have been made to improve the error bound (3.17), for example if A has one eigenvalue much larger than the others, say, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N-1} \ll \lambda_N$ or if the spectrum of A is well approximated by the union of two disjoint intervals. Some results can be found in Greenbaum [13, pp. 52–54]. In what follows, we want to introduce another approach that involves the fine structure of the spectrum of A with the help of a distribution function σ.

4 On the Convergence of Ritz Values

The expression 'a Ritz value has converged' is more a heuristic description than a mathematical term that can be defined exactly. We recall from Chapter 2 that the Ritz(m) values $\mu_{m,1}, \mu_{m,2}, \ldots, \mu_{m,m}$ of a matrix $A \in \mathbb{C}^{N\times N}$ are the eigenvalues of the unreduced upper Hessenberg matrix H_m generated by the Arnoldi process for an initial vector $\boldsymbol{b} \in \mathbb{C}^{N\times N}$. The Arnoldi process was given in Algorithm 2.14, page 29. The Ritz(m) polynomial χ_m is the characteristic polynomial of H_m, i.e., $\chi_m(z) = \det(zI - H_m)$. By Lemma 2.27 it is known that χ_m is the minimizing argument of $\|p(A)\boldsymbol{b}\|$ among all monic polynomials $p \in \mathcal{P}_m^\infty$. We will denote this by

$$\chi_m = \arg\min_{p\in\mathcal{P}_m^\infty} \|p(A)\boldsymbol{b}\|. \tag{4.1}$$

As before, $\|\cdot\|$ denotes the 2-norm of a vector or a matrix. Moreover, we retain the assumption that A is a normal matrix and thus can be written in the form

$$A =: UDU^H$$

with $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and an unitary matrix $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N]$. The vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N$ are the orthonormal eigenvectors of A associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$.

4.1 A Least Squares Problem

By the properties of the 2-norm, equation (4.1) can be rewritten as

$$\chi_m = \arg\min_{p\in\mathcal{P}_m^\infty} \sum_{i=1}^{N} |\langle \boldsymbol{u}_i, \boldsymbol{b}\rangle|^2 |p(\lambda_i)|^2, \tag{WLS}$$

which is a *weighted least squares problem* for the values of χ_m at the eigenvalues of A. This point of view gives an intuition how the Ritz values depend on the initial vector $\boldsymbol{b}$.

Example 4.1. *Let $\boldsymbol{b} \in \mathbb{C}^N$ be a linear combination of $m < N$ eigenvectors $\boldsymbol{u}_i$ of A, say without loss of generality*

$$\boldsymbol{b} := \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \cdots + \alpha_m \boldsymbol{u}_m, \qquad \text{where all } \alpha_i \neq 0.$$

Then

$$(z - \lambda_1)(z - \lambda_2) \cdots (z - \lambda_m) = \chi_m(z)$$

is the unique minimizer of (WLS) *because the polynomial χ_m is zero on all the eigenvalues λ_i of A that have nonzero weight $|\langle \boldsymbol{u}_i, \boldsymbol{b} \rangle|^2 = |\alpha_i|^2$. Such eigenvalues are said to be* active in $\boldsymbol{b}$. *We say an eigenvalue λ is* found *by a Ritz(m) value μ if the Ritz(m) polynomial χ_m has a root 'very close' to it. In our case, the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ are found since they are exactly the roots of χ_m. From $\|\chi_m(A)\boldsymbol{b}\| = 0$ and the fact that there is no monic polynomial of smaller degree with this property, it follows that χ_m is the minimal polynomial of $\boldsymbol{b}$ with respect to A.*

Now let $n > m$. It still makes sense to ask for a monic polynomial χ_n of degree n that minimizes (WLS)*, although the minimizer is no longer unique: every $\chi_n \in \mathcal{P}_n^\infty$ that is divided by χ_m minimizes* (WLS) *because $\lambda_1, \lambda_2, \ldots, \lambda_m$ are among its roots. Note that this is not a contradiction to Lemma 2.27 since the Ritz(n) polynomial does not exist. If we want the solution of* (WLS) *to be the Ritz(m) polynomial for all $m \leq N$, we have to assure the existence of the latter. Recall from Chapter 2 that we can run the Arnoldi process until $m = N$ if and only if the Krylov subspaces $\mathcal{K}_m(A, \boldsymbol{b})$ do not become stationary for $m < N$. This will not happen if and only if A is nonderogatory and $\boldsymbol{b}$ is cyclic for A.*

Assumption I. *Let A be a nonderogatory (and normal) matrix and $\boldsymbol{b}$ cyclic for A.*

Lemma 4.2. *This assumption assures $\langle \boldsymbol{u}_i, \boldsymbol{b} \rangle \neq 0$ for $i = 1, 2, \ldots, N$ and therefore the uniqueness of the solution of* (WLS) *for all degrees $m \leq N$.*

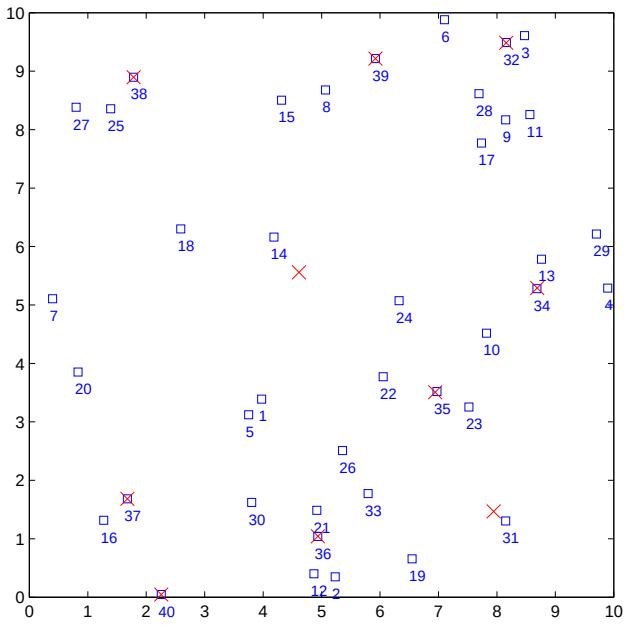
Proof. Assume the assertion is wrong, say $\langle \boldsymbol{u}_1, \boldsymbol{b} \rangle = 0$ without loss of generality. Then $\boldsymbol{b} = \sum_{i=2}^N \alpha_i \boldsymbol{u}_i$ and $U^H \boldsymbol{b} = [0, \alpha_2, \ldots, \alpha_N]^T$. For $m = 0, 1, \ldots, N-1$ we have

$$\begin{aligned}
\langle \boldsymbol{u}_1, A^m \boldsymbol{b} \rangle &= \langle \boldsymbol{u}_1, U D^m U^H \boldsymbol{b} \rangle \\
&= \left\langle \boldsymbol{u}_1, U \operatorname{diag}(\lambda_1^m, \lambda_2^m, \ldots, \lambda_N^m)[0, \alpha_2, \ldots, \alpha_N]^T \right\rangle \\
&= \left\langle \boldsymbol{u}_1, \sum_{i=2}^N \lambda_i^m \alpha_i \boldsymbol{u}_i \right\rangle \\
&= 0,
\end{aligned}$$

since $\{\boldsymbol{u}_i : i = 1, 2, \ldots, N\}$ is an orthonormal basis of $\mathbb{C}^N$. But this means that $\boldsymbol{0} \neq \boldsymbol{u}_1 \perp \mathcal{K}_N(A, \boldsymbol{b})$, hence $\mathcal{K}_N \neq \mathbb{C}^N$ and therefore $\boldsymbol{b}$ is not cyclic for A. This is a contradiction. □

We are still left with one problem. What happens if all the weights $|\langle \boldsymbol{u}_i, \boldsymbol{b} \rangle|^2$ are nonzero, but differ extremely in value?

In view of (WLS) it is necessary for the minimizing polynomial χ_m to be small at those eigenvalues λ_i which are associated with eigenvectors $\boldsymbol{u}_i$ that have a large weight $|\langle \boldsymbol{u}_i, \boldsymbol{b} \rangle|^2$ (relative to the eigenvectors with a small weight). This means that χ_m should 'prefer' to have its roots close those eigenvalues. Therefore we expect the eigenvalues associated with eigenvectors of a large weight to be found early (i.e., for small m) by Ritz(m) values. Figure 4.1 shows one example where this actually happens.



4weight

Figure 4.1:
The blue squares are the 40 eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{40}$ of a random normal matrix A, $A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_{40}) U^H$, where $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{40}]$. The vector $\boldsymbol{b} \in \mathbb{C}^{40}$ was chosen as $\boldsymbol{b} = \sum_{i=1}^{40} 2^i \boldsymbol{u}_i$. The red crosses show the Ritz(10) values of A. They lie close to the eigenvalues $\lambda_{40}, \lambda_{39}, \ldots$ because the associated eigenvectors $\boldsymbol{u}_{40}, \boldsymbol{u}_{39}, \ldots$ have a large component in $\boldsymbol{b}$.

Outside of such constructed scenarios we can assume that the weights $|\langle \boldsymbol{u}_i, \boldsymbol{b}\rangle|^2$ do not differ strongly in value. Moreover, we even should not assume that these weights are known, since their computation may be expensive in general. Therefore it is rather a must than a drawback to carry out our investigations without considering the influence of the starting vector $\boldsymbol{b}$. To get rid of $\boldsymbol{b}$ we consider the ideal case, that is

Assumption II.

$$|\langle \boldsymbol{u}_1, \boldsymbol{b}\rangle|^2 = |\langle \boldsymbol{u}_2, \boldsymbol{b}\rangle|^2 = \cdots = |\langle \boldsymbol{u}_N, \boldsymbol{b}\rangle|^2 = \text{const}\,.$$

At first glance this assumption seems to be too restrictive, and indeed, for the theory of Beckermann and Kuijlaars on the convergence of Ritz values it is sufficient to assume that the eigenvector components $|\langle \boldsymbol{u}_i, \boldsymbol{b}\rangle|$ do not vary exponentially in value (cf. Kuijlaars [18, p. 7]). Nevertheless we do not loose generality here. In fact, Beckermann and Kuijlaars make Assumption II implicitly but argue afterwards, see Kuijlaars [18, p. 23], that the theory holds more generally because small variations in $|\langle \boldsymbol{u}_i, \boldsymbol{b}\rangle|$ are not felt as $N \to +\infty$ and we will only obtain results in the asymptotic sense.

Under Assumption II the problem (WLS) reduces to

$$\chi_m = \arg\min_{p \in \mathcal{P}_m^\infty} \sum_{i=1}^{N} |p(\lambda_i)|^2, \tag{LS}$$

which is an *(unweighted) least squares problem* on the eigenvalues of A. If Assumption II does not hold, but the weights do not differ very strongly, we hope that the 'true' Ritz polynomial χ_m from (WLS) will have its roots 'close' to the roots of the polynomial from the ideal case (LS). Note that both polynomials are monic and therefore uniquely determined by their roots.

To motivate our further ongoing, we recall the Lanczos process from Chapter 2. We observed that the Ritz values produced by this algorithm lie in the real interval $[\lambda_{\min}, \lambda_{\max}]$ and fulfill the interlacing property, Theorem 2.31. But there is something more that can be observed concerning the location of the Ritz values over the course of the Lanczos process. We want to demonstrate this with the help of the following examples.


4Lanczosp

Example 4.3. *We consider a diagonal matrix A with 40 equidistant eigenvalues in the interval $[-1,1]$ and a vector $\boldsymbol{b} := [1,1,\ldots,1]^T$ of length 40. Note that the orthonormal eigenvectors of A are the unit coordinate vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N$. Thus, $\langle \boldsymbol{u}_i, \boldsymbol{b} \rangle = 1$ for $i = 1,2,\ldots,N$ and Assumption II is fulfilled. We run the Lanczos process until $m = 29$. In Figure 4.2 we plot the Ritz polynomial $\chi_{29}(z)$. The red squares indicate the values of χ_{29} at the eigenvalues of A. The remarkable fact is that χ_{29} has some of its roots very close to those eigenvalues of A that are located near the edge of the interval $[-1,1]$. These eigenvalues have been found by Ritz(29) values. It can be observed that, once an eigenvalue has been found by a Ritz value, it remains found as the degree m is increased further. Because of this 'convergence-like' behavior we say that this Ritz value* has converged *to an eigenvalue of A.*

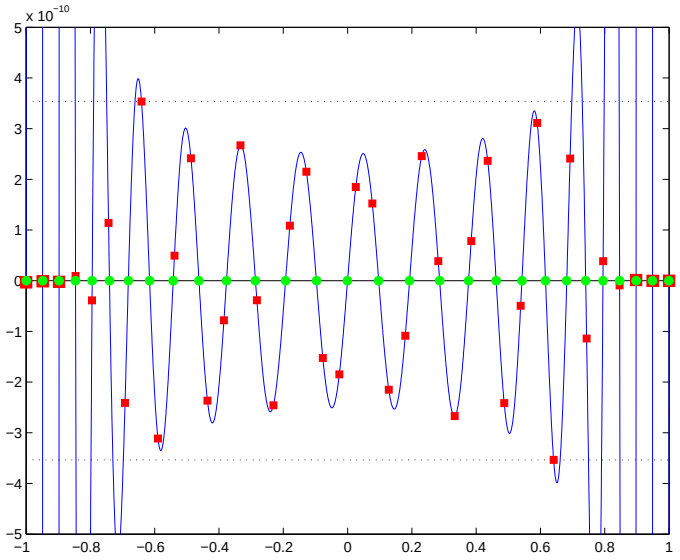


Figure 4.2:
The blue line is the graph of the Ritz(29) polynomial $\chi_{29}(z)$ on the interval $[-1,1]$. The red squares indicate the value of χ_{29} on the 40 equidistant eigenvalues of A. The green dots mark the roots of χ_{29}, which are the Ritz(29) values of A.


4interval

In Figure 4.3 one can see how the Ritz values begin to approximate the spectrum of A from the outside to the inside of the interval. The black circles indicate the location of the 40 eigenvalues of A. The colored disks indicate those eigenvalues that have been found by a particular Ritz(m) value (or: some Ritz(m) values have converged to those eigenvalues). The color of the disk encodes the value of m for which convergence sets in (see the figure's legend). For example, we plot a red disk

at a certain eigenvalue λ if there exists an index $\widetilde{m} \in \{30, 31, 32, 33\}$ such that each Ritz polynomial χ_m with $m \geq \widetilde{m}$ has a root within distance `tol` *to λ. Here we have chosen* `tol`$:= 10^{-3}$. *This convergence check is easily implemented and avoids misinterpreting 'lucky guesses' as converged Ritz values. Unfortunately, it can only be applied if the spectrum $\Lambda(A)$ is known a priori. The empty circles in the middle of the interval indicate the eigenvalues of A that are not found until $m = 40$.*

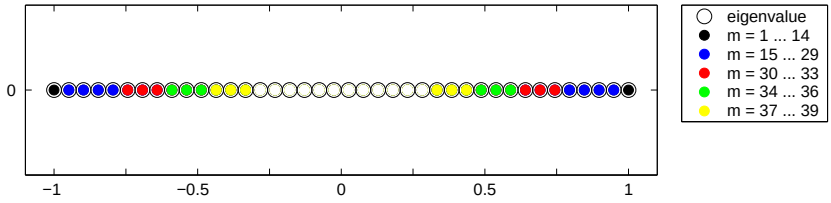


Figure 4.3:
The black circles indicate the eigenvalues of the matrix A. The colored disks encode from which index m on an eigenvalue is found by a Ritz(m) value.

The phenomenon that the eigenvalues on the *edge* of the spectrum are found first by Ritz values is not restricted to Hermitian matrices.

Example 4.4. *In Figure 4.4 we consider a complex random diagonal matrix A of size 100×100 (the real and imaginary parts of the diagonal entries are normally distributed with mean 0 and variance 1). We set $\boldsymbol{b} := [1, 1, \ldots, 1]^T$ and* `tol`$:= 10^{-3}$. *As above, the innermost eigenvalues of A are not found until the end of the Arnoldi process, whereas the outermost ones are approximated very early by a Ritz value.*


4random

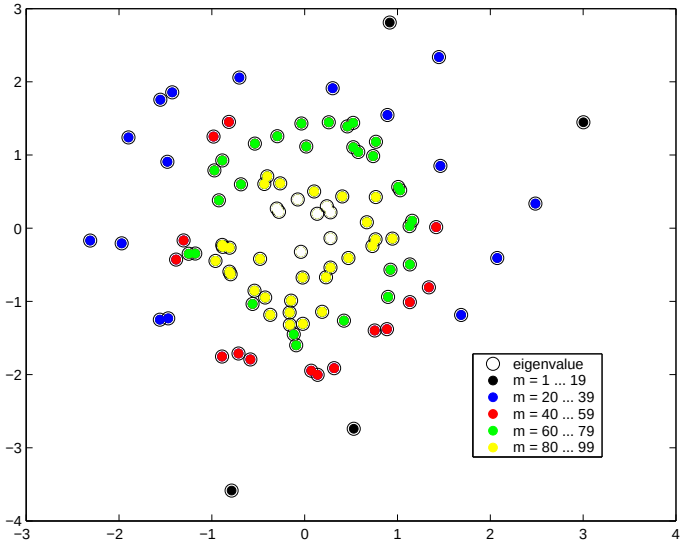


Figure 4.4: Convergence of the Ritz values of a non-Hermitian matrix.

The aim of this chapter can now be stated as follows: try to characterize the regions where eigenvalues are found by Ritz values and the regions where this is not the case. One way to accomplish this task is to use potential theoretic tools in the complex plane. In the recent years, a lot of effort has been put into this approach, mainly by Bernhard Beckermann and Arno B. J. Kuijlaars. Here we will shortly present their theory on the convergence of Ritz values and refer also to the articles [1, 18, 19].

4.2 The Theory of Beckermann and Kuijlaars I

Firstly, we will slightly modify the problem at hand. Instead of considering the minimizer χ_m of (LS), we consider the following problem:

$$\widehat{\chi}_m := \arg\min_{p \in \mathcal{P}_m^\infty} \max_{\lambda \in \Lambda(A)} |p(\lambda)|. \tag{4.2}$$

Here we replaced the L^2-norm by the uniform norm on $\Lambda(A)$. Note that $z^m - \widehat{\chi}_m$ is the uniform best approximating polynomial of degree $m-1$ to the function z^m on the discrete set $\Lambda(A)$.

We assume that moving from (LS) to (4.2) is indeed a slight modification in the sense that the roots of χ_m and $\widehat{\chi}_m$ are close to each other. Since χ_m is small on the eigenvalues of A, it should have some of its roots close to them. We expect also $\widehat{\chi}_m$ to be small on the eigenvalues and to have its roots close to them. In other words: since the roots of χ_m and $\widehat{\chi}_m$ are related to the spectrum of A, we hope that the roots of χ_m and $\widehat{\chi}_m$ are somehow related to each other:

$$\boxed{\text{roots of } \chi_m \ (= \text{Ritz}(m) \text{ values}) \leftsquigarrow \Lambda(A) \leftsquigarrow \text{roots of } \widehat{\chi}_m.}$$

In some cases this connection may fail and the roots of χ_m and $\widehat{\chi}_m$ are distributed completely differently. For example, one may think of minimizing polynomials that are very small on a particular eigenvalue (or a cluster of eigenvalues) but do not have any root in this region. In this case, the spectrum of A cannot explain any connection between the roots of χ_m and $\widehat{\chi}_m$.

In order to apply potential theoretic tools, it is necessary to replace the discrete set $\Lambda(A)$ by a 'larger' compact set $\Omega \supset \Lambda(A)$.[1] Recall that the same procedure was followed for the convergence investigation of the CG method, where we replaced the spectrum of A by an interval. We define

$$\widetilde{\chi}_m := \arg\min_{p \in \mathcal{P}_m^\infty} \max_{z \in \Omega} |p(z)|. \tag{4.3}$$

Example 4.5. *Again we consider a diagonal matrix A with 40 equidistant eigenvalues in $[-1, 1]$ and a vector $\boldsymbol{b} := [1, 1, \ldots, 1]^T$ of length 40. In Figure 4.5(a) we plot the polynomial $\widehat{\chi}_{29}$, which is the monic polynomial of degree 29 that minimizes $|\widehat{\chi}_{29}|$ on $\Lambda(A)$. For the computation of this polynomial we used an algorithm described in Stiefel [30]. Note that $p(z) := z^{29} - \widehat{\chi}_{29}(z) \in \mathcal{P}_{28}$ is the best uniform approximating polynomial to $f(z) = z^{29}$ on $\Lambda(A)$. The error $|f - p| = |\widehat{\chi}_{29}|$ takes on its maximum value in 30 points of $\Lambda(A)$, which is necessary and sufficient for the optimality of p by Theorem 3.4. In view of Figure 4.5(c), the roots of χ_{29} and $\widehat{\chi}_{29}$ are 'similarly distributed'.*


4minmax-discr

In Figure 4.5(b) we plot the polynomial $\widetilde{\chi}_{29}$ for the same matrix A. We have chosen $\Omega = [\lambda_{\min}, \lambda_{\max}] = [-1, 1]$ so that $\widetilde{\chi}_{29}$ equals the normalized Chebyshev polynomial $\widetilde{T}_{29}$. It can be observed that the roots of $\widetilde{\chi}_{29}$ are very closely spaced at the outer regions of the interval Ω, see also Figure 4.5(c). Apart from the extreme eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$, the polynomial $\widetilde{\chi}_{29}$ is independent of the spectrum of A. Hence it is easy to construct Hermitian matrices where the roots of $\widetilde{\chi}_{29}$ even fail to have the interlacing property satisfied by the Ritz(29) values: just concentrate most of the interior eigenvalues of A around the center of the interval. The idea to overcome this problem is to introduce a constraint that forces the interlacing property to be fulfilled.


4minmax-cont


4compare

[1] Later we will assume that the capacity of Ω is nonzero since this is sufficient for the existence of an unique equilibrium measure for Ω.

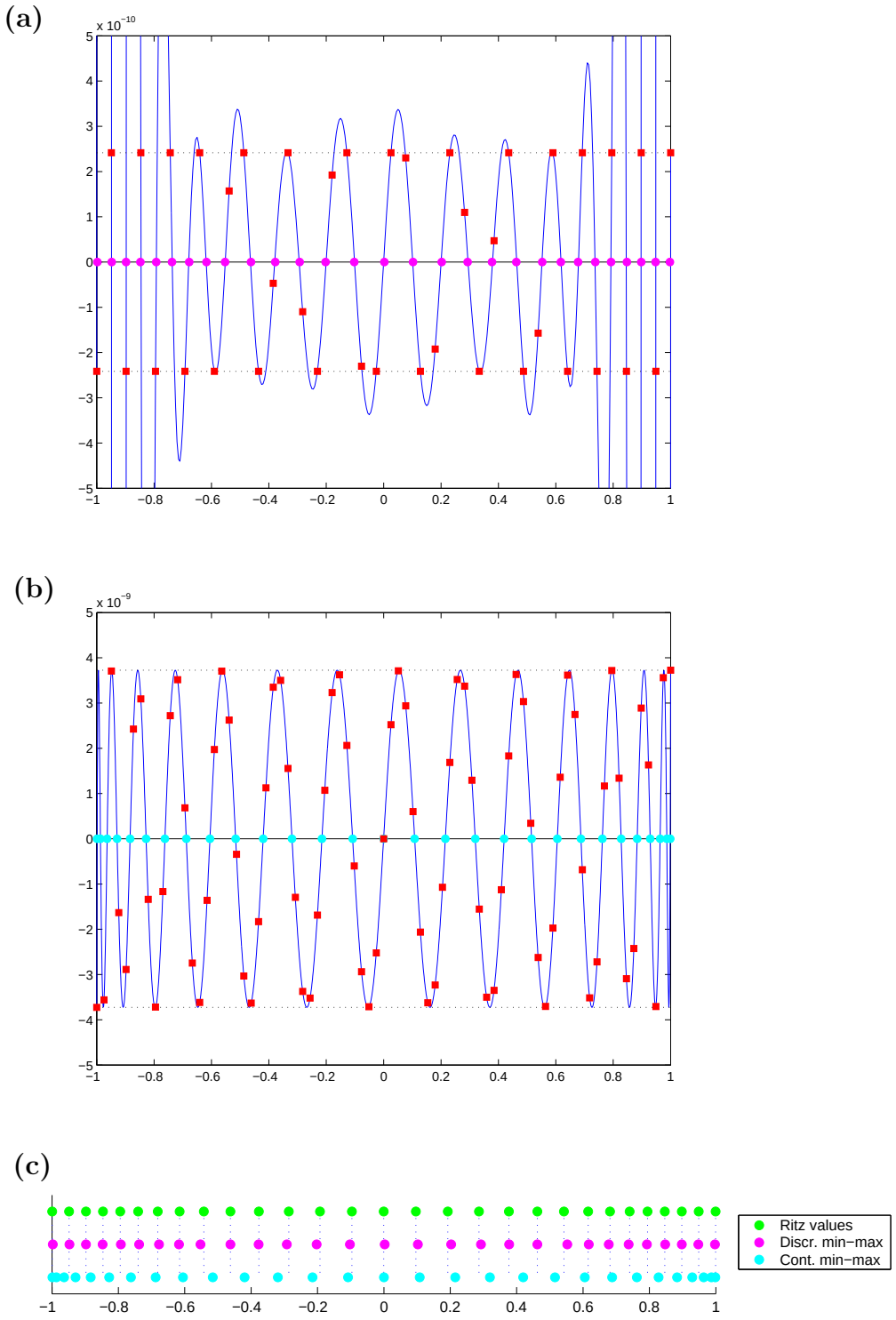


Figure 4.5:
Illustration to Example 4.5. (a) The blue line is the graph of $\widehat{\chi}_{29}(z)$. The red squares indicate the value of $\widehat{\chi}_{29}$ on the 40 equidistant eigenvalues of A. The magenta dots mark the roots of $\widehat{\chi}_{29}$. (b) The graph of $\widetilde{\chi}_{29}(z)$. The cyan dots mark the roots of $\widetilde{\chi}_{29}$. (c) The roots of the Ritz(29) polynomial χ_{29} (green), the polynomial $\widehat{\chi}_{29}$ (magenta) and $\widetilde{\chi}_{29}$ (cyan) in comparison.

4.3 Potential Theoretic Tools

On the next pages we introduce some potential theoretic tools at an introductory level as it is sufficient for our purposes. Here we follow the article [21] by Levin and Saad. For a more detailed exposition of this wide subject we refer to Ransford [25].

The *logarithmic potential* due to a unit charge placed at the point ζ in the complex plane is

$$u_\zeta(z) := -\log|z-\zeta|,$$

where we set

$$-\log 0 := +\infty,$$

so that $u_\zeta : \mathbb{C} \to \mathbb{R}\cup\{+\infty\}$. Due to the *superposition principle of electrostatics* (see Shadowitz [28]), the logarithmic potential caused by m particles $\zeta_1, \zeta_2, \ldots, \zeta_m \in \mathbb{C}$ each of charge $1/m$ is

$$\frac{1}{m}\left(u_{\zeta_1} + u_{\zeta_2} + \cdots + u_{\zeta_m}\right)(z) = -\frac{1}{m}\log|z-\zeta_1||z-\zeta_2|\cdots|z-\zeta_m|.$$

More generally, let the charges be distributed according to a measure $\mu \in \mathcal{M}(\Omega)$, where $\mathcal{M}(\Omega)$ denotes the set of *Borel probability measures* supported on Ω, i.e., their support is contained in a compact set $\Omega \subset \mathbb{C}$. The *support* $\operatorname{supp}(\mu)$ *of* μ is (in our case) the smallest closed subset of $\mathbb{C}$ with measure 1. We define the *(logarithmic) potential* U^μ *associated with* μ by

$$U^\mu(z) := -\int \log|z-\zeta|\, d\mu(\zeta). \tag{4.4}$$

This function $U^\mu : \mathbb{C} \to \mathbb{R}\cup\{+\infty\}$ is *harmonic* outside $\operatorname{supp}(\mu)$. Moreover, U^μ is *superharmonic* in $\mathbb{C}$, which means that it is lower semi-continuous and satisfies a *local supermean inequality*, i.e., for each $z \in \mathbb{C}$ there exists $\rho > 0$ such that

$$U^\mu(z) \geq \frac{1}{2\pi}\int_0^{2\pi} U^\mu(z + re^{it})\, dt \qquad \text{for all } 0 \leq r < \rho.$$

The *energy of* μ is defined by

$$I(\mu) := \int U^\mu(z)\, d\mu(z).$$

The energy is either finite or takes the value $+\infty$. We consider the following energy minimization problem

$$V(\Omega) := \inf\{I(\mu) : \mu \in \mathcal{M}(\Omega)\}$$

and define the *(logarithmic) capacity of* Ω by

$$\operatorname{cap}(\Omega) := \exp(-V(\Omega)).$$

If $V(\Omega) = +\infty$, we set $\operatorname{cap}(\Omega) := 0$. Such sets are called *polar*. Polar sets are thin in the sense that the 'area' (planar Lebesgue measure) and the 'length' (one-dimensional Hausdorff measure) of any polar set are equal to zero. For example, any countable set has capacity zero.

We assume from now on that

$$\operatorname{cap}(\Omega) > 0.$$

In this case the *Theorem of Frostman* asserts that there exists an unique measure $\mu_\Omega \in \mathcal{M}(\Omega)$ such that $I(\mu_\Omega) = V(\Omega)$. For a proof of this result see Ransford [25]. The measure μ_Ω is called *equilibrium measure for* Ω. Some important properties of μ_Ω, U^{μ_Ω} and the capacity will be required in what follows.


4equilibrium


4movie

- Let $\partial_\infty\Omega$ denote the *outer boundary of* Ω (that is, the boundary of the unbounded component of $\mathbb{C} \setminus \Omega$). Then μ_Ω is supported on $\partial_\infty\Omega$, i.e.,

  $$\operatorname{supp}(\mu_\Omega) \subseteq \partial_\infty\Omega.$$

- Since $\mathcal{M}(\partial_\infty\Omega) \subseteq \mathcal{M}(\Omega)$ and μ_Ω is unique, the above inclusion implies

  $$\operatorname{cap}(\Omega) = \operatorname{cap}(\partial_\infty\Omega).$$

- For all $z \in \mathbb{C}$ there holds

  $$U^{\mu_\Omega}(z) \leq V(\Omega)$$

  with equality holding *quasi-everywhere* on Ω; that is, except possibly for a set of capacity zero. There holds

  $$U^{\mu_\Omega}(z) = V(\Omega) = -\log \operatorname{cap}(\Omega) \qquad \text{quasi-everywhere on } \Omega.$$

  On the other hand, if the potential of some $\mu \in \mathcal{M}(\Omega)$ is constant quasi-everywhere on Ω and $I(\mu) < +\infty$, then $\mu = \mu_\Omega$.

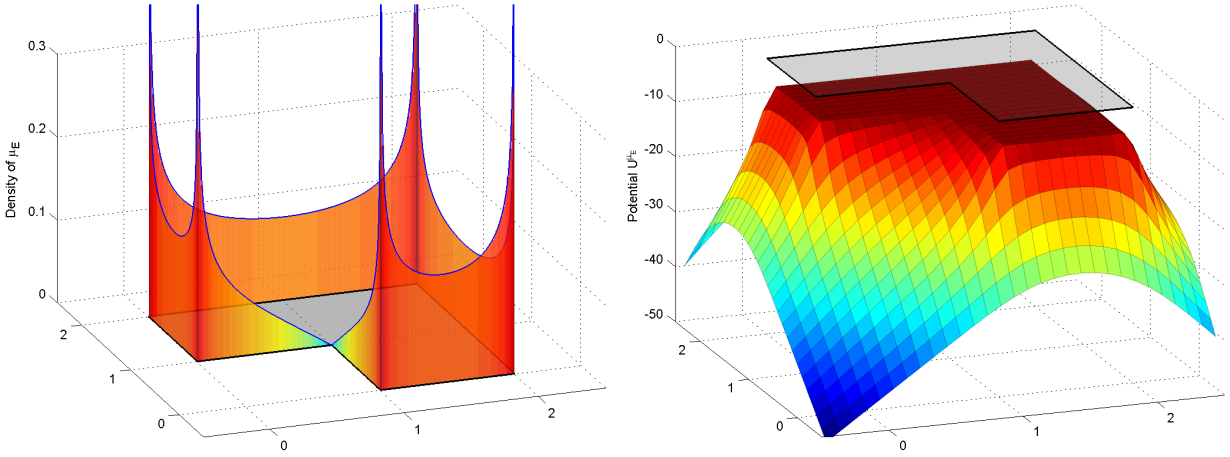


Figure 4.6: Density and potential U^{μ_Ω} of the equilibrium measure μ_Ω for the L-shaped domain.

4.4 The Theory of Beckermann and Kuijlaars II

In this section the potential theory comes in. Given a monic polynomial $p(z) = (z - z_1)(z - z_2)\cdots(z - z_m)$ of degree m, we define the associated *normalized zero counting measure*

$$\nu_p = \frac{1}{m}\sum_{i=1}^{m} \delta_{z_i},$$

where δ_z denotes the *unit Dirac measure at* $z \in \mathbb{C}$, i.e.,

$$\delta_z(S) = \begin{cases} 1, & z \in S; \\ 0, & z \notin S; \end{cases} \qquad \text{for all } S \subseteq \mathbb{C}.$$

The discrete measure ν_p assigns mass $1/m$ to each root of p and roots are counted by multiplicity. From the definition (4.4) of the potential it is easy to see that

$$\begin{aligned} U^{\nu_p}(z) &= -\int \log|z-\zeta|\, d\nu_p(\zeta) \\ &= -\frac{1}{m}\sum_{i=1}^{m} \log|z - z_i| \\ &= -\frac{1}{m}\log\prod_{i=1}^{m}|z - z_i| \\ &= -\frac{1}{m}\log|p(z)|, \end{aligned}$$

so that there is an immediate connection between the absolute value of a monic polynomial and the potential of its associated normalized zero counting measure.

Consequently, the minimizing problem (4.3) can be reformulated equivalently as

$$\text{Maximize } \min_{z\in\Omega} U^{\mu} \text{ among measures } \mu \text{ of the form } \mu = \frac{1}{m}\sum_{i=1}^{m} \delta_{z_i}.$$

The mass points of the maximizing measure of this problem are then exactly the roots of the minimizing polynomial $\widetilde{\chi}_m$ from problem (4.3). The advantage of measures in contrast to roots of polynomials is that measures need not be discrete. In the limit $m \to +\infty$ we ignore the fact that μ is discrete. This leads to

$$\text{Maximize } \min_{z\in\Omega} U^{\mu} \text{ among all probability measures } \mu. \tag{4.5}$$

A maximizer of this problem does not always exist. To overcome this problem we assume that $\operatorname{cap}(\Omega) > 0$. Then by Frostman's Theorem there exists a unique equilibrium measure μ_Ω for Ω. Moreover, we assume that Ω is *regular* in the sense that

$$U^{\mu_\Omega}(z) \leq V(\Omega) \tag{4.6}$$

with equality holding for all $z \in \Omega$ (not only quasi-everywhere!).

Now we show that μ_Ω is a maximizer of (4.5): Assume this is not the case and that there exists a probability measure $\tilde{\mu}$ such that $U^{\tilde{\mu}} > V(\Omega)$ on Ω. Integrating this inequality with respect to μ_Ω,

$$V(\Omega) < \int U^{\tilde{\mu}}(z)\, d\mu_\Omega(z) = -\iint \log|z-\zeta|\, d\tilde{\mu}(\zeta) d\mu_\Omega(z),$$

and, using Fubini's Theorem to interchange the order of integration, we obtain $V(\Omega) < \int U^{\mu_\Omega}\, d\tilde{\mu}$. But this is a contradiction to (4.6) because $\tilde{\mu}$ is a probability measure.

If Ω has nonzero capacity but is not regular, we may still determine the equilibrium measure μ_Ω for Ω as the minimizer of the *energy problem*

$$\boxed{\text{Minimize the energy } \int U^{\mu}(z)\, d\mu(z) \text{ among } \mu \in \mathcal{M}(\Omega).} \tag{4.7}$$

Problems arise because the maximizer of (4.5) may not be uniquely determined. However, in many important cases it is unique (and therefore coincides with μ_Ω), for example if Ω is a union of curves with simply connected complement (see Kuijlaars [18, p. 13]). This includes also real intervals, which is sufficient for the analysis of Hermitian matrices.

The Hermitian Case

Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian matrix with (real) eigenvalues

$$\lambda_{\min} := \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N =: \lambda_{\max}.$$

Moreover, we denote the Ritz(m) values of A by

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_m\,.$$

Actually, due to Theorem 2.31, the inequalities in the last string are even strict.

We define the following *normalized counting measures* (see Kuijlaars [18, p. 10])

$$\sigma_N(S) := \frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda_i}(S) \quad \text{and} \quad \mu_{N,m}(S) := \frac{1}{m}\sum_{i=1}^{m}\delta_{\theta_i}(S).$$

We restrict δ_z to the Borel sets of $\mathbb{R}$, so that σ_N and $\mu_{N,m}$ are Borel measures.

Definition 4.6. *The* distribution function $F_\nu : \mathbb{R} \to \mathbb{R}_+$ *of a positive Borel measure* ν *with support in* $\mathbb{R}$ *is*

$$F_\nu(x) := \nu\,(\,(-\infty, x]\,)\,.$$

Given two such measures ν_1, ν_2. *We say* ν_1 *is* smaller or equal (in the sense of inequality of measures) than ν_2, *if*

$$F_{\nu_1}(x) \leq F_{\nu_2}(x) \qquad \textit{for all } x \in \mathbb{R}.$$

We denote this by $\nu_1 \leq \nu_2$.

Note that

$$\sigma_N(S) = \frac{\#\{\text{ eigenvalues in } S\,\}}{N} \quad \text{and} \quad \mu_{N,m}(S) = \frac{\#\{\,\text{Ritz}(m) \text{ values in } S\,\}}{m}.$$

By taking into account Corollary 2.32, we obtain

$$m\,\mu_{N,m}(\,(-\infty, x]\,) \leq N\sigma_N(\,(-\infty, x]\,) \qquad \text{for all } x \in \mathbb{R},$$

or, equivalently, in our new notation

$$\boxed{m\,\mu_{N,m} \leq N\sigma_N.} \tag{GP1}$$

(GP1) is a *guiding principle* for the distribution of the Ritz(m) values. It serves as an upper *constraint* on the number of Ritz values. The intuition behind it is that it would be a 'waste of resources' to have more Ritz values than eigenvalues in some region (see Kuijlaars [18, p. 15]).

Now we let $N \to +\infty$. This is reasonable because the matrix A is of very large dimension. One may think of a sequence of matrices $(A_N)_{N\geq 1}$, where the eigenvalue distributions of the matrices A_N tend (in a weak sense, see below) to some limit distribution σ supported on Ω. If, for example, the matrices A_N arise from the dicretization of a partial differential equation (PDE), where N is determined by the discretization mesh size, the eigenvalues will follow some distribution which is related to the properties of the PDE (Kuijlaars [18, pp. 6].) To formalize this, we introduce the notion of *weak*-convergence*.

Definition 4.7. *By $C(\Omega)$ we denote the set of continuous functions $f : \Omega \to \mathbb{R}$. A sequence $(\nu_n)_{n\geq 1}$ in $\mathcal{M}(\Omega)$ is* weak*-convergent to $\nu \in \mathcal{M}(\Omega)$*, if*

$$\int f \, d\nu_n \to \int f \, d\nu \qquad \text{for all } f \in C(\Omega).$$

We write $\nu_n \xrightarrow{*} \nu$.

Assumption III. *We assume that $\sigma_N \xrightarrow{*} \sigma$, where σ is some Borel probability measure with* $\operatorname{supp}(\sigma) = \Omega$.

For a given real interval K it can be observed, that for a particular index m the number of Ritz(m) values in K is directly proportional to N if the eigenvalues of the matrices A_N are samples (in the sense of statistics) of one fixed distribution. We want to give an example to make this clear. See also Kuijlaars [18, pp. 8–9].

Example 4.8. *Let $A_N \in \mathbb{C}^{N\times N}$ have N equidistant eigenvalues in $[0, 1]$, namely*

$$0, 1/(N-1), 2/(N-1), \ldots, (N-2)/(N-1), 1.$$

By this we have

$$\sigma_N(S) = \frac{1}{N} \sum_{i=1}^{N} \delta_{(i-1)/(N-1)}(S).$$

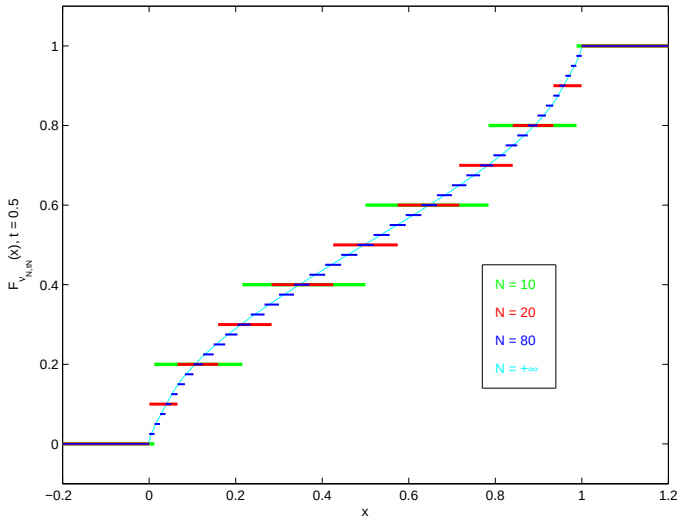
For $N \to +\infty$ this measure tends, in the weak-sense, to the uniform distribution on the interval $[0, 1]$, i.e., $\sigma_N \xrightarrow{*} \sigma$, where $\sigma(S)$ denotes the Lebesgue measure of $S \cap [0, 1]$.*

Now we consider a fixed interval K. It can be observed, that if we increase N – say, by a factor $\alpha > 1$ – then $\mu_{\alpha N,\alpha m}(K) \approx \mu_{N,m}(K)$. In the limit (weak-sense), the measure $\mu_{N,m}$ depends on the ratio $m/N =: t \in (0,1)$ only. Thus, we may denote the limit measure by μ_t and make the following assumption:*

Assumption IV.

$$\mu_{N,tN} \xrightarrow{*} \mu_t \qquad as \;\; N \to +\infty$$

for some Borel probability measure μ_t.



4ratio-t

Figure 4.7:
Here we plot the distribution functions of $\mu_{N,tN}$ for $N = 10, 20, 80, +\infty$ and $t = 0.5$. The matrix $A_N \in \mathbb{C}^{N\times N}$ has N equidistant eigenvalues in the interval $[0,1]$ and $\boldsymbol{b} = [1,1,\ldots,1]^T$.

We note that (GP1) can now be rewritten as

$$\frac{m}{N}\mu_{N,m} \leq \sigma_N. \tag{GP1'}$$

By taking $N \to +\infty$ and setting $t := m/N$ we obtain

$$\boxed{t\mu_t \leq \sigma.} \tag{4.8}$$

Finally, together with (4.7), we are led to the *constrained energy problem*

$$\boxed{\mu_t \text{ minimizes } \int U^{\mu}(z)\, d\mu(z) \text{ among } \mu \in \mathcal{M}(\Omega) \text{ with } t\mu \leq \sigma.} \tag{CEP}$$

We say μ_t is the *constrained equilibrium measure to* (CEP). It can be shown that μ_t exists and is uniquely determined if σ has a continuous and real-valued logarithmic potential U^σ. This is a smoothness condition on σ. It will be satisfied, for example, if σ has a density with respect to Lebesgue measure which is bounded, or which has power-type singularities near end-points. On the other hand, σ cannot have mass points, since the logarithmic potential would be infinite there.

It is clear that if $t\mu_\Omega \leq \sigma$, then the equilibrium measure μ_Ω also solves (CEP), i.e., $\mu_t = \mu_\Omega$. As for the equilibrium measure there is also a characterizing property in terms of the potential. We define the set

$$F_t := \operatorname{supp}(\sigma - t\mu_t),$$

which we call the *free region*. This is the set where the upper constraint (4.8) is not active. Under the above smoothness condition on σ, one can show that U^{μ_t} is equal to a constant C_t on F_t and smaller than or equal to C_t everywhere else. Moreover, the only probability measure μ that satisfies $0 \leq t\mu \leq \sigma$ and whose potential U^μ is constant on F_t and smaller everywhere else, is μ_t (see Helsen, Van Barel [15, p. 3]). On the complement

$$S_t := \Omega \setminus F_t$$

the measures σ and $t\mu_t$ agree. This is what we call the *saturated region*.

In our context, (CEP) has the following interpretation: Let $t \in (0, 1)$ and $A \in \mathbb{C}^{N\times N}$ be a Hermitian matrix (N large). Moreover, let Ω be a reasonable approximation to the spectrum of A, for example $\Omega := [\lambda_{\min}, \lambda_{\max}]$. Then the Ritz($tN$) values of A are distributed according to μ_t, i.e., they are distributed according to the equilibrium measure for Ω under the constraint that there should not be more Ritz values than eigenvalues in some region. In the saturated region S_t this constraint is active, i.e., the number of Ritz values is limited by (4.8). This is the region where the Ritz(tN) values have converged (see Kuijlaars [18, p. 16–17]).

4.5 Examples to the Constrained Energy Problem

The determination of μ_t is a non-trivial problem in general. For special eigenvalue distributions σ it can be calculated explicitly (cf. Kuijlaars [18, p. 17–19]). In other

cases, some properties of μ_t can be derived without being able to obtain an explicit solution. Hence it would be interesting to obtain a numerical approximation. This is what we do on the next pages. To solve the constrained energy problem we have used an algorithm by Helsen and Van Barel [15]. This algorithm computes μ_t if the constraint σ is given by a piecewise linear density function $f^\sigma(x)$, $x \in \mathbb{R}$.


4cep1D-1

Example 4.9. *Let $A \in \mathbb{R}^{N\times N}$ be a diagonal matrix with $N = 100$ equidistant eigenvalues in the interval $[0, 1]$ and $\boldsymbol{b} = [1, 1, \ldots, 1]^T$. Figure 4.8(a) shows, for which index m an eigenvalue $\lambda \in \Lambda(A)$ is found by a Ritz value. We plot a black dot at some eigenvalue λ and index $\tilde{m}$, if for every $m \geq \tilde{m}$ there exists a Ritz(m) value within distance $\texttt{tol}:=10^{-3}$ to λ. In (b) one can see the density of the measure σ (blue), which corresponds to the evenly distributed eigenvalues of A, and the associated potential U^σ (red). We computed the constrained equilibrium measure μ_t for $t = 0.4$ and $t = 0.7$. The resulting densities of μ_t are shown in (c) and (d) as a green line, as well as the associated potentials U^{μ_t} (red). The saturated regions S_t and the free regions F_t are shown below (light green and magenta). Note that S_t is exactly the region where $t\mu_t$ and σ agree (in our case, the density of μ_t is constant $1/t$ there), whereas the potential U^{μ_t} is constant on the free region F_t. In (e) we plot again the converged Ritz(m) values (black dots), and add the saturated regions S_t (light green), where $t = m/N$. We observe, that the saturated regions indicate very well the region where the eigenvalues of A have been found by Ritz values.*


4cep1D-2

Example 4.10. *The matrix $A \in \mathbb{R}^{N\times N}$ has $N = 100$ eigenvalues, which are distributed in the following way: 50 equidistant eigenvalues lie in the interval $[0, 0.5]$ and 50 equidistant eigenvalues in $[0.73, 1]$. See also Figure 4.9(a) and (b). In (c) and (d) we plot the density and potential of the constrained equilibrium measure μ_t for $t = 0.4$ and $t = 0.7$. The saturated regions S_t are a good indicator for the region where the eigenvalues of A have been found by Ritz values, see (e).*


4cep1D-3

Example 4.11. *This example corresponds to Figure 4.10. Here the matrix A is of dimension $N = 200$. Ten of its eigenvalues are clustered at the right end of the interval $[0, 1]$ and the others follow a random normal distribution with mean $1/3$ and standard deviation $1/9$. As above, the saturated regions S_t are a good predictor for the region where the Ritz values have converged. For example, we may read off from (e) that all the eigenvalues in the cluster at the right end of the interval are found for $t = 0.2$, i.e., after 40 iterations of the Lanczos process.*

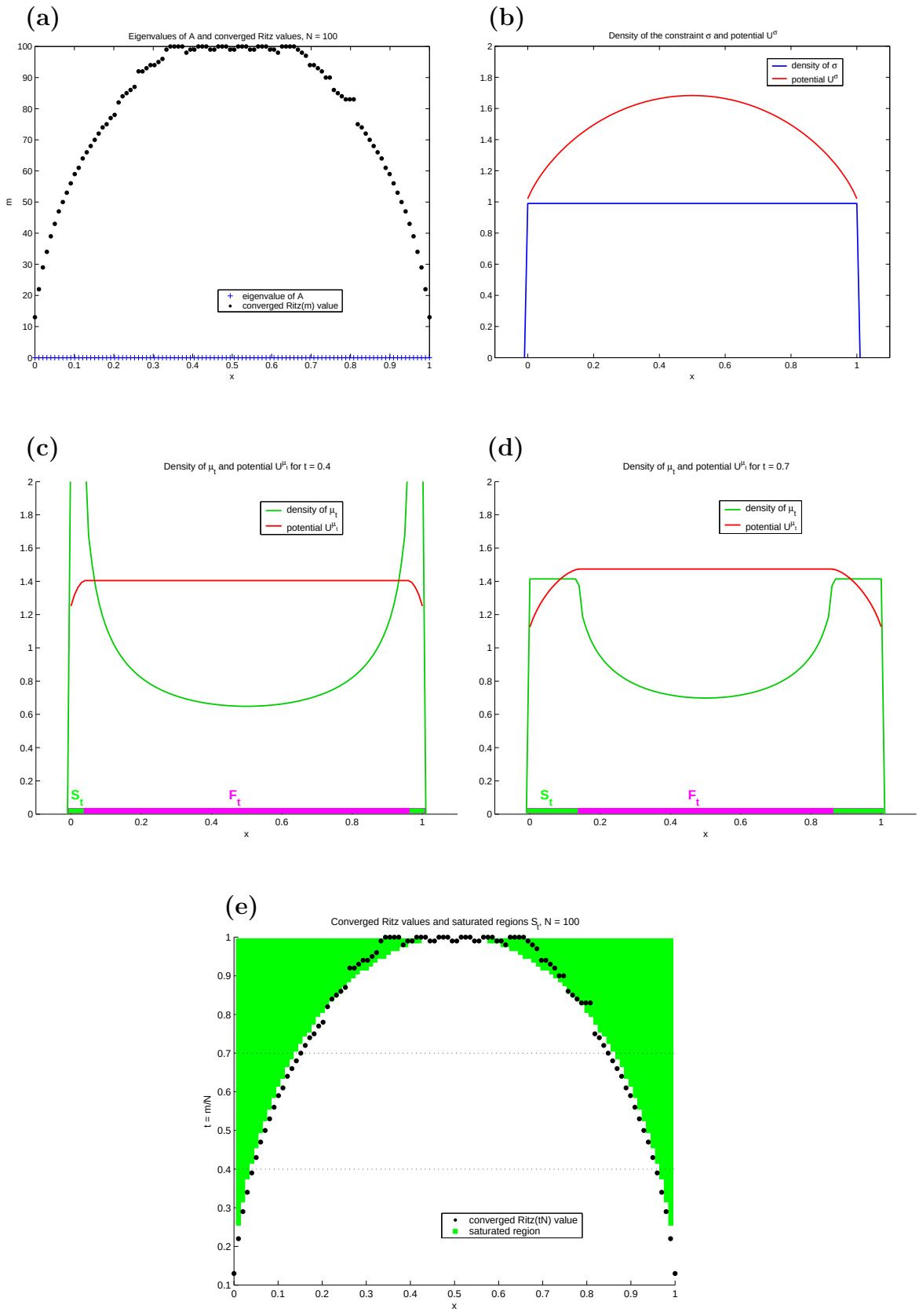


Figure 4.8: Convergence of Ritz values.

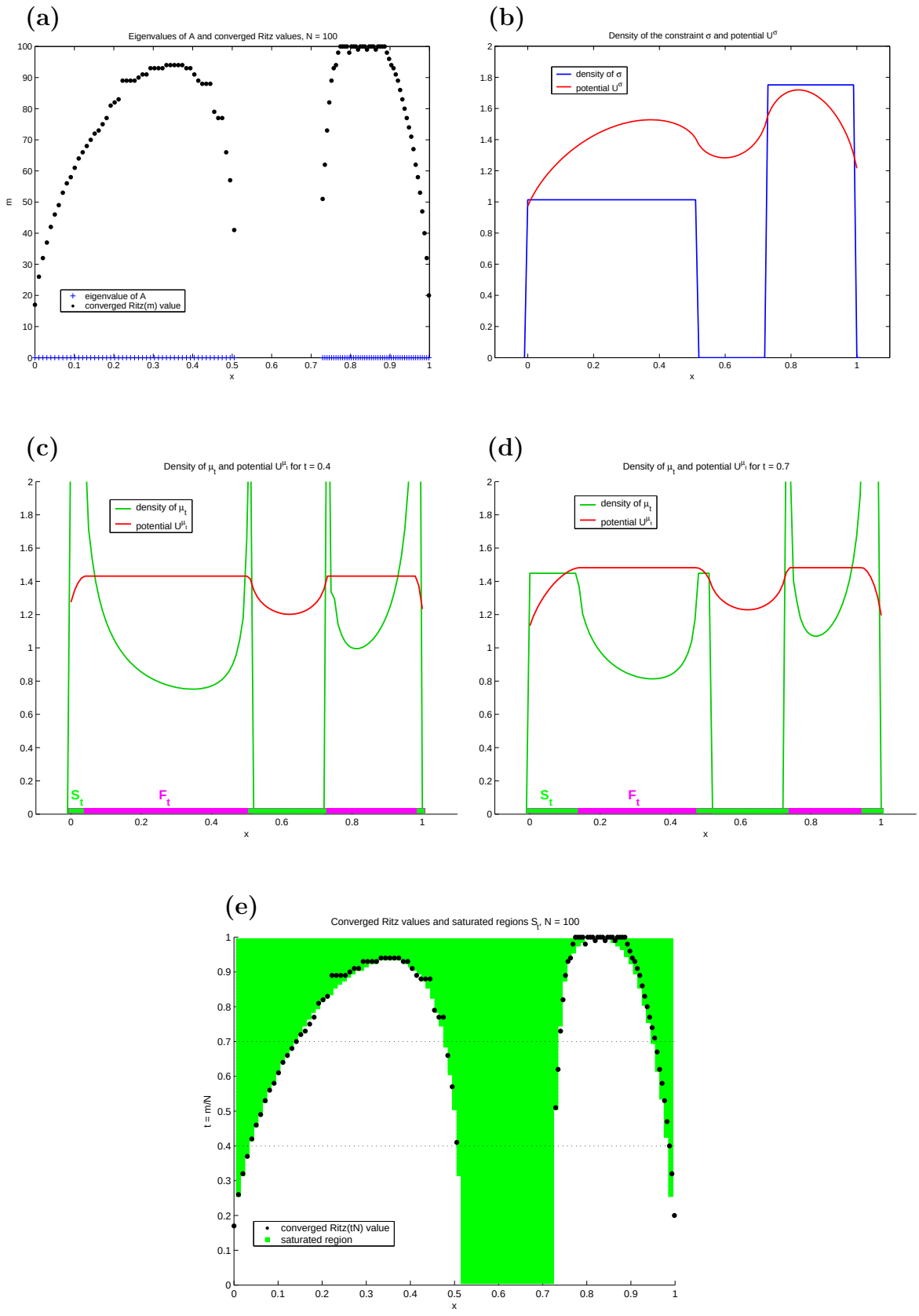


Figure 4.9: Convergence of Ritz values.

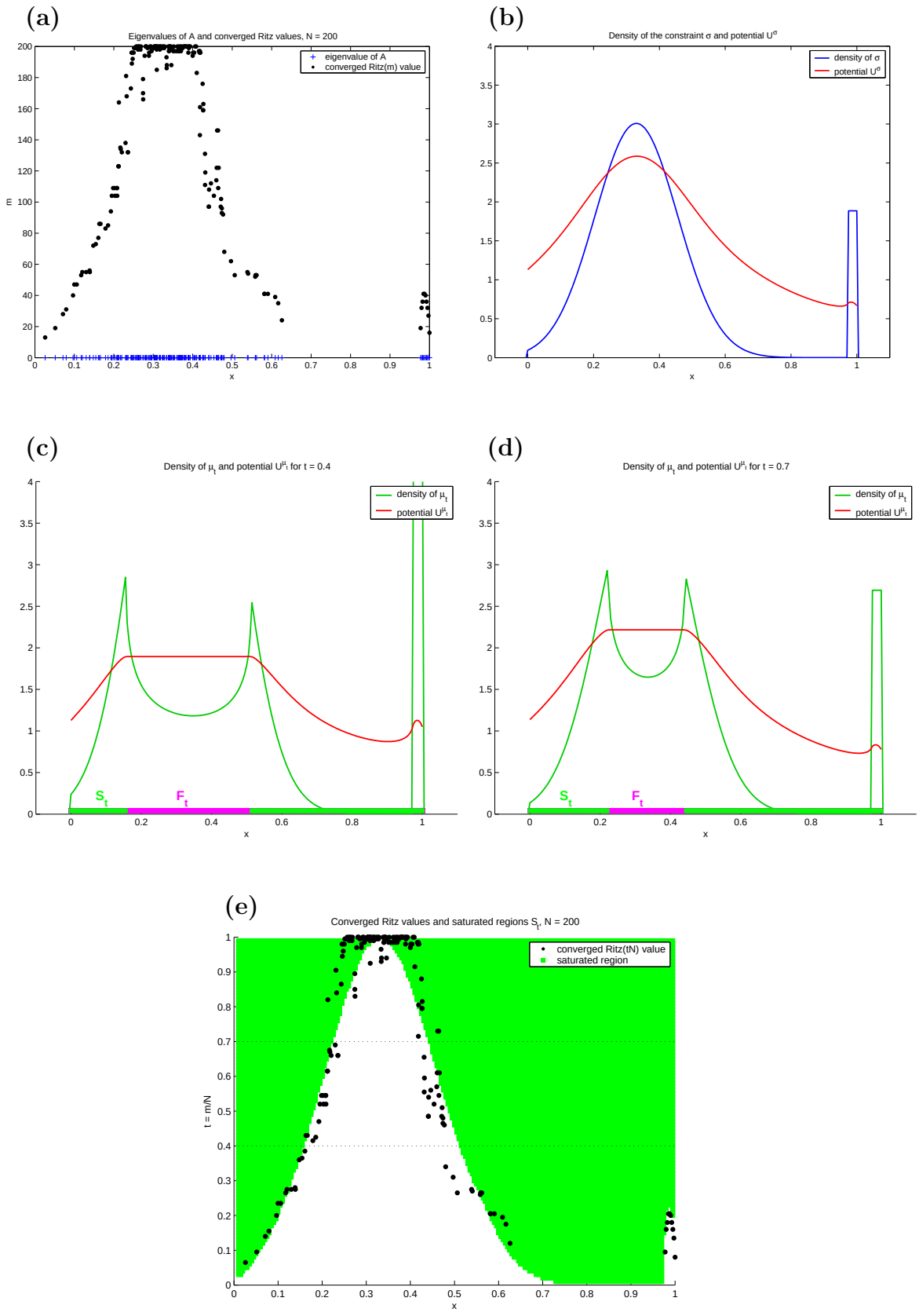


Figure 4.10: Convergence of Ritz values.

4.6 Fast Numerical Evaluation of Potentials in 2D

Let the measure $\mu \in \mathcal{M}(\Omega)$, $\Omega \subset \mathbb{R}^2$ have a *density function* $f^\mu(x,y)$, i.e.,

$$\mu(S) = \iint_S f^\mu(x,y)\,dxdy$$

for all Borel sets $S \subseteq \mathbb{R}^2$.

The associated logarithmic potential U^μ at a point (s,t) is

$$U^\mu(s,t) = -\iint \log \|[s,t]^T - [x,y]^T\| f^\mu(x,y)\,dxdy.$$

We assume that f^μ is of the form

$$f^\mu(x,y) = \alpha_1 f_1(x,y) + \alpha_2 f_2(x,y) + \cdots + \alpha_n f_n(x,y)$$

and evaluate U^μ at given node points $(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)$. Then for $i = 1,2,\ldots,m$ we have

$$\begin{aligned}\underbrace{U^\mu(x_i,y_i)}_{=:u_i} &= -\iint \log \|[x_i,y_i]^T - [x,y]^T\| f^\mu(x,y)\,dxdy \\ &= \sum_{j=1}^{n} \underbrace{\left(-\iint \log \|[x_i,y_i]^T - [x,y]^T\| f_j(x,y)\,dxdy\right)}_{=:P_{i,j}} \alpha_j,\end{aligned}$$

and in matrix-vector notation

$$P\boldsymbol{\alpha} = \boldsymbol{u}.$$

The j-th column of the matrix $P \in \mathbb{R}^{m\times n}$ contains the values of the potential U^μ evaluated at the points (x_i,y_i) for $i = 1,2,\ldots,m$, where μ has the density function f_j. The straightforward determination of P is very time-consuming since it involves the numerical evaluation of mn integrals. However, we can overcome this problem by choosing the densities f_j such that the generated potentials are invariant under certain rotations. In what follows, we present a method that allows to assemble P for arbitrary domains without evaluating an integral if only the potential of a reference density function f is given in sufficiently many *Gaussian points* (that are the points (x_i,y_i) where both x_i and y_i are integers).

We start with the node points (x_i,y_i) $(i = 1,2,\ldots,m)$ which are placed equidistantly according to a square grid with mesh size h. On this grid we introduce a

regular alternating triangulation, as shown in Figure 4.11 for the L-shaped domain.[2] We may divide the node points into two classes.

- **Type A:** a point (x_j, y_j) is adjacent to 8 triangles (see the red dots in Figure 4.11), the square region filled by these triangles is denoted by Q_j,
- **Type B:** a point (x_j, y_j) is adjacent to exactly 4 triangles (see the green dots), the (rotated) square region filled by these triangles is denoted by $\widetilde{Q}_j$.

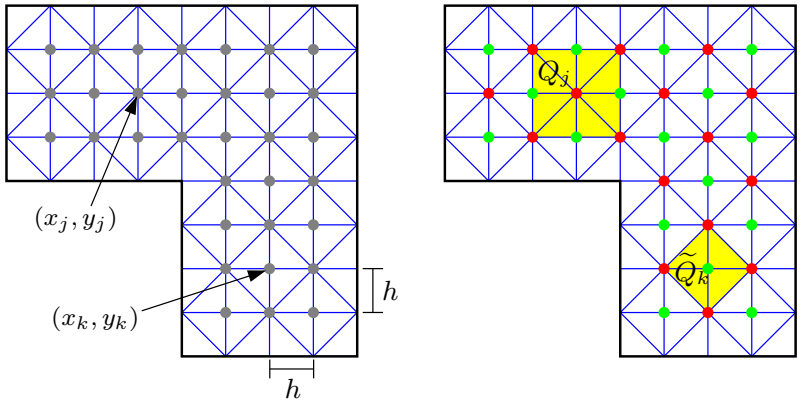


Figure 4.11: Constructing a piecewise linear density on the L-shaped domain.

Note that we do not have to place the nodes (x_i, y_i) inside the domain Ω. However, in the sequel we shall set the density outside Ω and on its boundary equal to zero, so that outlying points will become superfluous. We define m piecewise linear density functions f_i in the following way: Given a *reference density function* $f(u, v)$ with support $R := [-1, 1] \times [-1, 1]$,

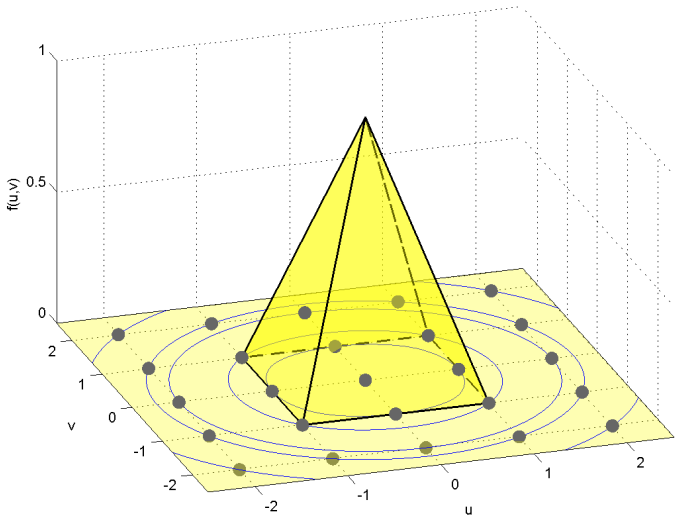
$$f(u,v) = \begin{cases} \min\{1-|u|, 1-|v|\}, & (u,v) \in R; \\ 0, & \text{otherwise.} \end{cases}$$

The image of f describes a square pyramid centered at 0 with height 1 and side-length 2 (see Figure 4.12). By $U(u, v)$ we denote the potential generated by this density.

If (x_j, y_j) is a point of Type A, we define

$$\Phi_j(u,v) := \begin{bmatrix} h & 0 \\ 0 & h \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} x_j \\ y_j \end{bmatrix}.$$

[2] Actually, this triangulation is not needed in the implementation of this method.



4refdensity

Figure 4.12:
Reference density $f(u, v)$ (yellow), certain level curves of the corresponding potential $U(u, v)$ (blue lines) and the Gaussian points (grey dots).

Φ_j is a bijective linear map on $\mathbb{R}^2$ that maps R onto Q. By Φ_j^{-1} we denote its inverse map. We set

$$f_j(x, y) := f(\Phi_j^{-1}(x, y)).$$

The image of f_i describes a square pyramid centered at z_j with height 1 and side-length $2h$.

If (x_j, y_j) is a point of Type B, we set

$$\widetilde{\Phi}_j(u, v) := \frac{1}{2} \begin{bmatrix} h & h \\ -h & h \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} x_j \\ y_j \end{bmatrix}$$

and

$$f_j(x, y) := f(\widetilde{\Phi}_j^{-1}(x, y)).$$

The image of f_i describes a square pyramid centered at (x_j, y_j) with height 1 and side-length $\sqrt{2}h$ that is rotated by the angle $-\pi/2$.

By the definition of Φ_j and $\widetilde{\Phi}_j$ it is obvious that each node point (x_i, y_i) is the image of a Gaussian point. In other words,

$$\Phi_j^{-1}(x_i, y_i) \quad \text{and} \quad \widetilde{\Phi}_j^{-1}(x_i, y_i) \quad \text{are Gaussian points for } i, j = 1, 2, \ldots, m.$$

Now we evaluate the potential U^j generated by a density f_j at the node (x_i, y_i). First we assume that (x_i, y_i) is of Type A.

$$\begin{aligned}
U^j(x_i,y_i) &= -\iint_{Q_j} \log\left\|[x_i,y_i]^T-[x,y]^T\right\| f_j(x,y)\,dxdy \\
&= -\iint_{\Phi_j(R)} \log\left\|[x_i,y_i]^T-[x,y]^T\right\| f_j(x,y)\,dxdy \\
&= -\iint_{R} \log\left\|[x_i,y_i]^T-\Phi_j(u,v)^T\right\| f_j(\Phi_j(u,v))|\det(\Phi_j'(u,v))|\,dudv \\
&= -\iint_{R} \log\left\|[x_i-x_j,y_i-y_j]^T-h(u,v)^T\right\| f(u,v)|h^2|\,dudv \\
&= -\iint_{R} \log\left(h\left\|[x_i-x_j,y_i-y_j]^T/h-(u,v)^T\right\|\right) f(u,v)h^2\,dudv \\
&= -\iint_{R} \left(\log h+\log\left\|\Phi_j^{-1}(x_i,y_i)-(u,v)^T\right\|\right) f(u,v)h^2\,dudv.
\end{aligned}$$

In the third line we used the *change of variables rule* of integral calculus and in the fourth line we applied the definition of Φ_j and f_j. Finally, we have

$$U^j(x_i,y_i) = -\frac{4}{3}h^2\log h + h^2U\left(\Phi_j^{-1}(x_i,y_i)\right).$$

Note that $U^j(x_i, y_i) = P_{i,j}$ and its evaluation involves the computation of U in a Gaussian point. Another remarkable fact is that U is independent of h, the discretization nodes (x_i, y_i) and the domain Ω.

If (x_i, y_i) is of Type B, a similar formula can be derived. There holds

$$P_{i,j} = U^j(x_i,y_i) = -\frac{4}{6}h^2\log\frac{h}{\sqrt{2}} + \frac{1}{2}h^2U\left(\widetilde{\Phi}_j^{-1}(x_i,y_i)\right).$$

As above, we have to evaluate U in a Gaussian point and U is independent of h, the discretization nodes (x_i, y_i) and the domain Ω. Therefore we will compute the values of U just once in sufficiently many Gaussian points and store them in a matrix S, which can be reused for each computation.

The assembly of the matrix P reduces to a look-up of some values in S and rescaling them using the above formulae depending on the node-type. Moreover, we

can exploit the 8-fold symmetry of the potential U:

$$\begin{aligned} U(x,y) &= U(x,-y) = U(-x,y) = U(-x,-y) \\ = U(-y,-x) &= U(y,-x) = U(-y,x) = U(y,x). \end{aligned}$$

Hence we will only evaluate U in sufficiently many Gaussian points in the first octant of the plane and store S as a triangular matrix:

$$S := \begin{bmatrix} U(0,0) & & & \\ U(1,0) & U(1,1) & & \\ U(2,0) & U(2,1) & U(2,2) & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The size of S that is necessary to carry out a computation depends on the mesh size h and the diameter of Ω.

The complete implementation of this method can be found on the CD-ROM.

4computeP

4.7 Outlook

With the above method for the fast evaluation of potentials associated with piecewise linear densities it is possible to solve the constrained energy problem (CEP) in the complex plane very efficiently. An extension of the 1D-algorithm by Helsen, Van Barel [15] has been obtained recently by M. Eiermann and the author. Even though for an arbitrary normal matrix A there is no interlacing property of the Ritz values (which ultimately led to the constraint in the Hermitian case), it still seems reasonable to assume that the number of Ritz values in a half-plane

$$S(x + \mathbb{i}y) := (-\infty, x] \times \mathbb{i}(-\infty, y]$$

does not exceed the number of eigenvalues of A in S for all $z = x + \mathbb{i}y$. This would be a 'waste of resources' otherwise. Under this assumption (which itself warrants closer investigation), the theory on the convergence of Ritz values by Beckermann and Kuijlaars (see [1, 18, 19]) may be carried over to non-Hermitian normal matrices. First numerical tests have been performed and are quite promising.


4cep2D-1

Example 4.12. *In Figure 4.13 we consider a normal matrix $A \in \mathbb{C}^{N\times N}$, where $N = 300$. The eigenvalues of A are evenly distributed in the L-shaped domain, i.e., the measure σ has constant density there; see the blue sheet in (a). In (b) we plot the associated potential U^σ. In (c) and (d) we show the density of the constrained equilibrium measure μ_t and the associated potential U^{μ_t} for $t = 0.8$. The saturated region S_t (light green) is exactly the region where $t\mu_t$ and σ agree (in our case, the density of μ_t is constant there), whereas the potential U^{μ_t} is constant on the free region F_t (magenta). In (e) we show the converged Ritz(m) values (colored disks), underlaid with the saturated regions S_t for $t = 0.2, 0.4, 0.6, 0.8, (N-1)/N$. For example, all eigenvalues that are found for $m \leq 0.2N$ (black disks) should belong to the black region ($t = 0.2$). In view of (e), the saturated regions are a good indicator for the converged Ritz values.*


4cep2D-2

Example 4.13. *In Figure 4.14 we consider a normal matrix $A \in \mathbb{C}^{300\times 300}$. The eigenvalues all lie in the domain $\Omega = [0, 2]^2 \setminus [0.5, 1.5]^2 \subset \mathbb{R}^2$. The distance of the eigenvalues in the upper left diagonal part is scaled by a factor $\sqrt{2}$ in comparison to the lower right part. Therefore the density of σ differs by a factor 2 between both diagonal parts.*


4cep2D-3

Example 4.14. *We consider a normal matrix $A \in \mathbb{C}^{300\times 300}$. The eigenvalues lie inside a circle, which is centered at -1 and has radius 1, and a triangle with vertices $\{-i, 2-i, 2+i\}$. The density of σ differs by a factor $4/3$ between both components, cf. Figure 4.15. Again, the regions of converged Ritz values are well predicted by the saturated regions.*

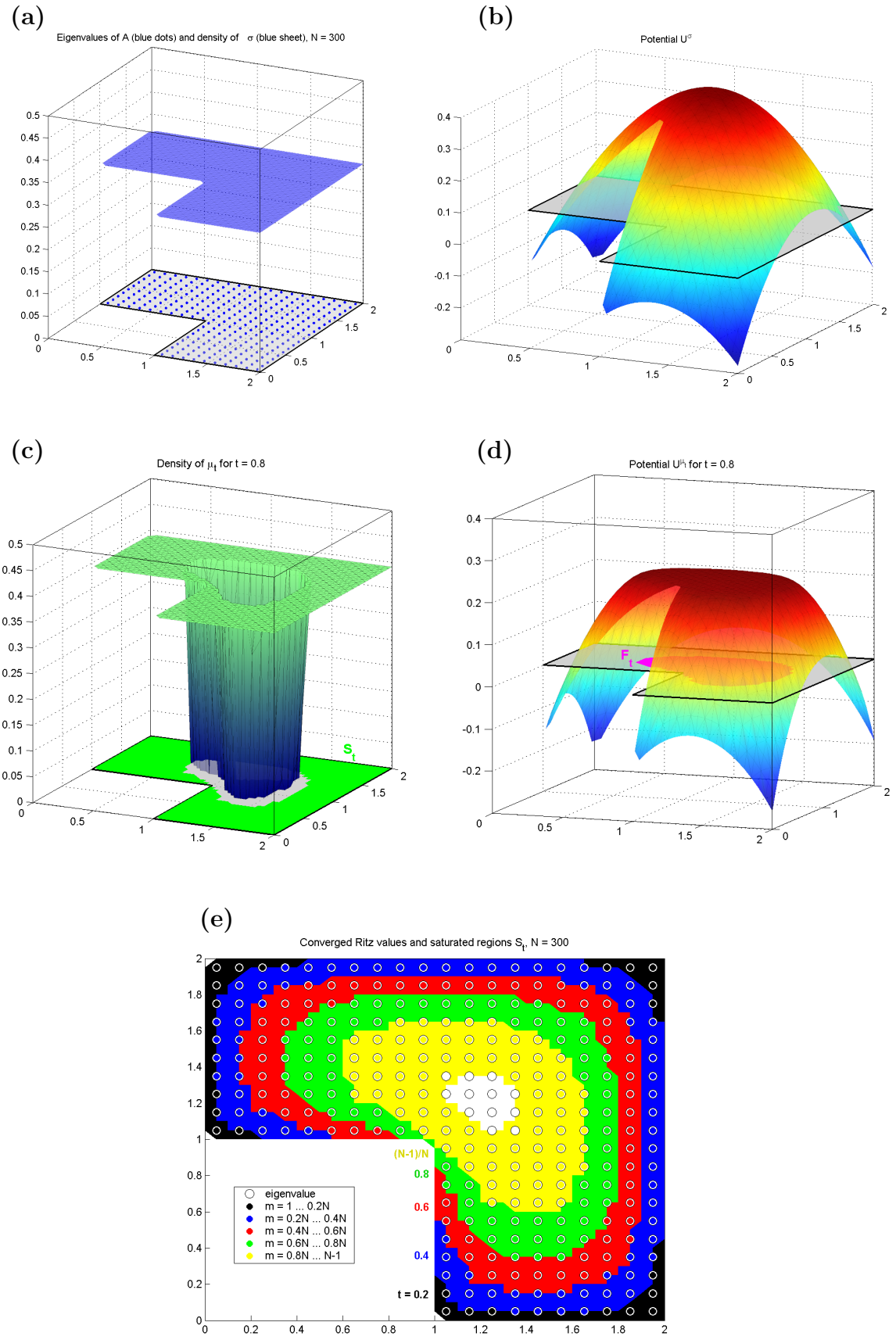


Figure 4.13: Convergence of Ritz values.

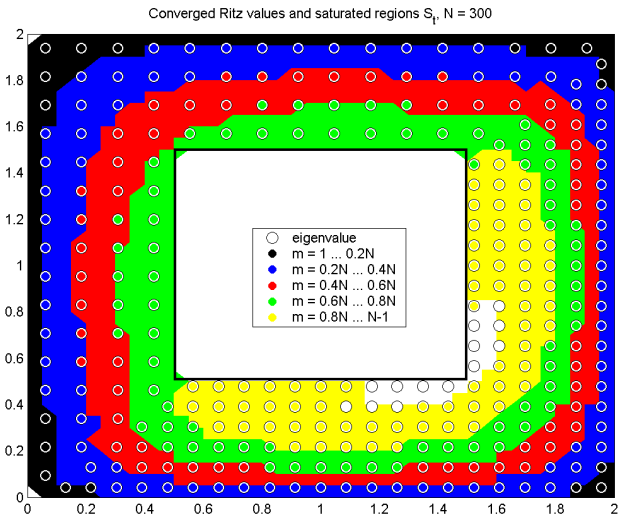


Figure 4.14: Convergence of Ritz values.

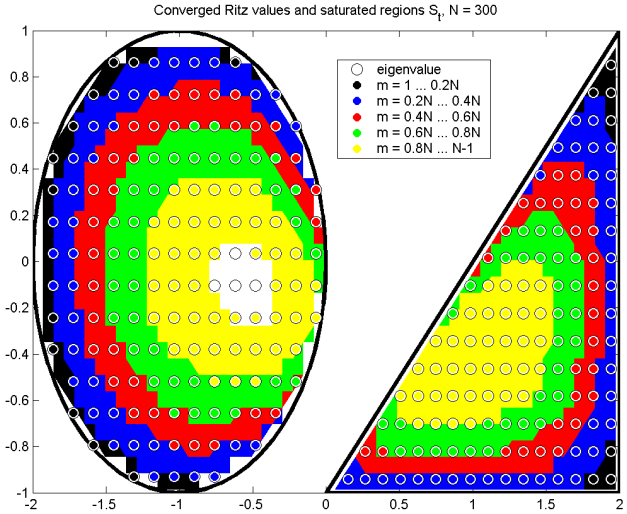


Figure 4.15: Convergence of Ritz values.

File List

- Folder **PDF**
  - file `diploma.pdf` – this document,
  - file `cyprus1.pdf` – presentation slides '*Matrix Functions and their Approximation using Krylov Subspaces*',
  - file `cyprus2.pdf` – presentation slides '*Matrix Functions and their Approximation by Polynomial Methods*',
- Folder **TEX**
  - subfolder **Diploma** – LaTeX-files of this document (main file: `main.tex`),
  - subfolder **Cyprus1** – LaTeX-files of `cyprus1.pdf` (main file: `main.tex`),
  - subfolder **Cyprus2** – LaTeX-files of `cyprus2.pdf` (main file: `main.tex`),
- Folder **FIG**
  - file `figX-Y.pdf` – *Figure X.Y* as `.pdf`-file for better view,
  - file `figX-Y.png` – *Figure X.Y* as `.png`-file for better view,
  - ...
- Folder **MAT**
  - file `guirun.m` – graphical user interface to run the examples,
  - subfolder **FILES** – `.m`-Files for direct access (main file: `rundemo.m`).

Notation

Symbol	Description	Page
I	identity matrix	5
O	null matrix	5
$\boldsymbol{\xi}_m$	m-th unit coordinate vector	5
$\mathrm{toep}(\underline{\cdot}, \cdot, \cdots)$	Toeplitz matrix, main diagonal is underlined	5
$\mathrm{diag}(\cdots)$	(block-)diagonal matrix	5
$\Lambda(A)$	spectrum of A	9
ψ_A	minimal polynomial of A	9
d	$d = \deg(\psi_A)$	9
d_λ, c_λ	multiplicity of the root λ in ψ_A, $\psi_{A,\boldsymbol{b}}$	9, 25
χ_A	characteristic polynomial of A	10
$p_{f,A}$	interpolates f at the roots of ψ_A	11
$C_{\lambda,i}$	components of A	15
γ, Γ	path, (Jordan) curve	17
wind_z	winding number around $z \in \mathbb{C}$	17
$\mathrm{int}(\Gamma)$	interior of the curve Γ	17
$\mathbb{i}$	imaginary unit, $\mathbb{i}^2 = -1$	18
$R_\zeta(A)$	resolvent of A to $\zeta \in \mathbb{C}$	18
ω, ω_m	nodal polynomial (of degree m)	19, 42
$p_{f,\omega}$, $q_{f,m}$	interpolates f at the roots of ω, ω_m	19, 42
$\lvert\!\lvert\!\lvert \cdot \rvert\!\rvert\!\rvert$	some arbitrary norm	20
$\varrho(A)$	spectral radius of A	21
$\mathcal{K}_m(A, \boldsymbol{b}) = \mathcal{K}_m$	m-th Krylov subspace	23
L	$L = \deg(\psi_{A,\boldsymbol{b}})$	24
$\psi_{A,\boldsymbol{b}}$	minimal polynomial of $\boldsymbol{b}$ with respect to A	25

Symbol	**Description**	**Page**
$C_{\boldsymbol{\alpha}}$	companion matrix	28
$\|\cdot\|$	2-norm of a matrix or a vector	29
H_m	Hessenberg matrix produced by Algorithm 2.14	29
χ_m	Ritz(m) polynomial	34
$p_{f,m}$	interpolates f at the roots of χ_m	35
$\mathcal{P}_m^\infty$	monic polynomials of degree m	34
$\mathcal{P}_m^0$	residual polynomials of degree m	40
$\boldsymbol{r}_m$	m-th residual vector	40
$\boldsymbol{e}_m$	m-th error vector	41
$\|\cdot\|_A$	A-norm	42
$\|\cdot\|_\Omega$	uniform norm on Ω	45
$C(\Omega)$	continuous functions on Ω	45
$f_m \Rightarrow f$	f_m converges uniformly to f	45
T_m	Chebyshev polynomial of degree m	47
$\widetilde{T}_m(x)$	normalized Chebyshev polynomial of degree m	48
T_m^K	shifted Chebyshev polynomial of degree m	50
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$	$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^H \boldsymbol{x}$, scalar product	68
U^μ	logarithmic potential associated with μ	77
$\mathcal{M}(\Omega)$	set of Borel probability measures on Ω	77
$I(\mu)$	energy of μ	77
$\mathrm{cap}(\Omega)$	logarithmic capacity of Ω	78
μ_Ω	equilibrium measure for Ω	78
δ_z	unit Dirac measure at $z \in \mathbb{C}$	79
σ_N, $\mu_{N,m}$	normalized counting measures	81
σ	measure, associated with the eigenvalues	82
μ_t	constrained equilibrium measure	84
F_t	free region	84
S_t	saturated region	84

References

[1] B. Beckermann, A. B. J. Kuijlaars. *Superlinear convergence of conjugate gradients.* SIAM J. Numer. Anal. 39 (2001), pp. 301-329.

[2] P. J. Davis. *Interpolation and approximation.* Dover Publications, New York, 1975.

[3] T. A. Driscoll. *A* MATLAB *toolbox for Schwarz-Christoffel mapping.* ACM Trans. Math. Softw. 22 (1996), pp. 168–186. Software available at `http://amath.colorado.edu/appm/faculty/tad/research/sc.html`.

[4] M. Eiermann, O. Ernst. *A collection of useful facts.* Unpublished notes, Freiberg, 2003.

[5] M. Eiermann, O. Ernst, O. Schneider. *Analysis of acceleration strategies for restarted minimal residual methods.* Journal of Computational and Applied Mathematics 123 (2000), pp. 261–292.

[6] M. Eiermann, O. Ernst. *Geometric aspects in the theory of Krylov subspace methods.* Acta Numerica 10 (2001), pp. 251–312.

[7] M. Eiermann. *Krylov-Verfahren und Potentialtheorie.* Seminar slides, Freiberg, 2004.

[8] M. Eiermann, O. Ernst. *Matrix functions and their numerical approximation.* Seminar slides, Freiberg, 2004.

[9] T. Ericsson. *Computing functions of matrices using Krylov subspace methods.* Report Numerical Analysis Group, Göteborg, 1990.

[10] D. Gaier. *Lectures on complex approximation.* Birkhäuser, Boston, 1987.

[11] F. R. Gantmacher. *Numerische Mathematik 1.* Springer, Berlin, 1994.

[12] G. H. Golub, C. F. Van Loan. *Matrix computations, 2nd ed.* The John Hopkins University Press, Baltimore, 1989.

[13] A. Greenbaum. *Iterative methods for solving linear systems.* SIAM, Philadelphia, 1997.

[14] A. Hadjidimos, N. Stylianopoulos. *Optimal semi-iterative methods for complex SOR with results from potential theory.* Numerische Mathematik 103 (2006), pp. 591–610.

[15] S. Helsen, M. Van Barel. *A numerical solution of the constrained energy problem.* submitted to Elsevier Science, Leuven, 2004.

[16] P. Henrici. *Applied and computational complex analysis.* Wiley and Sons, New York, 1988.

[17] M. R. Hestenes, E. L. Stiefel. *Methods of conjugate gradients for solving linear systems.* J. Res. Nat. Bur. Standards 49 (1952), pp. 409–436.

[18] A. B. J. Kuijlaars. *Convergence analysis of Krylov subspace iterations with methods from potential theory.* SIAM Review 48 (2006), pp. 3–40.

[19] A. B. J. Kuijlaars. *Which eigenvalues are found by the Lanczos method?* SIAM J. Matrix Anal. Appl. 22 (2000), pp. 306–321.

[20] P. Lancaster, M. Tismenetsky. *The theory of matrices, 2nd ed.* Academic Press, Boston, 1985.

[21] A. L. Levin, E. B. Saff. *Potential theoretic tools in polynomial and rational approximation.* Harmonic Analysis and Rational Approximation 327 (2006), to appear.

[22] C. D. Meyer. *Matrix analysis and applied linear algebra.* SIAM, Philadelphia, 2000.

[23] E. Mina-Diaz, E. B. Saff, N. Stylianopoulos. *Zero distribution for polynomials orthogonal with weights over certain planar regions.* CMFT 5 (2005), no. 1.

[24] P. Novati. *A polynomial method based on Fejér points for the computation of functions of unsymmetric matrices.* Appl. Numer. Math. 44 (2002), pp. 201–224.

[25] T. Ransford. *Potential theory in the complex plane.* Cambridge University Press, Cambridge, 1995.

[26] Y. Saad. *Iterative methods for sparse linear systems.* PWS Publishing Company, Boston, 1996.

[27] Y. Saad. *Numerical methods for large eigenvalue problems.* Manchester University Press, Manchester, 1991.

[28] A. Shadowitz. *The electromagnetic field.* Dover Publications, New York, 1975.

[29] B. Singer, S. Spilerman. *The representation of social processes by Markov models.* Amer. J. Sociology 8 (1976), pp. 1–54.

[30] E. L. Stiefel. *Numerical methods of Tchebycheff approximation.* Proc. Sympos. Math. Res. Center Madison (1959), pp. 217–232.

[31] J. Stoer. *Numerische Mathematik 1.* Springer, Berlin, 1994.

[32] J.L. Walsh. *Interpolation and approximation by rational functions in the complex domain.* American Mathematical Society, Providence, 1960.

Index

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich diese Diplomarbeit selbständig
und unter ausschließlicher Benutzung der angegebenen Quellen
und Hilfsmittel angefertigt habe.

Freiberg, den 19. Juni 2006 Stefan Güttel

Declaration of Academic Honesty

I hereby declare to have written this Diploma Thesis on my own,
having used only the listed resources and tools.

Freiberg, 19th of June 2006 Stefan Güttel