Bounds for the decay in matrix functions and its exploitation in matrix computations

Dissertation

Bergische Universität Wuppertal
Fakultät für Mathematik und Naturwissenschaften

eingereicht von
Claudia Schimmel, M. Ed.
zur Erlangung des Grades eines Doktors der Naturwissenschaften

Betreut durch Prof. Dr. Andreas Frommer

Wuppertal, 01.10.2019

The PhD thesis can be quoted as follows:

urn:nbn:de:hbz:468-20200121-115408-5
[http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20200121-115408-5]

DOI: 10.25926/sygx-rm09
[https://doi.org/10.25926/sygx-rm09]

Abstract

It is known that in many functions of large and sparse matrices the entries exhibit a rapid decay such that most of them are very small in magnitude. It is possible to give upper bounds for the magnitude of the entries which capture this decay without actually computing the matrix function, called *decay bounds*. In this thesis we derive new decay bounds for special types of matrices and functions, including the inverse as an important special case. In addition, based on the results for the inverse, we formulate decay bounds for Cauchy–Stieltjes functions of certain classes of matrices. The superiority of the new bounds compared to bounds from the literature is shown and illustrated in numerical experiments.

Furthermore, we discuss the practical relevance of this decay property. In particular, the decay in matrix functions reveals the existence of a sparse approximation. We exploit the decay property in order to compute sparse approximations and the trace of matrix functions, where decay bounds can be used for an error analysis. The resulting methods are compared to previous approaches from the literature and numerical examples show the effectiveness of the proposed methods.

Contents

Chapter 1

Introduction

For a given scalar function f and a square matrix A it is possible to define a matrix $f(A)$ such that it provides a useful generalization of the scalar function $f(z)$ with argument $z \in \mathbb{C}$. Those matrices $f(A)$ are called *matrix functions* or *functions of matrices*. One of the most important and popular special case is given by the matrix inverse A^{-1} generated by the scalar function $f(z) = z^{-1}$ which arises in many areas of numerical linear algebra, e.g., in the solution of systems of linear equations. Furthermore, lots of other matrix functions play an important role in plenty of applications. The matrix function $f(A) = \exp(A)$ is frequently used for the numerical solution of time-dependent differential equations or the analysis of dynamical systems [40]. In addition, it is an important tool for the analysis of networks, as well as the resolvent, generated by the scalar function $f(z) = (\alpha - z)^{-1}$ with $\alpha \in \mathbb{C}$ [29, 30, 31]. The matrix sign function $f(A) = \operatorname{sign}(A)$ has important applications in control theory [40, 82] and lattice quantum chromodynamics [14, 74, 103]. Inverse fractional powers $f(A) = A^{-\alpha}$ with $\alpha \in (0, 1)$ are strongly related to the matrix sign function and arise in generalized eigenvalue problems [75, Section 15.10] and fractional differential equations [16]. An abundance of further applications of matrix functions can be found in [53, Chapter 2].

In this thesis we assume that the considered matrices A are large and sparse. Since we only deal with square matrices, we denote with an n-dimensional matrix (or a matrix of dimension n) a matrix of size $n \times n$. While matrices of dimension n in general have up to n^2 nonzero entries, we say that an n-dimensional matrix is *sparse* if it only contains $\mathcal{O}(n)$ nonzero entries. In a less stricter definition, a matrix A is sparse if it is gainful to use a special storage format which only considers the nonzero entries of A and if it is profitable to make use of the sparsity of A in matrix computations like the computation of Av for a given vector v. If a matrix is not sparse, we call it dense. Sparse matrices naturally arise in plenty

of applications like in the discretization of continuous problems. For example, a D-dimensional partial differential equation of second order can be discretized by using a $(2D+1)$-point stencil on a uniform lattice, where every central point couples with its two nearest neighbors in each dimension of the lattice. This results in a discretized problem including a matrix with only $2D+1$ nonzero entries per row. Furthermore, lots of important types of graphs or networks result in sparse adjacency matrices, e.g., if the graph is planar or if the maximal degree of the graph is bounded. The consideration of large and sparse matrices leads to two crucial problems in the computation of matrix functions:

1. The computation of $f(A)$ has in general a cost of $\mathcal{O}(n^3)$ for an n-dimensional matrix A. Thus, the computation of $f(A)$ is very expensive for large matrices.

2. The matrix $f(A)$ is dense (apart from certain special cases) such that it is not possible to store $f(A)$ for large dimensional problems, while this was possible for the sparse matrix A.

For large and sparse matrices A most of the past and current research deals with the problem of computing $f(A)v$ for a given vector v without computing $f(A)$ explicitly, see, e.g, [36, 40, 53, 54]. This is motivated by the fact that the computation of $f(A)$ is way to costly and that in many applications only a vector $f(A)v$ is needed instead of of the whole matrix $f(A)$. For example, the solution of a system of linear equations $Ax = b$ is given by the vector $A^{-1}b$ and we do not explicitly require the matrix A^{-1}. On the other hand, even if we concede the high computational cost of computing $f(A)$, the storage problem still remains which makes the computation of $f(A)$ impossible. The goal of this thesis is to fix the occurring two problems for some types of large, sparse matrices by exploiting a phenomenon which is called *decay in matrix functions*. A possible consequence of such a decay in matrix functions is that most of the entries of $f(A)$ are very small in magnitude (and therefore negligible) and that the remaining important parts in $f(A)$ are only given by $\mathcal{O}(n)$ entries. Hence, this phenomenon reveals the possibility for the computation of a precise but sparse approximation of the dense matrix $f(A)$. In this thesis we will provide the theoretical basis for this phenomenon, given by *decay bounds of matrix functions*, which solves the illustrated storage problem in some cases. In addition we propose efficient ways for the computation of $f(A)$ (or more precisely a sparse approximation of $f(A)$) for large, sparse matrices A.

This thesis is organized as follows: In order to make our work as self-contained as possible we first present some basic material in Chapter 2. This especially includes the formal definition of matrix functions and some computational aspects associated with large and sparse matrices A. The results of this thesis are then given in Chapter 3 and Chapter 4. In Chapter 3 we introduce decay bounds for matrix functions which are the theoretical basis for the results in Chapter 4. In

particular, in Chapter 3 we introduce some known results from the literature, as well as new results for special types of functions and matrices. Some of our results in Chapter 3 have already been published in [38] and [39]. In Chapter 4 we then exploit the decay in $f(A)$ for certain matrix computations associated with matrix functions as the computation of a sparse approximation of $f(A)$ with low computational cost. Furthermore, an important part of this chapter is devoted to the use of the results of Chapter 3 for the computation of the trace of matrix functions without computing $f(A)$ explicitly. In Chapter 5 we summarize the results of this thesis and give concluding remarks on possible future research topics.

Chapter 2

Review of basic material

This chapter gives an overview of basic definitions and properties required for the development of the results in this thesis. We start with fundamental definitions and properties of general matrix functions, and we subsequently discuss important aspects and problems arising in the computation of functions of *large and sparse* matrices in particular. In Section 2.2 we review the classical terminology of graph theory and present the intimate relation between graphs and matrices which will be useful throughout this thesis. In addition, this relation yields an important application of functions of matrices in network analysis. Concluding this chapter, we introduce certain types of minimal polynomials which are fundamental for the development of the results in Chapter 3.

2.1 Functions of matrices

In this section we define the matrix function $f(A) \in \mathbb{C}^{n\times n}$ for a given scalar function f and a matrix $A \in \mathbb{C}^{n\times n}$ such that $f(A)$ represents a meaningful generalization of the scalar variable $f(z)$ for $z \in \mathbb{C}$. For example, for $f(z) = z^{-1}$ and the corresponding matrix function $f(A) = A^{-1}$ we want the matrix products AA^{-1} and $A^{-1}A$ to be the identity element I in $\mathbb{C}^{n\times n}$, i.e, $f(A)$ should result in the classical matrix inverse of A. Similarly, for a square root $A^{1/2}$ of a matrix A we expect the relation $(A^{1/2})^2 = A$. In addition, if f is a polynomial or in general if f has a power series expansion $f(z) = \sum_{k=0}^{\infty} a_k(z-\alpha)^k$ for $|z-\alpha| < r$, then it seems to be natural to define

$$f(A) = \sum_{k=0}^{\infty} a_k(A-\alpha I)^k$$

under some conditions on A with respect to the set of convergence. There are several equivalent definitions for functions of matrices fulfilling these requests and which give a meaningful definition for general scalar functions f. We present three of those definitions in Section 2.1.1. They directly provide ways for the computation of the matrix $f(A)$ which are especially not feasible for large and sparse matrices A. Hence, we discuss some computational aspects in particular for functions of large and sparse matrices in Section 2.1.2.

2.1.1 Basic definitions

The following definitions and results mainly follow the presentation in [53, Chapter 1]. An overview of functions of matrices can also be found in [40, 48, 55].

A matrix $A \in \mathbb{C}^{n\times n}$ can be expressed in Jordan canonical form $A = WJW^{-1}$, with a nonsingular matrix $W \in \mathbb{C}^{n\times n}$ and a block diagonal matrix $J = \operatorname{diag}(J_1, \ldots, J_p)$, with Jordan blocks J_k of the form

$$J_k = \begin{bmatrix} \lambda_{i_k} & 1 & & \\ & \lambda_{i_k} & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_{i_k} \end{bmatrix} \in \mathbb{C}^{m_k\times m_k},$$

where λ_{i_k} is an eigenvalue of A and $m_1 + m_2 + \cdots + m_p = n$. Let $\lambda_1, \ldots, \lambda_s$, $s \leq p$ be the distinct eigenvalues of A, then n_i denotes the size of the largest Jordan block corresponding to the eigenvalue λ_i and we call n_i the index of λ_i. Now, a function f is said to be defined on the spectrum of A if the values

$$f^{(j)}(\lambda_i), \quad j = 0, \ldots, n_i - 1, \quad i = 1, \ldots, s$$

exist. With these notations we give the following definition of matrix functions.

Definition 2.1. Let $A \in \mathbb{C}^{n\times n}$ be given in Jordan canonical form $A = WJW^{-1}$ and let f be defined on the spectrum of A. Then we define

$$f(A) := Wf(J)W^{-1} = W \operatorname{diag}(f(J_1), \ldots, f(J_p))W^{-1}, \tag{2.1}$$

with

$$f(J_k) := \begin{bmatrix} f(\lambda_{i_k}) & f'(\lambda_{i_k}) & \ldots & \frac{f^{(m_k-1)}(\lambda_{i_k})}{(m_k-1)!} \\ & f(\lambda_{i_k}) & \ddots & \vdots \\ & & \ddots & f'(\lambda_{i_k}) \\ & & & f(\lambda_{i_k}) \end{bmatrix} \in \mathbb{C}^{m_k\times m_k},$$

where λ_k is the eigenvalue corresponding to the Jordan block J_k.

Definition 2.1 reveals an important spectral property of matrix functions: The eigenvalues of $f(A)$ are given by $f(\lambda_i)$ since $f(A)$ is similar to the triangular matrix $f(J)$ with diagonal entries $f(\lambda_i)$. In addition, for a diagonalizable matrix A the Jordan canonical form reduces to the eigendecomposition $A = W\Lambda W^{-1}$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, thus

$$f(A) = W f(\Lambda) W^{-1} = W \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n)) W^{-1},$$

so the eigenvectors of A coincide with those of $f(A)$ for diagonalizable matrices A. For general matrices, every eigenvector of A is also an eigenvector of $f(A)$ which can be easily seen due to the following equivalent definition of matrix functions.

Definition 2.2. Let f be defined on the spectrum of $A \in \mathbb{C}^{n\times n}$ and let the minimal polynomial of A be of degree m. Then we define $f(A) := p(A)$, where p is the unique polynomial of degree less than m that satisfies the interpolation conditions

$$p^{(j)}(\lambda_j) = f^{(j)}(\lambda_j), \quad j = 1, \ldots, n_i - 1, \quad i = 1, \ldots, s,$$

and p is called the Hermite interpolating polynomial.

With Definition 2.2 we directly see a further important property: Every matrix function is a polynomial in A and only depends on the values of f (and derivatives of f) on the spectrum of A. If f has a power series of the form

$$f(z) = \sum_{k=0}^{\infty} a_k (z - \alpha)^k$$

for $|z - \alpha| < r$, then $f(A)$ can be expressed as

$$f(A) = \sum_{k=0}^{\infty} a_k (A - \alpha I)^k$$

if $\sigma(A) \subset \{z \in \mathbb{C} : |z - \alpha| < r\}$ (see, e.g., [53, Section 4.7]). However, an interesting consequence of Definition 2.2 is that $f(A)$ is also expressible as a polynomial of degree at most $n - 1$.

Our last definition of matrix functions is based on the Cauchy integral formula for analytic functions f. In this definition we have stricter conditions on the function f compared to Definitions 2.1 and 2.2 but in contrast to these definitions it can be directly generalized to operators on infinite-dimensional spaces.

Definition 2.3. Let $A \in \mathbb{C}^{n\times n}$ be a complex matrix and let f be analytic on a region $\Omega \subseteq \mathbb{C}$ containing $\sigma(A)$. Let Γ be a closed contour in Ω that encloses $\sigma(A)$, then we define

$$f(A) := \frac{1}{2\pi \mathbf{i}} \int_\Gamma f(t)(tI - A)^{-1} \mathrm{d}t.$$

This definition based on an integral expression of f is not restricted to the Cauchy integral formula. If f has another integral expression, e.g., a Stieltjes integral expression, then $f(A)$ can be defined in the same manner as in Definition 2.3. This will be an important fact at some points throughout this thesis.

The following result reveals the equivalence of the introduces definitions.

Theorem 2.4. *Let $A \in \mathbb{C}^{n\times n}$ be a complex matrix and let f be defined on $\sigma(A)$. Then the Definitions 2.1 and 2.2 are equivalent. If f fulfills the conditions of Definition 2.3, then Definition 2.3 is equivalent to Definitions 2.1 and 2.2.*

Proof. See, e.g., [53, Theorem 1.12] and [55, Theorem 6.2.28]. □

2.1.2 Computational aspects

An obvious way to compute the matrix $f(A)$ directly follows from Definition 2.1 of Section 2.1.1. If A can be expressed as $A = WBW^{-1}$, such that $f(B)$ is easily computable, we use the relation

$$f(A) = Wf(B)W^{-1}.$$

For diagonalizable matrices we can use the eigendecomposition $A = W\Lambda W^{-1}$, where $f(\Lambda)$ is just given by $f(\Lambda) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))$. If the matrix W is ill-conditioned or for general matrices, the Schur decomposition $A = WTW^H$ can be used instead and the problem of computing $f(A)$ reduces to the problem of computing $f(T)$ for an upper triangular matrix T where the entries $[f(T)]_{ij}$ can be computed recursively [53, Theorem 4.11]. This method is in general not feasible for large matrices, due to the high computational cost for determining a suitable decomposition of A. The high computational cost for large matrices A is also apparent for other direct techniques for computing $f(A)$. Further direct methods for the computation of $f(A)$ for small matrices A can be found in [53, Section 4] and [40].

Another well known problem in the computation of matrix functions for large and sparse matrices A is the (in general) full structure of $f(A)$. This results in the problem that we are not able to store the matrix $f(A)$ while this is possible for the sparse matrix A. However, in many applications only the vector $f(A)v$ or the bilinear form $v^H f(A)v$ for a given vector $v \in \mathbb{C}^n$ is needed instead of the whole matrix $f(A)$. A familiar application is, e.g., the solution of systems of linear equations $Ax = b$, where $x = f(A)b$, with $f(z) = z^{-1}$. Similar to iterative solvers for systems of linear equations, there exist iterative methods for the computation of $f(A)v$ and $v^H f(A)v$ for general functions f. The most important iterative methods for computing those quantities belong to the class of Krylov subspace methods. We now consider the so-called Arnoldi/Lanczos approximations of $f(A)v$ and $v^H f(A)v$, based on the Arnoldi/Laczos process.

The Arnoldi process provides a matrix $V_m \in \mathbb{C}^{n\times m}$ and an upper Hessenberg matrix $H_m \in \mathbb{C}^{m\times m}$ with $H_m = V_m^H A V_m$, $m \leq n$, where the columns of V_m form an orthonormal basis of the Krylov subspace $\mathcal{K}_m(A, v) := \operatorname{span}\{v, Av, \ldots, A^{m-1}v\}$. The Arnoldi algorithm is presented in Algorithm 2.1. In line 8 of Algorithm 2.1 the case $h_{j+1,j} = 0$ is considered and it can be shown that this condition is fulfilled if and only if the Krylov subspace $\mathcal{K}_j(A, v)$ is invariant under A, i.e., if $\mathcal{K}_k(A, v) = \mathcal{K}_j(A, v)$ for $k \geq j$; see [85, Proposition 6.6].

Algorithm 2.1: Arnoldi's method.

Input: Matrix A, vector v and number of iterations m.

1 $v_1 = \frac{1}{\|v\|_2} v$
2 **for** $j = 1, \ldots, m$ **do**
3 $\quad w_j = Av_j$
4 $\quad$ **for** $i = 1, \ldots, j$ **do**
5 $\quad\quad h_{i,j} = v_i^H w_j$
6 $\quad\quad w_j = w_j - h_{i,j} v_i$
7 $\quad$ **end**
8 $\quad h_{j+1,j} = \|w_j\|_2$
9 $\quad$ **if** $h_{j+1,j} = 0$ **then**
10 $\quad\quad$ stop
11 $\quad$ **end**
12 $\quad v_{j+1} = \frac{1}{h_{j+1,j}} w_j$
13 **end**

If A is Hermitian, the Arnoldi algorithm simplifies to the Lanczos process and only the last two computed basis vectors v_j and v_{j-1} are required for computing the current vector v_{j+1}, i.e., there is a three-term recurrence relation for the basis vectors. This can be easily seen by the fact that the Hessenberg matrix $H_m = V_m^H A V_m$ is Hermitian if A is Hermitian, hence, H_m is tridiagonal and

therefore $h_{i,j} = 0$ for $i < j - 1$. Because of the tridiagonal structure of H_m in the Lanczos process this matrix is sometimes denoted as T_m in the literature. In the following we maintain the notation H_m for ease of presentation. Algorithm 2.2 presents the Lanczos process for Hermitian matrices A.

Algorithm 2.2: Lanczos method.

Input: Hermitian matrix A, vector v and number of iterations m.

$v_1 = \frac{1}{\|v\|_2} v$
$h_{1,0} = 0$
for $j = 1, \ldots, m$ **do**
  $w_j = Av_j - h_{j,j-1}v_{j-1}$
  $h_{j,j} = v_j^H w_j$
  $w_j = w_j - h_{j,j}v_j$
  $h_{j+1,j} = \|w_j\|_2$
  if $h_{j+1,j} = 0$ **then**
    stop
  end
  $v_{j+1} = \frac{1}{h_{j+1,j}} w_j$
end

With the resulting matrices $V_m \in \mathbb{C}^{n \times m}$ and $H_m \in \mathbb{C}^{m \times m}$ and by denoting with e_1 the first canonical vector of evident size, the Arnoldi/Lanczos approximations of $f(A)v$ and $v^H f(A)v$ are given by

$$f(A)v \approx V_m f(H_m) V_m^H v = \|v\|_2 V_m f(H_m) e_1 \tag{2.2}$$

and

$$v^H f(A)v \approx v^H V_m f(H_m) V_m^H v = \|v\|_2^2 e_1^T f(H_m) e_1, \tag{2.3}$$

provided that f is defined on $\sigma(H_m)$. These approximations are exact if the Krylov subspace $\mathcal{K}_m(A, v)$ is invariant under A. Such an m can be very large in general (indeed at most n which immediately follows from the Cayley-Hamilton theorem), so we hope to obtain a good approximation of these quantities for a small number of iterations m. Then the large problem of size n is reduced to a problem of size m and $f(H_m)$ can be computed by a direct method.

For the inverse, the proposed approximation of $f(A)v$ leads to the Full Orthogonalization Method (FOM) for linear systems $Ax = b$. This choice for an m-th iterate realizes the condition that the m-th residual $r_m := b - Ax_m$ is orthogonal to the Krylov subspace $\mathcal{K}_m(A, r_0)$. If A is in addition Hermitian positive definite, we have the additional property that the m-th iterate minimizes the error in the A-norm over the subspace $\mathcal{K}_m(A, r_0)$, i.e., we obtain the conjugate gradients (CG) method. All these properties and further Krylov subspace methods for solving linear systems can be found in [85].

Before we discuss some interesting properties of these approximations for general functions f, we give some remarks on the choice of the number of iterations m.

Remark 2.5. The simplest way to abort the process is to stop, when two iterates differ less than a given threshold. In this case we need to compute the matrices $f(H_k)$ for $k = 1, \ldots, m$ in every iteration step. Of course, this is not desirable but this can be done without further restrictions. For Hermitian matrices, an iterative algorithm for determining $|e_1^T f(H_k) e_1 - e_1^T f(H_{k-1}) e_1|$ in step k of the Lanczos process is described in [18] without computing $f(H_k)$ or $f(H_{k-1})$ explicitly. $\diamond$

Remark 2.6. Another and computationally better possibility is to choose the number of iterations based on error bounds for the m-th iterate of the Arnoldi/Lanczos process. There exist numerous a priori and a posteriori bounds for the error $\|f(A)v - \|v\|_2 V_m f(H_m) e_1\|_2$ for certain classes of functions and matrices (see, e.g., [37, 40] for rational functions, [54, 84] for the exponential or [88] for Stieltjes functions). Of course, these bounds can be used for the bilinear forms $v^H f(A) v$ as well and we obtain error bounds for the m-th iterate based on error bounds for the Arnoldi/Lanczos approximation of $f(A)v$ since

$$\begin{aligned} |v^H f(A)v - \|v\|_2^2 e_1^H f(H_m) e_1| &= |v^H f(A)v - \|v\|_2 v^H V_m f(H_m) e_1| \\ &\leq \|v\|_2 \, \|f(A)v - \|v\|_2 V_m f(H_m) e_1\|_2. \end{aligned} \tag{2.4}$$

However, in some cases more accurate error bounds can be derived for the approximation of bilinear forms by using the relation to Gauss quadrature, which will be discussed in the following. Error bounds based on this relation are for example given in [18] for the Lanczos approximation of $v^H f(A) v$ for Hermitian matrices A. $\diamond$

Now, we review some interesting properties of the approximations (2.2) and (2.3).

Theorem 2.7. *Let $A \in \mathbb{C}^{n \times n}$ be a complex matrix and let V_m and H_m be the matrices obtained by m steps of the Arnoldi process. Then the approximation* (2.2) *is exact for polynomials of degree at most $m - 1$, i.e.,*

$$p_k(A)v = \|v\|_2 V_m p_k(H_m) e_1, \quad k = 1, \ldots, m-1.$$

In addition,

$$\|v\|_2 V_m f(H_m) e_1 = \widetilde{p}_{m-1}(A)v,$$

where $\widetilde{p}_{m-1}$ is the unique polynomial of degree at most $m - 1$ that interpolates f on the spectrum of H_m.

Proof. See [53, Lemma 13.4 and Theorem 13.5]. $\square$

With Theorem 2.7 the exactness for polynomials of degree $m-1$ for m steps of the Arnoldi/Lanczos process directly transfers to the Arnoldi/Lanczos approximation of $v^H f(A)v$. However, we can establish better exactness results in this case, especially for Hermitian matrices based on the relation between bilinear forms $v^H f(A)v$ and Riemann-Stieltjes integrals which we will discuss in the following. For a detailed overview of the computation of bilinear forms and the connection to Riemann-Stieltjes integrals and Gauss quadrature see [47] and the references therein.

For a Hermitian matrix $A \in \mathbb{C}^{n\times n}$, we have an eigendecomposition

$$A = U\Lambda U^H,$$

where $U \in \mathbb{C}^{n\times n}$ is a unitary matrix whose columns are the normalized eigenvectors of A and $\Lambda \in \mathbb{R}^{n\times n}$ a diagonal matrix with the eigenvalues of A ordered as $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ on the diagonal. We have

$$v^H f(A)v = v^H U f(\Lambda)U^H v = u^H f(\Lambda)u = \sum_{i=1}^{n} f(\lambda_i)|u_i|^2 \tag{2.5}$$

with $u = U^H v$. Now, the last expression can be interpreted as the Riemann-Stieltjes integral

$$\int_{\lambda_1}^{\lambda_n} f(\lambda)\mathrm{d}\alpha(\lambda) \tag{2.6}$$

where the function α is given as

$$\alpha(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_1 \\ \sum_{j=1}^{i} |u_j|^2 & \text{if } \lambda_i \leq \lambda < \lambda_{i+1} \\ \sum_{j=1}^{n} |u_j|^2 & \text{if } \lambda_n \leq \lambda \end{cases}.$$

Hence, (2.5) can be approximated by applying Gauss quadrature to the integral (2.6), i.e.,

$$v^H f(A)v = \int_{\lambda_1}^{\lambda_n} f(\lambda)\mathrm{d}\alpha(\lambda) \approx \sum_{i=1}^{N} w_i f(t_i) + \sum_{j=1}^{M} v_j f(z_j), \tag{2.7}$$

where the weights $[w_i]_{i=1}^N, [v_j]_{j=1}^M$ and the nodes $[t_i]_{i=1}^N$ are unknowns, and the nodes $[z_j]_{j=1}^M$ are prescribed. For $M = 0$ we just have the Gauss rule with no prescribed nodes. The choice $M = 1$ and $z_1 = \lambda_1$ or $z_1 = \lambda_n$ leads to the Gauss-Radau rule, and for $M = 2$ with $z_1 = \lambda_1$ and $z_2 = \lambda_n$ we obtain the Gauss-Lobatto rule. Since the eigenvalues of A and therefore α are not explicitly known, we use the following relation between Gauss quadrature and the Lanczos process.

Theorem 2.8. *Let A be Hermitian and H_m be the matrix obtained by m steps of the Lanczos process with respect to A and v, where $\|v\|_2 = 1$. Then the Gauss quadrature approximation of (2.6) with $N = m$ and $M = 0$ is given by*

$$\sum_{i=1}^{N} w_i f(t_i) = e_1^T f(H_m) e_1$$

Proof. See [47, Theorem 6.6]. □

The same statement holds for the Gauss-Radau and Gauss-Lobatto rule with modified matrices $\widetilde{H}_{m+1}$ (see [47, Section 6]). For example, for the Gauss-Radau rule we have to extend the matrix H_m in such a way that it has an additional prescribed eigenvalue λ_1 or λ_n. For having an additional eigenvalue λ, we need to solve the system $(H_m - \lambda I)x = \beta_m^2 e_m$, where β_m is the norm of the last orthogonalized vector in the Lanczos process before normalization. Then the matrix $\widetilde{H}_{m+1}$ is given by

$$\widetilde{H}_{m+1} = \begin{pmatrix} H_m & \beta_m e_m \\ \beta_m e_m^T & \lambda + x_m \end{pmatrix}.$$

For the Gauss-Lobatto rule we need to prescribe the two additional eigenvalues λ_1 and λ_n.

Under certain conditions, the approximation (2.7) is an upper or a lower bound for the bilinear form $v^H f(A)v$: If $f^{(2n)}(\xi) > 0$ for all n and $\xi \in (\lambda_1, \lambda_n)$ then the approximation (2.7) provides a lower bound if the Gauss rule is used and an upper bound in the case of the Gauss-Lobatto rule. For the Gauss-Radau rule an upper bound can be obtained with $z_1 = \lambda_1$ if $f^{(2n+1)}(\xi) < 0$ for all n and $\xi \in (\lambda_1, \lambda_n)$. With $z_1 = \lambda_n$ we compute a lower bound. For example, when f is completely monotonic, i.e., when f satisfies

$$(-1)^k f^{(k)}(z) \geq 0 \text{ for all } k \in \mathbb{N}_0 \text{ and } z \in \mathbb{R}^+,$$

then all these conditions are fulfilled if A is Hermitian positive definite. Examples of completely monotonic functions are obviously given by $f(z) = z^{-1}$ or by $f(z) = \exp(-az)$ with $a \geq 0$. In Section 3.3 we introduce the important class of Cauchy–Stieltjes and Laplace–Stieltjes functions. It can be shown that every Cauchy–Stieltjes function $f : \mathbb{R}^+ \to \mathbb{R}$ is completely monotonic and that a function $f : \mathbb{R}^+ \to \mathbb{R}$ is completely monotonic if and only if it can be expressed as a Laplace–Stieltjes integral [13]. Hence, it is possible to compute upper and lower bounds for the Lanczos approximation of $v^H f(A)v$ for a large class of functions.

Since the Gauss quadrature rule (2.7) is exact for polynomials of degree $2N-1$ (see [47, Section 6.2]) the approximation (2.3) is exact for $v^H f(A)v$ if f is a polynomial

of degree at most $2m-1$. Because of the additional prescribed quadrature nodes in the Gauss-Radau and Gauss-Lobatto rule, these quadrature rules are exact for polynomials of degree $2N$ and $2N+1$, respectively, which directly lead to exact approximations for polynomials of degree at most $2m$ and $2m+1$. These exactness results establish an approximately twice as fast convergence of the Lanczos approximation of $v^H f(A)v$ compared to the approximation of $f(A)v$. In addition, we do not need to store the whole (dense) matrix of basis vectors V_m since the basis vectors can be normalized with only two previous basis vectors and the matrix V_m does not appear in the Lanczos approximation of $v^H f(A)v$. For non-Hermitian matrices, an exact result can be obtained for polynomials of degree at most $2m-1$ by using the nonsymmetric Lanczos process with m iterations (see [47, Section 8.1]) but in return two matrix-vector products with respect to A and A^H are required in every iteration step, which overall leads to the same number of matrix-vector products as for the Arnoldi process with $2m$ steps. For the Arnoldi approximation (2.3) we obtain an exact result for polynomials of degree at most m, which directly follows from the equality

$$v^H g(A)^H f(A)v = e_1^T g(H_m)^H f(H_m)e_1 \text{ for all } (g,f) \in \mathbb{W}_m,$$

from [17, Theorem 2.2] where $\mathbb{W}_m = (\mathbb{P}_{m-1} \oplus \mathbb{P}_m) \cup (\mathbb{P}_m \oplus \mathbb{P}_{m-1})$. By setting $g(z) = 1$, we have the claimed exactness.

The computation of bilinear forms $v^H f(A)v$ will be an important part in Section 4.4 and is therefore summarized in Algorithm 2.3.

Algorithm 2.3: Arnoldi/Lanczos approximation of $v^H f(A)v$.

Input: Matrix A, vector v and number of iterations m.

Set $\widetilde{v} = \frac{1}{\|v\|_2} v$
if *A is Hermitian* **then**
    Run m iterations of the Lanczos process with respect to A and $\widetilde{v}$ and compute the tridiagonal matrix H_m for the Gauss rule or a tridiagonal matrix $\widetilde{H}_{m+1}$ for the Gauss-Radau or Gauss-Lobatto rule.
else
    Run m iterations of the Arnoldi process with respect to A and $\widetilde{v}$ and compute the Hessenberg matrix H_m.
end
Compute $\|v\|_2^2 e_1^T f(H_m)e_1$ (or $\|v\|_2^2 e_1^T f(\widetilde{H}_{m+1})e_1$, respectively).

2.2 Graph theory

Throughout this thesis we will often use connections between properties of a matrix A and the corresponding graph $G(A)$ or vice versa. Therefore, in this section

we introduce central definitions and results from graph theory. For an elaborate introduction into graph theory and graph algorithms see, e.g., [28, 32, 60, 100, 105]. Classical and well-known definitions for graphs are embedded in the running text while new or modified definitions are emphasized.

A graph is defined as an ordered pair $G = (V, E)$ where V is a set of nodes and E is a set of edges. We distinguish between directed and undirected graphs. In an undirected graph, the edges have no orientation and are therefore defined by an unordered pair of nodes $\{v, w\}$ where $v, w \in V$. For an edge $\{v, w\} \in E$, the nodes v and w are said to be adjacent and v and w are incident to the edge $\{v, w\}$ and vice versa. Similarly, two edges are called adjacent if they share a common vertex. The degree $\deg(v)$ of a node v is defined as the number of adjacent nodes and $\Delta(G) := \max\{\deg(v) : v \in V\}$ defines the maximum degree over all nodes v in G. In a directed graph (or digraph), the edges are defined by an ordered pair of nodes (v, w) so that now an edge between two nodes has a direction. Every undirected graph can be considered as an directed graph by replacing every (undirected) edge $\{v, w\}$ by two (directed) edges (v, w) and (w, v). In the undirected graph $|G|$ corresponding to a directed graph G, every edge (v, w) is replaced by an undirected edge $\{v, w\}$. For an undirected graph G we set $|G| = G$. The indegree $\deg_{\text{in}}(v)$ of a node v in a digraph $G = (V, E)$ is defined as the number of edges ending in v, i.e., the number of edges $(w, v) \in E$. Similarly, the outdegree $\deg_{\text{out}}(v)$ is defined as the number of edges beginning in v, i.e., the number of edges $(v, w) \in E$.

We define a walk of length k from node v_1 to v_{k+1} in a graph (digraph) G as a sequence of nodes $(v_1, \ldots, v_{k+1})$, where $\{v_i, v_{i+1}\} \in E$ ($(v_i, v_{i+1}) \in E$) for $i = 1, \ldots, k$. A walk is closed, if $v_1 = v_{k+1}$. A path of length k from v_1 to v_{k+1} in G is a walk $(v_1, \ldots, v_{k+1})$, where $v_i \neq v_j$ for $i, j = 1, \ldots, k+1$. The geodesic distance $\mathrm{d}(v, w)$ of two nodes $v, w \in V$ is the length of the shortest path from v to w in G. We set $\mathrm{d}(v, w) = 0$ for $v = w$ and $\mathrm{d}(v, w) = \infty$ if $v \neq w$ and there is no path from node v to node w. Of course, for undirected graphs we have $\mathrm{d}(v, w) = \mathrm{d}(w, v)$ for all $v, w \in V$ which in general is not true for directed graphs.

Definition 2.9. The undirected distance $\bar{\mathrm{d}}(v, w)$ of two nodes $v, w \in V$ in the graph $G = (V, E)$ is the length of the shortest path connecting nodes v and w in $|G|$.

For the defined distances the relation

$$\bar{\mathrm{d}}(v, w) \leq \min\{\mathrm{d}(v, w), \mathrm{d}(w, v)\} \tag{2.8}$$

holds, where equality always holds for undirected graphs since $\bar{\mathrm{d}}(v, w) = \mathrm{d}(v, w) = \mathrm{d}(w, v)$, while a strict inequality is possible for undirected graphs as demonstrated

in a simple example in Figure 2.1. The diameter of a graph $G = (V, E)$ is defined as

$$\operatorname{diam}(G) = \max_{i,j \in V} \mathrm{d}(i, j).$$

For a graph G we define the corresponding adjacency matrix $A(G)$ as follows.

Definition 2.10. Let G be a graph with nodes $V = \{1, \ldots, n\}$ and edges E. The adjacency matrix of G is a matrix $A(G) \in \{0, 1\}^{n \times n}$ with entries

$$[A(G)]_{ij} = \begin{cases} 1 \text{ if } (i, j) \in E \text{ or } i = j \\ 0 \text{ else .} \end{cases}$$

This definition slightly differs from most classical definitions of adjacency matrices of graphs. In Definition 2.10 the whole diagonal of $A(G)$ is nonzero while in classical definitions a diagonal entry is nonzero if and only if the corresponding node has a self loop. Of course, in most applications it is important to store this self loop information. In this thesis it will be more important to have the following relation between the powers of the matrix $A(G)$ and the distances of the nodes in G which requires an adjacency matrix as defined above.
Let $A := A(G)$ be the adjacency matrix of a graph G. Of course, an entry $a_{i,j}$, $i \neq j$ of A indicates whether the nodes i and j are adjacent, i.e., whether there is a walk of length one between these nodes. Similarly, an entry $[A^2]_{ij} = \sum_{k=1}^{n} a_{ik} a_{kj}$ of A^2 gives us the information whether there is a walk from i to j with length two. In general, we have

$$[A^d]_{ij} = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_{d-1}=1}^{n} a_{ik_1} a_{k_1 k_2} \ldots a_{k_{d-1} j} \neq 0$$

if and only if there is a walk from i to j with length d. If the adjacency matrix A is defined in the classical way, the entry $[A^d]_{ij}$ is in addition the number of walks between i and j with exactly length d. On the other hand, if we define the adjacency matrix A as in Definition 2.10 the entry $[A^d]_{ij}$ not only gives the information whether there is a walk between i and j with *exactly* length d, it indicates if there is a walk between i and j with length *at most* d, due to the nonzero diagonal entries which represent artificial self loops. Now, since a walk with length at most d between i and j also implies a path between i and j with length at most d, we have the following relations between the powers of the adjacency matrix $A(G)$ and the distances of the nodes in the graph G.

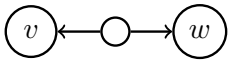

Figure 2.1: Directed Graph with $\mathrm{d}(v,w) = \mathrm{d}(w,v) = \infty$ and $\bar{\mathrm{d}}(v,w) = 2$.

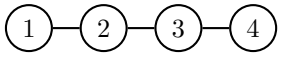

Figure 2.2: Undirected Graph from Example 2.12.

Lemma 2.11. *Let G be a graph and $A := A(G)$ the corresponding adjacency matrix as defined in Definition 2.10. Then for $d \geq 1$ the following equivalent statements hold:*

(i) $[A^d]_{i,j} = 0$ *if and only if* $\mathrm{d}(i,j) > d$.

(ii) $[A^d]_{ij} \neq 0$ *if and only if* $\mathrm{d}(i,j) \leq d$

These statements do not hold if the adjacency matrix A is defined in the classical way which is illustrated in the following example.

Example 2.12. Let $G = (V,E)$ be an undirected graph with $V = \{1,2,3,4\}$ and $E = \{\{1,2\},\{2,3\},\{3,4\}\}$. G is illustrated in Figure 2.2. For the distances of the nodes in G we have $\mathrm{d}(1,2) = \mathrm{d}(2,3) = \mathrm{d}(3,4) = 1$, $\mathrm{d}(1,3) = \mathrm{d}(2,4) = 2$ and $\mathrm{d}(1,4) = 3$. There are no self loops in G, hence the classical adjacency matrix $\tilde{A}(G)$ is given by

$$\tilde{A}(G) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

and therefore

$$\tilde{A}(G)^2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Now, $[\tilde{A}(G)^2]_{1,2} = [\tilde{A}(G)^2]_{2,3} = 0$ and $\mathrm{d}(1,2) = \mathrm{d}(2,3) = \mathrm{d}(3,4) \leq 2$, so Lemma 2.11 does not hold for $A = \tilde{A}(G)$ and $d = 2$. In contrast if $A = A(G)$ is given by

$$A(G) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

then

$$A(G)^2 = \begin{bmatrix} 2 & 2 & 1 & 0 \\ 2 & 3 & 2 & 1 \\ 1 & 2 & 3 & 2 \\ 0 & 1 & 2 & 2 \end{bmatrix},$$

and Lemma 2.11 holds for $A = A(G)$ and $d = 2$. ◇

For the classical adjacency matrix $A = \tilde{A}(G)$ we only have the implication $\mathrm{d}(i,j) \leq d$ if $[A^d]_{ij} \neq 0$ and $[A^d]_{ij} = 0$ if $\mathrm{d}(i,j) > d$, respectively, i.e., if $[A^d]_{ij} = 0$ we cannot guarantee that $\mathrm{d}(i,j) > d$.

However, the classical adjacency matrix is an important tool in network analysis [29]. A graph is sometimes called a network if it stems from the modeling of a real world problem. Examples are railway networks, road networks or social networks. Oftentimes, a problem is modeled by a weighted graph or network where weights are assigned to the edges representing, e.g., lengths, costs or capacities. Definitions like the length of a walk or distances between nodes can be adapted by considering the weights of the edges. In the analysis of networks, functions of matrices play an important role as we will see in the following.

In network analysis quantities of interest are, e.g., the centrality or connectivity of nodes. Those centrality or connectivity measures are used for rating the "importance" of nodes and links between them in a network. The simplest way of rating the centrality of a node is given by the degree centrality, where nodes with high degree are more important or central than nodes with low degree. Such a centrality measure is easily available but it does not capture the global structure of the network. Hence, further quantities for the centrality of a node considering the global structure of the graph were developed [29, 30, 31]. The so called *subgraph centrality* of a node i is given by

$$SC(i) := [\exp(A)]_{ii}$$

where A is the classical adjacency matrix of the network. To see why this measure might be an appropriate (global) quantity for rating the centrality of a node, we write

$$[\exp(A)]_{ii} = \sum_{k=1}^{\infty} \frac{[A^k]_{i,i}}{k!},$$

i.e., the i-th diagonal entry of the exponential of an adjacency matrix of a network is the sum over the number of closed walks with length k for node i in which the number of walks with length k is weighted by the factor $\frac{1}{k!}$. The scaling factor represents the fact that shorter walks are typically more important than longer walks since information can be passed more quickly and efficiently through short walks. The sum over all nodes of the subgraph centrality is known as the *Estrada Index*

$$EE(G) := \mathrm{tr}(\exp(A)) := \sum_{i=1}^{n} \sum_{k=0}^{\infty} \frac{[A^k]_{i,i}}{k!}$$

which is often used for normalization purposes. With this centrality measure based on the exponential, the structure of a network is represented more comprehensively than by just considering direct neighbors. Instead of using the scaling factor $\frac{1}{k!}$ it is possible to assign other decreasing weights $c_k \geq 0$. Then, if $\sum_{k=0}^{\infty} c_k z^k$ converges

for $|z| < \rho(A)$, where $\rho(A)$ is the spectral radius of the adjacency matrix A, the f-centrality of a node i is given by

$$[f(A)]_{ii}$$

for a function

$$f(z) = \sum_{k=0}^{\infty} c_k z^k. \tag{2.9}$$

An important function for the f-centrality of a node besides the exponential is given by the resolvent. The resolvent centrality of a node i is defined as

$$RC(i) := [(\alpha I - A)^{-1}]_{ii} = \sum_{k=0}^{\infty} \alpha^{-(k+1)} [A^k]_{ii},$$

where $\alpha > \max\{1, \rho(A)\}$.

Besides the centrality of a node, the communicability between two nodes i and j does also play an important role in the analysis of networks. Similar to the definition of the f-centrality, the f-communicability of two nodes is defined as

$$[f(A)]_{ij},$$

for a function f with power series (2.9). The (i, j)-entry then represents the sum over the number of walks with length k from node i to j in which the number of walks with length k is scaled by c_k. Again, important examples for f-communicability of two nodes are induced by the exponential and the resolvent. Hence, these measures in graph and network analysis illustrate an important application for functions of large, sparse matrices.

So far, we defined matrices $A(G)$ corresponding to a network or graph G. The other way around, we can define a graph $G(A)$ corresponding to a matrix A. The *(directed)* graph $G(A) = (V, E)$ corresponding to a matrix $A \in \mathbb{C}^{n \times n}$ is defined by the set of nodes $V = \{1, \ldots, n\}$ and the set of edges $E = \{(i, j) \in V \times V : a_{ij} \neq 0, i \neq j\}$. The corresponding *undirected* graph $|G(A)|$ is given by the tuple $(V, |E|)$ with $|E| = \{\{i, j\} : (i, j) \in E\}$.

Definition 2.13. We call a matrix A structurally symmetric if $G(A) = |G(A)|$.

Obvious examples of structurally symmetric matrices are Hermitian and skew-Hermitian matrices. In the following, we will sometimes work with the graph $G(A)$ of a matrix A and the corresponding adjacency matrix from Definition 2.10, since we are interested in the distances of the nodes in $G(A)$. Note that the adjacency matrix of $G(A)$ is given by a binary matrix which represents the off-diagonal sparsity pattern of A with full diagonal.

2.3 Minimal polynomials

In this section, we introduce certain types of minimal polynomials. We define a minimal (or optimal) polynomial $p^* \in \mathbb{P}_m$ as a polynomial which minimizes $\|p\|$ over all polynomials $p \in \mathbb{P}_m$ with an additional normalization condition which excludes the trivial choice $p^* = 0$. In the following, we consider the minimal polynomial p^* with respect to the supremum or Chebyshev norm on a set $\Omega \subseteq \mathbb{C}$ which is defined for real- or complex-valued functions f defined on Ω as

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

In addition, we claim that Ω is a compact set and $p^* \in \mathbb{P}_m^\gamma := \{p \in \mathbb{P}_m : p(\gamma) = 1\}$ for $\gamma \notin \Omega$. Summarizing, we are interested in a polynomial p^* which solves the minimization problem

$$\min_{p \in \mathbb{P}_m^\gamma} \|p(z)\|_\Omega = \min_{p \in \mathbb{P}_m^\gamma} \max_{z \in \Omega} |p(z)|. \tag{2.10}$$

Sometimes, the minimal polynomial solving (2.10) is called a *min-max polynomial* with respect to Ω. For some sets Ω the minimal polynomial p^* is explicitly known. If Ω is an interval $[a, b]$, then the *Chebyshev polynomials* lead to optimal polynomials for the minimization problem (2.10). These types of polynomials are discussed in detail in Section 2.3.1. For general sets Ω, the minimal polynomial is not explicitly known but there exist asymptotically optimal polynomials, i.e., polynomials that converge to the minimal polynomial for increasing m. For a large class of sets Ω those asymptotically optimal polynomials are given by the *Faber polynomials* introduced in Section 2.3.2.

2.3.1 Chebyshev polynomials

Chebyshev polynomials can be defined in several equivalent ways. In this section, we introduce some of those definitions and important properties of Chebyshev polynomials. For a detailed overview see [81] or [43, Chapter 3].

Definition 2.14. The m-th Chebyshev polynomial T_m is defined by

$$T_m(z) = \begin{cases} \cos(m\xi) \text{ where } \cos(\xi) = z \text{ for } z \in [-1, 1], \\ \cosh(m\xi) \text{ where } \cosh(\xi) = z \text{ for } z \notin [-1, 1]. \end{cases} \tag{2.11}$$

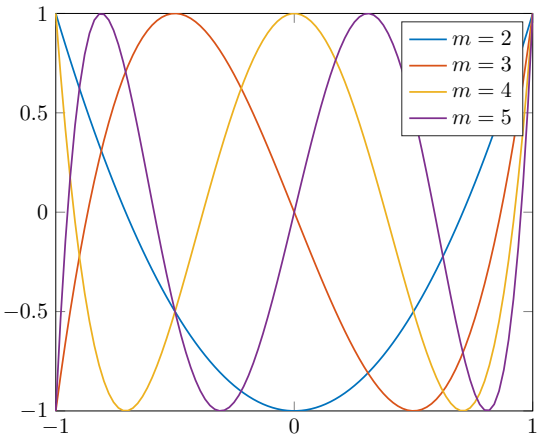

Figure 2.3: Chebyshev polynomials T_m for $m = 2, 3, 4, 5$.

Definition 2.14 does not immediately reveal T_m to be a polynomial of degree m. By induction, it can be shown that the three-term recurrence relation

$$T_{m+1}(z) = 2zT_m(z) - T_{m-1}(z), \tag{2.12}$$

with $T_1(z) = z$ and $T_0(z) = 1$, holds. The relation (2.12) is sometimes used as a definition of Chebyshev polynomials in the literature, since it immediately establishes the polynomial property. Based on (2.12) with $T_1(z) = z$ and $T_0(z) = 1$ we easily obtain the next few Chebyshev polynomials

$$\begin{aligned} T_2(z) &= 2z^2 - 1, \\ T_3(z) &= 4z^3 - 3z, \\ T_4(z) &= 8z^4 - 8z^2 + 1, \\ T_5(z) &= 16z^5 - 20z^3 + 5z \end{aligned}$$

illustrated in Figure 2.3.

With Definition 2.14 and the identities $\cos(z) = \frac{1}{2}(\mathrm{e}^{\mathbf{i}z} + \mathrm{e}^{-\mathbf{i}z})$ and $\cosh(z) = \frac{1}{2}(\mathrm{e}^{z} + \mathrm{e}^{-z})$, respectively, we obtain the alternative expression

$$T_m(z) = \frac{1}{2}(w^m + w^{-m}), \tag{2.13}$$

where

$$z = \frac{1}{2}(w + w^{-1}). \tag{2.14}$$

Equation (2.14) has two solutions which are inverses of each other, i.e., the value of $T_m(z)$ does not depend on which of these solutions is chosen. The equation (2.14) is also known as the Joukowski mapping which maps a circle centered at

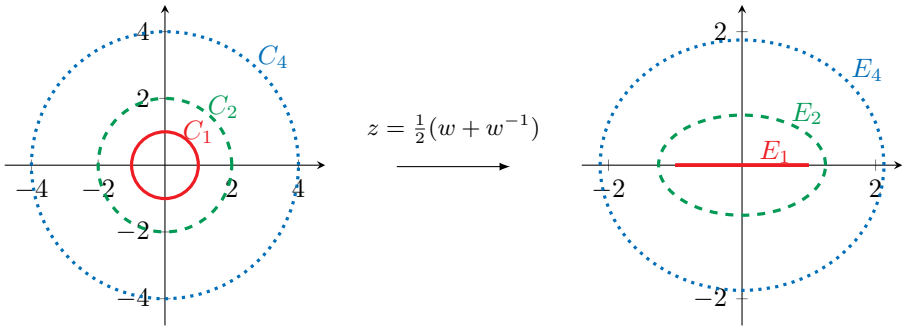

Figure 2.4: The Joukowski mapping $z = \frac{1}{2}(w + w^{-1})$.

the origin to an ellipse with focal points -1 and 1. This mapping will play an important role in the analysis of the results in Chapter 3 and is illustrated in Figure 2.4.

The most important property of these polynomials for this thesis is the minimization property of normalized Chebyshev polynomials. The polynomial

$$P_m(z) = \frac{T_m(z)}{T_m(\gamma)} \tag{2.15}$$

solves the minimization problem

$$\min_{p \in \mathbb{P}_m^\gamma} \max_{z \in [-1,1]} |p(z)|,$$

and since

$$\max_{z \in [-1,1]} |T_m(z)| = 1$$

we have

$$\min_{p \in \mathbb{P}_m^\gamma} \max_{z \in [-1,1]} |p(z)| = \frac{1}{|T_m(\gamma)|}.$$

This result can be easily generalized to any interval $[a, b]$ by using the transformation $t = 1 + 2\frac{z-b}{b-a}$ (see [85, Section 6.11] and the references therein). Hence, normalized Chebyshev polynomials are minimal polynomials with respect to intervals. There is also an interesting relation between Chebyshev polynomials and minimal polynomials with respect to ellipses.

Let C_ρ be a circle with radius ρ, centered at the origin. By applying the Joukowski mapping, we obtain an ellipse E_ρ with focal points -1 and 1 and semi-axes $\frac{\rho - \rho^{-1}}{2}$ and $\frac{\rho + \rho^{-1}}{2}$. Now, it is shown in [85, Theorem 6.5] that

$$\frac{\rho^m}{|w_\gamma|^m} \leq \min_{p_m \in \mathbb{P}_m^\gamma} \max_{z \in E_\rho} |p_m(z)| \leq \frac{\rho^m + \rho^{-m}}{|w_\gamma^m + w_\gamma^{-m}|}, \tag{2.16}$$

where w_γ is defined by $\gamma = \frac{1}{2}(w_\gamma + w_\gamma^{-1})$ and γ is chosen such that $p_m(\gamma) = 0$. The upper bound in (2.16) is achieved by the normalized Chebyshev polynomial (2.15). This can be seen by the fact that $T_m(\gamma) = w_\gamma^m + w_\gamma^{-m}$ and

$$\max_{z \in E_\rho} |T_m(z)| = \max_{w \in C(0,\rho)} \left| \frac{1}{2}(w^m + w^{-m}) \right| \leq \max_{w \in C(0,\rho)} \frac{1}{2}(|w|^m + |w|^{-m}) = \frac{1}{2}(\rho^m + \rho^{-m}), \tag{2.17}$$

and this upper bound is reached for $w = \rho$. Since the difference between the upper and lower bound for the min-max problem in (2.16) tends to zero for increasing m, Chebyshev polynomials are asymptotically optimal for ellipses E_ρ. By applying a variable transformation, this results holds for general ellipses.

The following lemma gives a useful lower bound with respect to Chebyshev polynomials which will be used for some of the results in Chapter 3.

Lemma 2.15. *Let T_m be the Chebyshev polynomial of degree m and let $z \in \mathbb{R}$ be of the form $z = 1 + 2x,\ x \in \mathbb{R}$. Then*

$$T_m(z) \geq \frac{1}{2}\left(\sqrt{x} + \sqrt{x+1}\right)^{2m}.$$

Proof. See [85, Section 6.11.3]. □

The existence of the three-term recurrence relation (2.12) also follows from the fact that the Chebyshev polynomials $\{T_m\}_{m=0}^{\infty}$ form a sequence of orthogonal polynomials on $[-1, 1]$ with respect to the weight function $(1 - z^2)^{-1/2}$, i.e.,

$$\int_{-1}^{1} T_m(x) T_k(x) \frac{\mathrm{d}x}{\sqrt{1 - x^2}} = N_m \delta_{mk},$$

with $N_0 = \pi$ and $N_m = \frac{\pi}{2}$ if $m \neq 0$. If f is a function which is continuous on $[-1, 1]$, then f has a Chebyshev series expansion

$$f(x) = \sum_{k=0}^{\infty} c_k T_k(x), \quad \text{for } x \in [-1, 1]$$

with coefficients

$$c_k = N_k^{-1} \int_{-1}^{1} \frac{f(x) T_k(x)}{\sqrt{1 - x^2}} \mathrm{d}x = N_k^{-1} \int_0^{\pi} f(\cos(t)) \cos(kt) \mathrm{d}t.$$

This result can be generalized to functions continuous on a line segment $[a, b]$ in the complex plane, i.e.,

$$[a, b] := \{z \in \mathbb{C} : z = ax + b(1 - x),\ 0 \leq x \leq 1,\ a, b \in \mathbb{C}\},$$

by considering the function $g = f \circ t^{-1}$ where t is a affine linear function which maps the line segment $[a, b]$ to the interval $[-1, 1]$. This directly follows from [15, Theorem 7], a convergence result of the Chebyshev series on ellipses.

2.3.2 Faber polynomials

Faber polynomials can be viewed as a generalization of Chebyshev polynomials which are asymptotically optimal for a larger class of sets Ω than ellipses. The following definitions and properties of Faber polynomials can be found in [96].

Let the extended complex plane be defined as $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$ and let G be a bounded simply connected domain with boundary Γ such that $D := \overline{\mathbb{C}} \setminus (G \cup \Gamma)$ is simply connected. Then by the Riemann mapping theorem there exists a unique function $w = \Phi(z)$ holomorphic in $D \setminus \{\infty\}$, which maps D conformally and univalently onto the domain $|w| > 1$, i.e., onto the outside of the unit disk, satisfying

$$\Phi(\infty) = \infty, \quad \Phi'(\infty) =: \gamma > 0.$$

The condition $\Phi'(\infty) = \gamma > 0$ is sometimes written as

$$\lim_{z\to\infty} \frac{\Phi(z)}{z} = \gamma > 0,$$

i.e., $w = \Phi(z)$ has a simple pole at $z = \infty$ and the Laurent expansion of $\Phi(z)$ in some neighborhood of $z = \infty$ is given by

$$\Phi(z) = \gamma z + \gamma_0 + \frac{\gamma_1}{z} + \frac{\gamma_2}{z^2} + \cdots + \frac{\gamma_k}{z^k} + \cdots .$$

For a non-negative integer m we consider

$$\begin{aligned} \Phi(z)^m &= \left(\gamma z + \gamma_0 + \frac{\gamma_1}{z} + \frac{\gamma_2}{z^2} + \cdots + \frac{\gamma_k}{z^k} + \cdots\right)^m \\ &= \gamma^m z^m + a_{m-1}^{(m)} z^{m-1} + a_{m-2}^{(m)} z^{m-2} + \cdots + a_1^{(m)} z + a_0^{(m)} \\ &+ \frac{b_1^{(m)}}{z} + \frac{b_2^{(m)}}{z^2} + \cdots + \frac{b_k^{(m)}}{z^k} + \cdots . \end{aligned} \tag{2.18}$$

Now, the Faber polynomial $\Phi_m(z)$ of degree m with respect to G is given by the polynomial part of (2.18), i.e.,

$$\Phi_m(z) = \gamma^m z^m + a_{m-1}^{(m)} z^{m-1} + a_{m-2}^{(m)} z^{m-2} + \cdots + a_1^{(m)} z + a_0^{(m)}.$$

Example 2.16. If the domain G is a disk $D(z_0, r) = \{z \in \mathbb{C} : |z - z_0| \le r\}$ then the Riemann mapping is given by

$$\phi(z) = \frac{z - z_0}{r}$$

and thus the Faber polynomials with respect to G are

$$\Phi_m(z) = \left(\frac{z - z_0}{r}\right)^m, \quad m = 0, 1, 2, \ldots .$$

$\diamond$

Example 2.17. It can be shown (see, e.g., [96, p. 36]) that the Faber polynomials with respect to the line segment $[-1, 1]$ are given by

$$\Phi_m(z) = \begin{cases} T_0(z) \text{ for } m = 0, \\ 2\,T_m(z) \text{ for } m \geq 1, \end{cases}$$

where T_m are the Chebyshev polynomials introduced in Section 2.3.1. ◇

Let Γ_R denote the line in the z-plane, which under the mapping $w = \Phi(z)$ goes onto the circle $|w| = R > 1$, i.e.,

$$\Gamma_R = \{z : |\Phi(z)| = R\}$$

and let G_R denote the interior of Γ_R. Then the representation

$$\Phi_m(z) = \frac{1}{2\pi\mathbf{i}} \int_{\Gamma_R} \frac{\Phi(t)^m}{t - z} \mathrm{d}t, \quad z \in G_R$$

holds and can be considered as a definition of Faber polynomials, as well. Let the function $z = \psi(w)$ be the inverse function of $w = \Phi(z)$ which maps the domain $|w| > 1$ conformally and univalently onto the domain D. Then with the substitution $x = \psi(t)$ we alternatively obtain

$$\Phi_m(z) = \frac{1}{2\pi\mathbf{i}} \int_{|x|=R} \frac{x^m \psi'(x)}{\psi(x) - z} \mathrm{d}x, \quad z \in G_R.$$

Using these expressions, one can show that a function f analytic on G_R can be expressed as Faber series

$$f(z) = \sum_{m=0}^{\infty} a_m \Phi_m(z), z \in G_R$$

with

$$a_m = \frac{1}{2\pi\mathbf{i}} \int_{|t|=r} \frac{f(\psi(t))}{t^{m+1}} \mathrm{d}t = \frac{1}{2\pi\mathbf{i}} \int_{\Gamma_r} \frac{f(x)\Phi'(x)}{\Phi^{m+1}(x)} \mathrm{d}x$$

where $1 < r < R$ (see [96, Chapter 3]).

An important property of normalized Faber polynomials $\Phi_m(z)/\Phi_m(\gamma)$ is the asymptotical optimality for a bounded simply connected set Ω in the sense that

$$\limsup_{m\to\infty} \left(\min_{p \in \mathbb{P}_m^\gamma} \|p\|_\Omega \right)^{1/m} = \lim_{m\to\infty} \left(\frac{\|\Phi_m\|_\Omega}{|\Phi_m(\gamma)|} \right)^{1/m},$$

which directly follows from [89, Proposition 3.6].

Chapter 3

Bounds for the decay in functions of matrices

The computation of matrix functions $f(A)$ is a challenging task especially for large and sparse matrices A. A major problem is the in general full structure of $f(A)$ which makes it impossible to store the matrix $f(A)$, although this was possible for the sparse matrix A. This problem is common knowledge for the inverse but it is also apparent for general matrix functions. In [11] the full structure of $f(A)$ is proven for irreducible matrices A and a large class of functions f. Based on this result, it seems to be impossible to compute $f(A)$ if A is a large, sparse matrix. However, even if $f(A)$ is dense, it sometimes has a special structure which is related to the sparsity pattern of A.

Example 3.1. Figure 3.1 shows the magnitude of the entries of the (dense) matrices A^{-1} with $A = \text{tridiag}(-1, 4, -1)$ and $B^{1/2}$ with $B = \left(\begin{smallmatrix} A & A \\ A & A \end{smallmatrix}\right)$, both of dimension $n = 100$. We observe that most of the entries are very small in magnitude and that the important, large entries are localized around the sparsity patterns of the matrices A and B, respectively. In particular, we have a *decay* of the entries away from the primal sparsity patterns. ◇

The decay of the entries of matrix functions was first discussed for the inverse of banded matrices in [23]. Example 3.1 shows that this phenomenon is not restricted to the inverse but is also apparent for other functions and general sparse (not necessarily banded) matrices. This decay behavior and localization property in matrix functions was noticed for many types of functions and matrices. A survey on this topic can be found in [8]. A definition of a decaying matrix is given in [8], where the considered matrix is not necessarily a matrix function. Thus, we give the following definition to specify the decay in matrix functions.

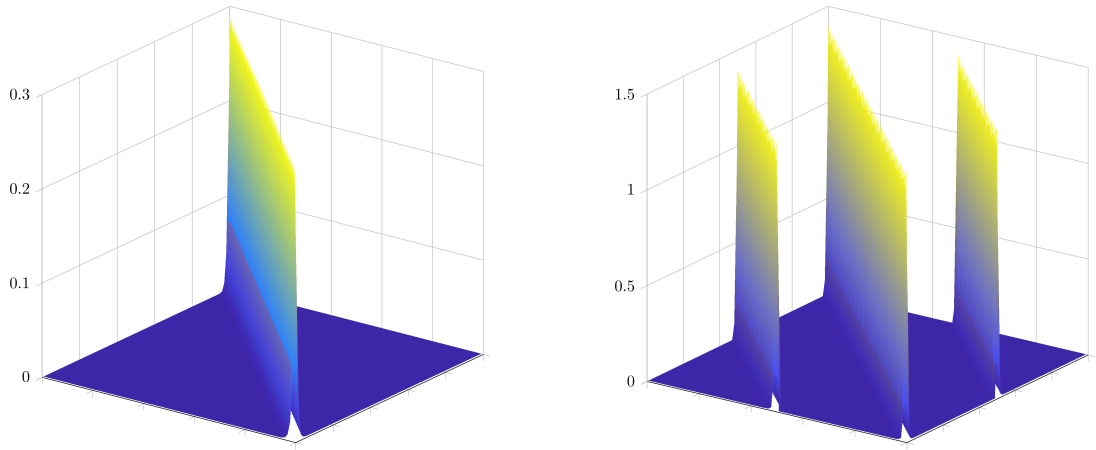

Figure 3.1: Magnitude of the entries of A^{-1} with $A = \text{tridiag}(-1, 4, -1)$, $A \in \mathbb{R}^{100\times 100}$ (left) and $B^{1/2}$, $B = \begin{pmatrix} A & A \\ A & A \end{pmatrix}$, $A \in \mathbb{R}^{50\times 50}$ (right).

Definition 3.2. Let $A \in \mathbb{C}^{n\times n}$ be a complex matrix, let f be a function defined on the spectrum of A and let $\mathrm{d}(i,j)$ denote the distance between the nodes i and j in $G(A)$, the graph of A. Then the matrix $f(A)$ is said to have decay away from the sparsity pattern of A if there exist a constant C and a function $\phi(x)$, both independent of n, such that

$$|[f(A)]_{ij}| \leq C\phi(\mathrm{d}(i,j)) \text{ for } i,j \in \{1,\ldots,n\}, \tag{3.1}$$

where the function $\phi(x)$ is defined and positive for $x \geq 0$ and $\phi(x) \to 0$ for $x \to \infty$.

We distinguish between certain modes of decay based on the function ϕ. For example, for $\phi(x) = q^x$ with $q < 1$ we have a decay which is termed *exponential* in [8] and which is faster than an algebraic decay where $\phi(x) = (1+x)^{-1}$. Especially an exponential decay is of great practical interest. In [8, 11] sequences of $n \times n$ matrices $\{A_n\}$ with *exponential off-diagonal* decay are considered (for off-diagonal decay we replace $\mathrm{d}(i,j)$ by $|i-j|$) and the following result was proven in [11].

Theorem 3.3. *Let $\{A_n\}$ be a sequence of $n \times n$ matrices satisfying an exponential off-diagonal decay, i.e.,*

$$|[A_n]_{ij}| \leq Cq^{|i-j|}$$

where C and $q < 1$ do not depend on n. Then for all $\epsilon > 0$ there exists $\tilde{m}$ independent of n such that

$$\|A_n - A_n^{(m)}\|_1 < \epsilon$$

for $m > \tilde{m}$ where $A_n^{(m)} \in \mathbb{C}^{n\times n}$ is a matrix with $[A_n^{(m)}]_{ij} = 0$ for $|i-j| > m$, containing only $\mathcal{O}(n)$ nonzeros.

Proof. Let $A_n^{(m)}$ be defined by

$$[A_n^{(m)}]_{ij} := \begin{cases} [A_n]_{ij}, & \text{if } |i-j| \leq m \\ 0 & \text{otherwise} \end{cases}.$$

then

$$\|A_n - A_n^{(m)}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |[A_n]_{ij} - [A_n^{(m)}]_{ij}| \leq \sum_{k=m+1}^{n} 2Cq^k \leq \sum_{k=m+1}^{\infty} 2Cq^k \leq \frac{2Cq^{m+1}}{1-q}.$$

Now for a given $\epsilon > 0$ one can find $\tilde{m}$ such that

$$\frac{2Cq^{m+1}}{1-q} \leq \epsilon$$

for every $m \geq \tilde{m}$. □

This result induces the possibility for computing a sparse approximation of a in general dense matrix $f(A)$ with an exponential (off-diagonal) decay. The assertion of Theorem 3.3 does not hold in full generality for exponential decaying matrices, i.e., matrices with $|[A_n]_{ij}| \leq Cq^{\mathrm{d}(i,j)}$ where the distance between the nodes is given with respect to a prescribed graph (see the discussion in Section 4.2). However, if most of the entries of $f(A)$ are very small in magnitude, as illustrated in Example 3.1, then it should be possible to compute $f(A)$ via a sparse approximation. For example, in this case an obvious sparse approximation of $f(A)$ is given by the matrix $\widetilde{f(A)}$ with

$$[\widetilde{f(A)}]_{ij} = \begin{cases} 0 & \text{if } |[f(A)]_{ij}| \text{ is sufficiently small,} \\ [f(A)]_{ij} & \text{else,} \end{cases}$$

i.e., by setting all those entries $[f(A)]_{ij}$ to zero which are (in absolute value) smaller than a prescribed threshold. Of course, it is not reasonable to first compute all entries of $f(A)$, e.g., by computing the bilinear forms $e_i^T f(A) e_j$, and then setting them to zero if they are small enough. Therefore it would be interesting to determine regions with small entries a priori without computing the entries of $f(A)$. There are already lots of publications dealing with the decay behavior of matrix functions and upper bounds for the entries of $f(A)$ for special types of functions and matrices. Especially for the matrix inverse there are results for Hermitian and certain types of tridiagonal matrices [23, 26, 34, 58, 71, 72, 76]. Also for the exponential, decay bounds were derived for special types of matrices

[12, 57, 91]. These results are restricted to quite special classes of matrices in order to obtain sharp bounds for the entries $|[f(A)]_{ij}|$. We also find bounds for more general settings, e.g., for the inverse of general matrices [23] or bounds for normal matrices and functions, which are analytic on a set containing the spectrum of A [10, 11]. These results are mainly of theoretical interest, since their decay bounds are rather pessimistic and/or very hard or impossible to compute in practice.

In this chapter we want to develop new decay bounds for special types of functions (especially the inverse and Cauchy–Stieltjes functions) including results which are intended for practical implementations as well as results which are mainly of theoretical interest. We first introduce the relation between decay in matrix functions and polynomial approximation, following the fundamental idea for the results in [23]. In Section 3.2 we use this relation for decay bounds for the inverse of special types of matrices. Results for functions which can be defined by an integral transform are discussed in Section 3.3.

Many of the results in this chapter have already been published in [38] and [39], while some are modified or completely new. We will refer to this at the corresponding results.

3.1 Relation between decay in matrix functions and polynomial approximation

In [23] a polynomial approximation of the inverse was used for decay bounds of Hermitian, positive definite, banded matrices. This approach can be easily transferred to general functions of sparse matrices. In this section we briefly sketch a generalization of the result in [23], yielding a relation between polynomial approximation and the decay in general matrix functions.

Let f be a function defined on a compact set $\Omega \subset \mathbb{C}$. We define the best polynomial approximation p_m^* of degree at most m for f on Ω as the polynomial which solves the min-max problem

$$\min_{p_m \in \mathbb{P}_m} \max_{z \in \Omega} |f(z) - p_m(z)|$$

and we define the corresponding error $E_m(f, \Omega)$ as

$$E_m(f, \Omega) := \max_{z \in \Omega} |f(z) - p_m^*(z)|. \tag{3.2}$$

Assume we have a function f which is defined on the spectrum of A and we are interested in an upper bound for the entries $|[f(A)]_{ij}|$. Based on the relation between powers of A and the distances of the nodes in $G(A)$ (see Section 2.2) we have $[p_m(A)]_{ij} = 0$ for every polynomial of degree $m < \mathrm{d}(i, j)$. Let p_m^* be the best

polynomial approximation of degree $m = \mathrm{d}(i,j) - 1$ for f on a set Ω containing the spectrum of A. Then

$$|[f(A)]_{ij}| = |[f(A)]_{ij} - [p_m^*(A)]_{ij}| \leq \|f(A) - p_m^*(A)\|_2. \tag{3.3}$$

If A is normal, the eigendecomposition $A = U\Lambda U^H$ with Λ diagonal and U unitary induces the relation

$$\begin{aligned} \|f(A) - p_m^*(A)\|_2 &= \|U(f(\Lambda) - p_m^*(\Lambda))U^H\|_2 \\ &= \|f(\Lambda) - p_m^*(\Lambda)\|_2 \\ &= \max_{z \in \sigma(A)} |f(z) - p_m^*(z)| \\ &\leq E_m(f, \Omega), \end{aligned} \tag{3.4}$$

i.e., every (i,j)-entry of the function of a normal matrix can be bounded by the error of a (not necessarily best) polynomial approximation of f of degree $m = \mathrm{d}(i,j) - 1$ on Ω.

For non-normal, but diagonalizable matrices $A = W\Lambda W^{-1}$ we obviously have the relation

$$\|f(A) - p_m(A)\|_2 \leq \|W\|_2 \|W^{-1}\|_2 E_m(f, \Omega) = \kappa(W) E_m(f, \Omega),$$

where $\kappa(W)$ is the condition number of W, i.e., it is possible to bound the entries of a function of a non-normal, diagonalizable matrix A with the error of a polynomial approximation as well. However, the additional constant $\kappa(W)$ might be very large for highly non-normal matrices, which makes this bound useless in practice. In addition the computation of $\kappa(W)$ is typically not feasible. Alternatively, for non-normal matrices A the following relation can be used.

For a compact set $\widetilde{\Omega} \subset \mathbb{C}$ containing $W(A)$, the field of values of A, and f continuous on $\widetilde{\Omega}$ and analytic on the interior of $\widetilde{\Omega}$, the relation

$$\|f(A)\|_2 \leq C \max_{z \in W(A)} |f(z)| \leq C \max_{z \in \widetilde{\Omega}} |f(z)| \tag{3.5}$$

holds for a constant C (the so called Crouzeix constant) with $2 \leq C \leq 1 + \sqrt{2}$ (see [20] and the reference therein). Hence, the entries of functions of non-normal matrices can be bounded by the error of a polynomial approximation with respect to a set $\widetilde{\Omega}$ containing $W(A)$. Again, for highly non-normal matrices such a bound might be too pessimistic since the set $\widetilde{\Omega}$ containing $W(A)$ can be very large compared to a set containing the spectrum of A and therefore the error $E_m(f, \widetilde{\Omega})$ might be very large as well.

Based on these relations, in some cases decay bounds immediately follow from well known results in approximation theory. We now present two examples from

the literature for decay bounds based on polynomial approximation. Most of the results in the literature were formulated for β-banded matrices. We say that a matrix A has bandwidth $\beta \ll n$ (or, as a shorthand, that A is β-banded) if $a_{ij} = 0$ for $|i-j| > \beta$. Using this definition, a tridiagonal matrix has bandwidth $\beta = 1$, a pentadiagonal matrix has bandwidth $\beta = 2$ and so on. A matrix is said to have upper bandwidth β and lower bandwidth γ if $a_{ij} = 0$ for $i - j > \beta$ or $j - i > \gamma$. Many of the results from the literature in the following sections are formulated for β-banded matrices, i.e., for $\beta = \gamma$ but they can be easily transferred to matrices with different upper and lower bandwidths as well. In addition, they can be generalized to sparse matrices by using the distance $\mathrm{d}(i,j)$ of the nodes i and j in $G(A)$.

For the first result we need to define the ellipse E_ρ with focal points -1 and 1 and semi-axes $\frac{\rho-\rho^{-1}}{2}$ and $\frac{\rho+\rho^{-1}}{2}$. Based on an upper bound for the error of a polynomial approximation on ellipses due to Bernstein (see, e.g., [67, Theorem 68]) we obtain the following result for a large class of functions f which was given in [10].

Theorem 3.4. *Let A be Hermitian and β-banded with spectrum contained in $[-1,1]$. Let f be analytic on the interior of an ellipse E_ρ and continuous on E_ρ for $\rho > 1$ and let $f(z)$ be real for real z. Then*

$$|[f(A)]_{ij}| \leq C \left(\frac{1}{\rho}\right)^{\frac{|i-j|}{\beta}}, \tag{3.6}$$

with

$$C = \frac{2\rho M(\rho)}{\rho - 1} \text{ and } M(\rho) = \max_{z \in E_\rho} |f(z)|.$$

The bound of Theorem 3.4 actually represents a family of bounds for different choices of $\rho > 0$. Clearly, for the decay rate it would be advantageous to choose ρ as large as possible but then the constant, including the quantity $\max_{z \in E_\rho} |f(z)|$, deteriorates, especially if f is not an entire function but has a singularity on the boundary of an ellipse $E_{\bar{\rho}}$, $\rho < \bar{\rho}$. Thus, it is not trivial to find the minimizer of the right-hand side of (3.6) as a function in ρ but it can be determined (sometimes only numerically) in many cases. This theorem is formulated for Hermitian matrices with spectrum in $[-1,1]$ but it can be generalized to normal matrices with spectrum on a real interval $[\lambda_{\min}, \lambda_{\max}]$ by using a transformation t which maps $[\lambda_{\min}, \lambda_{\max}]$ to the interval $[-1,1]$. This was already mentioned in [10]. We now formulate a generalized version of Theorem 3.4. With $E(\rho, f_1, f_2)$ we denote an ellipse with focal points f_1 and f_2 and semi-axes $\frac{1}{4}|f_1 - f_2|(\rho - \rho^{-1})$ and $\frac{1}{4}|f_1 - f_2|(\rho + \rho^{-1})$.

Corollary 3.5. *Let A be normal with spectrum contained in the real interval $[\lambda_{\min}, \lambda_{\max}]$. Let f be analytic on the interior of an ellipse $E(\rho, \lambda_{\min}, \lambda_{\max})$ and*

continuous on $E(\rho, \lambda_1, \lambda_2)$ *for* $\rho > 1$ *and let* $f(z)$ *be real for real* z*. Then*

$$|[f(A)]_{ij}| \leq C \left(\frac{1}{\rho}\right)^{\frac{|i-j|}{\beta}}, \tag{3.7}$$

with

$$C = \frac{2\rho M(\rho)}{\rho - 1} \quad \text{and} \quad M(\rho) = \max_{z \in E(\rho, \lambda_{\min}, \lambda_{\max})} |f(z)|.$$

Proof. The transformation

$$t(z) = \frac{\lambda_{\min} + \lambda_{\max} - 2z}{\lambda_{\max} - \lambda_{\min}}$$

maps the interval $[\lambda_{\min}, \lambda_{\max}]$ to $[-1, 1]$ and the ellipse $E(\rho, \lambda_{\min}, \lambda_{\max})$ to E_ρ. Thus, we can apply Theorem 3.4 to the matrix function $g(B)$, where $g = f \circ t^{-1}$ and $B = t(A)$ with $\sigma(B) \subset [-1, 1]$. The assertion then follows by the back transformation of the quantities. □

Note that this result cannot be generalized to matrices with spectrum on a complex line segment $[\lambda_1, \lambda_2]$ as the result from Bernstein only holds when $f(z)$ is real for real z and this is in general not the case for the function $f \circ t^{-1}$ if λ_1 and λ_2 are complex numbers. However, the result of Theorem 3.4 shows the existence of exponentially decaying bounds for a large class of functions and matrices. An even more general result is given in [78]. This result is formulated for non-normal matrices where the function f is analytic on a convex continuum containing the field of values. The result is based on Faber polynomials and the Faber series of f introduced in Section 2.3.2.

Theorem 3.6. *Let* A *be a matrix with upper bandwidth* β *and lower bandwidth* γ *and define*

$$\xi = \begin{cases} \lceil (i-j)/\beta \rceil, & \text{if } i \leq j \\ \lceil (j-i)/\gamma \rceil, & \text{if } j \leq i. \end{cases}$$

Let E *be a convex continuum containing the field of values of* A *and let* Φ *be the conformal mapping of* E*. If* f *is analytic on the set* $G_\tau := \{w : |\Phi(w)| < \tau\}$ *with* $\tau > 1$ *and bounded on the boundary of* G_τ*, then*

$$|[f(A)]_{ij}| \leq 2\frac{\tau}{\tau - 1} \max_{|z|=\tau} |f(\Phi^{-1}(z))| \left(\frac{1}{\tau}\right)^{\xi}. \tag{3.8}$$

With Theorem 3.6 we again obtain a family of bounds and we have a trade-off in the choice of τ similar to the choice of ρ in Theorem 3.4. The larger τ, the better is the decay rate but the worse is the corresponding constant. Theorem 3.6 is

formulated for general convex, compact sets E and we are faced with the problem of finding a conformal mapping of the set E. More concrete bounds can be easily obtained by choosing sets E for which the conformal mapping is explicitly known. We now formulate the corresponding results for line segments and disks as we need the resulting bounds for a comparison in Section 3.2.3. Note that matrices with field of values on a line segment $[\lambda_1, \lambda_2]$ are normal, which directly follows from [59, Theorem 5]. In addition, since for normal matrices A the field of values of A is the convex hull of $\sigma(A)$ (see, e.g., [55, Section 1.2]), we have $W(A) \subset [\lambda_1, \lambda_2]$ if and only if $\sigma(A) \subset [\lambda_1, \lambda_2]$. Thus we obtain the following result.

Corollary 3.7. *Let A be a normal matrix with $\sigma(A) \subset [\lambda_1, \lambda_2]$ and let f be analytic on the interior of the ellipse $E(\tau, \lambda_1, \lambda_2)$ and bounded on the boundary of $E(\tau, \lambda_1, \lambda_2)$ for $\tau > 1$. Then*

$$|[f(A)]_{ij}| \leq 2\frac{\tau}{\tau - 1} \max_{z \in E(\tau, \lambda_1, \lambda_2)} |f(z)| \left(\frac{1}{\tau}\right)^{\mathrm{d}(i,j)}.$$

Proof. We apply Theorem 3.6 to $E = [\lambda_1, \lambda_2]$. Let t be the transformation which maps $[\lambda_1, \lambda_2]$ to $[-1, 1]$ and J be the Joukowski mapping (2.14). Then the conformal mapping Φ of E is given by $\Phi = J^{-1} \circ t$. Thus, we have

$$G_\tau = \{w : |\Phi(w)| < \tau\} = \{\Phi^{-1}(z) : |z| < \tau\} = \{t^{-1}(J(z)) : |z| < \tau\}$$

which is just the interior of the ellipse $E(\tau, \lambda_1, \lambda_2)$. In addition, using

$$\max_{|z|=\tau} |f(\Phi^{-1}(z))| = \max_{z \in E(\tau, \lambda_1, \lambda_2)} |f(z)|$$

gives the result. □

Similarly, we can formulate the result of Theorem 3.6 if E is a disk:

Corollary 3.8. *Let A be a matrix with field of values in a disk $D(z_0, r)$ with center z_0 and radius r and let f be analytic on the interior of the disk $D(z_0, r\tau)$ and bounded on the boundary of $D(z_0, r\tau)$ for $\tau > 1$. Then*

$$|[f(A)]_{ij}| \leq 2\frac{\tau}{\tau - 1} \max_{z \in D(z_0, r\tau)} |f(z)| \left(\frac{1}{\tau}\right)^{\mathrm{d}(i,j)}.$$

Proof. The assertion follows by noticing that the conformal mapping of $D(z_0, r)$ if given by $\Phi(z) = (z - z_0)/r$. □

In [78] bounds for the exponential and the inverse square root are derived based on Theorem 3.6 with respect to a (horizontal) ellipse containing the field of values of A and then the optimal τ which minimizes the right-hand side of (3.8) is determined.

With Theorem 3.4 and Theorem 3.6 we have results for very general settings resulting in implicit bounds where lots of parameters have to be determined in order to obtain practical bounds for concrete problems. In the following we will restrict the class of functions to obtain sharper and more practical bounds, starting with decay bounds for the inverse of certain types of normal matrices.

3.2 Bounds for the inverse

As the probably most important matrix function, we first consider decay bounds for the inverse. There are already lots of results for the inverse of certain types of matrices. In Section 3.2.1 we will discuss known results which are formulated for banded matrices but can be easily generalized to arbitrary sparse matrices. We skip those results which are restricted to very special classes of matrices, e.g., tridiagonal matrices where the tridiagonal structure of the matrix is explicitly used [58, 71, 72, 76]. We introduce new results for the inverse in Section 3.2.2. In Section 3.2.3 we then compare them to the results from the literature.

3.2.1 Literature review

The pioneering result for the inverse of Hermitian, positive definite matrices from Demko, Moss and Smith [23] uses the approach described in Section 3.1 and the explicit knowledge of the error of the best polynomial approximation of the inverse on a positive interval. In addition, an extension to the non-Hermitian case was given in [23]. The results are summarized in the following theorem.

Theorem 3.9. *Let A be a Hermitian positive definite and β-banded matrix with smallest eigenvalue $\lambda_{\min}$, largest eigenvalue $\lambda_{\max}$ and condition number $\kappa(A) = \lambda_{\max}/\lambda_{\min}$. Then*

$$|[A^{-1}]_{ij}| \leq Cq^{\frac{|i-j|}{\beta}} \tag{3.9}$$

with

$$q = \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \text{ and } C = \max\left\{\frac{1}{\lambda_{\min}}, \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2\lambda_{\max}}\right\}. \tag{3.10}$$

If A is nonsingular and β-banded, then

$$|[A^{-1}]_{ij}| \leq C_1 q_1^{\frac{|i-j|}{\beta}} \tag{3.11}$$

with

$$q_1 = \sqrt{\frac{\kappa(A)-1}{\kappa(A)+1}}, \quad \kappa(A) = \|A\|_2\|A^{-1}\|_2 \tag{3.12}$$

and

$$C_1 = (2\beta+1)q_1^{-2}\|A^{-1}\|_2\kappa(A)\max\left\{1, \left(\frac{1+\kappa(A)}{\kappa(A)}\right)^2/2\right\}. \tag{3.13}$$

It was shown in [23] that the bound of Theorem 3.9 for Hermitian positive definite matrices is sharp in the sense that the decay rate is exact for certain types of tridiagonal Toeplitz matrices. This was not shown for the second statement, which follows from the first one by using the relation $A^{-1} = A^H(AA^H)^{-1}$. Note that the condition number of A appears as a multiplicative factor in the definition of the constant C_1 in the non-Hermitian case. Due to this factor, the entries of A^{-1} are highly overestimated by the bound (3.11) in many cases especially for ill-conditioned problems, so the quality of the results might get worse by enlarging the class of matrices.

In [35] the important class of matrices of the form

$$A = cI + dT \text{ where } T = T^H, c, d, \in \mathbb{C}, d \neq 0, \tag{3.14}$$

is considered, resulting in the following bounds for the entries of the inverse.

Theorem 3.10. *Let A be a matrix of type* (3.14) *and β-banded. Define $\lambda_1 = c + d\lambda_{\min}(T)$ and $\lambda_2 = c + d\lambda_{\max}(T)$. Let $a = (\lambda_2+\lambda_1)/(\lambda_2-\lambda_1)$ be represented as $a = \alpha_R\cos(\psi) + \mathbf{i}\beta_R\sin(\psi)$ with $0 \leq \psi < 2\pi$ and*

$$\alpha_R = \frac{1}{2}\left(R + \frac{1}{R}\right) \text{ and } \beta_R = \frac{1}{2}\left(R - \frac{1}{R}\right).$$

Then

$$|[A^{-1}]_{ij}| \leq \frac{2R}{|\lambda_1-\lambda_2|}B(a)\left(\frac{1}{R}\right)^{\frac{|i-j|}{\beta}}, \tag{3.15}$$

where R is the solution of

$$\frac{1}{2}\left(R + \frac{1}{R}\right) = \frac{|\lambda_1|+|\lambda_2|}{|\lambda_1-\lambda_2|} \text{ with } R > 1$$

and

$$B(a) = \beta_R^{-1}\frac{R}{\sqrt{\alpha_R^2-\cos(\psi)^2}(\alpha_R+\sqrt{\alpha_R^2-\cos(\psi)^2})}.$$

3.2.2 New results

In this section we introduce results which are based on the approach described in Section 3.1, i.e., we need a polynomial approximation p_m of the inverse on a set Ω containing the spectrum of A. Actually, we do not explicitly need the polynomial p_m but rather a bound for the error $\max_{z\in\Omega} |z^{-1} - p_m(z)|$. Therefore, instead of looking for a polynomial approximation of the inverse directly it is possible to consider a normalized polynomial approximation of the zero function. Those polynomials also play an important role in iterative methods for the solution of systems of linear equations. For instance, they are used as iteration polynomials for Krylov subspace methods (e.g., the Chebyshev iteration [48, Section 11.2]) or for an convergence analysis of the CG method or GMRES [85, Section 6.11]. Of course, the best polynomial approximation of the zero function on a compact set $\Omega \subset \mathbb{C}$ leads to the minimal polynomials defined in Section 2.3. We now use the polynomials introduced in Section 2.3 with respect to certain sets, especially ellipses and, as a special case, line segments. In the following $E(\rho, f_1, f_2)$ is an ellipse with $\rho \geq 1$, focal points f_1 and f_2 and semi-axes $\frac{1}{4}|f_1 - f_2|(\rho - \rho^{-1})$ and $\frac{1}{4}|f_1 - f_2|(\rho + \rho^{-1})$. As a short-hand, we use the notation $E_\rho := E(\rho, -1, 1)$. Note in particular that $E(1, f_1, f_2)$ is a line segment connecting f_1 and f_2, as this is an important special case.

We start with a result for matrices where the spectrum lies in an ellipse excluding the origin. As already mentioned in Section 2.3, Chebyshev polynomials are asymptotically minimal polynomials on ellipses $E(\rho, f_1, f_2)$, i.e., they can be viewed as a polynomial approximation of the zero function for $E(\rho, f_1, f_2)$. Before the corresponding result is shown, we need some auxiliary results with respect to the function $\cosh : \mathbb{C} \to \mathbb{C}$ due to Definition 2.14 of Chebyshev polynomials.

Lemma 3.11. *Let $z \in \mathbb{C}$ be a complex number, then $\operatorname{Re}(z) = 0$ if and only if $\cosh(z) \in [-1, 1]$.*

Proof. First we assume that $\operatorname{Re}(z) = 0$, i.e., $z = \alpha\mathbf{i}$ for $\alpha \in \mathbb{R}$. Then

$$\cosh(z) = \frac{1}{2}\left(\mathrm{e}^{\alpha\mathbf{i}} + \mathrm{e}^{-\alpha\mathbf{i}}\right) = \frac{1}{2}\left(\mathrm{e}^{\alpha\mathbf{i}} + \overline{\mathrm{e}^{\alpha\mathbf{i}}}\right) = \operatorname{Re}\left(\mathrm{e}^{\alpha\mathbf{i}}\right) \in \mathbb{R}$$

and

$$|\cosh(z)| = |\operatorname{Re}\left(\mathrm{e}^{\alpha\mathbf{i}}\right)| \leq |\mathrm{e}^{\alpha\mathbf{i}}| = 1,$$

i.e., $\cosh(z) \in [-1, 1]$.

Now we assume $\cosh(z) \in [-1, 1]$. Then we know that

$$\cosh(z) = \frac{1}{2}\left(\mathrm{e}^{\operatorname{Re}(z)}\,\mathrm{e}^{\operatorname{Im}(z)\mathbf{i}} + \mathrm{e}^{-\operatorname{Re}(z)}\,\mathrm{e}^{-\operatorname{Im}(z)\mathbf{i}}\right) = \frac{1}{2}\left(\mathrm{e}^{\operatorname{Re}(z)}\,\mathrm{e}^{\operatorname{Im}(z)\mathbf{i}} + \mathrm{e}^{-\operatorname{Re}(z)}\,\overline{\mathrm{e}^{\operatorname{Im}(z)\mathbf{i}}}\right)$$

is real if and only if $\operatorname{Re}(z) = -\operatorname{Re}(z)$, i.e., $\operatorname{Re}(z) = 0$ or $\mathrm{e}^{\operatorname{Im}(z)\mathbf{i}} \in \mathbb{R}$. Assume $\mathrm{e}^{\operatorname{Im}(z)\mathbf{i}} \in \mathbb{R}$, then $x := \mathrm{e}^z \in \mathbb{R}$ and the function $|\frac{1}{2}(x + x^{-1})|$ attains its minimum at $x = \pm 1$. Therefore $|\cosh(z)| = |\frac{1}{2}(x + x^{-1})| > 1$ if $x \neq \pm 1$. If $x = \mathrm{e}^z = \pm 1$, then $\operatorname{Re}(z) = 0$. Summarizing we have $\operatorname{Re}(z) = 0$ if $\cosh(z) \in [-1, 1]$. □

Lemma 3.12. *Let $\lambda_1, \lambda_2 \in \mathbb{C}$ be complex numbers and assume $0 \notin [\lambda_1, \lambda_2]$. Then $x := \frac{\lambda_1+\lambda_2}{\lambda_2-\lambda_1} \notin [-1, 1]$.*

Proof. We have

$$x = \frac{\lambda_1 + \lambda_2}{\lambda_2 - \lambda_1} = \frac{|\lambda_2|^2 - |\lambda_1|^2 + \lambda_1\overline{\lambda_2} - \overline{\lambda_1}\lambda_2}{|\lambda_2 - \lambda_1|^2}.$$

Since $\lambda_1\overline{\lambda_2} - \overline{\lambda_1}\lambda_2 = 2\operatorname{Im}(\lambda_1\overline{\lambda_2})\mathbf{i}$, x is real if and only if

$$\lambda_1\overline{\lambda_2} = \overline{\lambda_1}\lambda_2. \tag{3.16}$$

By multiplying (3.16) with $\lambda_1\lambda_2$ we see that for $\lambda_1, \lambda_2 \neq 0$ equation (3.16) is equivalent to $\lambda_1 = \alpha\lambda_2$, with $\alpha = \pm|\lambda_1/\lambda_2|$. Since $0 \notin [\lambda_1, \lambda_2]$ we have $\alpha = |\lambda_1/\lambda_2| > 0$ and therefore

$$x = \frac{1 + \alpha}{1 - \alpha} > 1$$

if x is real, hence $x \notin [-1, 1]$. □

Now we can show the following result for normal matrices, which is not published in [38] and which can be viewed as a generalization of Theorem 2 in [38].

Proposition 3.13. *Let A be a normal matrix with spectrum contained in an ellipse $E(\rho, f_1, f_2)$ excluding the origin. Define $x := \frac{f_1+f_2}{f_2-f_1}$. Then for $i \neq j$*

$$|[A^{-1}]_{ij}| \leq \|A^{-1}\|_2 \, \frac{1 + \rho^{-2\mathrm{d}(i,j)}}{1 - q^{-2\mathrm{d}(i,j)}} \left(\frac{\rho}{q}\right)^{\mathrm{d}(i,j)} \leq C \left(\frac{\rho}{q}\right)^{\mathrm{d}(i,j)} \tag{3.17}$$

with

$$C = \frac{2\,\|A^{-1}\|_2}{1 - q^{-2}} \quad \textit{and} \quad q = \mathrm{e}^{\operatorname{Re}(z)} > 1,$$

where z is the solution of

$$x = \cosh(z) \quad \textit{with} \quad \operatorname{Re}(z) \geq 0.$$

Proof. Let T_{m+1} be the Chebyshev polynomial of degree $m + 1$ and define the function $t(z) := \frac{f_1+f_2-2z}{f_2-f_1}$ which maps the ellipse $E(\rho, f_1, f_2)$ to the ellipse E_ρ, then the polynomial

$$P_{m+1}(z) = \frac{T_{m+1}(t(z))}{T_{m+1}(t(0))} = \frac{T_{m+1}(t(z))}{T_{m+1}(x)}$$

is the normalized Chebyshev polynomial of degree $m+1$ with respect to the ellipse $E(\rho, f_1, f_2)$. Due to the normalization, we have $P_{m+1}(z) = 1 - zp_m(z)$ where p_m is a polynomial of degree m. We use p_m as a polynomial approximation of the inverse on $E(\rho, f_1, f_2)$, and based on (3.3) and (3.4) we have

$$|[A^{-1}]_{ij}| \leq \max_{z\in\sigma(A)} \left|\frac{1}{z} - p_m(z)\right| = \max_{z\in\sigma(A)} \left|\frac{P_{m+1}(z)}{z}\right| \leq \|A^{-1}\|_2 \max_{z\in\sigma(A)} |P_{m+1}(z)| \tag{3.18}$$

for $m = \mathrm{d}(i,j) - 1$. Because of (2.17)

$$\max_{z\in\sigma(A)} |P_{m+1}(z)| \leq \max_{z\in E(\rho,f_1,f_2)} |P_{m+1}(z)| \leq \frac{\rho^{m+1} + \rho^{-(m+1)}}{2\,|T_{m+1}(x)|}. \tag{3.19}$$

Using the representation (2.11) with $x = \cosh(z)$, we find

$$\begin{aligned}
\frac{1}{|T_{m+1}(x)|} &= \frac{1}{|\cosh((m+1)\,z)|} \\
&= \frac{2}{|\mathrm{e}^{z(m+1)} + \mathrm{e}^{-z(m+1)}|} \\
&\leq \frac{2}{\left||\mathrm{e}^{z}|^{m+1} - |\mathrm{e}^{z}|^{-(m+1)}\right|} \\
&= \frac{2}{|q^{m+1} - q^{-(m+1)}|}.
\end{aligned}$$

Since we choose for z the solution of $x = \cosh(z)$ with $\mathrm{Re}(z) \geq 0$, we have $q = \mathrm{e}^{\mathrm{Re}(z)} \geq 1$. The assertion $q > 1$ follows by the fact that $0 \notin [f_1, f_2]$, thus $x = \cosh(z) \notin [-1, 1]$ and therefore $\mathrm{Re}(z) \neq 0$ (see Lemma 3.12 and Lemma 3.11). Thus

$$\frac{1}{|T_{m+1}(x)|} \leq \frac{2}{q^{m+1} - q^{-(m+1)}}. \tag{3.20}$$

Putting (3.18), (3.19) and (3.20) together, we obtain

$$|[A^{-1}]_{ij}| \leq \|A^{-1}\|\,\frac{\rho^{m+1} + \rho^{-(m+1)}}{q^{m+1} - q^{-(m+1)}} = \|A^{-1}\|\,\frac{1 + \rho^{-2(m+1)}}{1 - q^{-2(m+1)}}\left(\frac{\rho}{q}\right)^{m+1} \leq C\left(\frac{\rho}{q}\right)^{m+1}.$$

Using $m + 1 = \mathrm{d}(i,j)$ gives the desired result. □

It is not immediately clear, that the bound of Proposition 3.13 does represent an exponential decay bound since we need the condition $\rho < q$. The next result shows that this is actually the case due to the assumption $0 \notin E(\rho, f_1, f_2)$.

Lemma 3.14. *Let $E(\rho, f_1, f_2)$ be an ellipse excluding the origin and let q be defined as in Proposition 3.13. Then $\rho < q$, i.e., the bound of Proposition 3.13 does represent an exponential decay bound.*

Proof. First note, that $E(\rho, f_1, f_2)$ can be constructed, by mapping a disk with radius ρ and centered at 0 to the ellipse E_ρ and mapping E_ρ to $E(\rho, f_1, f_2)$, so we have

$$E(\rho, f_1, f_2) = \left\{ t \in \mathbb{C} : t = \frac{f_1 + f_2 - w(f_2 - f_1)}{2}, w = \frac{1}{2}(z + z^{-1}), |z| \leq \rho \right\}.$$

Hence, the condition $0 \in E(\rho, f_1, f_2)$ is equivalent to the existence of a $z \in \mathbb{C}$ with $|z| \leq \rho$ and $0 = (f_1 + f_2 - w(f_2 - f_1))/2$ which is equivalent to

$$\frac{1}{2}(z + z^{-1}) = \frac{f_1 + f_2}{f_2 - f_1}. \tag{3.21}$$

Now assume $q \leq \rho$. Note that by the substitution $z' = \mathrm{e}^z$ an alternative representation of q in Proposition 3.13 is given by $q = |z'|$, where z' is the solution of (3.21) with $|z'| > 1$. Hence, we have $z' \in \mathbb{C}$ with $|z'| = q \leq \rho$ and (3.21) which is equivalent to $0 \in E(\rho, f_1, f_2)$. This contradicts the assumption on $E(\rho, f_1, f_2)$. □

Remark 3.15. The right hand side in (3.17) is only given for clarifying the existence of an exponential decay bound with constant C. In practice it would be more suitable to compute the decreasing function $\|A^{-1}\| \frac{1+\rho^{-2\mathrm{d}(i,j)}}{1-q^{-2\mathrm{d}(i,j)}}$ in $\mathrm{d}(i, j)$, which tends to $\|A^{-1}\|$ for $\mathrm{d}(i, j) \to \infty$. For normal matrices the quantity $\|A^{-1}\|_2$ can be bounded by $1/\min_{z \in E(\rho, f_1, f_2)} |z|$. ◇

Remark 3.16. The result of Proposition 3.13 can be generalized to non-normal diagonalizable matrices $A = W \Lambda W^{-1}$, by adding the factor $\kappa(W)$ to the constant C or by considering the field of values. Then it is reasonable to modify the proof of Proposition 3.13 to avoid additional constants for a bound of $\|A^{-1}\|$. In equation (3.18) of the proof, we can directly bound the maximum over the discrete set $\sigma(A)$ (or $W(A)$, respectively) by the maximum over the set $E(\rho, f_1, f_2)$ such that the factor $\|A^{-1}\|$ is just replaced by $1/\min_{z \in E(\rho, f_1, f_2)} |z|$. ◇

Proposition 3.17. *Let A be a matrix with field of values contained in an ellipse $E(\rho, f_1, f_2)$ excluding the origin. Define $x := \frac{f_1+f_2}{f_2-f_1}$. Then for $i \neq j$*

$$|[A^{-1}]_{ij}| \leq C \left(\frac{\rho}{q} \right)^{\mathrm{d}(i,j)} \tag{3.22}$$

with

$$C = \frac{1 + \sqrt{2}}{\min_{z \in E(\rho, f_1, f_2)} |z|} \frac{2}{1 - q^{-2}} \quad \textit{and} \quad q = \mathrm{e}^{\mathrm{Re}(z)} > \rho \geq 1,$$

where z is the solution of

$$x = \cosh(z) \quad \textit{with} \quad \mathrm{Re}(z) \geq 0.$$

It is not immediately clear how to find an appropriate ellipse including $\sigma(A)$ or $W(A)$, respectively. If the spectrum of a normal matrix A lies in a line segment $[\lambda_1, \lambda_2]$, then the result of Proposition 3.13 can be applied to the ellipse $E(1, \lambda_1, \lambda_2)$ and we "only" need the two endpoints of the line segment λ_1 and λ_2. The resulting bounds for normal matrices are given in Theorem 3.18. Note that normal matrices with spectrum on a line segment are of the form (3.14) and important special cases are given by (shifted) Hermitian and skew-Hermitian matrices.

Theorem 3.18. *Let A be a normal matrix with spectrum in a complex line segment $[\lambda_1, \lambda_2]$ excluding the origin. Define $x := \frac{\lambda_1+\lambda_2}{\lambda_2-\lambda_1}$. Then for $i \neq j$ we obtain the bound*

$$|[A^{-1}]_{ij}| \leq C \left(\frac{1}{q}\right)^{\mathrm{d}(i,j)} \tag{3.23}$$

with

$$C = \frac{2\,\|A^{-1}\|}{1-q^{-2}} \text{ and } q = \mathrm{e}^{\mathrm{Re}(z)} > 1,$$

where z is the solution of

$$x = \cosh(z) \text{ with } \mathrm{Re}(z) \geq 0.$$

Proof. The bound follows by applying Proposition 3.13 to the ellipse $E(1, \lambda_1, \lambda_2) = [\lambda_1, \lambda_2]$. □

We already published a similar result in [38], with sightly different proof.

Now we consider several special cases of normal matrices with eigenvalues on a line segment for which the approach of Theorem 3.18 is either not applicable, or for which better bounds can be obtained by using different techniques. In particular, we consider the important class of shifted skew-Hermitian matrices whose eigenvalues lie in a set of the form $E = a + \big(\mathbf{i}[-b_2, -b_1] \cup \mathbf{i}[b_1, b_2]\big)$, as well as the case of Hermitian indefinite matrices and skew-Hermitian matrices with eigenvalues below and above the origin. For those the approach of Theorem 3.18 cannot be applied since then $0 \in [\lambda_1, \lambda_2]$ for any line segment $[\lambda_1, \lambda_2]$ containing $\sigma(A)$.

Theorem 3.19. *Let A be a nonsingular matrix of the form $A = S + aI$ with $a \in \mathbb{R}$ where S is skew-Hermitian and its spectrum is contained in a set of the form $\mathbf{i}[-b_2, -b_1] \cup \mathbf{i}[b_1, b_2]$ with $b_1, b_2 \in \mathbb{R}$ and $b_1 < b_2$. Then*

$$|[A^{-1}]_{ij}| \leq C \cdot \begin{cases} q^{\mathrm{d}(i,j)} & \text{for } \mathrm{d}(i,j) \text{ even} \\ q^{\mathrm{d}(i,j)-1} & \text{for } \mathrm{d}(i,j) \text{ odd} \end{cases} \tag{3.24}$$

where

$$C = \frac{2}{\sqrt{a^2+b_1^2}},\ q = \left(\sqrt{x} + \sqrt{x+1}\right)^{-1} \quad \text{and} \quad x = \frac{a^2+b_1^2}{b_2^2-b_1^2} \tag{3.25}$$

Proof. Define $\Omega := a + (\mathbf{i}[-b_2, -b_1] \cup \mathbf{i}[b_1, b_2])$, then $\sigma(A) \subset \Omega$. For bounds for the entries $|[A^{-1}]_{ij}|$ we will find bounds for $E_m(f, \Omega)$ with $f(z) = z^{-1}$ and $m = \mathrm{d}(i, j) - 1$. In the following we just write $E_m(\Omega)$. The set Ω is mapped onto the interval $[-1, 1]$ by the transformation $t = 1 + 2\frac{(z-a)^2+b_1^2}{b_2^2-b_1^2}$. We define the polynomial P_{2k} of degree $2k$ as

$$P_{2k}(z) = P_k(t) = \frac{T_k(t)}{T_k(t_0)} \text{ with } t_0 = 1 + 2\frac{a^2 + b_1^2}{b_2^2 - b_1^2}. \tag{3.26}$$

The polynomial P_{2k} approximates the zero function on Ω, so that the polynomial p_{2k-1} defined by $P_{2k}(z) = 1 - zp_{2k-1}(z)$ is a polynomial approximation of $f(z) = 1/z$ on Ω. For m odd, i.e., $m = 2k - 1$, we find

$$E_m(\Omega) \leq \max_{z\in\Omega} \left| \frac{1}{z} - p_m(z) \right| = \max_{z\in\Omega} \left| \frac{P_{m+1}(z)}{z} \right| \leq \frac{\max_{z\in\Omega} |P_{m+1}(z)|}{\min_{z\in\Omega} |z|} = \frac{|T_{\frac{m+1}{2}}(t_0)|^{-1}}{\sqrt{a^2 + b_1^2}}.$$

With Lemma 2.15 we obtain

$$T_{\frac{m+1}{2}}(t_0) = T_{\frac{m+1}{2}}\left(1 + 2\frac{a^2 + b_1^2}{b_2^2 - b_1^2}\right) \geq \frac{1}{2}\left(\sqrt{\frac{a^2 + b_1^2}{b_2^2 - b_1^2}} + \sqrt{\frac{a^2 + b_1^2}{b_2^2 - b_1^2} + 1}\right)^{m+1}.$$

Summarizing, we have

$$E_m(\Omega) \leq C\, q^{m+1} \tag{3.27}$$

for $m = 2p - 1$ and

$$E_m(\Omega) \leq E_{m-1}(\Omega) \leq C\, q^m \tag{3.28}$$

for $m = 2p$. Hence, for $i \neq j$ we bound the entries $|[A^{-1}]_{ij}|$ by using the bounds (3.27) and (3.28) for $m = \mathrm{d}(i, j) - 1$. Since the bound

$$|[A^{-1}]_{ij}| \leq \|A^{-1}\|_2 \leq \frac{1}{\sqrt{a^2 + b_1^2}}$$

holds for all i, j the bound (3.24) also holds for $i = j$. □

As a direct consequence, we also find a result for Hermitian indefinite matrices with an imaginary shift.

Corollary 3.20. *Let A be a nonsingular matrix of the form $A = T + \mathbf{i} \cdot aI$ with $a \in \mathbb{R}$ where T is Hermitian and its spectrum is contained in a set of the form $[-b_2, -b_1] \cup [b_1, b_2]$ with $b_1, b_2 \in \mathbb{R}$ and $b_1 < b_2$. Then the bound* (3.24) *of Theorem 3.19 holds for the entries $|[A^{-1}]_{ij}|$.*

Proof. We can write A as $A = \mathbf{i}B$ with $B = S + aI$, $S = -\mathbf{i}T$. Then

$$|[A^{-1}]_{ij}| = |[(\mathbf{i}B)^{-1}]_{ij}| = |-\mathbf{i}[B^{-1}]_{ij}| = |-\mathbf{i}|\,|[B^{-1}]_{ij}| = |[B^{-1}]_{ij}|. \tag{3.29}$$

Since B fulfills the assumptions of Theorem 3.19 and $\sigma(B) \subset a + (i[-b_2, -b_1] \cup i[b_1, b_2])$, we can apply the bound (3.24). □

As a special case, Theorem 3.19 and Corollary 3.20 yield results for indefinite Hermitian or skew-Hermitian matrices without a shift. In this case, the formula for the decay rate q greatly simplifies, as shown in the following result.

Corollary 3.21. *Let A be indefinite Hermitian or skew-Hermitian and nonsingular. Then the entries of A^{-1} can be bounded by*

$$|[A^{-1}]_{ij}| \leq 2 \|A^{-1}\|_2 \cdot \begin{cases} q^{\mathrm{d}(i,j)} & \text{for } \mathrm{d}(i,j) \text{ even} \\ q^{\mathrm{d}(i,j)-1} & \text{for } \mathrm{d}(i,j) \text{ odd} \end{cases} \tag{3.30}$$

where

$$q = \sqrt{\frac{\kappa(A)-1}{\kappa(A)+1}}.$$

Proof. The indefinite Hermitian and skew-Hermitian cases are special cases of Theorem 3.19 and Corollary 3.20 for $a = 0$. A straightforward calculation shows that

$$q^2 = \left(\sqrt{\frac{a^2+b_2^2}{a^2+b_1^2}} - 1\right) / \left(\sqrt{\frac{a^2+b_2^2}{a^2+b_1^2}} + 1\right)$$

for the decay rate q from Theorem 3.19. Now for $a = 0$, $b_1 = \min_{\lambda\in\sigma(A)} |\lambda|$ and $b_2 = \max_{\lambda\in\sigma(A)} |\lambda|$ we obtain

$$q^2 = \frac{\frac{b_2}{b_1} - 1}{\frac{b_2}{b_1} + 1} = \frac{\kappa(A)-1}{\kappa(A)+1}$$

and $C = 2\|A^{-1}\|_2$ which completes the proof. □

We published the results of Theorem 3.19, Corollary 3.20 and Corollary 3.21 in [38]. All the results in this section are based on (asymptotically) minimal polynomials with respect to a set containing the spectrum of A. So far we considered ellipses, line segments (as a special case of ellipses) and "splitted" line segments. The following result which was not given in [38] is a straightforward generalization of this approach based on the Faber polynomials introduced in Section 2.3.2.

Proposition 3.22. *Let A be normal and let Ω be a compact, simply connected set containing the spectrum of A and excluding the origin. Then*

$$|[A^{-1}]_{ij}| \leq \frac{\max_{z\in\Omega} |P_{\mathrm{d}(i,j)}(z)|}{\min_{z\in\Omega} |z|},$$

where $P_m(z) = \Phi_m(z)/\Phi_m(0)$ and $\Phi_m(z)$ is the m-th Faber polynomial with respect to Ω.

Now, by using other sets Ω than ellipses or line segments, additional results can be obtained. As an example, we give the corresponding result for $\Omega := D(z_0, r)$, i.e., Ω is a disk with center z_0 and radius r. We also give the generalized result for non-normal matrices, since we use these bounds for a comparison in Section 3.2.3.

Corollary 3.23. *Let $D(z_0, r)$ be a disk excluding the origin and containing the spectrum of a normal matrix A. Then*

$$|[A^{-1}]_{ij}| \leq C\, q^{\mathrm{d}(i,j)}$$

with $C = 1/\min_{z \in D(z_0.r)} |z|$ and $q = r/|z_0|$.

For non-normal matrices the bound

$$|[A^{-1}]_{ij}| \leq (1 + \sqrt{2})\, C\, q^{\mathrm{d}(i,j)}$$

holds, where $D(z_0, r)$ is a disk excluding the origin and containing the field of values of A.

Proof. The assertion directly follows using (3.5), Proposition 3.22 and the fact that the conformal map of $D(z_0, r)$ is given by $\Phi(z) = (z - z_0)/r$. □

As another example, the conformal map of so called "bratwurst" sets $s\Omega_\epsilon$ was given in [64] in order to develop a polynomial iterative method for solving systems of linear equations. The results in [64] can immediately be used to formulate decay bounds for matrices in cases where it is not possible to enclose the spectrum by a convex set excluding the origin.

Corollary 3.24. *Let $\gamma \in \delta D(0, 1)$ and $\chi \in (0, 2\pi)$. Let A be normal with $\sigma(A) \subseteq s\Omega_\epsilon$ where*

$$\begin{aligned} s\Omega_\epsilon &:= \{s\Psi_\epsilon(z) : z \in D(0,1)\}, \quad \Psi_\epsilon(z) := \frac{(z - \gamma N_\epsilon)(z - \gamma M_\epsilon)}{(N_\epsilon - M_\epsilon)z + \gamma(N_\epsilon M_\epsilon - 1)}, \\ N_\epsilon &:= \frac{1}{2}\left(\frac{P}{1+\epsilon} + \frac{1+\epsilon}{P}\right), \quad M_\epsilon := \frac{(1+\epsilon)^2 - 2}{2\tan(\phi/4)(1+\epsilon)}, \\ P &:= \tan(\chi/4) + (\cos(\chi/4))^{-1}, \quad \epsilon_{\max} := \tan(\chi/4)(1 + \tan(\chi/8) \end{aligned}$$

and $\epsilon \in [0, \epsilon_{\max})$. Then

$$|[A^{-1}]_{ij}| \leq \frac{V(s\Omega_\epsilon)}{\pi\left(N_\epsilon^{\mathrm{d}(i,j)} + M_\epsilon^{\mathrm{d}(i,j)} - (-S_\epsilon)^{\mathrm{d}(i,j)}\right)}$$

where $V(s\Omega_\epsilon)$ is the boundary rotation of $s\Omega_\epsilon$.

The sets $s\Omega_\epsilon$ generated by the mapping Ψ_ϵ are exemplarily illustrated in Figure 3.2 for $s = 1$, $\gamma = -1$, $\chi = \pi/2$ and different values of ϵ.

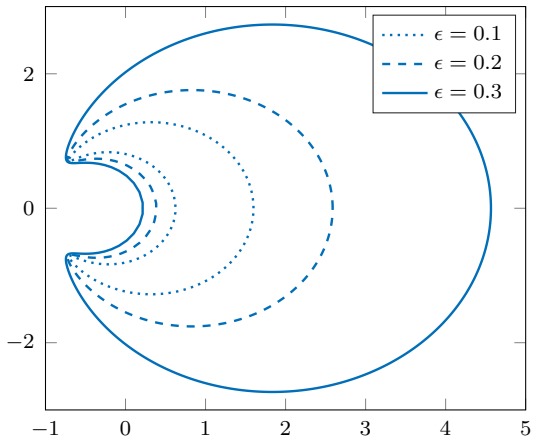

Figure 3.2: Sets $s\Omega_\epsilon$ with $s = 1$, $\gamma = -1$, $\chi = \pi/2$ and $\epsilon = 0.1, 0.2, 0.3$.

3.2.3 Comparison and numerical examples

In this section we compare various aspects of the introduced bounds from the literature and our new bounds from Section 3.2.2.

We start with the discussion of the results of Theorem 3.18 for matrices with spectrum on a line segment excluding the origin. In Section 3.1 we introduced Corollary 3.7 (as a consequence of Theorem 3.6 from [78]) which can be used for normal matrices with spectrum on a line segment and functions analytic on ellipses with focal points λ_1 and λ_2. Hence, it would be interesting to check whether there is an improvement by using the approach for the inverse of Section 3.2.2 instead of using the general result of Theorem 3.6 applied to a line segment and the inverse.

Since the inverse has a singularity in 0, the maximal τ such that f is analytic on the interior of the ellipse $E(\tau, \lambda_1, \lambda_2)$ and bounded on the boundary of $E(\tau, \lambda_1, \lambda_2)$ is smaller than the absolute value of the solution of

$$\frac{1}{2}(z + z^{-1}) = \frac{\lambda_1 + \lambda_2}{\lambda_2 - \lambda 1},$$

which is just q as defined in Theorem 3.18, i.e., $1 < \tau < q$. Therefore the decay rate of Theorem 3.18 is always better than the decay rate from Corollary 3.7 in the case of the inverse. At the same time, the constant from Corollary 3.7 is given by

$$\frac{2\tau}{\tau - 1} \max_{z \in E(\tau,\lambda_1,\lambda_2)} |z^{-1}| = \frac{2}{1 - \tau^{-1}} \frac{1}{\min_{z \in E(\tau,\lambda_1,\lambda_2)} |z|} =: I \cdot II,$$

whereas the constant from Theorem 3.18 can be bounded by

$$\frac{2}{1 - q^{-2}} \frac{1}{\min_{z \in [\lambda_1,\lambda_2]} |z|} =: \widetilde{I} \cdot \widetilde{II}.$$

As already mentioned, there is a trade-off in Corollary 3.7 for the choice of τ. For the decay rate and the factor I it would be advantageous to choose τ as large as possible. By setting $\tau = q$ we obtain the decay rate from Theorem 3.18 and a factor I which is still larger than $\widetilde{I}$. At the same time, the factor II tends to infinity for $\tau \to q$. For the factor II it would be advantageous to choose τ as small as possible. By setting $\tau = 1$, the factor II coincides with $\widetilde{II}$ but I is not defined and we would have no decay in the bound of Corollary 3.7. Summarizing, the bound of Theorem 3.18 is sharper than the bound of Corollary 3.7 even if we use the optimal choices of τ for the decay rate and the factors I and II, independently, implying that for any choice of τ the bound of Theorem 3.18 is smaller than that of Corollary 3.7. We conclude that the approach of Section 3.2.2 leads to better decay bounds for the inverse in comparison to results for general functions applied to the inverse and we do not need to solve any minimization problem.

For the next comparison we consider the classical results of Theorem 3.9 from [23]. The first result of Theorem 3.9 was formulated for the inverse of Hermitian positive definite matrices. Since Hermitian positive definite matrices belong to the class of normal matrices with spectrum on a line segment, we can use the approach of Section 3.2.2 for those types of matrices as well, i.e., we can bound the entries of A^{-1} by

$$|[A^{-1}]_{ij}| \leq \frac{\max_{z\in[\lambda_{\min},\lambda_{\max}]} |P_{m+1}(z)|}{\min_{z\in[\lambda_{\min},\lambda_{\max}]} |z|},$$

where P_{m+1} is a normalized Chebyshev polynomial of degree $m+1 = \mathrm{d}(i,j)$. We actually know that the result of Theorem 3.9 will lead to better results for these types of matrices, since it is based on the explicit knowledge of the error of the best polynomial approximation of the inverse on a positive interval, and it was shown in [23] that this bound is sharp for certain types of matrices. However, it is interesting to check the quality of the bounds based on minimal polynomials by applying the approach of Section 3.2.2 to Hermitian positive definite matrices (and in addition, we will need the resulting decay bound on a later point) and compare them to the results of Theorem 3.9 for Hermitian positive definite matrices.

Using the normalized Chebyshev polynomial P_{m+1} of degree $m+1 = \mathrm{d}(i,j)$ with respect to the line segment $[\lambda_{\min}, \lambda_{\max}]$ where $\lambda_{\min}, \lambda_{\max} > 0$ are the smallest and largest eigenvalues of A, we obtain the bound

$$|[A^{-1}]_{ij}| \leq \frac{\max_{z\in[\lambda_{\min},\lambda_{\max}]} |P_{m+1}(z)|}{\min_{z\in[\lambda_{\min},\lambda_{\max}]} |z|} = \frac{\left(T_{m+1}\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)\right)^{-1}}{\lambda_{\min}}. \tag{3.31}$$

Using Lemma 2.15 and the identity

$$\left(\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}} + \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}+1}\right)^2 = \frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)}-1} \tag{3.32}$$

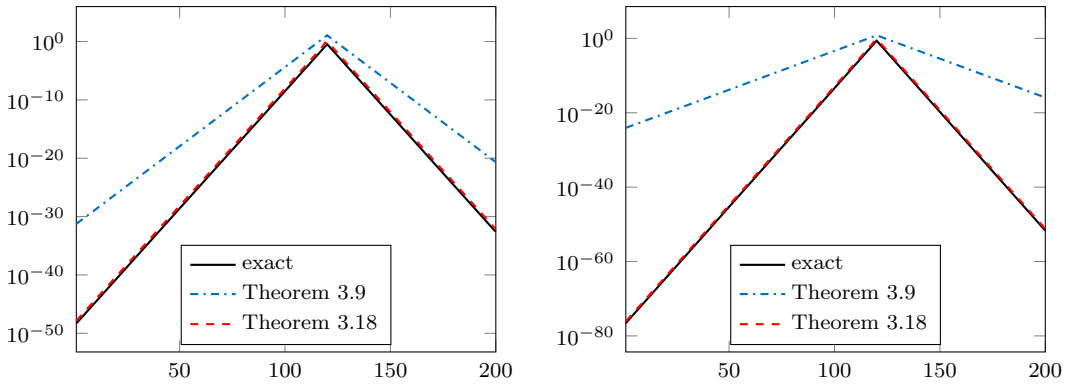

Figure 3.3: Exact values and bounds for $|[A^{-1}|_{ij}$, $j = 120$ of dimension $n = 200$ for $A = \text{tridiag}(1, \mathbf{i}, -1) + 2I$ (left) and $A = \text{tridiag}(-1, 4, -1) + 2\mathbf{i}I$ (right).

we find the bound

$$|[A^{-1}]_{ij}| \leq \frac{2}{a}\, q^{\mathrm{d}(i,j)}, \tag{3.33}$$

where q is the decay rate of Theorem 3.9, i.e., the decay rates coincide. For the constant C of Theorem 3.9 we have $\frac{1}{a} \leq C \leq \frac{2}{a}$ which shows that in the worst case the constant deteriorates by a factor of two if we apply the approach of Section 3.2.2 to Hermitian positive definite matrices instead of using the result of Theorem 3.9. Hence, in the Hermitian positive definite case we obtain an acceptable decay bound (also with sharp decay rate) when using the approach of Section 3.2.2 which can also be applied to general matrices with spectrum on a line segment. Note that we do not obtain this bound by just applying Theorem 3.18 to the interval $[\lambda_{\min}, \lambda_{\max}]$ since Lemma 2.15 can only be used in the real case. However, this comparison illustrates that a deterioration of the bounds caused by using the bound (3.31) instead of the the error of the best polynomial approximation is maintainable.

For non-Hermitian matrices with spectrum on a line segment excluding the origin, we can alternatively use the second result of Theorem 3.9 for general matrices. In this case we have rather different results for both, the constant and the decay rate. The decay rate (3.12) and constant (3.13) depend on the condition number $\kappa(A)$, whereas the decay rate and constant of Theorem 3.18 only depend on the two endpoints of the line segment which are not necessarily related to the condition number of A. In Figure 3.3 we see the exact absolute values of the 120-th column of the inverse of two tridiagonal matrices and the corresponding bounds of Theorem 3.9 and Theorem 3.18. While the bound of Theorem 3.18 perfectly captures the decay in A^{-1}, the decay rate (3.12) is too large compared to the actual decay rate. These examples show, that it is profitable to use the special structure of the spectrum of A.

Now we consider the result of Theorem 3.10 for matrices of the form (3.14). Note that this class of matrices coincides with the class of normal matrices with spectrum on a line segment, so Theorem 3.10 and Theorem 3.18 are intended for the same types of matrices. At the same time the decay rates given in these theorems coincide, which can be seen as follows.

The decay rate of Theorem 3.10 is given by q_1^{-1} where $q_1 > 1$ is defined as the solution of

$$\frac{1}{2}(z + z^{-1}) := \frac{|\lambda_1| + |\lambda_2|}{|\lambda_1 - \lambda_2|} =: x_1 \text{ with } z > 1,$$

whereas in the decay rate q_2^{-1} of Theorem 3.18, the value of $q_2 > 1$ can be written as the absolute value of the solution of

$$\frac{1}{2}(z + z^{-1}) = \frac{\lambda_1 + \lambda_2}{\lambda_2 - \lambda_1} =: x_2 \text{ with } |z| > 1. \tag{3.34}$$

Now, the condition $q_1 = q_2$ means that q_1 and the solution of (3.34) lie on the same circle C_{q_1} with radius q_1 which is equivalent to the condition that x_1 and x_2 lie on the same ellipse E with focal points -1 and 1, where E is the Joukowski mapping of C_{q_1}. This is equivalent to

$$|x_2 - 1| + |x_2 + 1| = 2x_1,$$

which is obviously true, so the decay rates q_1 and q_2 coincide. Hence we have no improvement here with respect to the decay rate, but with the approach of Section 3.2.2 we obtain a much simpler constant which will especially be useful for the results in Section 3.3. In addition we provided a detailed discussion and comparison of the decay rate to previous results, which was not given in [35].

In Theorem 3.19 we considered special types of shifted skew-Hermitian matrices, which e.g., represents an important class of matrices arising in lattice QCD (an example of such a matrix is given in Section 3.3.3). For those types of matrices the bounds of Theorem 3.18 can be used as well, hence, in the following we will check if it is advantageous to use this more extensive knowledge of the structure of the spectrum. For this, we compare the decay rate of Theorem 3.18 (which is equal to the decay rate from Theorem 3.10) when applied to those shifted skew Hermitian matrices to the decay rate from Theorem 3.19. Recall that the former can be written as q_2^{-1} where q_2 the solution of (3.34) whereas the latter is given by

$$q_3 = \left(\sqrt{x_3} + \sqrt{x_3 + 1}\right)^{-1} < 1 \text{ with } x_3 = \frac{a^2 + b_1^2}{b_2^2 - b_1^2}, \tag{3.35}$$

At first glance it might seem that the result of Theorem 3.18 yields a sharper bound as the decay rate q_2^{-1} does not depend on the distance of the spectrum of A to the real axis, i.e., on the conditioning of the skew-Hermitian matrix S.

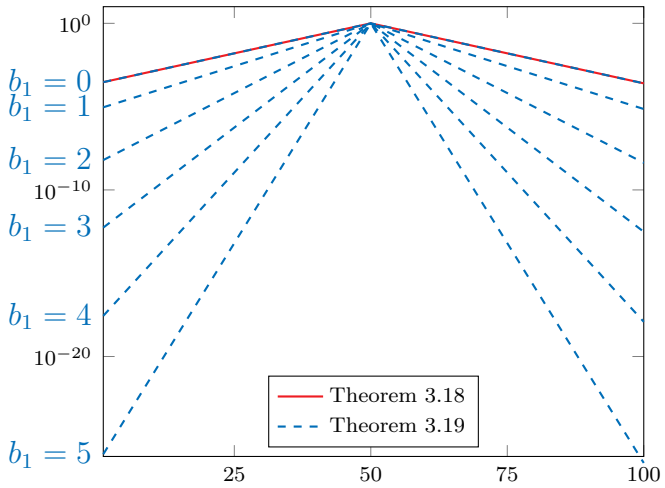

Figure 3.4: Decay rates predicted by Theorem 3.18 and Theorem 3.19 for a shifted skew-Hermitian matrix of dimension $n = 100$ with spectrum contained in the line segment $E = 1 + (\mathbf{i}[-6, -b_1] \cup \mathbf{i}[b_1, 6])$ for values of b_1 ranging from 0 to 5.

Interestingly, however, the opposite is the case: If S is non-singular, i.e., $b_1 > 0$, the bound of Theorem 3.19 gives the better decay rate, i.e.,

$$q_3 < q_2^{-1}.$$

To explore this, we set $\lambda_1 = a + b_2\,\mathbf{i}$ and $\lambda_2 = a - b_2\,\mathbf{i}$. Hence

$$x_2 = \frac{\lambda_1 + \lambda_2}{\lambda_2 - \lambda_1} = \frac{a}{b_2}\mathbf{i}$$

and therefore the solution of (3.34) is given by

$$z = \left(\frac{a}{b_2} + \sqrt{\left(\frac{a}{b_2}\right)^2 + 1}\right)\mathbf{i}$$

and we obtain

$$q_2 = |z| = \frac{a}{b_2} + \sqrt{\left(\frac{a}{b_2}\right)^2 + 1} = \sqrt{|x_2|^2} + \sqrt{|x_2|^2 + 1}.$$

Since $|x_2|^2 \leq x_3$ with equality if and only if $b_1 = 0$, this shows $q_3 < q_2^{-1}$ for $b_1 > 0$ and $q_2 = q_2^{-1}$ for $b_1 = 0$. In conclusion we find that the smaller the gap b_1 around the real axis is, the closer the decay rate of Theorem 3.18 is to that of Theorem 3.19. This is illustrated in Figure 3.4, in which we compare the predicted decay for a shifted skew-Hermitian matrix with $a = 1, b_2 = 6$ and different values of b_1.

As a last comparison for matrices with spectrum on a line segment, we consider the Hermitian indefinite case which is treated in Corollary 3.21. As the results of Theorem 3.18 and Theorem 3.10 are not applicable here, we compare our result to the general result of Theorem 3.9. Both results give the same decay rate, so that our new result does not give an improvement in this respect. It, however, yields a better constant. In Theorem 3.9, the constant for β-banded matrices is given by

$$C_1 = (2\beta+1)q_1^{-2}\|A^{-1}\|_2\kappa(A)\max\left\{1, \left(\frac{1+\kappa(A)}{\kappa(A)}\right)^2/2\right\},$$

while in the new bound from Corollary 3.21 it is given by

$$C = 2\|A^{-1}\|_2.$$

Therefore, their ratio is

$$\frac{C}{C_1} = \frac{q_1^2}{(\beta+\frac{1}{2})\kappa(A)\max\{1, (\frac{1+\kappa(A)}{\kappa(A)})^2/2\}} < 1.$$

For large condition numbers $\kappa(A)$, the terms q_1^2 and $\max\{1, (\frac{1+\kappa(A)}{\kappa(A)})^2/2\}$ both tend to one, so that in this case the ratio between the constants approximately becomes

$$\frac{C}{C_1} \approx \frac{1}{(\beta+\frac{1}{2})\kappa(A)},$$

showing that the constant C of Theorem 3.18 is smaller by a factor which depends both on the bandwidth *and* the condition number of A.

We concluded Section 3.2.2 with the observation that the use of normalized Faber polynomials leads to additional results for general matrices. As an example, we presented Proposition 3.23 for matrices with spectrum or field of values on a disk excluding the origin. As an additional comparison, we consider the result of Corollary 3.8 (which is a consequence of Theorem 3.6 from [78]) for general analytic functions applied to $f(z) = z^{-1}$. For non-normal matrices we obtained in Proposition 3.23 the bound

$$|(A^{-1})_{ij}| \leq \frac{1+\sqrt{2}}{\min_{z\in D(z_0,r)}|z|} \cdot \left(\frac{r}{|z_0|}\right)^{d(i,j)}, \tag{3.36}$$

where $D(z_0, r)$ is a disk containing the field of values of A. On the other hand, the bound of Corollary 3.8 is given by

$$\left|(A^{-1})_{i,j}\right| \leq \frac{2\tau}{\tau-1}\frac{1}{\min_{z\in D(z_0,r\tau)}|z|}\left(\frac{1}{\tau}\right)^{\mathrm{d}(i,j)}, \tag{3.37}$$

with $1 < \tau < |z_0|/r$, where the upper bound for τ follows from the singularity of the inverse in 0. So we have a similar situation as in the comparison of the bounds of Corollary 3.7 and Theorem 3.18 for matrices with spectrum on a line segment. There is a trade-off in the choice of τ in (3.37). For the decay rate it would be advantageous to choose τ as large as possible, but then the constant tends to infinity. For the second factor of the constant it would be optimal to choose τ as small as possible but then the decay rate tends to one and the first factor of the constant tends to infinity. So again, the bound (3.36) is similar to the bound (3.37) with optimal choices of τ independently for every factor. The additional fact that $1 + \sqrt{2} < 2\frac{\tau}{\tau-1}$ shows, that for any choice of τ the bound (3.36) is sharper than (3.37).

We conclude this section with a numerical example, which reveals an important drawback of the bounds developed so far.

Typically, the quality of decay bounds is illustrated by showing the actual decay in one row or column of the inverse and comparing it to the bound for this row or column. In the example below, we instead compare the whole matrix A^{-1} to a matrix containing the bounds for every entry $|[A^{-1}]_{i,j}|$. For example, if we apply the bounds of Theorem 3.9 to a tridiagonal matrix, then collecting all the bounds (3.9) in a matrix Q yields the symmetric Toeplitz structure

$$Q = C \cdot \begin{bmatrix} q^0 & q^1 & \cdots & \cdots & q^{n-1} \\ q^1 & q^0 & q^1 & \ddots & \vdots \\ \vdots & q_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & q^1 \\ q^{n-1} & \cdots & \cdots & q^1 & q^0 \end{bmatrix}, \tag{3.38}$$

which, by Theorem 3.9, fulfills

$$|A^{-1}| \leq Q, \tag{3.39}$$

where "$\leq$" and "$|\cdot|$" are understood component-wise. Similarly, all the results on decay bounds for banded matrices from [8, 11, 12, 35, 38, 71, 72, 76] also result in a Toeplitz structured matrix Q of bounds since for all entries with the same distance between the nodes the same bound is used. However, even when A is a banded Toeplitz matrix, A^{-1} is in general not Toeplitz which indicates that the bound (3.39) will typically not bound all the entries of A^{-1} equally well when going along a specific sub- or superdiagonal. We now give an example where this effect is particularly pronounced.

Example 3.25. We construct a symmetric, tridiagonal matrix with prescribed spectrum following the construction principle developed in [62, Section 6.1]. The idea is to start with a diagonal matrix containing the prescribed eigenvalues, and then to apply a series of (two-sided) Givens rotations in order to introduce

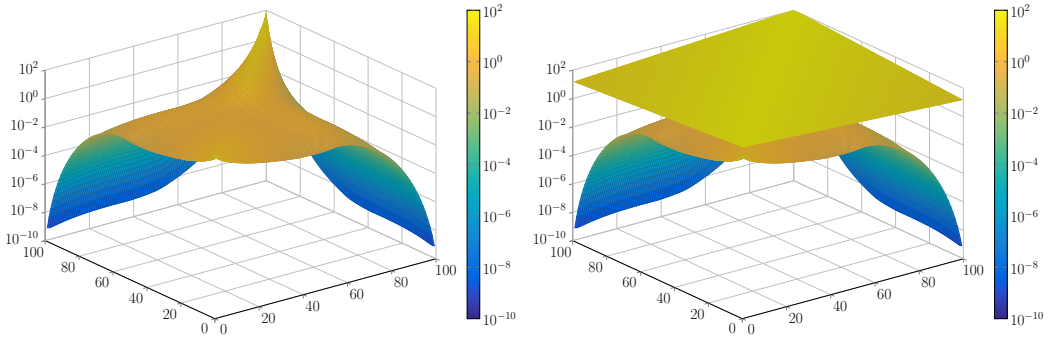

Figure 3.5: Magnitude of the entries of A^{-1} (left) and Q from (3.38) (right), where A is the matrix from Example 3.25.

nonzero elements on the sub- and superdiagonal. Nonzero elements that are introduced outside the band are immediately chased off the bottom right corner of the matrix, similarly to Schwarz band reduction [87]. Specifically, we construct a symmetric positive definite tridiagonal matrix $A \in \mathbb{R}^{100\times 100}$ with logarithmically spaced eigenvalues in the interval $[10^{-2}, 10^{2}]$. Figure 3.5 shows the magnitude of the entries of A^{-1} and of the Toeplitz matrix (3.38) containing the bounds from Theorem 3.9. Two problems with the bound are apparent: On the one hand, the magnitude of the entries is very severely overestimated due to a super-exponential decay in A^{-1}, and on the other hand, the structure of A^{-1} is far from being Toeplitz, in contrast to the matrix Q. ◇

In the next section we develop a theoretical framework for explaining and accurately predicting a decay behavior as observed in the example above for the Hermitian positive definite case.

3.2.4 Bounds with non-Toeplitz structure

In this section we want to find alternative ways to formulate decay bounds which capture the actual decay of a matrix function more accurately, even in extreme situations as given in Example 3.25.

First, we show how to obtain a family of decay bounds based on full spectral information of A, in contrast to the bound in Theorem 3.9 which is only based on the largest and smallest eigenvalue, i.e., the spectral interval. Although this information is typically not available in practical situations, it will help us to make a first step towards explaining the decay behavior observed for the matrix A from Example 3.25. As these bounds turn out to be still not necessarily accurate, we improve them further by relating the decay above and below the diagonal

of the kth column of A^{-1} to the eigenvalues of (slightly modified) $k \times k$ and $(n-k) \times (n-k)$ submatrices of A, respectively. By combining both approaches, we obtain sharp bounds for the entries of A^{-1}, even in extreme situations as the one from Example 3.25.

The decay bounds from Theorem 3.9 are obtained by using $\Omega = [\lambda_{\min}(A), \lambda_{\max}(A)]$ in the approach outlined in Section 3.1 and then using the best polynomial approximation for the inverse on a real positive interval. A drawback of the bounds of Theorem 3.9 is that they cannot accurately capture the actual decay behavior if the decay is super-exponential. This problem is comparable to what one observes for the classical textbook convergence bound for the conjugate gradient method; see, e.g., [85]: While for a given condition number one can always find a matrix such that the bound for step k is sharp, it is typically neither sharp for other matrices with the same condition number, nor for other steps of the iteration. In the same way, one cannot expect the decay bounds from Theorem 3.9 to be sharp for all matrices with a given condition number, or even just for all entries of one specific matrix. In particular, the classical CG convergence result only predicts linear convergence (and similarly, the bound of Theorem 3.9 only predicts exponential decay), while in practice one often observes super-linear convergence due to spectral adaptation. A simple approach for explaining the super-linear CG convergence is based on bounding the iteration polynomial from the CG method by other, so-called *composite* polynomials, which leads to a family of bounds. This approach is described in detail in [65, Chapter 5.6.4]. To transfer this approach, we cannot use the bounds of Theorem 3.9. Instead, we use the approach which leads to the bound (3.33) from Section 3.2.3 for Hermitian positive definite matrices, which gives the same decay rate as in Theorem 3.9 and a constant which, in the worst case, deteriorates by a factor of two. The advantage is that with this approach we may use the same idea as for the CG convergence analysis to explain a super-exponential convergence behavior. Altogether this gives the following family of bounds based on the *effective condition number*. This result is already published in [39].

Theorem 3.26. *Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian positive definite and β-banded matrix with eigenvalues $\lambda_{\min}(A) = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = \lambda_{\max}(A)$. Further, let us define*

$$\kappa_\ell(A) = \frac{\lambda_{n-\ell}}{\lambda_1}, \quad q_\ell = \frac{\sqrt{\kappa_\ell(A)} - 1}{\sqrt{\kappa_\ell(A)} + 1} \text{ and } C = \frac{2}{\lambda_1}. \tag{3.40}$$

Then the entries of A^{-1} can be bounded as

$$|[A^{-1}]_{i,j}| \leq C\, q_\ell^{\frac{|i-j|}{\beta} - \ell} \text{ for all } \ell = 0, 1, \ldots, \left\lfloor \frac{|i-j|}{\beta} \right\rfloor. \tag{3.41}$$

Proof. Instead of bounding the entries of the inverse by using the polynomial approximation on an interval containing the spectrum of A directly, we now first

work with the discrete set $\sigma(A) = \{\lambda_1, \dots, \lambda_n\}$. Because of (3.4) with $m = \mathrm{d}(i,j) - 1 = \lceil |i-j|/\beta \rceil - 1$ we can bound the entries of the inverse as

$$\begin{aligned}
|[A^{-1}]_{i,j}| &\leq \min_{p_m \in \mathbb{P}_m} \max_{z \in \{\lambda_1,\dots,\lambda_n\}} |z^{-1} - p_m(z)| \\
&= \min_{\substack{P_{m+1} \in \mathbb{P}_{m+1} \\ P_{m+1}(0)=1}} \max_{z \in \{\lambda_1,\dots,\lambda_n\}} \left| \frac{P_{m+1}(z)}{z} \right| \\
&\leq \frac{1}{\lambda_1} \min_{\substack{P_{m+1} \in \mathbb{P}_{m+1} \\ P_{m+1}(0)=1}} \max_{z \in \{\lambda_1,\dots,\lambda_n\}} |P_{m+1}(z)| \\
&\leq \frac{1}{\lambda_1} \min_{\substack{P_{m+1-\ell} \in \mathbb{P}_{m+1-\ell} \\ P_{m+1-\ell}(0)=1}} \max_{z \in \{\lambda_1,\dots,\lambda_n\}} |r_\ell(z) P_{m+1-\ell}(z)|
\end{aligned}$$

where

$$r_\ell(z) = \prod_{i=n-l+1}^{n} \left(1 - \frac{z}{\lambda_i}\right)$$

is a polynomial which satisfies $r_\ell(0) = 1$, $r_\ell(\lambda_i) < 1$ for $i = 1, \dots, n-\ell$ and $r_\ell(\lambda_i) = 0$ for $i = n-\ell+1, \dots, n$. Therefore, the entries of the inverse can further be bounded as

$$\begin{aligned}
|[A^{-1}]_{i,j}| &\leq \frac{1}{\lambda_1} \min_{\substack{P_{m+1-\ell} \in \mathbb{P}_{m+1-\ell} \\ P_{m+1-\ell}(0)=1}} \max_{z \in \{\lambda_1,\dots,\lambda_{n-\ell}\}} |r_\ell(z) P_{m+1-\ell}(z)| \\
&\leq \frac{1}{\lambda_1} \max_{z \in \{\lambda_1,\dots,\lambda_{n-\ell}\}} |r_\ell(z)| \min_{\substack{P_{m+1-\ell} \in \mathbb{P}_{m+1-\ell} \\ p_{m+1-\ell}(0)=1}} \max_{z \in \{\lambda_1,\dots,\lambda_{n-\ell}\}} |P_{m+1-\ell}(z)| \\
&\leq \frac{1}{\lambda_1} \min_{\substack{P_{m+1-\ell} \in \mathbb{P}_{m+1-\ell} \\ P_{m+1-\ell}(0)=1}} \max_{z \in [\lambda_1,\lambda_{n-\ell}]} |P_{m+1-\ell}(z)| \\
&= \frac{1}{\lambda_1} \left(T_{m+1-\ell} \left(\frac{\lambda_{n-\ell} + \lambda_1}{\lambda_{n-\ell} - \lambda_1} \right) \right)^{-1}
\end{aligned}$$

where $T_{m+1-\ell}$ is the Chebyshev polynomial of degree $m+1-\ell$. Again, by using Lemma 2.15 and equation (3.32) with $\lambda_{\min} = \lambda_1$ and $\lambda_{\max} = \lambda_{n-\ell}$ we obtain

$$\max_{z \in [\lambda_1, \lambda_{n-\ell}]} |T_{m+1-\ell}(z)| \leq 2\, q_\ell^{m+1-\ell},$$

so that the bound (3.41) follows by using $|i-j|/\beta \leq m+1$. □

The family (3.41) of bounds can potentially predict the decay behavior in A^{-1} much more accurately than (3.9) – which is contained as a special case for $\ell = 0$ (except for a factor of two in the constant) – if one chooses the value of ℓ which minimizes (3.41) for each entry $[A^{-1}]_{i,j}$.

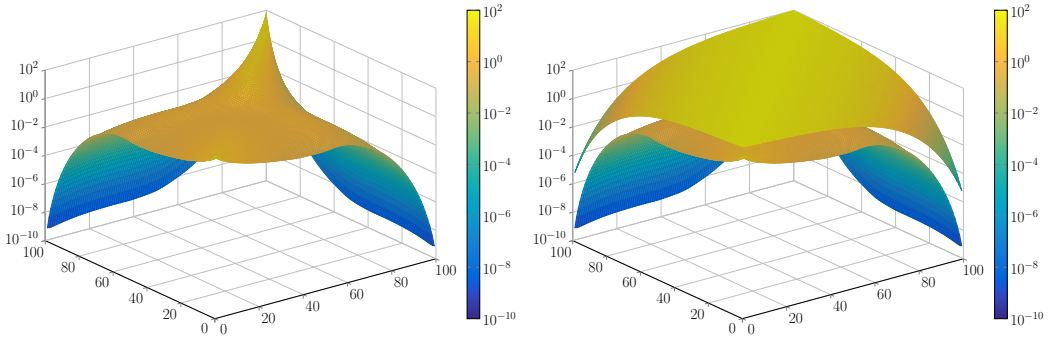

Figure 3.6: Magnitude of the entries of A^{-1} (left) and Q from (3.42) (right), where A is the matrix from Example 3.25.

Example 3.27. We consider the matrix from Example 3.25 again, but this time define a matrix Q of bounds such that

$$Q_{i,j} = \min_{\ell=0,\dots,\left\lfloor \frac{|i-j|}{\beta} \right\rfloor} C q_\ell^{\frac{|i-j|}{\beta}-\ell}, \tag{3.42}$$

where q_ℓ and C are defined as in (3.40). Note that the choice of ℓ which minimizes the right-hand side of (3.42) depends on the distribution of the eigenvalues of A. A larger number ℓ improves the effective condition number and therefore the decay rate q_ℓ, but at the same time the exponent $\frac{|i-j|}{\beta} - \ell$ decreases. Therefore, the distribution of the eigenvalues determines whether the improvement of the decay rate outweighs the smaller exponent.

The magnitude of the entries of the resulting matrix Q is given in Figure 3.6 (together with the magnitude of the absolute entries of A^{-1} for comparison). In contrast to what we observed in Figure 3.5, the entries are not overestimated as much as before (though still by several orders of magnitude), and at least in the first few rows and columns (which correspond to the front part of the plotted surfaces), the *qualitative decay behavior* is predicted quite accurately. For the other rows and columns, however, the decay behavior predicted by the bounds is still not satisfactory. To better understand why this is the case, we take a closer look at the decay behavior in two individual columns in Figure 3.7. On the left-hand side, the magnitude of the entries of the first column of A^{-1} and the corresponding bounds are given, and on the right-hand side, the same information is shown for the 50th column. The plot on the right-hand side nicely illustrates the main problem that still occurs with (3.42): While the bounds obtained by (3.42) give a better idea of the actual decay behavior, they still lead to a symmetric Toeplitz structure of Q. As a consequence, the bound for the 50th column in the right part of Figure 3.7 is symmetric with respect to the 50th entry, and the bounds for the first 50 entries of the first column plotted in the left part of Figure 3.7 agree with the bounds

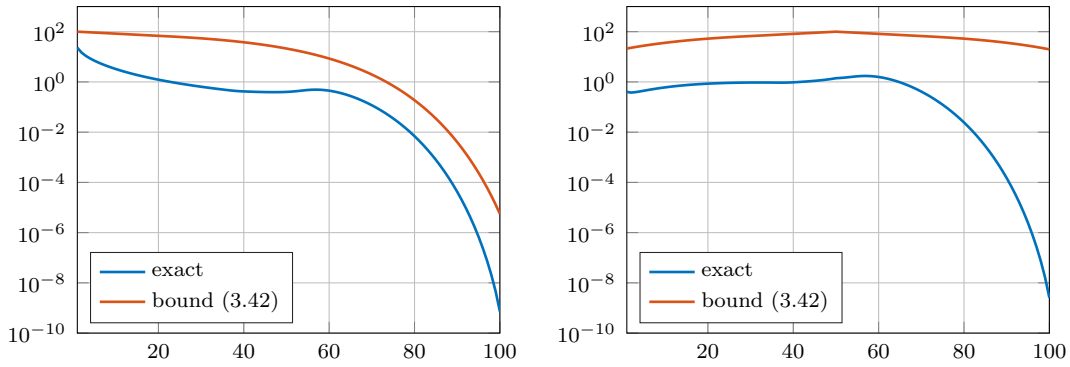

Figure 3.7: Magnitude of the entries of the first (left) and 50th (right) column of A^{-1} and corresponding bounds (3.42), where A is the matrix from Example 3.25.

for entries 51–100 of the 50th column. This means in particular that the decay predicted "above" the diagonal (i.e., for $i < j$) is the same as the decay predicted "below" the diagonal (i.e., for $i > j$), although the actual decay is very different for these two parts of the 50th column. Whenever this is the case, *any* symmetric Toeplitz-structured bound has this shortcoming as it must be valid for both parts of the column, and will thus be forced to follow the part of the column with the "slower decay". ◇

Theorem 3.26 can immediately be generalized to Hermitian and positve definite matrices with an arbitrary, not necessarily banded, sparsity pattern by replacing the quantity $|i-j|/\beta$ by the distance $\mathrm{d}(i,j)$ of the nodes i and j. Everything then works in a completely analogous manner and we can state the following result without proof, which is also published in [39].

Theorem 3.28. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and let $\lambda_{\min}(A) = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = \lambda_{\max}(A)$ be the eigenvalues of A. Let $\kappa_\ell(A), q_\ell$ and $C = \frac{2}{\lambda_1}$ be defined as in Theorem 3.26. Then the entries of A^{-1} can be bounded as*

$$|[A^{-1}]_{i,j}| \leq C\, q_\ell^{d(i,j)-\ell} \text{ for all } \ell = 0, 1, \ldots, d(i,j). \tag{3.43}$$

Motivated by Example 3.27, we now proceed to show how to obtain bounds that are not restricted to a Toeplitz structure any longer. In doing so we use the bound of Theorem 3.26 for submatrices of A. We again first consider the tridiagonal case. The basic idea is to perform a rank-one modification of a tridiagonal matrix A that reduces it to block diagonal form (similar to what is done in the divide and conquer algorithm for the symmetric tridiagonal eigenvalue problem [51]). By applying the Sherman–Morrison formula [93], the inverse of A can then be written as the sum of the inverse of a block diagonal matrix and a rank-one matrix. The

result then follows from the fact that *both* the inverse of the block diagonal matrix *and* the rank-one term exhibit off-diagonal decay.

We fix $k \in \{1, \dots, n-1\}$ and decompose

$$A = \begin{bmatrix} A_{11} & A_{22} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix} + uu^H, \quad u = \alpha(e_k + \frac{a_{k+1,k}}{|a_{k+1,k}|} e_{k+1}) \in \mathbb{C}^n, \quad (3.44)$$

with $\alpha = \sqrt{|a_{k+1,k}|}$, e_k, e_{k+1} the kth and $k+1$st canonical unit vector in $\mathbb{C}^n$. Note that $B_1 \in \mathbb{C}^{k \times k}$ is tridiagonal and that it differs from A_{11} only in its (k,k) entry, which is $[B_1]_{k,k} = a_{k,k} - |a_{k+1,k}|$. Similarly, $B_2 \in \mathbb{C}^{(n-k)\times(n-k)}$ is tridiagonal too, and it differs from A_{22} only in its $(1,1)$ entry, which is $[B_2]_{1,1} = a_{k+1,k+1} - |a_{k+1,k}|$.

Theorem 3.29. *Let $A \in \mathbb{C}^{n\times n}$ be a tridiagonal Hermitian positive definite matrix. Assume that B_1 and B_2 in (3.44) are positive definite, and define for $s = 1, 2$*

$$\kappa_s = \frac{\lambda_{\max}(B_s)}{\lambda_{\min}(B_s)}, \quad q_s = \frac{\sqrt{\kappa_s} - 1}{\sqrt{\kappa_s + 1}}, \quad c_s = \frac{2}{\lambda_{\min}(B_s)}.$$

Then the entries of A^{-1} can be bounded as

$$|[A^{-1}]_{i,j}| \le \begin{cases} c_1\, q_1^{|i-j|} + c_1^2\, \widetilde{c}\, q_1^{2k-j-i} & \text{for } i,j \le k \\ c_2\, q_2^{|i-j|} + c_2^2\, \widetilde{c}\, q_2^{i+j-2(k+1)} & \text{for } i,j > k \\ c_1\, c_2\, \widetilde{c}\, q_1^{k-i}\, q_2^{j-k-1} & \text{for } i \le k < j \\ c_1\, c_2\, \widetilde{c}\, q_1^{j-k}\, q_2^{i-k-1} & \text{for } j \le k < i \end{cases}$$

with the constant

$$\widetilde{c} = \frac{|a_{k+1,k}|}{1 + |a_{k+1,k}| \left(\frac{1}{\lambda_{\max}(B_1)} + \frac{1}{\lambda_{\max}(B_2)} \right)} \le |a_{k+1,k}|.$$

Proof. We set $B = \operatorname{diag}(B_1, B_2)$. The Sherman–Morrison formula gives

$$A^{-1} = B^{-1} - \frac{B^{-1} u u^H B^{-1}}{1 + u^H B^{-1} u} =: B^{-1} - R.$$

Obviously, we thus have

$$|[A^{-1}]_{i,j}| \le |[B^{-1}]_{i,j}| + |R_{i,j}|,$$

and as the inverse of B is given by $B^{-1} = \operatorname{diag}(B_1^{-1}, B_2^{-1})$, we can use Theorem 3.26 for $\ell = 0$ applied to B_1 and B_2 in order to bound the entries of B^{-1} as

$$|[B^{-1}]_{i,j}| = \begin{cases} |[B_1^{-1}]_{i,j}| \le c_1\, q_1^{|i-j|} & \text{for } i,j \le k \\ |[B_2^{-1}]_{i,j}| \le c_2\, q_2^{|i-j|} & \text{for } i,j > k \\ 0 & \text{otherwise} \end{cases}.$$

Denoting with $M_{\bullet,\ell}$ the ℓ-th column of a matrix M, we get

$$B^{-1}u = \alpha \begin{bmatrix} [B_1^{-1}]_{\bullet,k} \\ \frac{a_{k+1,k}}{|a_{k+1,k}|}[B_2^{-1}]_{\bullet,1} \end{bmatrix},$$

which shows that the absolute values of the entries of the rank-one term R are given by

$$|R_{i,j}| = \frac{\alpha^2}{1+u^H B^{-1}u} \cdot \begin{cases} |[B_1^{-1}]_{i,k}| \cdot |[B_1^{-1}]_{j,k}| & \leq c_1^2\, q_1^{k-i}\, q_1^{k-j} \quad \text{for } i,j \leq k \\ |[B_2^{-1}]_{i-k,1}| \cdot |[B_2^{-1}]_{j-k,1}| & \leq c_2^2\, q_2^{i-k-1}\, q_1^{j-k-1} \quad \text{for } i,j > k \\ |[B_1^{-1}]_{i,k}| \cdot |[B_2^{-1}]_{j-k,1}| & \leq c_1\, c_2\, q_1^{k-i}\, q_2^{j-k-1} \quad \text{for } i \leq k < j \\ |[B_1^{-1}]_{j,k}| \cdot |[B_2^{-1}]_{i-k,1}| & \leq c_1\, c_2\, q_1^{k-j}\, q_2^{i-k-1} \quad \text{for } j \leq k < i \end{cases},$$

where we applied the bounds for the inverse also to these terms. Due to the relation

$$u^H B^{-1} u = \alpha^2 \left([B_1^{-1}]_{k,k} + [B_2^{-1}]_{1,1} \right),$$

we further find

$$1 + u^H B^{-1} u \geq 1 + \alpha^2 \left(\frac{1}{\lambda_{\max}(B_1)} + \frac{1}{\lambda_{\max}(B_2)} \right).$$

Putting all these inequalities together gives the desired result. □

This theorem is also published in [39]. We discuss Theorem 3.29 in the following remarks.

Remark 3.30. The result of Theorem 3.29 is based on Theorem 3.26 with $\ell = 0$ for ease of presentation. It is possible to rewrite it in the spirit of (3.42) to obtain sharper decay bounds. We will state and illustrate the resulting bounds in Example 3.33 below, but refrain from giving a formal proof, because it is essentially the same as that of Theorem 3.29. ◇

Remark 3.31. A crucial assumption in Theorem 3.29 is that B_1 and B_2 are positive definite. One situation in which this assumption is guaranteed to be fulfilled is when A is strictly diagonally dominant, as in this case B_1 and B_2 inherit this property and must therefore also be positive definite. ◇

Remark 3.32. Theorem 3.29 only applies to the case of tridiagonal matrices, i.e., matrices with bandwidth $\beta = 1$. To modify it in order to account for matrices with bandwidth $\beta > 1$, one can use a rank-β modification that again reduces it to block diagonal form and then proceed in an analogous manner. However, the quality of the bounds obtained this way will deteriorate more and more the larger the bandwidth is. Another possible generalization of Theorem 3.29 is to consider general sparse matrices, where we can find a low-rank modification of A such that the graph of the resulting matrix is disconnected. Then, by renumbering the nodes,

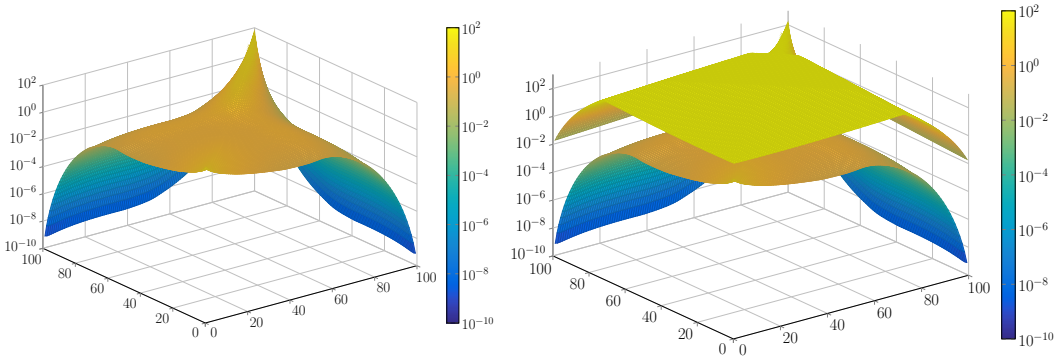

Figure 3.8: Magnitude of the entries of A^{-1} (left) and the bounds from (3.45) (right), where A is the matrix from Example 3.25.

we also obtain a decomposition which is given by the sum of a block diagonal matrix and a low-rank matrix. In this case, the exponent $|i-j|$ is replaced by $\mathrm{d}(i,j)$, the graph distance between the nodes i and j. Again, the resulting bounds can be combined with the super-exponential bounds for a general sparsity pattern, according to Theorem 3.28. ◇

Using Theorem 3.29 (modified accordingly to Remark 3.30), we are now in a position to compute bounds that accurately predict the decay behavior of the inverse for the matrix A from Example 3.25.

Example 3.33. In this example, we illustrate the bounds arising from the block-partitioning approach of Theorem 3.29. As Theorem 3.29 gives possibly different bounds for an entry of $|[A^{-1}]_{i,j}|$ for each value of k, the best possible bounds are obtained by computing bounds for every $k = 1, \dots, n-1$ and then taking the smallest among those bounds, i.e.,

$$|[A^{-1}]_{i,j}| \leq \min_{k=1,\dots,n-1} Q_{i,j}^{(k)}, \tag{3.45}$$

where

$$Q_{i,j}^{(k)} = \begin{cases} c_1\, q_1^{|i-j|} + c_1^2\, \widetilde{c}\, q_1^{2k-j-i} & \text{for } i,j \leq k \\ c_2\, q_2^{|i-j|} + c_2^2\, \widetilde{c}\, q_2^{i+j-2(k+1)} & \text{for } i,j > k \\ c_1\, c_2\, \widetilde{c}\, q_1^{k-i}\, q_2^{j-k-1} & \text{for } i \leq k < j \\ c_1\, c_2\, \widetilde{c}\, q_1^{j-k}\, q_2^{i-k-1} & \text{for } j \leq k < i \end{cases}. \tag{3.46}$$

Note that all quantities on the right-hand side of (3.46) depend on k. The resulting matrix of bounds (3.45) arising for the matrix A from Example 3.25 is illustrated on the right-hand side of Figure 3.8 with the magnitude of the entries of A^{-1} given on the left-hand side again. We observe a considerable improvement compared to the bounds (3.42) shown in Figure 3.6. In particular, the matrix of bounds is not

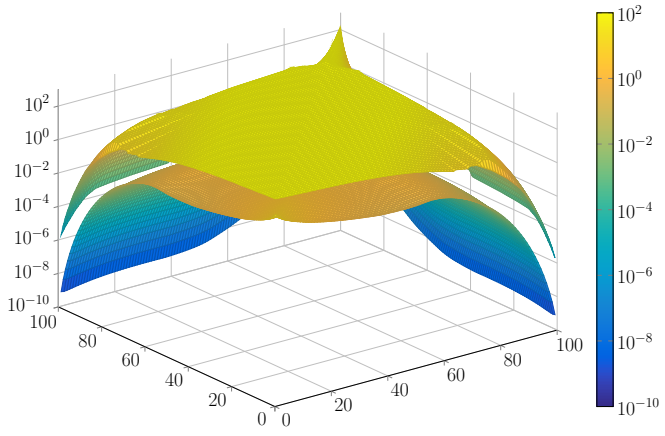

Figure 3.9: Bounds for A^{-1} obtained by combining Theorem 3.26 with Theorem 3.29, where A is the matrix from Example 3.25.

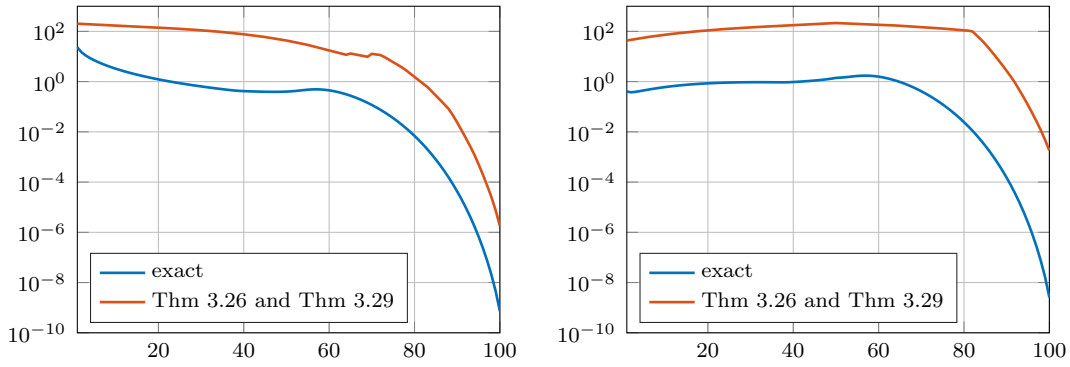

Figure 3.10: Magnitude of the entries of the first (left) and 50th (right) column of A^{-1} and corresponding bounds obtained by combining Theorem 3.26 with Theorem 3.29, where A is the matrix from Example 3.25.

a Toeplitz matrix any longer and can thus better capture the different decay rates above and below the diagonal.

Further improvements can be obtained by combining the approach of Theorem 3.29 with that of Theorem 3.26, i.e., replacing the bounds in (3.45) by bounds involving the *effective* condition numbers of the diagonal blocks B_1 and B_2 for suitable values of ℓ. The bounds that are obtained this way (when always selecting the best possible value ℓ for each entry $|[A^{-1}]_{i,j}|$) are depicted in Figure 3.9. These bounds resemble the actual decay behavior in A^{-1} even better, and in order to allow another comparison to the results presented in Example 3.27, we show a comparison of the exact values of the first and 50th column of A^{-1} with our bounds in Figure 3.10. While still overestimating the entries by about two orders

of magnitude, the qualitative decay behavior is resolved quite well by these new bounds. ◇

Remark 3.34. In contrast to most bounds from the literature, these bounds can predict a super-exponential decay behavior and do not obey a Toeplitz structure, hence, they more accurately predict the actual decay in A^{-1}. However, the bounds in the presented form are not meant to be used for practical computations, as they require complete spectral information on several submatrices of A. ◇

3.3 Bounds for functions defined by an integral transform

So far we directly obtained decay bounds for a matrix $f(A)$ by using the error of a polynomial approximation of f on a set containing the spectrum of A. Another approach is given by using an integral representation of f and by applying known, sharp decay bounds for the functions arising in the integrand in order to obtain sharp bounds for $f(A)$ this way. For example, based on Definition 2.3 of matrix functions, the entries of a matrix $f(A)$ can be written as

$$[f(A)]_{ij} = \frac{1}{2\pi\mathbf{i}} \int_\Gamma f(t)\, [(tI - A)^{-1}]_{ij}\, \mathrm{d}t$$

for every function f that is analytic on and inside a closed contour Γ that encloses $\sigma(A)$. Thus, the entries can by bounded by

$$|[f(A)]_{ij}| \leq \frac{1}{2\pi\mathbf{i}} \int_\Gamma |f(t)|\, |[(tI - A)^{-1}]_{ij}|\, \mathrm{d}t.$$

Now, by using decay bounds for $|[(tI - A)^{-1}]_{ij}|$ we immediately obtain proper (but also only implicit and not practical) bounds for a large class of functions.

In the following we will use this approach for special types of functions which can be expressed as an integral transform and where decay bounds for functions in the integrand were already obtained. In some cases, we even obtain explicit bounds, i.e., bounds where no integral appears, by further bounding the resulting integral. Mainly, we will consider Cauchy–Stieltjes functions which are defined as functions $f : \mathbb{C} \setminus \mathbb{R}_0^- \to \mathbb{C}$ with

$$f(z) = \int_0^\infty \frac{\mathrm{d}\mu(\tau)}{z + \tau}, \tag{3.47}$$

where μ is a monotonically increasing, real-valued and non-negative function on $[0, \infty)$. Hence, the bounds for the inverse obtained in Section 3.2 can be used to

obtain decay bounds for Cauchy–Stieltjes functions of matrices. Interestingly, the inverse itself is a Cauchy–Stieltjes function, generated by the step function

$$\mu(t) = \begin{cases} 0 & \text{for } t = 0 \\ 1 & \text{for } t > 1. \end{cases}$$

Other important examples of functions which can be expressed by an integral of the form (3.47) for $z \in \mathbb{C} \setminus \mathbb{R}_0^-$ are given by

$$z^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi} \int_0^\infty \frac{t^{-\alpha}}{z+t} \mathrm{d}t$$

for $\alpha \in (0, 1)$ and

$$\frac{\log(1+z)}{z} = \int_1^\infty \frac{t^{-1}}{z+t} \mathrm{d}t.$$

This approach was already used in [12] to obtain bounds for Cauchy–Stieltjes and Laplace–Stieltjes functions of banded, Hermitian positive definite matrices. Laplace–Stieltjes functions can be defined by the integral representation

$$f(z) = \int_0^\infty \exp(-\tau z) \mathrm{d}\alpha(\tau)$$

for $z \in \mathbb{C}$ with $\mathrm{Re}(z) > 0$, where α is a real-valued, nondecreasing function. Examples of Laplace–Stieltjes functions are given by

$$\frac{1 - \exp(-z)}{z} = \int_0^1 \exp(-tz) \mathrm{d}t$$

and

$$z^{-\frac{1}{2}} = \int_0^\infty \frac{\exp(-tz)}{\sqrt{\pi t}} \mathrm{d}t,$$

where the representation of $z^{-\frac{1}{2}}$ can be seen by substituting $tz = x$ and using the identity

$$\int_0^\infty \frac{\exp(-x)}{\sqrt{x}} \mathrm{d}x = \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},$$

where Γ is the gamma function (see, e.g., [2]). In [12] decay bounds for the exponential where used to obtain decay bounds for the entries of Laplace–Stieltjes functions of Hermitian matrices.

3.3.1 Literature review

In this section we review decay bounds for Cauchy–Stieltjes and Laplace–Stieltjes functions of banded, Hermitian positive definite matrices using the approach outlined above. For Cauchy–Stieltjes functions, Theorem 3.9 from [23] can be used and we immediately obtain the following result from [12].

Theorem 3.35. *Let A be a β-banded and Hermitian positive definite matrix and let f be a Cauchy–Stieltjes function. Then the entries of $f(A)$ can be bounded by*

$$|[f(A)]_{ij}| \leq \int_0^\infty C(\tau) q(\tau)^{\frac{|i-j|}{\beta}} \, \mathrm{d}\mu(\tau) \tag{3.48}$$

with $q(\tau) = (\sqrt{\kappa(\tau)} - 1)/(\sqrt{\kappa(\tau)} + 1)$, $\kappa(\tau) = (\lambda_{\max} + \tau)/(\lambda_{\min} + \tau)$ and

$$C(\tau) = \max \left\{ \frac{1}{\lambda_{\min} + \tau}, \frac{\left(1 + \sqrt{\kappa(\tau)}\right)^2}{2(\lambda_{\max} + \tau)} \right\}. \tag{3.49}$$

In order to use these bounds in practice, the integral (3.48) has to be evaluated numerically and it is not even clear whether its value is finite. Hence, it would be advantageous to find a finite upper bound of (3.48) which shows the existence of the integral in the first place and which can be used as an explicit bound for the entries of $f(A)$. For the special case $f(z) = z^{-1/2}$ the bound

$$|[A^{-\frac{1}{2}}]_{ij}| \leq (C(0) + C_2) q(0)^{\frac{|i-j|}{\beta}} \tag{3.50}$$

with

$$C_2 = \frac{\left(1 + \sqrt{\kappa(0)}\right)^2}{2}$$

was derived in [12].

Using the same approach for Laplace–Stieltjes matrix functions, we first need results for the exponential. Based on convergence results of the Lanczos approximation of the exponential in [54], the following decay bounds for the exponential were derived in [12].

Theorem 3.36. *Let A be a Hermitian, β-banded matrix with eigenvalues in the interval $[0, 4\rho]$. Then for $i \neq j$ and $\xi = \lceil |i-j|/\beta \rceil$ we have*

$$|[\exp(-\tau A)]_{ij}| \leq \Phi(\xi, \rho)$$

with

$$\Phi(\xi, \rho) = \begin{cases} 10 \exp\left(-\frac{\xi^2}{5\rho\tau}\right) & \text{for } \sqrt{4\rho\tau} \leq \xi \leq 2\rho\tau \\ 10 \frac{\exp(-\rho\tau)}{\rho\tau} \left(\frac{\mathrm{e}\,\rho\tau}{\xi}\right)^\xi & \text{for } \xi \geq 2\rho\tau. \end{cases} \tag{3.51}$$

Note that this results holds for general Hermitian matrices, as for any Hermitian matrix A the spectrum of the matrix $A - \lambda_{\min} I$ is contained in an interval $[0, 4\rho]$. Thus, for $\exp(-\tau A)$ the bound of Theorem 3.36 holds with the additional factor $\exp(-\lambda_{\min}\tau)$ (since clearly the matrices A and $\lambda_{\min} I$ commute). Using the results for the exponential, the following results have been obtained in [12].

Theorem 3.37. *Let A be a Hermitian, β-banded matrix and let $\widehat{A} = A - \lambda_{\min} I$ be a matrix whose spectrum is contained in $[0, 4\rho]$. Then for a Laplace–Stieltjes function f and $\xi = \lceil |i-j|/\beta \rceil$ the entries of $f(A)$ can be bounded by*

$$\begin{aligned} |[f(A)]_{ij}| &\leq \int_0^\infty \exp(-\lambda_{\min}\tau) |[\exp(-\tau\widehat{A})]_{ij}| \mathrm{d}\alpha(\tau) \\ &\leq 10 \int_0^{\frac{\xi}{2\rho}} \exp(-\lambda_{\min}\tau) \frac{\exp(-\rho\tau)}{\rho\tau} \left(\frac{\mathrm{e}\,\rho\tau}{\xi}\right)^\xi \mathrm{d}\alpha(\tau) \\ &\quad + 10 \int_{\frac{\xi}{2\rho}}^{\frac{\xi^2}{4\rho}} \exp(-\lambda_{\min}\tau) \exp\left(-\frac{\xi^2}{5\rho\tau}\right) \mathrm{d}\alpha(\tau) \\ &\quad + \int_{\frac{\xi^2}{4\rho}}^\infty \exp(-\lambda_{\min}\tau) |[\exp(-\tau\widehat{A})]_{ij}]| \mathrm{d}\alpha(\tau). \end{aligned} \tag{3.52}$$

Note again that the results of Theorem 3.36 and Theorem 3.37 can be generalized to sparse matrices just by replacing ξ by $\mathrm{d}(i,j)$. However, in [12] bounds for the important class of matrices with Kronecker structure of the form

$$\mathcal{A} = A \oplus A := A \otimes I + I \otimes A \tag{3.53}$$

were derived explicitly, by using the relation

$$\exp(-\tau\mathcal{A}) = \exp(-\tau A) \otimes \exp(-\tau A). \tag{3.54}$$

Using relation (3.54) results in the following bounds for the exponential, where the indices i and j are written as the pairs of coordinates $i = (i_1, i_2)$ and $j = (j_1, j_2)$ in the related two-dimensional grid.

Theorem 3.38. *Let A be Hermitian with bandwidth β and spectrum contained in $[0, 4\rho]$. Let $\mathcal{A}$ be of structure* (3.53)*, then it holds*

$$[\exp(-\tau\mathcal{A})]_{ij} = [\exp(-\tau A)]_{i_1 j_1} [\exp(-\tau A)]_{i_2 j_2},$$

and therefore, for $\tau > 0$ and $\xi_1, \xi_2 \geq \sqrt{4\rho\tau}$,

$$|[\exp(-\tau\mathcal{A})]_{ij}| \leq \Phi(\xi_1, \rho)\Phi(\xi_2, \rho),$$

where $\xi_k = \lceil |i_k - j_k|/\beta \rceil$, $k = 1, 2$ and Φ is defined by (3.51)*.*

For Laplace–Stieltjes functions the authors from [12] further presented the bound

$$|[f(\mathcal{A})]_{ij}| \leq \int_0^\infty \exp(-2\lambda_{\min}\tau)\ |[\exp(-\tau\widehat{A})]_{i_1 j_1}|\ |[\exp(-\tau\widehat{A})]_{i_2 j_2}|\ \mathrm{d}\alpha(\tau), \tag{3.55}$$

where $\mathcal{A}$ is defined as in Theorem 3.38, $\widehat{A} = A - \lambda_{\min} I$ and $\lambda_{\min}$ the smallest eigenvalue of A. Using the bound for the exponential from Theorem 3.36 for the integrand in (3.55) directly would lead to an excessive fragmentation of the integral due to the definition of the function Φ. Hence, it was suggested in [12] to bound (3.55) by

$$\left(\int_0^\infty \exp(-\lambda_{\min}\tau)|[\exp(-\tau\widehat{A})]_{i_1 j_1}|^2 \mathrm{d}\alpha(\tau)\right)^{\frac{1}{2}} \cdot \left(\int_0^\infty \exp(-\lambda_{\min}\tau)|[\exp(-\tau\widehat{A})]_{i_2 j_2}|^2 \mathrm{d}\alpha(\tau)\right)^{\frac{1}{2}} \tag{3.56}$$

such that the two integrals can now be bounded according to (3.52) from Theorem 3.37 which requires the evaluation of six integrals. It was additionally noted in [12] that this approach can be generalized to the case where $\mathcal{A}$ is the Kronecker sum of three and more banded matrices. At this point we want to emphasize that in this case again a bound like (3.56) is necessary and the number of integrals increases with the number of matrices in the Kronecker sum.

3.3.2 New results

In this section we present improvements of the results introduced in the previous section and we extend the results for Cauchy–Stieltjes functions to a larger class of matrices using the results from Section 3.2.

Recalling the definition of a Stieltjes function from (3.47), we can obtain bounds for Cauchy–Stieltjes matrix functions of sparse, normal matrices by exploiting decay bounds for $(A + \tau I)^{-1}$, the inverses of shifted versions of A. The following theorem shows that it is possible to obtain explicit bounds, i.e., bounds in which no integrals appear anymore, for *any* Stieltjes function of a Hermitian positive definite matrix. This is similar to what was done in [12] for the special case of the inverse square root.

Theorem 3.39. *Let f be a Cauchy–Stieltjes function of the form (3.47) and let A be Hermitian positive definite. Then the entries of $f(A)$ can be bounded by*

$$|[f(A)]_{ij}| \leq 2f(\lambda_{\min})\, q^{\mathrm{d}(i,j)} \text{ with } q = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}. \tag{3.57}$$

Proof. Using Theorem 3.35 for general sparse matrices, we have

$$|[f(A)]_{ij}| \leq \int_0^\infty C(\tau) q(\tau)^{\mathrm{d}(i,j)} \, \mathrm{d}\mu(\tau). \tag{3.58}$$

Since for $\tau \geq 0$ we have

$$q(\tau) = \frac{\sqrt{\kappa(\tau)}-1}{\sqrt{\kappa(\tau)}+1} = \frac{\sqrt{\lambda_{\max}+\tau}-\sqrt{\lambda_{\min}+\tau}}{\sqrt{\lambda_{\max}+\tau}+\sqrt{\lambda_{\min}+\tau}}$$
$$= \frac{\lambda_{\max}-\lambda_{\min}}{\left(\sqrt{\lambda_{\max}+\tau}+\sqrt{\lambda_{\min}+\tau}\right)^2} \leq \frac{\lambda_{\max}-\lambda_{\min}}{\left(\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}\right)^2} = q$$

we obtain the bound

$$|[f(A)]_{ij}| \leq q^{\mathrm{d}(i,j)} \int_0^\infty C(\tau)\, \mathrm{d}\mu(\tau). \tag{3.59}$$

The integrand $C(\tau)$ is defined in (3.49), where the second argument of the maximum can be bounded as

$$\frac{(1+\sqrt{\kappa(\tau)})^2}{2(\lambda_{\max}+\tau)} = \frac{1}{2(\lambda_{\max}+\tau)} + \frac{1}{\sqrt{(\lambda_{\min}+\tau)(\lambda_{\max}+\tau)}} + \frac{1}{2(\lambda_{\min}+\tau)} \leq \frac{2}{\lambda_{\min}+\tau},$$

which obviously is also an upper bound for the first argument in (3.49), such that

$$\int_0^\infty C(\tau)\, \mathrm{d}\mu(\tau) \leq \int_0^\infty \frac{2}{\lambda_{\min}+\tau}\, \mathrm{d}\mu(\tau) = 2f(\lambda_{\min}).$$

□

With the following lemma we can give similar bounds for other classes of normal matrices with $\sigma(A) \subset \mathbb{C} \setminus \mathbb{R}_0^-$, namely (shifted) skew-Hermitian matrices.

Lemma 3.40. *Let $A = M + sI$ be a matrix with $M^H = -M$ and $s \geq 0$. Define*

$$\widehat{\lambda} := \operatorname*{argmin}_{\lambda \in \sigma(A)} |\lambda|.$$

Then

$$\|(A+\tau I)^{-1}\|_2 \leq \frac{\sqrt{2}}{|\widehat{\lambda}|+\tau} \text{ for } \tau \in \mathbb{R}_0^+. \tag{3.60}$$

Proof. Since A and thus $A + \tau I$ is (shifted) skew-Hermitian, we have

$$\|(A+\tau I)^{-1}\|_2 = \frac{1}{\min_{\lambda\in\sigma(A)} |\lambda+\tau|} = \frac{1}{|\widehat{\lambda}+\tau|}.$$

The function $g(\tau) = \frac{|\widehat{\lambda}|+\tau}{|\widehat{\lambda}+\tau|}$ attains its maximum at $\tau = |\widehat{\lambda}|$. Hence, we obtain

$$g(\tau) \leq \frac{2|\widehat{\lambda}|}{|\widehat{\lambda}+|\widehat{\lambda}|\,|} = \frac{2}{\left|\frac{\widehat{\lambda}}{|\widehat{\lambda}|}+1\right|} \leq \frac{2}{\sqrt{2}} = \sqrt{2},$$

where the second inequality holds due to $\mathrm{Re}(\widehat{\lambda}) \geq 0$. The assertion then follows because $g(\tau) \leq \sqrt{2}$ implies $\frac{1}{|\widehat{\lambda}+\tau|} \leq \frac{\sqrt{2}}{|\widehat{\lambda}|+\tau}$. □

Theorem 3.41. *Let A be a nonsingular, skew-Hermitian matrix and let f be a Cauchy–Stieltjes function of the form* (3.47)*. Then the entries of $f(A)$ can be bounded by*

$$|[f(A)]_{ij}| \leq 2\sqrt{2}\, f(\|A^{-1}\|_2^{-1}) \cdot \begin{cases} q^{\mathrm{d}(i,j)} & \text{for } \mathrm{d}(i,j) \text{ even} \\ q^{\mathrm{d}(i,j)-1} & \text{for } \mathrm{d}(i,j) \text{ odd} \end{cases}$$

with

$$q = \sqrt{\frac{\kappa(A)-1}{\kappa(A)+1}}.$$

Proof. We know from Corollary 3.21 that

$$|[f(A)]_{ij}| = \left| \int_0^\infty [(A+\tau I)^{-1}]_{ij}\, \mathrm{d}\mu(\tau) \right| \leq \int_0^\infty 2\, \|(A+\tau I)^{-1}\|_2\, q(\tau)^{\mathrm{d}(i,j)}\, \mathrm{d}\mu(\tau) \tag{3.61}$$

with

$$q(\tau) = \sqrt{\frac{\kappa(A+\tau I)-1}{\kappa(A+\tau I)+1}}$$

holds for $\mathrm{d}(i,j)$ even. Define $a := \min_{\lambda\in\sigma(A)} |\lambda|$ and $b := \max_{\lambda\in\sigma(A)} |\lambda|$. Then for $\tau \geq 0$ the inequality

$$\frac{\kappa(A+\tau I)-1}{\kappa(A+\tau I)+1} = \frac{\sqrt{\frac{\tau^2+b^2}{\tau^2+a^2}}-1}{\sqrt{\frac{\tau^2+b^2}{\tau^2+a^2}}+1} = 1 - \frac{2}{\sqrt{\frac{\tau^2+b^2}{\tau^2+a^2}}+1} \leq 1 - \frac{2}{\frac{b}{a}+1} = \frac{\kappa(A)-1}{\kappa(A)+1},$$

holds because $\frac{\tau^2+b^2}{\tau^2+a^2} = 1 + \frac{b^2-a^2}{\tau^2+a^2}$ is monotonically decreasing in τ. This gives $q \geq q(\tau)$ for all $\tau \geq 0$, and from (3.61) we therefore obtain

$$|[f(A)]_{ij}| \leq q^{\mathrm{d}(i,j)} \int_0^\infty 2\, \|(A+\tau I)^{-1}\|_2\, \mathrm{d}\mu(\tau).$$

With Lemma 3.40 it follows

$$\int_0^\infty 2\, \|(A+\tau I)^{-1}\|_2\, \mathrm{d}\mu(\tau) \leq \int_0^\infty \frac{2\sqrt{2}}{a+\tau}\, \mathrm{d}\mu(\tau) = 2\sqrt{2} f(a) = 2\sqrt{2} f(\|A^{-1}\|_2^{-1}).$$

The case where $\mathrm{d}(i,j)$ is odd can be treated in an analogous manner. □

A similar result can be obtained for general normal matrices where the spectrum is contained in a line segment $[\lambda_1, \lambda_2]$.

Theorem 3.42. *Let A be a normal matrix with $\sigma(A) \subset [\lambda_1, \lambda_2]$ and $[\lambda_1, \lambda_2] \cap \mathbb{R}_0^- = \emptyset$. Then there exist real numbers $\gamma \in [1, \infty)$ and $\tau^* \in \mathbb{R}_0^+$ such that for all $i \neq j$ the entries of $f(A)$ can be bounded by*

$$|[f(A)]_{ij}| \leq \int_0^\infty \|(A + \tau I)^{-1}\|_2 \, \frac{2}{1 - q(\tau)^{-2}} \, q(\tau)^{-\mathrm{d}(i,j)} \, \mathrm{d}\mu(\tau) \tag{3.62}$$

$$\leq \gamma \, f(\|A^{-1}\|_2^{-1}) \, \frac{2}{1 - q(\tau^*)^{-2}} \, q(\tau^*)^{-\mathrm{d}(i,j)} \tag{3.63}$$

with

$$q(\tau) = \mathrm{e}^{\mathrm{Re}(z)} > 1,$$

where z is the solution of $x(\tau) = \cosh(z)$ with

$$x(\tau) = \frac{\lambda_1 + \lambda_2 + 2\tau}{\lambda_2 - \lambda_1}.$$

For $i = j$ we obtain

$$|[f(A)]_{ii}| \leq \gamma \, f(\|A^{-1}\|_2^{-1}). \tag{3.64}$$

Proof. The inequality (3.62) immediately follows by applying Theorem 3.18 to $A + \tau I$. In particular, $[\lambda_1, \lambda_2] \cap \mathbb{R}_0^- = \emptyset$ implies that $0 \notin [\lambda_1 + \tau, \lambda_2 + \tau]$ for all $\tau \in \mathbb{R}_0^+$, so with Lemma 3.11 and 3.12 we have $q(\tau) > 1$ for all such τ.
Postponing the proof to the end, let us assume that we already know that $q(\tau)$ has a minimum on $\mathbb{R}_0^+$ and let $\tau^* = \mathrm{argmin}_{\tau \in \mathbb{R}_0^+} q(\tau)$ denote the corresponding minimizer. Then (3.62) can be bounded by

$$\int_0^\infty \|(A + \tau I)^{-1}\|_2 \, \mathrm{d}\mu(\tau) \, \frac{2}{1 - q(\tau^*)^{-2}} \, q(\tau^*)^{-\mathrm{d}(i,j)}.$$

Defining $|\widehat{\lambda}| := \min_{\lambda \in \sigma(A)} |\lambda| = \|A^{-1}\|_2^{-1}$, we then obtain

$$\|(A + \tau I)^{-1}\|_2 = \frac{1}{\min_{\lambda \in \sigma(A)} |\lambda + \tau|} \leq \frac{\gamma}{|\widehat{\lambda}| + \tau}$$

for some $\gamma \in [1, \infty)$ which can be seen from the equivalent formulation

$$g(\tau) := \frac{|\widehat{\lambda}| + \tau}{\min_{\lambda \in \sigma(A)} |\lambda + \tau|} \leq \gamma,$$

where this upper bound γ exists since g is a continuous function in $\tau \geq 0$, $g(0) = 1$ and $\lim_{\tau \to \infty} g(\tau) = 1$.

Overall, we have the estimate

$$\begin{aligned} \int_0^\infty \|(A + \tau I)^{-1}\|_2 \, \mathrm{d}\mu(\tau) &\leq \int_0^\infty \frac{\gamma}{\|A^{-1}\|_2^{-1} + \tau} \, \mathrm{d}\mu(\tau) \\ &= \gamma \, f(\|A^{-1}\|_2^{-1}). \end{aligned}$$

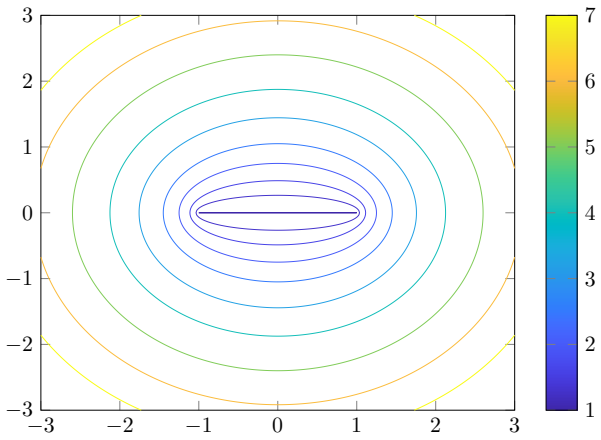

Figure 3.11: Contour lines of the function $h(z)$.

It remains to show that the minimizer τ^* of $q(\tau)$ on $\mathbb{R}_0^+$ exists. We define the function

$$h : \mathbb{C} \to [1, \infty) \text{ with } h(z) = \mathrm{e}^{\mathrm{Re}(w)}$$

where w is the solution of $z = \cosh(w)$ with $\mathrm{Re}(w) \geq 0$. Then it holds $q(\tau) = h(x(\tau))$, such that we can now investigate the function h and the argument $x(\tau)$ for $\tau \in \mathbb{R}_0^+$ instead.

The contour lines of h on levels $\rho \in [1, \infty)$ defined by $\{z \in \mathbb{C} : h(z) = \rho\}$ are given by confocal ellipses E_ρ with focal points -1 and 1 and semi-axes $\frac{\rho - \rho^{-1}}{2}$ and $\frac{\rho + \rho^{-1}}{2}$. This can be seen as follows: For a given $\rho \in [1, \infty)$ the equation

$$\rho = h(z) = \mathrm{e}^{\mathrm{Re}(w)} = |\, \mathrm{e}^w \,| \tag{3.65}$$

is fulfilled for all those $z \in \mathbb{C}$, for which e^w lies on a circle with radius ρ centered at the origin. Since w is defined as the solution of $z = \cosh(w)$ and $\cosh(w)$ can be interpreted as the Joukowski mapping of e^w, equation (3.65) is fulfilled for all $z \in \mathbb{C}$ that lie on the ellipse E_ρ. Hence, the contour lines of the function h on levels $\rho \geq 1$ are confocal ellipses E_ρ with focal points -1 and 1, as illustrated in Figure 3.11. Note in particular, that the values of h on these ellipses E_ρ are trivially monotonically increasing with increasing ρ.

At the same time $x(\mathbb{R}_0^+) := \{x(\tau) : \tau \in \mathbb{R}_0^+\}$ is a half-line with endpoint $x(0)$ not intersecting the interval $[-1, 1]$ (see Lemma 3.12). Hence, $h(x(\tau)) > 1$ and $h(x(\tau)) \to \infty$ for $\tau \to \infty$, thus $h(x(\tau)) = q(\tau)$ must attain a minimum for $\tau \in \mathbb{R}_0^+$. The minimum on the half line $x(\mathbb{R}_0^+)$ is either given by the endpoint $x(0)$ or a point of tangency $x(\tau^*)$ of the half line $x(\mathbb{R}_0^+)$ and an ellipse E_{ρ^*} with minimal $\rho^* > 1$. Such a point $x(\tau^*)$ is exemplarily illustrated in Figure 3.12 for a shifted Hermitian matrix with spectrum contained in the line segment $[\lambda_1, \lambda_2] = [-3 + \mathbf{i}, 1 + \mathbf{i}]$. □

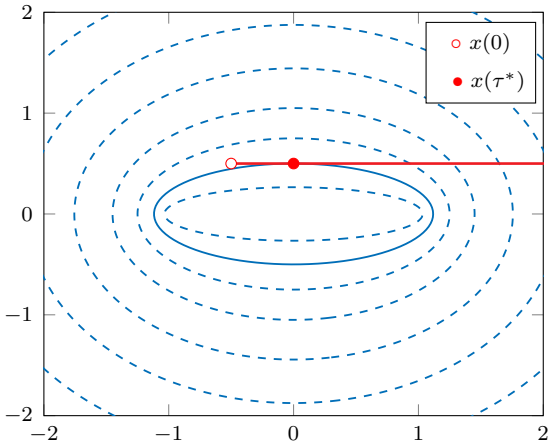

Figure 3.12: Contour lines of $h(z)$ (blue lines) and $x(\mathbb{R}_0^+)$ (red line) for the line segment $[\lambda_1, \lambda_2] = [-3 + \mathbf{i}, 1 + \mathbf{i}]$.

Calculating γ and τ^* explicitly might be too expensive in general, but Theorem 3.42 shows that the integral (3.62) exists and can thus, e.g., be approximated numerically. In the next theorem we show that τ^* can be easily calculated for shifted skew-Hermitian matrices and Hermitian matrices with a complex shift.

Theorem 3.43. *Let A be of the form $A = M + \mathbf{i} \cdot sI$, where $M = M^H$, or of the form $A = M + sI$, where $M = -M^H$, with $s \in \mathbb{R}$. Further let $\sigma(A) \subset [\lambda_1, \lambda_2]$ with $[\lambda_1, \lambda_2] \cap \mathbb{R}_0^- = \emptyset$. Then there exists a real number $\gamma \in [1, \infty)$ such that for all $i \neq j$ we have*

$$|[f(A)]_{ij}| \leq \gamma \, f(\|A^{-1}\|_2^{-1}) \, \frac{2}{1 - q(\tau^*)^{-2}} \, q(\tau^*)^{-\mathrm{d}(i,j)} \tag{3.66}$$

with

$$\tau^* = \max\left\{0, -\tfrac{\mathrm{Re}(\lambda_1)+\mathrm{Re}(\lambda_2)}{2}\right\}, \tag{3.67}$$

where $q(\tau)$ is defined as in Theorem 3.42. For $i = j$ we obtain

$$|[f(A)]_{ij}| \leq \gamma \, f(\|A^{-1}\|_2^{-1}). \tag{3.68}$$

Proof. The assertion is proven by showing that

$$\operatorname*{argmin}_{\tau \in \mathbb{R}_0^+} q(\tau) = \max\left\{0, -\tfrac{\mathrm{Re}(\lambda_1)+\mathrm{Re}(\lambda_2)}{2}\right\}.$$

First let A be Hermitian with complex shift $\mathbf{i} \cdot s$, i.e., $\mathrm{Im}(\lambda_1) = \mathrm{Im}(\lambda_2) = s$. Then the imaginary part of $x(\tau)$ is constant and

$$\mathrm{Re}(x(\tau)) = \frac{\mathrm{Re}(\lambda_1) + \mathrm{Re}(\lambda_2) + 2\tau}{\mathrm{Re}(\lambda_2) - \mathrm{Re}(\lambda_1)}.$$

Hence, $x(\mathbb{R}_0^+)$ is a half-line which is parallel to the real axis. Therefore, if $x(\mathbb{R}_0^+)$ does not cross the imaginary axis, $\mathrm{Re}(\operatorname{arcosh}(x(\tau)))$ attains it minimal value for $\tau = 0$. Otherwise $q(\tau)$ is minimal for the intersection with the imaginary axis, i.e., for $\mathrm{Re}(x(\tau)) = 0$ (see the example of Figure 3.12), which is the case when $\tau = -(\mathrm{Re}(\lambda_1) + \mathrm{Re}(\lambda_2))/2$ holds. Thus, $x(\mathbb{R}_0^+)$ intersects the imaginary axis if and only if $-(\mathrm{Re}(\lambda_1) + \mathrm{Re}(\lambda_2))/2 \geq 0$.

Now consider the shifted skew-Hermitian case. Then we have $\mathrm{Re}(\lambda_1) = \mathrm{Re}(\lambda_2) = s$ and the real part of $x(\tau)$ is constant. For the imaginary part we have

$$\mathrm{Im}(x(\tau)) = \frac{2s + 2\tau}{\mathrm{Im}(\lambda_1) - \mathrm{Im}(\lambda_2)}.$$

With similar arguments as above, $q(\tau)$ is either minimal if $\tau = 0$ or at the intersection of $x(\mathbb{R}_0^+)$ with the real axis, i.e., if $\tau = -s = -(\mathrm{Re}(\lambda_1) + \mathrm{Re}(\lambda_2))/2 \geq 0$. □

By imposing further conditions on A, it is possible to specify the constant γ. For instance, for $A = M + sI$ with $M^H = -M$ and $s \in \mathbb{R}^+$, we know from Lemma 3.40 that $\gamma = \sqrt{2}$, hence, the entries of A are bounded by

$$|[f(A)]_{i,j}| \leq f(\|A^{-1}\|_2^{-1}) \, \frac{2\sqrt{2}}{1 - q^{-2}} \, q^{-\mathrm{d}(i,j)}, \tag{3.69}$$

where $q = q(0)$. For Cauchy–Stieltjes functions of the form $f(z) = z^{-\alpha}$ with $\alpha \in (0, 1)$ the bound (3.69) provides bounds for shifted Hermitian matrices $A = M + \mathbf{i} \cdot sI$, with $M^H = M$, as well, since there exists a shifted skew-Hermitian matrix B with $A = \mathbf{i}B$ and

$$|[A^{-\alpha}]_{ij}| = |[(\mathbf{i}B)^{-\alpha}]_{ij}| = |\mathbf{i}^{-\alpha}[B^{-\alpha}]_{ij}| = |[B^{-\alpha}]_{ij}|.$$

The presented new results for matrices of Cauchy–Stieltjes functions are already published in [38] with a slightly different proof of Theorem 3.42. We now consider bounds for Laplace–Stieltjes functions of matrices with Kronecker structure based on the graph distance. The following results and the corresponding discussion in Section 3.3.3 for Laplace–Stieltjes functions are not mentioned in our publication [38]. In [12] the relation

$$\exp(-\tau\mathcal{A}) = \exp(-\tau A) \otimes \exp(-\tau A) \tag{3.70}$$

is used to find bounds for the exponential and Laplace–Stieltjes functions of matrices $\mathcal{A}$ of the form (3.53). Alternatively, a generalized version of Theorem 3.36 for sparse matrices can be applied to matrices of the form (3.53) where the distances of the nodes $\mathrm{d}(i, j)$ is explicitly known, resulting in the following corollary. We inherit the conditions of Theorem 3.38 in order to make a comparison between the following bounds and the bounds of Theorem 3.38 easier.

Corollary 3.44. *Let A be Hermitian with bandwidth β and spectrum contained in $[0, 4\rho]$. Let further $\mathcal{A}$ be defined as $\mathcal{A} = I \otimes A + A \otimes I$, then for $\xi_1 + \xi_2 \geq \sqrt{8\rho\tau}$ we obtain*

$$|[\exp(-\tau\mathcal{A})]_{ij}| \leq \Phi(\xi_1 + \xi_2, 2\rho),$$

where $\xi_k = \lceil |i_k - j_k|/\beta \rceil$, $k = 1, 2$ and Φ defined by (3.51).

Proof. The bound directly follows by using the generalization of Theorem 3.36 with $\xi = \mathrm{d}(i, j)$ and by noticing that $\mathrm{d}(i, j) = \xi_1 + \xi_2$ in $G(\mathcal{A})$ and that $\sigma(\mathcal{A}) \subset [0, 8\rho]$ if $\sigma(A) \subset [0, 4\rho]$ (see, e.g., [55, Theorem 4.4.5]). □

This result for the exponential was already mentioned incidentally in [12]. Using Corollary 3.44 for the exponential, we can formulate the following result for Laplace–Stieltjes functions of matrices with Kronecker structure based on the graph distance.

Corollary 3.45. *Let $\mathcal{A} = I \otimes A + A \otimes I$ be a matrix with A Hermitian and $\lambda_{\min}$ the smallest eigenvalue of A. Let the spectrum of $\widehat{\mathcal{A}} := \mathcal{A} - 2\lambda_{\min} I$ be contained in $[0, 4\rho]$ and let f be a Laplace–Stieltjes function. Then the bound*

$$\begin{aligned}
|[f(\mathcal{A})]_{ij}| \leq 10 \int_0^{\frac{\xi_1+\xi_2}{2\rho}} & \exp(-2\lambda_{\min}\tau) \frac{\exp(-\rho\tau)}{\rho\tau} \left(\frac{\mathrm{e}\,\rho\tau}{\xi_1 + \xi_2} \right)^{\xi_1+\xi_2} \mathrm{d}\alpha(\tau) \\
& + 10 \int_{\frac{\xi_1+\xi_2}{2\rho}}^{\frac{(\xi_1+\xi_2)^2}{4\rho}} \exp(-2\lambda_{\min}\tau) \exp\left(-\frac{(\xi_1 + \xi_2)^2}{5\rho\tau} \right) \mathrm{d}\alpha(\tau) \\
& + \int_{\frac{(\xi_1+\xi_2)^2}{4\rho}}^{\infty} \exp(-2\lambda_{\min}\tau) |[\exp(-\tau\widehat{\mathcal{A}})]_{ij}]| \mathrm{d}\alpha(\tau),
\end{aligned}$$

holds, where the last term can be bounded by

$$\int_{\frac{(\xi_1+\xi_2)^2}{4\rho}}^{\infty} \exp(-2\lambda_{\min}\tau) \mathrm{d}\alpha(\tau).$$

Proof. We clearly have

$$|[f(\mathcal{A})]_{ij}| \leq \int_0^{\infty} |[\exp(-\tau\mathcal{A})]_{ij}| \mathrm{d}\alpha(\tau).$$

Since $\lambda_{\min}$ is the smallest eigenvalue of A, $2\lambda_{\min}$ is the smallest eigenvalue of $\mathcal{A}$ (see, e.g., [55, Theorem 4.4.5]) and therefore the spectrum of $\widehat{\mathcal{A}} := \mathcal{A} - 2\lambda_{\min} I$ is contained in an interval $[0, 4\rho]$. Hence,

$$|[f(\mathcal{A})]_{ij}| \leq \int_0^{\infty} \exp(-2\lambda_{\min}\tau) |[\exp(-\tau\widehat{\mathcal{A}})]_{ij}| \mathrm{d}\alpha(\tau)$$

can be bounded via Corollary 3.44 with an adapted ρ. In addition we have

$$|[\exp(-\tau\widehat{\mathcal{A}})]_{ij}| \leq \|\exp(-\tau\widehat{\mathcal{A}})\|_2 \leq 1$$

for all $\tau \geq 0$, such that the bound for the last term holds. □

The advantages of using the results based on the graph distance given by Corollary 3.44 and Corollary 3.45 compared to the results based on relation (3.70) are pointed out in the next section.

3.3.3 Comparison and numerical examples

In this section we illustrate some of the bounds obtained in Section 3.3.2 and compare them to the results from the literature presented in Section 3.3.1 if possible. We begin by investigating the case of Cauchy–Stieltjes functions of Hermitian positive definite matrices, where explicit bounds are given in Theorem 3.39. In [12] the explicit bound (3.50) was obtained for the inverse square root so we now compare the bounds for the special case $f(z) = z^{-1/2}$. The bound of Theorem 3.39 is then given by

$$|[A^{-\frac{1}{2}}]_{ij}| \leq \frac{2}{\sqrt{\lambda_{\min}}}\, q^{\mathrm{d}(i,j)} \tag{3.71}$$

with

$$q = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1},$$

whereas the bound from [12] is given by

$$|[A^{-\frac{1}{2}}]_{ij}| \leq C\, q^{\frac{|i-j|}{\beta}} \tag{3.72}$$

with

$$C = C(0) + \frac{\left(1 + \sqrt{\kappa(A)}\right)^2}{2}$$

and $C(\tau)$ as defined in (3.49). Therefore we only need to compare the constants $2/\sqrt{\lambda_{\min}}$ and C. First we consider the case when $C(\tau)$ is given by the first quantity of the maximum (3.49), i.e.,

$$C = \frac{1}{\lambda_{\min}} + \frac{\left(1 + \sqrt{\kappa(A)}\right)^2}{2} \geq \frac{1}{\lambda_{\min}} + 2.$$

Since we have

$$f(x) = \frac{1}{x} - \frac{2}{\sqrt{x}} + 2 = \frac{1 - 2\sqrt{x} + 2x}{x} = \frac{\left(1 - \sqrt{x}\right)^2 + x}{x} > 0$$

for $x > 0$, we obtain $f(\lambda_{\min}) > 0$, and therefore

$$C > \frac{2}{\sqrt{\lambda_{\min}}}.$$

On the other hand, if $C(\tau)$ is given by the second quantity of the maximum, then the constant C reads

$$C = \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2\lambda_{\max}} + \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2} = \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2\lambda_{\min}\kappa(A)} + \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2}.$$

Hence, we now investigate the function

$$\begin{aligned} f(x,y) &= \frac{\left(1+\sqrt{y}\right)^2}{2xy} + \frac{\left(1+\sqrt{y}\right)^2}{2} - \frac{2}{\sqrt{x}} \\ &= \frac{1+2\sqrt{y}+y+xy+2xy\sqrt{y}+xy^2-2\sqrt{x}y}{2xy} \end{aligned}$$

with $x > 0$ and $y \geq 1$. For $y \geq 1$ we have $\sqrt{y} \geq 1$ and $y^2 \geq y$, so that for $x > 0$ we obtain

$$\begin{aligned} f(x,y) &> \frac{y+xy+2xy+xy-2\sqrt{x}y}{2xy} \\ &= \frac{1+4x-2\sqrt{x}}{x} = \frac{\left(1-\sqrt{x}\right)^2+3x}{x} > 0. \end{aligned}$$

Therefore we have $f(\lambda_{\min}, \kappa(A)) > 0$, i.e.,

$$C > \frac{2}{\sqrt{\lambda_{\min}}}.$$

Summarizing, the bound of Theorem 3.35 can be used for any Cauchy-Stieltjes function and in addition it is sharper than the explicit bound from [12] for the special case $f(z) = z^{-1/2}$.

For the matrix

$$A = \operatorname{tridiag}(-1, 4, -1) \in \mathbb{C}^{200\times 200}$$

the exact values of the 120th column of $A^{-1/2}$ as well as the bounds from Theorem 3.35 of [12] (evaluated by numerical quadrature), (3.71) and (3.72) are depicted in Figure 3.13. Of course, all approaches give the same decay rate, but the constant obtained in Theorem 3.39 is slightly smaller than the one in (3.72). This gives us a sharper bound, which almost agrees with the quadrature based bound from Theorem 3.39 (which is the sharpest of the bounds, as the explicit bounds are obtained by bounding the terms appearing in the integral in (3.48)). Figure 3.14 shows the same experiment for the function $f(z) = z^{-1/4}$. In this case,

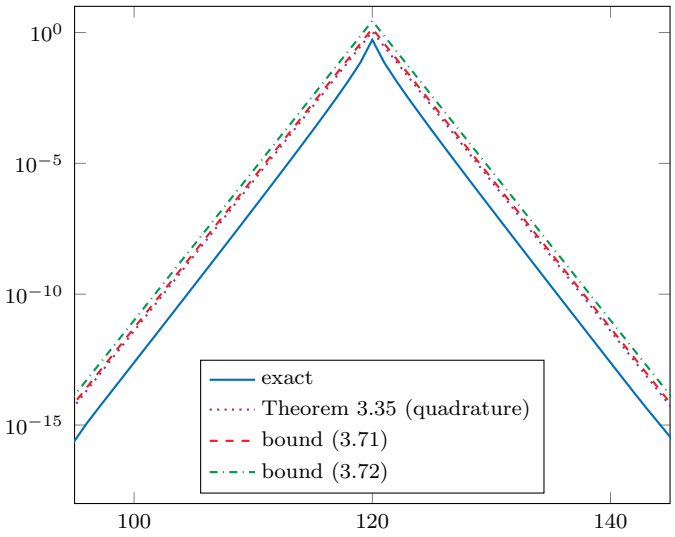

Figure 3.13: Exact values and bounds for $|[A^{-1/2}]_{ij}|$ of column $j = 120$ with $A = \text{tridiag}(-1, 4, -1)$.

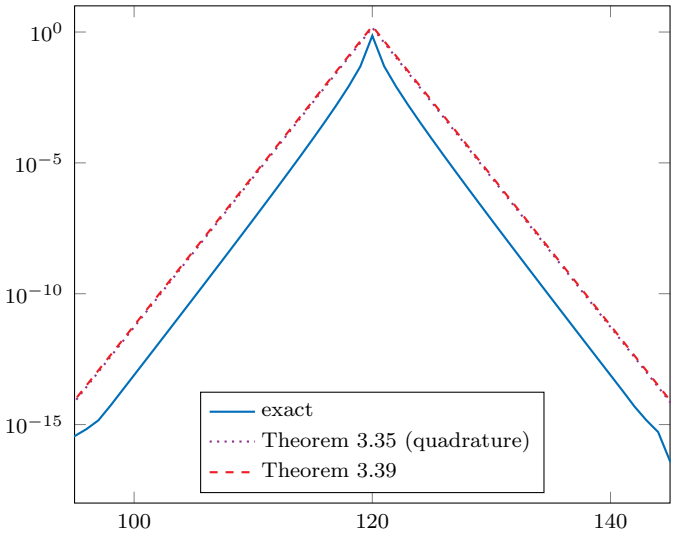

Figure 3.14: Exact values and bounds for $|[A^{-1/4}]_{ij}|$ of column $j = 120$ with $A = \text{tridiag}(-1, 4, -1)$.

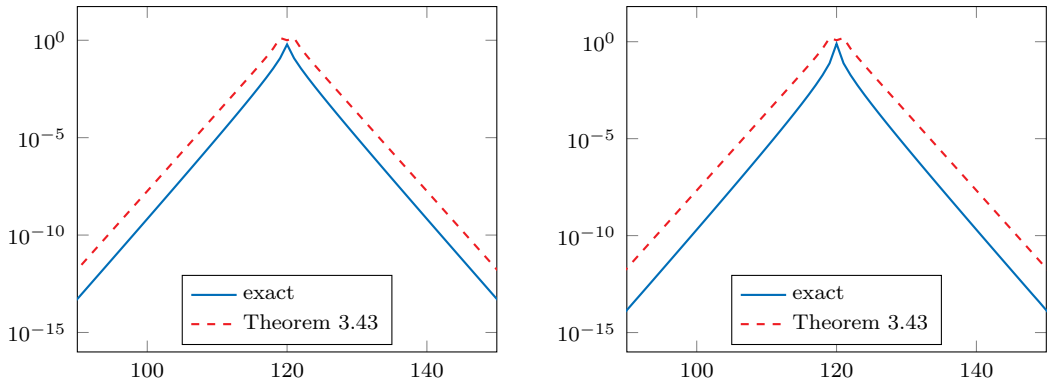

Figure 3.15: Exact values and bounds for $[A^{-1/2}]_{ij}$ (left) and $[A^{-1/4}]_{ij}$ (right) of column $j = 120$ with $A = \text{tridiag}(1, \mathbf{i}, -1) + 2 \cdot I$ of dimension $n = 200$.

no explicit bound was obtained in [12]. For that reason we are just able to compare the bound of Theorem 3.39 to that of Theorem 3.35 from [12] which has to be evaluated by numerical quadrature. Of course, the quadrature based bound is sharper than the bound from Theorem 3.39 but the difference between the bounds is not visible in Figure 3.14. Hence, with the explicit bound of Theorem 3.39 we obtain sharp decay bounds without needing to use numerical quadrature.

In a second series of experiments, we compute bounds for the entries of the matrix functions $A^{-1/2}$ and $A^{-1/4}$, where A is now the shifted skew-Hermitian matrix

$$A = \text{tridiag}(1, \mathbf{i}, -1) + 2 \cdot I \in \mathbb{C}^{200 \times 200}.$$

For the shifted skew-Hermitian case, no bounds were provided in [12], so that we only compare our bound from Theorem 3.43 to the exact value. The results of these experiments are shown in Figure 3.15. We again observe a very good approximation of the actual decay rate and the magnitude of the entries is only slightly overestimated, giving sharp bounds overall.

In order to also show a numerical example for a non-banded sparse matrix, we consider the staggered Schwinger discretization arising in quantum electrodynamics, the basic quantum field theory for the interaction of electrons and photons according to the standard model of Theoretical Physics. The discretization we are considering here uses a periodic two-dimensional lattice, where at each lattice site $x = (i, j)$ the unknown $\psi_{i,j}$ couples with its direct neighbors as

$$\begin{aligned} s\,\psi_{i,j} + u^1_{i,j}\psi_{i+1,j} + \eta_{i,j}u^2_{i,j}\psi_{i,j+1} - \overline{u}^1_{i-1,j}\psi_{i-1,j} - \eta_{i,j}\overline{u}^2_{i,j-1}\psi_{i,j-1} = \phi_{i,j}, \\ i, j = 1, \ldots, N, \;\; \eta_{i,j} = (-1)^i. \end{aligned} \tag{3.73}$$

Herein, the indices $i-1, i+1, j-1, j+1$ are to be understood modulo N to account for the periodicity. The numbers $u^1_{i,j}$ and $u^2_{i,j}$ represent the $SU(1)$ background

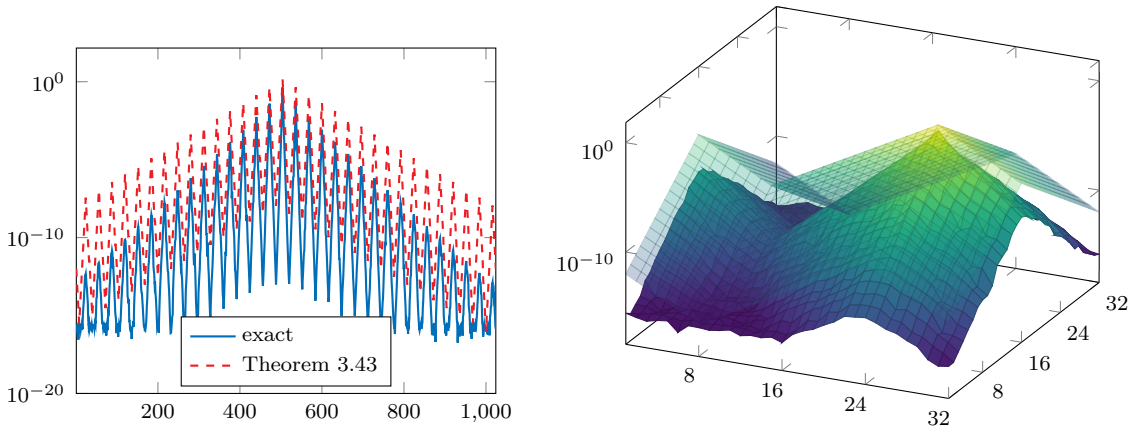

Figure 3.16: Exact values and bounds for $|[(sI + D)^{-1/2}]_{ij}|$ of column $j = 504$ for the staggered Schwinger discretization of dimension $n = 1024$ with lexicographic ordering of the nodes (left) and on a two-dimensional grid (right).

field, i.e. they are randomly distributed complex numbers of modulus 1. Clearly, (3.73) results in a system

$$(sI + D)\psi = \phi \tag{3.74}$$

with a skew-Hermitian matrix D. The spectrum of $D = -D^H$ is not only purely imaginary, but also symmetric with respect to the origin due to the odd-even structure of the coupling. The graph underlying D is the periodic $N \times N$ lattice, so that here we are in the presence of the general case where the graph distance contributes to the decay bounds. For $N = 32$, Figure 3.16 shows the exact decay for the 504th column of the inverse square root corresponding to the point $(16, 24)$ on the lattice and the bounds from Theorem 3.43. The left panel arranges the values according to a one-dimensional, lexicographic ordering of the lattice points, whereas the right panel gives the same information arranged on the underlying two-dimensional lattice. Again, the proposed decay bounds give a good impression of the actual decay in $(sI + D)^{-1/2}$.

A similar discussion and comparison of the new decay bounds for Cauchy–Stieltjes function can also be found in [38]. We now continue with a discussion of the results for the exponential and Laplace–Stieltjes functions of matrices with Kronecker structure of the form (3.53), not given in [38]. The bound of Theorem 3.38 for the exponential was obtained by using the relation (3.70) resulting in the bound

$$|[\exp(-\tau\mathcal{A})]_{ij}| \leq \Phi(\xi_1, \rho)\Phi(\xi_2, \rho), \tag{3.75}$$

whereas the bound of Corollary 3.44, given by

$$|[\exp(-\tau\mathcal{A})]_{ij}| \leq \Phi(\xi_1 + \xi_2, 2\rho), \tag{3.76}$$

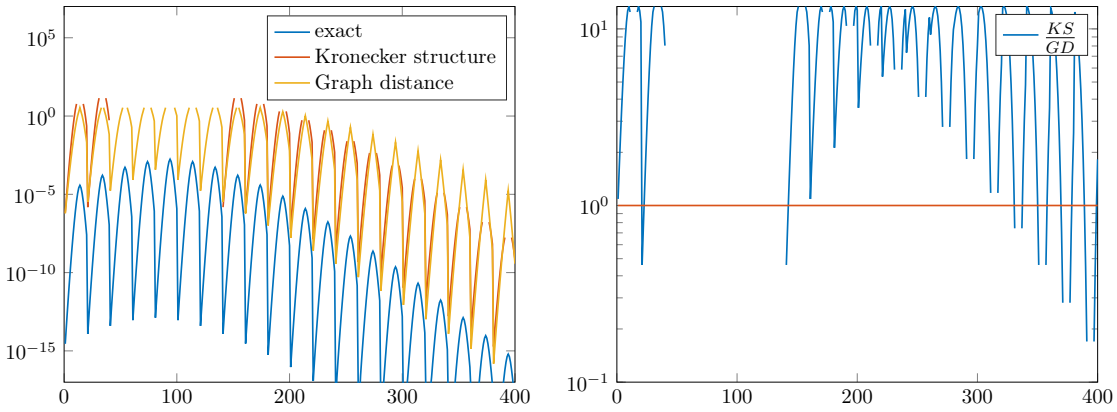

Figure 3.17: Exact values and bounds for $|[\exp(-\mathcal{A})]_{ij}|$ of column $j = 94$ for the matrix $\mathcal{A} = A \oplus A$ with $A = \text{tridiag}(-1, 4, -1)$ (left) and the corresponding ratio between the bounds based on the Kronecker structure (KS) and the graph distance (GD) (right).

is only based on the graph distance in $G(\mathcal{A})$. An apparent advantage of (3.76) in contrast to (3.75) is that it is defined for more entries of $\exp(-\tau\mathcal{A})$, since the first bound is only defined for $i, j \in \{1, \dots, n\}$ with $\xi_1 \geq \sqrt{4\rho\tau}$ and $\xi_2 \geq \sqrt{4\rho\tau}$ while the second is defined for all $i, j \in \{1, \dots, n\}$ with $\xi_1 + \xi_2 \geq \sqrt{8\rho\tau}$. Furthermore, experiments show that the bound (3.76) is slightly sharper than (3.75) for many entries of $\exp(-\tau\mathcal{A})$. In Figure 3.17 we see the exact absolute values of the 94th column of $\exp(-\mathcal{A})$, where $\mathcal{A} = A \oplus A$ with $A = \text{tridiag}(-1, 4, -1)$, as well as bound (3.75) based on the Kronecker structure and the bound (3.76) based on the graph distance. Here, more entries of $\exp(-\mathcal{A})$ can be bounded by the approach based on the graph distance. For entries where (3.75) and (3.76) are both defined, Figure 3.17 also shows the corresponding ratio between the bounds, where KS denotes the bound (3.76) based on the Kronecker structure and GS denotes the bound (3.76) based on the graph distance. We see that for most of the entries the ratio is larger than one (illustrated by the red line), i.e., the bound (3.76) is sharper than (3.75) for most of the entries.

Since the bounds for Laplace–Stieltjes functions immediately follow from the bounds for the exponential, we here have similar results for matrices with Kronecker structure (3.53). The bound of Corollary 3.45 based on the graph distance is defined for more entries of the matrix function than the bound (3.55) based on the Kronecker structure. In Figure 3.18 we have the exact values and the bounds for the 94th column of $f(\mathcal{A})$, where $f(z) = (1 - \exp(-z))/z$ and $\mathcal{A} = A \oplus A$ with $A = \text{tridiag}(-1, 4, -1)$. Again, the ratio between the proposed bounds shows that bound (3.76) is sharper than (3.75) for most of the entries.

As a second example for a Laplace–Stieltjes function, we consider the function $f(z) = z^{-1/2}$. The results for the 94th column of $f(A)$ are depicted in Figure 3.19

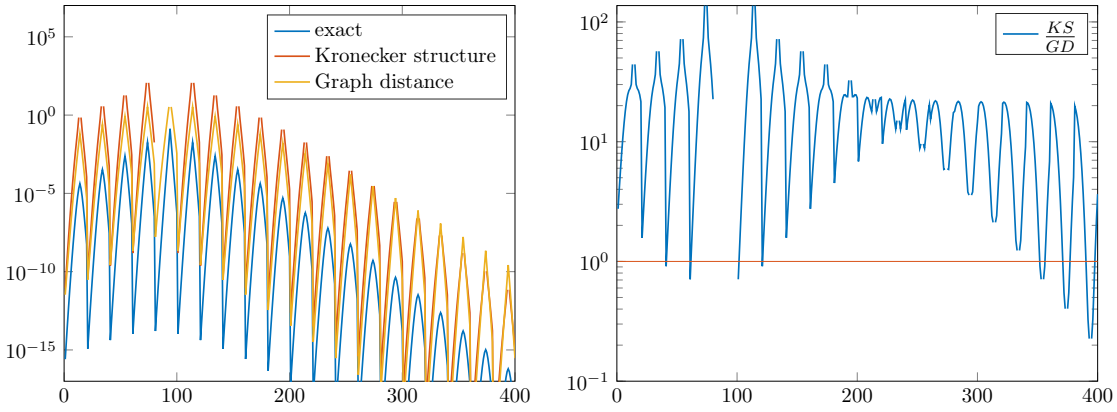

Figure 3.18: Exact values and bounds for $|[(I - \exp(-\mathcal{A}))\mathcal{A}^{-1}]_{ij}|$ of column $j = 94$ for the matrix $\mathcal{A} = A \oplus A$ with $A = \text{tridiag}(-1, 4, -1)$ (left) and the corresponding ratio between the bounds based on the Kronecker structure (KS) and the graph distance (GD) (right).

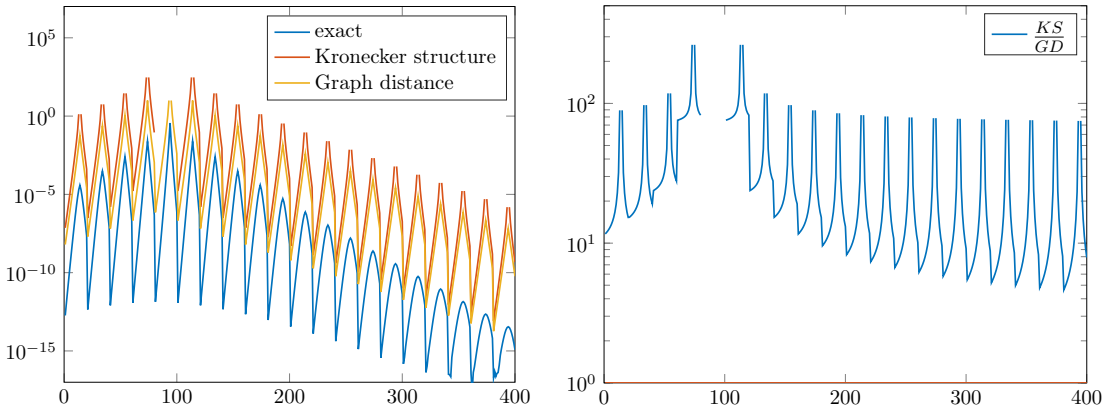

Figure 3.19: Exact values and bounds for $|[\mathcal{A}^{-1/2}]_{ij}|$ of column $j = 94$ for the matrix $\mathcal{A} = A \oplus A$ with $A = \text{tridiag}(-1, 4, -1)$ (left) and the corresponding ratio between the bounds based on the Kronecker structure (KS) and the graph distance (GD) (right).

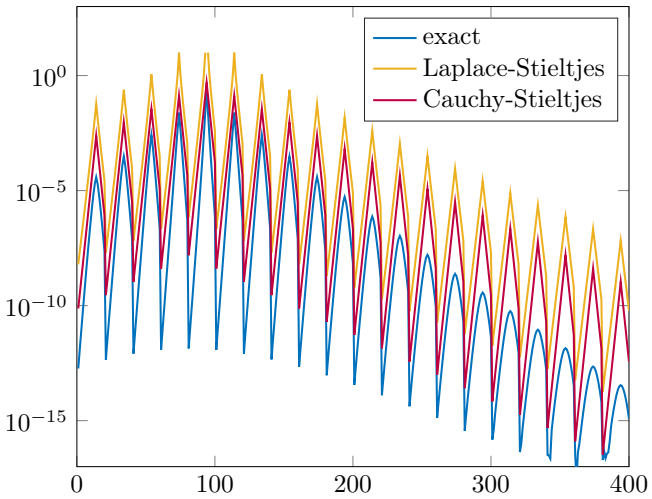

Figure 3.20: Exact values and bounds for $|[\mathcal{A}^{-\frac{1}{2}}]_{ij}|$ of column $j = 94$ for the matrix $\mathcal{A} = A \oplus A$ with $A = \text{tridiag}(-1, 4, -1)$ based on the Laplace–Stieltjes integral expression and the Cauchy–Stieltjes integral expression.

and the ratio between the bounds shows that the bound based on the graph distance is sharper than the bound based on the Kronecker structure for all entries of the 94th column.

Since $f(z) = z^{-1/2}$ is also a Cauchy–Stieltjes function, in Figure 3.20 the bound based on the Cauchy–Stieltjes integral expression of f from Theorem 3.39 is compared to the bound based on the Laplace–Stieltjes expression of f from Corollary 3.45. We see that in this case, we obtain better results with the bound for Cauchy–Stieltjes functions, where in addition no integral has to be computed.

A further advantage of the bounds for Laplace–Stieltjes functions of matrices with Kronecker structure based on the graph distance is the easy generalization to matrices of the form $\mathcal{A} = A_1 \oplus A_2 \oplus \cdots \oplus A_k$. Then the quantity $\xi_1 + \xi_2$ just needs to be replaced by $\xi_1 + \xi_2 + \cdots + \xi_k$ in Corollary 3.45 and we still only need to compute three integrals. On the other hand, using the Kronecker structure explicitly as suggested in [12] leads to a bound where $3k$ integrals have to be computed, and the numerical examples show that this additional effort does not necessarily result in sharper decay bounds. Summarizing, it seems to be sufficient and even more effective to use the general graph distance approach instead of considering any structure of the matrix (besides for determining the distance of the nodes).

Chapter 4

Exploiting the decay in matrix functions

In this chapter we develop methods which explicitly use the decay property in matrix functions for associated, important matrix computations. We are especially interested in methods which scale linearly with the matrix dimension. The problems we are considering include

- computing sparse approximations of matrix functions (Section 4.2),
- approximating $f(A)v$ for a given vector v (Section 4.3) and
- approximating traces of matrix functions (Section 4.4).

Besides giving some theoretical results, all these sections basically consider two approaches for the computations of the quantities of interest. The first one is based on the Chebyshev series expansion as introduced in Section 2.3 for some types of matrices, while the second one uses a special coloring of the nodes in $G(A)$, the graph of A. Computing such a *distance-d coloring* of a graph is a crucial task for the development of these approaches and is therefore discussed in detail in a preliminary section. Some of the methods we introduce are already used heuristically for certain applications by assuming a decay in the corresponding matrix. In these cases we supply a more extensive theoretical analysis by using decay bounds of matrix functions as developed in Chapter 3.

In the following sections we often need to compute the quantities $v^T f(A)v$ and $f(A)v$ for several vectors v. To this purpose we can use the Lanczos or Arnoldi approximation introduced in Section 2.1.2. We want to emphasize that the conditions under which we can guarantee a rapid decay in $f(A)$ are essentially the same as those for a fast convergence of the Lanczos/Arnoldi approximations for $v^T f(A)v$ and $f(A)v$. This will be discussed in detail in Section 4.3. Hence, we can assume that an accurate approximation can be reached with a small number of iterates, i.e., we have $m \ll n$ in Algorithm 2.1 and Algorithm 2.2, respectively.

Since the Arnoldi process requires $\mathcal{O}(m^2 n)$ operations for a sparse matrix A and since we assume that an accurate approximation can be reached for $m \ll n$ we consider that the quantities $v^T f(A)v$ and $f(A)v$ can be computed with a cost of $\mathcal{O}(n)$. We want to clarify that most of the error results in the following sections are formulated for the exact quantities $v^T f(A)v$ and $f(A)v$. It is straight forward to formulate results which consider additional errors caused by the computations of $v^T f(A)v$ and $f(A)v$ via the Lanczos or Arnoldi process. For this we refer to the remarks given in Section 2.1.2, especially for the Lanczos approximation of $v^T f(A)v$. A formulation of an error bound which considers the error caused by the computation of $v^T f(A)v$ is exemplarily given in Section 4.4.2.

In order to check the quality of the proposed methods, we mainly consider problems of moderate size. For a comparison, we compute the "exact" quantities of interest via built-in MATLAB functions associated with matrix functions, such as the functions `inv` or `sqrtm` for the computation of the inverse or the square root of matrices, respectively. In Section 4.4.2 we then exemplarily consider problems of large size where we cannot compute the exact quantities via built-in MATLAB functions and the quality of the approximations can only be checked with the proposed error bounds.

4.1 Preliminaries: The computation of a distance-d coloring

The coloring of graphs is an extensive field in graph theory. Starting with the familiar *Four Color Problem* (see, e.g., [106]), first mentioned in 1852, the number of further coloring problems for graphs massively increased over the recent years. The *distance-d coloring* of a graph is a crucial tool for the development of the methods and results described in the following sections. Therefore, we will discuss some aspects and methods related to this special coloring problem. A distance-d coloring of a graph is defined as follows.

Definition 4.1. For a graph $G = (V, E)$, a distance-d coloring is a mapping $\mathrm{col} : V \to \{1, \dots, k\}$ such that $\mathrm{col}(i) \neq \mathrm{col}(j)$ if $\mathrm{d}(i, j) \leq d$. A distance-d coloring is optimal if the number k of colors is minimal among all distance-d colorings of G.

Even for $d = 1$ the problem of finding an optimal coloring is NP-hard for general graphs [63]. The number of colors in an optimal coloring is called the *chromatic*

number. Besides the problem of determining an optimal distance-d coloring, the distance-d coloring problem also includes the determination of the chromatic number or deciding whether a graph is distance-d colorable with k colors [92]. There are numerous results on upper and lower bounds for the chromatic number of general and specific graphs, see, e.g., [61, 68, 69, 104].

In the following, we are interested in a low-cost method for computing a distance-d coloring of a graph for a given distance d. Ideally, the computation should not exceed a work of $\mathcal{O}(|V| + |E|)$. Of course, with this restriction we can not expect an optimal coloring. Instead, our goal is to produce a coloring with a sufficiently small number of colors. Low-cost methods for the coloring of a graph are well studied for the case $d = 1$. In the literature, a distance-1 coloring is called the classical (vertex-)coloring of G. Some heuristics for a classical coloring of a graph can be found in [63, Chapter 1] for undirected graphs. In the following we will discuss some heuristics for general distance-d colorings of undirected graphs. For directed graphs G, these heuristics can be applied to the corresponding undirected graph $|G|$. Since the two conditions $\mathrm{d}(i,j) > d$ and $\mathrm{d}(j,i) > d$ need to be fulfilled for $\mathrm{col}(i) = \mathrm{col}(j)$ in a distance-d coloring of a directed graph G and since $\bar{\mathrm{d}}(v,w) \leq \min\{\mathrm{d}(v,w), \mathrm{d}(w,v)\}$ (see Section 2.2), we obtain a distance-d coloring for G if we have a distance-d coloring for the graph $|G|$.

An efficient way for computing a coloring of a graph is based on a greedy strategy. For a sequence of nodes $K = (w_1, \ldots, w_n)$, Algorithm 4.1 produces a distance-1 coloring for an undirected graph G. Depending on how one orders the nodes in the sequence K one obtains an abundance of distance-1 coloring algorithms, like the random sequential (RS), the larges first (LF), or the smallest last (SL) method which all have a cost of $\mathcal{O}(|V| + |E|)$ [63, Chapter 1]. It is possible to extend Algorithm 4.1 to a greedy algorithm for general distance-d colorings, given in Algorithm 4.2.

Algorithm 4.1: Greedy algorithm for a distance-1 coloring.

Input: Graph $G = (V, E)$ and sequence of nodes K.
Output: Distance-1 coloring col.

for $i = 1 : n$ **do**
  $\mathrm{col}(w_i) = 0$
end
for $i = 1 : n$ **do**
  $\mathrm{col}(w_i) = \min\{k > 0 : k \neq \mathrm{col}(w) \text{ if } \{w_i, w\} \in E\}$
end

The most crucial point in Algorithm 4.2 is the computation of the sets W_i, which are the sets of nodes with illegal colors for w_i, i.e., all nodes with distance of at most d. An easy way to define those sets is given by the relation between the

Algorithm 4.2: Greedy algorithm for a distance-d coloring.

Input: Graph $G = (V, E)$, sequence of nodes K and distance d.
Output: Distance-d coloring col.

```
for i = 1 : n do
    col(w_i) = 0
end
for i = 1 : n do
    W_i := {w ∈ V : w ≠ w_i and d(w_i, w) ≤ d}
    col(w_i) = min{k > 0 : k ≠ col(w) for all w ∈ W_i}
end
```

distances of the nodes in G and powers of the adjacency matrix $A(G)$ as discussed in Section 2.2. Since in the graph of $A(G)^d$ the nodes i and j are adjacent if and only if $\mathrm{d}(i, j) \leq d$, the sets W_i can by defined by

$$\begin{aligned} W_i &= \{w \in V : w \neq w_i \text{ and } [A(G)^d]_{w,w_i} \neq 0\} \\ &= \{w \in V : w \neq w_i \text{ and } \{w, v_i\} \in E^d\}, \end{aligned}$$

where E^d is the set of edges in the graph of the matrix $A(G)^d$. The computation of the matrix $A(G)^d$ requires at most $2\lfloor \log_2 d \rfloor$ matrix-matrix products [53, Section 4.1]. If $A(G)$ is large and sparse, fast and efficient sparse matrix-matrix multiplications are, e.g., discussed in [22, 73, 77].

Alternatively, we can solely work with the graph G and obtain the following result.

Proposition 4.2. *Let G be a graph with maximal degree $\Delta(G)$. Then Algorithm 4.2 computes a distance-d coloring with at most $\Delta(G)^d + 1$ colors and can be implemented with cost $\mathcal{O}(\Delta(G)^d n)$.*

Proof. First note that for a node w_i, $i \in \{1, \ldots, n\}$ we have

$$\begin{aligned} |W_i| &\leq \Delta(G) + \Delta(G)\left(\Delta(G) - 1\right) + \cdots + \Delta(G)\left(\Delta(G) - 1\right)^{d-1} \\ &= \Delta(G) \sum_{k=0}^{d-1} \left(\Delta(G) - 1\right)^k \leq \Delta(G) \sum_{k=0}^{d-1} \binom{d-1}{k} \left(\Delta(G) - 1\right)^k \\ &= \Delta(G)^d, \end{aligned}$$

i.e., there are at most $\Delta(G)^d$ nodes with distance at most d. Thus, there must be at least one available color for w_i in $\{1, \ldots, \Delta(G)^d + 1\}$ and therefore Algorithm 4.2 assigns a color $\mathrm{col}(w_i)$ with $\mathrm{col}(w_i) \leq \Delta(G)^d + 1$. Since this holds for all nodes w_i, Algorithm 4.2 computes a distance-d coloring with at most $\Delta(G)^d + 1$ colors.

The set W_i with $|W_i| \leq \Delta(G)^d$ can be computed with a work of at most $\mathcal{O}(\Delta(G)^d)$ by following all paths from w_i with length at most d. In addition, the minimal

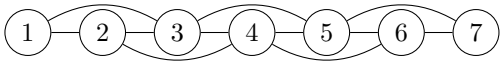

Figure 4.1: Graph with a full 2-banded adjacency matrix.

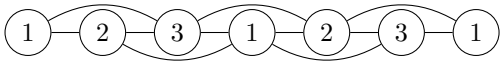

Figure 4.2: Optimal distance-1 coloring of the graph of Figure 4.1.

available color for w_i can be found in $\mathcal{O}(\Delta(G)^d + 1)$ since we know that $|W_i| \leq \Delta(G)^d$ and $\text{col}(w_i) \in \{1, \ldots, \Delta(G)^d + 1\}$. Overall this gives a cost of $\mathcal{O}(\Delta(G)^d n)$. □

In general, if $A(G)$ is large and sparse, the maximal degree $\Delta(G)$ is small and independent of the problem size n. Thus, solely working with the graph with cost $\mathcal{O}(\Delta(G)^d n)$ might be cheaper than computing the matrix $A(G)^d$ if $A(G)$ is large and sparse.

We now discuss special classes of graphs where a distance-d coloring can be given explicitly, i.e., a distance-d coloring can be obtained in $\mathcal{O}(n)$ work.

Let G be a graph with full, β-banded adjacency matrix, i.e., $[A(G)]_{i,j} = 0$ for $|i-j| > \beta$ and $[A(G)]_{i,j} \neq 0$ for $|i-j| \leq \beta$. Such a graph is illustrated in Figure 4.1 with $n = 7$ and $\beta = 2$. Then it is easy to see that an optimal distance-1 coloring is given by $c : \{1, \ldots, n\} \to \{1, \ldots, \beta + 1\}$ with

$$\text{col}(i) = (i - 1) \bmod (\beta + 1) + 1, \quad i = 1, \ldots, n,$$

and the $\beta + 1$ color classes are given by

$$\text{col}^{-1}(j) = \left\{ i : i = j + k(\beta + 1),\, k = 0, \ldots, \left\lfloor \frac{n - j}{\beta + 1} \right\rfloor \right\}, \quad j = 1, \ldots, \beta + 1.$$

Such an optimal coloring is shown in Figure 4.2, again for a 2-banded matrix. Now we know that $A(G)^d$ is a full, $d\beta$-banded matrix, and since a distance-d coloring of G is equivalent to a distance-1 coloring of the graph of $A(G)^d$, an optimal distance-d coloring for G is accordingly given by $\text{col} : \{1, \ldots, n\} \to \{1, \ldots, d\beta + 1\}$ with

$$c(i) = (i - 1) \bmod (d\beta + 1) + 1, \quad i = 1, \ldots, n$$

and

$$\text{col}^{-1}(j) = \left\{ i : i = j + k(d\beta + 1),\, k = 0, \ldots, \left\lfloor \frac{n - j}{d\beta + 1} \right\rfloor \right\}, \quad j = 1, \ldots, d\beta + 1.$$

Of course, such a coloring might not be optimal for graphs with banded adjacency matrices with zero entries within the bandwidth. However, we can use this approach as a heuristic for a distance-d coloring of graphs with banded adjacency matrices. If the matrix is just sparse and not necessarily banded (we only call

a matrix banded, if $\beta \ll n$), one could first determine a sequence of nodes $K = (w_1, \ldots, w_n)$ such that the bandwidth of the corresponding permuted adjacency matrix is reduced and preferably small. The problem of finding such a permutation is an important topic in the context of direct solvers for systems of linear equations, and lots of low-cost methods have been proposed over the last years; see, e.g., [19, 21, 42, 66, 80, 94]. The probably most familiar and fundamental heuristic for determining such a permutation is the algorithm of Cuthill and McKee proposed in [21] with cost $\mathcal{O}(|V| + |E|)$. An overview and comparison of various recent low-cost heuristics is given in [50]. Hence, if the adjacency matrix is banded with low bandwidth or if the nodes can be permuted such that the corresponding permuted adjacency matrix has a low bandwidth, then Algorithm 4.3 should provide a satisfactory distance-d coloring of the graph. The cost of Algorithm 4.3 is dominated by the computation of the sequence of nodes $K = (w_1, \ldots, w_n)$.

Algorithm 4.3: Distance-d coloring algorithm.

Input: Graph $G = (V, E)$ and distance d.
Output: Distance-d coloring col.

```
1 Compute a sequence of nodes K = (w_1, ..., w_n) such that the
  corresponding permuted adjacency matrix has a small bandwidth β
2 for i = 1 : n do
3 |  col(w_i) = (i − 1) mod (dβ + 1) + 1
4 end
```

Now assume that the graph $G = (V, E)$ is a regular D-dimensional lattice for $D > 1$. For $D = 1$ we obtain a tridiagonal adjacency matrix, so we have a special case of the banded case discussed above and an optimal distance-d coloring is known anyway, requiring $d+1$ colors. First, we notice that the greedy coloring algorithm can be specified for those graphs. If G is a regular D-dimensional lattice, then every node w can be naturally defined by coordinates $w = (w^{[1]}, \ldots, w^{[D]}) \in \mathbb{Z}^D$ such that $d(v, w) = \|v - w\|_1 = |v^{[1]} - w^{[1]}| + \cdots + |v^{[D]} - w^{[D]}|$ for $w, v \in V$. An example for such a numbering of the nodes is shown in Figure 4.3 for a two-dimensional 7×7 lattice with $0 \leq w^{[1]}, w^{[2]} \leq 6$. Hence, the sets W_i in Algorithm 4.2 can be defined by

$$W_i = \{w \in V : w \neq w_i \text{ and } \|w - w_i\|_1 \leq d\},$$

i.e., the sets W_i are all those nodes in V which lie in a ball around w_i with radius d with respect to the $\|\cdot\|_1$-norm. In addition, the maximal number of nodes which have to be examined to compute a set W_i can be specified.

Let us define $L_D(d) := |\{z \in \mathbb{Z}^D : \|z\|_1 \leq d\}|$. If w_i is a central node in the lattice, i.e., if $d \leq w_i^{[k]} \leq n - d$ for $k = 1, \ldots, D$, then $|W_i| = L_D(d) - 1$. Otherwise

$L_D(d) - 1$ represents an upper bound for $|W_i|$. It is known that

$$L_D(d) = \sum_{k=0}^{D} \binom{D}{k} \binom{d+D-k}{D}$$

where we set $\binom{n}{k} = 0$ if $n < k$ [5, Theorem 2.7]. In addition, L_D also provides a lower bound for the number of colors needed for a distance-d coloring. Since the distances between all points $z \in \mathbb{Z}^D$ in a disk with radius $\lfloor d/2 \rfloor$ is lower than or equal to d (distance and radius has to be understood with respect to the $\|\cdot\|_1$-norm), a lower bound for the chromatic number of a distance-d coloring of a D-dimensional lattice is given by $L_D(\lfloor d/2 \rfloor)$.

Instead of using the greedy approach, it is also possible to give an explicit distance-d coloring for regular D-dimensional lattices. For $D = 2$ even an optimal distance-d coloring is explicitly known.

Theorem 4.3. *Let $G = (V, E)$ be a 2-dimensional $N_1 \times N_2$ lattice. Let any node $w \in V$ be defined by its coordinates $w = (w^{[1]}, w^{[2]})$, with $0 \leq w^{[1]} \leq N_1 - 1$ and $0 \leq w^{[2]} \leq N_2 - 1$. Then an optimal distance-d coloring with $\lceil \frac{1}{2}(d+1)^2 \rceil$ colors is given by*

$$\mathrm{col}(w) = \begin{cases} (w^{[1]} + (d+1)w^{[2]}) \bmod (2m^2 + 2m + 1) & \text{if } d \text{ even}, \\ (w^{[1]} + dw^{[2]}) \bmod (2m^2 + 4m + 2) & \text{if } d \text{ odd}, \end{cases} \tag{4.1}$$

with $m = \lceil \frac{d}{2} \rceil$.

Proof. See the proof of Theorem 6 in [33]. □

For general D-dimensional lattices, an explicit *hierarchical* distance-d coloring was recursively computed in [95] for distances $d = 2^i$, $i = 0, 1, \ldots$, producing $2^{Di+1} = 2d^D$ colors. Such a distance-2^i coloring is obtained by coloring sublattices generated by a distance-2^{i-1} coloring of the lattice, starting with a classical Red-Black ordering of the nodes which represents the coloring for $i = 0$. An apparent drawback of this coloring is that for distances $d = 2^{i-1}+1, \ldots, 2^i - 1$ we necessarily need to compute a distance-2^i coloring which can result in more colors than necessary. For the applications in [95] it is actually useful to have such a *nested* coloring of the grid (which will be discussed in detail in Section 4.4), but for our purpose we are only interested in a distance-d coloring with a preferably small number of colors for arbitrary distances d. So alternatively we give the following explicit coloring for D-dimensional lattices:

Theorem 4.4. *Let $G = (V, E)$ be a D-dimensional $N_1 \times N_2 \cdots \times N_D$ lattice. Let any node $w \in V$ be defined by its coordinates $w = (w^{[1]}, \ldots, w^{[D]})$, with*

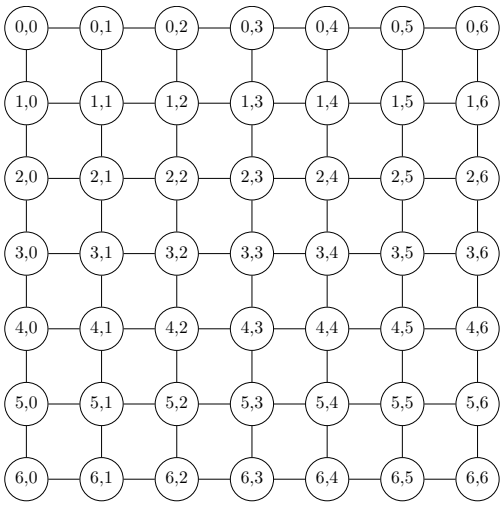

Figure 4.3: Two-dimesional 7×7 lattice, where each node is defined by two coordinates $0 \leq w_1, w_2 \leq 6$.

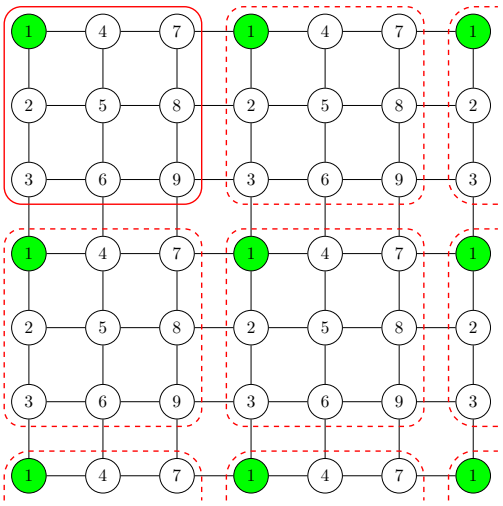

Figure 4.4: Distance-2 coloring produced by (4.2).

$0 \leq w^{[i]} \leq N_i - 1$, $i \in 1, \ldots, D$. *Then a distance-d coloring with* $(d+1)^D$ *colors is given by*

$$\mathrm{col}(w) = \left(\sum_{k=0}^{D-1} \widetilde{w^{[k]}} (d+1)^k \right) + 1 \tag{4.2}$$

where

$$\widetilde{w^{[k]}} = w^{[k]} \bmod (d+1). \tag{4.3}$$

Proof. Since for every node $w = (w^{[1]}, \ldots, w^{[D]})$ we have $\widetilde{w^{[k]}} \in \{0, \ldots, d\}$ for $k = 1, \ldots, D$, we know that (4.2) produces at most $(d+1)^D$ colors. Now assume $\mathrm{col}(w) = \mathrm{col}(v)$ for nodes $w \neq v$. We want to show that $\mathrm{d}(v, w) = \|v - w\|_1 > d$. Because of (4.3), we have $w^{[k]} = (d+1)a + \widetilde{w^{[k]}}$ and $v^{[k]} = (d+1)b + \widetilde{v^{[k]}}$ for some integers $a, b \geq 0$ and since $\mathrm{col}(w) = \mathrm{col}(v)$ we have $\widetilde{w^{[k]}} = \widetilde{v^{[k]}}$ for all $k = 1, \ldots, D$. Since $w \neq v$ there exists at least one k such that $w^{[k]} = (d+1)a + \widetilde{w^{[k]}} \neq (d+1)b + \widetilde{v^{[k]}} = v^{[k]}$ which is equivalent to $a \neq b$ for $d \geq 0$. By fixing such a k we obtain

$$\mathrm{d}(w, v) = \|w - v\|_1 \geq |w^{[k]} - v^{[k]}| = (d+1)|a - b| \geq d + 1.$$

□

For $D = 2$ the coloring of Theorem 4.4 needs twice as many colors as the optimal coloring. However, the coloring can be used for general D-dimensional lattices where we did not find an optimal coloring in the literature for $D \geq 3$. In addition, the special structure of the coloring (4.2) will be useful at a later point since it

d	optimal	greedy	hierar.	(4.2)
1	2	2	2	4
2	5	7	8	9
3	8	8	32	16
4	13	19	32	25
5	18	23	128	36
6	25	33	128	49
7	32	36	128	64
8	41	52	128	81

d	greedy	hierar.	(4.2)
1	2	2	8
2	12	16	27
3	20	128	64
4	42	128	125
5	60	1024	216
6	102	1024	343
7	133	1024	523
8	202	1024	729

Table 4.1: Number of colors for the various distance-d colorings of D-dimensional lattices with $D = 2$ (left) and $D = 3$ (right).

has two important characteristics. First of all, the construction is based on the fact that we color the nodes in the cube defined by all nodes w with $0 \leq w^{[k]} \leq d$ for all $k = 1, \ldots, D$ with $(d+1)^D$ colors as illustrated in Figure 4.4 (red, solid square). This coloring is then repeated by shifting the initial cube through the entire grid (red, dashed squares). Secondly, with this coloring every color class represents a coarse grid, where the distances between the nodes in one color class are multiples of the distance $d + 1$. This is exemplarily illustrated in Figure 4.4 where the green nodes represent one color class, defining the coarse grid. Thus, based on the structure of this special coloring, we have a more detailed information about the distances of the nodes within one color class which will be helpful for an error analysis in one of the following sections.

We will now briefly compare the number of colors obtained by the greedy approach and the explicit colorings of Theorem 4.4 and the hierarchical coloring from [95], where we consider $D = 2, 3$ and different values of d. Of course, the number of colors for the hierarchical coloring and for the coloring (4.2) is independent of the lattice sizes $N_1, \ldots, N_D$. For the greedy approach we used a lexicographic ordering of the nodes. For large lattices, the number of colors only slightly changes if we vary the lattice sizes. This can be explained by the fact that the greedy approach assigns colors by only using local information of the graph. Hence, we can assume that the number of color obtained by the greedy approach is also independent of the lattice sizes if $d \ll N_k$ for $k = 1, \ldots, D$.

Table 4.1 shows the number of colors obtained by the different approaches. For $D = 2$ the chromatic number is explicitly known such that the other approaches (which can be used for general D) can be compared to an optimal coloring in this case. We see that the greedy approach produces a surprisingly good coloring where the number of colors only slightly differs from the chromatic number. The coloring (4.2) produces twice as many colors than necessary, but it performs much better compared to the hierarchical coloring from [95], due to the fact that with the

hierarchical coloring only distance-2^i colorings can be produced. This is even more conspicuous for three dimensional lattices. Although the coloring (4.2) produces around three times as many colors as the greedy approach, we again want to emphasize, that the coloring (4.2) can be produced with a cost of $\mathcal{O}(n)$ (and in parallel), and that the special structure of the coloring is useful at a later point.

4.2 Sparse approximations of functions of matrices

The decay in functions of matrices establishes the possibility for computing $f(A)$ or, more precisely, a sparse approximation of $f(A)$. In Chapter 3 we provided an exponential decay property of matrix functions for a large class of functions f and matrices A. As a motivation for the existence of a sparse approximation for exponentially decaying matrices, we have the following essential result from [9].

Theorem 4.5. *Let $\{A_n\}$ be a sequence of $n \times n$ matrices having an exponential decay with respect to a sequence of graphs $\{G_n\}$ with bounded maximal degree $\Delta(G_n) \leq c$ for all n. Then for every $\epsilon > 0$, A_n contains at most $\mathcal{O}(n)$ entries greater than ϵ in magnitude.*

In addition, for matrices with exponential off-diagonal decay, we already know from Theorem 3.3 that there exists a banded approximation of such matrices with $\mathcal{O}(n)$ entries such that the error in the $\|\cdot\|_1$-norm does not depend on the matrix dimension. We now want to find a generalization of Theorem 3.3 for matrices with general exponential (not necessarily off-diagonal) decay. For this we define the level sets of a node $j \in \{1, \ldots, n\}$ in a graph as

$$\begin{aligned} L_n^{(k)}(j) &:= \{i : 1 \leq i \leq n, \mathrm{d}(i,j) = k\}, \quad k = 0, \ldots, n-1, \\ L_n^{(\infty)}(j) &:= \{i : 1 \leq i \leq n, \mathrm{d}(i,j) = \infty\}. \end{aligned}$$

Note that for any node j we have

$$\{1, \ldots, n\} = \cup_{k=0}^{n-1} L_n^{(k)}(j) \cup L_n^{(\infty)}(j),$$

since every node i has either a distance smaller than $n-1$ from j or cannot be reached from j in which case $i \in L_n^{(\infty)}(j)$. With these notation we can give the following generalization of Theorem 3.3.

Theorem 4.6. *Let $\{A_n\}$ be a sequence of $n \times n$ matrices with exponential decay property*

$$|[A_n]_{ij}| \leq C q^{\mathrm{d}(i,j)}, \qquad q < 1$$

with respect to a sequence of graphs $\{G_n\}$ with polynomial bounded level sets, i.e., for each node $j \in \{1, \ldots, n\}$ we have

$$|L_n^{(k)}(j)| \leq K\, k^{\alpha}$$

for $K > 0$ and $\alpha > 0$. For $m > 0$ define the matrix $A_n^{[m]}$ via

$$[A_n^{[m]}]_{ij} = \begin{cases} [A_n]_{ij} & \text{if } \mathrm{d}(i,j) \le m \\ 0 & \text{otherwise} \end{cases}. \tag{4.4}$$

Then for $\epsilon > 0$ there exists an $\tilde{m}$ independent of n such that $\|A_n - A_n^{[m]}\|_1 < \epsilon$ for all $m \ge \tilde{m}$.

Proof. Let $m_1 := m_1(q, \alpha)$ be such that $k^\alpha q^{\frac{k}{2}} < 1$ holds for $k > m_1$. Then for $m > m_1$ we obtain

$$\begin{aligned}
\|A_n - A_n^{[m]}\|_1 &= \max_{j=1}^{n} \sum_{i=1}^{n} \left|[A_m]_{ij} - [A_n^{[m]}]_{ij}\right| = \max_{j=1}^{n} \sum_{i:\,\mathrm{d}(i,j)>m} \left|[A_n^{[m]}]_{ij}\right| \\
&\le \max_{j=1}^{n} \sum_{i:\,\mathrm{d}(i,j)>m} C q^{\mathrm{d}(i,j)} \le \max_{j=1}^{n} C \sum_{k=m+1}^{n-1} |L_n^{(k)}(j)| q^k \\
&\le CK \sum_{k=m+1}^{n-1} k^\alpha q^k \le CK \sum_{k=m+1}^{\infty} k^\alpha q^{\frac{k}{2}} q^{\frac{k}{2}} \\
&\le CK \sum_{k=m+1}^{\infty} q^{\frac{k}{2}} \le CK \, \frac{q^{\frac{m+1}{2}}}{1-\sqrt{q}}.
\end{aligned}$$

Let $m_2 := m_2(q, \epsilon)$ be such that

$$CK \, \frac{q^{\frac{m+1}{2}}}{1-\sqrt{q}} < \epsilon$$

holds for $m > m_2$. Then the assertion holds for $\tilde{m} = \max\{m_1, m_2\}$. □

Clearly, for $\alpha = 0$ we just obtain the statement of Theorem 3.3 and now we have a similar results for more general important cases, e.g., when the graphs $\{G_n\}$ are regular D-dimensional lattices, where $\alpha = D-1$ (see Lemma 4.33 in Section 4.4.2).

Note that the condition on the sequence of graphs G_n of Theorem 4.5, i.e., a bounded maximal degree, is not sufficient to formulate an analogous result. This was done in [9] but the following example disproves this assertion.

Example 4.7. Let $0 < q < 1$, let $t \in \mathbb{N}$ be such that $tq > 1$ holds, and let G_n be the full t-ary tree with height p, which gives $n = 1 + t + \cdots t^p = (t^{p+1} - 1)/(t-1)$. Then the maximal degree of the graphs G_n can be bounded by $\Delta(G_n) \le t$. Let j be the root of this tree so that the level set $L_n^{(k)}(j)$ is formed exactly by all nodes at depth k in the tree, implying

$$|L_n^{(k)}(j)| = t^k, k = 0, \ldots, p, \;\; L_n^{(\infty)}(j) = \emptyset.$$

Let A_n be the matrix with $[A_n]_{ij} = q^{\mathrm{d}(i,j)}$. Then A_n has exponential decay with respect to G_n, and for all m we have

$$\begin{aligned}\|A_n - A_n^{[m]}\|_1 &\geq \sum_{i \in D_n^m(j)} |[A_n^{[m]}]_{ij}| \\ &= \sum_{k=m+1}^{p} |L_n^{(k)}(j)| q^k \\ &= \sum_{k=m+1}^{p} t^k q^k \\ &\geq (p-m)(tq)^{m+1},\end{aligned}$$

where the last inequality holds because of $tq > 1$. Thus, the first m for which $\|A_n - A_n^{[m]}\|_1 < 1$ holds is $m = p = \Omega(\log n)$, in which case we have $A_n^{[m]} = A_n$. $\diamond$

Obviously, the problem with this example is the exponential growth of the level sets while "only" having an exponential decay in A_n. For graphs $\{G_n\}$ with exponentially growing level sets, one needs to make sure that $tq < 1$ holds, where $L_n^{(k)}(j) \leq K\, t^k$ for all nodes $j = 1, \ldots, n$, and then the result holds in this case as well.

Summarizing, for a large class of functions f and matrices A, we motivated the existence of an approximation of a decaying matrix functions $f(A)$ with $\mathcal{O}(n)$ entries and with an error which is independent of the matrix size. We discuss efficient low-cost methods for the computation of such sparse approximations in the next section.

4.2.1 Computing a sparse approximation

So far we only established the *existence* of a sparse approximation of exponentially decaying matrix functions but it is not immediately clear how to compute those approximations. A naive idea is to use decay bounds as given in Chapter 3 to determine all entries $[f(A)]_{ij}$ which are larger than a given threshold ϵ (which are at most $\mathcal{O}(n)$, based on Theorem 4.5) and then compute only these entries via the Lanczos or Arnoldi process introduced in Section 2.1.2. However, the approximation of $\mathcal{O}(n)$ bilinear forms leads to a complexity of at least $\mathcal{O}(n^2)$. Hence, for computing an approximation in linear cost, decay bounds for matrix functions must be exploited in a different way. In this section we will discuss two approaches for the computation of a sparse approximation. The first is based on a polynomial series expansion of the function f. The chosen polynomial series is motivated by the decay property in $f(A)$. The second approach requires a

distance-d coloring of $G(A)$, the graph of A, as discussed in Section 4.1 where a suitable distance d can be chosen with respect to decay bounds for $f(A)$.

Let f be a function which can be expressed as a polynomial series expansion

$$f(z) = \sum_{k=0}^{\infty} c_k p_k(z) \text{ for } z \in \mathbb{E}, \tag{4.5}$$

where $p_k(z)$ are polynomials of degree k. If A is a normal matrix with $\sigma(A) \subset \mathbb{E}$ then

$$f(A) = \sum_{k=0}^{\infty} c_k p_k(A).$$

Hence an approximation of $f(A)$ is naturally given by the truncated series

$$\sum_{k=0}^{m} c_k p_k(A), \tag{4.6}$$

which is a polynomial in A of degree at most m, since the polynomials p_k, $k = 1, \ldots, m$ are polynomials of degree k. Thus, the number m should be rather small and independent of n in order to guarantee the sparsity of this approximation. Depending on the properties of the function f on the set $\mathbb{E}$, there could be several series expansions of the form (4.5), e.g., the Taylor series for infinitely differentiable functions at $a \in \mathbb{C}$ with $\mathbb{E} = \{z \in \mathbb{C} : |z - a| < R\}$ and $p_k(z) = (z - a)^k$. In the following we are interested in a polynomial series expansion of f such that $f(A)$ is well approximated by (4.6) for a small number m (, i.e., such that (4.6) converges rapidly to $f(A)$ for increasing m) with easy computable matrices $p_k(A)$, $k = 1, \ldots, m$. We naturally have a fast convergence of the approximation (4.6) for fast decaying coefficients c_k. Hence, we are looking for an expansion such that the corresponding coefficients c_k rapidly become small in magnitude for increasing k. In the following we concentrate on normal matrices with spectrum on a line segment.

In Section 2.3.1 we introduced Chebyshev polynomials and the Chebyshev series expansion for functions f that are continuous on a line segment. If the spectrum of A lies on a line segment $[\lambda_1, \lambda_2]$, then $f(A)$ can be approximated by

$$f(A) \approx P_m(A) := \sum_{k=0}^{m} c_k T_k(t(A)) \tag{4.7}$$

with

$$c_k = \frac{2}{\pi} \int_{-1}^{1} \frac{f \circ t^{-1}(x) T_k(x)}{\sqrt{1 - x^2}} \mathrm{d}x, k \geq 1, \tag{4.8}$$

where t is the affine linear transformation which maps the line segment $[\lambda_1, \lambda_2]$ to the interval $[-1, 1]$. Based on the three-term recurrence relation (2.12) we accordingly have

$$T_{k+1}(A) = 2AT_k(A) - T_{k-1}(A), \quad k = 1, 2, \ldots \tag{4.9}$$

with $T_0(A) = I$ and $T_1(A) = A$. Hence, $m-1$ matrix-matrix products are required to compute the approximation $P_m(A)$. We will now investigate the coefficients c_k in order to see if the truncated Chebyshev series is a good choice for approximating a decaying matrix function $f(A)$:

Let us define

$$p_m^*(z) := \operatorname*{argmin}_{p_m \in \mathbb{P}_m} \max_{z \in [\lambda_1, \lambda_2]} |f(z) - p_m(z)|.$$

Then $p_m^* \circ t^{-1}$ is a polynomial of degree at most m since t (and therefore t^{-1}) is of the form $t(z) = az + b$, $a, b \in \mathbb{C}$. Because of the orthogonality of Chebyshev polynomials with respect to the weight function $(1 - z^2)^{-1/2}$ we obtain for $k \geq 1$

$$\begin{aligned} |c_k| &= \left| \frac{2}{\pi} \int_{-1}^{1} \left(f \circ t^{-1}(x) - p_{k-1}^* \circ t^{-1}(x) \right) \frac{T_k(x)}{\sqrt{1-x^2}} \mathrm{d}x \right| \\ &\leq \frac{2}{\pi} \int_{-1}^{1} \frac{|T_k(x)|}{\sqrt{1-x^2}} \mathrm{d}x \max_{x \in [-1,1]} |f \circ t^{-1}(x) - p_{k-1}^* \circ t^{-1}(x)| \\ &= \frac{4}{\pi} \max_{z \in [\lambda_1, \lambda_2]} |f(x) - p_{k-1}^*(x)| = \frac{4}{\pi} E_{k-1}(f, [\lambda_1, \lambda_2]), \end{aligned} \tag{4.10}$$

using

$$\int_{-1}^{1} \frac{|T_k(x)|}{\sqrt{1-x^2}} \mathrm{d}x = 2.$$

This means that the magnitude of the coefficients can be bounded by the error of the best polynomial approximation on a line segment $[\lambda_1, \lambda_2]$ which is also an upper bound for the entries of $f(A)$, as presented in Section 3.1. Hence, a fast decay in $f(A)$ indicates a fast convergence of $P_m(A)$ to $f(A)$. Note that actually the Chebyshev series expression of f can be viewed as an explanation for a decay in $f(A)$. If there exists a good polynomial approximation of f on a set containing the spectrum of A, then the coefficients c_k decay fast for increasing k, resulting in a fast decay in $f(A)$ based on the spreading sparsity pattern of the powers of A.

The fast convergence of the Chebyshev series is frequently used in numerical computations associated with functions on real intervals. The Chebfun software package in Matlab exploits lots of useful properties of Chebyshev polynomials for tools working with functions on intervals [99]. For example, it provides a polynomial approximation based on the truncated Chebyshev series and based on polynomial interpolation with Chebyshev nodes [99, Chapter 4].

The coefficients c_k of the Chebyshev series can be computed numerically by

$$c_k \approx \frac{2}{M} \sum_{j=1}^{M} f \circ t^{-1}(\cos(t_j)) \cos(kt_j), \quad k \geq 1, \tag{4.11}$$

where $t_j = \pi(j - \frac{1}{2})/M$ and M is the number of quadrature points. The coefficient c_0 can be approximated accordingly to the coefficients c_k with the factor $1/M$ instead of $2/M$. This is also implemented in the Chebfun toolbox for the real case.

It was already suggested in [11] to approximate a decaying matrix $f(A)$ in $\mathcal{O}(n)$ complexity by a truncated Chebyshev series. The complexity of $\mathcal{O}(n)$ was established by the fact that the number m is chosen with respect to decay bounds for $f(A)$ which are independent of the problem size n. In addition, it was recommended to increase the accuracy of the approximation while keeping the cost linear by using a dropping strategy, where for $k > m$ all entries in $T_k(t(A))$ outside the sparsity pattern of $T_m(t(A))$ are dropped and the resulting matrices are used for computing the consecutive matrices via relation (4.9). This strategy is only reasonable if the matrices $T_k(t(A))$ exhibit a decay which is comparable to that of $f(A)$ since only in this case the dropped matrices are related to the matrices $T_k(t(A))$, $k > m$. Numerical examples showed that this is actually not the case and that dropping is not necessary (and even rather detrimental) for fast decaying coefficients. In addition, when using this dropping strategy it seems to be impossible to give a practical error bound for the resulting approximation. In contrast, using the approximation $P_m(A)$ without dropping, we easily obtain the error bound

$$\|f(A) - P_m(A)\|_2 \leq \sum_{k=m+1}^{\infty} |c_k| \tag{4.12}$$

which immediately follows from $\|T_k(t(A))\|_2 \leq 1$ for normal matrices A and which was also given in [11]. This bound can be refined for functions analytic inside an ellipse $E(\rho, \lambda_1, \lambda_2)$ with focal points λ_1 and λ_2 and semi-axis $\frac{1}{4}|\lambda_1 - \lambda_2|(\rho - \rho^{-1})$ and $\frac{1}{4}|\lambda_1 - \lambda_2|(\rho + \rho^{-1})$. Note that there are already results from Bernstein which deal with the error of the truncated Chebyshev series. In [67, Theorem 68] the error $\max_{z \in E(\rho,-1,1)} |f(z) - P_m(z)|$ was bounded in order to obtain an error bound for the best polynomial approximation. This bound was already used for the decay bounds of Theorem 3.4 from [10]. As already discussed in Section 3.1, this results can only be generalized for $\lambda_1, \lambda_2 \in \mathbb{R}$, so for general line segments $[\lambda_1, \lambda_2]$ we need another approach which is based on the error of the truncated Faber series.

Proposition 4.8. *Let A be normal with spectrum contained in $[\lambda_1, \lambda_2]$ and let f be analytic in the interior of an ellipse $E(\rho, \lambda_1, \lambda_2)$ and continuous on $E(\rho, \lambda_1, \lambda_2)$ for $\rho > 1$. Let $P_m(A)$ be the truncated Chebyshev series* (4.7)*. Then it holds*

$$\|f(A) - P_m(A)\|_2 \leq C \left(\frac{1}{\rho}\right)^m,$$

with

$$C = \frac{4}{\pi(\rho-1)}\frac{2M(\rho)}{1-\rho^{-1}} \text{ and } M(\rho) = \max_{z\in E(\rho,\lambda_1,\lambda_2)} |f(z)|.$$

Proof. Based on (4.10) we will first find a bound for the error of the best polynomial approximation of f on $[\lambda_1, \lambda_2]$ of degree at most $k-1$. Since f is analytic inside of the ellipse $E(\rho, \lambda_1, \lambda_2)$ it can be expressed as a Faber series

$$\sum_{\ell=1}^{\infty} a_\ell \Phi_\ell(z) \text{ for } z \in E(\rho, \lambda_1, \lambda_2),$$

where the coefficients a_ℓ and Faber polynomials Φ_ℓ are defined in Section 2.3.2. Let

$$p_{k-1}(z) := \sum_{\ell=1}^{k-1} a_\ell \Phi_\ell(z)$$

be the truncated Faber series which is a polynomial of degree at most $k-1$. Then

$$E_{k-1}(f, [\lambda_1, \lambda_2]) \leq \max_{z\in[\lambda_1,\lambda_2]} |f(z) - p_{k-1}(z)| \leq \max_{z\in E(\rho,\lambda_1,\lambda_2)} |f(z) - p_{k-1}(z)|.$$

Based on well known error results for the truncated Faber series (see, e.g., [27] or [6] and the references therein) we obtain

$$\max_{z\in E(\rho,\lambda_1,\lambda_2)} |f(z) - p_{k-1}(z)| \leq \widetilde{C}\left(\frac{1}{\rho}\right)^k$$

where

$$\widetilde{C} = \frac{2M(\rho)}{1-\rho^{-1}}.$$

Using (4.12) we obtain

$$\|f(A) - P_m(A)\|_2 \leq \sum_{k=m+1}^{\infty} \frac{4}{\pi}\widetilde{C}\left(\frac{1}{\rho}\right)^k = \frac{4}{\pi}\frac{1}{\rho-1}\widetilde{C}\left(\frac{1}{\rho}\right)^m.$$

□

Note that the truncated Faber series was also used for obtaining the decay bound of Corollary 3.7 in Chapter 3 with the decay rate ρ^{-1}. We showed in Section 3.2 that it is possible to obtain a better decay rate for the inverse. Accordingly we can give the following bound for the inverse with faster convergence rate than the one from Proposition 4.8 for general analytic functions.

Proposition 4.9. *Let A be a normal matrix with spectrum in the line segment $[\lambda_1, \lambda_2]$ excluding the origin. Define $x := \frac{\lambda_1+\lambda_2}{\lambda_2-\lambda_1}$. Let $P_m(A)$ be the truncated Chebyshev series (4.7) with $f(z) = z^{-1}$. Then it holds*

$$\|A^{-1} - P_m(A)\|_2 \leq C \left(\frac{1}{q}\right)^m,$$

with

$$C = \frac{4}{\pi(q-1)} \frac{2}{1-q^{-2}} \max_{z\in[\lambda_1,\lambda_2]} |z^{-1}| \text{ and } q = \mathrm{e}^{\mathrm{Re}(z)} > 1,$$

where z is the solution of

$$x = \cosh(z) \text{ with } \mathrm{Re}(z) \geq 0.$$

Proof. Using normalized Chebyshev polynomials and the arguments for the proofs of Proposition 3.13 and Theorem 3.18 of Section 3.2.2 we obtain for $f(z) = z^{-1}$

$$E_{k-1}(f, [\lambda_1, \lambda_2]) \leq \widetilde{C} \left(\frac{1}{q}\right)^k$$

with

$$\widetilde{C} = \frac{2}{1-q^{-1}} \max_{z\in[\lambda_1,\lambda_2]} |z^{-1}|.$$

Thus

$$\|f(A) - P_m(A)\|_2 \leq \sum_{k=m+1}^{\infty} \frac{4}{\pi} \widetilde{C} \left(\frac{1}{q}\right)^k = \frac{4}{\pi} \frac{1}{q-1} \widetilde{C} \left(\frac{1}{q}\right)^m.$$

□

With those (a priori) error bounds, we know that a prescribed accuracy ϵ of the approximation can be reached at the latest for $m = \lfloor \log_{\frac{1}{q}} C/\epsilon \rfloor$. The other way round, if the maximal possible m is restricted due to the computational cost and storage, an upper bound for the accuracy of a computable approximation can be provided. Note that the error bounds are independent of the dimension n and that we can assume that $P_m(A)$ is sparse, since m does not depend on n, either.

For general matrices, i.e., not necessarily normal matrices with spectrum on a line segment, one could use a polynomial series expansion with respect to the Faber polynomials introduced in Section 2.3.2. An error bound can easily be obtained by using classical results for the error of the truncated Faber series which was already used in the proof of Proposition 4.8. In general, Faber polynomials Φ_k do not satisfy any recurrence relation such that it is not possible to formulate a simple update formula for the polynomials Φ_k. This is a major drawback for the

computation of a sparse approximation via Faber polynomials and in [97] ways to overcome this problem are discussed.

We again refer to the results in [22, 73, 77] for computing sparse matrix-matrix products. However, even if the number of matrix-matrix products is rather small for fast decaying matrices $f(A)$, the computation of the polynomials $T_k(A)$ might be expensive for large matrices A. Thus, we now introduce an approach which is based on a distance-d coloring of $|G(A)|$, where d is chosen with respect to decay bounds for $f(A)$, i.e., d is independent of n. We already discussed ways for computing such a coloring in Section 4.1. We showed in Proposition 4.2 that Algorithm 4.2 can be implemented with a work of $\mathcal{O}(\Delta(G)^d n)$ for a graph G with maximal degree $\Delta(G)$. If we assume that we have a sequence of matrices A_n such that $\Delta(G(A_n)) \leq c$ for all n, then the maximal degree of the graphs of A_n and the distance d are both independent of n. Hence, a distance d coloring of $G(A_n)$ can be obtained with complexity $\mathcal{O}(\Delta(G(A_n))^d n) = \mathcal{O}(n)$. In addition, we introduced important classes of graphs for which a distance-d coloring is explicitly known. Thus, we now assume that a distance-d coloring of $|G(A)|$ is cheaply available.

A method for computing a sparse approximation which uses a distance-d coloring is based on the following idea: Even if we cannot compute $f(A)$, we can approximate $f(A)v$ for a vector v via the Arnoldi/Lanczos process as introduced in Section 2.1.2. So a naive idea is to compute the i-th column $f(A)e_i$, drop small entries and use the remaining entries for the i-th column of our sparse approximation. Similar to the computation of $\mathcal{O}(n)$ entries $e_i^T f(A) e_j$, this is not feasible since we have a cost of at least $\mathcal{O}(n)$ for each computation of the form $f(A)v$. Since most of the entries of the i-th column $f(A)e_i$ are dropped anyway, it would be advantageous to choose a small number of vectors $v_1, \ldots, v_m$ where $m \ll n$, such that each vector $f(A)v_\ell$ contains information about several columns of $f(A)$ and such that $f(A)$ can be approximated by computing only m vectors $f(A)v_\ell$. In the following we discuss how to obtain suitable vectors $v_1, \ldots, v_m$ for a decaying matrix $f(A)$.

Assume we have decay bounds for the matrix $f(A)$ depending on the distance of the nodes in the graph of A as described in Chapter 3. Then for a threshold ϵ we can determine a distance d such that

$$|[f(A)]_{ij}| \leq \epsilon \text{ for } \mathrm{d}(i,j) > d. \tag{4.13}$$

Let $G(A) = (V, E)$ be the graph corresponding to A, then we can partition the set of nodes V into non-empty subsets $V_1, \ldots, V_m$, i.e.,

$$V = V_1 \cup \ldots \cup V_m \text{ and } V_\ell \cap V_p = \emptyset \text{ for } \ell, p = 1, \ldots, m,$$

such that

$$\bar{\mathrm{d}}(i,j) > 2d \text{ if } i \neq j \text{ and } i, j \in V_\ell \text{ for } \ell = 1, \ldots, m. \tag{4.14}$$

Note that we consider such a partitioning of the nodes with respect to the undirected distance $\bar{\mathrm{d}}$, and it can be realized by a distance-$2d$ coloring of the undirected graph $|G(A)|$. Let $\mathrm{col} : V \to \{1, \dots, m\}$ be the corresponding coloring, then

$$V_\ell = \{i \in V : \mathrm{col}(i) = \ell\} \quad \ell = 1, \dots, m. \tag{4.15}$$

For

$$v_\ell := \sum_{i \in V_\ell} e_i, \ \ell \in \{1, \dots, m\},$$

we define the approximation $f(A)^{[d]}$ of $f(A)$ with entries

$$[f(A)^{[d]}]_{ij} := \begin{cases} [f(A)v_\ell]_i \text{ for } j \in V_\ell & \text{if } \bar{\mathrm{d}}(i,j) \le d \\ 0 & \text{if } \bar{\mathrm{d}}(i,j) > d \end{cases}. \tag{4.16}$$

Since $\mathrm{d}(i,j) \ge \bar{\mathrm{d}}(i,j) > d$, we know that the dropped entries are smaller than the threshold ϵ, and the error of the remaining entries can also be bounded by a multiple of ϵ as shown in the following result.

Proposition 4.10. *Let $|[f(A)]_{ij}| \le \epsilon$ for $\mathrm{d}(i,j) > d$ and let $f(A)^{[d]}$ be the matrix defined by* (4.16)*. Then for every (i,j)-entry we have the error bound*

$$|[f(A)]_{ij} - [f(A)^{[d]}]_{ij}| \le \begin{cases} (|V_\ell| - 1)\epsilon \text{ for } j \in V_\ell & \text{if } \bar{\mathrm{d}}(i,j) \le d \\ \epsilon & \text{if } \bar{\mathrm{d}}(i,j) > d \end{cases}.$$

Proof. The assertion is clear for $\bar{\mathrm{d}}(i,j) > d$. So consider an (i,j)-entry with $\bar{\mathrm{d}}(i,j) \le d$. Then we have

$$[f(A)^{[d]}]_{ij} = [f(A)v_\ell]_i = \sum_{k \in V_\ell} f(A)_{ik},$$

for $j \in V_\ell$. Thus,

$$[f(A)^{[d]}]_{ij} - [f(A)]_{ij} = \sum_{\substack{k \in V_\ell \\ k \ne j}} f(A)_{ik}. \tag{4.17}$$

Now assume that $\bar{\mathrm{d}}(i,k) \le d$ holds for $k \in V_\ell$ with $k \ne j$. Then we obtain

$$\bar{\mathrm{d}}(j,k) \le \bar{\mathrm{d}}(i,k) + \bar{\mathrm{d}}(i,j) \le 2d \tag{4.18}$$

which is a contradiction to $j, k \in V_\ell$. Thus $\mathrm{d}(i,k) \ge \bar{\mathrm{d}}(i,k) > d$, and therefore we have

$$|[f(A)]_{ij} - [f(A)^{[d]}]_{ij}| \le \sum_{\substack{k \in V_\ell \\ k \ne j}} \epsilon = (|V_\ell| - 1)\epsilon.$$

□

The relation (4.18) explains why we needed to use the undirected distance $\bar{\mathrm{d}}$, since otherwise we cannot guarantee that $|[f(A)]_{ik}| \leq \epsilon$ holds with $k \in V_\ell, k \neq j$ for a structurally non-symmetric matrix A.

Proposition 4.10 immediately leads to error bounds for $\|f(A) - f(A)^{[d]}\|$ for certain matrix norms. For example, we easily obtain

$$\|f(A) - f(A)^{[d]}\|_2 \leq \|f(A) - f(A)^{[d]}\|_F \leq n(s-1)\epsilon$$

or

$$\|f(A) - f(A)^{[d]}\|_1 \leq n(s-1)\epsilon$$

for $s = \max_\ell |V_\ell|$. In addition, for some types of graphs we explicitly know the size of the color classes, hence we obtain a priori error bounds without applying a coloring algorithm. Since in general already the size of the color classes depends on the dimension n, we will always obtain error bounds which depend on the the problem size if we use the result of Proposition 4.10 for error bounds. For matrices with polynomially bounded level sets we can give an error result which does not depend on the dimension n, similar to the theoretical result for exponential decaying matrices of Theorem 4.6.

Theorem 4.11. *Let $f(A) \in \mathbb{C}^{n\times n}$ be a matrix with exponential decay away from the sparsity pattern of A. Assume for $G(A)$ that*

$$|L^{(k)}(j)| \leq K\, k^\alpha$$

holds for all $j = 1, \ldots, n$, where $L^{(k)}(j) = \{i : 1 \leq i \leq n, \bar{\mathrm{d}}(i,j) = k\}$, i.e., there is a polynomial bound for the size of the (undirected) level sets for each node in $G(A)$. Let $f(A)^{[d]}$ be the matrix defined by (4.16). Then for every $\epsilon > 0$ there exists $\tilde{d}$ independent of n such that

$$\|f(A) - f(A)^{[d]}\|_1 \leq \epsilon$$

holds for $d \geq \tilde{d}$.

Proof. Since $f(A)$ has an exponential decay property, there is a constant C and a decay rate $q < 1$, both independent of n, such that we have

$$|[f(A)]_{ij}| \leq C q^{\bar{\mathrm{d}}(i,j)}.$$

Now, for every $d > 0$ we have

$$\begin{aligned}
\|f(A) - f(A)^{[d]}\|_1 &= \max_{j=1}^{n} \sum_{i=1}^{n} |[f(A)]_{ij} - [f(A)^{[d]}]_{ij}| \\
&= \max_{j=1}^{n} \left(\sum_{\substack{i \\ \bar{\mathrm{d}}(i,j)>d}} |[f(A)]_{ij}| + \sum_{\substack{i \\ \bar{\mathrm{d}}(i,j)\leq d}} \sum_{\substack{k\in V_\ell \\ k\neq j}} |[f(A)]_{ik}| \right) \\
&\leq \max_{j=1}^{n} \left(\sum_{i=d+1}^{n-1} |L^{(i)}(j)| C q^i + \sum_{i=1}^{d} |L^{(i)}(j)| \sum_{k=d+1}^{n-1} |L^{(k)}(i)| C q^k \right) \\
&\leq \sum_{i=d+1}^{\infty} K i^\alpha C q^i + \sum_{i=1}^{d} K i^\alpha \sum_{k=d+1}^{\infty} K k^\alpha C q^k,
\end{aligned} \tag{4.19}$$

where we used (4.17) for the second equality and also again used the fact that we have $\mathrm{d}(i,k) > d$ for $k \in V_\ell, k \neq j$. As shown in the proof of Theorem 4.6 there exists a d_1 such that

$$\sum_{i=d+1}^{\infty} K i^\alpha C q^i \leq \hat{C} q^{\frac{d+1}{2}}$$

for $d > d_1$ with $\hat{C} = \frac{CK}{1-\sqrt{q}}$. Hence, we obtain

$$\|f(A) - f(A)^{[d]}\|_1 \leq \left(1 + \sum_{i=1}^{d} K i^\alpha \right) \hat{C} q^{\frac{d+1}{2}}$$

for $d > d_1$. Since $\sum_{i=1}^{d} i^\alpha < d^{\alpha+1}$, we can find d_2 such that

$$\left(1 + \sum_{i=1}^{d} K i^\alpha \right) \hat{C} q^{\frac{d+1}{2}} < \epsilon \tag{4.20}$$

for $d \geq d_2$. Then the assertion holds for $\tilde{d} = \max\{d_1, d_2\}$. □

Again, this error result includes lots of important cases, e.g., when A is banded or $G(A)$ is a regular D-dimensional grid. In practice, one can either use an error bound based on Proposition 4.10 or a bound which follows from (4.19) by using more detailed information about the level sets of the graph. This is exemplarily illustrated in the next section.

In this section we discussed two approaches. The Chebyshev approach is simple and intuitive, and it was already considered in the literature. However, the

class of matrix functions for which we can use this approach is restricted and the computation of several matrix-matrix computations might be a problem in practice for large matrices A. The second approach is based on a distance-d coloring of the graph of A. This approach can be applied to general matrices and it does not require the computation of matrix-matrix products. On the other hand we need to compute vectors of the form $f(A)v$ for several vectors v and an additional distance-d coloring of the graph of A. In the following section, several numerical examples demonstrate the accuracy of the proposed approximations.

4.2.2 Numerical examples

For demonstrating the quality of the results in Section 4.2.1, we apply the discussed approaches to two test matrices.

Our first test matrix is the Hermitian, positive definite matrix

$$A^{\text{tridiag}} := \text{tridiag}(-1, 4, -1), \tag{4.21}$$

where we know that $\sigma(A) \subset [2, 6]$, independent of the matrix dimension n which, e.g., directly follows from the Gershgorin circle theorem (see, e.g., [48, Theorem 7.2.1]). This is an important information for us since we need this spectral information for the error results introduced in Section 4.2.1.

As a second test matrix we consider the shifted skew-Hermitian matrix

$$A^{\text{schwing}} := sI + D \tag{4.22}$$

from (3.74) in Section 3.3.3 with shift $s = 5$. This matrix stems from a staggered Schwinger discretization on a periodic two-dimensional lattice and we know that the spectrum of A^{schwing} is symmetric with respect to the real axis. Using the Gershgorin circle theorem and the fact that all off-diagonal entries of A^{schwing} have modulus one, we know that the spectrum of A^{schwing} is contained in the line segment $[s + 4\mathbf{i}, s - 4\mathbf{i}]$ independent of the lattice size.

For these test matrices we first consider the computation of a sparse approximation of $f(A)$ based on the Chebyshev approach.

Results for the Chebyshev approach

In the following numerical examples we compute sparse approximations with respect to the functions $f(z) = z^{-1}$ and $f(z) = z^{-1/2}$ for each of our two test matrices A^{tridiag} and A^{schwing}.

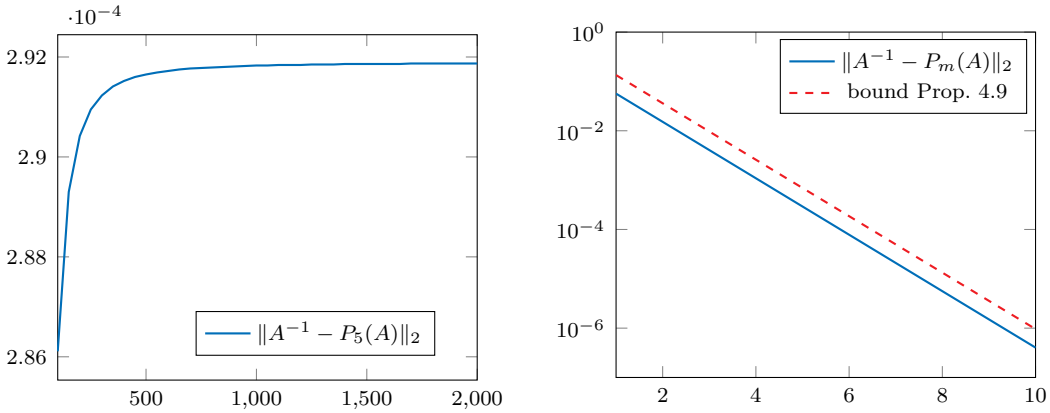

Figure 4.5: Chebyshev approach: Error $\|A^{-1} - P_m(A)\|_2$ for $A = A^{\text{tridiag}}$ for fixed $m = 5$ and different dimensions $n = 100, \ldots, 2000$ (left) and for fixed dimension $n = 1000$ and different degrees $m = 1, \ldots, 10$ (right).

Example 4.12. We consider the matrix $A = A^{\text{tridiag}}$ and the function $f(z) = z^{-1}$. For the Chebyshev approximation of A^{-1} we can use the (dimension independent) error bound of Proposition 4.9 with $\lambda_1 = 2$ and $\lambda_2 = 6$. If we fix the degree m of the approximation $P_m(A)$, then for $m = 5$ we obtain with Proposition 4.9 the error bound $\|f(A) - P_m(A)\|_2 \leq 6.9 \cdot 10^{-4}$, independent of n. The actual error of the approximation $P_m(A)$ for $m = 5$ and increasing dimensions $n = 100, \ldots, 2000$ is illustrated on the left-hand side of Figure 4.5. Here we see the convergence of the error for increasing n and the independence of the matrix size for large n. On the right-hand side of Figure 4.5, the matrix dimension is fixed while the degree m of the approximation $P_m(A)$ increases. Of course, the approximation $P_m(A)$ has bandwidth m and is therefore sparse for $m \ll n$. The predicted convergence rate of Proposition 4.9, which is also the predicted decay rate of A^{-1}, coincides with the actual convergence rate. Hence, we again see the strong relation between the decay behavior of A^{-1} and the convergence rate for the approximation $P_m(A)$ of $f(A)$. ◇

Example 4.13. We consider the matrix $A = A^{\text{tridiag}}$ and the function $f(z) = z^{-1/2}$. For the Chebyshev approximation of $A^{-1/2}$ we can use the (dimension independent) family of bounds of Proposition 4.8 with $\lambda_1 = 2$ and $\lambda_2 = 6$. We computed the minimum over these family of bounds numerically by the discrete variation of the parameter $\rho > 1$ and by considering the singularity of f in $z = 0$. If we fix the degree m of the approximation $P_m(A)$, then for $m = 5$ we obtain with Proposition 4.8 the error bound $\|f(A) - P_m(A)\|_2 \leq 1,9 \cdot 10^{-3}$, independent of n. The actual error of the approximation $P_m(A)$ for $m = 5$ and increasing dimensions $n = 100, \ldots, 2000$ is illustrated on the left-hand side of Figure 4.6. Here, the error bound overestimates the actual error by around one order of magnitude.

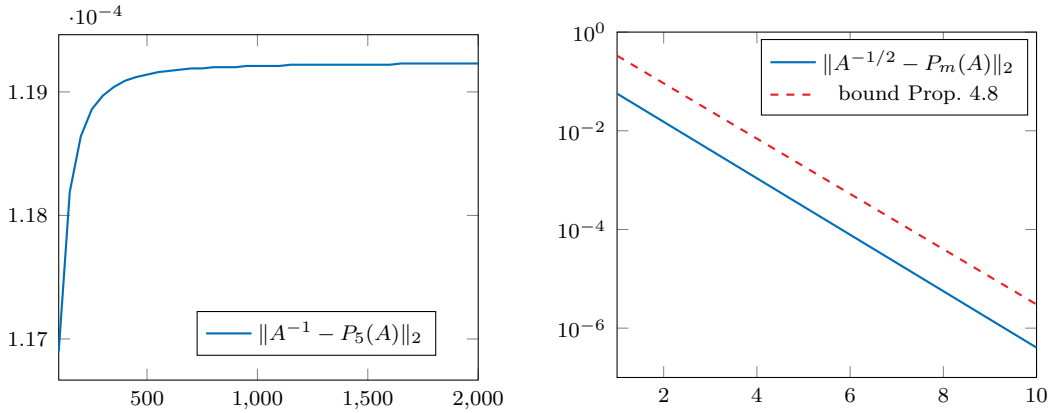

Figure 4.6: Chebyshev approach: Error $\|A^{-1/2} - P_m(A)\|_2$ for $A = A^{\text{tridiag}}$ for fixed $m = 5$ and different dimensions $n = 100, \dots, 2000$ (left) and for fixed dimension $n = 1000$ and different degrees $m = 1, \dots, 10$ (right).

On the right-hand side of Figure 4.6, the matrix dimension is fixed while the degree m of the approximation $P_m(A)$ increases. Again we see that the error bound overestimates the actual error by one order of magnitude and is worse compared to the results for the inverse. However, the predicted convergence rate of Proposition 4.8 well captures the actual convergence rate of the error. The approximation $P_m(A)$ has bandwidth m and is therefore sparse for $m \ll n$. ◇

For the same functions, we repeat the experiments for our second test matrix.

Example 4.14. We consider the matrix $A = A^{\text{schwing}}$ of dimension $n = 32^2$ and the function $f(z) = z^{-1}$. For the Chebyshev approximation of A^{-1} we can use the error bound of Proposition 4.9 applied to the line segment $[s + 4\mathbf{i}, s - 4\mathbf{i}]$. In Figure 4.7 we see the exact error $\|f(A) - P_m(A)\|_2$ and the error predicted by Proposition 4.9. Note again, that the error is independent of the matrix size, thus, the same error bound holds for a discretization with an arbitrary number of grid points. In Table 4.2 we see the error for $m = 1, \dots, 5$ and the number of nonzeros of our approximation $P_m(A)$ compared to the number of nonzeros in the sparse matrix A. Based on the fast convergence of the coefficients c_k (and of the entries of A^{-1}, respectively) we obtain a satisfying sparse approximation of A^{-1}, even for small numbers m. ◇

Example 4.15. We consider the matrix $A = A^{\text{schwing}}$ of dimension $n = 32^2$ and the function $f(z) = z^{-1/2}$. For the Chebyshev approximation of $A^{-1/2}$ we use the family of error bound of Proposition 4.8 for the line segment $[s + 4\mathbf{i}, m - 4\mathbf{i}]$. Again, we computed the minimum over these family of bounds numerically. In Figure 4.8 we see the exact error $\|f(A) - P_m(A)\|_2$ and the error predicted by

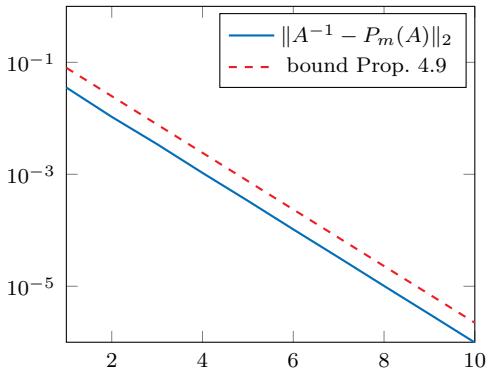

Figure 4.7: Chebyshev approach: Error $\|A^{-1} - P_m(A)\|_2$, $m = 1, \ldots, 10$ where $A = A^{\text{schwing}}$ of dimension $n = 32^2$.

m	$\|A^{-1} - P_m(A)\|_2$	$\text{nnz}(P_m(A))$
1	$3.5 \cdot 10^{-2}$	$\text{nnz}(A)$
2	$1.1 \cdot 10^{-2}$	$2.6 \cdot \text{nnz}(A)$
3	$3.5 \cdot 10^{-3}$	$5 \cdot \text{nnz}(A)$
4	$1.1 \cdot 10^{-3}$	$8.2 \cdot \text{nnz}(A)$
5	$3.4 \cdot 10^{-4}$	$12.2 \cdot \text{nnz}(A)$

Table 4.2: Chebyshev approach: Error $\|A^{-1} - P_m(A)\|_2$ where $A = A^{\text{schwing}}$ of dimension $n = 32^2$ and number of nonzeros in $P_m(A)$ ($\text{nnz}(P_m(A))$) in relation to the number of nonzeros in A ($\text{nnz}(A)$).

Proposition 4.8, which is again independent of the matrix dimension. The quality of the error bound is worse compared to that of the inverse but again the actual convergence rate is well captured by the bound of Proposition 4.8. In Table 4.3 we see the error for $m = 1, \ldots, 5$ and the number of nonzeros of our approximation $P_m(A)$ compared to the number of nonzeros in the sparse matrix A. ◇

The numerical examples show that the Chebyshev approach leads to good sparse approximations of matrix functions $f(A)$ for normal matrices A with spectrum on a line segment. The proposed error bounds (especially the bounds of Proposition 4.9 for the inverse) give us a good impression of the actual convergence of the Chebyshev approximation of $f(A)$. However, the class of matrix functions where we can apply the Chebyshev approach is restricted and the computation of several matrix-matrix products might be too expensive in practice. Thus, we now repeat the numerical examples for our second approach for the computation of $f(A)$ where no matrix-matrix products are required and which can be used for general matrix functions

Results for the coloring approach

We now consider the approach based on a graph coloring which can be applied to general matrices. In the following experiments we compute approximations of $f(A)$ for the functions $f(z) = z^{-1}$ and $f(z) = z^{-1/2}$ for each of our two test matrices A^{tridiag} and A^{schwing}.

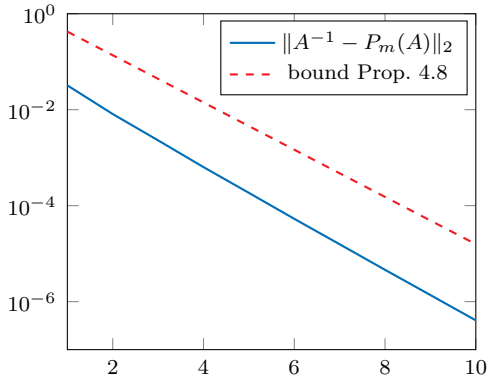

Figure 4.8: Chebyshev approach: Error $\|A^{-1/2} - P_m(A)\|_2$, $m = 1, \dots, 10$ where $A = A^{\text{schwing}}$ of dimension $n = 32^2$.

m	$\|A^{-1} - P_m(A)\|_2$	$\text{nnz}(P_m(A))$
1	$3.2 \cdot 10^{-2}$	$\text{nnz}(A)$
2	$8.1 \cdot 10^{-3}$	$2.6 \cdot \text{nnz}(A)$
3	$2.3 \cdot 10^{-3}$	$5 \cdot \text{nnz}(A)$
4	$6.4 \cdot 10^{-4}$	$8.2 \cdot \text{nnz}(A)$
5	$1.8 \cdot 10^{-4}$	$12.2 \cdot \text{nnz}(A)$

Table 4.3: Chebyshev approach: Error $\|A^{-1/2} - P_m(A)\|_2$ where $A = A^{\text{schwing}}$ of dimension $n = 32^2$ and number of nonzeros in $P_m(A)$ ($\text{nnz}(P_m(A))$) in relation to the number of nonzeros in A ($\text{nnz}(A)$).

Example 4.16. We consider the matrix $A = A^{\text{tridiag}}$ and the function $f(z) = z^{-1}$. By using Theorem 3.9 from [23] with $\lambda_{\min} = 2$ and $\lambda_{\max} = 4$ we obtain decay bounds defined by the constant C and a decay rate q. Using (4.19) and the fact that each level set in $G(A)$ has at most 2β elements if A is β-banded we obtain the error bounds

$$\|f(A) - f(A)^{[d]}\|_1 \leq 2\beta C \sum_{i=d+1}^{n-1} q^i(1 + 2d\beta) \leq 2\beta C \frac{1 + 2d\beta}{1 - q} q^{d+1}. \tag{4.23}$$

Thus, we have a (dimension independent) error bound for the matrix $A = A^{\text{tridiag}}$ with $\beta = 1$.

In the left panel of Figure 4.9 we see the error $\|f(A) - f(A)^{[d]}\|_2$ for the fixed distance $d = 5$ and increasing dimension $n = 100, \dots, 2000$. Similar to the Chebyshev approximation, the error in the $\|\cdot\|_2$-norm converges for increasing dimension, so that for large n we can consider it to be independent of the matrix size. The error bound is formulated for the error in the $\|\cdot\|_1$-norm and with (4.23) we obtain for $d = 5$ the error bound

$$\|f(A) - f(A)^{[5]}\|_1 \leq 5.6 \cdot 10^{-3},$$

whereas the actual error in the $\|\cdot\|_1$-norm is approximately given by

$$\|f(A) - f(A)^{[5]}\|_1 \approx 5.8 \cdot 10^{-4}$$

for $n = 100, \dots, 2000$, i.e., the actual error is overestimated by one order of magnitude. The same is observable if we fix the dimension and vary the distance

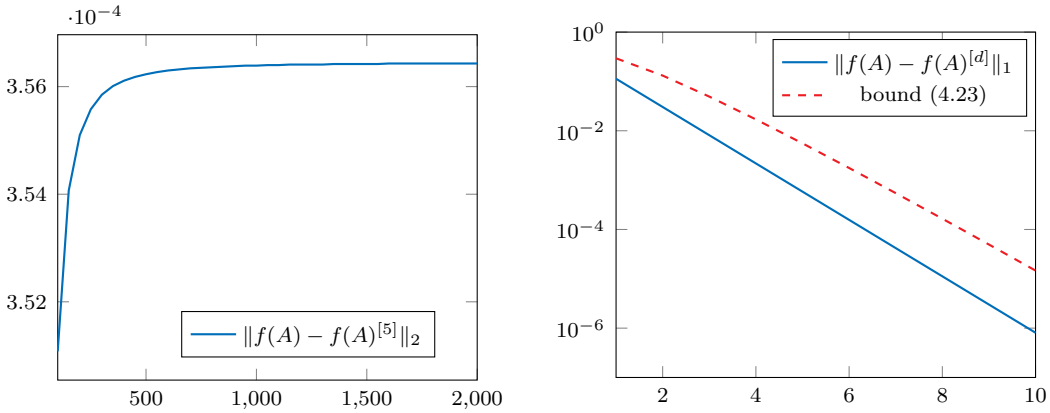

Figure 4.9: Coloring approach: Error for the approximation $f(A)^{[d]}$ with $f(z) = z^{-1}$ and $A = A^{\text{tridiag}}$ for fixed $d = 5$ and different dimensions $n = 100, \dots, 2000$ (left) and for fixed dimension $n = 1000$ and different distances $d = 1, \dots, 10$ (right).

d. In the right panel of Figure 4.9 we see the error in the $\|\cdot\|_1$-norm and the error bound (4.23) which overestimates the actual error by around one order of magnitude. The actual convergence rate is well captured by the bound but not as accurate as in the Chebyshev approach. Note that for a given distance d we need to compute a distance-$2d$ coloring resulting in $2d+1$ colors for a tridiagonal matrix, i.e., we need to compute $2d+1$ vectors $f(A)v$ for constructing the sparse approximation with bandwidth d. Hence, for $d = 10$ we obtain an accuracy of around 10^{-6} with only 21 vector computations of the form $f(A)v$. ◇

Example 4.17. We consider the matrix $A = A^{\text{tridiag}}$ and the function $f(z) = z^{-1/2}$. For the Cauchy–Stieltjes function $f(z) = 1/\sqrt{z}$ we can use the decay bound from Theorem 3.39 for obtaining an error bound with (4.23). The results are illustrated in Figure 4.10. The error bound (4.23) overestimates the actual error by around two orders of magnitude for $d = 10$ and the predicted convergence rate is slightly too slow compared to the actual convergence rate. Again, for $d = 10$ we see that it is possible to reach an accuracy smaller than 10^{-6} with only 21 vector computations.

◇

We repeat the experiments for the same functions and the staggered Schwinger matrix.

Example 4.18. We consider the matrix $A = A^{\text{schwing}}$ of dimension $n = 32^2$ and the function $f(z) = z^{-1}$. For error bounds of our approximation we use the decay bounds of Theorem 3.18 for the inverse of matrices with spectrum on a line segment. The actual error of an approximation for the inverse (in the $\|\cdot\|_1$-norm) is

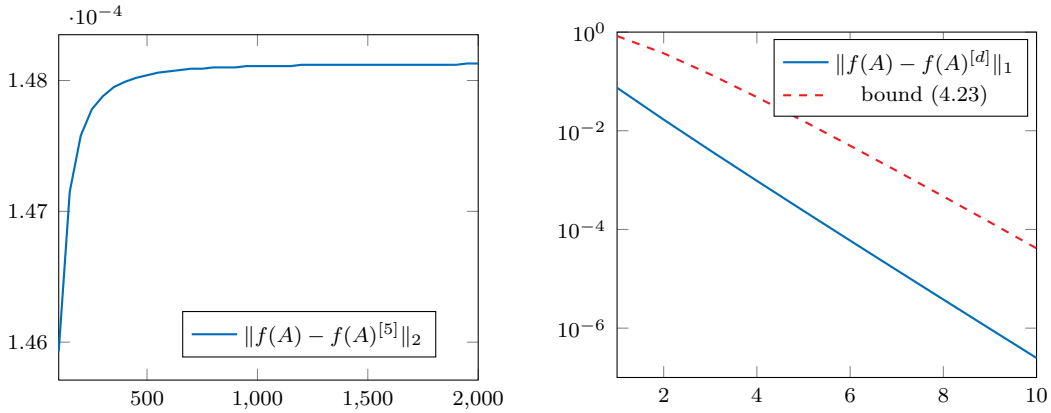

Figure 4.10: Coloring approach: Error for the approximation $f(A)^{[d]}$ with $f(z) = z^{-1/2}$ and $A = A^{\text{tridiag}}$ for fixed $d = 5$ and different dimensions $n = 100, \dots, 2000$ (left) and for fixed dimension $n = 1000$ and different distances $d = 1, \dots, 10$ (right).

illustrated in Figure 4.11 together with the error bounds based on Proposition 4.10 and based on (4.19) using the fact that the number of nodes with exactly length i in $G(A)$ is $4i$, since $G(A)$ is a periodic two-dimensional lattice. The bound based on Proposition 4.10 captures the convergence rate of the error well but the constant is way too large. The bound based on (4.19) predicts a too slow convergence, since the linear part of the bound (based on the linear growing of the level sets) initially distorts the exponential decay of the error. For growing d, the exponential part will dominate such that the predicted decay of the error then captures the actual decay for large d. In Table 4.4 we vary the distances $d = 1, \dots, 5$ resulting in m colors for the distance-$2d$ coloring of the graph, i.e., m vectors of the form $f(A)v$ have to be computed for obtaining the corresponding approximation. Note that the resulting number of vector computations is independent of the matrix dimension. In Table 4.4 we also see the number on nonzero elements in the approximation compared to the number of nonzeros in A. Hence, the full matrix $f(A)$ can be well approximated by the sparse matrix $f(A)^{[d]}$. ◇

Example 4.19. We consider the matrix $A = A^{\text{schwing}}$ of dimension $n = 32^2$ and the function $f(z) = z^{-1/2}$. For the function $f(z) = z^{-1/2}$ we use the decay bounds of Theorem 3.42 for Cauchy–Stieltjes functions of matrices with spectrum on a line segment and obtain error bounds based on Proposition 4.10 and based on (4.19). The results, which are illustrated in Figure 4.12, are comparable to the results for the inverse. In Table 4.5 we vary the distances $d = 1, \dots, 5$ resulting in m colors for the distance-$2d$ coloring of the graph, i.e., m vectors of the form $f(A)v$ have to be computed for obtaining the corresponding approximation. Note again that the number of necessary vector computations is independent of the matrix dimension. ◇

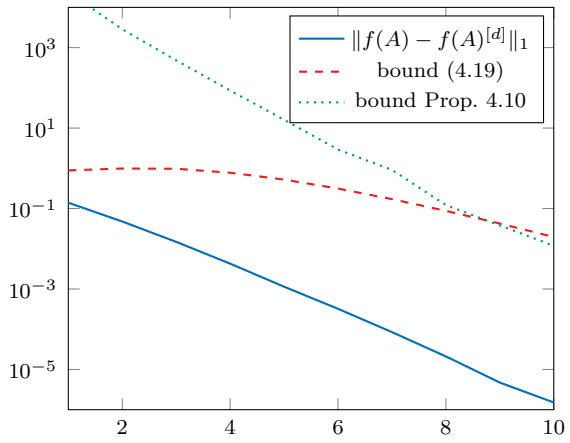

Figure 4.11: Coloring approach: Error $\|f(A) - f(A)^{[d]}\|_1$, $d = 1, \ldots, 10$ with $f(z) = z^{-1}$ where $A = A^{\text{schwing}}$ of dimension $n = 32^2$.

d	m	$\|f(A) - f(A)^{[d]}\|_2$	$\text{nnz}(f(A)^{[d]})$
1	9	$5.5 \cdot 10^{-2}$	$\text{nnz}(A)$
2	25	$1.6 \cdot 10^{-2}$	$2.6 \cdot \text{nnz}(A)$
3	44	$4.4 \cdot 10^{-3}$	$5 \cdot \text{nnz}(A)$
4	77	$1.2 \cdot 10^{-3}$	$8.2 \cdot \text{nnz}(A)$
5	131	$3 \cdot 10^{-4}$	$12.2 \cdot \text{nnz}(A)$

Table 4.4: Coloring approach: Error $\|f(A) - f(A)^{[d]}\|_2$, $d = 1, \ldots, 5$ with $f(z) = z^{-1}$ where $A = A^{\text{schwing}}$ and number of nonzeros in $f(A)^{[d]}$ ($\text{nnz}(f(A)^{[d]})$) in relation to the number of nonzeros in A ($\text{nnz}(A)$).

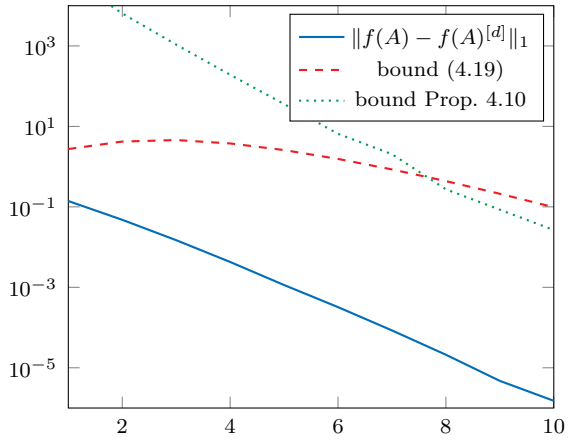

Figure 4.12: Coloring approach: Error $\|f(A) - f(A)^{[d]}\|_1$, $d = 1, \ldots, 10$ with $f(z) = z^{-1/2}$ where $A = A^{\text{schwing}}$ of dimension $n = 32^2$.

d	m	$\|f(A) - f(A)^{[d]}\|_2$	$\mathrm{nnz}(f(A)^{[d]})$
1	9	$1.2 \cdot 10^{-1}$	$\mathrm{nnz}(A)$
2	25	$3.5 \cdot 10^{-2}$	$2.6 \cdot \mathrm{nnz}(A)$
3	44	$9.7 \cdot 10^{-3}$	$5 \cdot \mathrm{nnz}(A)$
4	77	$2.5 \cdot 10^{-3}$	$8.2 \cdot \mathrm{nnz}(A)$
5	131	$6.2 \cdot 10^{-4}$	$12.2 \cdot \mathrm{nnz}(A)$

Table 4.5: Coloring approach: Error $\|f(A) - f(A)^{[d]}\|_2$, $d = 1, \ldots, 5$ with $f(z) = z^{-1/2}$ where $A = A^{\text{schwing}}$ and number of nonzeros in $f(A)^{[d]}$ ($\mathrm{nnz}(f(A)^{[d]})$) in relation to the number of nonzeros in A ($\mathrm{nnz}(A)$).

Summarizing, the Chebyshev approach leads to good approximations of $f(A)$ and provides satisfying error bounds. However, the class of matrix functions is restricted and we need to compute several matrix-matrix products for constructing the approximation. On the other hand, the approximation based on a coloring of $G(A)$ can be applied to general matrices. Instead of matrix-matrix products we need to compute several matrix-vector products for the computation of vectors of the form $f(A)v$ via the Lanczos or Arnoldi process. We saw that the coloring approach also results in accurate approximations but the proposed error bounds are inferior compared to the Chebyshev approach.

4.3 Approximation of $f(A)v$

The computation of vectors of the form $f(A)v$ for matrices $A \in \mathbb{C}^{n \times n}$, functions f defined on $\sigma(A)$ and vectors $v \in \mathbb{C}^n$ is an important task in matrix computations. One of the most important example is the solution of linear systems $Ax = b$, given by $x = A^{-1}b$, and in the previous section we needed to compute $f(A)v$ for several vectors v in order to obtain a sparse approximation of $f(A)$. In Section 2.1.2 we introduced the Arnoldi/Lanczos approximation $f(A)v \approx \|v\|_2 V_m f(H_m) e_1$, where $H_m = V_m^H A V_m$ and the columns of V_m are an orthonormal basis of the Krylov subspace $\mathcal{K}_m(A, v)$ computed by the Arnoldi/Lanczos process. Numerical examples show that a rapid decay in $f(A)$ indicates a fast converges of the Arnoldi/Lanczos approximation. This intimate relation is obvious for the inverse and Hermitian positive definite matrices: In this case, the Lanczos approximation leads to the conjugate gradients approximation for the solution of linear systems. The classical conjugate gradients convergence rate (see, e.g., [85, Section 6.11.4]) coincides with the decay rate from Theorem 3.9 for the inverse of Hermitian positive definite matrices. For general matrices A and functions f, we already know (see Section 3.1) that the entries of $f(A)$ can be bounded by

$$|[f(A)]_{ij}| \leq C \max_{z \in \Omega} |f(z) - p_m(z)|$$

where $2 \leq C \leq 1 + \sqrt{2}$ and Ω is a set containing $W(A)$, the field of values of A, for every polynomial p_m of degree $m < \mathrm{d}(i,j)$. At the same time for a vector v with $\|v\|_2 = 1$ we have the bound

$$\|f(A)v - V_m f(H_m)e_1\|_2 \leq 2\,C \max_{z\in\Omega} |f(z) - p_{m-1}(z)| \tag{4.24}$$

for every polynomial p_{m-1} of degree at most $m-1$ which can be seen as follows.

Because of Theorem 2.7 the Arnoldi approximation $V_m f(H_m)e_1$ of $f(A)$ is exact if f is a polynomial of degree at most $m-1$, i.e., we have $p_{m-1}(A)v = V_m p_{m-1}(H_m)e_1$. Thus,

$$\begin{aligned}\|f(A)v - V_m f(H_m)e_1\|_2 &= \|f(A)v - p_{m-1}(A)v + V_m p_{m-1}(H_m)e_1 - V_m f(H_m)e_1\|_2 \\ &\leq \|f(A) - p_{m-1}(A)\|_2 + \|p_{m-1}(H_m)e_1 - f(H_m)e_1\|_2.\end{aligned}$$

Using relation (3.5) and the fact that $W(H_m) \subset W(A) \subset \Omega$ gives the bound (4.24). The bound (4.24) was already used in [6] for the development of error bounds of the Arnoldi approximation of $f(A)v$.

Summarizing, a fast convergence of

$$E_m(f,\Omega) := \min_{p_m\in\mathbb{P}_m} \max_{z\in\Omega} |f(z) - p_m(z)|$$

is a sufficient condition for both, a fast decay in $f(A)$ and a fast convergence of the Arnoldi/Laczos approximation. Therefore we expect a fast convergence of the Arnoldi/Lanczos approximation if we identify a rapid decay in $f(A)$. In addition, in some cases we can exploit the decay in $f(A)$ for the computation of $f(A)v$ in order to obtain reliable error bounds and/or to save computational cost and storage. Those alternatives to the Arnoldi/Lanczos approximation are introduced and discussed in the following section.

4.3.1 Exploiting the decay for the computation of $f(A)v$

If A is normal with spectrum on a line segment, then we can use an approximation of $f(A)v$ which is based on the truncated Chebyshev series $P_m(A)$ of $f(A)$ introduced in Section 4.2. We obtain

$$f(A)v \approx P_m(A)v = \sum_{k=0}^{m} c_k T_k(t(A))v$$

where c_k are the Chebyshev coefficients (4.8), T_k are the Chebyshev polynomials of degree k and t is the linear transformation which maps the line segment containing the spectrum of A onto the interval $[-1,1]$. Based on the three term recurrence relation (2.12), we automatically have a three term recurrence relation for the

Algorithm 4.4: Chebyshev approximation of $f(A)v$.

Input: Matrix A with $\sigma(A) \subset [\lambda_1, \lambda_2]$, function f, vector v and number of iterates m.

Output: Approximation $x_m = P_m(A)v$ of $f(A)v$

Define $t(z) := \frac{\lambda_1+\lambda_2-2z}{\lambda_2-\lambda_1}$.
$\tilde{A} = t(A)$
$v_0 = v$
$v_1 = \tilde{A}v$
$x_0 = c_0 v_0$
for $k = 1, \ldots, m$ **do**
    Compute c_k via (4.11).
    $x_k = x_{k-1} + c_k v_k$
    $v_{k+1} \leftarrow 2\tilde{A}v_k - v_{k-1}$
end

vectors $v_k := T_k(t(A))v$, similar to the three term recurrence relation for the basis vectors of the Krylov subspace in the Lanczos process. The resulting algorithm is given in Algorithm 4.4.

Such a Chebyshev approximation of $f(A)v$ was already considered in the paper [25] from 1989 for symmetric matrices A where also the approximation based on the Lanczos process was discussed. Today, the Lanczos or Arnoldi process seems to be the first choice for the computation of $f(A)v$. However, a fast decay in $f(A)$ indicates a fast convergence of the Chebyshev series, hence, the Chebyshev approximation of $f(A)v$ seems to be a good choice in this case as well. In contrast to the Lanczos/Arnoldi approximation, we need information about the spectrum of A and the numerical computation of the coefficients c_k via (4.11) is necessary. On the other hand, we do not need to store an $n \times m$ matrix or to compute a matrix $f(H_m)$, and the three term recurrence relation holds for general normal matrices with spectrum on a line segment. In addition, with Proposition 4.8 and Proposition 4.9 we immediately obtain error bounds by using $\|f(A)v - P_m(A)v\|_2 \leq \|f(A) - P_m(A)\|_2 \|v\|_2$.

Besides the Chebyshev approach we now consider an approach which can only be applied to the inverse since it uses the residual $r = b - Ax$. Thus, we will call it *the approach based on the residual*. A similar approach was already used for the construction of preconditioners, which can be found in [7, Section 5] and the references therein. Now we will use this idea to approximate a decaying vector $x = A^{-1}b$. In the following we use MATLAB notation to refer to specific rows or columns of a matrix or to certain entries of a vector, respectively.

If we have exponential decay bounds for A^{-1}, given by a constant C and a decay

rate q, then the entries of x can be bounded by

$$|x_i| = \left| \sum_{j=1}^{n} b_j A_{ij} \right| \leq C \sum_{j=1}^{n} |b_j| q^{\mathrm{d}(i,j)}. \tag{4.25}$$

Clearly, based on (4.25) we can only guarantee a decay in x, if b is either sparse or if there is a rapid decay in b as well. With (4.25) we can identify whether there is a decay in x and, as a consequence, whether most of the entries in x are negligible. In this case we can find a sparse approximation $\hat{x}$ of x. Using (4.25) we can determine a set

$$\mathcal{H} := \{i \in \{1, \ldots, n\} : |x_i| \leq \epsilon\}$$

for a given threshold ϵ. If we define $\mathcal{J} := \{1, \ldots, n\} \setminus \mathcal{H}$, then an approximation $\hat{x}$ of x is obviously given by

$$\hat{x}_j := \begin{cases} x_j & \text{if } j \in \mathcal{J} \\ 0 & \text{else} \end{cases} \tag{4.26}$$

and we immediately obtain the error bound $\|x - \hat{x}\|_2 \leq \sqrt{|\mathcal{H}|}\,\epsilon$. Of course, $\hat{x}$ can be approximated by using the Arnoldi/Lanczos approximation and setting

$$\hat{x}(\mathcal{J}) \approx \|b\|_2 V_m(\mathcal{J}, :) f(H_m) e_1,$$

which can be done for general functions f. For the inverse we can use the residual and approximate $\hat{x}$ by the vector

$$\tilde{x} := \operatorname*{argmin}_{x \in \mathbb{C}^n_{\mathcal{J}}} \|b - Ax\|_2, \tag{4.27}$$

where $\mathbb{C}^n_{\mathcal{J}} := \{x \in \mathbb{C}^n : x_j = 0 \text{ for all } j \notin \mathcal{J}\}$, i.e., we are looking for the solution $\tilde{x}$ with the same sparsity pattern as $\hat{x}$, minimizing the residual. Since $Ax = A(:, \mathcal{J})x(\mathcal{J})$ holds for a vector $x \in C^n_J$, the least squares problem (4.27) is equivalent to the problem of finding a solution $\tilde{y}$ with

$$\tilde{y} = \operatorname*{argmin}_{y \in \mathbb{C}^{|\mathcal{J}|}} \|b - A(:, \mathcal{J})y\|_2. \tag{4.28}$$

Since A is sparse and we only consider a few columns of A, we can assume that lots of rows in $A(:, \mathcal{J})$ are zero. If we define the set

$$\mathcal{I} := \{i \in \{1, \ldots, n\} : \text{ there exists a } j \in \mathcal{J} \text{ such that } A(i, j) \neq 0\},$$

the least squared problem (4.28) is equivalent to

$$\tilde{y} = \operatorname*{argmin}_{x \in \mathbb{C}^{|\mathcal{J}|}} \|b(\mathcal{I}) - A(\mathcal{I}, \mathcal{J})y\|_2 \tag{4.29}$$

and $\tilde{x}$ can be computed by setting $\tilde{x}(\mathcal{J}) = \tilde{y}$. If most of the entries in the solution x are very small in magnitude, then the least squares problem (4.29) is much smaller than the original problem and can be solved by direct methods. For constructing a preconditioner as described in [7, Section 5], this approach was used for $b = e_i$ with $i = 1, \ldots, n$ and the sparsity pattern of powers of A was used for determining the set $\mathcal{J}$ instead of decay bounds for A^{-1}.

For an error bound of the approximation $\tilde{x}$ we can use the error of the sparse approximation $\hat{x}$ defined in (4.26).

Proposition 4.20. *Let $\tilde{x}$ be the approximation defined by* (4.27) *with respect to a given threshold ϵ. Then it holds*

$$\|x - \tilde{x}\|_2 \leq \kappa(A)\sqrt{|\mathcal{H}|}\,\epsilon.$$

Proof. Because of the minimization property (4.27) and $\hat{x} \in \mathbb{C}^n_{\mathcal{J}}$ we have $\|b - A\tilde{x}\|_2 \leq \|b - A\hat{x}\|_2$. Thus,

$$\begin{aligned}
\|x - \tilde{x}\|_2 &\leq \|A^{-1}\|_2 \|b - A\tilde{x}\|_2 \\
&\leq \|A^{-1}\|_2 \|b - A\hat{x}\|_2 \\
&\leq \|A^{-1}\|_2 \|A\|_2 \|x - \hat{x}\|_2 \\
&\leq \kappa(A)\sqrt{|\mathcal{H}|}\,\epsilon.
\end{aligned}$$

□

In the following section we will test the Chebyshev approach for general functions and the approach based on the residual for the inverse for two test matrices A.

4.3.2 Numerical examples

For testing the approaches introduced in Section 4.3.1 we use the matrices A^{tridiag} from (4.21) of dimension $n = 1000$ and A^{schwing} from (4.22) of dimension $n = 32^2$.

Results for the Chebyshev approach

In the following examples we illustrate the quality of the Chebyshev approximation of $f(A)v$ for the function $f(z) = z^{-1/2}$ and vector $v = e_1 + e_n$ compared to that of the Lanczos and Arnoldi approximation.

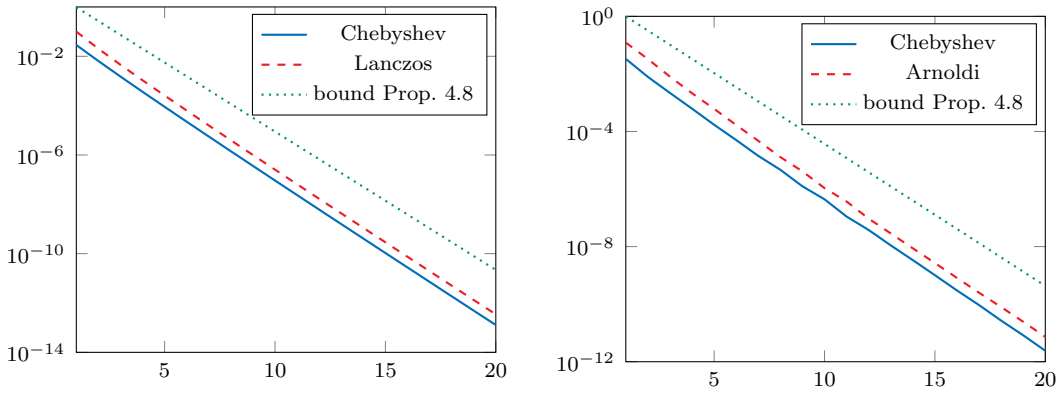

Figure 4.13: Chebyshev approach: Error of the approximations of $f(A)v$ in $\|\cdot\|_2$-norm with $f(z) = z^{-1/2}$ and $v = e_1 + e_n$ for $A = A^{\text{tridiag}}$ (left) and $A = A^{\text{schwing}}$ (right).

Example 4.21. For the matrix $A = A^{\text{tridiag}}$ we obtain error bounds for the Chebyshev approximation of $f(A)v$ based on Proposition 4.8. The minimum over the family of bounds is computed numerically, where we need to consider the singularity of f at $z = 0$. The left-hand side of Figure 4.13 shows the error of the Chebyshev approximation $P_m(A)v$, the error of the Lanczos approximation $\|v\|_2 V_m f(H_m)e_1$ and the error bound based on Proposition 4.8 for $m = 1, \ldots, 20$. The number m is in addition the number of necessary matrix-vector products for both, the computation of $P_m(A)v$ and $\|v\|_2 V_m f(H_m)e_1$. Figure 4.13 demonstrates the fast convergence of the Lanczos approximation of $f(A)v$. Other numerical examples, not shown here, also show that the Lanczos approximation with at most $m = 20$ iterations leads to an accurate approximation of $f(A)v$, independent of the matrix dimension. However, we see that the approximation based on the Chebyshev approach is slightly better than the Lanczos approximation and we obtain an error bound that captures the actual convergence of the methods. Note that the error bound does not depend on the matrix dimension n and we obtain the same error bound for $A = A^{\text{tridiag}}$ and vector $v = e_1 + e_n$ for arbitrary n. ◇

Example 4.22. We repeat the experiment for $A = A^{\text{schwing}}$ and obtain similar results as in the previous example, illustrated on the right-hand side of Figure 4.13. We have a fast convergence of the Arnoldi approximation but the Chebyshev approximation results in slightly better approximations. The convergence rate in the bound of Proposition 4.8 well captures the actual convergence. ◇

In these two examples, the Chebyshev approach leads to slightly better approximations than the Lanczos and Arnoldi approximations while needing the same number of matrix-vector products. For the Chebyshev approximation we need to find an interval containing the spectrum A but we do not need to store an

$n \times m$ matrix or to compute a matrix $f(H_m)$. In addition, we have a three term recurrence relation for general normal matrices with spectrum on a line segment.

We intentionally did not show numerical examples for the inverse, since there are already an abundance of iterative methods for the solution of systems of linear equations (see, e.g, [85]). A strongly related iterative method for the solution of systems of linear equations is given by Chebyshev iteration (see, e.g., [48, Section 11.2]). Instead of the Chebyshev series this approach uses a normalized Chebyshev polynomial similar to what was done in Section 3.2 for the development of decay bounds for the inverse.

For the inverse we now show numerical examples for the (non-iterative) method based on the residual.

Results for the approach based on the residual

We again consider our two test matrices and the right-hand side $b = e_1 + e_n$. With decay bounds for A^{-1} given by a constant C and a decay rate $0 < q < 1$ we then obtain the bound

$$|x_i| \leq C(q^{\mathrm{d}(1,i)} + q^{\mathrm{d}(n,i)})$$

for the entries of the solution $x = A^{-1}b$. For our examples, the distances between the nodes are easy available.

Example 4.23. For $A = A^{\mathrm{tridiag}}$ we can use the decay bounds of Theorem 3.9 from [23] for Hermitian positive definite matrices. The results for the approximation $\tilde{x}$ for a given threshold ϵ are shown in Table 4.6. For the computation of $\tilde{x}$ we need to solve a least-squares problem of size $|\mathcal{J}| \times |\mathcal{I}|$. Since the sets $\mathcal{J}$ and $\mathcal{I}$ are computed with respect to decay bounds for A^{-1}, the size of the least squares problem is independent of the matrix dimension. Based on the fast decay in A^{-1} and accurate decay bounds, we only have to solve very small-dimensional least-squared problems in order to obtain a good sparse approximation of the vector $x = A^{-1}b$. The resulting error lies one order of magnitude below the threshold ϵ, but the error bound based on Theorem 4.20 overestimates the error by about two orders of magnitude. ◇

Example 4.24. For the matrix $A = A^{\mathrm{schwing}}$ we can use the decay bounds from Theorem 3.18 for matrices with spectrum on a line segment. Table 4.7 shows the results for the approximation $\tilde{x}$ of $x = A^{-1}b$. For a given threshold ϵ, the dimensions $|\mathcal{J}|$ and $|\mathcal{I}|$ of the least-squares problem are larger compared to the previous example, but note again that the size of the least-squares problem is independent of the matrix dimension for our test matrix $A = A^{\mathrm{schwing}}$. Again, the resulting error lies one order of magnitude below the threshold ϵ, but the error

ϵ	$\lvert\mathcal{J}\rvert$	$\lvert\mathcal{I}\rvert$	$\lVert x - \tilde{x}\rVert_2$	error bound
10^{-1}	4	6	$2.8 \cdot 10^{-2}$	9.5
10^{-2}	6	8	$7.6 \cdot 10^{-3}$	$9.4 \cdot 10^{-1}$
10^{-3}	10	12	$5.4 \cdot 10^{-4}$	$9.4 \cdot 10^{-2}$
10^{-4}	14	16	$3.8 \cdot 10^{-5}$	$9.4 \cdot 10^{-3}$

Table 4.6: Results for the approximation $\tilde{x}$ of $x = A^{-1}b$ for $A = A^{\text{tridiag}}$ of dimension $n = 1000$.

ϵ	$\lvert\mathcal{J}\rvert$	$\lvert\mathcal{I}\rvert$	$\lVert x - \tilde{x}\rVert_2$	error bound
10^{-1}	4	24	$6.7 \cdot 10^{-2}$	3.9
10^{-2}	24	40	$5.3 \cdot 10^{-3}$	$3.8 \cdot 10^{-1}$
10^{-3}	60	112	$3.3 \cdot 10^{-4}$	$3.8 \cdot 10^{-2}$
10^{-4}	112	144	$2.8 \cdot 10^{-5}$	$3.7 \cdot 10^{-3}$

Table 4.7: Results for the approximation $\tilde{x}$ of $x = A^{-1}b$ for $A = A^{\text{schwing}}$ of dimension $n = 32^2$.

bound based on Theorem 4.20 overestimates the error by about two orders of magnitude. ◇

The two examples showed that it is possible to obtain a good sparse approximation of a decaying vector $f(A)v$ by solving a least squares problem of small and n-independent size. For both examples the factor $\sqrt{|\mathcal{H}|}$ in the error bound of Theorem 4.20 depends on the dimension n and it would be advantageous to formulate an n-independent error bound, but we couldn't find one. By now, the threshold ϵ based on decay bounds for A^{-1} gives us a better impression of the actual error than the proposed error bound.

The numerical examples for the two approaches showed that it can be profitable to consider the decay property in $f(A)$ for the computation of $f(A)v$ and that it is even possible to obtain better results than the classical used Lanczos or Arnoldi approximation of $f(A)v$.

4.4 Approximation of the trace of matrix functions

The computation of the trace of matrix functions is an important task in many applications. As examples, the trace of the inverse is needed in the study of

fractals [86], generalized cross-validation and its applications [46, 49], or in lattice quantum chromodynamics (QCD) [24, 90]. In graph theory, the Estrada index, a total centrality measure for networks, is defined as the trace of the exponential of the adjacency matrix of a graph [30, 45] and an analogous measure is given by the trace of the resolvent [29, Section 8.1]. For Hermitian positive definite matrices A, one can compute the log-determinant $\log(\det(A))$ as the trace of the logarithm of the matrix A. Amongst others, the log-determinant is needed in machine learning and related fields [79, 83]. For non-Hermitian matrices, the absolute value of the determinant can also be computed with the help of the trace of the matrix logarithm [3], which is also useful in many applications, e.g., if the sign of the determinant is irrelevant or explicitly known. Plenty of further applications can be found in [101, 102].

In this section, we consider the problem of computing

$$\operatorname{tr}(f(A)) = \sum_{i=1}^{n} [f(A)]_{ii}$$

for large, sparse matrices $A \in \mathbb{C}^{n\times n}$ and functions f defined on the spectrum of A. Of course, for small matrices $A \in \mathbb{C}^{n\times n}$, it is possible to explicitly compute $f(A)$ and extract the diagonal entries to determine $\operatorname{tr}(f(A))$. Because of the computational cost and the storage required for determining the (dense) matrix $f(A)$, this is not feasible for large, sparse matrices A. Alternatively, the computation of n bilinear forms $e_i^T f(A) e_i$, e.g., with the Lanczos or Arnoldi process, avoids the storage problem but leads to computational complexity comparable to that of a dense matrix computation. Thus, we are looking for a method where a much smaller number of bilinear forms $v^T f(A) v$ have to be computed, but where we still obtain a good approximation of $\operatorname{tr}(f(A))$. The central question is how to choose this small number of suitable vectors v. Bai et al. [3] proposed a Monte Carlo approach based on a result from [56] to estimate $\operatorname{tr}(f(A))$. Most of the trace estimators currently used are based on this Monte Carlo approach [47, 52, 101] which are sometimes called *stochastic trace estimators*. The Monte Carlo estimator for the trace of matrix functions is based on the following result from [56]:

Proposition 4.25. *Let A be an $n \times n$ matrix and let V be the discrete random variable which takes the values 1 and -1 each with probability 0.5. Let v be a vector of n independent samples from V. Then the random variable $X := v^T A v$ is an unbiased estimator of $\operatorname{tr}(A)$, i.e.,*

$$E[X] = \operatorname{tr}(A),$$

and we have

$$\operatorname{var}[X] = 2 \sum_{i \neq j} |a_{ij}|^2.$$

In the following we denote with $\mathbb{Z}_2^n$ the set of sample vectors defined in Proposition 4.25. Based on the results of Proposition 4.25 we define the Monte Carlo estimator of $\operatorname{tr}(f(A))$ as

$$\frac{1}{r}\sum_{i=1}^{r} v_i^T f(A) v_i \tag{4.30}$$

for r sample vectors $v_1, \ldots, v_r \in \mathbb{Z}_2^n$. The law of large numbers establishes the convergence of (4.30) to $\operatorname{tr}(f(A))$ for increasing r. However, we want to determine just a few expressions of the form $v^T f(A) v$ for obtaining a feasible trace estimator, i.e., we want $r \ll n$. To check the quality of this stochastic trace estimator for a number of samples r, we can use the corresponding confidence interval of the unknown parameter $E[X]$; see, e.g., [70, Chapter 9]:

For a random variable X, the $(1-\alpha)$ percent confidence interval for the unknown parameter $E[X]$ and samples $x_1, \ldots, x_r$ is given by

$$\left[\bar{x} - z_{(1-\frac{\alpha}{2})}\frac{\sigma[X]}{\sqrt{r}}, \bar{x} + z_{(1-\frac{\alpha}{2})}\frac{\sigma[X]}{\sqrt{r}}\right], \tag{4.31}$$

where $\bar{x}$ is the sample mean, $\sigma[X] = \sqrt{\operatorname{var}[X]}$ is the standard deviation of X, and $z_{(1-\frac{\alpha}{2})}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. In our setting, the samples are given by $x_i := v_i^T f(A) v_i$, $i = 1, \ldots, r$ and $\bar{x}$ is the proposed trace estimator. Unfortunately, the standard deviation $\sigma[X]$ cannot be determined in practice, since we know from Proposition 4.25 that it is given by the off diagonal entries of the unknown matrix $f(A)$. Therefore, the sample standard deviation

$$s = \left(\frac{1}{r-1}\sum_{i=1}^{r} |\bar{x} - x_i|^2\right)^{1/2}$$

can be used instead. In this case, for a small sample size we have to use the $(1-\frac{\alpha}{2})$-quantile of the t-distribution instead of the normal distribution for the $(1-\alpha)$ percent confidence interval. The confidence interval (4.31) establishes a convergence rate of $r^{-1/2}$ for the error $|E[X] - \bar{x}|$ which might be a problem in practice, as illustrated in the following example.

Example 4.26. Consider the matrix $A = A^{\text{tridiag}}$ from (4.21) of dimension $n = 1000$. We want to compute $\operatorname{tr}(A^{-1}) \approx 288,63$ with the Monte Carlo method. The left panel of Figure 4.14 shows the exact trace $\operatorname{tr}(A^{-1})$ together with the estimates and the corresponding 99% confidence interval for increasing number of sample vectors. The right hand side of Figure 4.14 illustrates the relative error for increasing number of sample vectors and an error bound based on the confidence interval. Already for a small number of samples, the relative error lies between 10^{-2} and 10^{-3}. However, the convergence of the estimator is typically very slow such that even if we reach the dimension of the matrix (where it would be possible

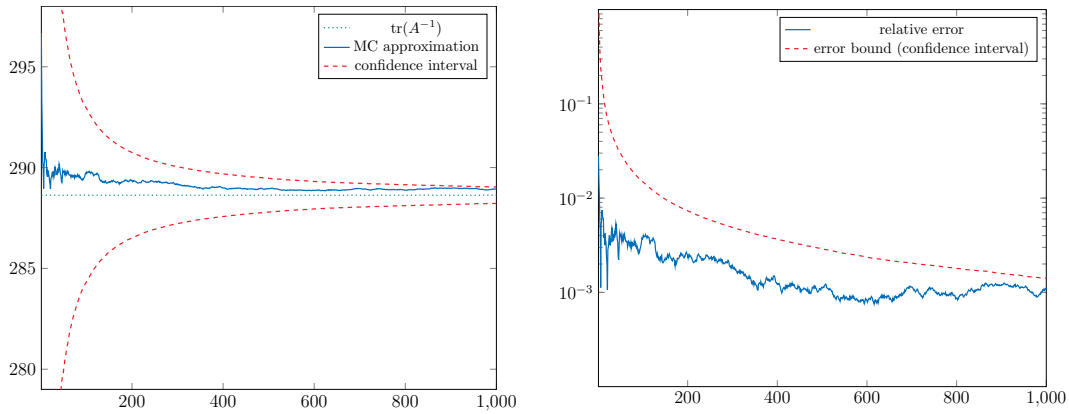

Figure 4.14: Monte Carlo approximation of $\mathrm{tr}(A^{-1})$.

to specify the exact trace with the same number of vectors) the relative error still ranges between 10^{-2} and 10^{-3}. This example illustrates, that we can not reach a certain accuracy, say a relative error of 10^{-4}, with a feasible number of samples. $\diamond$

We want to emphasize that the Monte Carlo approach is a black box method, i.e., no information about A or f is required to obtain an approximation of $\mathrm{tr}(f(A))$. However, we saw in Example 4.26 that increasing the effort (e.g., by increasing the number of sample vectors) does not necessarily result in a significant better approximation. An advancement of the classical Monte Carlo approach is given by the Multilevel Monte Carlo method [44, Section 1.2] which additionally requires approximations of the considered random variable:

Assume we have a sequence of approximations $X_0, \dots X_L$ for a random variable X with increasing accuracy. If $E[X_L]$ is a good approximation of $E[X]$ but X and X_L have similar variances, do not benefit from estimating $E[X_L]$ instead of $E[X]$. Therefore, we use the linearity of the expectation operator and obtain the identity

$$E[X_L] = E[X_0] + \sum_{\ell=1}^{L} E[X_\ell - X_{\ell-1}]. \tag{4.32}$$

Now, a Monte Carlo approach can be used to determine all or just some of the occurring expected values in (4.32) and we obtain a Multilevel Monte Carlo method. The sequence $X_0, \dots X_L$ should fulfill two requirements to make the Multilevel Monte Carlo method profitable: For small $\ell \in \{0, \dots, L\}$ the required expected values should be either known, be easily computable by a direct method, or an evaluation with large number of samples should be realizable in an efficient way. For large ℓ, it should be possible to estimate the expected values accurately with just a few samples, i.e., we expect small variances of the random variables $(X_\ell - X_{\ell-1})$. As an example, for a Multilevel Monte Carlo method for an estimator of

$\operatorname{tr}(f(A))$ we assume we have a sequence of matrices $A_0, \ldots, A_L$ which approximate the matrix $f(A)$. Then we define $X_\ell := v^T A_\ell v$, where $v \in \mathbb{Z}_2^n$. If we have $f(A) \approx A_L$, we obtain $\operatorname{tr}(f(A)) = E[X] \approx E[X_L] = \operatorname{tr}(A_L)$. Then we write

$$\operatorname{tr}(f(A)) \approx \operatorname{tr}(A_L) = E[v^T A_L v] = E[v^T A_0 v] + \sum_{\ell=1}^{L} E[v^T (A_\ell - A_{\ell-1}) v]$$

where we see that the sequence of matrices $A_0, \ldots, A_L$ should be chosen such that the sum of off diagonal entries in $A_\ell - A_{\ell-1}$ is small for large ℓ, since then the variance of $(X_\ell - X_{\ell-1})$ is small and just a few samples are needed for computing a good approximation of the corresponding expected values. The rest of the required expected values should be computable in an efficient way. Of course, the most challenging task in the Multilevel Monte Carlo method is to determine suitable variables $X_0, \ldots, X_L$, i.e., matrices $A_0, \ldots, A_L$ in our situation.

We want to find alternatives for the classical stochastic trace estimator by exploiting the decay in matrix functions. To overcome the problem of the slow convergence of the stochastic trace estimator as illustrated in Example 4.26 we have basically two possibilities. First, we try to reduce the variance such that a better estimator can be obtained for a small number of samples vectors. In Section 4.4.1 we will use the decay property in matrix functions to implement this. Note that reducing the variance does not change the convergence of the method, given by $r^{-1/2}$, where r is the number of sample vectors. Hence, a second idea is to find a method which has a faster convergence compared to the Monte Carlo approach. Such (non-stochastic) methods will be discussed in Section 4.4.2.

4.4.1 Acceleration of the basic Monte Carlo method

In this section we use spectral information of A for obtaining an improved stochastic estimator of $\operatorname{tr}(f(A))$. Similar to the previous sections we will discuss two approaches: The first is based on a Chebyshev expansion of f, while the second is based on a coloring of $G(A)$.

As discussed in Section 4.2, for normal matrices A with spectrum on a line segment and functions f such that $f(A)$ has a fast decay, $f(A)$ can be well approximated by the truncated Chebyshev series $P_m(A)$ of degree m. Consequently, the trace of $f(A)$ can be well approximated by the trace of $P_m(A)$. A suitable m can be determined with the a priori error bounds given in Proposition 4.8 and Proposition 4.9. Although the determined m is independent of the matrix dimension, the computation of $P_m(A)$ might be a problem due to the required matrix-matrix multiplications. However, we can use the truncated Chebyshev series to formulate a Multilevel Monte Carlo method for the computation of $\operatorname{tr}(f(A))$. Here, we define the sequence of matrices approximating $f(A)$ by $A_\ell = P_\ell(A)$, $\ell = 0, \ldots, L$, where

L is chosen such that we have $f(A) \approx P_L(A)$. It is clear that $A_\ell - A_{\ell-1} = c_\ell T_\ell(A)$ holds. Since the coefficients c_ℓ decay exponentially, the off diagonal entries of $c_\ell T_\ell(A)$ should be rather small for large ℓ and thus, the variance of the estimator $v^T c_\ell T_\ell(A) v$, where v is a $\mathbb{Z}_2^n$-random vector, is small such that $\operatorname{tr}(c_\ell T_\ell(A))$ can be well approximated with a small number of sample vectors. This motivates the usage of Chebyshev polynomials for a Multilevel Monte Carlo method. So for $0 \leq s \leq L$ we have

$$\operatorname{tr}(f(A)) \approx \operatorname{tr}(P_L(A)) = \operatorname{tr}(P_s(A)) + \sum_{s+1}^{L} \operatorname{tr}(c_\ell T_\ell(A)),$$

where s is chosen such that $\operatorname{tr}(P_s(A))$ can be computed exactly and $\operatorname{tr}(c_\ell T_\ell(A))$, $\ell = s+1, \ldots, L$ can be approximated by the stochastic trace estimator. The computation of $v^T c_\ell T_\ell(A) v$ can be realized by ℓ matrix-vector multiplications, based on the three term recurrence relation (4.9). Using p sample vectors for each estimation of $\operatorname{tr}(c_\ell T_\ell(A))$, $\ell = s+1, \ldots, L$ leads to $(L-s)p$ sample vectors, in total. Based on the exponential decay of the coefficients c_ℓ it could be also reasonable to decrease the number of sample vectors for increasing ℓ, based on the decreasing variance of the random variables. The resulting algorithm (with a fixed number of sample vectors per trace) is given in Algorithm 4.5.

Algorithm 4.5: Stochastic Chebyshev estimator of $\operatorname{tr}(f(A))$.

Input: Matrix A with $\sigma(A) \subset [\lambda_1, \lambda_2]$ and function f.
Output: Estimator t of $\operatorname{tr}(f(A))$.

1 Determine suitable L via Proposition 4.8 or Proposition 4.9.
2 Choose $0 \leq s \leq L$ such that $\operatorname{tr}(P_s(A))$ is computable.
3 Set $t = \operatorname{tr}(P_s(A))$.
4 **for** $\ell = s+1, \ldots, L$ **do**
5 | Produce p sample vectors $v_1, \ldots, v_p \in \mathbb{Z}_2^n$.
6 | $t \leftarrow t + \frac{1}{p} \sum_{i=1}^{p} v_i^T c_\ell T_\ell(A) v_i$
7 **end**

For general decaying matrices $f(A)$ one can use a method which is based on a coloring of $G(A)$. If $f(A)$ has an exponential decay property, then we can assume that many entries in $f(A)$ are very small in magnitude. Since the variance of the stochastic trace estimator is determined by the magnitude of the off-diagonal entries, it would be advantageous if the variance would be only given by those small off-diagonal entries. Hence, we divide the whole trace of $f(A)$ into *subtraces* which can be computed stochastically with corresponding small variances. This can be done as follows.

Let d be a preferably small distance such that

$$|[f(A)]_{ij}| \leq \epsilon \text{ for } \mathrm{d}(i,j) > d \tag{4.33}$$

for a given threshold ϵ. Such a d can be determined by appropriate decay bounds for $f(A)$. Let $V_1, \dots, V_m$ be a partition of $V = \{1, \dots, n\}$ such that

$$\mathrm{d}(i,j) > d \text{ if } i \neq j \text{ and } i,j \in V_\ell \text{ for } \ell = 1, \dots, m \tag{4.34}$$

in $G(A)$ which can be found by a distance-d coloring of $G(A)$ and using (4.15). Then clearly we have

$$\operatorname{tr}(f(A)) = \sum_{\ell=1}^{m} \operatorname{tr}_\ell(f(A))$$

where

$$\operatorname{tr}_\ell(f(A)) := \sum_{i \in V_\ell} [f(A)]_{ii} \tag{4.35}$$

are subtraces of $f(A)$ which now can be estimated stochastically by using sample vectors

$$v^{[\ell]} = \sum_{i \in V_\ell} x_i e_i \tag{4.36}$$

where $x_i \in \mathbb{Z}_2^1$. Using p sample vectors for each subtrace, we overall need to compute mp bilinear forms. Based on the small variance of the estimator for the subtraces, we expect a small number of samples to be necessary in order to obtain a good approximation. Note that for each subtrace the variance of the estimator is given by

$$2 \sum_{\substack{i,j \in V_\ell \\ i \neq j}} |[f(A)]_{ij}|^2$$

which can be easily bounded by

$$2|V_\ell|(|V_\ell| - 1)\epsilon^2,$$

i.e., for some types of matrices, where we know $|V_\ell|$, we can formulate an a priori confidence interval for our estimator without computing a distance-d coloring. The algorithm for such an estimator is summarized in Algorithm 4.6 for a given distance d.

This approach is already used in lattice QCD for the inverse (see, e.g., [4] or [41] and the references therein), where it is called *dilution*. Motivated by the decay in matrix functions, this approach can be used for general matrices and general functions. In addition, using decay bounds for matrix functions, we can formulate upper bounds for the variances of the estimators, leading to a priori error bounds for the estimator based on the confidence interval (4.31) in some cases. Amongst others, the deviation of such bounds is discussed in the next section for an easy example.

Algorithm 4.6: Stochastic estimator of $\operatorname{tr}(f(A))$ based on a distance-d coloring.

Input: Matrix A, function f and distance d.
Output: Estimator t for $\operatorname{tr}(f(A))$.

```
Compute the sets V_1, ..., V_m with a distance-d coloring of G(A).
Set t = 0.
for ℓ = 1, ..., m do
    Compute sample vectors v_1^[ℓ], ..., v_p^[ℓ] given by (4.36).
    t ← t + (1/p) Σ_{i=1}^{p} v_i^[ℓ]T f(A) v_i^[ℓ]
end
```

4.4.1.1 Numerical examples

In this Section we demonstrate the quality of the proposed trace estimators for our test matrices A^{tridiag} from (4.21) and A^{schwing} from (4.22) for $f(z) = z^{-1}$.

Results for the Chebyshev approach

Example 4.27. We consider the matrix $A = A^{\text{schwing}}$ of dimension $n = 32^2$. The exact trace is approximately given by $\operatorname{tr}(A^{-1}) \approx 178.95$. For the Multilevel Monte Carlo method based on the Chebyshev series, we choose $L = 15$ and $s = 3$, i.e., we need to compute $\operatorname{tr}(P_3(A))$ directly and the remaining $L - s$ traces are computed by the stochastic trace estimator. The left panel of Figure 4.15 shows the relative error of the classical Monte Carlo approach compared to the Multilevel Monte Carlo approach for increasing number of sample vectors. Note that we need to compute $L - s = 12$ remaining traces stochastically, hence, the total number of sample vectors are multiples of 12 for the Multilevel Monte Carlo method if we use the same number of samples per trace. Additionally, the relative error $|\operatorname{tr}(A^{-1}] - \operatorname{tr}(P_3(A))| / \operatorname{tr}(A^{-1})$ is shown, which is comparable to the relative error of the Monte Carlo estimator. Hence, we see that the additional (stochastically) computed traces improve the relative error by around one order of magnitude. $\diamond$

Example 4.28. We consider the matrix from Example 4.26, i.e., $A = A^{\text{tridiag}}$ of dimension $n = 1000$. The right-hand side of Figure 4.15 shows the results for $L = 15$ and $s = 1$, i.e., the trace of $P_1(A)$ is computed directly and the remaining traces are computed stochastically. We see that the Chebyshev approach leads to slightly better estimators compared to the classical Monte Carlo method. However, we also plotted the relative error for $\operatorname{tr}(P_1(A))$ which does already has an accuracy of around 10^{-4} because of the fast convergence of the Chebyshev series. Hence,

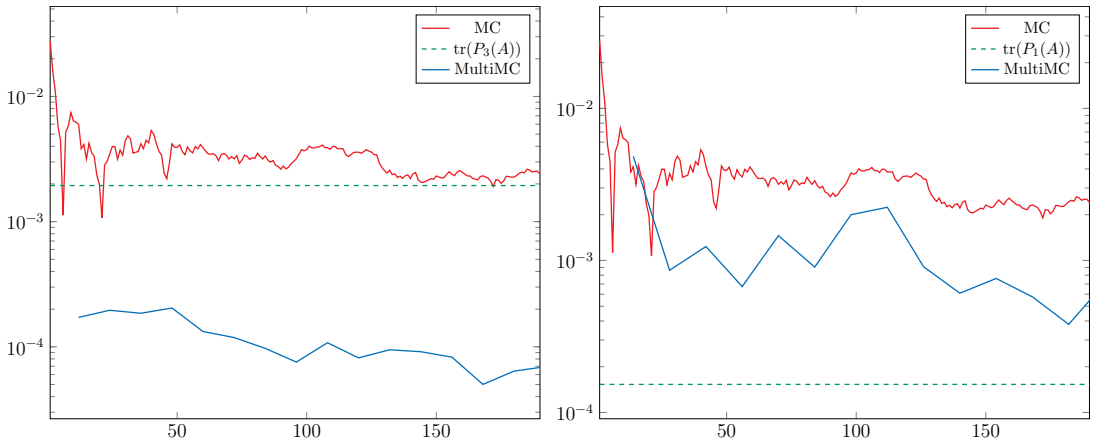

Figure 4.15: Relative errors for the estimators of $\mathrm{tr}(A^{-1})$ based on the Chebyshev series (MultiMC), for the classical Monte Carlo estimator (MC) and for the trace of the matrix $P_s(A)$ with $A = A^{\mathrm{schwing}}$ (left) and $A = A^{\mathrm{tridiag}}$ (right).

we do not benefit from the additional stochastically computed traces (not even for $s = 1$) but quite the contrary, so the Multilevel approach is not profitable here. ◇

The two examples indicates that the Multilevel Monte Carlo based on the Chebyshev series leads to better results compared to the classical Monte Carlo method. However, for fast decaying matrices already the trace of $P_s(A)$ with small s can be an accurate approximation of the actual trace and the additional stochastically computed traces can be destructive as illustrated in Example 4.28.

Results for the coloring approach

We repeat the experiments for the approach based on a distance-d coloring for the same matrices and the same function $f(z) = z^{-1}$.

Example 4.29. We consider the matrix $A = A^{\mathrm{schwing}}$ of dimension $n = 32^2$. We choose the distance $d = 3$, resulting in eight color classes, i.e., the total number of sample vectors are multiples of eight if we use the same number of sample vectors for each subtrace induced by the color classes. In the left panel of Figure 4.16 we see the relative error of the classical Monte Carlo method compared to the approach based on a distance-d coloring for increasing (total) number of sample vectors. Already for the small distance $d = 3$ we have an improvement of around two orders of magnitude. ◇

Example 4.30. We repeat the experiments for the matrix $A = A^{\mathrm{tridiag}}$ of dimension $n = 1000$ considered in Example 4.26. We use the distance $d = 5$, resulting

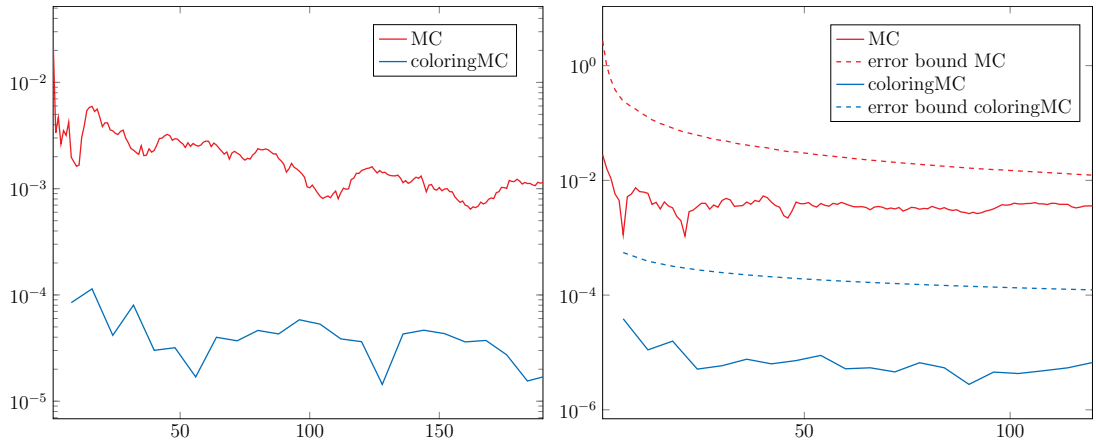

Figure 4.16: Relative errors for the estimators of $\mathrm{tr}(A^{-1})$ based on the coloring approach (coloringMC) and for the classical Monte Carlo estimator (MC) for $A = A^{\mathrm{schwing}}$ with $d = 3$ (left) and $A = A^{\mathrm{tridiag}}$ with $d = 5$ (right).

in six color classes, i.e., we need to compute six subtraces. The right-hand side of Figure 4.16 depicts the results for the relative error compared to the classical Monte Carlo estimator for increasing (total) number of sample vectors, where we used the same increasing number of sample vectors for each subtrace. In this example, already the distance $d = 5$ leads to an improvement of around three orders of magnitude. Note again that for each subtrace, the variance of the Monte Carlo estimator is given by

$$2 \sum_{\substack{i,j \in V_\ell \\ i \neq j}} |[f(A)]_{ij}|^2$$

which can be easily bounded by using decay bounds for $f(A)$. We now derive a priori error bounds for the approximation based on the coloring approach for $A = A^{\mathrm{tridiag}}$.

For the matrix $A = A^{\mathrm{tridiag}}$ we can use the decay bounds of Theorem 3.9 from [23]. Using an obvious coloring of $G(A)$ with $d + 1$ colors, we obtain

$$\begin{aligned} 2 \sum_{\substack{i,j \in V_\ell \\ i \neq j}} |[A^{-1}]_{ij}|^2 &\leq 2 \sum_{\substack{i,j \in V_\ell \\ i \neq j}} C^2 q^{2\mathrm{d}(i,j)} \\ &\leq 4C^2 |V_\ell| \sum_{i=1}^{\infty} q^{2di} \\ &\leq 4C^2 \left\lceil \frac{n}{d+1} \right\rceil \frac{q^{2d}}{1 - q^{2d}} \end{aligned}$$

for each color class V_ℓ, i.e., we obtain an upper bound for the variance of the stochastic trace estimator for each subtrace. Let $\mathrm{tr}_\ell(f(A))$ be a subtrace defined

in (4.35) and let $\bar{x}_\ell$ be the Monte Carlo estimator of $\mathrm{tr}_\ell(f(A))$ with p samples. Then the error of the approximation $\tilde{x} := \sum_{\ell=1}^{d+1} \bar{x}_\ell$ based on the $(1-\alpha)$ percent confidence interval (4.31) for the estimation of each subtrace can be bounded by

$$\begin{aligned}
|\tilde{x} - \mathrm{tr}(f(A))| &= \left| \sum_{\ell=1}^{d+1} \bar{x}_\ell - \sum_{\ell=1}^{d+1} \mathrm{tr}_\ell(f(A)) \right| \leq \sum_{\ell=1}^{d+1} |\bar{x}_\ell - \mathrm{tr}_\ell(f(A))| \\
&\leq \sum_{\ell=1}^{d+1} \frac{z_{(1-\frac{\alpha}{2})}}{\sqrt{p}} \sqrt{2 \sum_{\substack{i,j \in V_\ell \\ i \neq j}} |[f(A)]_{ij}|^2} \\
&\leq (d+1) \frac{z_{(1-\frac{\alpha}{2})}}{\sqrt{p}} 2C \sqrt{\left\lceil \frac{n}{d+1} \right\rceil} \frac{q^d}{\sqrt{1-q^{2d}}}
\end{aligned}$$

This error bound together with the error bound of the classical Monte Carlo Method is shown in Figure 4.16 as well. Both bounds are normalized in order to compare them to the exact errors. The improvement of the approximation based on the coloring compared to the classical Monte Carlo method is well captured by the bounds. Note again that the bound for the error of the Monte Carlo method is an a posteriori error bound since it uses the sample standard deviation as an estimator for the actual standard deviation. ◇

Concluding, these examples illustrate that it is possible to increase the accuracy of the classical Monte Carlo method by using the decay property in $f(A)$. We considered an example where it is possible to formulate a priori error bounds for the estimator by using decay bounds for the matrix $f(A)$ in order to bound the variance of the estimator. Those bounds can be formulated for all types of matrices A where a coloring of $G(A)$ is explicitly known. However, even if the accuracy of the estimator can be improved for decaying matrices $f(A)$, we still have a slow convergence for increasing number of sample vectors. Vividly, the error curve is only shifted downwards by a few orders of magnitude if we fix the numbers s or d, respectively. Hence, in the next section we are looking for (non-stochastic) methods which exhibit a faster convergence of the error for increasing numbers of vectors.

4.4.2 Non-stochastic approximation

Similar to the classical Monte Carlo approach, we want to develop a method which induces an approximation of $\mathrm{tr}(f(A))$ as the sum of a small number of bilinear forms $v^T f(A) v$. In order to obtain a good accuracy, we again choose the vectors based on the decay property in $f(A)$ and a coloring of $G(A)$, the graph of A.

For matrices with spectrum on a line segment we already noticed the fast convergence of $P_m(A)$ to $f(A)$ for decaying matrices $f(A)$. Hence the trace of the approximation $P_m(A)$ should result in a good estimator of $\mathrm{tr}(f(A))$. The computation of m matrix-matrix products might be to expensive in practice. Thus, we are now looking for a way to compute $\mathrm{tr}(P_m(A))$ without computing $P_m(A)$ explicitly.
For a given degree m, we can compute a partition $V_1, \dots, V_k$ such that

$$\mathrm{d}(i,j) > m \text{ if } i \neq j \text{ and } i,j \in V_\ell \text{ for } \ell = 1, \dots, k$$

via a distance-m coloring of $|G(A)|$. If we define

$$v_\ell := \sum_{i \in V_\ell} e_i,\ \ell \in \{1, \dots, k\}$$

then we see that

$$v_\ell^T P_m(A) v_\ell = \sum_{i,j \in V_\ell} [P_m(A)]_{ij} = \sum_{i \in V_\ell} [P_m(A)]_{ii}$$

holds, since P_m is a polynomial of degree m and therefore we have $[P_m(A)]_{ij} = 0$ if $d(i,j) > m$. Thus,

$$\mathrm{tr}(P_m(A)) = \sum_{\ell=1}^{k} v_\ell^T P_m(A) v_\ell,$$

i.e., $\mathrm{tr}(P_m(A))$ can be computed via km matrix-vector multiplications.

Note that this approach can be combined with the Multilevel Monte Carlo method based on the Chebyshev series introduced in the previous section. If the computation of $P_s(A)$ is too costly due to the required matrix-matrix products, then one can compute $\mathrm{tr}(P_s(A))$ with the approach described above. This is for example reasonable, when the graph of A has no regular structure, i.e., when we need to use the greedy approach for a coloring with cost $\mathcal{O}(\Delta(G(A))^m n)$. To keep the constant $\Delta(G)^m$ small, one can choose a small degree m in order to obtain a cheaply available approximation. If the resulting approximation is not accurate enough, then one can add stochastically computed traces with small variances. This *hybrid method* combines a non-stochastic approximation based on a distance-d coloring and a Monte Carlo method.

For general matrices we can use a similar approach based on a distance-d coloring of $G(A)$. Assume we have exponential decay bounds for $f(A)$. Then we can determine a distance d such that

$$|[f(A)]_{ij}| \leq C q^{\mathrm{d}(i,j)} \leq \epsilon \text{ for } \mathrm{d}(i,j) > d \tag{4.37}$$

for a given threshold $\epsilon > 0$. With a distance-d coloring of $G(A) = (V, E)$ and using (4.15) we find a partition $V_1, \ldots, V_k$ of V with

$$\mathrm{d}(i,j) > d \text{ if } i \neq j \text{ and } i,j \in V_\ell \text{ for } \ell = 1, \ldots, k \tag{4.38}$$

and we define

$$v_\ell := \sum_{i \in V_\ell} e_i, \ \ell \in \{1, \ldots, k\}. \tag{4.39}$$

Then we have

$$v_\ell^H f(A) v_\ell = \sum_{i \in V_\ell} [f(A)]_{ii} + \sum_{\substack{i,j \in V_\ell \\ i \neq j}} [f(A)]_{ij},$$

and therefore

$$\mathrm{tr}(f(A)) = \sum_{i=1}^{n} [f(A)]_{ii} = \sum_{\ell=1}^{k} v_\ell^H f(A) v_\ell - \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} [f(A)]_{ij}.$$

Now, we define our approximation for the trace of $f(A)$ as

$$\mathcal{T}(f(A)) := \sum_{\ell=1}^{k} v_\ell^H f(A) v_\ell \tag{4.40}$$

with error

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| = \left| \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} [f(A)]_{ij} \right|. \tag{4.41}$$

The computation of the proposed approximation $\mathcal{T}(f(A))$ is summarized in Algorithm 4.7.

Of course, the choice for the vectors v_ℓ is motivated by the error (4.41) since with this choice the addends in (4.41) are very small in magnitude (smaller than ϵ) and the error can be easily bounded by

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} \epsilon = \sum_{\ell=1}^{k} |V_\ell|(|V_\ell| - 1)\epsilon. \tag{4.42}$$

For matrices for which an explicit coloring of the graph is known, the bound (4.42) represents an a priori error bound without actually computing a distance-d coloring. If we assume that the size of the color classes is asymptotically given by $\mathcal{O}(n/k)$, i.e., if the nodes are distributed uniformly among the color classed, and if the number of colors k is independent of n, then the error bound (4.42) is of

Algorithm 4.7: Non-stochastic approximation of $\mathrm{tr}(f(A))$ based on a distance-d coloring.

Input: Matrix A and exponential decay bounds defined by a constant C and a decay rate q.
Output: Approximation $\mathcal{T}(f(A))$ of $\mathrm{tr}(f(A))$.

```
Select ε > 0.
Set d = log_q(ε/C).
Compute a distance-d coloring col : V → {1, ..., k} of G(A).
Set T_f(A) = 0.
for ℓ = 1, ..., k do
    Compute x_ℓ = v_ℓ^T f(A) v_ℓ with v_ℓ as in (4.39).
    T(f(A)) ← T_f(A) + x_ℓ
end
```

order $\mathcal{O}(n^2\epsilon)$. The explicit representation (4.41) of the error is of great practical interest and in the following we give a more extensive error analysis where we discuss cases in which better error bounds than the trivial bound (4.42) can be obtained.

If the decay bounds for computing the distance d stem from a bound for the error of a polynomial approximation, e.g., as in the results in Section 3.2, then it is possible to obtain a sharper error bound.

Theorem 4.31. *Let $A \in \mathbb{C}^{n\times n}$ and let f be defined on the spectrum of A. Assume*

$$|[f(A)]_{ij}| \leq \|f(A) - p_m(A)\|_2 \leq Cq^{\mathrm{d}(i,j)}$$

for a polynomial p_m of degree $m = \mathrm{d}(i,j) - 1$. Let $Cq^{\mathrm{d}(i,j)} \leq \epsilon$ for $\mathrm{d}(i,j) > d$. Then for the approximation $\mathcal{T}(f(A))$ defined by (4.40) we have

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{k} |V_\ell|\sqrt{|V_\ell| - 1}\,\epsilon. \tag{4.43}$$

Proof. First note that for a vector $x \in \mathbb{C}^n$ and a set $V_\ell \subseteq \{1, \ldots, n\}$ we have the relation

$$\sum_{i\in V_\ell} |x_i| \leq \sqrt{|V_\ell|}\left(\sum_{i\in V_\ell} |x_i|^2\right)^{1/2} \leq \sqrt{|V_\ell|}\|x\|_2.$$

Hence,

$$\begin{aligned}
\sum_{\substack{i,j\in V_\ell \\ i\neq j}} |[f(A)]_{ij}| &= \sum_{\substack{i,j\in V_\ell \\ i\neq j}} |[f(A)]_{ij} - [p_m(A)]_{ij}| \\
&= \sum_{j\in V_\ell} \sum_{\substack{i\in V_\ell \\ i\neq j}} |[f(A)]_{ij} - [p_m(A)]_{ij}| \\
&\leq \sum_{j\in V_\ell} \sqrt{|V_\ell|-1}\, \|f(A)e_j - p_m(A)e_j\|_2 \\
&\leq \sum_{j\in V_\ell} \sqrt{|V_\ell|-1}\, \|f(A) - p_m(A)\|_2 \\
&\leq \sum_{j\in V_\ell} \sqrt{|V_\ell|-1}\, \epsilon \\
&= |V_\ell| \sqrt{|V_\ell|-1}\, \epsilon.
\end{aligned}$$

Using

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} |[f(A)]_{ij}|$$

gives the result. □

If we again assume a uniform distribution of the nodes among the color classes, this improved error bound of Theorem 4.31 is of order $\mathcal{O}(n^{3/2}\epsilon)$. Depending on the coloring of $G(A)$ it is possible to give $\mathcal{O}(n\epsilon)$ error bounds based on more detailed information of the distances between the nodes in one color class. If we use the coloring of Algorithm 4.3 in Section 4.1 for matrices which can be permuted to a matrix with small bandwidth, we obtain the following improved error bound depending on the distance d.

Theorem 4.32. *Let $f(A) \in \mathbb{C}^{n\times n}$ be a matrix with exponential decay property with respect to $G(A)$, i.e.,*

$$[f(A)]_{ij} \leq Cq^{\mathrm{d}(i,j)}$$

for a constant C and a decay rate $0 < q < 1$. Let $\mathcal{T}(f(A))$ be defined by (4.40), where the vectors v_ℓ are computed with respect to a coloring produced by Algorithm 4.3 for a given distance d. Then

$$|\,\mathrm{tr}(f(A) - \mathcal{T}(f(A))| \leq 2nC\frac{q^d}{1-q^d}.$$

Proof. Based on (4.41) and using the decay bounds, the error can be bounded by

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} |[f(A)]_{ij}| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} C q^{\mathrm{d}(i,j)}, \tag{4.44}$$

with $k = d\beta+1$ where β is the bandwidth of the permuted matrix of Algorithm 4.3.

For ease of presentation, we assume that A does already have the bandwidth β from line 1 of Algorithm 4.3, i.e., $K = (1, \ldots, n)$, otherwise we first apply the corresponding permutation on A. Using Algorithm 4.3 we obtain the sets

$$V_\ell = \left\{ \ell + k(d\beta+1), k = 0, \ldots, \left\lfloor \frac{n-\ell}{d\beta+1} \right\rfloor \right\}, \quad \ell = 1, \ldots, d\beta+1.$$

Hence, a set of nodes V_ℓ can be written as

$$V_\ell = \{w_1, \ldots, w_{|V_\ell|}\}, \quad \ell = 1, \ldots, d\beta+1$$

with

$$w_k := \ell + (k-1)(d\beta+1).$$

Within a set V_ℓ, we have the distance relations

$$\mathrm{d}(w_i, w_j) > |i-j|d, \quad i,j \in \{1, \ldots, |V_\ell|\},\ i \neq j \tag{4.45}$$

which can be seen as follows.

Based on Lemma 2.11 of Section 2.2, for $i, j \in \{1, \ldots, |V_\ell|\}$ with $i \neq j$ the condition (4.45) is equivalent to $[\mathcal{A}^{|i-j|d}]_{w_i,w_j} = 0$, where $\mathcal{A}$ is the adjacency matrix of $G(A)$, i.e., $\mathcal{A}$ is a binary matrix with the same off-diagonal sparsity pattern as A and full diagonal. In particular, $\mathcal{A}$ is β-banded, thus $\mathcal{A}^{|i-j|d}$ is $|i-j|d\beta$-banded. Hence, due to

$$|w_i - w_j| = |\ell + (i-1)(b\beta+1) - (\ell + (j-1)(\beta d+1)| = |i-j|(d\beta+1) > |i-j|d\beta,$$

we have $[\mathcal{A}^{|i-j|d}]_{w_i,w_j} = 0$, i.e., $\mathrm{d}(w_i, w_j) > |i-j|d$.

Because of (4.45) we know that for every node $w \in V_\ell$ we have $\mathrm{d}(w, w_i) > 2d$ for at least $|V_\ell| - 2$ nodes $w_i \in V_\ell \setminus \{w\}$, $\mathrm{d}(w, w_i) > 3d$ for at least $|V_\ell| - 4$ nodes $w_i \in V_\ell \setminus \{w\}$ and so on. Hence we obtain

$$\sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} C q^{\mathrm{d}(i,j)} \leq \sum_{\ell=1}^{k} |V_\ell| \sum_{i=1}^{\infty} 2C q^{id} = 2Cn \frac{q^d}{1-q^d}.$$

□

A similar $\mathcal{O}(n\epsilon)$ error bound can be formulated if $G(A)$ is a regular D-dimensional lattice and if the coloring of Theorem 4.4 is used. For this we need the following lemma in order to bound the level sets in regular D-dimensional lattices.

Lemma 4.33. *Let* $L_D^=(d) := |\{z \in \mathbb{Z}^D : \|z\|_1 = d\}|$, *then* $L_d^= \leq 2Dd^{D-1}$.

Proof. We already introduced the identity

$$L_D(d) := |\{z \in \mathbb{Z}^D : \|z\|_1 \leq d\}| = \sum_{k=0}^{D} \binom{D}{k} \binom{d+D-k}{D}$$

from [5, Theorem 2.7] in Section 4.1. With this formula we obtain

$$L_D^=(d) = L_D(d) - L_D(d-1) = \sum_{k=0}^{D} \binom{D}{k} \binom{d+D-k-1}{D-1},$$

where we used $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$.

We will now use a proof technique called *double counting* (see, e.g., [1, Section 20]) to prove that

$$\sum_{k=0}^{D} \binom{D}{k} \binom{d+D-k-1}{D-1} \tag{4.46}$$

is equal to

$$\sum_{k=0}^{D-1} \binom{D}{k} \binom{d-1}{D-1-k} 2^{D-k}. \tag{4.47}$$

For this, we first give a combinatorial interpretation of equation (4.46), then formulate an equivalent statement which at last results in equation (4.47).

Let $X = \{X_1, \ldots, X_D\}$ be a set with D elements and let $Y = \{Y_1, \ldots, Y_{d-1}\}$ be a set with $d-1$ elements with $X \cap Y = \emptyset$. Then (4.46) counts the number of ways for choosing subsets $A \subseteq X$ and $B \subseteq X \cup Y$ with $|B| = D-1$ and $A \cap B = \emptyset$. This can be seen as follows.

If $0 \leq k \leq D$ is the number of elements in A, then $\binom{D}{k}$ counts the number of ways for choosing A. Since $A \cap B = \emptyset$ there are $D + (d-1) - k$ elements left for the set B. Thus, the number of ways for choosing B with $|B| = D-1$ is given by $\binom{d+D-k-1}{D-1}$. The sum over the number of elements in A gives (4.46).

Now, choosing such a $B \subseteq X \cup Y$ with $|B| = D-1$ and $A \cap B = \emptyset$ is equivalent to choosing subsets $N \subseteq X$ and $M \subseteq Y$ such that $|M| + |N| = D-1$ and $(N \cup M) \cap A = \emptyset$. Hence, we now count the number ways of choosing subsets $A \subseteq X$, $N \subseteq X$ and $M \subseteq Y$ with $|M| + |N| = D-1$ and $(N \cup M) \cap A = \emptyset$. If $1 \leq k \leq D-1$ is the number of elements in M, then there a $\binom{D}{k}$ ways for choosing

M. The number of ways for choosing the left $D-1-k$ elements of N out of Y is given by $\binom{d-1}{D-1-k}$. Since $(N \cup M) \cap A = \emptyset$ there are $D-k$ elements left for A, i.e., there are 2^{D-k} ways for choosing A. The sum over the number of elements in M gives (4.47).

As a last step, we need to bound (4.47), where we use

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n}{n-k}\frac{(n-1)!}{(n-k-1)!k!} = \frac{n}{n-k}\binom{n-1}{k},$$

$\binom{n}{k} \leq \frac{n^k}{k!}$ and $2^n \leq (n+1)!$ which can be easily shown by induction. We have

$$\begin{aligned}
\sum_{k=0}^{D-1} \binom{D}{k}\binom{d-1}{D-1-k}2^{D-k} &= \sum_{k=0}^{D-1} \frac{D}{D-k}\binom{D-1}{k}\binom{d-1}{D-1-k}2^{D-k} \\
&\leq D\sum_{k=0}^{D-1} \frac{1}{D-k}\binom{D-1}{k}\frac{(d-1)^{D-1-k}}{(D-1-k)!}2^{D-k} \\
&= 2D\sum_{k=0}^{D-1} \binom{D-1}{k}\frac{(d-1)^{D-1-k}}{(D-k)!}2^{D-k-1} \\
&\leq 2D\sum_{k=0}^{D-1} \binom{D-1}{k}(d-1)^{D-k-1} \\
&= 2Dd^{D-1},
\end{aligned}$$

where the last equality stems from the binomial theorem for $((d-1)+1)^{D-1}$. □

As a consequence of Lemma 4.33, we now know that for each node in a regular D-dimensional lattice the number of nodes with exactly distance d can be bounded by $2Dd^{D-1}$. We use this for the following result.

Theorem 4.34. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix where $G(A)$ is a regular D-dimensional lattice. Let $f(A)$ have an exponential decay with respect to $G(A)$, i.e.,*

$$[f(A)]_{ij} \leq Cq^{\mathrm{d}(i,j)}$$

for a constant C and a decay rate $0 < q < 1$. Let $\mathcal{T}(f(A))$ be defined by (4.40), where the vectors v_ℓ are computed with respect to the coloring of Theorem 4.4 for a given distance $d > 1$. Then

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| \leq 2CDn\left(\frac{D-1}{\mathrm{e}\ln(1/q)}\right)^{D-1}\frac{q^{d-1}}{1-q^{d-1}}.$$

Proof. Again, the error can obviously be bounded by

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} |[f(A)]_{ij}| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} C q^{\mathrm{d}(i,j)}, \tag{4.48}$$

where the sets V_ℓ are computed with respect to the coloring (4.2) of Theorem 4.4. One characteristic of the coloring (4.2) is that each color class represents a coarse grid where the distances between the nodes in one color class are multiples of d (see Figure 4.4). From Lemma 4.33 we know that for a D-dimensional grid, for each node the number of nodes with exactly distance i can be bounded by $2D\, i^{D-1}$, thus

$$\begin{aligned} \sum_{\ell=1}^{k} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} C q^{\mathrm{d}(i,j)} &\leq \sum_{\ell=1}^{k} |V_\ell| \sum_{i=1}^{\infty} 2D\, i^{D-1}\, C q^{id} \\ &\leq 2CDn \sum_{i=1}^{\infty} i^{D-1} q^i\, q^{i(d-1)}. \end{aligned}$$

The function $g(i) = i^{D-1}q^i$ with $i \geq 1$ attains its maximum at $i = (D-1)/\ln(1/q)$. Thus, it follows that

$$i^{D-1} q^i \leq \left(\frac{D-1}{\mathrm{e} \ln(1/q)} \right)^{D-1}$$

and

$$\begin{aligned} 2CDn \sum_{i=1}^{\infty} i^{D-1} q^i\, q^{i(d-1)} &\leq 2CDn \left(\frac{D-1}{\mathrm{e} \ln(1/q)} \right)^{D-1} \sum_{i=1}^{\infty} q^{i(d-1)} \\ &\leq 2CDn \left(\frac{D-1}{\mathrm{e} \ln(1/q)} \right)^{D-1} \frac{q^{d-1}}{1 - q^{d-1}}. \end{aligned}$$

□

Note that the bound $2D\, i^{D-1}$ for the number of nodes with exactly distance i is sharp for $D = 1, 2$ but deteriorates for increasing dimension, so for large dimensions a sharper bound can be used in order to obtain improved error bounds. The numerical examples in the next section illustrate that the error of the proposed approximations linearly scales with the dimension n of the matrix, i.e., $\mathcal{O}(n\epsilon)$ error bounds are optimal for our approximation.

We now briefly discuss some aspects with respect to additional errors caused by the computation of the bilinear forms $v^T f(A) v$ via the Arnoldi or Lanczos process.

For the computation of the approximation $\mathcal{T}(f(A))$ we need to compute k bilinear forms where k is the number of colors in the distance-d coloring. In this case we obtain an approximation $\widetilde{\mathcal{T}}_f(A)$ of $\mathcal{T}_f(A)$, i.e.,

$$\mathcal{T}(f(A)) = \sum_{\ell=1}^{k} v_\ell^H f(A) v_\ell \approx \sum_{\ell=1}^{k} \|v_\ell\|_2^2 e_1^T f(H_{m_\ell}^{(\ell)}) e_1 =: \widetilde{\mathcal{T}}(f(A)),$$

where $H_{m_\ell}^{(\ell)}$ are the matrices obtained by the Lanczos or Arnoldi process with respect to A and v_ℓ, $\ell = 1, \ldots, k$. If we want to compute bounds for the error of our approximation, we need to take into account that the bounds presented so far only represent upper bounds for the error $|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))|$. The actual error of our approximation can be bounded by

$$\begin{aligned} |\operatorname{tr}(f(A)) - \widetilde{\mathcal{T}}_f(A)| &= |\operatorname{tr}(f(A)) - \mathcal{T}_f(A) + \mathcal{T}_f(A) - \widetilde{\mathcal{T}}_f(A)| \\ &\leq |\operatorname{tr}(f(A)) - \mathcal{T}_f(A)| + |\mathcal{T}_f(A) - \widetilde{\mathcal{T}}_f(A)|, \end{aligned}$$

i.e., we have an additional error of up to $|\mathcal{T}_f(A) - \widetilde{\mathcal{T}}_f(A)|$. If the conditions are fulfilled to compute upper and lower bounds for the bilinear forms $v^T f(A) v$ with the Gauss, Gauss-Radau or Gauss-Lobatto rule (see Section 2.1.2), we can give upper and lower bounds $T^U(f(A))$ and $\mathcal{T}^L(f(A))$ for $\mathcal{T}(f(A))$ as the sums over the upper and lower bounds of the bilinear forms. Then, for an approximation $\widetilde{\mathcal{T}}(f(A)) \in [\mathcal{T}^L(f(A)), \mathcal{T}^U(f(A))]$ we have

$$|\mathcal{T}(f(A)) - \widetilde{\mathcal{T}}(f(A))| \leq T^U(f(A)) - T^L(f(A)).$$

In the next section, we give an example where we consider the additional error $|\mathcal{T}_f(A) - \widetilde{\mathcal{T}}_f(A)|$. In practice it should be sufficient to use the bounds for the error $|\operatorname{tr}(f(A)) - \mathcal{T}_f(A)|$ to get a good impression of the convergence of the approximation.

4.4.2.1 Numerical examples and comparison to previous approaches

In this section we demonstrate the efficiency of the proposed approximation of the trace of matrix function based on a distance-d coloring, and then we discuss the relation of the proposed method to other methods from the literature.

Numerical examples

It is clear that the approximation $\operatorname{tr}(P_m(A))$ of $\operatorname{tr}(f(A))$ is accurate if $P_m(A)$ is a good approximation of $f(A)$, and several numerical examples for this approximation were already given in Section 4.2. Hence, in the following we will concentrate on the coloring approach which can be applied to general matrices.

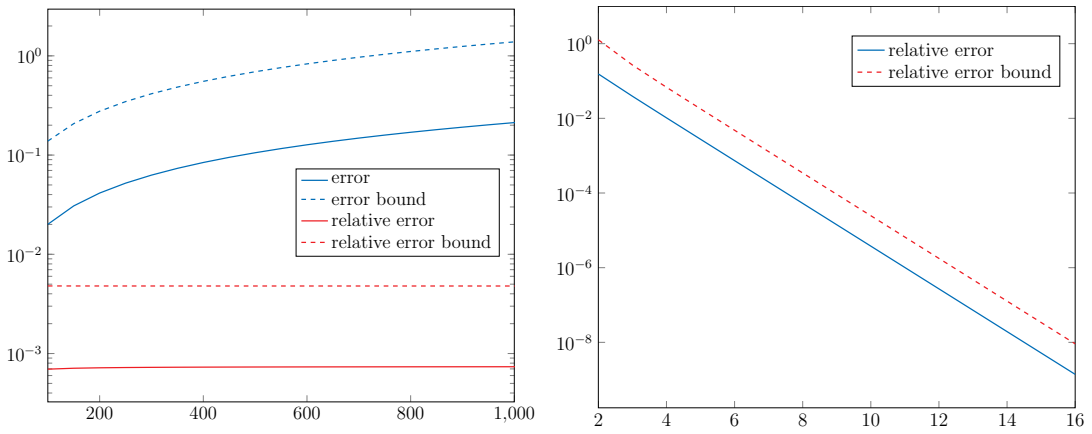

Figure 4.17: Error $|\operatorname{tr}(A^{-1}) - \mathcal{T}_f(A)|$ and error bound of Theorem 4.32 for $A = A^{\text{tridiag}}$ for increasing dimensions $n = 100, \dots, 1000$ and distance $d = 5$ (left) and for increasing distances $d = 1, \dots, 15$ and dimension $n = 1000$ (right).

We will first discuss results for the matrix $A = A^{\text{tridiag}}$ of moderate dimensions in order to check the quality of the approximation and the corresponding error bounds.

Example 4.35. We consider the matrix $A = A^{\text{tridiag}}$ and $f(z) = z^{-1}$. The left panel of Figure 4.17 shows the results of the approximation for increasing dimension and fixed distance $d = 5$. The (absolute) error bound of Theorem 4.32 perfectly captures the exact error curve of the approximation. We also depicted the relative error and the relative error bound of the approximation and we see that the relative error is almost constant for increasing dimension. Thus, the error scales linearly with the dimension n, so it seems to be natural to have error bounds which depend on the dimension n. On the right panel of Figure 4.17 we fix the dimension to $n = 1000$ and increase the distances d. The x-axis represents the number of colors in the corresponding distance-d coloring, which is just given by $d + 1$ in the tridiagonal case, so we need to determine $d + 1$ bilinear forms for the approximation of the trace. Note that this example was already considered in Example 4.26 for illustrating the convergence of the classical Monte Carlo approach, where the relative errors of the approximations linger between 10^{-2} and 10^{-3} for increasing sample vectors. In contrast, using the approach based on a distance-d coloring leads to a fast convergence of the approximation for increasing distances d corresponding to increasing number of vectors. ◇

We will now consider two examples where we cannot compute the exact traces due to the large dimensions of the problems, hence, we can only use the error bounds to check the quality of the approximation.

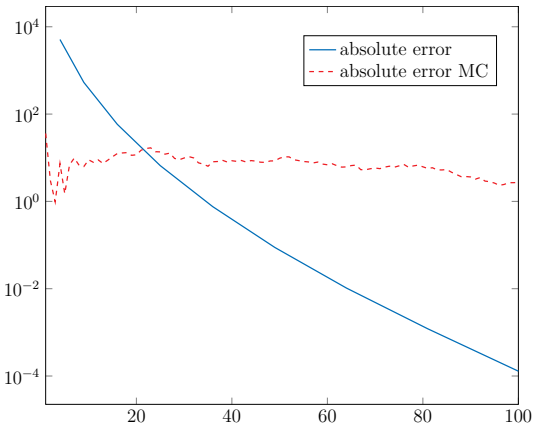

Figure 4.18: Error of the approximation of $\mathrm{tr}((\alpha I - A)^{-1})$ for $A = A^{\mathrm{adj}}$ based on a distance-d coloring compared to the Monte Carlo approach depending on the number of vectors k.

d	k	error bound	approx.
2	9	$2.1 \cdot 10^4$	$1.1 \cdot 10^5$
4	25	$7.2 \cdot 10^2$	$1.05 \cdot 10^5$
6	49	31	$1.04 \cdot 10^5$
8	81	1.4	$1.04 \cdot 10^5$
10	121	$5.9 \cdot 10^{-2}$	$1.04 \cdot 10^5$

Table 4.8: Results for the approximation of $\mathrm{tr}((\alpha I - A)^{-1})$ (approx.) with $A = A^{\mathrm{adj}}$ for increasing distances d and the error bound of Theorem 4.34.

Example 4.36. We consider a regular $10^3 \times 10^3$ grid with lexicographic ordering of the nodes $\{1, \dots, n\}$, $n = 10^6$ and remove the edges $(1, 2)$ and $(n-1, n)$. The corresponding adjacency matrix $A = A^{\mathrm{adj}}$ is of dimension $n = 10^6$. In Section 2.2 we introduced the resolvent

$$f(A) = (\alpha I - A)^{-1}$$

with $\alpha > 1$ and $\alpha \notin \sigma(A)$. This matrix function is often used for the analysis of networks, because of the relation

$$(\alpha I - A)^{-1} = \sum_{i=0}^{\infty} \alpha^{-(i+1)} A^i,$$

i.e., the factor α scales walks with certain length in order to clarify that typically short walks are more important than long walks. We now consider $\alpha = 10$ (note that $\sigma(A) \subset [-4, 4]$) i.e. walks with length i are scaled with $\left(\frac{1}{10}\right)^i$. In network analysis the trace of the resolvent is used as a normalization factor for centrality and communicability measures of graphs.

For this graph we can apply the coloring of Theorem 4.4 for $d = 1, \dots, 10$ resulting in $(d+1)^2$ color classes, such that we can use Theorem 4.34 for an error bound. For the approximation we need to compute the bilinear forms $v^T(\alpha I - A)^{-1}v$ for $(d+1)^2$ vectors v. For the computation of the bilinear forms we use the Lanczos approximation with respect to the matrix $\alpha I - A$ and the function $f(z) = z^{-1}$. For this matrix and function we can compute upper and lower bounds for the error of the approximation of $v^T(\alpha I - A)^{-1}v$‘ (see Section 2.1.2), i.e., we can compute

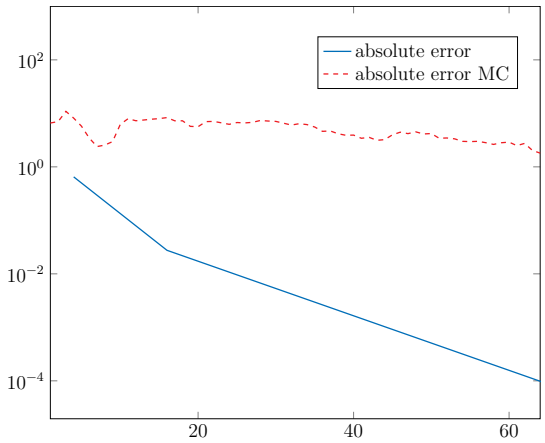

Figure 4.19: Error of the approximation of $\mathrm{tr}(A^{-1})$ with $A = A^{\mathrm{schwing}}$ of dimension $n = 2^{16}$ based on a distance-d coloring compared to the Monte Carlo approach for increasing number of vectors k.

d	k	error bound	approx.
1	4	-	$1.15 \cdot 10^4$
3	16	$9.3 \cdot 10^2$	$1.15 \cdot 10^4$
7	64	8.5	$1.15 \cdot 10^4$
15	256	$8.6 \cdot 10^{-4}$	$1.15 \cdot 10^4$

Table 4.9: Results for the approximation of $\mathrm{tr}(A^{-1})$ (approx.) with $A = A^{\mathrm{schwing}}$ of dimension $n = 2^{16}$ for increasing distances d and the error bound of Theorem 4.34.

error bounds for the approximations of the bilinear forms. In this example we stop the process if the computed error bound is smaller that 10^{-10} and we use the arithmetic mean of the computed upper and lower bound as our approximation. Hence, the additional error of the approximation of the trace caused by the Lanczos approximation is smaller than 10^{-10} times the number of vectors. For this example the number of iterates in the Lanczos process lies between 9 and 10. In Table 4.8 we see the results for increasing distances d, resulting in $k = (d+1)^2$ color classes if we use the coloring of Theorem 4.34. The error bounds in Table 4.8 are based on Theorem 4.34 and, e.g., for $d = 10$ we obtain the approximation $\mathrm{tr}(A^{-1}) \approx 1.04 \cdot 10^5$ with an absolute error which is smaller than $5.9 \cdot 10^{-2}$, i.e., the approximation is pretty accurate for an approximation with only 121 vectors. Using this approximation as exact trace, we can compute the error of the approximations for smaller distances d. Figure 4.18 shows the error of the approximations for increasing number of vectors. For comparison, we also depicted the error of the classical Monte Carlo approach, again using the approximation for $d = 10$ as exact value. We once again observe the classical convergence behavior of the Monte Carlo approach: We obtain a pretty good approximation for a small number of sample vectors but the error stagnates for an increasing number of samples. In contrast, we have a fast convergence of the error for the approximation based on the distance-d coloring of $G(A)$. ◇

Example 4.37. As a next example we again consider the staggered Schwinger discretization on a $2^8 \times 2^8$ lattice, which results in the matrix A^{schwing} from (4.22) of dimension $n = 2^{16}$. In order to apply the error bounds of Theorem 4.34, we need to use the coloring of Theorem 4.4 for D-dimensional lattices. Since now $G(A)$

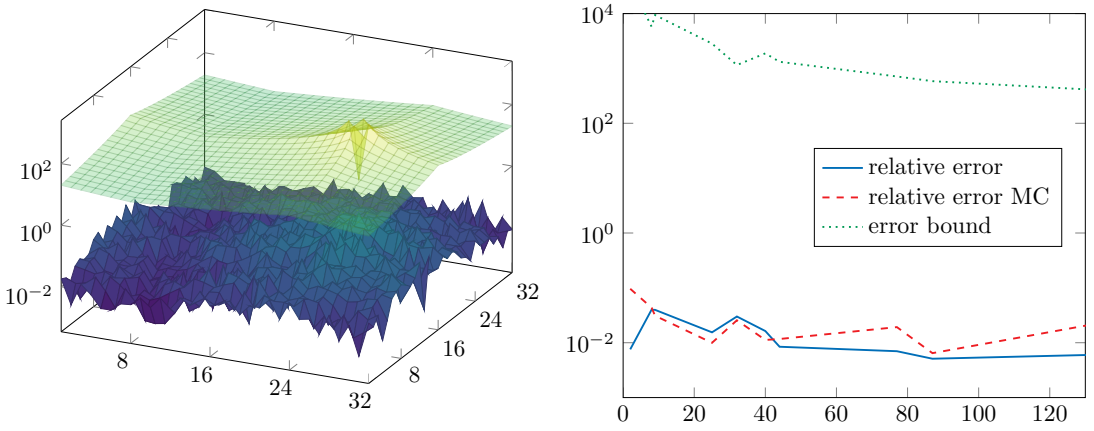

Figure 4.20: Cautionary example: Decay bounds of Theorem 3.18 for the 504th column of A^{-1} (left) and error of the approximation of $\operatorname{tr}(A^{-1})$ compared to the Monte Carlo approach (right) with $A = A^{\text{schwing}}$ of dimension $n = 32^2$.

is a *periodic* two-dimensional lattice, the coloring of Theorem 4.4 only leads to a distance-d coloring for $G(A)$ if d is of the form $d = 2^i - 1$. In Table 4.9 we have the results for distances $d = 1, 5, 7, 15$ resulting in $k = (d+1)^2$ color classes. The k bilinear forms are computed with the Arnoldi approximation where the process stops if two iterates differ by less than 10^{-10}. For this example, the number of iterations lies between 13 and 15. For $d = 15$ we need to compute 256 bilinear forms which results in the approximation $\operatorname{tr}(A^{-1}) \approx 1.15 \cdot 10^4$ with an absolute error less than $8.6 \cdot 10^{-4}$, i.e., we obtain a pretty accurate approximation of the trace. Using this approximation as exact value for $\operatorname{tr}(A^{-1})$, we see in Figure 4.19 the error of the coloring approach for $d = 1, 3, 7$ compared to the Monte Carlo approach for an increasing number of sample vectors. The approach based on a distance-d coloring immediately results in better approximations than the Monte Carlo approach. ◇

We conclude the numerical experiments with a cautionary example in order to illustrate the limits of this approach.

Example 4.38. Let $A = A^{\text{schwing}}$ be the staggered Schwinger discretization of size $n = 32^2$ with shift $s = 0.1$. The left panel of Figure 4.20 shows the exact absolute values of the 504th column of A^{-1} on the underlying grid, together with decay bounds based on Theorem 3.18. Two problems with this example are apparent: The constant of the decay bound is way to large compared to the exact entries of $f(A)$ and actually there is no decay in A^{-1} which is at least well captured by the decay bound. If we now use the approach based on a distance-d coloring, we see in the right panel of Figure 4.20 that we do not benefit from this approach compared to the Monte Carlo approach. In addition, the error bound is useless based on the

overrated bound for the entries of A^{-1}. Hence, this example illustrates that this approach is only reasonable and superior to the Monte Carlo approach, if there is a substantial decay in $f(A)$. In addition, for obtaining useful error bounds we indeed need sharp bounds for the entries of $f(A)$. ◇

Comparison to previous approaches

We will now discuss the relation of this method to some results from the literature. The proposed approach is strongly related to a method for computing an approximation of the diagonal of the inverse of Hermitian matrices from [98], where it is called *probing*. This method in [98] is based on the following proposition, where $\mathcal{D}(A)$ denotes a diagonal matrix whose diagonal entries coincide with those of A.

Proposition 4.39. *Let $A \in \mathbb{C}^{n\times n}$ and let $V_s \in \mathbb{R}^{n\times s}$ be a matrix with no zero rows. Then*

$$\mathcal{D}(A) = \mathcal{D}(AV_sV_s^T)\mathcal{D}(V_sV_s^T)^{-1} \tag{4.49}$$

holds if each i-th row of V_s is orthogonal to all those rows j of V_s for which $a_{ij} \neq 0$ holds.

Hence, the diagonal of a matrix A can be computed via the right-hand side of (4.49) if A has an appropriate sparsity pattern. This proposition is used in [98] for computing the diagonal of A^{-1} by using the sparsity pattern of a sparse approximation of A^{-1}. Motivated by the Neumann expansion, the sparsity pattern of the matrix A^p is used where a suitable p is determined by computing and analyzing one column of the inverse. The matrix V_s is computed by using the greedy graph coloring algorithm applied to the the graph $G(A^p)$. As a last step, for computing $A^{-1}V_s$, a sequence of s linear systems has to be solved, where s is the number of colors needed to color the graph $G(A^p)$. It is proposed in [98] to use Krylov subspace methods to solve the sequence of linear systems. This approach is closely related to the proposed trace estimator: As mentioned in Section 4.1, the distance-d coloring problem for $G(A)$ is equivalent to the distance-1 coloring problem of $G(A^d)$. In our approach, the distance d was chosen with respect to decay bounds for $f(A)$, i.e., it is the smallest number such that $Cq^d \leq \epsilon$ for a preassigned threshold $\epsilon > 0$. Now, if we have $d = p$, where p is the determined number of the approach in [98], the columns of the matrix $V_s \in \mathbb{R}^{n\times s}$ coincide with the vectors v_ℓ if the distance-d coloring of $G(A)$ and the distance-1 coloring of $G(A^p)$ coincide. In addition, we clearly have

$$\operatorname{tr}(A^{-1}V_sV_s^T) = \sum_{\ell=1}^{s} v_\ell^H A^{-1} v_\ell$$

if v_ℓ are the columns of $V_s \in \mathbb{R}^{n\times s}$. Hence, we conclude that the methods result in the same approximation if $d = p$ and if we use the method from [98] for computing

the trace of the inverse of Hermitian matrices. In [98] no bound for the error of the estimator of the diagonal was given, such that there is no possibility to check the quality of the approximation. Based on the fact that the approximation was motivated by the error (4.41) instead of Proposition 4.39 we were able to formulate error bounds based on decay bounds for $f(A)$. Using the approach in [98] for the computation of the trace of matrix functions was also considered in [95] for matrices A, where $G(A)$ is a regular D-dimensional lattice. The authors in [95] reveal the problem that if the distance d is changed then the vectors v_ℓ and the corresponding bilinear forms must be recomputed from scratch. Hence, they developed a method based on a hierarchical distance-d coloring of D-dimensional lattices (see Section 4.1) where the graph is successive colored until all nodes have different colors. The advantage of this approach is that one can increase the distance d until a certain accuracy of the trace estimator is reached, and in each step one can reuse previous computations. However, no criterion for the accuracy of the computed estimator was given in [95] and therefore it is not clear which distance d finally leads to a sufficient trace estimator. In contrast, since we motivated the proposed approximation by the error (4.41), we were able to formulate error bounds depending on the distance d. In addition, we have a priori error bounds, i.e., bound without actually computing a distance-d coloring, for D-dimesional lattices in Theorem 4.34. Thus, we can determine a suitable distance d in advance such that it is not necessary to compute the approximation for several distances d.

Chapter 5

Conclusions

In the introduction of this thesis, we emphasized the two major problems in the computation of matrix functions $f(A)$ for large and sparse matrices A: The high computational cost and the impossibility of storing the dense matrix $f(A)$. In the thesis we showed that the storage problem can be solved in several cases, motivated by the existence of a sparse approximation of $f(A)$ based on an exponential decay of the entries in $f(A)$. We then proposed ways for the computation of a sparse approximation of $f(A)$ with linear cost. In detail, we obtained to following results.

In Chapter 3 we provided conditions under which we can guarantee an exponential decay in $f(A)$ for special functions f and important classes of matrices A. The results gave us sharp bounds for the entries $|[f(A)]_{ij}|$ in many cases and we illustrated the superiority of the newly developed bounds over the general results from the literature in the considered cases. We also discussed the limitation of the proposed bounds: The bound of an entry of $f(A)$ only depends on the corresponding graph distance in the graph of A and we presented an example where this fact results in the overestimation of many entries of A^{-1} for a specially constructed tridiagonal matrix A. The classical decay bounds result in Toeplitz-type bounds for banded matrices while it is not necessarily the case that $f(A)$ is a Toeplitz matrix. Hence, we gave a theoretical framework for the computation of non-Toeplitz bounds in order to fix this problem. This approach requires the knowledge of the complete spectrum of several submatrices of A, which is not feasible in practice. Finding ways to efficiently compute non-Toeplitz bounds in practical situations remains an interesting topic for future research.

The results of Chapter 3 reveal the existence of a sparse approximation of $f(A)$ with a dimension-independent error in lots of important special cases. This was proven in Chapter 4 where we also discussed ways to compute such a sparse approximation with linear cost. In addition, the decay in matrix functions was

used for further important matrix computations, such as the computation of $f(A)v$ for a given vector v. We illustrated the strong relation between the decay property in $f(A)$ and the convergence of classical Krylov subspace approximations of $f(A)v$. Although the computation of $f(A)v$ is well studied for lots of functions f (especially for the inverse), we discussed some approaches where the decay in matrix functions could be further exploited for saving cost and/or storage. It would be interesting to find other approaches which consider the decay in $f(A)$ for the efficient computation of the vector $f(A)v$.

Finally, we considered the computation of the trace of matrix functions. We pointed out that the decay in $f(A)$ can be used for Monte Carlo trace estimators, as well as for approximations of the trace based on non-stochastic methods. Some of the introduced methods were already considered in a similar manner in the literature. With decay bounds, as presented in Chapter 3, we were able to supply an error analysis for the proposed approximations. Improving decay bounds for $f(A)$ for certain types of functions f and matrices A automatically improves the proposed error bounds for these approximations. On the other hand, in order to obtain a more accurate error prediction, it would be interesting to see if one can supply a more extensive analysis of the occurring errors for general matrices similar to what was done for banded matrices or matrices, where the corresponding graph is a regular lattice. This probably requires a more extensive analysis of the distance-d coloring problem. Most of the proposed methods are based on a distance-d coloring of $G(A)$, the graph of A. There are lots of theoretical results for this optimization problem but only few results are known concerning the development of low-cost methods for the computation of a distance-d coloring and an analysis of the resulting coloring. Hence, it would be of great practical interest to fill this gap for general graphs, as this thesis illustrates the importance of this special optimization problem in numerical linear algebra.

The decay of the entries of matrix functions is an interesting phenomenon from the theoretical point of view. I addition, this thesis demonstrates that this topic is practically relevant for important computational problems associated with matrix functions. We hope that this thesis will help to rise the awareness of this frequently appearing phenomenon and that the exploitation of this decay property and of decay bounds of matrix functions will develop into a widely used and helpful tool for further matrix computations and applications.

Acknowledgements

I thank all the people who supported me during the work on this thesis and I would like to thank some of them in particular.

First of all, I thank my supervisor Prof. Dr. Andreas Frommer who gave me the opportunity to write this thesis. He has always taken time to help me with occurring problems and gave me helpful advice during my work. I would also like to thank him for encouraging us to attend conferences all over the world from which I benefited both professionally and personally.

In addition I thank Marcel Schweitzer for supervising me, for his time and patience to answer my (not always meaningful) questions and for being a supportive friend. Right from the start he encouraged me to accomplish this work and supported me with his great mathematical expertise.

I thank all my current and past colleges from the Applied Computer Science Group for the pleasant years of working together. Many of them became very good friends and I will always remember the amazing conferences and vacations with them. In particular, I thank Jan Hahne for proofreading this work, for the organizational help over the last years and for his great personality and sense of humor, which have made the last few years even more enjoyable.

Finally, I thank my family and friends for their emotional support. I particularly thank my parents who always support me unconditionally.

List of Figures

List of Tables

List of Algorithms

List of Notations

Throughout this thesis, scalars and vectors are denoted by lower-case letters and matrices are denoted by upper-case letters. In addition, the following notations are used:

$a_{i,j}$	the (i,j)-th entry of the matrix A
A^H	the complex adjoint of a matrix A
$A(G)$	the adjacency matrix of a graph G
$\mathbb{C}$	the field of complex numbers
$\mathrm{d}(i,j)$	distance between nodes i and j in a graph
e_i	the i-th column of the identity matrix
$[f(A)]_{i,j}$	the (i,j)-th entry of the matrix $f(A)$
$G(A)$	the graph of a matrix A
$\mathbf{i}$	the the imaginary unit
I	the identity matrix
$\mathrm{Im}(z)$	the imaginary part of a complex number z
$\lambda_{\min}(A)$	the smallest eigenvalue of a Hermitian matrix A
$\lambda_{\max}(A)$	the largest eigenvalue of a Hermitian matrix A
p_m	a polynomial of degree m
$\mathbb{P}_m$	the set of polynomials of degree at most m
$\mathbb{R}$	the field of real numbers
$\mathbb{R}_0^-$	the negative real axis including 0
$\mathrm{Re}(z)$	the real part of a complex number z
$\sigma(A)$	the spectrum of the matrix A
v^H	the complex adjoint of the vector v
v^T	the transposed of the vector v
$W(A)$	the field of values of the matrix A
$\bar{z}$	the complex conjugate of a complex number z

Bibliography

[1] M. Aigner and G. M. Ziegler, *Proofs from THE BOOK*, Springer Publishing Company, Incorporated, 4th ed., 2009.

[2] E. Artin and M. Butler, *The Gamma Function*, Dover Books on Mathematics, Dover Publications, 2015.

[3] Z. Bai, G. Fahey, and G. Golub, *Some large-scale matrix computation problems*, Journal of Computational and Applied Mathematics, 74 (1996), pp. 71–89.

[4] G. Bali, S. Collins, and A. Schäfer, *Effective noise reduction techniques for disconnected loops in Lattice QCD*, Computer Physics Communications, 181 (2010), pp. 1570–1583.

[5] M. Beck and S. Robins, *Computing the Continuous Discretely: Integer-Point Enumeration in Polyhedra*, Undergraduate Texts in Mathematics, Springer New York, 2015.

[6] B. Beckermann and L. Reichel, *Error Estimates and Evaluation of Matrix Functions via the Faber Transform*, SIAM J. Numerical Analysis, 47 (2009), pp. 3849–3883.

[7] M. Benzi, *Preconditioning techniques for large linear systems: A survey*, J. Comput. Phys., 182 (2002), pp. 418–477.

[8] M. Benzi, *Localization in Matrix Computations: Theory and Applications*, in Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications, M. Benzi and V. Simoncini, eds., vol. 2173 of C.I.M.E. Foundation Subseries, Springer, New York, 2016, pp. 211–317.

[9] M. Benzi, P. Boito, and N. Razouk, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Review, 55 (2013), pp. 3–64.

[10] M. Benzi and G. H. Golub, *Bounds for the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417–438.

[11] M. Benzi and N. Razouk, *Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices*, Electron. Trans. Numer. Anal., 28 (2007), pp. 16–39.

[12] M. Benzi and V. Simoncini, *Decay bounds for functions of Hermitian matrices with banded or Kronecker structure*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1263–1282.

[13] C. Berg, *Stieltjes-Pick-Bernstein-Schoenberg and their connection to complete monotonicity*, 2008, pp. 15–45.

[14] J. Bloch, A. Frommer, B. Lang, and T. Wettig, *An iterative method to compute the sign function of a non-Hermitian matrix and its application to the overlap Dirac operator at nonzero chemical potential*, Computer Physics Communications, 177 (2007), pp. 933–943.

[15] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover Books on Mathematics, Dover Publications, second ed., 2001.

[16] K. Burrage, N. Hale, and D. Kay, *An Efficient Implicit FEM Scheme for Fractional-in-Space Reaction-Diffusion Equations*, SIAM Journal on Scientific Computing, 34 (2012).

[17] D. Calvetti, S.-M. Kim, and L. Reichel, *Quadrature Rules Based on the Arnoldi Process*, SIAM Journal on Matrix Analysis and Applications, 26 (2005), pp. 765–781.

[18] J. Chen and Y. Saad, *A posteriori error estimate for computing* $\mathrm{tr}(f(A))$ *by using the Lanczos method*, Numerical Linear Algebra with Applications, (2018).

[19] K. Y. Cheng, *Minimizing the bandwidth of sparse symmetric matrices*, Computing, 11 (1973), pp. 103–110.

[20] M. Crouzeix and C. Palencia, *The numerical range is a* $(1+\sqrt{2})$*-spectral set*, SIAM J. Matrix Anal., 38 (2017), pp. 649–655.

[21] E. Cuthill and J. McKee, *Reducing the Bandwidth of Sparse Symmetric Matrices*, in Proceedings of the 1969 24th National Conference, ACM '69, New York, NY, USA, 1969, ACM, pp. 157–172.

[22] S. Dalton, L. Olson, and N. Bell, *Optimizing Sparse Matrix-Matrix Multiplication for the GPU*, ACM Trans. Math. Softw., 41 (2015), pp. 25:1–25:20.

[23] S. Demko, W. F. Moss, and W. Smith, *Decay rates for inverses of banded matrices*, Math. Comput., 43 (1984), pp. 491–499.

[24] S. Dong and K. Liu, *Stochastic Estimation with Z_2 Noise*, Phys. Lett. B, 328 (1994), pp. 130–136.

[25] V. L. Druskin and L. A. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, Comput.Maths.math.Phys, 29 (1989), pp. 112–121.

[26] V. Eijkhout and B. Polman, *Decay rates of inverses of banded M-matrices that are near to Toeplitz matrices*, Linear Algebra Appl., 109 (1988), pp. 247–277.

[27] S. W. Ellacott, *Computation of Faber series with application to numerical polynomial approximation in the complex plane*, Math. Comp., 40 (1983), pp. 575–587.

[28] K. Erciyes, *Guide to Graph Algorithms: Sequential, Parallel and Distributed*, Texts in Computer Science, Springer International Publishing, 2018.

[29] E. Estrada, *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, Inc., New York, NY, USA, 2011.

[30] E. Estrada and D. Higham, *Network properties revealed through matrix functions*, SIAM Review, 52 (2010), pp. 696–714.

[31] E. Estrada and J. A. Rodríguez-Velázquez, *Subgraph centrality in complex networks*, Phys. Rev. E, 71 (2005), p. 056103.

[32] S. Even and G. Even, *Graph Algorithms*, Cambridge University Press, 2011.

[33] G. Fertin, E. Godard, and A. Raspaud, *Acyclic and k-distance coloring of the grid*, Information Processing Letters, 87 (2003), pp. 51 – 58.

[34] N. J. Ford, D. V. Savostyanov, and N. L. Zamarashkin, *On the decay of the elements of inverse triangular Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1288–1302.

[35] R. Freund, *On polynomial approximations to $f_a(z) = (z-a)^{-1}$ with complex a and some applications to certain non-Hermitian matrices*, Approximation Theory Appl., 5 (1989), pp. 15–31.

[36] A. Frommer, S. Güttel, and M. Schweitzer, *Efficient and Stable Arnoldi Restarts for Matrix Functions Based on Quadrature*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 661–683.

[37] A. Frommer, K. Kahl, T. Lippert, and H. Rittich, *2-norm Error Bounds and Estimates for Lanczos Approximations to Linear Systems and Rational Matrix Functions*, SIAM J. Matrix Analysis Applications, 34 (2013), pp. 1046–1065.

[38] A. Frommer, C. Schimmel, and M. Schweitzer, *Bounds for the decay of the entries in inverses and Cauchy–Stieltjes functions of certain sparse, normal matrices*, Numerical Linear Algebra with Applications, 25 (2018), p. e2131.

[39] A. Frommer, C. Schimmel, and M. Schweitzer, *Non-Toeplitz decay bounds for inverses of Hermitian positive definite tridiagonal matrices*, Electron. Trans. Numer. Anal., 48 (2018), pp. 362–372.

[40] A. Frommer and V. Simoncini, *Matrix Functions*, in Model Order Reduction: Theory, Research Aspects and Applications, W. H. A. Schilders, H. A. van der Vorst, and J. Rommes, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 275–303.

[41] A. Gambhir, A. Stathopoulos, K. Orginos, B. Yoon, R. Gupta, and S. Syritsyn, *Algorithms for Disconnected Diagrams in Lattice QCD*, (2016).

[42] N. E. Gibbs, W. G. Poole, and P. K. Stockmeyer, *An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix*, SIAM Journal on Numerical Analysis, 13 (1976), pp. 236–250.

[43] A. Gil, J. Segura, and N. Temme, *Numerical Methods for Special Functions*, Society for Industrial and Applied Mathematics, 2007.

[44] M. Giles, *Multilevel Monte Carlo methods*, Acta Numerica, 24 (2015), pp. 259–328.

[45] Y. Ginosar, I. Gutman, T. Mansour, and M. Schork, *Estrada index and Chebyshev polynomials*, 454 (2008), pp. 145–.

[46] G. H. Golub, M. Heath, and G. Wahba, *Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter*, Technometrics, 21 (1979), pp. 215–223.

[47] G. H. Golub and G. Meurant, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton, NJ, USA, 2009.

[48] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 2013.

[49] G. H. Golub and U. von Matt, *Generalized Cross-Validation for Large Scale Problems*, J. Comput. Graph. Stat, 6 (1995), pp. 1–34.

[50] S. L. Gonzaga de Oliveira, J. A. B. Bernardes, and G. O. Chagas, *An evaluation of low-cost heuristics for matrix bandwidth and profile reductions*, Computational and Applied Mathematics, 37 (2018), pp. 1412–1471.

[51] M. Gu and S. C. Eisenstat, *A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 172–191.

[52] I. Gutman and A.Graovac, *Estrada index of cycles and paths*, 436 (2007), pp. 294–296.

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

[54] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[55] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1994.

[56] M. Hutchinson, *A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines*, J. Commun. Statist. Simula., (1990), pp. 433–450.

[57] A. Iserles, *How large is the exponential of a banded matrix*, N. Z. J. Math., 29 (2000), pp. 177–192.

[58] D. Kershaw, *Inequalities on the elements of the inverse of a certain tridiagonal matrix*, Math. Comput., (1970), pp. 155–158.

[59] R. Kippenhahn, *On the numerical range of a matrix*, Linear and Multilinear Algebra, 56 (2008), pp. 185–225. Translated by P. F. Zachlin and M. E. Hochstenbach.

[60] K. Koh, F. Dong, and E. Tay, *Introduction to Graph Theory: H3 Mathematics*, World Scientific, 2007.

[61] F. Kramer and H. Kramer, *On the generalised chromatic number*, Ann. Discrete Math., 30 (1986), pp. 275 – 284.

[62] D. Kressner and A. Šušnjara, *Fast computation of spectral projectors of banded matrices*, SIAM J. Matrix Anal. Appl., 38 (2007), pp. 984–1009.

[63] M. Kubale, *Graph Colorings*, Contemporary mathematics (American Mathematical Society) v. 352, American Mathematical Society, 2004.

[64] J. Liesen, *Construction and analysis of polynomial iterative methods for non-Hermitian systems of linear equations*, PhD thesis, University of Bielefeld, 1998.

[65] J. Liesen and Z. Strakos, *Krylov Subspace Methods: Principles and Analysis*, Oxford University Press, Oxford, 2013.

[66] A. Lim, B. Rodrigues, and F. Xiao, *A fast algorithm for bandwidth minimization*, International Journal on Artificial Intelligence Tools, 16 (2007), pp. 537–544.

[67] G. Meinardus, *Approximation von Funktionen und ihre numerische Behandlung*, Springer Tracts in Natural Philosophy, New York, 1964.

[68] L. Y. Miao and Y. Z. Fan, *The distance coloring of graphs*, Acta Mathematica Sinica, English Series, 30 (2014), pp. 1579–1587.

[69] M. Molloy and M. R. Salavatipour, *A bound on the chromatic number of the square of a planar graph*, Journal of Combinatorial Theory, Series B, 94 (2005), pp. 189 – 213.

[70] H. Mulholland and C. Jones, *Fundamentals of statistics*, Plenum Press, 1968.

[71] R. Nabben, *Decay rates of the inverse of nonsymmetric tridiagonal and band matrices*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 820–837.

[72] ——, *Two-sided bounds on the inverses of diagonally dominant tridiagonal matrices*, Linear Algebra Appl., 287 (1999), pp. 289–305.

[73] Y. Nagasaka, S. Matsuoka, A. Azad, and A. Buluç, *High-Performance Sparse Matrix-Matrix Products on Intel KNL and Multicore Architectures*, in Proceedings of the 47th International Conference on Parallel Processing Companion, ICPP '18, New York, NY, USA, 2018, ACM, pp. 34:1–34:10.

[74] H. Neuberger, *The Overlap Dirac Operator*, in Numerical Challenges in Lattice Quantum Chromodynamics. Lecture Notes in Computational Science and Engineering, M. B. S. K. Frommer A., Lippert T., ed., vol. 15, Springer, Berlin, Heidelberg, 2000.

[75] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall Series in Computational Mathematics, Pearson Education Canada, 1980.

[76] R. Peluso and T. Politi, *Some improvements for two-sided bounds on the inverse of diagonally dominant tridiagonal matrices*, Linear Algebra Appl., 330 (2001), pp. 1–14.

[77] L. Polok, V. Ila, and P. Smrz, *Fast Sparse Matrix Multiplication on GPU*, in Proceedings of the Symposium on High Performance Computing, HPC '15, San Diego, CA, USA, 2015, Society for Computer Simulation International, pp. 33–40.

[78] S. Pozza and V. Simoncini, *Inexact Arnoldi residual estimates and decay properties for functions of non-Hermitian matrices*, BIT Numerical Analysis, 59 (2019), pp. 969–986.

[79] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

[80] J. Reid and J. Scott, *Reducing the Total Bandwidth of a Sparse Unsymmetric Matrix*, SIAM Journal on Matrix Analysis and Applications, 28 (2006), pp. 805–821.

[81] T. Rivlin, *The Chebyshev Polynomials*, A Wiley-Interscience publication, Wiley, 1974.

[82] J. D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, International Journal of Control, 32 (1980), pp. 677–687.

[83] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*, Chapman & Hall/CRC, 2005.

[84] Y. Saad, *Analysis of Some Krylov Subspace Approximations to the Matrix Exponential Operator*, SIAM Journal on Numerical Analysis, 29 (1992), pp. 209–228.

[85] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2nd edition ed., 2003.

[86] B. Sapoval, T. Gobron, and A. Margolina, *Vibrations of fractal drums*, Phys. Rev. Lett., 67 (1991), pp. 2974–2977.

[87] H. R. Schwarz, *Handbook series linear algebra: Tridiagonalization of a symmetric band matrix*, Numer. Math., 12 (1968), pp. 231–241.

[88] M. Schweitzer, *Restarting and error estimation in polynomial and extended Krylov subspace methods for the approximation of matrix functions*, PhD thesis, Bergische Universität Wuppertal, 2015.

[89] O. Sète and J. Liesen, *Properties and Examples of Faber–Walsh Polynomials*, Computational Methods and Function Theory, 17 (2017), pp. 151–177.

[90] J. Sexton and D. Weingarten, *Systematic Expansion for Full QCD Based on the Valence Approximation*, tech. rep., Watson Research Center, 1994.

[91] M. Shao, *On the finite section method for computing exponentials of doubly-infinite skew-Hermitian matrices*, Linear Algebra and Its Applications, 451 (2014), pp. 65–96.

[92] A. Sharp, *Distance coloring*, in Algorithms – ESA 2007, L. Arge, M. Hoffmann, and E. Welzl, eds., Berlin, Heidelberg, 2007, Springer Berlin Heidelberg, pp. 510–521.

[93] J. Sherman and W. J. Morrison, *Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix*, Ann. Math. Stat., 21 (1950), pp. 124–127.

[94] W. Smyth, *Algorithms for the reduction of matrix bandwidth and profile*, Journal of Computational and Applied Mathematics, 12-13 (1985), pp. 551 – 561.

[95] A. Stathopoulos, J. Laeuchli, and K. Orginos, *Hierarchical Probing for Estimating the Trace of the Matrix Inverse on Toroidal Lattices*, SIAM J. Scientific Computing, 35 (2013).

[96] P. Suetin and E. Pankratiev, *Series of Faber Polynomials*, Analytical Methods and Special Functions, Taylor & Francis, 1998.

[97] H. Tal-Ezer, *Polynomial approximation of functions of matrices and applications*, Journal of Scientific Computing, 4 (1989), pp. 25–60.

[98] J. M. Tang and Y. Saad, *A probing method for computing the diagonal of a matrix inverse*, Numerical Linear Algebra with Applications, 19 (2012), pp. 485–501.

[99] L. N. Trefethen, *Approximation Theory and Approximation Practice*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2012.

[100] R. Trudeau, *Introduction to Graph Theory*, Dover Books on Mathematics, Dover Publications, 2013.

[101] S. Ubaru, J. Chen, and Y. Saad, *Fast estimation of* $\mathrm{tr}(f(A))$ *via stochastic lanczos quadrature*, SIAM J. Matrix Analysis Applications, 38 (2017), pp. 1075–1099.

[102] S. Ubaru and Y. Saad, *Applications of trace estimation techniques*, in HPCSE, 2017.

[103] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. Van der Vorst, *Numerical Methods for the QCD Overlap Operator: I. Sign-function and Error Bounds*, Computer Physics Communications, 146 (2002), pp. 203–224.

[104] J. van den Heuvel and S. McGuinness, *Coloring the square of a planar graph*, Journal of Graph Theory, 42 (2003), pp. 110–124.

[105] D. West, *Introduction to Graph Theory*, Math Classics, Pearson, 2017.

[106] R. Wilson, *Four Colors Suffice: How the Map Problem was Solved*, Princeton Paperbacks, Princeton University Press, 2002.