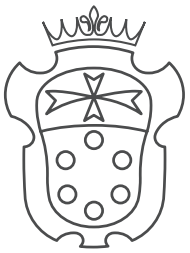

SCUOLA
NORMALE
SUPERIORE

Classe di Scienze

Corso di perfezionamento in
Metodi Computazionali e Modelli Matematici per le Scienze
e la Finanza

35º ciclo

Approximation of Matrix Functions Arising in Physics and Network Science: Theoretical and Computational Aspects

Settore Scientifico Disciplinare **MAT/08**

Candidato
dr. Michele Rinelli

Relatore
Prof. Michele Benzi

Supervisore
Prof.ssa Paola Boito

Anno accademico 2023–2024

Contents

List of Figures

List of Tables

Abstract

Many applications in physics and network science require the computation of quantities related to certain matrix functions. In many cases, a straightforward way to proceed is by diagonalization. However, the cost scales cubically with the size, hence this method becomes prohibitive for large dimensions. The aim of this thesis is to provide techniques with much lower computational cost that exploit the (approximate) sparse structure of the matrices involved.

Spectral projectors associated with Hermitian matrices play a key role in applications such as electronic structure computations in quantum chemistry and physics. Linear scaling methods in the gapped case, that corresponds to nonmetallic systems, are based on the fact that such projectors are localized, which means that the entries decay rapidly away from the main diagonal or with respect to more general sparsity patterns. The relation with the sign function, together with an integral representation of the latter, is used to obtain new decay bounds, which turn out to be optimal in an asymptotic sense. The influence of isolated extremal eigenvalues on the decay properties is also investigated and a superexponential behavior is predicted. Using similar techniques, we extend our results to related matrix functions, such as the Fermi-Dirac function and Cauchy-Stieltjes functions.

Another problem of interest is the computation of the trace of a matrix function $f(A)$. In certain situations, in particular if $f(A)$ cannot be well approximated by a low-rank matrix, combining probing methods based on graph colorings with stochastic trace estimation techniques can yield accurate approximations at moderate cost. Until recently, such methods had not been thoroughly analyzed, however, but were rather used as efficient heuristics by practitioners. We perform a detailed analysis of stochastic probing methods and, in particular, expose conditions under which the expected approximation error in the stochastic probing method scales more favorably with the dimension of the matrix than the error in non-stochastic probing.

A quantity that often appears in quantum physics and network science is the von Neumann entropy $\mathrm{Tr}(f(A))$, where $f(x) = -x \log x$ and A is a density matrix, i.e., a symmetric positive semidefinite matrix with unit trace. As an alternative to diagonalization, probing techniques or stochastic trace estimators can be used to obtain approximations of the entropy. Both methods are based on the computation of several quadratic forms $\boldsymbol{b}^T f(A)\boldsymbol{b}$ and matrix-vector products $f(A)\boldsymbol{b}$, which can in turn be approximated efficiently using polynomial and rational Krylov subspace methods. With this approach, the matrix A is only accessed via matrix-vector products $A\boldsymbol{v}$ and via the solution of shifted linear systems $(A + \xi I)^{-1}\boldsymbol{v}$ with

$\xi > 0$. For the probing approach, theoretical bounds and heuristic estimates are provided for the error on the entropy, which can be used to select the number of quadratic forms required to reach a certain accuracy. Moreover, a posteriori error bounds are given for the Krylov approximations. Our results are validated by several numerical experiments on a number of test problems arising in network analysis.

Chapter 1

Introduction

Matrix functions have been extensively researched in numerical linear algebra and find versatile applications across numerous disciplines in the applied sciences. They are written in the form $f(A)$, where A is a square matrix and $f(x)$ is a scalar function defined over the spectrum of A. Prominent examples include the matrix inverse A^{-1}, widely encountered in numerical linear algebra, e.g., in the solution of linear systems, and the matrix exponential e^A, renowned for its role in solving differential equations.

Another significant case is the spectral projector of a Hermitian matrix H, which acts as the orthogonal projector onto the subspace spanned by the eigenvectors associated with eigenvalues below a certain value $\mu \in \mathbb{R}$ [14]. Here H denotes the Hamiltonian operator of the system. This projector, also referred to as the density matrix in the chemistry and physics literature [97, Chapter 4], is of central importance in electronic structure computations at zero temperature [24, 70]. For positive temperatures, the electron density matrix can be expressed by the Fermi-Dirac function $(I + e^{\beta(H-\mu I)})^{-1}$, where $\beta \in \mathbb{R}$ is inversely proportional to T. In quantum statistical mechanics, the density matrix is the Gibbs state, which takes the form $\exp(-\beta H)/Z$, where $Z = \mathrm{Tr}(\exp(-\beta H))$.

Chapter 3 of this thesis is devoted to a study of decay properties of certain matrix functions of interest in physics and chemistry. The a priori knowledge of decay bounds for functions of localized matrices allows to approximate them with banded or sparse matrices and develop linear scaling methods. This is crucial, for example, in electronic structure computations [14, 24, 70, 83]. An exponential decay holds in general for $f(A)$, where A is Hermitian and banded (or sparse) and f is analytic over an ellipse containing the spectrum of A [15]. Specific bounds are known for important matrix functions, like the matrix inverse [9, 35] or entire functions, like the matrix exponential, which exhibit superexponential decay [19, 69]. Further results for classes of functions defined by an integral transform, such as Laplace-Stieltjes and Cauchy-Stieltjes functions, are given in [19, 45], where the analysis makes use of results for the inverse or the exponential.

In [14] one can find rigorous proofs of exponential decay of the entries of spectral projectors for gapped systems, like insulators, and for the Fermi-Dirac function associated with general systems. For the latter, the results of [15] based on Bernstein's Theorem are used, while, for the projector, another approach is inspired by [65] and makes use of the polynomial

approximation of a piecewise constant function over the union of disjoint intervals. We enhance this approach by exploiting an integral representation of the sign function and obtain refined decay bounds [17]. We apply the same idea to the Fermi-Dirac function.

Most of the existing bounds for $f(A)$ depend only on partial information about the spectrum of A, for example, the spectral interval $[\lambda_{\min}(A), \lambda_{\max}(A)]$ if A is Hermitian and positive definite [19, 35] or the field of values in the general case [12, 87]. For the spectral projector, a key role is played by the spectral gap. However, numerical experiments show that the bounds are often pessimistic and do not capture the actual decay behavior, which seems to depend also on the distribution of the eigenvalues within the spectral sets. A first step in this direction is taken in [45], where the authors show a connection between the decay in the inverse of a positive definite Hermitian matrix and the distribution of the eigenvalues near the upper end of the spectrum. We refine and generalize these results for the inverse and other matrix functions defined by integral transforms.

Another important task in numerical linear algebra is to estimate the trace of an implicitly given matrix $B \in \mathbb{R}^{n\times n}$,

$$\operatorname{Tr}(B) = \sum_{i=1}^{n} [B]_{ii}. \tag{1.1}$$

This has applications in, e.g., machine learning and data science [63, 89], theoretical particle physics [96, 101] and analysis of complex networks [13, 40]; see [102] for an extensive survey. In many of these applications, we have $B = f(A)$, where $A \in \mathbb{R}^{n\times n}$ is a large and sparse or structured matrix. In this case, B cannot be formed explicitly and often the only feasible operations with A are matrix-vector products (and sometimes linear system solves). This then translates into methods for approximating (1.1) that must also rely on matrix-vector products or quadratic forms with B, which are, e.g., performed by applying a polyomial (or rational) Krylov subspace method or Chebyshev expansion for approximating $f(A)\boldsymbol{v}$ or $\boldsymbol{v}^T f(A)\boldsymbol{v}$ for a given vector $\boldsymbol{v}$; hence the name *implicit* or *matrix-free* trace estimation.

Popular methods for this task come in two flavors: on the one hand, there are stochastic estimators, with the most important ones being the classical Hutchinson estimator as established workhorse algorithm [68] and the recently proposed Hutch++ algorithm [81] and its refined variants; see [26, 37, 86]. These estimators are black-box methods for (1.1) which do not exploit any inherent structure in B and mainly rely on applying B to randomly sampled vectors. On the other hand, when $B = f(A)$ with sparse A, another popular class of methods includes estimators based on probing [47, 98, 99], a technique which aims to reduce the number of quadratic forms that are needed for a given accuracy by exploiting structure in A to carefully craft specific probing vectors instead of just randomly sampling them.

A goal of this thesis is to analyze the combination of probing techniques with stochastic estimators, which is obtained filling the nonzero entries of the probing vectors with random samples from an appropriate distribution, leading to the stochastic probing estimation method [44]. The combination of these two approaches is algorithmically quite straightforward and was used before by practitioners; see, e.g., [5, 7, 82, 105]. However, an in-depth analysis is lacking so far. We provide such an analysis in Chapter 4. Our analysis explains and highlights many of the important properties of the approach and reveals in particular for which matrix functions f and matrices A large gains can not only be expected, but actually

be guaranteed. Our theoretical findings are illustrated and confirmed by a variety of numerical experiments. As a by-product of our analysis, we also refine classical results on sign patterns in the entries of $f(A)$.

An important example is the von Neumann entropy [103] of a symmetric positive semidefinite A with $\mathrm{Tr}(A) = 1$. It is defined as $S(A) = \mathrm{Tr}(-A \log A)$, where $\log A$ is the matrix logarithm. With the usual convention that $0 \cdot \log 0 = 0$, the von Neumann entropy is given by

$$S(A) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i \,,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the $n \times n$ matrix A. The von Neumann entropy plays an important role in several fields including quantum statistical mechanics [74], quantum information theory [10], and network science [25]. For example, computing the von Neumann entropy is used in order to determine the ground state of many-electron systems at finite temperature [1]. The von Neumann entropy of graphs is also an important tool in the structural characterization and comparison of complex networks [33, 64].

If the size of A is large, the computation of $S(A)$ by means of explicit diagonalization can be too expensive, so it becomes necessary to resort to cheaper methods that compute approximations of the entropy. In recent years, a few papers have appeared devoted to this problem; see, e.g., [27, 36, 77, 107]. These papers investigate different approaches based on quadratic (Taylor) approximants, the global Lanczos algorithm, Gaussian quadrature and Chebyshev expansion. In general, the problem of computing the von Neumann entropy of a large matrix is difficult because the underlying matrix function is not analytic in a neighbourhood of the spectrum when the matrix is singular; difficulties can also be expected when A has eigenvalues close to zero, which is usually the case. In particular, polynomial approximation methods may converge slowly. In these cases, the use of rational Krylov methods is often recommended.

In Chapter 5 we propose to approximate the von Neumann entropy using either the probing approach developed in [47] or a stochastic trace estimator [68, 81, 86]. First, we obtain an integral expression for the entropy function $f(x) = -x \log x$ that relates it with a Cauchy-Stieltjes function [106, Chapter VIII], and we use it to derive error bounds for the polynomial approximation of f, which in turn lead to a priori bounds for the approximation of $S(A)$ with deterministic probing methods. In order to also have a practical stopping criterion alongside the theoretical bounds, we propose some heuristics to estimate the error of probing methods and demonstrate their reliability with numerical experiments. We also use properties of symmetric M-matrices to show that, in the case of the graph entropy, the approximation obtained with a probing method is always a lower bound for the exact entropy.

Both probing methods and stochastic trace estimators require the computation of a large number of quadratic forms with $f(A)$, which can be efficiently approximated using Krylov subspace methods [48, 58]. We propose to combine polynomial Krylov iterations with rational Krylov iterations that use asymptotically optimal poles for Cauchy-Stieltjes functions [79]. Next, we obtain new a posteriori error bounds and estimates for this task, building on the ones presented in [57], and we discuss methods to compute them efficiently. For the case

of the graph entropy, we make use of a desingularization technique introduced in [20] that exploits properties of the graph Laplacian to compute quadratic forms more efficiently.

The resulting algorithm can be seen as a black box method that only requires an input tolerance ϵ and computes an approximation of $S(A)$ with relative accuracy ϵ. While this accuracy is not guaranteed when the rigorous bounds are replaced by heuristics, we found the algorithm to be quite reliable in practice.

Our implementation of the probing algorithm is compared to a state-of-the-art randomized trace estimator developed in [86] with several numerical experiments, in which we approximate the graph entropy of various complex networks. The performance of stochastic probing is discussed as well.

To summarize, the thesis is organized as follows: In Chapter 2 we review basic definitions and results for matrix functions and graph theory. In Chapter 3 we study the decay properties of matrix functions. In particular, we prove new decay bounds for spectral projectors and the Fermi-Dirac function and show the connection between the eigenvalue distribution and the decay properties of such matrix function. In Chapter 4, we consider the estimation of the trace of matrix functions. We recall the two basic trace estimation methods, namely probing and the Hutchinson estimator. Then we analyze the stochastic probing estimator and derive formulas for the variance as well as more sophisticated tail bounds. We show that in certain situations, the expected value of the error increases only with the square root of the matrix dimension n, whereas in deterministic probing the error scales linearly with n. We present a number of numerical experiments that illustrate the theoretical results and the performance of the stochastic probing estimator compared to other state-of-the-art methods. In Chapter 5, we recall some properties of the von Neumann entropy and we obtain bounds for the polynomial approximation of $f(x) = -x \log x$, which are applied to derive bounds for the convergence of probing methods. Then we describe the computation of quadratic forms using Krylov subpsace methods and we get a posteriori error bounds and estimates. We summarize the overall algorithm and discuss heuristics and stopping criteria. We test the performance of the methods on density matrices of several complex networks. Finally, we conclude by summarizing the presented work and exploring possible future research directions.

Chapter 2

Basics on Matrix Functions and Graphs

In this chapter we review the basic definitions and properties of matrix functions and graphs.

2.1 Background on Matrix Functions

In this section we define a matrix function $f(A)$, where A is a square matrix and $f(x)$ is a scalar function. Hence, we discuss some of the main properties we are interested in, such as the connections with polynomials and basic ways to approximate them. For more details on the theory of matrix functions, see the book [67].

2.1.1 Definition

There are various equivalent ways to define the matrix function $f(A)$. We first give the definition based on the Jordan canonical form [67, Definition 1.2].

Recall that every matrix A can be written in Jordan canonical form:

$$A = Z^{-1}JZ, \quad J = \operatorname{diag}(J_1, \dots, J_p), \quad J_k = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix}, \tag{2.1}$$

where $J_k \in \mathbb{C}^{m_k \times m_k}$ is a *Jordan block* associated with the *eigenvalue* λ_k of A. Note that the representation is unique up to the order of the blocks and $m_1 + \cdots + m_p = n$.

Denote with $\sigma(A)$ the *spectrum* of A, i.e., the set of all the eigenvalues. Note that, in general, there can be more than one Jordan block associated with each eigenvalue, hence $\sigma(A)$ contains at most p distinct numbers. The *index* $n(\lambda)$ of $\lambda \in \sigma(A)$ is the size of the largest Jordan block associated with λ.

In order to define $f(A)$ we need f to be *defined on the spectrum of* A, i.e. the values

$$f^{(j)}(\lambda), \quad j = 1, \dots, n(\lambda) - 1, \quad \lambda \in \sigma(A)$$

must exist. With these notations, we are able to give the following definition.

Definition 2.1. Let f be defined over the spectrum of $A \in \mathbb{C}^{n\times n}$, and let A have the Jordan canonical form (2.1). Then we define

$$f(A) := Z^{-1} f(J) Z, \quad f(J) = \operatorname{diag}(f(J_1), \dots, f(J_p)),$$

with

$$f(J_k) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \cdots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix} \in \mathbb{C}^{m_k\times m_k}, \quad k = 1, \dots, p. \tag{2.2}$$

A direct consequence of (2.2) is that $\sigma(f(A)) = \{f(\lambda) : \lambda \in \sigma(A)\}$, since $f(J)$ is triangular and only the values $f(\lambda_k)$, $k = 1, \dots, p$, appear on the diagonal. Moreover, this definition simplifies if A is diagonalizable. In this case each Jordan block J_k in (2.1) has size $m_k = 1$, hence $s = n$ and $J = \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Therefore f is defined on the spectrum of A if $f(\lambda_k)$ is defined for $k = 1, \dots, n$, and

$$f(A) = Z^{-1} f(\Lambda) Z, \quad f(\Lambda) = \operatorname{diag}(f(\lambda_1), \dots, f(\lambda_n)).$$

In particular, we do not require f to be differentiable or even continuous at any point. In the following chapters we will only deal with Hermitian matrices. Hence, A will always be diagonalizable and $Z^{-1} = Z^*$, so $f(A) = Z^* f(\Lambda) Z$ is Hermitian as well.

2.1.2 Main Properties

Matrix functions are strictly connected to polynomials, as explained by the following proposition.

Proposition 2.1. *Let $A \in \mathbb{C}^{n\times n}$ be diagonalizable with eigenvalues $\lambda_1, \dots, \lambda_n$ (not necessarily distinct). Let $f(x)$ be defined on the spectrum of A and let $p(x) = \sum_{j=0}^{d} c_j x^j$ be a polynomial. Then*

- *$p(x)$ is defined over the spectrum of A and $p(A) = \sum_{j=0}^{d} c_k A^k$.*
- *If $f(\lambda_k) = p(\lambda_k)$ for all $k = 1, \dots, n$, then $f(A) = p(A)$.*

A polynomial that interpolates the values of f on $\lambda_1, \dots, \lambda_n$ always exists and an example is the Lagrange interpolating polynomial, which has degree $n - 1$ (or less). In particular, every matrix function is a polynomial evaluated at the matrix argument. This argument can be generalized to non-diagonalizable matrices by using the Hermite interpolation, and, if used as the definition of a matrix function, it is equivalent to Definition 2.1; see [67, Chapter 1].

Many important matrix functions, such as the matrix exponential, are commonly defined via their power series. The following result shows the equivalence of these two definitions.

Theorem 2.2 (Theorem 4.7 in [67]). *Let $f(x)$ have the Taylor series*

$$f(x) = \sum_{j=0}^{\infty} c_j (x-\alpha)^j$$

that converges absolutely over the complex disk $\mathbb{D} = \{z \in \mathbb{C} : |z-r| < r\}$. If $A \in \mathbb{C}^{n\times n}$ is such that $\sigma(A) \subset \mathbb{D}$, then

$$f(A) = \sum_{j=0}^{\infty} c_j (A-\alpha I)^j.$$

Theorem 2.2 includes classical examples in the matrix functions setting. For instance,

$$\begin{aligned} e^A &= \sum_{j=0}^{\infty} \frac{1}{j!} A^j, \\ (I-A)^{-1} &= \sum_{j=0}^{\infty} A^j, \\ \log(I+A) &= \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} A^j, \end{aligned}$$

where the last two identities hold if $|\lambda| < 1$ for all $\lambda \in \sigma(A)$.

Another classical way to define matrix functions is via the Cauchy integral formula.

Theorem 2.3. *Let $A \in \mathbb{C}^{n\times n}$, and let f be analytic on and inside a closed contour Γ that encloses $\sigma(A)$, oriented clockwise. Then*

$$f(A) = \frac{1}{2\pi \boldsymbol{i}} \int_{\Gamma} f(z)(zI-A)^{-1}\,\mathrm{d}z. \tag{2.3}$$

The representation (2.3) can be used for many different purposes, especially in the context of computing $f(A)\boldsymbol{b}$, where $\boldsymbol{b} \in \mathbb{C}^n$; see, e.g., [60]. We will see an application in Section 5.3.3.

Approximation Theory

Here we discuss basic ways to approximate matrix functions by means of polynomials or rational functions.

Let $A \in \mathbb{C}^{n\times n}$ be Hermitian with spectrum $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$. Recall that the most widely used matrix norms (p-norm for $p = 1, 2, \infty$ and Frobenius norm) are given by

$$\|A\|_1 = \|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |[A]_{ij}|, \quad \|A\|_2 = \max_{k=1,\ldots,n} |\lambda_k|, \quad \|A\|_F = \sqrt{\sum_{k=1}^{n} \lambda_k^2}. \tag{2.4}$$

The following result is useful to analyze the approximations of a matrix function.

Proposition 2.4. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian with spectral interval $[\lambda_{\min}, \lambda_{\max}]$, and let f be defined over $[\lambda_{\min}, \lambda_{\max}]$. Then*

$$\|f(A)\|_2 \leq \max_{x\in[\lambda_{\min},\lambda_{\max}]} |f(x)|.$$

Proposition 2.4 is a consequence of the spectral theorem for Hermitian matrices and of the unitary invariance of the 2-norm. This inequality can also be generalized to functions of non-Hermitian matrices, due to an important result by Crouzeix [30, 31]. In this case, the set on which f is maximized is the numerical range (or field of values) of A, and an additional constant factor $K = 1 + \sqrt{2}$ must be added in front of the right-hand side, although it is conjectured that the result is true with $K = 2$.

In view of Proposition 2.4, for any function $g(x)$ defined over $[\lambda_{\min}, \lambda_{\max}]$ we get

$$\|f(A) - g(A)\|_2 \leq \max_{x\in[\lambda_{\min},\lambda_{\max}]} |f(x) - g(x)|.$$

Hence, in order to approximate $f(A)$, we can choose a function g that approximates f over $[\lambda_{\min}, \lambda_{\max}]$ and is easy to compute (e.g., a polynomial or rational function).

2.2 Background on Graph Theory

Here we recall some notions and notations for graphs. For a detailed survey see, e.g., [40].

2.2.1 Definitions

A *graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of *nodes* and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of *edges*. Here we deal with finite graphs, so $\mathcal{V}$ is a finite set. A graph is called *undirected* if $(v, w) \in \mathcal{E}$ if and only if $(w, v) \in \mathcal{E}$ for all edges $(v, w) \in \mathcal{E}$, otherwise $\mathcal{G}$ is *directed*. Given a numbering $\mathcal{V} = \{v_1, \ldots, v_n\}$ of the nodes, the *adjacency matrix* A has entries

$$[A]_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E}, \\ 0 & \text{if } (v_i, v_j) \notin \mathcal{E}. \end{cases}$$

Note that A is symmetric if and only if the graph is undirected.

For an undirected graph, the *degree* of a node is the number of edges incident to v, that is,

$$\deg(v) = |\{w \in \mathcal{V} \mid (v, w) \in \mathcal{E}\}|.$$

The degree can also be expressed in terms of the adjacency matrix: $\deg(v_i) = [A\mathbf{1}]_i$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones. Through the degree we can define the *graph Laplacian* as

$$L = D - A, \quad D = \operatorname{diag}(\deg(v_1), \ldots, \deg(v_n)).$$

A *walk* of length ℓ is a sequence $w_0, w_1, \dots, w_\ell \in \mathcal{V}$ such that $(w_{k-1}, w_k) \in \mathcal{E}$ for $k = 1, \dots, \ell$. The relation between walks and the entries of the powers of the adjacency matrix is explained by the following formula:

$$[A^\ell]_{ij} = |\{\text{walks of length } \ell \text{ starting at } v_i \text{ and ending at } v_j\}|. \tag{2.5}$$

We say that two nodes $v, w \in \mathcal{V}$ are *connected* if there is a walk that *connects* v to w, i.e. it starts with v and ends with w. We say that an undirected graph is *connected* if, for each pair of vertices $v, w \in \mathcal{V}$, v and w are connected.

The *geodesic distance* $\mathrm{d}(v, w)$ between two nodes $v, w \in \mathcal{V}$ is the minimal length of a walk that connects v and w. We let $\mathrm{d}(v, w) = \infty$ if no such walk exists. If $v = v_i$ and $w = v_j$, in view of (2.5), we get the formula

$$\mathrm{d}(v_i, v_j) = \min\{\ell \in \mathbb{N} \mid [A^\ell]_{ij} \neq 0\}.$$

The *diameter* of a graph is the maximum distance between any two nodes $v, w \in \mathcal{V}$. Finally, the *average path length* of a graph is

$$\frac{1}{n(n-1)} \sum_{v \in \mathcal{V}} \sum_{w \in \mathcal{V}} \mathrm{d}(v, w).$$

A small average path length (with respect to the number of nodes) is typical of *small-world graphs*; see [3].

Chapter 3

Decay of Spectral Projectors and related Matrix Functions

In this chapter, we explore new decay bounds on the entries of matrix functions related to spectral projectors. In Section 3.1, we introduce the class of matrices with decay away from the diagonal or more general sparsity patterns. In Section 3.2, we delve into the correlation between the decay properties and polynomial approximations. We also illustrate the application of this technique to some of the bounds present in the literature for the inverse, general analytic functions and integral transforms. In Section 3.3, we focus on spectral projectors and the sign function, reporting the refined decay bounds obtained in [17]. In Section 3.4, we investigate the connection between decay bounds and the distribution of the eigenvalues. We refine the result in [46] for the matrix inverse and report the bounds in [17] for spectral projectors and the sign function. Bounds of this kind are also given for Cauchy-Stieltjes functions and the Fermi-Dirac function.

3.1 Matrices with Decay

Localization in numerical linear algebra is intended as the possibility to approximate a dense matrix with a sparse matrix, which can lead to huge computational savings in matrix computations. The storage can go from $\mathcal{O}(n^2)$ down to $\mathcal{O}(n)$, and the cost of many $\mathcal{O}(n^3)$ algorithms can also decrease substantially. See the survey [11] for more insights into this topic.

An important example of sparsity is provided by the class of banded matrices.

Definition 3.1. Let $A \in \mathbb{C}^{n\times n}$, and let m be a positive integer. We say that A is m-banded if $[A]_{ij} = 0$ for any i, j such that $|i - j| > m$.

The storage required for a banded matrix is $\mathcal{O}(mn)$, that corresponds to the entries $[A]_{ij}$ with $|i - j| \leq m$. The cost of elementary operations on banded matrices, such as matrix sums and multiplication, or the solution of linear systems, grows also linearly with the size. See, for example, [56].

Now we introduce the concept of *off-diagonal decay*.

Definition 3.2. Let $\{F_n\}_{n\in\mathbb{N}}$ be a sequence of matrices, $F_n \in \mathbb{C}^{n\times n}$ for all n. The matrix sequence has the *off-diagonal decay property* if

$$|[F_n]_{ij}| \leq \phi(|i-j|) \quad \text{for all } i \neq j,$$

where $\phi(x)$ is a real scalar function independent of n such that $\phi(x) \to 0$ as $x \to \infty$.

A typical example is the exponential decay, associated with $\phi(x) = K\, e^{-\alpha x}$ with $K, \alpha > 0$, or equivalently $\phi(x) = K\,\zeta^x$ with $K > 0$ and $0 < \zeta < 1$. Frequently, other decay behaviors emerge, such as algebraic (or power-law) decay associated with $\phi(x) = 1/(1+x)^\alpha$, $\alpha > 0$, or even more complex patterns, as we will see in Section 3.3.

Remark 3.1. An off diagonal decay property allows us to truncate the matrices in the sequence in the following way: corresponding to each F_n we define $F_n^{(m)}$, for a positive integer m, as follows:

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

Each matrix $F_n^{(m)}$ is m-banded. Moreover, if F_n has an exponential or algebraic decay property, then for all $\epsilon > 0$ and for $p = 1, 2, \infty$ there is an $\bar{m}$ independent of n such that $\|F_n - F_n^{(m)}\|_p \leq \epsilon$, for $m \geq \bar{m}$. See [16] for more details.

The foregoing considerations can be extended to matrices with more general decay patterns.

Definition 3.3. Let $\{F_n\}_{n\in\mathbb{N}}$ be a sequence of matrices, $F_n \in \mathbb{C}^{n\times n}$ for all n, and let $\mathcal{G}_n$ be a sequence of graphs with nodes $\{1, 2, \ldots, n\}$ and graph distances $\mathrm{d}_n(i,j)$. The matrices have the *decay property relative to the graph* $\mathcal{G}_n$ if

$$|[F_n]_{ij}| \leq \phi(\mathrm{d}_n(i,j)) \quad \text{for all } i \neq j,$$

where $\phi(x)$ is a real scalar function independent of n such that $\phi(x) \to 0$ as $x \to \infty$.

In this more general setting, we can still define the truncation $F_n^{(m)}$ as

$$[F_n^{(m)}]_{ij} = \begin{cases} [F_n]_{ij} & \text{if } \mathrm{d}(i,j) \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

However, some restrictions on the graphs must be imposed in order for these approximations to be sparse. For instance, if the degree of a node is $n-1$, then all the entries of $F_n^{(m)}$ are nonzero when $m \geq 2$. In general, a necessary hypothesis is that the degree of all the nodes in $\mathcal{G}_n$ is uniformly bounded in n. A more technical sufficient condition, based on the level sets of the graphs $\mathcal{G}_n$, is given in [47].

In the following sections, we consider decay bounds for matrix functions of Hermitian matrix arguments.

3.2 Decay Properties via Polynomial Approximations

A key ingredient to derive decay bounds for a matrix function $f(A)$ is the error of the best uniform polynomial approximation of f over a suitable set containing the spectrum of A. Denote with Π_k the set of all polynomials with degree at most k. Denote the error of the best uniform approximation in Π_k of a function f over a set $\mathcal{S}$ as

$$E_k(f, \mathcal{S}) = \inf_{p_k \in \Pi_k} \sup_{z \in \mathcal{S}} |f(z) - p_k(z)|. \tag{3.1}$$

Notice that if $\mathcal{S}$ is a real compact interval and f is real valued over $\mathcal{S}$ and continuous, then there exists a unique solution to the minimization problem (3.1), which becomes a minimum [80]. In general, (3.1) is not a minimum.

The following result [17], implicitly used in [11, 14, 15, 45], relates the polynomial approximations with the off-diagonal decay properties.

Proposition 3.1. *Let $A \in \mathbb{C}^{n \times n}$ be Hermitian and m-banded with $\sigma(A) \subset \mathcal{S}$, and let $f(x)$ be defined over $\mathcal{S}$. Let i, j be two indices such that $i \neq j$, and let $k := \left\lfloor \frac{|i-j|}{m} \right\rfloor$. Then*

$$|[f(A)]_{ij}| \leq E_k(f, \mathcal{S}). \tag{3.2}$$

If $\phi(x)$ is nonincreasing and $E_k(f, \mathcal{S}) \leq \phi(k)$, then

$$|[f(A)]_{ij}| \leq \phi\left(\left\lfloor \frac{|i-j|}{m} \right\rfloor\right) \leq \phi\left(\frac{|i-j|}{m} - 1\right).$$

Proof. Let $p_k \in \Pi_k$. Then $[p_k(A)]_{ij} = 0$ since $p_k(A)$ is km-banded and $|i-j| > km$. Therefore,

$$\begin{aligned} |[f(A)]_{ij}| = |[f(A)]_{ij} - [p_k(A)]_{ij}| \leq \|f(A) - p_k(A)\|_2 &= \max_{x \in \sigma(A)} |f(x) - p_k(x)| \\ &\leq \max_{x \in \mathcal{S}} |f(x) - p_k(x)|. \end{aligned}$$

Since the inequality holds for any $p_k \in \Pi_k$, by the definition of $E_k(f, \mathcal{S})$ we conclude that (3.2) holds. The second part follows from the inequality $\left\lfloor \frac{|i-j|}{m} \right\rfloor \geq \frac{|i-j|}{m} - 1$ and (3.2). □

Remark 3.2. The result of Proposition 3.1 implies that if we are given a sequence of $n \times n$ matrices $\{A_n\}$ of increasing size, all Hermitian, uniformly m-banded and such that $\sigma(A_n) \subset \mathcal{S}$ for all n, then $f(A_n)$ is well defined for all n and the bound (3.2) holds for all the matrices in the sequence, since it depends only on the set $\mathcal{S}$ and on the bandwidth, and not on n.

The same approach works more generally for a sparse matrix A; see [14, 16, 47].

Proposition 3.2. *Let $A \in \mathbb{C}^{n \times n}$ be Hermitian with $\sigma(A) \subset \mathcal{S}$, and let $f(x)$ be defined over $\mathcal{S}$. Let i, j be two indices such that $i \neq j$, and let $k := \mathrm{d}(i, j) - 1$, where $\mathrm{d}(i, j)$ is the geodesic distance in $\mathcal{G}(A)$. Suppose that $E_k(f, \mathcal{S}) \leq \phi(k)$. Then*

$$|[f(A)]_{ij}| \leq \phi\left(\mathrm{d}(i, j) - 1\right).$$

Proof. Let $k = \mathrm{d}(i, j) - 1$ be fixed. We have $k \geq 0$ since $i \neq j$. Then $[p_k(A)]_{ij} = 0$ for any $p_k \in \Pi_k$ since $\mathrm{d}(i, j) > k$. We conclude by proceeding as in Proposition 3.1. □

3.2.1 Decay Bounds for the Inverse

The error for the best uniform polynomial approximation of the function $f(x) = x^{-1}$ over an interval is explicitly known [80].

Theorem 3.3. *Let $f(x) = 1/x$ be defined over $[a, b]$, where $0 < a < b$. Let*

$$r = \frac{b}{a}, \quad K = \frac{(1+\sqrt{r})^2}{2b}, \quad \zeta = \frac{\sqrt{r}-1}{\sqrt{r}+1}. \tag{3.3}$$

Then

$$E_k(1/x, [a, b]) = K\zeta^{k+1} \tag{3.4}$$

for all k.

Theorem 3.3 has been used in [35] to obtain the following result regarding the entries of the matrix inverse.

Theorem 3.4. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian, positive definite and m-banded. Let $a = \lambda_{\min}(A)$, $b = \lambda_{\max}(A)$. Let r, K, ζ be defined as in* (3.3)*. Then, for any i and j such that $i \neq j$,*

$$|[A^{-1}]_{ij}| \leq K\,\zeta^{\frac{|i-j|}{m}}. \tag{3.5}$$

Theorem 3.4 follows directly from Proposition 3.1. In [35] the authors choose a different constant in order to capture the case $i = j$. In fact, $|[A^{-1}]_{ii}| \leq \|A^{-1}\|_2 = 1/a$ for any i, so if we choose the maximum between $1/a$ and the value of K in (3.3) we get a bound that holds for all i, j. Since it is important to keep the constant factors as small as possible, we think that it is more convenient to distinguish the two cases.

Theorem 3.4 holds only for Hermitian matrices. For the case of normal matrices, see [45].

A reader familiar with Krylov methods will recognize in the expression for ζ given in (3.3) the geometric rate of the bound for the error reduction (measured in the A-norm) of the conjugate gradient method applied to a linear system $Ax = b$ with a positive definite A; see, for example, [76, Section 5.6]. It is also well known that this bound can be overly pessimistic, and that much faster convergence can occur for certain distributions of the eigenvalues of A, for instance when most of the eigenvalues are clustered in the lower end of the spectrum. The following result [46] shows that this phenomenon holds also for the decay in the entries of the inverse.

Theorem 3.5. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian, positive definite and m-banded with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Let*

$$r_\ell = \frac{\lambda_{n-\ell}}{\lambda_1}, \quad \zeta_\ell = \frac{\sqrt{r_\ell}-1}{\sqrt{r_\ell}+1}, \quad K = \frac{2}{\lambda_1}.$$

Then the entries of A^{-1} are bounded by

$$|[A^{-1}]_{ij}| \leq K\zeta_\ell^{\frac{|i-j|}{m}-\ell} \quad \text{for all } \ell = 0, 1, \ldots, \left\lfloor \frac{|i-j|}{m} \right\rfloor. \tag{3.6}$$

The family of bounds given in Theorem 3.5 implies that we can remove some eigenvalues in the upper end of the spectrum and obtain a bound like the one in (3.5) with a smaller geometric rate that depends on the "effective" condition number r_ℓ, but paying the price of a smaller exponent. This means that, if some of the largest eigenvalues are isolated, the decay can be predicted much more accurately than (3.5), which is a special case of (3.6) with $\ell = 0$ up to a constant factor. We will return on this type of bounds in Section 3.4.

3.2.2 Analytic Functions and Bernstein's Theorem

In [15] the authors have shown that $f(A)$ has the exponential decay property whenever f is analytic on a suitable region containing $\sigma(A)$ and A is symmetric and banded. To analyze the polynomial approximations of f over the spectral interval, they used the following classical result [80, Theorem 73].

Theorem 3.6 (Bernstein's Theorem)**.** *Let $\mathcal{E}_\chi$ be the unique ellipse with foci at -1 and 1, with semiaxes $\kappa_1 > 1$ and $\kappa_2 > 0$, and $\chi = \kappa_1 + \kappa_2$. Let f be analytic in the interior of $\mathcal{E}_\chi$ and continuous on $\mathcal{E}_\chi$. Assume that $f(x)$ is real for real x. Then*

$$E_k(f, [-1, 1]) \leq \frac{2M(\chi)}{\chi^k(\chi - 1)}, \quad M(\chi) := \max_{z \in \mathcal{E}_\chi} |f(z)|.$$

Theorem 3.7 ([15])**.** *Let A be Hermitian and m-banded with spectrum $\sigma(A)$ contained in $[-1, 1]$, and let f be analytic in the interior of $\mathcal{E}_\chi$, $\chi > 1$, and continuous on $\mathcal{E}_\chi$. Let*

$$M(\chi) = \max_{z \in \mathcal{E}_\chi} |f(z)|, \quad K = \frac{2\chi M(\chi)}{\chi - 1}.$$

Then

$$|f(A)_{ij}| \leq K \left(\frac{1}{\chi}\right)^{\frac{|i-j|}{m}}, \quad \text{for all } i, j.$$

Remark 3.3. Theorem 3.7 gives us a family of bounds, depending on a parameter $\chi > 1$. Notice that for $\chi \to 1$ we have that $K \to \infty$, and if $\overline{\chi}$ is the minimal value such that $\mathcal{E}_{\overline{\chi}}$ contains a pole of f, we also have that $K \to \infty$ for $\chi \to \overline{\chi}$. A way to effectively use this bound is to optimize among the admissible values of χ for every entry in $f(A)$:

$$|[f(A)]_{ij}| \leq \inf_{1 < \chi < \overline{\chi}} \left[\frac{2\chi M(\chi)}{\chi - 1} \left(\frac{1}{\chi}\right)^{\frac{|i-j|}{2m}} \right].$$

This minimization can be carried numerically, by choosing a finite number of admissible values of χ and minimizing among them, or analytically if the expression is simple. See, for example, [14, Section 8].

Remark 3.4. The assumption $\sigma(A) \subset [-1, 1]$ is not restrictive. Indeed, if $\sigma(A) \subset [a, b]$, the shifted and scaled matrix

$$\tilde{A} = \frac{2}{b-a}A - \frac{a+b}{a-b}I$$

has spectrum contained in $[-1, 1]$. Then we can consider

$$\tilde{f}(x) = f(u), \quad u = \frac{b-a}{2}x + \frac{a+b}{2}$$

and get $f(A) = \tilde{f}(\tilde{A})$, so we can replace f by $\tilde{f}$ in our analysis.

Theorem 3.7 has been used, for example, in [14] to derive the bounds of Theorem 3.11 and Theorem 3.20 for spectral projectors and the Fermi-Dirac function, respectively. We will return to this in the following sections.

3.2.3 Functions Defined by an Integral Transform

We need the following continuity result for the polynomial of best uniform approximation [80].

Theorem 3.8 (Theorem 24 in [80])**.** *Let $f(x), \tilde{f}(x)$ be continuous functions over $[a, b]$, and let $p_k(x), \tilde{p}_k(x)$ be their polynomials of best uniform approximation over $[a, b]$ in Π_k. There exists a constant $M(f) > 0$ depending only on f such that*

$$|p_k(x) - \tilde{p}_k(x)| \leq M\,|f(x) - \tilde{f}(x)|$$

for all $x \in [a, b]$.

The following result comes from [18] and analyzes the polynomial approximations of a function expressed as an integral transform.

Lemma 3.9. *Let $f(x)$ be defined for $x \in [a, b]$ by*

$$f(x) = \int_0^\infty g_t(x)\,\mathrm{d}t, \tag{3.7}$$

where $g_t(x)$ is continuous for $(t, x) \in (0, \infty) \times [a, b]$ and the integral (3.7) *is absolutely convergent. Then*

$$E_k(f, [a, b]) \leq \int_0^\infty E_k(g_t(x), [a, b])\,\mathrm{d}t, \tag{3.8}$$

for all $k \geq 0$.

Proof. If the right-hand side of (3.8) is infinite, then the thesis is trivially true. From now on, we assume that the integral in (3.8) converges.

For all $t \in (0, \infty)$ and for any degree $k \geq 0$, since $g_t(x)$ is continuous over the compact interval $[a, b]$ there exists a unique polynomial $p_k^{(t)}(x) \in \Pi_k$ such that

$$E_k(g_t(x), [a, b]) = \max_{x \in [a,b]} \left| g_t(x) - p_k^{(t)}(x) \right|.$$

For all $x \in [a,b]$ we can define

$$p_k(x) := \int_0^\infty p_k^{(t)}(x)\,\mathrm{d}t. \tag{3.9}$$

This is well defined: for all $x \in [a,b]$, the mapping $t \mapsto p_k^{(t)}(x)$ is continuous for all $t \in (0,\infty)$ in view of Theorem 3.8, since $g_t(x)$ is uniformly continuous for (t,x) in the compact sets contained in $(0,\infty)\times[a,b]$. Moreover, the integral is absolutely convergent, since

$$\begin{aligned}\int_0^\infty \left|p_k^{(t)}(x)\right|\mathrm{d}t &\le \int_0^\infty |g_t(x)|\,\mathrm{d}t + \int_0^\infty \left|g_t(x) - p_k^{(t)}(x)\right|\mathrm{d}t \\ &\le \int_0^\infty |g_t(x)|\,\mathrm{d}t + \int_0^\infty E_k(g_t(x),[a,b])\mathrm{d}t < +\infty.\end{aligned} \tag{3.10}$$

We want to show that $p_k(x)$ is also a polynomial. Consider the expression

$$p_k^{(t)}(x) = \sum_{i=0}^k a_i(t)\,x^i,$$

which defines the coefficients $a_i(t)$ as functions of t. Formally, we have

$$p_k(x) = \sum_{i=0}^k \left(\int_0^\infty a_i(t)\,\mathrm{d}t\right)\cdot x^i = \sum_{i=0}^k \bar{a}_i x^i,$$

where $\bar{a}_i := \int_0^\infty a_i(t)\,\mathrm{d}t$. To conclude, we need to show that this expression is well defined, that is, the functions $a_i(t)$ are integrable in t. Consider $k+1$ distinct points $x_0,\dots,x_k$ in the interval $[a,b]$, and let $\boldsymbol{a}(t) = [a_0(t),\dots,a_k(t)]^T$, $\boldsymbol{p}(t) = [p_k^{(t)}(x_0),\dots,p_k^{(t)}(x_k)]^T$. We have that $V\boldsymbol{a}(t) = \boldsymbol{p}(t)$, where

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^k \\ 1 & x_1 & x_1^2 & \cdots & x_1^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^k \end{bmatrix}$$

is a Vandermonde matrix. Since V is nonsingular, we have that $\boldsymbol{a}(t) = V^{-1}\boldsymbol{p}(t)$. Then, if we let $V^{-1} = (c_{ij})_{i,j=0}^k$, we obtain the expression

$$a_i(t) = \sum_{j=0}^k c_{ij} p_k^{(t)}(x_j), \quad i = 0,\dots,k,$$

where c_{ij} is independent of t for all i,j. This shows that, for all i, $a_i(t)$ is continuous for $t \in (0,\infty)$, and we have

$$|\bar{a}_i| \le \int_0^\infty |a_i(t)|\,\mathrm{d}t \le \sum_{j=0}^k |c_{ij}| \int_0^\infty \left|p_k^{(t)}(x_j)\right|\mathrm{d}t < +\infty, \quad i = 0,\dots,k,$$

hence $\bar{a}_i$ is well defined for $i = 0, \dots, k$ and $p_k(x)$ is a polynomial of degree at most k. Finally, we have

$$\begin{aligned} E_k(f, [a,b]) &\le \max_{x\in[a,b]} |f(x) - p_k(x)| \\ &\le \max_{x\in[a,b]} \int_0^\infty \left| g_t(x) - p_k^{(t)}(x) \right| \, \mathrm{d}t \\ &\le \int_0^\infty E_k(g_t(x), [a,b]) \, \mathrm{d}t. \end{aligned}$$

This concludes the proof. □

Remark 3.5. The hypothesis of $g_t(x)$ being continuous for $(t,x) \in (0,\infty) \times [a,b]$ can be weakened since it is needed only to prove that the mapping $t \mapsto p_k^{(t)}(x)$ is measurable, so that the integrals in (3.9) and (3.10) are defined. For instance, we can assume that $g_t(x)$ is continuous over $((0,\infty) \setminus \mathcal{D}) \times [a,b]$, where $\mathcal{D} \subset (0,\infty)$ is a discrete set. In this way, Theorem 3.8 shows that $t \mapsto p_k^{(t)}(x)$ is continuous almost everywhere, hence it is measurable.

Many functions in the applications can be expressed as in (3.7). For instance, if $g_t(x) = h(t)/(x+t)$, then $f(x)$ belongs to the class of Cauchy-Stieltjes functions. The following bound generalizes and refines the results in [19, 45].

Theorem 3.10. *Let $f(x)$ be a Cauchy-Stieltjes function of the form*

$$f(x) = \int_0^\infty \frac{h(t)}{x+t} \, \mathrm{d}t$$

defined for $x \in [a,b]$, $0 < a < b$. Then

$$E_k(f, [a,b]) \le \frac{1}{2} \left(\sqrt{f(a)} + \sqrt{f(b)} \right)^2 \left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \right)^{k+1}.$$

Proof. By applying Lemma 3.9 and Theorem 3.3, we get

$$\begin{aligned} E_k(f(x), [a,b]) \le \int_0^\infty h(t) \, E_k(1/(x+t), [a,b]) \, \mathrm{d}t &= \int_0^\infty h(t) \, K(t) \, \zeta(t)^{k+1} \, \mathrm{d}t \\ &\le \left(\int_0^\infty h(t) \, K(t) \, \mathrm{d}t \right) \zeta(0)^{k+1}, \end{aligned}$$

where

$$K(t) = \frac{(1 + \sqrt{r(t)})^2}{2(b+t)}, \quad \zeta(t) = \frac{\sqrt{r(t)} - 1}{\sqrt{r(t)} + 1}, \quad r(t) = \frac{b+t}{a+t}$$

for $t \ge 0$. Proceeding as in the proof of [45, Theorem 4], we get

$$\int_0^\infty h(t) K(t) \, \mathrm{d}t = \frac{1}{2} \int_0^\infty \frac{h(t)}{a+t} \, \mathrm{d}t + \frac{1}{2} \int_0^\infty \frac{h(t)}{b+t} \, \mathrm{d}t + \int_0^\infty \frac{h(t)}{\sqrt{(a+t)(b+t)}} \, \mathrm{d}t.$$

The first two terms are equal to $f(a)/2$ and $f(b)/2$, respectively, in view of the representation of f. The third term can be bounded by using the Cauchy-Schwarz inequality:

$$\int_0^\infty \frac{h(t)}{\sqrt{(a+t)(b+t)}}\,\mathrm{d}t \le \left(\int_0^\infty \frac{h(t)}{a+t}\,\mathrm{d}t\right)^{\frac{1}{2}} \left(\int_0^\infty \frac{h(t)}{b+t}\,\mathrm{d}t\right)^{\frac{1}{2}} = \sqrt{f(a)f(b)}.$$

Finally,

$$\begin{aligned} E_k(f(x),[a,b]) &\le \frac{1}{2}\left(f(a)+f(b)+2\sqrt{f(a)f(b)}\right)\zeta(0)^{k+1} \\ &= \frac{1}{2}\left(\sqrt{f(a)}+\sqrt{f(b)}\right)^2 \left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}\right)^{k+1}. \end{aligned}$$

This concludes the proof. □

The application of Theorem 3.10 to derive decay bounds is straightforward in view of Proposition 3.1 or Proposition 3.2.

Although this class is wider and includes transforms associated with more general measures, many important examples have this simple representation. For instance,

$$\begin{aligned} x^{-\alpha} &= \frac{\sin(\alpha\pi)}{\pi}\int_0^\infty \frac{t^{-\alpha}}{x+t}\,\mathrm{d}t, \\ \frac{\log(1+x)}{x} &= \int_1^\infty \frac{1}{t\,(x+t)}\,\mathrm{d}t. \end{aligned}$$

Other functions that belong to the case (3.7) are Laplace-Stieltjes functions, characterized by $g_t(x) = h(t)\,e^{-tx}$ [19, 79], the sign function (Section 3.3), the Fermi-Dirac function (Section 3.3.7) and $-x\log x$, which is involved in the definition of the von Neumann entropy (Chapter 5).

3.3 Refined Bounds for Spectral Projectors and the Sign Function

The ability to approximate spectral projectors associated with banded or sparse matrices is crucial for the development of linear scaling methods in electronic structure computations, as discussed in [14, 24, 70, 83].

3.3.1 Spectral Projectors in Electronic Structure Computations

In electronic structure computations, we deal with a quantum systems of electrons at a certain temperature $T \ge 0$. A continuous Hamiltonian is discretized via a finite difference scheme or a Galarkin projection, leading to the Hermitian matrix H_n of size $n = n_b \cdot n_e$, where n_b is the number of basis functions and depends only on the projection scheme, while n_e is the number of electrons, so we consider this as an increasing parameter.

Let $\lambda_1 \leq \cdots \lambda_{n_e} < \lambda_{n_e+1} \leq \cdots \leq \lambda_n$ be the eigenvalues of H_n, with associated orthonormal eigenvectors $\boldsymbol{v}_i$, $i = 1, \ldots, n$. The first n_e eigenvalues are called *occupied levels*, and the associated eigenvectors are called *occupied states*. For systems with temperature $T = 0$, a key role in the computations is played by the spectral projector associated with the occupied states [70, 83],

$$P_n = \boldsymbol{v}_1 \boldsymbol{v}_1^* + \cdots + \boldsymbol{v}_{n_e} \boldsymbol{v}_{n_e}^* = \sum_{i=1}^{n_e} \boldsymbol{v}_i \boldsymbol{v}_i^*.$$

We can also write it as the matrix function $P_n = h(H_n)$, where h is the Heaviside function

$$h(x) = \begin{cases} 1 & \text{if } x < \mu, \\ \frac{1}{2} & \text{if } x = \mu, \\ 0 & \text{if } x > \mu, \end{cases} \tag{3.11}$$

and $\lambda_{n_e} < \mu < \lambda_{n_e+1}$ (μ is also called *Fermi level* or *Fermi energy*).

Let us consider the sequences of matrices $\{H_n\}$ and $\{P_n\}$. In order to establish a common decay on the entries for all the projectors P_n, the authors in [14] identified the following reasonable assumptions:

- the matrices H_n have uniformly bounded bandwidth;
- there exist four parameters $b_1 < a_1 < a_2 < b_2$ independent of n such that $\sigma(H_n) \subset [b_1, a_1] \cup [a_2, b_2]$ and, for all n_e, $[b_1, a_1]$ contains the first n_e eigenvalues of H_n while $[a_2, b_2]$ contains the remaining $n - n_e$.

The key quantity here is the relative spectral gap $\gamma = (a_2 - a_1)/(b_2 - b_1)$. If this quantity is not too small (e.g., insulators or semiconductors) then the projector exhibits exponential decay, while for small or vanishing gap (e.g., metallic systems) the decay can be very slow [14].

For systems with positive temperature $T > 0$, the role previously played by the projector is taken by $f_{FD}(H_n)$, where f_{FD} is the *Fermi-Dirac function*

$$f(H_n) = \frac{1}{1 + e^{\beta(x-\mu)}}, \tag{3.12}$$

where $\beta = (\kappa_B T)^{-1}$ and κ_B is the Boltzmann's constant. Note that, for $T \to 0$, we get $f_{FD}(H_n) \to P_n$. Hence, for large values of β, $f_{FD}(H_n)$ can be used as an approximation of P_n; see [14]. On the other hand, when T is sufficiently large, a decay is present also in the vanishing gap case corresponding to metallic systems.

3.3.2 Previous Work

Throughout this section, we consider a single Hermitian, banded $H \in \mathbb{C}^{n \times n}$ with spectrum contained in $[b_1, a_1] \cup [a_2, b_2]$ and its spectral projector $P = h(H)$. We keep in mind that, in the applications, H is an element of a matrix sequence with uniformly bounded bandwidth and spectrum, as described above.

Since the Heaviside function is discontinuous, the quantity $E_k(h, [b_1, b_2])$ does not converge to 0, so we cannot use directly Proposition 3.1 to obtain decay bounds for P. In [14] the authors considered the approximation of $h(x)$ over the spectrum of H with the Fermi-Dirac function, which is analytic over a family of ellipses containing $[b_1, b_2]$, so the entries of $f_{FD}(H)$ decay exponentially in view of Bernstein's Theorem; see Theorem 3.20. Usually, a large value of β is required to have a good approximation $f_{FD}(x) \approx h(x)$ over $[b_1, a_1] \cup [a_2, b_2]$, and this often leads to pessimistic decay bounds.

Since the spectral projector of H depends only on the eigenvectors associated with the first n_e eigenvalues, a scaled and shifted modification such as

$$\hat{H} = c\,H + d\,I, \quad c > 0, \quad d \in \mathbb{R},$$

has the same spectral projector. This allows us to make convenient assumptions on the spectrum. For instance, if $d = -(a_2 + a_1)/2$ then $\sigma(\hat{H}) \subset [\hat{b}_1, -a] \cup [a, \hat{b}_2]$, with $a = (a_2 - a_1)/2$, so we can choose $\mu = 0$ in (3.11). Then, by putting $b = \max\{\hat{b}_2, -\hat{b}_1\}$, we have $\sigma(\hat{H}) \subset [-b, -a] \cup [a, b]$. When dealing with a sequence, the transformation must be the same for all the matrices. This can be done if all the spectra are contained in the same union of intervals.

Another approach, which turns out to be better than the one based on the Fermi-Dirac approximation, consists in estimating directly the error of the polynomial approximation of the Heaviside function over $[-b, -a] \cup [a, b]$. If $\mu = 0$, then the identity $h(x) = (1 - \mathrm{sign}(x))/2$ holds, where

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

hence $[P]_{ij} = -\frac{1}{2}[\mathrm{sign}(H)]_{ij}$ for $i \neq j$. So it is equivalent to study the decay properties of $\mathrm{sign}(H)$ instead of P. Moreover, $\mathrm{sign}(x) = x/|x| = x/(x^2)^{\frac{1}{2}}$ for any $x \neq 0$. The main idea is to consider a polynomial $q_k(x) \in \Pi_k$ which approximates $x^{-\frac{1}{2}}$ over $[a^2, b^2]$, and then construct $p_{2k+1}(x) = \frac{1}{2}(1 - x q_k(x^2))$ to approximate $h(x)$. This, using Theorem 3.6, leads to the following result [14].

Theorem 3.11. *Let H be Hermitian and m-banded with $\sigma(H) \subset [-b, -a] \cup [a, b]$, and let $P = h(H)$ be the spectral projector associated with the negative eigenvalues of H. Then, for $1 < \xi < \bar{\xi} := \frac{b+a}{b-a}$, we have*

$$|[P]_{ij}| \leq \frac{2b\xi M(\xi)}{\xi - 1} \left(\frac{1}{\xi}\right)^{\frac{|i-j|}{2m}} \quad \textit{for all } i, j,$$

where

$$M(\xi) = \frac{1}{\sqrt{z_0}}, \quad z_0 = \left[\frac{b^2 + a^2}{b^2 - a^2} - \frac{\xi^2 + 1}{2\xi}\right] \frac{b^2 - a^2}{2}.$$

As discussed in Remark 3.3, we can optimize the bound among the admissible values of ξ, so

$$|[P]_{ij}| \leq \inf_{1<\xi<\bar{\xi}} \left[\frac{2b\xi M(\xi)}{\xi - 1} \left(\frac{1}{\xi} \right)^{\frac{|i-j|}{2m}} \right]. \tag{3.13}$$

Remark 3.6. Other features of the spectral projector follow from the fact that $P = P^2$. An important consequence of this identity is that $\sigma(P) \subset \{0, 1\}$, so $|[P]_{ij}| \leq \|P\|_2 = 1$ for any i, j. This means that any bound for the entries of P is useless unless it is less than 1 too.

3.3.3 Exploiting an Integral Representation of the Sign Function

In this section we report the new decay bounds in [17] for the spectral projector $P = h(H)$, where H is banded, Hermitian and with spectrum contained in the union of two symmetric intervals and $h(x)$ is the Heaviside function defined as in (3.11) with $\mu = 0$. For this purpose, since $P = \frac{1}{2}(I - \operatorname{sign}(H))$, we study the decay properties of $\operatorname{sign}(H)$ instead of P. Numerical validation of the results is also given with some experiments.

Consider the following integral representation of $\operatorname{sign}(x)$ [67, Chapter 5]:

$$\operatorname{sign}(x) = \frac{2}{\pi} \int_0^\infty \frac{x}{x^2 + t^2} \, \mathrm{d}t. \tag{3.14}$$

In view of Lemma 3.9 we can reduce the problem to study the polynomial approximations of $f_t(x) = x(x^2 + t^2)^{-1}$. Notice that $f_t(x) = x g_t(x^2)$ where $g_t(x) = (x + t^2)^{-1}$. The following result is specific for odd functions of this form.

Lemma 3.12. *Let $f(x) = x g(x^2)$ be defined for $x \in [-b, -a] \cup [a, b]$, where $g(x)$ is defined for $x \in [a^2, b^2]$. Let $k \geq 0$ and $s = \left\lceil \frac{k-1}{2} \right\rceil$. Then*

$$E_k(f, [-b, -a] \cup [a, b]) \leq b \, E_s(g, [a^2, b^2]).$$

Proof. Suppose that k is odd, so $k = 2s + 1$ with $s \geq 0$. Let $q_s(x) \in \Pi_s$ and $p_k(x) = x q_s(x^2) \in \Pi_k$. Then

$$\begin{aligned} E_k(f, [-b, -a] \cup [a, b]) &\leq \sup_{x \in [-b, -a] \cup [a, b]} |f(x) - p_k(x)| \\ &= \sup_{x \in [-b, -a] \cup [a, b]} |x| \cdot |g(x^2) - q_s(x^2)| \\ &\leq b \sup_{x \in [a^2, b^2]} |g(x) - q_s(x)|. \end{aligned}$$

Since this holds for all $q_s(x) \in \Pi_s$, by the definition in (3.1) we get

$$E_k(f, [-b, -a] \cup [a, b]) \leq b \, E_s(g, [a^2, b^2]).$$

If k is even, then $k - 1 = 2s + 1$ is odd, and we can use the inequality

$$E_k(f, [-b, -a] \cup [a, b]) \leq E_{k-1}(f, [-b, -a] \cup [a, b]) \leq E_s(g, [a^2, b^2]).$$

This concludes the proof. □

Now we can bound $E_k(\operatorname{sign}(x), [-b,-a]\cup[a,b])$.

Theorem 3.13. *Let $0 < a < b$, and let, for any $t \geq 0$,*

$$r(t) := \frac{b^2+t^2}{a^2+t^2}, \quad K(t) := \frac{(1+\sqrt{r(t)})^2}{2(b^2+t^2)}, \quad \zeta(t) := \frac{\sqrt{r(t)}-1}{\sqrt{r(t)}+1}. \tag{3.15}$$

Then

$$E_k(\operatorname{sign}(x), [-b,-a]\cup[a,b]) \leq \frac{2b}{\pi}\int_0^\infty K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t, \tag{3.16}$$

and

$$E_k(\operatorname{sign}(x), [-b,-a]\cup[a,b]) \leq \frac{1}{2}\left(1+\sqrt{\frac{b}{a}}\right)^2 \cdot \left(\frac{b-a}{b+a}\right)^{\frac{k}{2}}. \tag{3.17}$$

Proof. Let $k \geq 0$. Let $f_t(x) = x(x^2+t^2)^{-1} = xg_t(x^2)$ for $t \geq 0$, where $g_t(x) = (x+t^2)^{-1}$, so that

$$\operatorname{sign}(x) = \frac{2}{\pi}\int_0^\infty f_t(x)\,\mathrm{d}t.$$

From (3.4), we get that $E_k(g_t(x), [a^2,b^2]) = E_k(1/x, [a^2+t^2, b^2+t^2]) = K(t)q(t)^{k+1}$. Therefore, in view of (3.14), Lemma 3.9, and Lemma 3.12, since $\lceil \frac{k-1}{2} \rceil \geq \frac{k-2}{2}$,

$$\begin{aligned} E_k(\operatorname{sign}(x), [-b,-a]\cup[a,b]) &\leq \frac{2}{\pi}\int_0^\infty E_k(f_t(x), [-b,-a]\cup[a,b]\mathrm{d}t \\ &\leq \frac{2b}{\pi}\int_0^\infty K(t)\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t. \end{aligned}$$

For (3.17), since $\frac{k}{2} \geq 0$ and $\zeta(t) \leq \zeta(0)$ for all $t \geq 0$, it holds that

$$\frac{2b}{\pi}\int_0^\infty K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t \leq \frac{2b}{\pi}\int_0^\infty K(t)\,\mathrm{d}t\cdot\zeta(0)^{\frac{k}{2}} = \frac{2b}{\pi}\int_0^\infty K(t)\,\mathrm{d}t\cdot\left(\frac{b-a}{b+a}\right)^{\frac{k}{2}}. \tag{3.18}$$

In order to estimate the integral, we can expand it as follows:

$$\begin{aligned} &\int_0^\infty K(t)\,\mathrm{d}t = \int_0^\infty \frac{(1+\sqrt{r(t)})^2}{2(b^2+t^2)}\,\mathrm{d}t \\ &= \frac{1}{2}\left(\int_0^\infty \frac{1}{b^2+t^2}\,\mathrm{d}t + \int_0^\infty \frac{1}{a^2+t^2}\,\mathrm{d}t + 2\int_0^\infty \frac{1}{\sqrt{(b^2+t^2)(a^2+t^2)}}\,\mathrm{d}t\right). \end{aligned}$$

The first two terms can be explicitly computed:

$$\int_0^\infty \frac{1}{b^2+t^2}\,\mathrm{d}t = \frac{\pi}{2b}, \quad \int_0^\infty \frac{1}{a^2+t^2}\,\mathrm{d}t = \frac{\pi}{2a}.$$

The third can be bounded by using the Cauchy-Schwarz inequality:

$$\int_0^\infty \frac{1}{\sqrt{(b^2+t^2)(a^2+t^2)}}\,\mathrm{d}t \le \left(\int_0^\infty \frac{1}{b^2+t^2}\,\mathrm{d}t\right)^{\frac{1}{2}} \cdot \left(\int_0^\infty \frac{1}{a^2+t^2}\,\mathrm{d}t\right)^{\frac{1}{2}} = \frac{\pi}{2\sqrt{ab}}.$$

Therefore

$$\frac{2b}{\pi}\int_0^\infty K(t)\,\mathrm{d}t \le \frac{1}{2}\left(1+\frac{b}{a}+2\sqrt{\frac{b}{a}}\right) = \frac{1}{2}\left(1+\sqrt{\frac{b}{a}}\right)^2. \tag{3.19}$$

This concludes the proof. □

By using these results, we can derive bounds for the entries of $\mathrm{sign}(H)$ and spectral projectors.

Theorem 3.14. *Let H be Hermitian and m-banded with $\sigma(H) \subset [-b,-a]\cup[a,b]$. Let $\mathrm{sign}(H)$ be the matrix sign function and $P = h(H)$ be the spectral projector associated with the negative eigenvalues. Let*

$$\hat{K} := \frac{1}{4}\left(1+\sqrt{\frac{b}{a}}\right)^2, \quad \hat{\zeta} := \frac{b-a}{b+a}.$$

Then

$$|[\mathrm{sign}(H)]_{ij}| \le 2\hat{K}\,\hat{\zeta}^{\frac{|i-j|}{2m}-\frac{1}{2}} \quad \text{for } |i-j| \ge m, \tag{3.20}$$

$$|[P]_{ij}| \le \hat{K}\,\hat{\zeta}^{\frac{|i-j|}{2m}-\frac{1}{2}} \quad \text{for all } i,j. \tag{3.21}$$

Proof. The inequality (3.20) follows from Theorem 3.13 combined with Lemma 3.1. Regarding (3.21), if $|i-j| \ge m$, the inequality follows directly from (3.20) and the identity $|[P]_{ij}| = |[\mathrm{sign}(H)]_{ij}|/2$. For $|i-j| < m$, note that $|[P]_{ij}| \le 1$ for all i,j (see Remark 3.6) and the right-hand side of (3.21) is greater than 1. □

Remark 3.7. This result improves the bound in (3.13) since all the geometric rates are smaller than the one in (3.21). Moreover, one does not need to choose the better estimate among a family of bounds.

The general (non-banded) case follows by using Proposition 3.2 with essentially the same proof of Theorem 3.14.

Theorem 3.15. *Let H be Hermitian with $\sigma(H) \subset [-b,-a]\cup[a,b]$. Then,*

$$|[P]_{ij}| \le \hat{K}\hat{\zeta}^{\frac{\mathrm{d}(i,j)-1}{2}} \quad \text{for all } i,j,$$

where $\mathrm{d}(i,j)$ is the geodesic distance in $\mathcal{G}(H)$ and $\hat{K}$, $\hat{\zeta}$ are as in Theorem 3.14.

3.3.4 An Asymptotically Optimal Bound

Although the bound given in Theorem 3.14 behaves well in practice, it is not optimal from an asymptotic point of view. Hasson showed in [65] that there exists $K > 0$ such that

$$E_k(\operatorname{sign}(x), [-b, -a] \cup [a, b]) \leq \frac{K}{\sqrt{k}} \left(\frac{b-a}{b+a} \right)^{\frac{k}{2}}.$$

By Proposition 3.1, this leads to

$$|[\operatorname{sign}(H)]_{ij}| \leq \frac{K}{\sqrt{\frac{|i-j|}{m} - 1}} \left(\frac{b-a}{b+a} \right)^{\frac{|i-j|}{2m} - \frac{1}{2}}, \tag{3.22}$$

that is asymptotically faster than the bound in (3.20). This is actually the best result we can obtain by using polynomial approximations of the sign function if the only available spectral information is that $\sigma(H) \subset [-b, -a] \cup [a, b]$, since

$$\sqrt{k} \left(\frac{b+a}{b-a} \right)^{\frac{k}{2}} \cdot E_k(\operatorname{sign}(x), [-b, -a] \cup [a, b]) = \mathcal{O}(1) \quad \text{as } k \to \infty.$$

See [38] for more details. Here the disadvantage is that we do not know an explicit formula for K. In what follows we will obtain, by manipulating the integral in (3.16), a decay that is asymptotically equivalent to (3.22) but with computable parameters.

In order to obtain better bounds, we start with the inequality (3.16). In the proof of Theorem 3.13, a key argument used in (3.18) is the inequality

$$\zeta(t)^{\frac{k}{2}} \leq \zeta(0)^{\frac{k}{2}} \quad \text{for any } t \geq 0. \tag{3.23}$$

The next result gives us a better estimate of the left-hand side in (3.23).

Lemma 3.16. *Let $\zeta(t)$ be defined as in* (3.15)*, and let $\alpha > 0$ be real. Then*

$$\zeta(t)^{\alpha} \leq e^{-\alpha(c_1 - t^2 c_2)t^2} \zeta(0)^{\alpha} \quad \textit{for all } t \geq 0,$$

where

$$c_1 = \frac{1}{2ab}, \quad c_2 = \frac{a^2 + ab + b^2}{8a^3b^3}.$$

Moreover, for any fixed τ such that $0 < \tau < \sqrt{\frac{c_1}{c_2}}$ and $\sigma(\tau) := c_1 - \tau^2 c_2$, we have

$$\zeta(t)^{\alpha} \leq e^{-\alpha\sigma(\tau)t^2} \zeta(0)^{\alpha} \quad \textit{for } 0 \leq t \leq \tau. \tag{3.24}$$

Proof. Let $t \geq 0$ be fixed. We have

$$\zeta(t)^{\alpha} = \left(\frac{\sqrt{b^2 + t^2} - \sqrt{a^2 + t^2}}{\sqrt{b^2 + t^2} + \sqrt{a^2 + t^2}} \right)^{\alpha} \leq \left(\frac{1}{b+a} \right)^{\alpha} \cdot \left(\sqrt{b^2 + t^2} - \sqrt{a^2 + t^2} \right)^{\alpha},$$

so

$$\frac{\zeta(t)^\alpha}{\zeta(0)^\alpha} \le \frac{\left(\sqrt{b^2+t^2}-\sqrt{a^2+t^2}\right)^\alpha}{(b-a)^\alpha}. \tag{3.25}$$

Consider the function $s(x) = \sqrt{b^2+x}-\sqrt{a^2+x}$ defined for $x > 0$, and its Taylor expansion with Lagrange remainder centered in 0:

$$s(x) = b - a - \frac{1}{2}\frac{b-a}{ab}x + \frac{1}{2}s''(\xi)x^2,$$

for some ξ between 0 and x. Since

$$\begin{aligned} s''(\xi) = \frac{1}{4}\left(\frac{1}{(a^2+\xi)^{3/2}} - \frac{1}{(b^2+\xi)^{3/2}}\right) &\le \frac{1}{4}\left(\frac{1}{a^3}-\frac{1}{b^3}\right) \\ &= \frac{1}{4}\frac{(b-a)(a^2+ab+b^2)}{a^3b^3}, \end{aligned}$$

we have

$$s(x) \le (b-a)\left(1 - \frac{1}{2ab}x + \frac{1}{8}\frac{a^2+ab+b^2}{a^3b^3}x^2\right) = (b-a)(1-c_1x+c_2x^2). \tag{3.26}$$

Note that $s(x)$ is a positive function, so $1 - c_1x + c_2x^2$ is also positive in view of (3.26). Since the numerator in (3.25) is $s(t^2)^\alpha$, we have

$$\frac{\zeta(t)^\alpha}{\zeta(0)^\alpha} \le (1-c_1t^2+c_2t^4)^\alpha = e^{\alpha\log(1-c_1t^2+c_2t^4)} \le e^{\alpha(-c_1t^2+c_2t^4)} = e^{-\alpha(c_1-c_2t^2)t^2},$$

where for the last inequality we have used that $\log(x) \le x-1$ for all $x > 0$.

For (3.24), observe that if $0 \le t \le \tau$ then $c_1 - t^2c_2$ achieves its minimum in τ, so $c_1 - t^2c_2 \ge c_1 - \tau^2c_2 = \sigma(\tau)$ for $0 \le t \le \tau$. This concludes the proof. □

Remark 3.8. The inequality (3.24) is uniform in t as long as $0 \le t \le \tau$. Moreover, for $\tau < \sqrt{\frac{c_1}{c_2}}$, we have that $\sigma(\tau) > 0$. This means that $\zeta(t)^\alpha$ is bounded by $\zeta(0)^\alpha$ times a Gaussian function in the variable t.

Theorem 3.17. *Let $0 < a < b$, and let $c_1 = \frac{1}{2ab}$, $c_2 = \frac{a^2+ab+b^2}{8a^3b^3}$ and $0 < \tau < \bar\tau := \sqrt{\frac{c_1}{c_2}}$. Then, for all k,*

$$E_k(\operatorname{sign}(x), [-b,-a]\cup[a,b]) \le \frac{K_1(\tau)}{\sqrt{k}}\zeta(0)^{\frac{k}{2}} + K_2\zeta(\tau)^{\frac{k}{2}},$$

where $\zeta(t)$ is defined as in (3.15) *and*

$$\begin{aligned} K_1(\tau) &= \frac{1}{b\sqrt{2\pi\sigma(\tau)}}\left(1+\frac{b}{a}\right)^2, \quad \sigma(\tau) = c_1 - \tau^2c_2, \\ K_2 &= \frac{1}{2}\left(1+\sqrt{\frac{b}{a}}\right)^2 \end{aligned} \tag{3.27}$$

for all $k \ge 0$.

Proof. From Theorem 3.13 we have

$$E_k(\operatorname{sign}(x), [-b,-a]\cup[a,b]) \leq \frac{2b}{\pi}\int_0^\infty K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t, \tag{3.28}$$

where $K(t)$ is defined in (3.15). We split the integral in two terms as follows:

$$\int_0^\infty K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t = \int_0^\tau K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t + \int_\tau^\infty K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t.$$

The first term can be bounded by using Lemma 3.16, as t ranges from 0 to τ, and the inequality $K(t) \leq K(0) = (1+b/a)^2/2b$:

$$\begin{aligned}
\frac{2b}{\pi}\int_0^\tau K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t &\leq \frac{2b}{\pi}C(0)\int_0^\tau \zeta(t)^{\frac{k}{2}}\,\mathrm{d}t \\
&\leq \frac{1}{b\pi}\left(1+\frac{b}{a}\right)^2 \zeta(0)^{\frac{k}{2}}\int_0^\tau e^{-\frac{k}{2}\sigma(\tau)t^2}\,\mathrm{d}t \\
&\leq \frac{1}{b\pi}\left(1+\frac{b}{a}\right)^2 \zeta(0)^{\frac{k}{2}}\int_0^\infty e^{-\frac{k}{2}\sigma(\tau)t^2}\,\mathrm{d}t \\
&= \frac{1}{2b\pi}\left(1+\frac{b}{a}\right)^2 \zeta(0)^{\frac{k}{2}}\cdot\sqrt{\frac{2\pi}{k\,\sigma(\tau)}} \\
&= \frac{1}{b\sqrt{2\pi\sigma(\tau)}}\left(1+\frac{b}{a}\right)^2 \frac{1}{\sqrt{k}}\zeta(0)^{\frac{k}{2}}.
\end{aligned}$$

For the second term:

$$\begin{aligned}
\frac{2b}{\pi}\int_\tau^\infty K(t)\,\zeta(t)^{\frac{k}{2}}\,\mathrm{d}t &\leq \frac{2b}{\pi}\left(\int_0^\infty K(t)\,\mathrm{d}t\right)\zeta(\tau)^{\frac{k}{2}} \\
&\leq \frac{1}{2}\left(1+\sqrt{\frac{b}{a}}\right)^2 \zeta(\tau)^{\frac{k}{2}},
\end{aligned}$$

where we have used (3.19). Combining these two inequalities with (3.28), we conclude. □

Remark 3.9. Since $\zeta(\tau) < \zeta(0)$ for any $\tau > 0$, the second term in (3.29) decays faster than the first. Hence, the asymptotic behavior of this bound is equal to the one in (3.22), but with computable parameters.

The combination of Theorem 3.17 and Proposition 3.1 yields the following result.

Theorem 3.18. *Let H be Hermitian and m-banded with $\sigma(H) \subset [-b,-a]\cup[a,b]$. Let $\operatorname{sign}(H)$ be the matrix sign function and $P = h(H)$ be the spectral projector associated with the negative eigenvalues. Let $c_1 = \frac{1}{2ab}$, $c_2 = \frac{a^2+ab+b^2}{8a^3b^3}$ and $0 < \tau < \bar{\tau} := \sqrt{\frac{c_1}{c_2}}$. Let $\zeta(t)$ be defined as in* (3.15)*, and let $K_1(\tau), K_2$ and $\sigma(\tau)$ be defined as in* (3.27)*. Then, for $|i-j| \geq m$,*

$$|[\operatorname{sign}(H)]_{ij}| \leq \frac{K_1(\tau)}{\sqrt{\frac{|i-j|}{m}-1}}\zeta(0)^{\frac{|i-j|}{2m}-\frac{1}{2}} + K_2\zeta(\tau)^{\frac{|i-j|}{2m}-\frac{1}{2}},$$

and

$$|[P]_{ij}| \leq \frac{1}{2}\left(\frac{K_1(\tau)}{\sqrt{\frac{|i-j|}{m}-1}}\zeta(0)^{\frac{|i-j|}{2m}-\frac{1}{2}} + K_2\zeta(\tau)^{\frac{|i-j|}{2m}-\frac{1}{2}}\right). \tag{3.29}$$

Remark 3.10. The bounds in Theorem 3.18 depend on the choice of τ, which ranges between 0 and $\bar{\tau}$. As in 3.11, we have a whole family of bounds which can be optimized among the admissible values of τ.

We also have the counterpart of Theorem 3.18 for general sparse matrices.

Theorem 3.19. *Let H be as in Theorem 3.18 but without the hypothesis that it is banded. Then, for $i \neq j$,*

$$|[\mathrm{sign}(H)]_{ij}| \leq \frac{K_1(\tau)}{\sqrt{\mathrm{d}(i,j)-1}}\zeta(0)^{\frac{\mathrm{d}(i,j)-1}{2}} + K_2\zeta(\tau)^{\frac{\mathrm{d}(i,j)}{2}},$$

and

$$|[P]_{ij}| \leq \frac{1}{2}\left(\frac{K_1(\tau)}{\sqrt{\mathrm{d}(i,j)-1}}\zeta(0)^{\frac{\mathrm{d}(i,j)-1}{2}} + K_2\zeta(\tau)^{\frac{\mathrm{d}(i,j)}{2}}\right),$$

where all the parameters are defined as in Theorem 3.18.

3.3.5 Comparison of Existing Bounds

For the next experiments, we will assume that $\sigma(H) \subset [-1,-a] \cup [a,1]$, so $b = 1$. This is not restrictive, since we can scale and shift the matrix in order to satisfy the condition.

Since any bound for a generic entry $[P]_{ij}$ of the spectral projector associated with a banded H depends only on the value $|i-j|$, in order to study the exact decay of P we can consider the quantities

$$D_P(k) := \max_{|i-j|=k} |[P]_{ij}|, \quad \text{for any } k \geq 0.$$

Let us denote the bounds for $D_P(k)$ induced by (3.13), (3.21) and (3.29) for $|i-j| = k$ as $B_1(k)$, $B_2(k)$ and $B_3(k)$, respectively. The third is optimized among the admissible values of τ as described in Remark 3.10. The components of P satisfy $|[P]_{ij}| \leq 1$ for all i, j, so it is convenient to use the following bound:

$$D_P(k) \leq \min\{1, B_s(k)\},$$

for any $s = 1, 2, 3$.

For the tests we have constructed Hermitian matrices with prescribed size, bandwidth, and spectrum with the following procedure.

Algorithm 1 Matrix with prescribed bandwidth and spectrum

Input: Eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$, bandwidth $m > 0$
Output: $A \in \mathbb{R}^{n\times n}$ symmetric, m-banded such that $\sigma(A) = \{\lambda_1, \dots, \lambda_n\}$

1: Sample $\Omega \in \mathbb{R}^{n\times n}$ random matrix, compute the QR factorization $\Omega = QR$
2: Set $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $A_m = Q\Lambda Q^*$
3: **for** $k = m, \dots, n-1$ **do**
4: $\quad \tilde{\boldsymbol{x}} = A_k(k:n, k-m+1) \in \mathbb{R}^{n-k+1}$, $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{x}} - \|\tilde{\boldsymbol{x}}\|_2 \boldsymbol{e}_1$, $\boldsymbol{u} = [0, \dots, 0, \tilde{\boldsymbol{u}}]$
5: $\quad H_k = I - \beta \boldsymbol{u}\boldsymbol{u}^T$, $\beta = 2/\boldsymbol{u}^T\boldsymbol{u}$
6: $\quad A_{k+1} = H_k A_k H_k$ ▷ Implicit operation, H_k is never formed
7: **end for**
8: **return** $A = A_n$

Algorithm 1 is a variant of [56, Algorithm 7.4.2], which returns a tridiagonal matrix. At each step of the loop, the principal submatrix $A_k(1:k+1, 1:k+1)$ is m-banded, so at the final step we get an $n \times n$ m-banded matrix similar to A_m.

With Algorithm 1, we have constructed a 2000×2000 Hermitian matrix H which is 20-banded and such that $\sigma(H) \subset [-1, -0.3] \cup [0.3, 1]$. In Figure 3.1 the exact decay is compared with the bounds. We can see that the decay rate seems to be captured by all the bounds, although $B_3(k) < B_2(k) < B_1(k)$ for almost all the values of k.

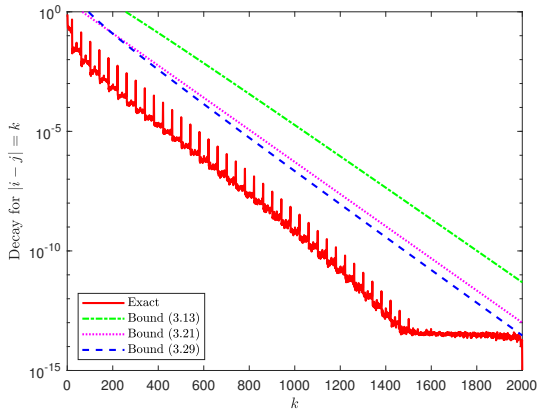

Figure 3.1: Logarithmic plot of the bounds (3.13), (3.21), and (3.29) compared with the exact decay for the spectral projector associated with the negative eigenvalues of a 20-banded, 2000×2000 Hermitian matrix with uniformly distributed eigenvalues in $[-1, -0.3] \cup [0.3, 1]$.

The bounds can be used to truncate the projector to a banded matrix with a small error. Let us see how the bounds behave in order to truncate P to $P^{(m)}$ where m is such that $|[P]_{ij}| < \epsilon$ for $|i-j| > m$, for a fixed threshold ϵ. Let us define

$$m_i(\epsilon) = \min\{\bar{k} : B_i(k) \le \epsilon \text{ for all } k \ge \bar{k}\}, \quad i = 1, 2, 3,$$
$$m_P(\epsilon) = \min\{\bar{k} : D_P(k) \le \epsilon \text{ for all } k \ge \bar{k}\}.$$

The value $m_i(\epsilon)$ is the first for which the bound B_i becomes definitively smaller than a threshold ϵ, and $m_P(\epsilon)$ does the same with D_p. The values of $m_i(\epsilon)$, $i = 1, 2, 3$, and $m_P(\epsilon)$ associated with the previous example are displayed in Table 3.1. We note that $m_3(\epsilon)$ provides the best estimate in all cases. Once again, it should be emphasized that for a given accuracy ϵ, the estimated bandwidth is independent of n. We can also notice that the exact decay is oscillatory with period equal to the original bandwidth m. This is reflected in the fact that $m_P(\epsilon)$ is always a multiple of m.

	$\epsilon = 1e-1$	$\epsilon = 1e-2$	$\epsilon = 1e-3$	$\epsilon = 1e-4$	$\epsilon = 1e-5$
$m_1(\epsilon)$	419	577	733	887	1041
$m_2(\epsilon)$	270	419	568	717	865
$m_3(\epsilon)$	218	347	483	623	764
$m_P(\epsilon)$	60	180	300	420	540

Table 3.1: Estimated bandwidth needed to achieve a prescribed error in the approximate spectral projector using different bounds. In the last row of the table, we report the actual bandwidth needed to achieve the prescribed error levels.

3.3.6 Nonsymmetric Spectrum

Our approach strongly relies on the fact that $\sigma(H)$ is contained in the union of two symmetric intervals. Although this hypothesis is always satisfied by choosing suitable values of a and b, the bound does not behave like the real decay when b (or $-b$) is not close to the maximum (resp., minimum) eigenvalue. If $\sigma(H) \subset [-b_1, -a] \cup [a, b_2]$ with $b_1 \neq b_2$, it would be preferable to consider this domain instead of $[-b, -a] \cup [a, b]$ with $b = \max\{b_1, b_2\}$.

It is shown in [39, 49] that there exist positive constants C_1, C_2 and η such that

$$C_1 k^{-\frac{1}{2}} e^{-\eta k} \leq E_k(\mathrm{sign}(x), [-b_1, -a] \cup [a, b_2]) \leq C_2 k^{-\frac{1}{2}} e^{-\eta k}. \tag{3.30}$$

The rate η is computable and given by

$$\eta = \int_{-1}^{K} \frac{K - x}{\sqrt{(1 - x^2)(x + b_1/a)(x - b_2/a)}} \, \mathrm{d}x,$$

where

$$K = \frac{\int_{-1}^{1} x((1 - x^2)(x + b_1/a)(x - b_2/a))^{-1/2} \, \mathrm{d}x}{\int_{-1}^{1} ((1 - x^2)(x + b_1/a)(x - b_2/a))^{-1/2} \, \mathrm{d}x}.$$

However, the arguments in [49] do not give the values of C_1 and C_2, so they are unknown.

As an example, we constructed a 300×300, tridiagonal, Hermitian matrix H such that the spectrum is uniformly distributed over $[-0.5, -0.1] \cup [0.1, 1]$. In Figure 3.2 the real decay is compared with the asymptotic rate in (3.30) and the rate in (3.22), which is asymptotically equivalent to the bound (3.29). All the constant factors are set to 1 to compare only the asymptotic behavior.

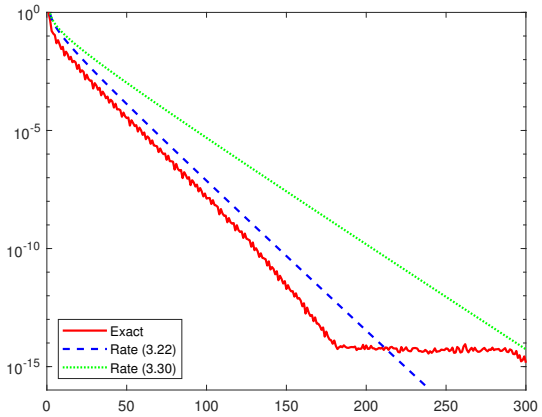

Figure 3.2: Logarithmic plot of the decay rates (3.30) and (3.22) compared with the exact decay of the spectral projector associated with the negative eigenvalues of a 300×300 tridiagonal matrix with spectrum in $[-0.5, -0.1] \cup [0.1, 1]$.

3.3.7 Bounds for the Fermi-Dirac Function

Using the same techniques introduced in this section, we can derive new decay bounds for the Fermi-Dirac matrix function.

Let us recall a previous result based on Bernstein's Theorem 3.6.

Theorem 3.20 (Theorem 8.1 in [14]). *Let H be Hermitian and m-banded with $\sigma(H) \subset [-1, 1]$. Let $f_{FD}(x) = (1 + e^{\beta(x-\mu)})^{-1}$ be the Fermi-Dirac function, where $\beta > 0$ and $\mu \in [-1, 1]$. Then*

$$|[f_{FD}(H)]_{ij}| \leq \frac{2\chi M(\chi)}{\chi - 1} \left(\frac{1}{\chi}\right)^{\frac{|i-j|}{m}}, \quad M(\chi) = \max_{z \in \mathcal{E}_\chi} |f_{FD}(z)|, \tag{3.31}$$

for any $1 < \chi < \overline{\chi}$, where

$$\overline{\chi} = \frac{\sqrt{\sqrt{(\beta^2(1-\mu^2) - \pi^2)^2 + 4\pi^2\beta^2} - \beta^2(1-\mu^2) + \pi^2}}{\sqrt{2}\beta} + \frac{\sqrt{\sqrt{(\beta^2(1-\mu^2) - \pi^2)^2 + 4\pi^2\beta^2} + \beta^2(1+\mu^2) + \pi^2}}{\sqrt{2}\beta},$$

and $\mathcal{E}_\chi$ is the unique ellipse with foci in -1 and 1, with semiaxes $\kappa_1 > 1$ and $\kappa_2 >$, and $\chi = \kappa_1 + \kappa_2$.

Remark 3.11. The hypothesis $\sigma(H) \subset [-1, 1]$ is not restrictive. In fact, if $\sigma(H) \subset [a, b]$, then

$$(I + \exp(\beta(H - \mu)))^{-1} = (I + \exp(\tilde{\beta}(\tilde{H} - \tilde{\mu})))^{-1},$$

where $\tilde{H} = \frac{2}{b-a}H - \frac{b+a}{b-a}I$, $\tilde{\mu} = \frac{2}{b-a}\mu - \frac{b+a}{b-a}$, $\tilde{\beta} = \frac{b-a}{2}\beta$ and $\sigma(\tilde{H}) \subset [-1, 1]$.

The parameters of Theorem 3.20 simplify when $\mu = 0$.

Corollary 3.21. *Let H be m-banded with $\sigma(H) \subset [-1, 1]$. Let $f_{FD}(x) = (1 + e^{\beta x})^{-1}$ with $\beta > 0$. Then* (3.31) *holds with $\overline{\chi}$ and $M(\chi)$ given by*

$$\overline{\chi} = \frac{\pi + \sqrt{\beta^2 + \pi^2}}{\beta}, \quad M(\chi) = \left| 1/(1 + e^{\beta z^*}) \right|, \quad z^* = \mathbf{i}\frac{\chi^2 - 1}{2\chi}. \tag{3.32}$$

In the same spirit of the previous sections, we can give a unified bound for the entries of $f_{FD}(H)$ in the case where $\mu = 0$ and $\sigma(H)$ is contained in the union of two symmetric intervals. We consider the identity

$$f_{FD}(x) = \frac{1}{1 + \mathrm{e}^{\beta x}} = \frac{1}{2}\left(1 - \frac{\mathrm{e}^{\frac{\beta}{2}x} - \mathrm{e}^{-\frac{\beta}{2}x}}{\mathrm{e}^{\frac{\beta}{2}x} + \mathrm{e}^{-\frac{\beta}{2}x}}\right) = \frac{1}{2}\left(1 - \tanh\left(\frac{\beta}{2}x\right)\right), \tag{3.33}$$

so here $\tanh\left(\frac{\beta}{2}x\right)$ plays with $f_{FD}(x)$ the same role as $\operatorname{sign}(x)$ with the Heaviside function (3.11). We have the following relation between hyperbolic and trigonometric tangent:

$$\tanh(z) = -\mathbf{i}\tan(\mathbf{i}z). \tag{3.34}$$

Moreover, $\tan(z)$ has the pole expansion [4, Example 11.7.3]

$$\tan(z) = \sum_{s=0}^{\infty} \frac{2z}{z_s^2 - z^2}, \quad z_s = \frac{2s+1}{2}\pi. \tag{3.35}$$

Combining (3.34) with (3.35) gives us

$$\tanh(z) = \sum_{s=0}^{\infty} \frac{2z}{z_s^2 + z^2}. \tag{3.36}$$

Due to the relation (3.33), we can reduce the problem to that of bounding the entries of the hyperbolic tangent matrix function.

Theorem 3.22. *Let $H \in \mathbb{C}^{n\times n}$ be Hermitian and m-banded with $\sigma(A) \subset [-b, -a] \cup [a, b]$. Let*

$$\tilde{K} = \begin{cases} \frac{b}{2}\left[\left(\frac{\tanh(b)}{b}\right)^{\frac{1}{2}} + \left(\frac{\tanh(a)}{a}\right)^{\frac{1}{2}}\right]^2 & \text{if } a > 0, \\ \frac{b}{2}\left[\left(\frac{\tanh(b)}{b}\right)^{\frac{1}{2}} + 1\right]^2 & \text{if } a = 0, \end{cases}$$

and

$$\tilde{\zeta} = \frac{\sqrt{b^2 + \pi^2/4} - \sqrt{a^2 + \pi^2/4}}{\sqrt{b^2 + \pi^2/4} + \sqrt{a^2 + \pi^2/4}} = \frac{b^2 - a^2}{(\sqrt{b^2 + \pi^2/4} + \sqrt{a^2 + \pi^2/4})^2}.$$

Then

$$|[\tanh(A)]_{ij}| \le \tilde{K}\tilde{\zeta}^{\frac{|i-j|}{2m} - \frac{1}{2}},$$

for $|i - j| \ge m$.

Proof. The sum in (3.36) can be interpreted as the integral of a piecewise continuous function. Hence, in view of Lemma 3.9 and Remark 3.5 and applying Lemma 3.12, for $k \geq 0$ we get

$$\begin{aligned} E_k(\tanh(x), [-b,-a]\cup[a,b]) &\leq \sum_{s=0}^{\infty} E_k(2x/(z_s^2+x^2), [-b,-a]\cup[a,b]) \\ &\leq 2b\sum_{s=0}^{\infty} K(z_s)\zeta(z_s)^{\frac{k}{2}}. \\ &\leq \left(2b\sum_{s=0}^{\infty} K(z_s)\right)\zeta(z_0)^{\frac{k}{2}}, \end{aligned}$$

where $r(t), K(t), \zeta(t)$ are as in (3.15). Proceeding similarly as in Theorem 3.14 we expand $K(z_s)$ as follows:

$$\begin{aligned} &\sum_{s=0}^{\infty} 2\,K(z_s) = \sum_{s=0}^{\infty} \frac{(1+\sqrt{r(z_s)})^2}{b^2+z_s^2} \\ &= \sum_{s=0}^{\infty} \frac{1}{b^2+z_s^2} + \sum_{s=0}^{\infty} \frac{1}{a^2+z_s^2} + 2\sum_{s=0}^{\infty} \frac{1}{\sqrt{(b^2+z_s^2)(a^2+z_s^2)}}. \end{aligned}$$

Consider the case $a > 0$. In view of (3.36),

$$\sum_{s=0}^{\infty} \frac{1}{b^2+z_s^2} = \frac{\tanh(b)}{2b}, \quad \sum_{s=0}^{\infty} \frac{1}{a^2+z_s^2} = \frac{\tanh(a)}{2a}, \tag{3.37}$$

while the third can be bounded by using the Cauchy-Schwarz inequality:

$$\begin{aligned} \sum_{s=0}^{\infty} \frac{1}{\sqrt{(b^2+z_s^2)(a^2+z_s^2)}} &\leq \left(\sum_{s=0}^{\infty} \frac{1}{b^2+z_s^2}\right)^{\frac{1}{2}} \left(\sum_{s=0}^{\infty} \frac{1}{a^2+z_s^2}\right)^{\frac{1}{2}} \\ &= \left(\frac{\tanh(b)}{2b}\,\frac{\tanh(a)}{2a}\right)^{\frac{1}{2}}. \end{aligned}$$

Finally,

$$\begin{aligned} \sum_{s=0}^{\infty} K(z_s) &\leq \frac{\tanh(b)}{2b} + \frac{\tanh(a)}{2a} + 2\left(\frac{\tanh(b)}{2b}\,\frac{\tanh(a)}{2a}\right)^{\frac{1}{2}} \\ &= \frac{1}{2}\left[\left(\frac{\tanh(b)}{b}\right)^{\frac{1}{2}} + \left(\frac{\tanh(a)}{a}\right)^{\frac{1}{2}}\right]^2. \end{aligned}$$

If $a = 0$, observe that the second identity in (3.37) becomes

$$\sum_{s=0}^{\infty} \frac{1}{z_s^2} = \sum_{s=0}^{\infty} \lim_{a\to 0} \frac{1}{a^2+z_s^2} = \lim_{a\to 0} \sum_{s=0}^{\infty} \frac{1}{a^2+z_s^2} = \lim_{a\to 0} \frac{\tanh(a)}{2a} = \frac{1}{2},$$

where we have used the Monotone Convergence Theorem [91, Theorem 1.26]. The rest of the proof proceeds as before. □

We can use Theorem 3.22 to obtain a new bound for the entries of the Fermi-Dirac function.

Corollary 3.23. *Let $f_{FD}(x) = (1 + e^{\beta x})^{-1}$, and let H be Hermitian and m-banded with $\sigma(H) \subset [-b, -a] \cup [a, b]$. Let*

$$\tilde{\zeta} = \frac{\beta^2(b^2 - a^2)}{(\sqrt{\beta^2 b^2 + \pi^2} + \sqrt{\beta^2 a^2 + \pi^2})^2},$$

and

$$\tilde{K} = \frac{b}{4}\left[\left(\frac{\tanh\left(\frac{\beta}{2}b\right)}{b}\right)^{\frac{1}{2}} + \left(\frac{\tanh\left(\frac{\beta}{2}a\right)}{a}\right)^{\frac{1}{2}}\right]^2 \quad \text{if } a > 0,$$

$$\tilde{K} = \frac{b}{4}\left[\left(\frac{\tanh\left(\frac{\beta}{2}b\right)}{b}\right)^{\frac{1}{2}} + \left(\frac{\beta}{2}\right)^{\frac{1}{2}}\right]^2 \quad \text{if } a = 0.$$

Then

$$|[f_{FD}(H)]_{ij}| \leq \tilde{K}\tilde{\zeta}^{\frac{|i-j|}{2m} - \frac{1}{2}}, \tag{3.38}$$

for all i, j such that $|i - j| \geq m$.

Proof. Since $|[f_{FD}(H)]_{ij}| = -\frac{1}{2}|[\tanh(\frac{\beta}{2}H)]$ for all $i \neq j$, we can apply Theorem 3.22 by observing that $\sigma\left(\frac{\beta}{2}H\right) \subset \left[-\frac{\beta}{2}b, -\frac{\beta}{2}a\right] \cup \left[\frac{\beta}{2}a, \frac{\beta}{2}b\right]$. □

Remark 3.12. Let us compare this result with the previous bound given by Corollary 3.21. In (3.31) the rate is $(1/\chi)^{\frac{|i-j|}{m}}$, where $1 < \chi < \overline{\chi}$ and $\overline{\chi}$ is given in (3.32), while in (3.38) the rate is $\tilde{\zeta}^{\frac{|i-j|}{2m}}$. When $a = 0$ and $b = 1$, so the only assumption on the spectrum of H is $\sigma(H) \subset [-1, 1]$, we have that $\tilde{\zeta}^{\frac{1}{2}} = 1/\overline{\chi}$, hence $\tilde{\zeta}^{\frac{|i-j|}{2m}} = (1/\overline{\chi})^{\frac{|i-j|}{m}} < (1/\chi)^{\frac{|i-j|}{m}}$ for $\chi < \overline{\chi}$, so the geometric rate in (3.38) decays faster than all the rates in (3.31). Moreover, in Corollary 3.21 no attention is paid to a possible gap around μ in the spectrum of H. Hence, when such a gap is present, the bound in (3.38) can be much smaller.

Remark 3.13. Theorem 3.22 and Corollary 3.23 optimize the spectral information only if $\sigma(H)$ is symmetric with respect to μ. The case $\mu \neq 0$ can be handled by considering $\tilde{H} = H - \mu I$ so that $(I + \exp(\beta(H - \mu I))^{-1} = (I + \exp(\beta\tilde{H}))^{-1}$. However, if $\sigma(\tilde{H})$ is not symmetric, we expect the bounds to be pessimistic, so the same considerations of Section 3.3.6 hold also in this case. This problem does not hold for Theorem 3.20, in which, however, no benefits are obtained from a possible gap around μ. Specific bounds for the nonsymmetric, gapped case are yet to be found.

Example 3.14. As a numerical example, we constructed two 100×100 tridiagonal matrices with uniformly distributed spectrum over $[-1,-a] \cup [a,1]$, $a = 0, 0.3$. We then computed $f_{FD}(H) = (I + \exp(10H))^{-1}$, so $\beta = 10$ and $\mu = 0$. We compared the off-diagonal decay

$$\max_{|i-j|=k} |[f_{FD}(H)]_{ij}|$$

with the bounds (3.38) and (3.31) optimized in χ. The results are shown in Figure 3.3. When $a = 0$, (3.38) yields a small refinement, while, for $a = 0.3$, the improvement is notable.

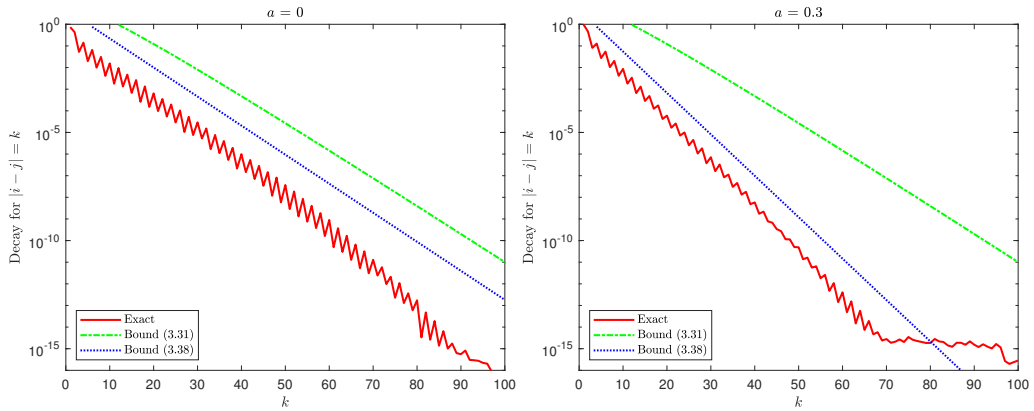

Figure 3.3: Logarithmic plot of the exact off-diagonal decay of $f_{FD}(H) = (I + e^{10H})^{-1}$ compared with the bounds (3.38) and (3.31) optimized in χ, where H is symmetric, tridiagonal with $\sigma(H) \subset [-1,-a] \cup [a,1]$. Left: $a = 0$. Right: $a = 0.3$.

3.4 Bounds Related to the Eigenvalue Distribution

In this section we will give bounds for the entries of some matrix functions which take account of more spectral information than the previous techniques.

3.4.1 Inverse Function

The result of Theorem 3.24 can be refined by directly working on the best polynomial approximation.

Theorem 3.24. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and m-banded with distinct eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_\nu$, with $\nu \leq n$. For $\ell < \nu$ define $\sigma_\ell(A) := \{\lambda_1, \ldots, \lambda_{\nu-\ell}\}$ and*

$$r_\ell = \frac{\lambda_{\nu-\ell}}{\lambda_1}, \quad \hat{\zeta}_\ell = \frac{\sqrt{r_\ell} - 1}{\sqrt{r_\ell} + 1}, \quad K_\ell = \frac{\left(1 + \sqrt{r_\ell}\right)^2}{2\lambda_{\nu-\ell}}. \tag{3.39}$$

Then

$$E_k(1/x, \sigma(A)) \leq E_{k-\ell}(1/x, \sigma_\ell(A)) \leq K_\ell \hat{\zeta}_\ell^{k+1-\ell}, \tag{3.40}$$

for all $k \geq 0$ *and* $\ell = 0, 1, \ldots, k$. *Moreover, we have that*

$$|[A^{-1}]_{ij}| \leq K_\ell \hat{\zeta}_\ell^{\frac{|i-j|}{m}-\ell}, \tag{3.41}$$

for $|i-j| > m$ *and* $\ell = 0, 1, \ldots, \left\lfloor \frac{|i-j|}{m} \right\rfloor$.

Proof. Let $p_{k-\ell}$ be a polynomial of degree $k-\ell$. Define

$$R_\ell(x) = \prod_{i=\nu-\ell+1}^{\nu} \left(1 - \frac{x}{\lambda_i}\right)$$

and let

$$\overline{p}_k(x) = \frac{1}{x}(1 - R_\ell(x)) + R_\ell(x)\, p_{k-\ell}(x). \tag{3.42}$$

Since $R_\ell(0) = 1$, we have that $1 - R_\ell(x)$ is a multiple of x so the first term in (3.42) is a polynomial of degree $\ell - 1$. Then $\overline{p}_k(x)$ has degree k. From the identity

$$\frac{1}{x} - \overline{p}_k(x) = R_\ell(x) \left(\frac{1}{x} - p_{k-\ell}(x)\right),$$

and by using that $R_\ell(\lambda_i) = 0$ for $i = \nu - \ell + 1, \ldots, n$, and $|R_\ell(x)| \leq 1$ for any $x \in \sigma_\ell(A)$, we have

$$\begin{aligned} E_k(1/x, \sigma(A)) \leq \max_{x \in \sigma(A)} \left|\frac{1}{x} - \overline{p}_k(x)\right| &= \max_{x \in \sigma(A)} \left[|R_\ell(x)| \cdot \left|\frac{1}{x} - p_{k-\ell}(x)\right|\right] \\ &\leq \max_{x \in \sigma_\ell(A)} \left|\frac{1}{x} - p_{k-\ell}(x)\right|. \end{aligned}$$

The inequality holds for any $p_{k-\ell}(x) \in \Pi_{k-\ell}$ and by using Theorem 3.3, we have that

$$\begin{aligned} E_k(1/x, \sigma(H)) \leq E_k(1/x, \sigma_\ell(A)) &\leq E_k(1/x, [a, \lambda_{\nu-\ell}]) \\ &= K_\ell \hat{\zeta}_\ell^{k+1-\ell}, \end{aligned}$$

so (3.40) holds. Finally, for (3.41) it is sufficient to apply Proposition 3.1. □

Remark 3.15. With Theorem 3.24 we have refined the result of Theorem 3.5, since we can now exclude a multiple eigenvalue by increasing ℓ by 1, while by using Theorem 3.5 we would need to increase ℓ by the algebraic multiplicity of that eigenvalue. We also slightly decreased the constant factor, since $K_\ell \leq 2/\lambda_1$ for any ℓ. This is not a big improvement since K_ℓ increases with ℓ and the values of these two constants do not differ much in practice.

3.4.2 Cauchy-Stieltjes Functions

Theorem 3.24 extends to Cauchy-Stieltjes functions through Lemma 3.9.

Theorem 3.25. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and m-banded with distinct eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_\nu$, with $\nu \leq n$. Define $\sigma_\ell(A+tI) := \{\lambda_1 + t, \ldots, \lambda_{\nu-\ell} + t\}$ for $\ell < \nu$ and $t \geq 0$, and let $\hat{\zeta}_\ell$ be as in* (3.39)*. Consider a Cauchy-Stieltjes function of the form*

$$f(x) = \int_0^\infty \frac{h(t)}{x+t}\,\mathrm{d}t, \tag{3.43}$$

with $h(t) \geq 0$ for all t. Then

$$E_k(f(x), \sigma(A)) \leq \frac{1}{2}\left(\sqrt{f(\lambda_1)} + \sqrt{f(\lambda_{\nu-\ell})}\right)^2 \hat{\zeta}_\ell^{k+1-\ell}, \tag{3.44}$$

for all $k \geq 0$ and $\ell = 0, \ldots, k$. Moreover, we have that

$$|[f(A)]_{ij}| \leq \frac{1}{2}\left(\sqrt{f(\lambda_1)} + \sqrt{f(\lambda_{\nu-\ell})}\right)^2 \hat{\zeta}_\ell^{\frac{|i-j|}{m}-\ell}, \tag{3.45}$$

for $|i-j| \geq m$ and $\ell = 0, 1, \ldots, \left\lfloor \frac{|i-j|}{m} \right\rfloor$.

Proof. Let $k \geq 0$. By applying Lemma 3.9, we get

$$E_k(f, \sigma(A)) \leq \int_0^\infty h(t) E_k(1/(x+t), \sigma(A))\,\mathrm{d}t = \int_0^\infty h(t) E_k(1/x, \sigma(A+tI))\,\mathrm{d}t.$$

Let $\ell = 0, \ldots, \lfloor k \rfloor$ be fixed. In view of Theorem 3.24, by defining

$$r_\ell(t) = \frac{\lambda_{\nu-\ell}+t}{\lambda_1+t}, \quad \zeta_\ell(t) = \frac{\sqrt{r_\ell(t)}-1}{\sqrt{r_\ell(t)}+1}, \quad K_\ell(t) = \frac{\left(1+\sqrt{r_\ell(t)}\right)^2}{2(\lambda_{\nu-\ell}+t)},$$

we get

$$E_k(1/x, \sigma(A+tI)) \leq K_\ell(t)\zeta_\ell(t)^{k+1-\ell} \leq K_\ell(t)\hat{\zeta}_\ell,$$

Since $k+1-\ell \geq 0$ and $\zeta_\ell(t) \leq \zeta_\ell(0) = \hat{\zeta}_\ell$. Hence,

$$E_k(f, \sigma(A)) \leq \hat{\zeta}_\ell \int_0^\infty h(t) K_\ell(t)\,\mathrm{d}t. \tag{3.46}$$

In order to estimate the integral, we split it as follows:

$$\begin{aligned}
&\int_0^\infty h(t)K_\ell(t)\,\mathrm{d}t = \int_0^\infty h(t)\frac{(1+\sqrt{r(t)})^2}{2(\lambda_{\nu-\ell}+t)}\,\mathrm{d}t \\
&= \frac{1}{2}\left(\int_0^\infty \frac{h(t)}{\lambda_1+t}\,\mathrm{d}t + \int_0^\infty \frac{h(t)}{\lambda_{\nu-\ell}+t}\,\mathrm{d}t + 2\int_0^\infty \frac{h(t)}{\sqrt{(\lambda_1+t)(\lambda_{\nu-\ell}+t)}}\,\mathrm{d}t\right).
\end{aligned}$$

The first two terms are equal to $f(\lambda_1)$ and $f(\lambda_{\nu-\ell})$, respectively, in view of (3.43). The third can be bounded by using the Cauchy-Schwartz inequality:

$$\begin{aligned}\int_0^\infty \frac{h(t)}{\sqrt{(\lambda_1+t)(\lambda_{\nu-\ell}+t)}}\,\mathrm{d}t &\le \left(\int_0^\infty \frac{h(t)}{\lambda_1+t}\,\mathrm{d}t\right)^{\frac{1}{2}} \cdot \left(\int_0^\infty \frac{h(t)}{\lambda_{\nu-\ell}+t}\,\mathrm{d}t\right)^{\frac{1}{2}} \\ &= \sqrt{f(\lambda_1)f(\lambda_{\nu-\ell})}.\end{aligned}$$

Finally,

$$\begin{aligned}\int_0^\infty h(t)K_\ell(t)\,\mathrm{d}t &\le \frac{1}{2}(f(\lambda_1)+f(\lambda_{\nu-\ell})+2\sqrt{f(\lambda_1)f(\lambda_{\nu-\ell})}) \\ &= \frac{1}{2}(\sqrt{f(\lambda_1)}+\sqrt{f(\lambda_{\nu-\ell})})^2.\end{aligned}$$

In view of this and (3.46), we get (3.44). The bound (3.45) follows from (3.44) and Proposition 3.1. □

3.4.3 Spectral Projector and Sign Function

As in Section 3.3, we analyze the polynomial approximations of the sign function by exploiting the integral representation

$$\operatorname{sign}(x) = \frac{2}{\pi}\int_0^\infty \frac{x}{x^2+t^2}\,\mathrm{d}t,$$

so we can reduce the problem to analyzing $x(x^2+t^2)^{-1}$ for $t \ge 0$. Our goal is to extend the results valid for the inverse function concerning the decay with respect to the effective condition number r_ℓ to the case of the spectral projector and the matrix sign.

Lemma 3.26. *Let $f_t(x) := x(x^2+t^2)^{-1}$ for $t \ge 0$. Let $H \in \mathbb{C}^{n\times n}$ be a Hermitian nonsingular matrix. Let $a = \mu_1 < \mu_2 < \cdots < \mu_\nu = b$, with $\nu \le n$, be the distinct values of $|\lambda|$ for $\lambda \in \sigma(H)$, and let $b_\ell = \mu_{\nu-\ell}$. For any $t \ge 0$, let*

$$r_\ell(t) = \frac{b_\ell^2+t^2}{a^2+t^2}, \quad K_\ell(t) = \frac{(1+\sqrt{r_\ell(t)})^2}{2(b_\ell^2+t^2)}, \quad \zeta_\ell(t) = \frac{\sqrt{r_\ell(t)}-1}{\sqrt{r_\ell(t)}+1}. \tag{3.47}$$

Then

$$E_k(f_t(x), \sigma(H)) \le b_\ell\, K_\ell(t)\, \zeta_\ell(t)^{\frac{k}{2}-\ell},$$

for any $\ell = 0, 1, \ldots, \lfloor \frac{k}{2} \rfloor$.

Proof. From the definition of μ_i, we have $\sigma(H^2+t^2I) = \{\mu_1^2+t^2, \ldots, \mu_\nu^2+t^2\}$. We denote $\sigma_\ell(H^2+t^2I) = \{\mu_1^2+t^2, \ldots, \mu_{\nu-\ell}^2+t^2\}$. Consider the function $1/x$ defined over $\sigma(H^2+t^2I)$. By proceeding as in Theorem 3.24, we can construct $\overline{p}_k(x) \in \Pi_k$ such that

$$\frac{1}{x} - \overline{p}_k(x) = R_\ell(x)\left(\frac{1}{x} - p_{k-\ell}(x)\right), \tag{3.48}$$

where $R_\ell(x) \in \Pi_\ell$ satisfies $|R_\ell(\mu_i^2+t^2)| < 1$ for $i = 1, \dots, \nu-\ell$ and $R_\ell(\mu_i^2+t^2) = 0$ for $i = \nu-\ell+1, \dots, \nu$, and $p_{k-\ell}(x) \in \Pi_{k-\ell}$ is the polynomial of best uniform approximation for $1/x$ over the interval $[\mu_1^2+t^2, \mu_{\nu-\ell}^2+t^2]$. Then

$$\begin{aligned}\max_{x\in\sigma(H^2+t^2I)} \left|\frac{1}{x} - \overline{p}_k(x)\right| &\le \max_{x\in\sigma_\ell(H^2+t^2I)} \left|\frac{1}{x} - p_{k-\ell}(x)\right| \\ &\le K_\ell(t)\,\zeta_\ell(t)^{k+1-\ell}.\end{aligned}$$

In order to approximate $f_t(x)$ over $\sigma(H)$, consider $S_{2k+1}(x) := x\,\overline{p}_k(x^2+t^2) \in \Pi_{2k+1}$. In view of (3.48), we have

$$\begin{aligned}f_t(x) - S_{2k+1}(x) &= x\left(\frac{1}{x^2+t^2} - \overline{p}_k(x^2+t^2)\right) \\ &= xR_\ell(x^2+t^2)\left(\frac{1}{x^2+t^2} - p_{k-\ell}(x^2+t^2)\right).\end{aligned}$$

Therefore,

$$\begin{aligned}E_{2k+1}(f_t, \sigma(H)) &\le \max_{x\in\sigma(H)} |f_t(x) - S_{2k+1}(x)| \\ &= \max_{x\in\sigma(H)} \left(|x|\cdot|R_\ell(x^2+t^2)|\cdot\left|\frac{1}{x^2+t^2} - p_{k-\ell}(x^2+t^2)\right|\right) \\ &\le \max_{x\in\{\mu_1,\dots,\mu_{\nu-\ell}\}} \left(|x|\cdot\left|\frac{1}{x^2+t^2} - p_{k-\ell}(x^2+t^2)\right|\right) \\ &\le b_\ell \cdot \max_{x\in\{\mu_1,\dots,\mu_{\nu-\ell}\}} \left|\frac{1}{x^2+t^2} - p_{k-\ell}(x^2+t^2)\right| \\ &\le b_\ell\, K_\ell(t)\,\zeta_\ell(t)^{k+1-\ell}.\end{aligned}$$

By proceeding as in Lemma 3.12, we obtain that

$$E_k(f_t(x), \sigma(H)) \le b_\ell\, K_\ell(t)\,\zeta_\ell(t)^{\frac{k}{2}-\ell}.$$

This concludes the proof. □

Now we can state the analogue of Theorem 3.14.

Theorem 3.27. *Let H be Hermitian with $\sigma(H) \subset [-b,-a] \cup [a,b]$. Let $a = \mu_1 < \mu_2 < \cdots < \mu_\nu = b$, with $\nu \le n$, be the distinct values of $|\lambda|$ for $\lambda \in \sigma(H)$, and let $b_\ell = \mu_{\nu-\ell}$. Let*

$$\hat{K}_\ell := \frac{1}{4}\left(1 + \sqrt{\frac{b_\ell}{a}}\right)^2, \quad \hat{\zeta}_\ell := \frac{b_\ell - a}{b_\ell + a}.$$

Then, for all $k \ge 0$ and $\ell = 0, \dots, \lfloor \frac{k}{2} \rfloor$,

$$E_k(\mathrm{sign}(x), \sigma(H)) \le 2b_\ell \hat{K}_\ell \hat{\zeta}_\ell^{\frac{k}{2}-\ell}. \tag{3.49}$$

Moreover, if H is m-banded and $\ell = 0, \dots, \left\lfloor \frac{|i-j|}{2m} - \frac{1}{2} \right\rfloor$,

$$|[\operatorname{sign}(H)]_{ij}| \leq 2b_\ell \hat{K}_\ell \hat{\zeta}_\ell^{\frac{|i-j|}{2m} - \frac{1}{2} - \ell} \quad \textit{for } |i-j| \geq m,$$
$$|[P]_{ij}| \leq b_\ell \hat{K}_\ell \hat{\zeta}_\ell^{\frac{|i-j|}{2m} - \frac{1}{2} - \ell} \quad \textit{for all } i, j. \tag{3.50}$$

Proof. Let k be fixed, and let $\ell = 0, \dots, \left\lfloor \frac{k}{2} \right\rfloor$. By applying (3.16) and Lemma 3.26, we have

$$\begin{aligned} E_k(\operatorname{sign}(x), \sigma(H)) &\leq \frac{2}{\pi} \int_0^\infty b_\ell \, K_\ell(t) \zeta_\ell(t)^{\frac{k}{2} - \ell} \, \mathrm{d}t \\ &\leq \left(\frac{2}{\pi} \int_0^\infty b_\ell \, K_\ell(t) \, \mathrm{d}t \right) \zeta_\ell(0)^{\frac{k}{2} - \ell}, \end{aligned}$$

where $K_\ell(t), \zeta_\ell(t)$ are as in (3.47). By proceeding as in Theorem 3.13, we get (3.49). The other bounds follow by applying Proposition 3.1 and Proposition 3.2 together with (3.49), also noting that $|P_{ij}| \leq 1$ for all i, j. □

Theorem 3.27 gives us a family of bounds parametrized by ℓ. Hence, for fixed i, j, the corresponding entry of the projector is bounded by

$$|[P]_{ij}| \leq \min_{\ell = 0, \dots, \left\lfloor \frac{|i-j|}{2m} - \frac{1}{2} \right\rfloor} \hat{C}_\ell \hat{q}_\ell^{\frac{|i-j|}{2m} - \frac{1}{2} - \ell}. \tag{3.51}$$

Depending on the eigenvalue distribution of H, this can predict a much faster decay than the results of Section 3.3. For instance, increasing ℓ gives a smaller geometric rate $\hat{q}_\ell$ but also a smaller exponent. If some of the eigenvalues that are largest in magnitude are isolated, $\hat{q}_\ell$ becomes much smaller even for moderate values of ℓ. In case of a cluster of eigenvalues near the spectral gap, we can also predict a superexponential decay. We will see some examples in the next section.

In all the results in this section, a special attention is given to the case where the absolute value $|\lambda|$ of an eigenvalue $\lambda \in \sigma(H)$ appears more than once. For instance, in electronic structure computations it is usual to have isolated eigenvalues with the largest absolute value that have high multiplicity; see [14, Section 8.2].

3.4.4 Numerical Experiments

Here we see how the bound (3.50) works on some examples. The matrices are generated with the method described in Section 3.3.5.

For the first example, we consider a 3000×3000, 20-banded matrix H for which -1 is an eigenvalue with multiplicity 10 and all the other eigenvalues are uniformly distributed over $[-0.5, -0.1] \cup [0.1, 0.5]$. In order to apply the results of Section 3.3 we must consider the inclusion $\sigma(H) \subset [-1, -0.1] \cup [0.1, 1]$. However, if we apply Theorem 3.27 with $\ell = 1$ we obtain $b_1 = 0.5$ that leads to a much more accurate bound, as we can see in Figure 3.4.

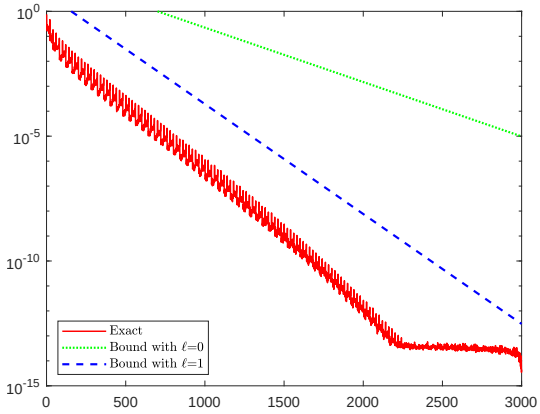

Figure 3.4: Logarithmic plot of the bounds given by Theorem 3.27 applied with $\ell = 0, 1$ compared with the exact decay of the spectral projector associated with the negative eigenvalues of a 20-banded, 3000×3000 Hermitian matrix with spectrum contained in $\{-1\} \cup [-0.5, -0.1] \cup [0.1, 0.5]$.

Now we show that Theorem 3.27 can predict a superexponential decay behavior if the eigenvalues are clustered near the spectral gap. We first consider the case where the spectrum is symmetric with respect to the origin, so, in the notation of Lemma 3.26, any μ_i corresponds to two eigenvalues, one positive and one negative. More precisely, we consider a 300×300, tridiagonal matrix H with eigenvalues

$$\lambda_i^{(j)} = (-1)^j \left[1 + 0.9\left(1 - \frac{i-1}{149} - 2\sqrt{1 - \frac{i-1}{149}}\right)\right] \in [-1, -0.1] \cup [0.1, 1],$$

for $i = 1, \ldots, 150$ and $j = 0, 1$. In the notation of Lemma 3.26 we have that $\nu = 150$ and $\mu_i = \lambda_i^{(0)} = |\lambda_i^{(1)}|$ for $i = 1, \ldots, 150$. So, increasing ℓ by 1 removes two eigenvalues of equal modulus.

In Figure 3.5 the decay of the spectral projector is compared with the bounds given by Theorem 3.27 for $\ell = 0, \ldots, 50$, and with a bound that is optimized among the values of ℓ. Due to the eigenvalue distribution, the geometric rate $\hat{q}_\ell$ decreases rapidly in ℓ. Hence, the optimized bound (3.51) describes a superexponential decay.

The situation is different if the eigenvalues are not symmetrically distributed. For instance, consider a 300×300, tridiagonal, Hermitian matrix H with eigenvalues

$$\lambda_i = (-1)^i \left[1 + 0.9\left(1 - \frac{i-1}{299} - 2\sqrt{1 - \frac{i-1}{299}}\right)\right] \in [-1, -0.1] \cup [0.1, 1],$$

for $i = 1, \ldots, 300$.

For this case, the comparison is shown in Figure 3.6. We can see that the optimized bound has a superexponential decay, but does not quite capture the exact behavior.

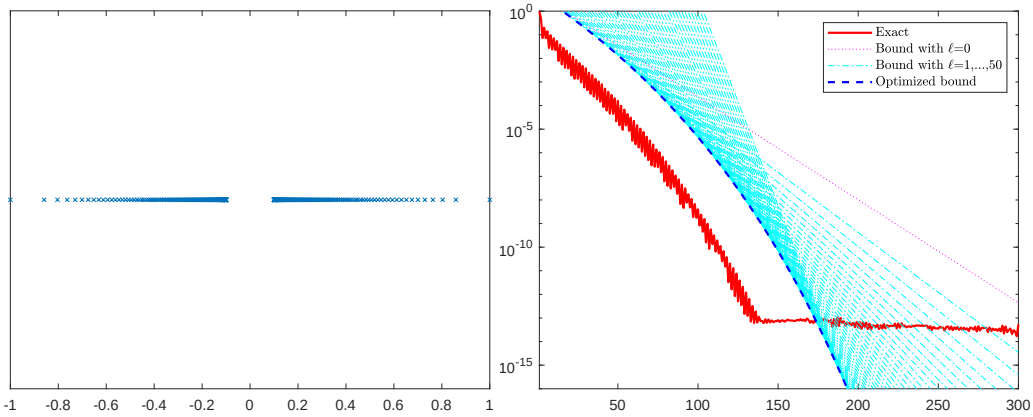

Figure 3.5: Left: Plot of the spectrum of H. Note that it is symmetric with respect to the origin. The eigenvalues tend to form a dense concentration near the spectral gap from the left and from the right, while they are increasely isolated tending to the extremes. Right: Exact decay of the entries of the projector compared with the bounds (3.50) for $\ell = 0, 1, \ldots, 50$. The dotted line, which corresponds to $\ell = 0$, is the bound (3.21). The dashed line corresponds to the best bound among the values of ℓ. We see that the decay behavior is captured by the optimized bound.

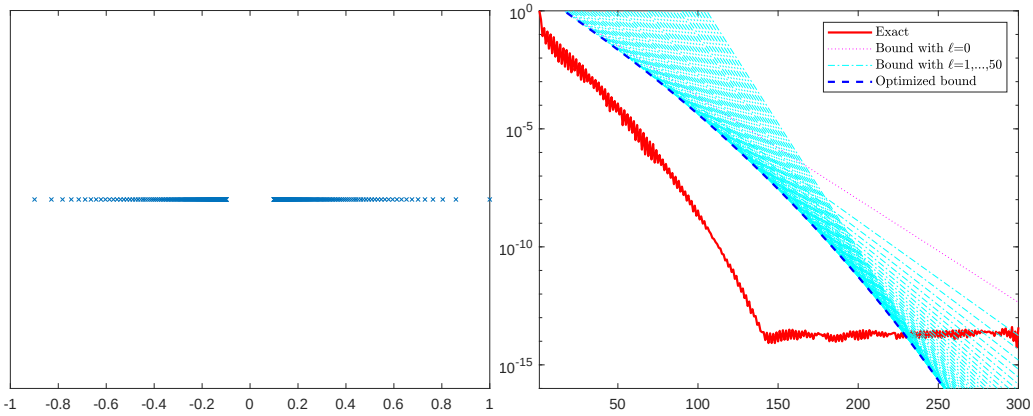

Figure 3.6: Left: Plot of the spectrum of H. In this case, no symmetry is present. Again, the eigenvalues tend to form a dense concentration near the spectral gap from the left and from the right, while they are increasely isolated tending to the extremes. Right: Exact decay of the entries of the projector compared with the bounds in (3.50) for $\ell = 0, 1, \ldots, 50$.

3.4.5 Fermi-Dirac Function

As discussed in Section 3.3.7, due to the relation $(1+e^{\beta x})^{-1} = \frac{1}{2}(1-\tanh(\frac{\beta}{2}x)$ and the pole expansion (3.36), in the case where $\mu = 0$ and the spectrum of H is contained in the union of two symmetric intervals, the Fermi-Dirac function and the spectral projector share many properties. We give the following result without proof, since it is essentially the same as Theorem 3.27 and follows by a combination of Theorem 3.22 and 3.26.

Theorem 3.28. *Let $f_{FD}(x) = (1+e^{\beta x})$, $\beta > 0$. Let H be Hermitian and m-banded with $\sigma(H) \subset [-b,-a] \cup [a,b]$. Let $a = \mu_1 < \mu_2 < \cdots < \mu_\nu = b$, with $\nu \leq n$, be the distinct values of $|\lambda|$ for $\lambda \in \sigma(H)$, and let $b_\ell = \mu_{\nu-\ell}$. Let, for $\ell \geq 0$,*

$$\hat{\zeta}_\ell = \frac{\beta^2(b_\ell^2 - a^2)}{(\sqrt{\beta^2 b_\ell^2 + \pi^2} + \sqrt{\beta^2 a^2 + \pi^2})^2},$$

and

$$\tilde{K}_\ell = \frac{b_\ell}{4}\left[\left(\frac{\tanh\left(\frac{\beta}{2}b_\ell\right)}{b_\ell}\right)^{\frac{1}{2}} + \left(\frac{\tanh\left(\frac{\beta}{2}a\right)}{a}\right)^{\frac{1}{2}}\right]^2 \quad \text{if } a > 0,$$

$$\tilde{K}_\ell = \frac{b_\ell}{4}\left[\left(\frac{\tanh\left(\frac{\beta}{2}b_\ell\right)}{b_\ell}\right)^{\frac{1}{2}} + \left(\frac{\beta}{2}\right)^{\frac{1}{2}}\right]^2 \quad \text{if } a = 0.$$

Then, for $|i-j| \geq m$ and $\ell = 0, \ldots, \left\lfloor \frac{|i-j|}{2m} - \frac{1}{2} \right\rfloor$,

$$|[f_{FD}(H)]_{ij}| \leq \hat{K}_\ell \hat{\zeta}_\ell^{\frac{|i-j|}{2m} - \frac{1}{2} - \ell} \quad \text{for all } i,j.$$

3.5 Conclusions and Further Developments

We have developed new bounds for the entries of spectral projectors, Cauchy-Stieltjes function and the Fermi-Dirac function, which improve and refine the existing ones under suitable hypotheses. Integral representations proved to be a powerful tool to describe well the decay rate. Manipulating such integrals can yield asymptotically optimal bounds in the sense of polynomial approximation, as shown in Theorem 3.17 for the sign function.

We have also shown that, like for the matrix inverse, the decay properties of many matrix functions is connected to the full spectral information. The situation here is similar to that arising in the analysis of the convergence of the conjugate gradient method for solving linear systems: this is not surprising, since in both cases we are dealing with polynomial approximations of matrix functions on the spectrum of a matrix. As a result, we are able to delete a few isolated eigenvalues from the spectral information and predict superexponential decay behavior in the presence of clusters.

Our results for spectral projectors and the Fermi-Dirac function can be improved by developing bounds without assuming that the spectrum is contained in the union of two symmetric intervals. In this sense, we can reduce to study functions of the form $x/(x^2 + t^2)$ in view of the representations (3.4.3) and (3.36).

Numerical results also show that the constant factors of the bounds remain pessimistic, especially when the condition number is large, so other techniques need to be explored as well. Moreover, as illustrated in Section 3.3.6, specific bounds for the projector are needed when the eigenvalues of the matrix argument are not symmetric with respect to the origin, in order to replicate the behavior captured by the purely asymptotic results.

Chapter 4

Estimating the Trace of Matrix Functions

In this chapter we consider methods for approximating the trace of functions of sparse symmetric matrices. We introduce and analyze the stochastic probing approach, based on a combination of probing [47] with the classical Hutchinson's estimator [68], and show both theoretically and experimentally that it can outperform existing techniques. In Section 4.1 we review some existing stochastic trace estimators, namely Hutchinson, Hutch++ and XTrace. In Section 4.2, we introduce the deterministic probing approach. In Section 4.3 we present a theoretical analysis of the stochastic probing estimation that reveals in particular for which matrix functions f and matrices A large gains with respect to the deterministic approach can not only be expected, but can actually be guaranteed. As a by-product of our analysis, we also refine classical results on sign patterns in the entries of $f(A)$. Section 4.4 includes a variety of numerical experiments that validate the theoretical results and highlight the performance of stochastic probing compared to the deterministic counterpart and the stochastic estimators introduced in Section 4.1. This chapter is based in part on [44].

4.1 Literature Review on Randomized Trace Estimators

In this section, we consider the task of approximating the trace of an implicitly given matrix $B \in \mathbb{R}^{n\times n}$ by means of stochastic estimators. We assume that B is accessible only via matrix-vector products $B\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{R}^n$. Recall that we are instered in the case $B = f(A)$, where $A \in \mathbb{R}^{n\times n}$ is large and sparse, hence the operations with B can be performed by applying a polynomial (or rational) Krylov subspace method; see Section 5.3. Here we review some literature about Hutchinson's estimator and its recent variants, Hutch++ and XTrace.

4.1.1 Hutchinson's Estimator

The Hutchinson estimator was first proposed in [54] for estimating the trace of influence matrices in penalized least squares problems and in [68] for estimating $\mathrm{Tr}(A^{-1})$, the trace of the

inverse of a matrix. It can be applied in the same way for any implicitly given matrix, however. It is based on the following fact: If $\boldsymbol{x}$ is a random vector with independent and identically distributed (i.i.d.) components such that $\mathbb{E}[\boldsymbol{x}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = I$, then $\mathbb{E}[\boldsymbol{x}^T B \boldsymbol{x}] = \operatorname{Tr}(B)$. One obtains Hutchinson's estimator by sampling N independent random vectors with this property, and then averaging the results:

$$\operatorname{Tr}_N^{(H)}(B) = \frac{1}{N} \sum_{s=1}^{N} (\boldsymbol{x}^{(s)})^T B \boldsymbol{x}^{(s)}. \tag{4.1}$$

From the linearity of the expected value it follows that $\mathbb{E}[\operatorname{Tr}_N^{(H)}(B)] = \operatorname{Tr}(B)$. The accuracy of the approximation $\operatorname{Tr}_N^{(H)}(B)$ depends on N and the distribution of the entries of the sample vectors. Two important examples are *Rademacher vectors*, whose entries are i.i.d., taking the values ± 1 with equal probability $\frac{1}{2}$, or *Gaussian vectors*, whose entries are i.i.d. according to the standard normal distribution. For these particular cases, the variance of the estimator is explicitly known; see, e.g., [43, Corollary 3.2]. For convenience, we from now on we use the notation $\operatorname{Off}(B)$ for the off-diagonal part of a matrix B.

Theorem 4.1. *Let $B \in \mathbb{R}^{n \times n}$ be symmetric, and let $\boldsymbol{x}^{(s)}, s = 1, \ldots, N$ be independent random vectors with i.i.d. components.*

1. *If $\boldsymbol{x}^{(s)}$ are Rademacher vectors, then*

$$\mathbb{V}[\operatorname{Tr}_N^{(H)}(B)] = \frac{2}{N} \| \operatorname{Off}(B) \|_F^2 = \frac{2}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} |[B]_{ij}|^2. \tag{4.2}$$

2. *If $\boldsymbol{x}^{(s)}$ are Gaussian vectors, then*

$$\mathbb{V}[\operatorname{Tr}_N^{(H)}(B)] = \frac{2}{N} \|B\|_F^2 = \frac{2}{N} \sum_{i,j=1}^{n} |[B]_{ij}|^2. \tag{4.3}$$

Note that Rademacher vectors can result in estimates with much smaller variance, especially if B is diagonally dominant. Similar results hold for non-symmetric B, involving $\operatorname{Off}(B + B^T)$.

A more accurate analysis of the Hutchinson approximation can be achieved by deriving tail bounds of the following form: given a target accuracy $\varepsilon > 0$ and a failure probability δ, find N such that the (ε, δ) approximation

$$\mathbb{P}(|\operatorname{Tr}(B) - \operatorname{Tr}_N^{(H)}(B)| \geq \varepsilon) \leq \delta \tag{4.4}$$

holds. A vast literature on the topic has been developed in the past years; see for instance [6, 28, 90]. The following recent result is currently the one with the tightest tail bounds.

Theorem 4.2 (Theorem 1 and Corollary 1 in [28])**.** *Let $B \in \mathbb{R}^{n \times n}$ be symmetric and let $\boldsymbol{x}^{(s)}, s = 1, \ldots, N$ be independent random vectors.*

1. *If $\boldsymbol{x}^{(s)}$ are Rademacher vectors, then*

$$\mathbb{P}(|\operatorname{Tr}(B) - \operatorname{Tr}_N^{(H)}(B)| \geq \varepsilon) \leq 2\exp\left(-\frac{N\varepsilon^2}{8\|\operatorname{Off}(B)\|_F^2 + 8\varepsilon\|\operatorname{Off}(B)\|_2}\right) \quad (4.5)$$

   for every $\varepsilon > 0$. In particular, for $N \geq \frac{8}{\varepsilon^2}(\|\operatorname{Off}(B)\|_F^2 + \varepsilon\|\operatorname{Off}(B)\|_2)\log\frac{2}{\delta}$ it holds that $\mathbb{P}(|\operatorname{Tr}(B) - \operatorname{Tr}_N^{(H)}(B)| \geq \varepsilon) \leq \delta$.

2. *If $\boldsymbol{x}^{(s)}$ are Gaussian vectors, then*

$$\mathbb{P}(|\operatorname{Tr}(B) - \operatorname{Tr}_N^{(H)}(B)| \geq \varepsilon) \leq 2\exp\left(-\frac{N\varepsilon^2}{4\|B\|_F^2 + 4\varepsilon\|B\|_2}\right)$$

   for all $\varepsilon > 0$. In particular, for $N \geq \frac{4}{\varepsilon^2}(\|B\|_F^2 + \varepsilon\|B\|_2)\log\frac{2}{\delta}$ it holds that $\mathbb{P}(|\operatorname{Tr}(B) - \operatorname{Tr}_N^{(H)}(B)| \geq \varepsilon) \leq \delta$.

As a consequence of these tail bounds, to achieve an (ε, δ) approximation, we require $N = \mathcal{O}(\varepsilon^{-2}\log(\frac{1}{\delta}))$ samples. This asymptotic result is sharp, as discussed in [108]. For favorable distributions of the eigenvalues of B it is possible to obtain better asymptotic results by resorting to the Hutch++ estimator [81].

4.1.2 Hutch++

The Hutch++ method from [81] is essentially based on the observation that the trace of B might be well approximated by the trace of a low-rank approximation of B if its eigenvalues decay rapidly. Specifically, consider the eigendecomposition $B = W\Lambda W^T$ such that $W = [W_1, W_2], \Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2)$, where $\Lambda_1 \in \mathbb{R}^{p\times p}$ is associated to the p largest eigenvalues, and the columns of $W_1 \in \mathbb{R}^{n\times p}$, $W_2 \in \mathbb{R}^{n\times(n-p)}$ are eigenvectors associated with the diagonal entries of Λ_1, Λ_2, respectively. Then

$$\operatorname{Tr}(B) = \operatorname{Tr}(W_1\Lambda_1 W_1^T) + \operatorname{Tr}(W_2\Lambda_2 W_2^T), \quad (4.6)$$

and if the eigenvalues in Λ_1 are significantly larger than those in Λ_2, the first term in (4.6) might already serve as a good approximation, or at least the trace of the second term will be small and easier to approximate using the Hutchinson method. Precomputing and "deflating" eigenvectors associated with large eigenvalues is a well-known technique in the context of trace estimation; see, e.g., [50].

The key idea in Hutch++ is that it is enough to perform deflation using very rough approximations to eigenvectors of B which result from a single step of power iteration (i.e., a multiplication with B). The usage of more steps is also discussed and employed in [93]. We recall here the basic Hutch++ algorithm, that is, the version where the parameters p (rank of the low rank approximation) and q (number of vectors for the standard Hutchinson estimation) are given *a priori*. We use the notation $Q = \operatorname{orth}(M)$ to denote a matrix whose columns form an orthonormal basis of $\operatorname{range}(M)$. This can be retrieved from a reduced QR factorization of M.

Algorithm 2 Basic Hutch++ trace estimator

Input: $B \in \mathbb{R}^{n\times n}$ symmetric, $p, q \in \mathbb{N}$
Output: $\mathrm{Tr}_{p,q}^{(H++)}(B) \approx \mathrm{Tr}(B)$
1: Sample $\Omega \in \mathbb{R}^{n\times p}$ with random i.i.d. Gaussian or Rademacher entries
2: Compute $Q = \mathrm{orth}(B\Omega)$
3: Compute $\mathrm{Tr}_1 = \mathrm{Tr}(Q^TBQ) = \sum_{i=1}^{p}(q^{(i)})^TBq^{(i)}$ ▷ exact trace of low rank approx
4: Sample $X \in \mathbb{R}^{n\times q}$ with random i.i.d. Gaussian or Rademacher entries
5: Compute $Y = (I - QQ^T)X = \left[y^{(1)}|\cdots|y^{(q)}\right]$
6: Compute $\mathrm{Tr}_2 = \frac{1}{q}\mathrm{Tr}(Y^TBY) = \frac{1}{q}\sum_{i=1}^{q}(y^{(i)})^TBy^{(i)}$ ▷ Hutchinson estimator on remainder
7: **return** $\mathrm{Tr}_{p,q}^{(H++)}(B) = \mathrm{Tr}_1 + \mathrm{Tr}_2$

Algorithm 2 requires $2p + q$ matrix-vector multiplications with B: p for computing $B\Omega$ at line 2, p for computing $\mathrm{Tr}_1 = \mathrm{Tr}(Q^TBQ)$ at line 3, and q to compute BY and form Tr_2 at line 6. Notice that $\mathrm{Tr}(Q^TBQ) = \mathrm{Tr}(QQ^TB)$ and QQ^TB coincides with the *randomized rangefinder* of B and is close, up to a failure probability, to the best rank p approximation of B; see [78, Chapter 11]. A variant of Hutch++ for the trace of positive definite matrices, called Nyström++, replaces the randomized low rank approximation technique with the Nyström approximation [55] $B_{\mathrm{Nys}} = B\Omega(\Omega^TB\Omega)^{\dagger}(B\Omega)^T$, where $\Omega \in \mathbb{R}^{n\times p}$ is a sample matrix and $(\Omega^TB\Omega)^{\dagger}$ is the Moore-Penrose pseudoinverse of $\Omega^TB\Omega$; see [86] for more details.

Hutch++ performs particularly well if the eigenvalues of B drop off rapidly. If they don't, Hutch++ will not significantly improve over the standard Hutchinson estimator. We will return to this in Section 4.1.3.

An adaptive variant has been developed in [86]. This allows the user to give a tolerance ϵ and a failure probability δ as input and chooses adaptively the parameters p and q needed to get the tail bound

$$\mathbb{P}\{|\mathrm{Tr}(B) - \mathrm{Tr}_{p,q}^{(H++)}(B)| \geq \epsilon\} \leq \delta.$$

Further strategies, based on Krylov subspace iterations, can be found in [26].

4.1.3 XTrace

XTrace was recently suggested in [37] as a modification of Hutch++ using bootstrap resampling, for which any permutation of the test vectors involved in the computation yields the same result. In view of the *exchangeability principle* [62] this property must be satisfied by a trace estimator with optimal variance. A basic pseudocode is the following.

The quantities $\mathrm{Tr}(Q_{(k)}^TBQ_{(k)}), k = 1, \ldots, N/2$ can all be recovered by $Q = \mathrm{orth}(Z)$ with cheap operations. Hence, an efficient implementation requires $N/2$ matvecs for computing Z at line 2 and $N/2$ quadratic forms for computing $(\boldsymbol{\omega}^{(k)})^TB\boldsymbol{\omega}^{(k)}$, resulting in N total matvecs. See [37] for more details.

Another powerful trace estimator that follows the same principles as XTrace comes from the Nyström approximation, similarly to what was done in [86] for Nyström++. The resulting

Algorithm 3 Basic XTrace trace estimator

Input: $B \in \mathbb{R}^{n\times n}$ symmetric, number N of matvecs, where N is even
Output: $\mathrm{Tr}_N^{(XT)}(B) \approx \mathrm{Tr}(B)$
1: Sample $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(N/2)} \in \mathbb{R}^n$ with random i.i.d. Gaussian or Rademacher entries
2: Compute $Z = B\big[\boldsymbol{\omega}^{(1)} | \cdots | \boldsymbol{\omega}^{(N/2)}\big]$
3: **for** $k = 1, \ldots, N/2$ **do**
4: $\quad Q_{(k)} = \mathrm{orth}(Z_{-k})$ ▷ Z_{-k} obtained by removing the k-th column of Z
5: $\quad \widehat{\mathrm{Tr}} = \mathrm{Tr}(Q_{(k)}^T B Q_{(k)}) + (\boldsymbol{\omega}^{(k)})^T B \boldsymbol{\omega}^{(k)}$
6: **end for**
7: **return** $\mathrm{Tr}_N^{(XT)}(B) = \frac{2}{N}\sum_{k=1}^{N/2} \hat{\mathrm{Tr}}_k$

procedure is called XNysTrace and is summarized by the following pseudocode.

Algorithm 4 Basic XNysTrace trace estimator

Input: $B \in \mathbb{R}^{n\times n}$ symmetric positive definite, number N of matvecs
Output: $\mathrm{Tr}_N^{(XNT)}(B) \approx \mathrm{Tr}(B)$
1: Sample $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(N)} \in \mathbb{R}^n$ with random i.i.d. Gaussian or Rademacher entries
2: Compute $Z = B\big[\boldsymbol{\omega}^{(1)} | \cdots | \boldsymbol{\omega}^{(N/2)}\big]$
3: **for** $k = 1, \ldots, N$ **do**
4: $\quad \widehat{B}_k = Y_{-k}(\Omega_{-k}^T Z_{-k})^\dagger Z_{-k}^T$
5: $\quad \widehat{\mathrm{Tr}_k} = \mathrm{Tr}(\widehat{B}_k) + (\boldsymbol{\omega}^{(k)})(B - \widehat{B}_k)\boldsymbol{\omega}^{(k)}$
6: **end for**
7: **return** $\mathrm{Tr}_N^{(XNT)}(B) = N^{-1}\sum_{k=1}^{N} \widehat{\mathrm{Tr}_k}$

A careful implementation of this method requires N matvecs with B and other cheaper operations; see [37] for more details.

The excellent performance brought by Hutch++, XTrace and XNysTrace in case of exponentially decaying eigenvalues is explained by the following result [37].

Theorem 4.3. *Let $B \in \mathbb{R}^{n\times n}$ be symmetric positive definite whose eigenvalues have the exponential decay $\lambda_i(B) \le \alpha^i$ for $i = 1, 2, \ldots$, where $\alpha \in (0, 1)$. With Gaussian test vectors, the standard deviations of the Hutch++, XTrace and XNysTrace estimators satisfy*

$$\begin{aligned}(\mathbb{V}[\mathrm{Tr}_{N/3,N/3}^{(H++)}(B)])^{1/2} &\le C_1(\alpha)\alpha^{N/3};\\ (\mathbb{V}[\mathrm{Tr}_N^{(XT)}(B)])^{1/2} &\le C_2(\alpha)\alpha^{N/2};\\ (\mathbb{V}[\mathrm{Tr}_N^{(XNT)}(B)])^{1/2} &\le C_3(\alpha)\alpha^{N},\end{aligned}$$

where $C_j(\alpha)$, $j = 1, 2, 3$, depend only on α.

In our setting, where $B = f(A)$, Theorem 4.3 means that one can expect good performance of the three methods when f is large on a small part of the spectral interval of A and

has a fast decay everywhere else, e.g., when $f(A) = \exp(-A)$ and A is positive definite, but for a function like $f(A) = \sqrt{A}$ with $\sigma(A) \subseteq [0, 1]$, performance will generally not be satisfactory, as we will see by means of an example in Section 4.4.3.

4.2 Deterministic Probing Approach

Probing methods for trace (or diagonal) estimation were introduced in [99] and later refined, extended and analyzed in, e.g., [47, 73, 98]. The basic idea of probing methods is based on the observation that the magnitude of the entries of most matrix functions $f(A)$ exhibit a (often exponential) decay away from the sparsity pattern of A; see, e.g., [15, 17, 35, 45, 47, 95] and the references therein.

This observation motivates the construction of *probing vectors*

$$\boldsymbol{v}_\ell = \sum_{i \in C_\ell} \boldsymbol{e}_i, \quad \ell = 1, \ldots, n_\mathrm{c}, \tag{4.7}$$

where $\boldsymbol{e}_i$ denotes the ith vector of the canonical basis of $\mathbb{R}^n$ and the sets $C_\ell, \ell = 1, \ldots, n_\mathrm{c}$, form a partitioning of the index set $\mathcal{V} = \{1, \ldots, n\}$. We denote their sizes by $|C_\ell| =: n_\ell$. The (deterministic) probing approximation corresponding to (4.7) is then given by

$$\mathcal{T}(f(A)) = \sum_{\ell=1}^{n_\mathrm{c}} \boldsymbol{v}_\ell^T f(A) \boldsymbol{v}_\ell \tag{4.8}$$

with error

$$\mathrm{Tr}(f(A)) - \mathcal{T}(f(A)) = -\sum_{\ell=1}^{n_\mathrm{c}} \sum_{\substack{i,j \in C_\ell \\ i \neq j}} [f(A)]_{ij}. \tag{4.9}$$

One can see from (4.9) that the approximation is accurate when the entries $[f(A)]_{ij}$ are small in modulus for i and j in the same index set C_ℓ. Therefore, if the entries of $f(A)$ decay away from the sparsity pattern of A, a typical approach is to construct the sets $C_1, \ldots, C_{n_\mathrm{c}}$ via a distance-d coloring of $\mathcal{G}(A)$, according to the following definition.

Definition 4.1. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. If the mapping $\mathrm{col} : \mathcal{V} \to \{1, \ldots, n_\mathrm{c}\}$ is such that $\mathrm{col}(i) \neq \mathrm{col}(j)$ if $\mathrm{d}(i, j) \leq d$, then the corresponding partition of $\mathcal{V}$ given via $C_\ell = \{i \in \mathcal{V} \mid \mathrm{col}(i) = \ell\}, \ell = 1, \ldots, n_\mathrm{c}$, is called a *distance-$d$ coloring* (with n_c colors) of $\mathcal{G}$.

Note that a distance-1 coloring would be the standard node coloring of a graph, where any two adjacent nodes have different colors. The computation of a distance-d coloring (in particular for large values of d) is a computationally demanding task in general, even if one does not aim for an optimal coloring. It is common practice to use the greedy approach described in the following subsection for obtaining a suboptimal coloring with affordable computation cost [18, 47, 94].

Algorithm 5 Greedy algorithm for a distance-d coloring

Input: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, \dots, n\}$ and distance d
Output: Distance-d coloring col

1: $\mathrm{col}(1) = 1$
2: **for** $i = 2 : n$ **do**
3: $\quad W_i = \{j \in \{1, \dots, i-1\} : \mathrm{d}(i, j) \leq d\}$
4: $\quad \mathrm{col}(i) = \min\{k > 0 : k \neq \mathrm{col}(j) \text{ for all } j \in W_i\}$
5: **end for**

4.2.1 Computation of the Coloring

The computation of the coloring is often the most expensive part of the overall probing method; see, e.g., the discussion in [18, Section 6].

One can obtain different colorings depending on the order of the nodes; sorting the nodes by descending degree usually leads to good performance [94, 72]. The number of colors obtained with this algorithm is at most $\Delta^d + 1$, where Δ is the maximum degree of a node in the graph induced by A, and the cost is at most $\mathcal{O}(n\,\Delta^d)$ if the d-neighbors W_i of each node i are computed with a traversal of the graph. Alternatively, the greedy coloring can be obtained by computing A^d, which can be done using at most $2\lfloor \log_2 d \rfloor$ matrix-matrix multiplications [67, Section 4.1]. In our Matlab implementation of the probing method, we construct the greedy distance-d coloring by computing A^d, as we found it to be faster than the graph-based approach; this is likely due to the more efficient Matlab implementation of matrix-matrix operations.

Remark 4.1. The efficiency of Algorithm 5 is significantly influenced by the sparsity structure of $\mathcal{G}$. For instance, good performance is expected if $\mathcal{G}$ is a large-world network, such as road networks, characterized by a long average path length. On the other hand, small-world graph, such as social networks and scientific collaboration networks, tend to exhibit large values of Δ and a considerable fill-in of the entries of A^d even for small values of d, resulting in a high computational effort of Algorithm 5.

For special graph structures, colorings with a small but not necessarily minimal number of colors are known in closed form and therefore available cheaply. As these will play a role in our subsequent analysis, we recall them here.

Proposition 4.4 (from Section 2 in [47])**.** *1. Let A be m-banded, i.e., $[A]_{ij} = 0$ for $|i - j| > m$. Then a distance-d coloring for $\mathcal{G}(A)$ with $n_\mathrm{c} = dm + 1$ colors is given by*

$$\mathrm{col}(i) = (i-1) \mod (dm+1) + 1, \quad i = 1, \dots, n, \tag{4.10}$$

and this coloring is optimal if all entries within the band of A are nonzero.

2. Let $\mathcal{G}(A) = (V, E)$ be a regular D-dimensional lattice of size $n_1 \times \cdots \times n_D$ and let its nodes be labeled as $V = \{v \in \mathbb{Z}^D : 0 \leq [v]_k \leq n_k - 1 \text{ for } k = 1, \dots, D\}$. Then

a distance-d coloring for $\mathcal{G}(A)$ with $n_c = (d+1)^D$ colors is given by

$$\mathrm{col}(v) = \left(\sum_{k=0}^{D-1} \widetilde{[v]}_k (d+1)^k \right) + 1, \quad \widetilde{[v]}_k = [v]_k \mod (d+1). \tag{4.11}$$

Remark 4.2.

1. The coloring (4.11) is not optimal. For $D = 2$, the optimal distance-d coloring is explicitly known and requires $\lceil \frac{1}{2}(d+1)^2 \rceil$ colors. The regular structure of the coloring (4.11) allows for an easier analysis of corresponding probing approximations, though, as discussed in [47, 94]. For $D > 2$, optimal distance-d colorings of the lattice are not explicitly known for general $d > 1$.

2. A different way of coloring regular lattices is used in the *hierarchical probing* approach from [73, 98], which starts with a distance-1 coloring and recursively subdivides it. This typically results in a non-optimal coloring that uses more colors than a coloring computed by Algorithm 5, but allows us to reuse certain computational results if one detects that the probing approximation is not accurate enough and the distance for the coloring thus needs to be increased.

4.2.2 Error Bounds

Based on the colorings from Proposition 4.4, one can derive bounds for the error (4.9) under the assumption that $f(A)$ has the exponential decay property

$$|[f(A)]_{ij}| \leq c q^{\mathrm{d}(i,j)} \tag{4.12}$$

for all i, j, where $c > 0$, $0 \leq q < 1$ are constants independent of i, j.

Theorem 4.5 (Theorems 4.1 and 4.2 in [47])**.** *Let $A \in \mathbb{R}^{n \times n}$ and let $f(A)$ be defined.*

1. *Suppose that A is m-banded such that $f(A)$ fulfills* (4.12)*, and let $\mathcal{T}_d(f(A))$ be the probing approximation* (4.8) *of $\mathrm{Tr}(f(A))$ associated with the distance-d coloring* (4.10)*. Then*

$$|\operatorname{Tr}(f(A)) - \mathcal{T}_d(f(A))| \leq n \cdot 2c \frac{q^d}{1 - q^d}. \tag{4.13}$$

2. *Suppose that A is such that $\mathcal{G}(A)$ is a regular D-dimensional lattice and $f(A)$ fulfills* (4.12)*. Let $\mathcal{T}_d(f(A))$ be the probing approximation of $\mathrm{Tr}(f(A))$ associated with the distance-d coloring* (4.11)*. Then*

$$|\operatorname{Tr}(f(A)) - \mathcal{T}_d(f(A))| \leq n \cdot 2cD \operatorname{Li}_{1-D}(q^d), \tag{4.14}$$

where $\mathrm{Li}_s(z) = \sum_{k=1}^{\infty} \frac{z^i}{i^s}$ is the polylogarithm.

Remark 4.3. Note that, when D is a positive integer, $\mathrm{Li}_{1-D}(z)$ is a rational function, and explicit representations are known, e.g.

$$\mathrm{Li}_0(z) = \frac{z}{1-z}, \quad \mathrm{Li}_{-1}(z) = \frac{z}{(1-z)^2}, \quad \mathrm{Li}_{-2}(z) = \frac{z+4z^2+z^3}{(1-z)^4}.$$

In all these cases, $\mathrm{Li}_{1-D}(q^d) = \mathcal{O}(q^d)$ for large d.

It is also possible to obtain an error bound that holds for general graphs $\mathcal{G}(A)$ (i.e., without relying on a specific structure or the use of a specific coloring), based on polynomial approximation of f over an interval $[a,b] \supset \sigma(A)$. To this purpose, we consider the polynomial approximation error defined in (3.1). Theorem 4.4 in [47] gives a bound for the case when $E_k(f,[a,b])$ decays exponentially in k. Using essentially the same proof, one can obtain the following general result.

Theorem 4.6. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with $\sigma(A) \subset [a,b]$. Let $\mathcal{T}_d(f(A))$ be the probing approximation of $\mathrm{Tr}(f(A))$ associated with a distance-d coloring of $\mathcal{G}(A)$. Then*

$$|\,\mathrm{Tr}(f(A)) - \mathcal{T}_d(f(A))| \leq 2n \cdot E_d(f,[a,b]). \tag{4.15}$$

The numerical examples in Section 4.4 and in [47, Section 6] illustrate that the error of the deterministic probing approximation indeed scales linearly with the size n, suggesting that $\mathcal{O}(n)$ error bounds as in Theorem 4.6 are the best we can achieve with this method.

Since we expect $\mathrm{Tr}(f(A))$ to grow with n, the linear scaling of the absolute error is usually a minor issue when we aim for a certain relative accuracy. However, one needs to be careful when dealing with large-scale problems in which $\mathrm{Tr}(f(A))$ is much smaller than n. In such cases, it is necessary to increase d to achieve a small relative error. Nonetheless, we will see that better scaling with the size is achieved via stochastic probing in Section 4.3.

4.3 Stochastic Probing

Stochastic probing methods combine the probing approach discussed in Section 4.2 with the Hutchinson estimator from Section 4.1.1. Stochastic probing has already been used under the name "dilution"—restricted to distance $d = 1$—for approximating the trace of the inverse in lattice quantum chromodynamics computations [7, 82] and for more general d in [5, 50], but without a theoretical analysis.

4.3.1 Description of the Method

We assume that $C_1, \ldots, C_{n_\mathrm{c}}$ (with $|C_\ell| = n_\ell$) is a partition of $V = \{1,\ldots,n\}$ associated with a distance-d coloring of $\mathcal{G}(A)$. We define a stochastic probing vector associated with C_ℓ as

$$\boldsymbol{w}_\ell = \sum_{i\in C_\ell} X_i \boldsymbol{e}_i, \quad \ell = 1,\ldots,n_\mathrm{c}, \tag{4.16}$$

where X_i, $i \in C_\ell$, are i.i.d. random variables such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$ for all $i \in C_\ell$. In the case of Rademacher variables, we call $\boldsymbol{w}_\ell$ a *Rademacher probing vector*, while in the case of Gaussian variables, we call $\boldsymbol{w}_\ell$ a *Gaussian probing vector*. Basic properties of these stochastic probing vectors are summarized in the following proposition.

Proposition 4.7. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Let a partition $C_1, \dots, C_{n_c}$ of $V = \{1, \dots, n\}$ associated with a distance-d coloring of $\mathcal{G}(A)$ be given and let $\boldsymbol{w}_\ell$ be stochastic probing vectors associated with C_ℓ, $\ell = 1, \dots, n_c$, as defined in* (4.16)*. Then*

$$\boldsymbol{w}_\ell^T f(A)\boldsymbol{w}_\ell = \sum_{k\in C_\ell} X_k^2 [f(A)]_{kk} + \sum_{\substack{i,j\in C_\ell \\ i\neq j}} X_i X_j [f(A)]_{ij}, \tag{4.17}$$

and thus,

$$\mathbb{E}[\boldsymbol{w}_\ell^T f(A)\boldsymbol{w}_\ell] = \operatorname{Tr}([f(A)]_{C_\ell}) = \sum_{k\in C_\ell} [f(A)]_{kk}. \tag{4.18}$$

Proof. The identity (4.17) follows trivially from the expression (4.16) that defines $\boldsymbol{w}_\ell$. Since X_i and X_j are independent for $i \neq j$, we get $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = 0$, and since $\mathbb{E}[X_k^2] = 1$ for all k, we obtain (4.18). □

In the spirit of the Hutchinson estimator, we sample N_ℓ probing vectors $\boldsymbol{w}_\ell^{(1)}, \dots, \boldsymbol{w}_\ell^{(N_\ell)}$ associated with C_ℓ and obtain the approximation

$$\frac{1}{N_\ell} \sum_{s=1}^{N_\ell} (\boldsymbol{w}_\ell^{(s)})^T f(A) \boldsymbol{w}_\ell^{(s)} \approx \operatorname{Tr}([f(A)]_{C_\ell}) \tag{4.19}$$

for the "partial trace" belonging to color ℓ. To approximate the full trace, we sum estimates for all the partial traces according to the following definition.

Definition 4.2. Let $C_\ell, \ell = 1, \dots, n_c$, be a partitioning of $V = \{1, \dots, n\}$ and let the number of samples for each ℓ be collected in the n_c-tuple $\mathcal{N} = (N_1, \dots, N_{n_c})$. Then the *stochastic probing estimator* of $\operatorname{Tr}(f(A))$ is given by

$$\mathcal{T}_d^{\mathcal{N}}(f(A)) := \sum_{\ell=1}^{n_c} \frac{1}{N_\ell} \sum_{s=1}^{N_\ell} (\boldsymbol{w}_\ell^{(s)})^T f(A) \boldsymbol{w}_\ell^{(s)}. \tag{4.20}$$

In view of (4.18), we have

$$\mathbb{E}[\mathcal{T}_d^{\mathcal{N}}(f(A))] = \operatorname{Tr}(f(A)). \tag{4.21}$$

4.3.2 Variance of the Stochastic Probing Estimator

From basic properties of the variance, we have

$$\mathbb{V}[\mathcal{T}_d^{\mathcal{N}}(f(A))] = \sum_{\ell=1}^{n_c} V_\ell / N_\ell, \quad \text{where } V_\ell := \mathbb{V}[\boldsymbol{w}_\ell^T f(A)\boldsymbol{w}_\ell]. \tag{4.22}$$

The individual variances V_ℓ can be given explicitly.

Theorem 4.8. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric, let a partition $C_1, \dots C_{n_c}$ of $V = \{1, \dots, n\}$ be given and let $\boldsymbol{w}_\ell$ be stochastic probing vectors associated with $C_\ell, \ell \in \{1, \dots, n_c\}$, as defined in* (4.16)*.*

1. *If $\boldsymbol{w}_\ell$ is a Rademacher probing vector, then*

$$V_\ell = 2\| \operatorname{Off}([f(A)]_{C_\ell})\|_F^2 = 2 \sum_{\substack{i,j\in C_\ell \\ i\neq j}} |[f(A)]_{ij}|^2. \tag{4.23}$$

2. *If $\boldsymbol{w}_\ell$ is a Gaussian probing vector, then*

$$V_\ell = 2\|[f(A)]_{C_\ell}\|_F^2 = 2 \sum_{i,j\in C_\ell} |[f(A)]_{ij}|^2. \tag{4.24}$$

Proof. For any $\ell = 1, \dots, n_c$ we have $\boldsymbol{w}_\ell^T f(A)\boldsymbol{w}_\ell = ([\boldsymbol{w}_\ell]_{C_\ell})^T [f(A)]_{C_\ell}[\boldsymbol{w}_\ell]_{C_\ell}$. In view of Theorem 4.1, if $\boldsymbol{w}_\ell$ is a Rademacher probing vector, we get

$$\mathbb{V}[\boldsymbol{w}_\ell^T f(A)\boldsymbol{w}_\ell] = \mathbb{V}[([\boldsymbol{w}_\ell]_{C_\ell})^T [f(A)]_{C_\ell}[\boldsymbol{w}_\ell]_{C_\ell}] = \| \operatorname{Off}([f(A)]_{C_\ell})\|.$$

Similarly, if $\boldsymbol{w}_\ell$ is a Gaussian probing vector, then

$$\mathbb{V}[\boldsymbol{w}_\ell^T f(A)\boldsymbol{w}_\ell] = \mathbb{V}[([\boldsymbol{w}_\ell]_{C_\ell})^T [f(A)]_{C_\ell}[\boldsymbol{w}_\ell]_{C_\ell}] = \|[f(A)]_{C_\ell}\|_F^2.$$

This concludes the proof. □

Remark 4.4. The diagonal entries of $f(A)$ are present in the summation in (4.24), but not in (4.23). This suggests—and is actually confirmed by numerical experiments not reported here—that if a decay away from the sparsity pattern of A is present in $f(A)$, then the variance is much smaller in the Rademacher case, since the diagonal entries will dominate in (4.24). From now on, we will therefore only consider the Rademacher distribution for our analysis and experiments.

From (4.22) and Theorem 4.8, we obtain the variance of the Rademacher probing approximation. We immediately state it for distance-d colorings, although it also holds for a general partitioning $C_1, \dots, C_{n_c}$.

Lemma 4.9. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and let $C_\ell, \ell = 1, \dots, n_c$, be a distance-$d$ coloring of $\mathcal{G}(A)$. Further, let $\mathcal{N} = (N_1, \dots, N_{n_c})$, let $\boldsymbol{w}_\ell^{(1)}, \dots, \boldsymbol{w}_\ell^{(N_\ell)}$ be Rademacher probing vectors associated with C_ℓ, and let $\mathcal{T}_d^{\mathcal{N}}(f(A))$ be the corresponding Rademacher probing approximation* (4.20)*. Then, with V_ℓ from* (4.23)*,*

$$\mathbb{V}[\mathcal{T}_d^{\mathcal{N}}(f(A))] = \sum_{\ell=1}^{n_c} \frac{V_\ell}{N_\ell} = 2\sum_{\ell=1}^{n_c} \frac{1}{N_\ell} \| \operatorname{Off}([f(A)]_{C_\ell})\|_F^2. \tag{4.25}$$

Remark 4.5. Lemma 4.9 tells us that the variance becomes much smaller as d increases if a decay on the entries is present. However, as in the deterministic case with the formula (4.9), the exact value of $V_\ell = \| \operatorname{Off}([f(A)]_{C_\ell})\|_F^2$ on the right-hand side is not known, since $f(A)$ is not explicitly available. Bounds on V_ℓ for special cases will be given in Section 4.3.4.

Assuming that a distance-d coloring of $\mathcal{G}(A)$ is already given, the computational cost of the stochastic probing method is proportional to the number of matrix-vector products (or rather quadratic forms) that need to be evaluated. In the following, we therefore only count the number of quadratic forms to gauge the efficiency of the method. For example, the cost for computing the stochastic probing estimator $\mathcal{T}_d^{\mathcal{N}}(f(A))$ covered in Lemma 4.9 is taken to be $N_1 + \cdots + N_{n_c}$ quadratic forms with $f(A)$.

It is in general not advisable to use the same number of samples for each C_ℓ: When the partition comes from a distance-d coloring of $\mathcal{G}(A)$, it is often the case that the variances V_ℓ vary widely, e.g., if some sets C_ℓ in the coloring are much smaller than the others. A cost optimal approach will thus use different sample sizes for different colors, as we develop now.

Suppose that we aim to obtain an estimator with overall variance at most ε^2. Then, taking into account the expression (4.22) for the variance and assuming that we know the individual variances V_ℓ, the overall lowest number of quadratic forms is obtained by solving the optimization problem

$$\min_{N_1,\ldots,N_{n_c}\in\mathbb{N}} \quad N_1 + \cdots + N_{n_c} \qquad \text{s. t.} \sum_{\ell=1}^{n_c} V_\ell/N_\ell = \varepsilon^2. \tag{4.26}$$

While the discrete problem (4.26) is difficult to solve, it is easy to do so if one relaxes it by allowing $N_\ell \geq 0$ to be real. Similar optimization problems occur in general multi-level Monte Carlo methods and have been solved before; see, e.g., [43, 53, 61]. The solution of the relaxed version of (4.26) is given by

$$N_\ell = \mu\sqrt{V_\ell}, \quad \mu = \varepsilon^{-2}\sum_{\ell=1}^{n_c}\sqrt{V_\ell}. \tag{4.27}$$

To obtain integer numbers of samples, one can simply pick the ceiling of the values, $\lceil N_\ell \rceil$ in (4.27).

As discussed before, the variances V_ℓ required for computing (4.27) are generally unknown in practice. In some cases it is possible to bound them a priori—as we will discuss in Section 4.3.4—or they can be estimated on-the-fly during the computation—as is, e.g., described in [61] in the context of stochastic multilevel trace estimation methods.

4.3.3 Tail Bounds for the Stochastic Probing Estimator

According to the central limit theorem, for large values of the N_ℓ, the stochastic estimator $\mathcal{T}_d^{\mathcal{N}}(f(A))$ for the trace is approximately normally distributed with mean $\operatorname{Tr}(f(A))$ and standard deviation $\sigma = (\sum_{\ell=1}^{n_c} V_\ell/N_\ell)^{1/2}$. This gives the approximate tail bound

$$\mathbb{P}\left(|\operatorname{Tr}(f(A)) - \mathcal{T}_d^{\mathcal{N}}(f(A))| \geq \varepsilon\right) \lesssim 1 - \operatorname{erf}\left(\frac{\varepsilon}{\sqrt{2}\sigma}\right),$$

with the error function $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$. Solving for σ^2 results in

$$\mathbb{P}\left(|\operatorname{Tr}(f(A)) - \mathcal{T}_d^{\mathcal{N}}(f(A))| \geq \varepsilon\right) \leq \delta \quad \text{if} \quad \sum_{\ell=1}^{n_c} V_\ell/N_\ell \lesssim \frac{\varepsilon^2}{2(\operatorname{inverf}(1-\delta))^2}, \tag{4.28}$$

where $\operatorname{inverf}(z)$ denotes the inverse of $\operatorname{erf}(z)$.

The relation (4.28) is not entirely satisfactory, since it does not quantify how "inexact" the inequality is and how large N_ℓ should be. It has the advantage, though, that it applies for any probability distribution used for the stochastic probing vectors. For Rademacher vectors, we now derive tail bounds for the stochastic probing estimator which do not rely on the central limit theorem and which hold for any choice of the sample sizes N_ℓ. To this end we recall the following result from [28] which, actually, is at the basis of theorem 4.2.

Theorem 4.10. *Let $\boldsymbol{x}$ be a Rademacher vector of length n and let $M \in \mathbb{R}^{n\times n}$ be a nonzero matrix such that $[M]_{ii} = 0$ for $i = 1, \dots, n$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\left(|\boldsymbol{x}^T M \boldsymbol{x}| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{8\|M\|_F^2 + 8\varepsilon\|M\|_2}\right).$$

Using this theorem, we can derive an analogue of Theorem 4.2 for the stochastic probing method.

Theorem 4.11. *Let A be symmetric, and let f and A be such that $f(A)$ is nonzero. Let $C_1, \dots, C_{n_c}$ be a distance-d coloring of $\mathcal{G}(A)$. Let $\mathcal{N} = (N_1, \dots, N_{n_c})$, and let $\boldsymbol{w}_\ell^{(s)}$ be Rademacher probing vectors, with $\mathcal{T}_d^{\mathcal{N}}(f(A))$ the corresponding trace estimate* (4.20)*. Then*

$$\mathbb{P}\left(|\operatorname{Tr}(f(A)) - \mathcal{T}_d^{\mathcal{N}}(f(A))| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{8\eta_1 + 8\varepsilon\eta_2}\right) \tag{4.29}$$

for every $\varepsilon > 0$, where

$$\eta_1 := \sum_{\ell=1}^{n_c} \frac{1}{N_\ell}\|\operatorname{Off}([f(A)]_{C_\ell})\|_F^2 = \sum_{\ell=1}^{n_c} V_\ell/N_\ell, \quad \eta_2 := \max_{\ell=1,\dots,n_c} \frac{1}{N_\ell}\|\operatorname{Off}([f(A)]_{C_\ell})\|_2, \tag{4.30}$$

with V_ℓ the variances from (4.23)*. In particular, if we choose $N_1, \dots, N_{n_c}$ such that*

$$\eta_1 + \varepsilon\eta_2 \leq \frac{\varepsilon^2}{8\log\frac{2}{\delta}}, \tag{4.31}$$

we have $\mathbb{P}\left(|\operatorname{Tr}(f(A)) - \mathcal{T}_d^{\mathcal{N}}(f(A))| \geq \varepsilon\right) \leq \delta$.

Proof. Define the block diagonal matrices

$$M_\ell := \frac{1}{N_\ell}\begin{bmatrix} \operatorname{Off}([f(A)]_{C_\ell}) & & \\ & \ddots & \\ & & \operatorname{Off}([f(A)]_{C_\ell}) \end{bmatrix} \in \mathbb{R}^{n_\ell N_\ell \times n_\ell N_\ell}, \quad \ell = 1, \dots, n_c,$$

and

$$M := \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_{n_c} \end{bmatrix} \in \mathbb{R}^{\widehat{N}\times\widehat{N}},$$

where $\widehat{N} = \sum_{\ell=1}^{n_c} n_\ell N_\ell$. Then, $\mathcal{T}_d^{\mathcal{N}}(f(A)) - \operatorname{Tr}(f(A)) = \boldsymbol{x}^T M \boldsymbol{x}$ with $\boldsymbol{x}$ a Rademacher vector of length $\widehat{N}$ of the form

$$\boldsymbol{x} = \begin{bmatrix} \widehat{\boldsymbol{w}}_1 \\ \vdots \\ \widehat{\boldsymbol{w}}_{n_c} \end{bmatrix} \in \mathbb{R}^{\widehat{N}}, \quad \text{where } \widehat{\boldsymbol{w}}_\ell = \begin{bmatrix} [\boldsymbol{w}_\ell^{(1)}]_{C_\ell} \\ \vdots \\ [\boldsymbol{w}_\ell^{(N_\ell)}]_{C_\ell} \end{bmatrix} \in \mathbb{R}^{n_\ell N_\ell}, \ell = 1, \dots, n_c,$$

and each $[\boldsymbol{w}_\ell^{(s)}]_{C_\ell} \in \mathbb{R}^{n_\ell}$ being a Rademacher vector of length n_ℓ. We conclude by applying Theorem 4.10, noting that $\eta_1 = \|M\|_F^2$ and $\eta_2 = \|M\|_2$. □

Similarly to the discussion at the end of Section 4.3.2, we now elaborate on how to choose the number of samples $N_\ell, \ell = 1, \dots, n_c$, in order to obtain a cost-efficient (ε, δ) approximation of the trace, based on inequality (4.31). To minimize the number $N_1 + \cdots + N_\ell$ of quadratic forms involved in the computation of (4.20), one can solve

$$\min_{N_1,\dots,N_{n_c} \in \mathbb{N}} N_1 + \cdots + N_{n_c} \qquad \text{s.t.} \quad \eta_1 + \varepsilon\eta_2 \leq \varepsilon^2 / 8 \log(2/\delta), \tag{4.32}$$

where η_1, η_2 are defined in (4.30). A way to solve this approximately is to ignore the term $\varepsilon\eta_2$, due to the small factor ε and since we expect $\varepsilon\eta_2 \ll \eta_1$; a similar reasoning in a slightly different context is also used in [86]. With this assumption, noting also that $\eta_1 = \mathbb{V}[\mathcal{T}_d^{\mathcal{N}}(f(A))]$, the problem (4.32) reduces to

$$\min_{N_1,\dots,N_{n_c} \in \mathbb{N}} N_1 + \cdots + N_{n_c} \qquad \text{s.t.} \quad \mathbb{V}[\mathcal{T}_d^{\mathcal{N}}(f(A))] = \varepsilon^2 / 8 \log(2/\delta). \tag{4.33}$$

This is the same as (4.26), where ε^2 is replaced by $\varepsilon^2/8\log(2/\delta)$. Hence, by adapting (4.27) to this case, we find

$$N_\ell = \lceil \mu \sqrt{V_\ell} \rceil, \quad \mu = 8\varepsilon^{-2} \log \frac{2}{\delta} \sum_{\ell=1}^{n_c} \sqrt{V_\ell}, \quad V_\ell = \| \operatorname{Off}([f(A)]_{C_\ell}) \|_F^2. \tag{4.34}$$

4.3.4 A Priori Bounds for the Variance in Specific Cases

We now discuss three specific situations in which we can give a priori bounds on the variances V_ℓ and thus, according to Lemma 4.9, on the variance of the stochastic probing estimator $\mathcal{T}_d^{\mathcal{N}}(f(A))$. These bounds can then, in turn, be used to determine suitable numbers N_ℓ of samples based on the tail bounds of Section 4.3.3. An important feature of all three bounds is that they depend *only linearly* on the dimensions n_ℓ. Interpreting the square root of the variance as a measure for the accuracy of the Rademacher probing method, we thus have a *sublinear* dependence, proportional to the square root of the dimension. This is an "order $\frac{1}{2}$" improvement over the non-stochastic probing method, where the accuracy depends linearly on the dimension; see Theorem 4.5.

Proposition 4.12. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric, let $\mathcal{G}(A)$ be its graph, and assume that $f(A)$ has the exponential decay property* (4.12) *with constants $c > 0$ and $0 < q < 1$.*

(i) If $\mathcal{G}(A)$ is m-banded and $\mathcal{T}_d^{\mathcal{N}}(f(A))$ is the Rademacher probing approximation (4.20) *associated with the distance-d coloring* (4.10)*, then the variances V_ℓ from* (4.23) *satisfy*

$$V_\ell \leq n_\ell \cdot 4c^2 \frac{q^{2d}}{1-q^{2d}}, \tag{4.35}$$

(ii) If $\mathcal{G}(A)$ is a regular D-dimensional lattice and $\mathcal{T}_d^{\mathcal{N}}(f(A))$ is the Rademacher probing approximation (4.20) *associated with the distance-d coloring* (4.11)*, then the variances V_ℓ from* (4.23) *satisfy*

$$V_\ell \leq n_\ell \cdot 4c^2 D \operatorname{Li}_{1-D}(q^{2d}). \tag{4.36}$$

Proof. Let $C_1, \ldots, C_{n_\mathrm{c}}$ be the colors of the coloring in (i) or (ii). From the exponential decay property (4.12), we obtain

$$V_\ell = 2 \sum_{\substack{i,j \in C_\ell \\ i \neq j}} |[f(A)]_{ij}|^2 \leq 2 \sum_{\substack{i,j \in C_\ell \\ i \neq j}} c^2 q^{2\,\mathrm{d}(i,j)}. \tag{4.37}$$

In the case of (i), the distance-d coloring (4.10) has the property that for given i all other nodes in C_ℓ have a distance from i which is a multiple γd of d, and there are at most two such nodes for each $\gamma \in \mathbb{N}$; cf. [47, Section 4] for a detailed discussion. This is why for each i we can write $\sum_{j \in C_\ell, j \neq i} q^{2\,\mathrm{d}(i,j)} \leq 2 \sum_{\gamma=1}^{\infty} q^{2d\gamma} = 2q^{2d}/(1-q^{2d})$, which gives (4.35).

In the case of (ii) and the coloring (4.11), for a given node $i \in C_\ell$ it was shown in [47, Section 4] that all other nodes in C_ℓ have again a distance which is a multiple γd of d, and for each γ there are this time at most $2D\gamma^{D-1}$ nodes which have this distance. Thus,

$$V_\ell = 2 \sum_{i \in C_\ell} \sum_{j \in C_\ell, j \neq i} |[f(A)]_{ij}|^2 \leq 2 n_\ell \sum_{\gamma=1}^{\infty} 2D\gamma^{D-1} \cdot c^2 q^{2d\gamma} = n_\ell \cdot 4c^2 \operatorname{Li}_{1-D}(q^{2d}),$$

which is (4.36). □

The third situation that we consider is when $[f(A)]_{ij}$ has constant sign for $\mathrm{d}(i,j) > d$. Here, we can establish a bound using the polynomial approximation error $E_d(f, [a,b])$.

Proposition 4.13. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric with $\sigma(A) \subset [a,b]$ and let $C_1, \ldots, C_{n_\mathrm{c}}$ be a distance-d coloring of $\mathcal{G}(A)$. Let $\mathcal{T}_d^{\mathcal{N}}(f(A))$ be the Rademacher probing approximation* (4.20) *associated with this coloring. Suppose that $[f(A)]_{ij}$ has constant sign for all i, j with $\mathrm{d}(i,j) > d$. Then*

$$V_\ell \leq n_\ell \cdot 4\big(E_d(f, [a,b])\big)^2. \tag{4.38}$$

Proof. Denote by $\boldsymbol{v}_\ell$ the deterministic probing vector associated with the color C_ℓ, $\ell = 1, \ldots, n_\mathrm{c}$, so that we have

$$\operatorname{Tr}([f(A)]_{C_\ell}) - \boldsymbol{v}_\ell^T f(A) \boldsymbol{v}_\ell = - \sum_{\substack{i,j \in C_\ell \\ i \neq j}} [f(A)]_{ij}.$$

By assumption, all terms on the right-hand side have the same sign, which gives

$$\sum_{\substack{i,j\in C_\ell\\ i\neq j}} |[f(A)]_{ij}| = \Big|\sum_{\substack{i,j\in C_\ell\\ i\neq j}} [f(A)]_{ij}\Big| = \left|\mathrm{Tr}([f(A)]_{C_\ell}) - \boldsymbol{v}_\ell^T f(A)\boldsymbol{v}_\ell\right|,$$

and thus

$$\begin{aligned} V_\ell = 2\sum_{\substack{i,j\in C_\ell\\ i\neq j}} |[f(A)]_{ij}|^2 &\leq 2\max_{\substack{i,j\in C_\ell\\ i\neq j}} |[f(A)]_{ij}| \sum_{\substack{i,j\in C_\ell\\ i\neq j}} |[f(A)]_{ij}| \\ &\leq 2\max_{\mathrm{d}(i,j)>d} |[f(A)]_{ij}| \cdot \left|\mathrm{Tr}([f(A)]_{C_\ell}) - \boldsymbol{v}_\ell^T f(A)\boldsymbol{v}_\ell\right|. \end{aligned} \tag{4.39}$$

The first factor in (4.39) is bounded by a quantity that does not depend on ℓ,

$$\max_{\mathrm{d}(i,j)>d} |[f(A)]_{ij}| \leq E_d(f,[a,b]); \tag{4.40}$$

see, e.g., [45, Section 2.1]. For the second term, let p_d denote the polynomial of best approximation of degree d. Then $\mathrm{Tr}(p_d(A)) = v_\ell^T p_d(A) v_\ell$, since $[p_d(A)]_{ij} = 0$ if $\mathrm{d}(i,j) > d$, and thus

$$\begin{aligned} \left|\mathrm{Tr}([f(A)]_{C_\ell}) - \boldsymbol{v}_\ell^T f(A)\boldsymbol{v}_\ell\right| &\leq |\mathrm{Tr}([f(A) - p_d(A)]_{C_\ell})| + \left|\boldsymbol{v}_\ell^T (f(A) - p_d(A))\boldsymbol{v}_\ell\right| \\ &\leq \sum_{k\in C_\ell} |[f(A) - p_d(A)]_{kk}| + \left|\boldsymbol{v}_\ell^T (f(A) - p_d(A))\boldsymbol{v}_\ell\right|. \end{aligned}$$

Herein, each term in the sum is bounded by $E_d(f,[a,b])$, and for the second term we have

$$\left|v_\ell^T (f(A) - p_d(A)) v_\ell\right| \leq \|v_\ell\|_2^2 \cdot E_d(f,[a,b]) = n_\ell \cdot E_d(f,[a,b]).$$

Using this in (4.39) gives (4.38). □

Remark 4.6. When given a fixed budget N of quadratic forms that one wants to invest for trace estimation via stochastic probing, our results lead to a simple heuristic for distributing these across the different colors C_ℓ: According to (4.27), the optimal number of samples should be chosen proportionally to $\sqrt{V_\ell}$, and under the assumptions of Proposition 4.12 or Proposition 4.13 we further know that V_ℓ essentially scales as $\mathcal{O}(n_\ell)$. Thus, it is sensible to choose the numbers of samples as $N_\ell = \mathrm{round}(\nu\sqrt{n_\ell})$, where

$$\nu = \frac{N}{\sum_{\ell=1}^{n_\mathrm{c}} \sqrt{n_\ell}},$$

and $\mathrm{round}(\nu\sqrt{n_\ell})$ denotes the nearest integer to $\nu\sqrt{n_\ell}$.

The following result gives us bounds on the standard deviation (i.e., the sqare root of the variance) for the stochastic probing estimator.

Theorem 4.14. *Let $f(A)$ be defined for $A \in \mathbb{R}^{n\times n}$ symmetric. Let $C_1,\ldots,C_{n_\mathrm{c}}$ be a distance-d coloring of $\mathcal{G}(A)$, and let $\mathcal{T}_d^{\mathcal{N}}(f(A))$ be the Rademacher probing approximation* (4.20) *associated with the coloring.*

1. *If A is m-banded and $f(A)$ fulfills* (4.12)*, and the distance-d coloring is chosen as* (4.10)*, then*
$$\left(\mathbb{V}\left[\mathcal{T}_d^{\mathcal{N}}(f(A))\right]\right)^{\frac{1}{2}} \leq 2\sqrt{n}\, c \frac{q^d}{\sqrt{1-q^{2d}}};$$
2. *If $\mathcal{G}(A)$ is a regular D-dimensional lattice and the distance-d coloring is chosen as* (4.11)*, then*
$$\left(\mathbb{V}\left[\mathcal{T}_d^{\mathcal{N}}(f(A))\right]\right)^{\frac{1}{2}} \leq 2\sqrt{n}\, c \sqrt{D} \sqrt{\mathrm{Li}_{1-D}(q^{2d})};$$
3. *If $[f(A)]_{ij}$ has constant sign for $\mathrm{d}(i,j) > d$ and $\sigma(A) \subset [a,b]$, then*
$$\left(\mathbb{V}\left[\mathcal{T}_d^{\mathcal{N}}(f(A))\right]\right)^{\frac{1}{2}} \leq 2\sqrt{n}\, c\, E_d(f,[a,b]).$$

Proof. The variance can be bounded in general by $\mathbb{V}\left[\mathcal{T}_d^{\mathcal{N}}(f(A))\right] \leq \mathbb{V}\left[\mathcal{T}_d^{\mathbb{1}}(f(A))\right]$ in view of (4.22), where $\mathbb{1} = (1,1,\ldots,1)$ (i.e., only one sample per color is used). In all the inequalities (4.35), (4.36), and (4.38), the right-hand side has the form $n_\ell \cdot \phi(d)$, where $\phi(d)$ does not depend on ℓ. Hence, we get
$$\mathbb{V}\left[\mathcal{T}_d^{\mathbb{1}}(f(A))\right] = \sum_{\ell=1}^{n_\mathrm{c}} V_\ell \leq \sum_{\ell=1}^{n_\mathrm{c}} n_\ell\, \phi(d) = n\, \phi(d).$$
Thus, by replacing $\phi(d)$ with the correct formula, we conclude. □

Remark 4.7. Let us compare Theorem 4.14 with the error bounds of Section 4.2.2. Under suitable assumptions, the behavior of the bounds is the same with respect to d. This is easy to see for the banded and constant sign case, while for the lattice sparsity pattern note that $\sqrt{\mathrm{Li}_{1-D}(q^{2d})} = \mathcal{O}(q^d)$ for large d, as discussed in Remark 4.3. On the other hand, the factor n present in the deterministic case is replaced by $\sqrt{n}$ in the stochastic probing case. This can yield a huge gain, especially for large-scale problems.

4.3.5 Matrix Functions with Constant Sign Patterns

Here we characterize classes of functions and matrices for which $[f(A)]_{ij}$ has a fixed sign for $\mathrm{d}(i,j) \geq d$, i.e., situations in which Proposition 4.13 applies. We first state a classical result from [42] which treats the case $d = 0$, i.e., it gives conditions such that $f(A) \geq 0$ for all i and j. Recall that a matrix $A \in \mathbb{R}^{n\times n}$ is called a (possibly singular) *M-matrix* if it can be written as $A = \theta I - B$, where all entries in B are non-negative, $B \geq 0$, and the spectral radius satisfies $\rho(B) \leq \theta$. If A and thus B is symmetric, $\rho(B) = \|B\|_2 = \lambda_{\max}(B)$, the largest eigenvalue of B; see, e.g., [23].

Lemma 4.15 ([42])**.** *Let $A = \theta I - B \in \mathbb{R}^{n\times n}$ with $B \geq 0$ and $\|B\|_2 \leq \theta$ be a symmetric M-matrix. Suppose that $f(x)$ is continuous over $[\theta - \|B\|_2, \theta + \|B\|_2]$, analytic over $(\theta - \|B\|_2, \theta + \|B\|_2 + \varepsilon)$, for some $\varepsilon > 0$, and that*
$$(-1)^k f^{(k)}(\theta) \geq 0 \quad \textit{for all } k \geq 0. \tag{4.41}$$
Then $f(A) \geq 0$.

This gives rise to the following proposition, where we assume that the derivatives of f alternate their sign only beyond the dth derivative.

Proposition 4.16. *Let A and f be as in Lemma 4.15, except that condition* (4.41) *is replaced by*

$$(-1)^k f^{(k)}(\theta) \geq 0 \quad \text{for } k \geq d, \tag{4.42}$$

where d is a positive integer. Then

$$[f(A)]_{ij} \geq 0 \quad \text{for } \mathrm{d}(i,j) \geq d.$$

Proof. We consider the Taylor series centered at θ, writing it in terms of $x \in (\theta - \|B\|, \theta + \|B\|)$. Then

$$f(x) = \sum_{k=0}^{\infty} c_k x^k = p_{d-1}(x) + g(x),$$

where $p_{d-1}(x) = \sum_{k=0}^{d-1} c_k x^k$ and $g(x) = \sum_{k=d}^{\infty} c_k x^k$. Since $[A^k]_{ij} = 0$ for $k < \mathrm{d}(i,j)$, we have that $[p_{d-1}(A)]_{ij} = 0$ if $d \leq \mathrm{d}(i,j)$. Moreover,

$$g^{(k)}(\theta) = \begin{cases} 0 & \text{if } k < d \\ f^{(k)}(\theta) & \text{if } k \geq d \end{cases},$$

so from Lemma 4.15 we get that $g(A) \geq 0$. Finally,

$$[f(A)]_{ij} = [p_{d-1}(A)]_{ij} + [g(A)]_{ij} = [g(A)]_{ij} \geq 0$$

for $\mathrm{d}(i,j) \geq d$. □

Of course, if the assumptions on f in Proposition 4.16 hold for $-f$ instead of f, we obtain $[f(A)]_{ij} \leq 0$ for $\mathrm{d}(i,j) \geq d$.

Remark 4.8. Lemma 4.15 or Proposition 4.16 can be applied for the following functions f of a symmetric M-matrix A:

1. Completely monotone functions, i.e., function whose derivatives satisfy

   $$(-1)^k f^{(k)}(x) \geq 0, \quad k \geq 0,$$

   for all $x \in (0, \infty)$. In that case, Lemma 4.15 implies that $f(A) \geq 0$.

2. Bernstein functions, i.e., nonnegative functions f on $(0, \infty)$ with completely monotone derivative f'. By applying Proposition 4.16 to $-f$ we get $[f(A)]_{ij} \leq 0$ for $\mathrm{d}(i,j) \geq 1$, i.e. for $i \neq j$.

3. The entropy function $f(x) = -x \log x$. In fact, for $\theta > 0$ we have

   $$f^{(k)}(\theta) = (-1)^{k-1} \theta^{1-k} (k-2)!, \quad k \geq 2.$$

   If $A = \theta I - B$ with $B \geq 0$ is an M-matrix, by applying Proposition 4.16 to $-f$ we get $[f(A)]_{ij} \leq 0$ for $\mathrm{d}(i,j) \geq 2$. If $\theta \leq \exp(-1)$ we also get $f'(\theta) = -\log\theta - 1 \geq 0$, so that in this case we even have $[f(A)]_{ij} \leq 0$ for $\mathrm{d}(i,j) \geq 1$, i.e., for $i \neq j$.

4. Fractional powers with exponent larger than 1, i.e., $f(x) = x^\alpha$ with $\alpha > 1$. If $d = \lceil \alpha \rceil$, we get

$$\begin{cases} f^{(k)}(x) & \geq 0 \quad \text{for } k < d, \\ (-1)^{k+d} f^{(k)}(x) & \geq 0 \quad \text{for } k \geq d. \end{cases}$$

   Hence, if d is even, we get $[f(A)]_{ij} \geq 0$ for $\mathrm{d}(i,j) \geq d$, and if d is odd we get $[f(A)]_{ij} \leq 0$ for $\mathrm{d}(i,j) \geq d$.

5. $f(x) = x \exp(-x)$, for which

$$f^{(k)}(x) = (-1)^k (x-k) \exp(-x), \quad k \geq 0.$$

   For $d = \lceil \theta \rceil$, we get that $(-1)^{k+d+1} f^{(k)}(\theta) \geq 0$ for $k \geq d$. Therefore, if d is even, we get $[f(A)]_{ij} \leq 0$ for $\mathrm{d}(i,j) \geq d$, and if d is odd we get $[f(A)]_{ij} \geq 0$ for $\mathrm{d}(i,j) \geq d$.

In all situations just outlined, we also have that stochastic probing with only one sample per color is never worse than deterministic probing (which has the same computational cost).

Proposition 4.17. *Assume that A and f satisfy the assumptions of Proposition 4.13 with $C_1, \ldots, C_{n_c}$ a distance-d coloring of $\mathcal{G}(A)$. Let $\mathcal{T}_d^{\mathbb{1}}(f(A))$ the stochastic probing approximation of the trace using just one Rademacher vector for each color, i.e., $\mathcal{N} = (1, \ldots, 1)$ and by $\mathcal{T}_d(f(A))$ the deterministic probing approximation defined in* (4.8) *with respect to the same distance-d coloring. Then*

$$|\,\mathcal{T}_d^{\mathbb{1}}(f(A)) - \mathrm{Tr}(f(A))| \leq |\,\mathcal{T}_d(f(A)) - \mathrm{Tr}(f(A))|.$$

Proof. We use the notation and the result from Proposition 4.7 to obtain

$$\big|\,\mathcal{T}_d^{\mathbb{1}}(f(A)) - \mathrm{Tr}(f(A))\big| = \Big|\sum_{\substack{i,j \in C_\ell \\ i \neq j}} X_i X_j [f(A)]_{ij}\Big| \leq \sum_{\substack{i,j \in C_\ell \\ i \neq j}} |[f(A)]_{ij}|,$$

and the assumptions of Proposition 4.13 together with (4.9) to conclude,

$$\sum_{\substack{i,j \in C_\ell \\ i \neq j}} |[f(A)]_{ij}| = \Big|\sum_{\substack{i,j \in C_\ell \\ i \neq j}} [f(A)]_{ij}\Big| = |\,\mathcal{T}_d(f(A)) - \mathrm{Tr}(f(A))|.$$

□

Note that this result assumes constant sign in $[f(A)]_{ij}$ only for $\mathrm{d}(i,j) \geq d$. For special cases and $d = 1$ the result has been observed before, for example in [5, Section 2.3] for the specific case of the log-determinant of certain precision matrices.

Remark 4.9. One might think that the stochastic probing method that we discussed in this chapter could be further enhanced by using Hutch++ or XTrace instead of standard Hutchinson on each of the colors C_ℓ where we estimate $\mathrm{Tr}([f(A)]_{C_\ell})$. This, however, turns out to be counter-productive, since these methods then invest too much effort in computing the low

rank approximation for each color block. What we would need for the advanced estimators to be efficient is that, for each color C_ℓ, the block $[f(A)]_{C_\ell}$ can be well approximated by a matrix with small rank compared to the block size n_ℓ. The coloring approach aims at making the off-diagonal entries of $[f(A)]_{C_\ell}$ small, so if the diagonal elements are not comparably small, there cannot be a low rank approximation to $[f(A)]_{C_\ell}$. And even if there is a rapid decay in the eigenvalues of $f(A)$, we can at best expect to have a similar decay in the eigenvalues of *each* color block $[f(A)]_{C_\ell}, l = 1, \ldots, n_\mathrm{c}$. But this means that, compared to no probing, we have an effort which is n_c times as large to invest before we retrieve a good low rank approximation for each of the color blocks.

4.4 Numerical Experiments

We now illustrate our theoretical analysis with several numerical examples. All experiments are done in MATLAB R2020b on a computer with Intel® Core™ i7-7700HQ CPU and 16 GB RAM.

4.4.1 Scaling with the Size for Random Geometric Graphs

We start with a series of experiments illustrating the results in Section 4.3.4 on the more favorable scaling of the error of stochastic vs. deterministic probing. The matrices used are Laplacians L or adjacency matrices A of random geometric graphs with n nodes. These were obtained via the command `random_geometric_graph(n,radius)` from the package `NetworkX` [59] in Python with n ranging from 200 to $5,000$ in steps of 200. In order to keep similar properties for all the matrices, in particular to keep the number n_c of colors of the distance-d coloring similar for all n, we used $\texttt{radius} = \sqrt{\frac{\log n}{\pi n}}$ for all n, motivated by the results in [85, Section 6].

We test four different matrix functions: the shifted inverse $f_1(L) = (L + 2I)^{-1}$, the square root $f_2(L) = \sqrt{L}$, the exponential $f_3(L) = \exp(-10L)$, where L is the graph Laplacian, and the absolute value $f_4(A) = |A|$ of the adjacency matrix A. The trace of $|A|$ is the *graph energy*; cf. [75]. Graph Laplacians are M-matrices, since they can be written as $\theta I - B$ with θ the maximum degree in the graph and $B \geq 0$ with $\rho(B) = \theta$; cf. [23, Chapter 6]. Hence, in view of Remark 4.8 the assumptions of Proposition 4.13 are satisfied for the functions $f_1(L), f_2(L), f_3(L)$, for all distance-d colorings. On the other hand, we have no information on the signs in $f_4(A)$.

In this experiment, we use a distance-3 coloring for all graphs, which is computed with Algorithm 5. With f a generic symbol for one of the functions $f_1, \ldots, f_4$ and $M^{(n)}$ a generic symbol for the Laplacian or the adjacency matrix of the random geometric graph with n nodes,

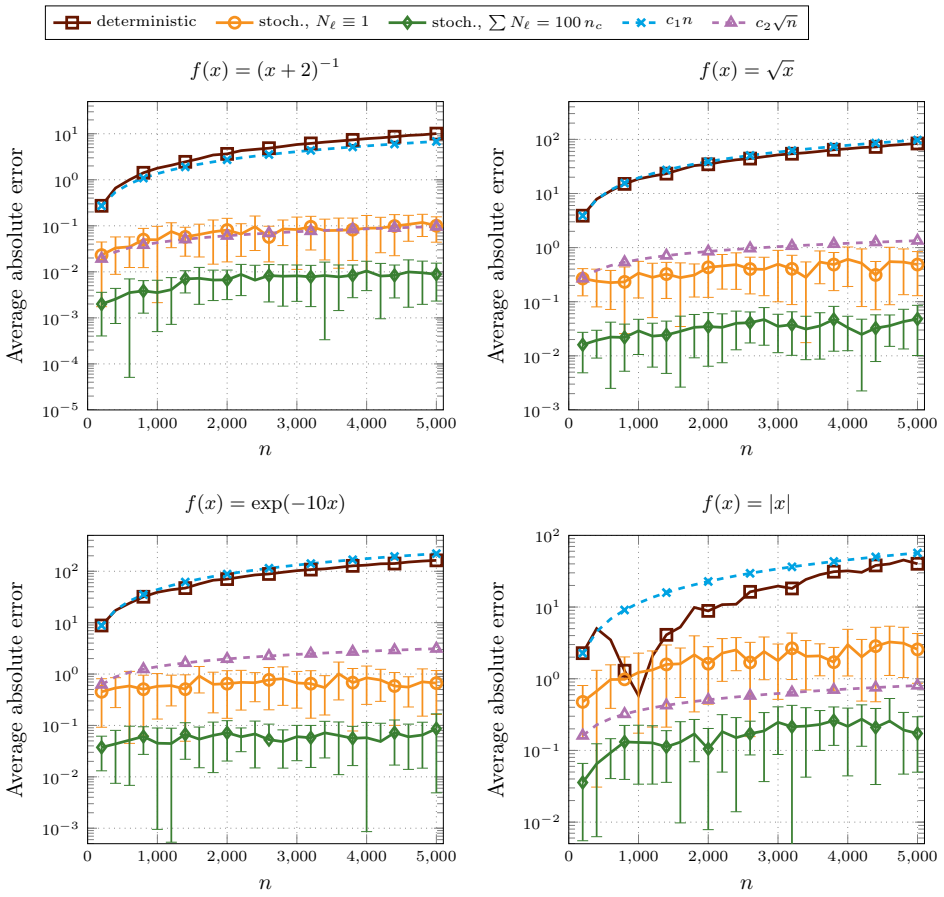

Figure 4.1: Absolute errors and asymptotic behavior for increasing n for the three methods in (4.43). For stochastic methods, we show the average error over 20 runs, together with a confidence interval.

Figure 4.1 reports the quantities

$$
\begin{aligned}
&|\operatorname{Tr}(f(M^{(n)})) - \mathcal{T}_d(f(M^{(n)}))| && \text{(deterministic probing)},\\
&|\operatorname{Tr}(f(M^{(n)})) - \mathcal{T}_d^{\mathbb{1}}(f(M^{(n)}))| && \text{(stochastic probing, one vector per color)}\\
&|\operatorname{Tr}(f(M^{(n)})) - \mathcal{T}_d^{\mathcal{N}}(f(M^{(n)}))| && \text{(stochastic probing with } \mathcal{N} = (N_1, \dots, N_{n_c}),\\
& && \text{where } N_\ell = \nu\sqrt{|C_\ell|} \text{ and } \textstyle\sum_{\ell=1}^{n_c} N_\ell = 100\, n_c).
\end{aligned}
\tag{4.43}
$$

The choice of N_ℓ in the third case is motivated by Remark 4.6, and We expect the error given by $\mathcal{T}_d^{\mathcal{N}}(f(M^{(n)}))$ to be smaller by one order of magnitude as compared to the error of $\mathcal{T}_d^{\mathbb{1}}(f(M^{(n)}))$. For the stochastic methods, we actually show an average of these errors over 20 runs, together with the confidence interval obtained by adding and subtracting the empirical standard deviation of the samplings. We also display the functions $g_1(n) = c_1 n$ and $g_2(n) = c_2\sqrt{n}$ with suitably chosen constants c_1 and c_2 which allow to easily appreciate the expected scaling behavior.

Figure 4.1 shows that stochastic probing with just one vector per color performs indeed better—and this by one to two orders of magnitude—than deterministic probing. Moreover, the expected decrease by one order of magnitude for stochastic probing with N_ℓ samples per color is clearly visible in all four examples. In addition, for f_1, f_2 and f_4 we see that the error of stochastic probing (with one vector) scales with the square root of the dimension whereas it scales linearly with the dimension for deterministic probing. This is exactly the scaling of the bounds on the error (or the variance) that we obtained in our analysis in Section 4.3.4, and this analysis applies to f_1, f_2 and f_3. For f_1 and f_2 the actual error is very close to the bounds, whereas for f_3 the bounds are more pessimistic: the accuracy of stochastic probing appears to almost not depend on the dimension, whereas the growth of the error for deterministic probing appears to be less than linear. For the graph energy, i.e., function f_4, the analysis of Section 4.3.4 cannot be applied. Interestingly, though, we observe again linear growth of the error for deterministic probing and growth with the square root of the dimension for stochastic probing.

4.4.2 Scaling with the Distance

In the next experiment, we analyze the scaling with the distance d. As a test graph, we consider the road network DC of the District of Columbia[1], more specifically its biggest connected component which has $n = 9,522$ nodes. Road networks exhibit a large-world structure, hence we need only few colors for a distance-d coloring; see Remark 4.1. Again, we consider the functions $(L + 2I)^{-1}$, $\sqrt{L}$, $\exp(-10L)$ and $|A|$, where L and A are the graph Laplacian and the adjacency matrix, respectively. We again compare deterministic probing, stochastic probing with just one vector per color (i.e., n_c vectors in total) and stochastic probing with $\sum_{\ell=1}^{n_\mathrm{c}} N_\ell = 100\, n_\mathrm{c}$ with each N_ℓ proportional to $\sqrt{|C_\ell|}$ as motivated in Remark 4.6. This time, we report the average relative error over 20 runs, displayed together with the confidence interval obtained by adding and subtracting the empirical standard deviation of the samplings; cf. Figure 4.2.

As expected, for the first three functions the error for stochastic probing scales similarly as a function of d as for deterministic probing while being smaller by a few orders of magnitude. This is consistent with Remark 4.7. In the fourth example, the error in stochastic probing is smaller for the smaller values of d but becomes increasingly closer to the error of deterministic probing for larger values of d. Note that none of the assumptions of Theorem 4.14 hold in this case.

4.4.3 Comparison with Hutch++ and XTrace

In a last series of experiments, we compare stochastic probing with the other variance reduction techniques outlined in Section 4.1. Since these techniques use low-rank approximations of $f(A)$, their performance depends on the eigenvalue distribution in $f(A)$ and is heavily enhanced by an exponential decay, as shown by Theorem 4.3. We compare these methods with plain Hutchinson and with stochastic probing with distances $d = 1, 3, 5$. The test matrices

[1] available at `http://www.diag.uniroma1.it/challenge9/data/tiger/`

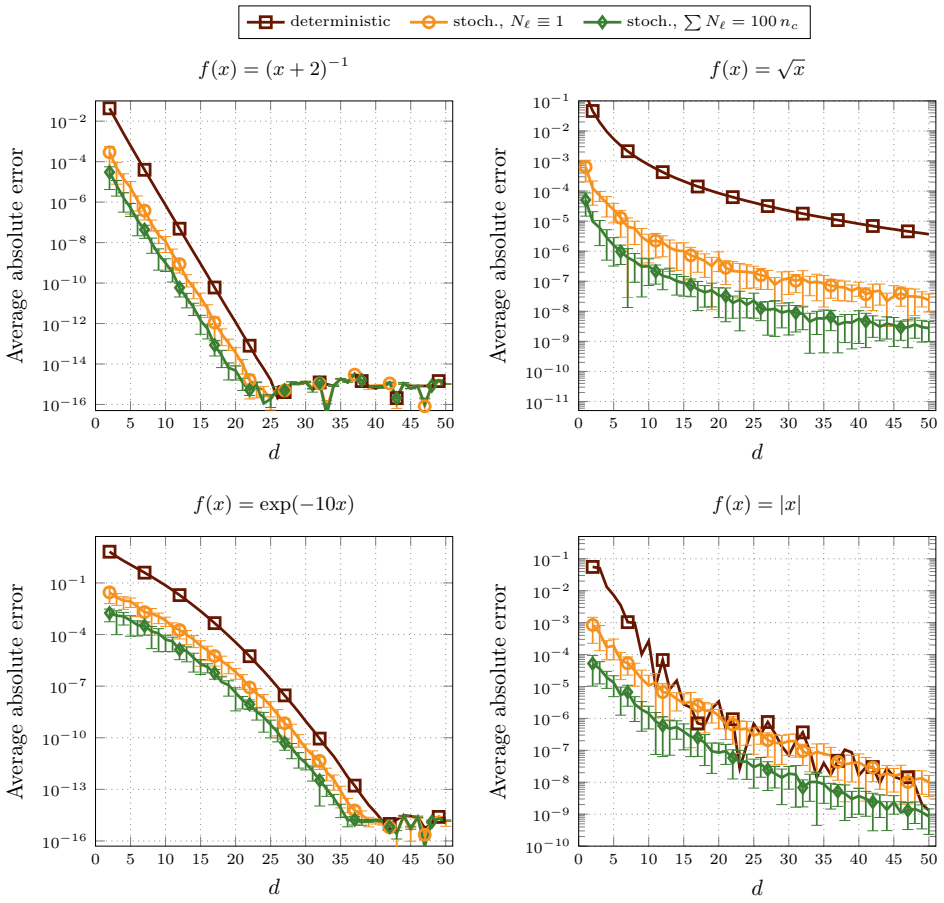

Figure 4.2: Absolute errors and asymptotic behavior for increasing d for the three methods in (4.43). For stochastic methods, we show the average error over 20 runs, together with a confidence interval.

are $f_1(L) = (L + 2I)^{-1}$, $f_2(L) = \sqrt{L}$ and $f_3(L) = \exp(-10L)$, where L is again the Laplacian L of the DC graph. For $d = 1, 3, 5$, the colorings computed with Algorithm 5 have $n_c = 4, 15, 35$ colors, respectively.

The largest 2500 eigenvalues of all three test matrices are depicted in the top left plot of Figure 4.3. The remaining three plots report the error of the different trace estimation methods as a function of the total number of matrix-vector multiplications. Since these evaluations are by far the most costly component in any of the methods, their number quite accurately reflects the computational cost. The essence of Figure 4.3 is that for situations where the eigenvalues do not show a pronounced decay (functions f_1 and f_2), the techniques using low rank approximations perform significantly worse than stochastic probing. The situation is reversed, though, when the eigenvalues decay rapidly, as in $f_3(L)$. The plots for the shifted inverse and the square root also illustrate that using a larger distance in the coloring gives more accurate results in stochastic probing for the same cost, an observation which should

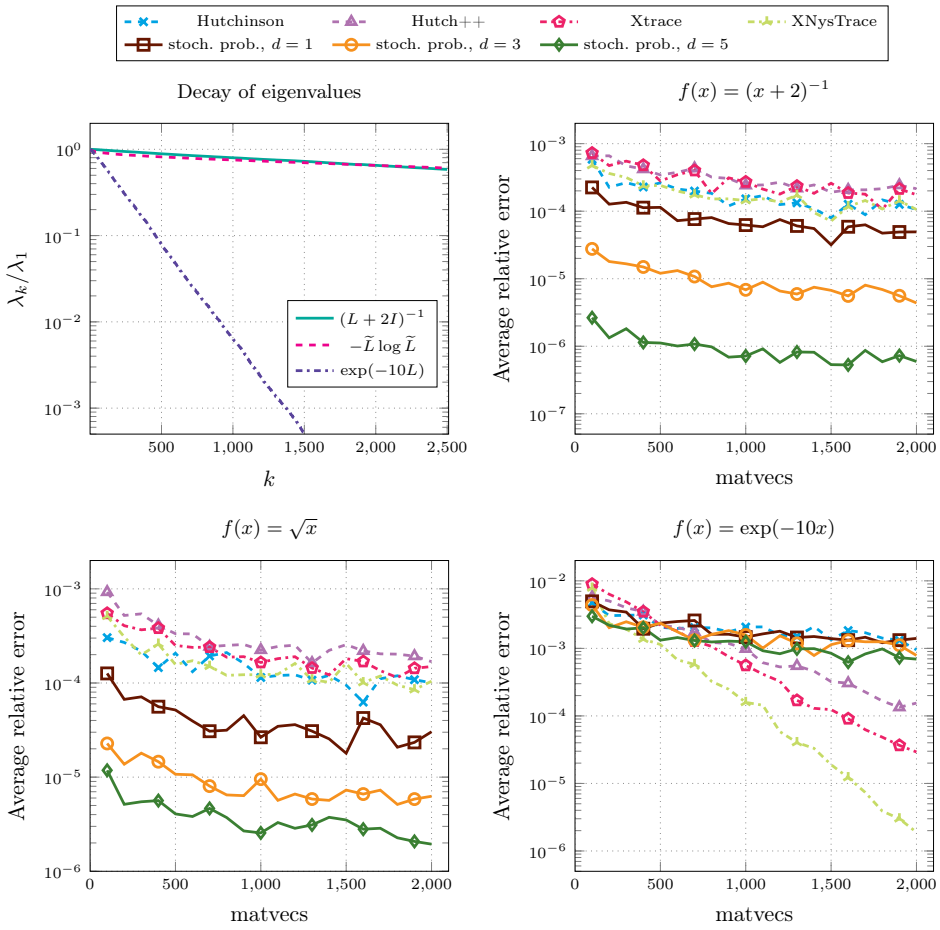

Figure 4.3: Comparison of the relative accuracy of different trace estimators as a function of matvecs $f(L)\boldsymbol{x}$. For each method, we show the average error over 20 runs.

not come as a big surprise.

4.5 Conclusions and Further Developments

We analyzed the performance of stochastic probing methods for estimating the trace of functions of sparse matrices. We derived formulas for the variance and tail bounds of this combination between a stochastic approach (Hutchinson's estimator) and probing estimators. For some common cases, for instance when the matrix argument is banded, when the associated graph is a regular grid, or when the entries $[f(A)]_{ij}$ have constant sign for $\mathrm{d}(i,j) > d$, we derived bounds on the variance showing that the error scales on average with the square root of the size, in contrast to the linear scaling exhibited by the deterministic method. Our theory is validated by several numerical experiments where we observed the scaling of the error with the size and compared the performance with other known estimators, indicating

when stochastic probing can be the method of choice. Further developments could consist on finding more efficient ways to compute the coloring, depending on the structure of the graph associated with the matrix argument.

Chapter 5

Von Neumann Entropy and its Computation

In this chapter we consider the computation of the von Neumann entropy, which is defined as the trace of the matrix function $f(A) = -A \log A$. We analyze its approximation through trace estimators and deal with the computation of quadratic forms and matrix-vector products. In Section 5.1, we study the analytical properties of the function $f(x) = -x \log x$. In Section 5.2, we discuss the approximation of the entropy by means of the probing approach. In Section 5.3, we analyze the computation of matrix-vector multiplications and quadratic forms via polynomial and rational Krylov methods, by developing a pole selection strategy and a posteriori error bounds. In Section 5.4, we give an overview of the algorithms used for the numerical experiments of Section 5.5, where we compare the described methods. This chapter is based in part on [18].

5.1 Properties of the von Neumann Entropy

A *density matrix* is a positive semi-definite linear operator ρ with unit trace acting on a complex Hilbert space. In quantum mechanics, density operators describe mixed states of quantum mechanical systems, which are convex combinations of *pure states* (i.e., density matrices with rank 1). Here we consider only finite dimensional Hilbert spaces, so ρ is an $n \times n$ Hermitian matrix. For simplicity, we focus on the real symmetric case.

The *von Neumann entropy* of a system described by the density matrix ρ [103] is given by

$$S(\rho) = -\sum_{\lambda \in \sigma(\rho)} \lambda \log(\lambda) = -\operatorname{Tr}(\rho \log \rho), \tag{5.1}$$

under the convention that $0 \cdot \log(0) = 0$. Here $\log(x)$ is the natural logarithm. Note that in the literature the entropy is sometimes defined using $\log_2(x)$ instead, but, since $\log_2(x) = \log(x)/\log(2)$, the two definitions are equivalent up to a constant scaling factor. We also mention that in the original definition the entropy includes the factor κ_B (the Boltzmann constant), which we omit.

The function $f(x) = -x \log x$ is defined over the spectrum of any Hermitian positive semidefinite matrix A. Hence, even though it has no physical meaning, the quantity $S(A)$ is defined without the additional hypothesis of unit trace. Moreover, if ρ is given in the form $\rho = \gamma A$ with $\gamma > 0$, we have the relation

$$S(\rho) = -\gamma \operatorname{Tr}(A \log(A)) - \gamma \log(\gamma) \operatorname{Tr}(A) = \gamma S(A) - \gamma \log(\gamma) \operatorname{Tr}(A). \tag{5.2}$$

Hence, we can easily recover $S(\rho)$ from $S(A)$.

In quantum mechanics, the von Neumann entropy of a density matrix gives a measure of how far the system is from being in a pure state, and therefore it measures the uncertainty in our knowledge of the state of the system. For any density matrix of size n, we have $0 \leq S(\rho) \leq \log(n)$. Moreover, $S(\rho) = 0$ if and only if ρ is a pure state (i.e., ρ is a rank 1 matrix with one eigenvalue equal to 1 and all the others equal to 0) and $S(\rho) = \log(n)$ if and only if $\rho = \frac{1}{n} I$ [103, 104].

The von Neumann entropy also has applications in network theory. Let $\mathcal{L}$ be the Laplacian of an undirected graph. Then $\mathcal{L}$ is a singular positive semidefinite matrix and the eigenvalue 0 is simple if and only if $\mathcal{G}$ is connected, in which case the associated one-dimensional eigenspace is spanned by $\mathbf{1}$. Given the density matrix $\rho = \mathcal{L}/\operatorname{Tr}(\mathcal{L})$, the von Neumann entropy of $\mathcal{G}$ is defined as $S(\rho)$ [25, 27].

It should be mentioned that, strictly speaking, the graph entropy defined in this manner is not a "true" entropy, since it does not satisfy the sub-additivity requirement. In other words, the graph entropy could decrease when an edge is added to the graph, see [34]. For this reason, different notions of graph entropy have been proposed in the literature; see, e.g., [51, 52]. These entropies still have the form of a von Neumann entropy, $S = -\operatorname{Tr}(\rho \log \rho)$, but with a different definition of the density matrix ρ which, however, is still expressed as a function of a matrix (Hamiltonian) associated to the graph.

Throughout this chapter, we use the notation A to denote general matrices, and ρ to denote density matrices, i.e., $\operatorname{Tr}(\rho) = 1$. Most of our results are applicable in general for symmetric positive semidefinite matrices and not only to density matrices.

A straightforward way to compute the von Neumann entropy is by diagonalization. However, this approach is unfeasible when the dimension is very large. Here we propose some approaches to compute approximations to the von Neumann entropy based on the trace estimation of matrix functions, and propose methods for the computation of the quadratic forms associated with the estimation scheme via Krylov methods.

5.1.1 Integral Representation and Polynomial Approximation

A quantity that plays an important role in our analysis is the error of the best uniform polynomial approximation of a continuous function, which we already introduced in detail in Section 3.2. Here we briefly recall some notations. Let Π_k be the set of all polynomials with degree at most k. For a continuous function $f : [a, b] \to \mathbb{R}$, the error of the best uniform polynomial approximation in Π_k is

$$E_k(f, [a, b]) = \min_{p \in \Pi_k} \max_{x \in [a,b]} |f(x) - p(x)|. \tag{5.3}$$

The function $f(x) = x\log(x)$ is not analytic on any neighborhood of 0, so we cannot expect to find a geometric decay as for the case of the inverse (cf., Theorem 3.3) or for general analytic functions (cf., Theorem 3.6) if we consider $[0, b]$, $b > 0$, as the interval of definition. However, since f is continuous, by the Weierstrass Approximation Theorem $E_k(f, [0, b])$ must go to 0 as $k \to \infty$. A more precise estimate is the following [107], derived by computing the coefficients of the Chebyshev expansion of $f(x)$:

$$E_k(f, [0, b]) \leq \frac{b}{2k(k+1)} \quad \text{for all } k \geq 1. \tag{5.4}$$

This shows that the decay rate of the error is algebraic in k. Our approach is based on an integral representation and leads to sharper bounds.

Recall that a Cauchy-Stieltjes function has the form (3.43). An example is given by the function $\log(1+z)/z$ that has the expression

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

It is easy to check that the above identity also holds for $z \in (-1, 0)$. With the change of variable $x = 1+z$ and some simple rearrangements, we get the following integral expression for the entropy:

$$-x\log(x) = \int_1^\infty \frac{x(1-x)}{s(s+x-1)} \mathrm{d}s, \qquad x \in [0, 1]. \tag{5.5}$$

With the additional change of variable $s = t + 1$, we can rewrite (5.5) in the form

$$-x\log(x) = x(1-x)\int_0^\infty \frac{1}{(t+x)(t+1)} \mathrm{d}t, \qquad x \in [0, 1]. \tag{5.6}$$

Note that the above identities also hold for $x = 0$ and $x = 1$, because of the factor $x(1-x)$ in front of the integral. This shows that although the entropy function $-x\log(x)$ is not itself a Cauchy-Stieltjes function, we can recognize a factor of the form (3.43) in its integral representation. This observation will be important later for the selection of poles in a rational Krylov method; see Section 5.3.2.

Since the integral representation (5.6) has the form (3.7), we can apply Lemma 3.9 to derive a new bound for $E_k(x\log x, [a, b])$.

Theorem 5.1. *Let $0 \leq a < b$ and $\gamma = a/b$. We have*

$$E_k(x\log(x), [a, b]) \leq b\,(1-\sqrt{\gamma})\frac{1+\gamma+2k\sqrt{\gamma}}{4(k^2-1)}\left(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right)^k, \tag{5.7}$$

for all $k \geq 2$.

Proof. Let $f(x) = x\log(x)$. Notice that $f(x) = bf(b^{-1}x) + \log(b)x$, so

$$E_k(f(x), [a, b]) = E_k(bf(b^{-1}x), [a, b]) = b\,E_k(f(x), [\gamma, 1])$$

since $k \geq 2$ and we can ignore terms of degree 1 for the polynomial approximation. For $x \in [\gamma, 1]$ we can use the representation (5.6). Let $g_t(x) := \frac{x(1-x)}{(1+t)(x+t)}$ be the integrand in (5.6) for all $t > 0$. Note that $g_t(x)$ is a continuous fuction in the variables $(t, x) \in (0, \infty) \times [a, b]$, so we can apply Lemma 3.9. Then we can write $g_t(x)$ as

$$g_t(x) = \frac{x}{x+t} - \frac{x}{1+t} = 1 - \frac{t}{x+t} - \frac{x}{1+t}, \tag{5.8}$$

and since $k \geq 2$ and $1 - \frac{x}{1+t}$ has degree 1, we get that

$$E_k(g_t(x), [\gamma, 1]) = E_k(t/(x+t), [\gamma, 1]) = tE_k(1/x, [\gamma + t, 1 + t]).$$

Hence, by using Theorem 3.3 and Lemma 3.9, we get

$$E_k(x \log(x), [a, b]) \leq b \int_0^\infty \frac{t\left(\sqrt{\kappa(t)} + 1\right)^2}{2(1+t)} \left(\frac{\sqrt{\kappa(t)} - 1}{\sqrt{\kappa(t)} + 1}\right)^{k+1} \mathrm{d}t, \tag{5.9}$$

where $\kappa(t) = (1+t)/(\gamma+t)$. In order to bound the integral, consider the identities

$$\frac{\sqrt{\kappa(t)} - 1}{\sqrt{\kappa(t)} + 1} = \frac{\left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^2}{1-\gamma}, \quad \left(\sqrt{\kappa(t)} + 1\right)^2 = \frac{\left(\sqrt{1+t} + \sqrt{\gamma+t}\right)^2}{\gamma+t}.$$

We have

$$\begin{aligned}
&\int_0^\infty \frac{t\left(1 + \sqrt{\kappa(t)}\right)^2}{2(1+t)} \left(\frac{\sqrt{\kappa(t)} - 1}{\sqrt{\kappa(t)} + 1}\right)^{k+1} \mathrm{d}t \\
&= \frac{1}{2} \int_0^\infty \frac{t}{(\gamma+t)(1+t)} \cdot \frac{\left(\sqrt{1+t} + \sqrt{\gamma+t}\right)^2 \left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^{2k+2}}{(1-\gamma)^{k+1}} \mathrm{d}t \\
&\leq \frac{1}{2(1-\gamma)^{k-1}} \int_0^\infty \left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^{2k} \mathrm{d}t.
\end{aligned} \tag{5.10}$$

By checking the derivative, it can be shown that

$$F(t) = \frac{\sqrt{1+t}\sqrt{\gamma+t}\left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^{2k} \left(\frac{\left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^4}{2k+2} - \frac{(1-\gamma)^2}{2k-2}\right)}{2\left(\sqrt{1+t}\sqrt{\gamma+t} - \gamma - t\right)\left(1 + t - \sqrt{1+t}\sqrt{\gamma+t}\right)}$$

is an antiderivative of $\left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^{2k}$. Since $\lim_{t\to\infty} F(t) = 0$, we deduce that

$$\begin{aligned}
\int_0^\infty \left(\sqrt{1+t} - \sqrt{\gamma+t}\right)^{2k} \mathrm{d}t &= \frac{\sqrt{\gamma}\left(1 - \sqrt{\gamma}\right)^{2k} \left(\frac{(1-\gamma)^2}{2k-2} - \frac{\left(1-\sqrt{\gamma}\right)^4}{2k+2}\right)}{2\left(1 - \sqrt{\gamma}\right)\left(\sqrt{\gamma} - \gamma\right)} \\
&= 2(1 - \sqrt{\gamma})^{2k} \frac{1 + \gamma + 2\sqrt{\gamma}\, k}{2(k^2 - 1)}.
\end{aligned}$$

This, combined with (5.10), concludes the proof. □

Remark 5.1. The result of Theorem 5.1 is an improvement of (5.4). When $a > 0$, the right-hand side in (5.7) is the product of an algebraic and a geometric factor, so the decay is asymptotically faster. When $a = 0$, we have $\gamma = 0$ and (5.7) becomes

$$E_k(x \log(x), [0, b]) \leq \frac{b}{4(k^2 - 1)},$$

which is better than (5.4) for all $k \geq 2$. The new bound (5.7) is compared with (5.4) in Figure 5.1.

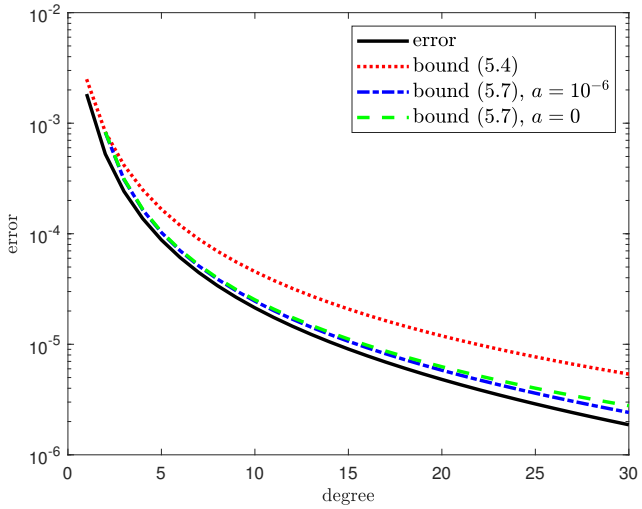

Figure 5.1: Comparison of the bounds (5.4) and (5.7) with the error of the polynomial approximation of the entropy function $x \log x$ on the interval $[10^{-6}, 10^{-2}]$.

5.2 Computation via Deterministic Probing

In this section we consider the deterministic probing approach, analyzed in [47] and reviewed in Section 4.2, for the computation of $S(A) = \mathrm{Tr}(f(A))$, where A is a symmetric positive semidefinite matrix and $f(x) = -x \log x$. Recall that the probing estimation is associated with a distance-d coloring $C_1, \ldots, C_{n_c}$ of the graph $\mathcal{G}(A)$, and the probing estimator is given by

$$\mathcal{T}_d(f(A)) = \sum_{\ell=1}^{n_c} \boldsymbol{v}_\ell^T f(A) \boldsymbol{v}_\ell,$$

where $\boldsymbol{v}_\ell$ is the probing vector (4.7) associated with the color ℓ, $\ell = 1, \ldots, n_c$.

5.2.1 A Priori Error Bounds

We have the following error bound for the approximation of $S(A)$ via probing.

Corollary 5.2. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with $\sigma(A) \subset [a, b]$, $0 \leq a < b$, and let $\gamma = a/b$. Then*

$$|S(A) - \mathcal{T}_d(-A\log(A))| \leq n\, b\, (1-\sqrt{\gamma})\frac{1+\gamma+2d\sqrt{\gamma}}{2(d^2-1)}\left(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right)^d, \tag{5.11}$$

for all $d \geq 2$. In particular, if $a = 0$, we have

$$|S(A) - \mathcal{T}_d(-A\log(A))| \leq \frac{n\, b}{2(d^2-1)}, \tag{5.12}$$

for all $d \geq 2$.

Proof. The inequality (5.11) follows from Theorem 5.1 and Theorem 4.5. The inequality (5.12) follows from (5.11) with $\gamma = a/b = 0$. □

Remark 5.2. The bound (5.11) can be pessimistic in practice, especially for large values of d. We will see this in Example 5.3 and in Section 5.5.2. A priori bounds based on polynomial approximations often fail to catch the exact convergence behavior in many problems related to matrix functions, since a minimization problem over the spectrum of a matrix is relaxed to the whole spectral interval. This occurs in the convergence of polynomial Krylov methods [76, Section 5.6] and in the decay bounds on the entries of matrix functions (cf., Section 3.4). Also, the coloring we get via Algorithm 5 can return far more colors than needed for a distance-d coloring, and this can benefit the convergence in a way that is not predicted by the bound. We will see in Section 5.4.1 a more practical heuristic to predict the error with higher accuracy.

Example 5.3. Let us see how the bound and the convergence of the probing method perform in practice. We use the density matrix $\rho = \mathcal{L}/\operatorname{Tr}(\mathcal{L})$, where $\mathcal{L}$ is the Laplacian of the graph `minnesota` from the SuiteSparse Matrix Collection [32]. More precisely, we consider the biggest connected component whose graph Laplacian has size $n = 2640$ and 9244 nonzero entries.

For several values of d, we compute the approximation $\mathcal{T}_d(-\rho\log\rho)$ associated with two different distance-d colorings: the first is obtained by the greedy coloring (Algorithm 5) after sorting the nodes by descending degree, while for the second we use the reverse Cuthill-McKee algorithm to get a 67-banded matrix and then apply (4.10).

In Figure 5.2 we compare $\mathcal{T}_d(-\rho\log(\rho))$ with the value of $S(\rho)$ obtained by diagonalizing ρ, considered as exact. The left plot shows the error in terms of the number of colors (i.e. the number of probing vectors) associated with the distance-d colorings. In the right plot the errors are shown in terms of d together with bound (5.12), where $b = \lambda_{\max}(\rho)$.

For a fixed value of d, the coloring based on the bandwidth provides a smaller error than the greedy one, but it also uses a larger number of colors. Indeed, we can see from the left plot that with the same computational effort the greedy algorithm obtains a smaller error. On the right we can observe that bound (5.12) is close to the error given by the greedy coloring for small values of d, while it fails to catch the convergence behavior for large values of d.

For the graph entropy, i.e. when $\rho = \mathcal{L}/\operatorname{Tr}(\mathcal{L})$ and $\mathcal{L}$ is a graph Laplacian of a graph $\mathcal{G}$, we can show that $-\mathcal{T}_d(\rho\log\rho)$ is a lower bound for $S(\rho)$. In the proof we use the fact that ρ is a symmetric M-matrix and Proposition 4.16.

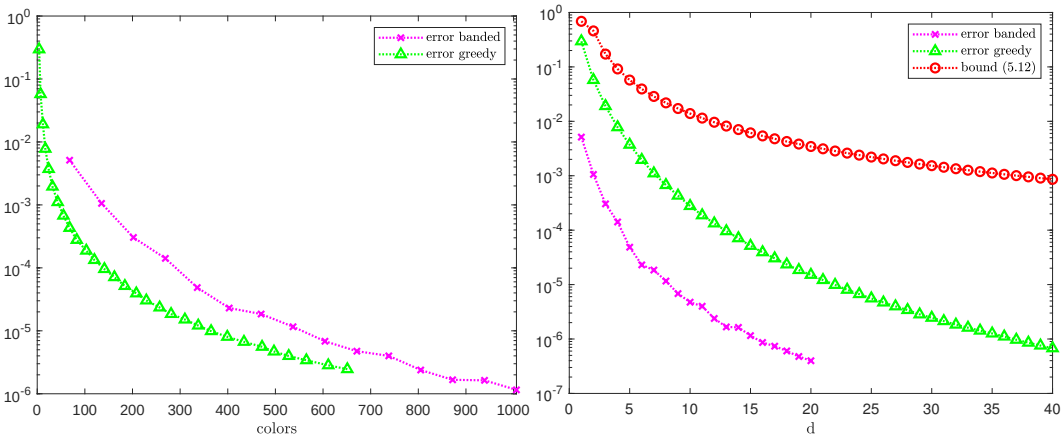

Figure 5.2: Absolute errors of the probing approximation of $S(\rho)$ where the distance-d coloring is obtained either by the greedy procedure (Algorithm 5) or with the reverse Cuthill-McKee algorithm and the coloring (4.10) for banded matrices. On the left the abscissa represents the number of colors used for the coloring. On the right the errors are compared with bound (5.12) in terms of d.

Proposition 5.3. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric M-matrix, and let $\mathcal{T}_d(-A\log(A))$ be the approximation* (4.8) *of $S(A)$ induced by a distance-d coloring of $\mathcal{G}(A)$ with $d \geq 1$. Then $\mathcal{T}_d(-A\log(A)) \leq S(A)$.*

Proof. If $V_1, \ldots, V_{n_c}$ is the graph partitioning associated with a distance-d coloring, the error of the approximation can be written as

$$S(A) - \mathcal{T}_d(-A\log(A)) = -\sum_{\ell=1}^{n_c} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} [-A\log(A)]_{ij}; \tag{5.13}$$

see Section 4.2. By definition of a distance-d coloring, for all $i, j \in V_\ell$, $i \neq j$, we have $\mathrm{d}(i,j) \geq d+1 \geq 2$. Then, in view of Proposition 4.16 (see also Remark 4.8), the right-hand side of (5.13) is nonnegative. □

5.2.2 Density Matrices Expressed as Matrix Functions

In this chapter, we mainly consider the entropy of sparse density matrices. However, an important case is given by $\rho = g(H)$, where H is the Hamiltonian of a certain quantum system and g is a function defined on the spectrum of H. Notable examples are the Gibbs state [29, 104] and the Fermi-Dirac state [1]. Despite ρ being a dense matrix in general, a sparse structure of H implies that ρ exhibits decay properties and can be well approximated by a sparse matrix [11, 14]. Moreover, since $S(\rho) = -g(H)\log(g(H))$ one can apply the techniques described in Chapter 4 to the composition $-g(x)\log(g(x))$. Although the detailed investigation of this problem is outside the scope of this work, we give some insight and possible future directions.

Gibbs Entropy

Let $H \in \mathbb{C}^{N\times N}$ be the Hamiltonian of a certain system (i.e., a symmetric matrix). The *Gibbs state* [29, 104] associated with H is the density matrix

$$\rho_G = Z^{-1}\exp(-\beta H), \quad Z = \operatorname{Tr}(\exp(-\beta H)),$$

where β is an inverse temperature factor. Without loss of generality, we can assume that H is positive semidefinite, since

$$\frac{e^{-\beta H}}{\operatorname{Tr}(e^{-\beta H})} = \frac{e^{-\beta(H+\alpha I)}}{\operatorname{Tr}(e^{-\beta(H+\alpha I)})},$$

and we can choose α such that $H + \alpha I$ is positive semidefinite.

The *Gibbs entropy* of a system with Hamiltonian H is $S(\rho_G)$. The importance of this entropy comes from the following fact [104]: ρ_G has the maximal von Neumann entropy among all the density matrices with the same expectation. In other words:

$$S(\rho_G) = \max\{S(\rho) : \rho \text{ density matrix, } \operatorname{Tr}(\rho H) = \operatorname{Tr}(\rho_G H)\}.$$

Moreover, if H is the Laplacian of a certain graph, $S(\rho_G)$ is used as an alternative definition for the entropy of a graph; see [51, 52].

In view of (5.2), $S(\rho_G)$ can be written as follows:

$$S(\rho_G) = Z^{-1}S(e^{-\beta H}) + \log(Z^{-1}) = \frac{\operatorname{Tr}(-\beta H e^{-\beta H})}{\operatorname{Tr}(e^{-\beta H})} - \log(\operatorname{Tr}(e^{-\beta H})). \tag{5.14}$$

Hence, the problem reduces to compute $\operatorname{Tr}(e^{-\beta H})$ and $\operatorname{Tr}(-\beta H e^{-\beta H})$. These two matrix functions typically exhibit a fast (in particular, exponential) decay of the eigenvalues. In view of the experiments in Section 4.4, in which the computation of $\operatorname{Tr}(e^{-\beta H})$ is also present, we expect that stochastic estimators based on low rank approximations, such as Hutch++, XTrace, and XNysTrace, are able to accurately compute these quantities with low effort.

Fermi-Dirac Entropy

The expression (5.14) allows to reduce the problem to much simpler matrix functions when the density matrix is a Gibbs state. However, this is not the case in general.

Let us consider $\phi(x) = -g(x)\log(g(x))$, so that $S(g(H)) = \operatorname{Tr}(\phi(H))$. In view of Theorem 4.5, the error of the probing approximation can be bounded by using $E_k(\phi(x), [a, b])$. Since we don't have, in general, an integral representation of $\phi(x)$, we rely on Bernstein's Theorem 3.6 to get bounds. In order to simplify the analysis, we consider the case where $[a, b] = [-1, 1]$. The discussion in Section 3.2.2 shows that this is not restrictive.

Let us recall some properties of the complex logarithm. Its principal branch is analytic over the set

$$\mathcal{S} = \mathbb{C} \setminus \{z \in \mathbb{R} : z \leq 0\},$$

which is the complex plane without the real nonpositive numbers. We also have

$$\log(z) = \log(|z|) + \mathbf{i}\arg(z),$$

where $\arg(z)$ is the principal branch of the argument function. In particular, $|\arg(z)| \leq \pi$ for all $z \in \mathcal{S}$, and

$$|\log(z)| \leq \sqrt{\log(|z|)^2 + \pi^2} \leq |\log(|z|)| + \pi. \tag{5.15}$$

Bernstein's Theorem makes use of the notation $\mathcal{E}_\chi$ to denote an ellipse with foci at ± 1 and sum of semiaxes $\chi > 1$ [80, Theorem 73]. We have the following result.

Theorem 5.4. *Let g be continuous over $\mathcal{E}_\chi$ and analytic over the interior of $\mathcal{E}_\chi$. Suppose that $g(z) \in \mathcal{S}$ for all $z \in \mathcal{E}_\chi$. Let $\phi(z) = g(z)\log(g(z))$. Then*

$$E_k(\phi, [-1, 1]) \leq K(\chi)\left(\frac{1}{\chi}\right)^k, \quad K(\chi) = \max_{z \in \mathcal{E}_\chi} |\phi(z)| \tag{5.16}$$

for all $k \geq 0$. Moreover,

$$K(\chi) \leq |M(\chi)\log(M(\chi))| + \pi M(\chi), \quad M(\chi) = \max_{z \in \mathcal{E}_\chi} |g(z)|. \tag{5.17}$$

Proof. Since $\phi(z)$ is the composition of two analytic functions, it is analytic itself over the interior of $\mathcal{E}_\chi$. Hence, in view of Bernstein's Theorem 3.6 we get (5.16).

For (5.17), in view of (5.15), we get

$$|g(z)\log(g(z))| \leq |g(z)|(\log(|g(z)|) + \pi)| \leq |g(z)|\log(|g(z)|) + \pi|g(z)|.$$

By taking the maximum of $|g(z)|$, we conclude. □

Theorem 5.4 applies when $g(z)$ is the Fermi-Dirac function: $g(z) = f_{FD}(z) = (1 + e^{\beta(z-\mu)})^{-1}$.

Proposition 5.5. *Let $f_{FD}(z) = (1 + e^{\beta(z-\mu)})^{-1}$, where $\beta > 0$ and $\mu \in \mathbb{R}$. Let $1 < \chi < \overline{\chi}$, where $\overline{\chi}$ is defined in Theorem 3.20. Then $f_{FD}(z) \in \mathcal{S}$ for all $z \in \mathcal{E}_\chi$.*

Proof. Suppose that there exists $z^* \in \mathcal{E}_\chi$ such that $f_{FD}(z^*) \in \mathbb{R}$ and $f_{FD}(z^*) \leq 0$. Then, $e^{\beta(z^*-\mu)} \in \mathbb{R}$ and $e^{\beta(z^*-\mu)} < -1$. Since $e^{\beta(z^*-\mu)} \in \mathbb{R}$ and $e^{\beta(z^*-\mu)} < 0$, we have that $\mathrm{Imag}(z^*) = \frac{1}{\beta}(k + \frac{1}{2})\pi$ for some integer k. Moreover, since $|e^{\beta(z^*-\mu)}| = e^{\mathrm{Real}(\beta(z^*-\mu))} > 1$, we have that $\mathrm{Real}(z^*) > \mu$. Since $\mathcal{E}_\chi$ is convex and symmetric with respect to both the real and immaginary axes, this implies that $\alpha + \frac{\pi}{2}\boldsymbol{i} \in \mathcal{E}_\chi$ for all$|\alpha| \leq \mathrm{Real}(z^*)$. In particular, $\tilde{z} = \mu + \frac{\pi}{2}\boldsymbol{i} \in \mathcal{E}_\chi$. Since $\tilde{z}$ is a pole of $f_{FD}(z)$, we get a contradiction. □

Proposition 5.5 tell us that the composition $\phi(z) = g(z)\log(g(z))$ satisfies the hypotheses of Theorem 5.4 when $g(z) = f_{FD}(z)$. Moreover, the same holds for $g(z) = 1 - f_{FD}(z)$. This is important since the electronic entropy [1] of a system with Hamiltonian $H \in \mathbb{R}^{n\times n}$ is

$$\mathrm{Tr}(f_{FD}(H)\log(f_{FD}(H)) + (I - f_{FD}(H))\log(I - f_{FD}(H))).$$

In view of the previous results, the probing approach is applicable, in principle, for the computation of the electronic entropy. However, one still needs a way to compute the matrix-vector products and the quadratic forms with the matrix function $\phi(H) = g(H)\log(g(H))$. This can be the subject of future research.

5.3 Computation of Quadratic Forms with Krylov Methods

Although the content of this section has been revised by all the authors of [18], the original results Lemma 5.7, Proposition 5.8, Theorem 5.9, and all the content of Section 5.3.3, are to be attributed to Igor Simunec.

The approximate computation of $\mathrm{Tr}(f(A))$ with probing methods or stochastic trace estimators can be reduced to the computation of several quadratic forms with $f(A)$, i.e. expressions of the form $\boldsymbol{b}^T f(A)\boldsymbol{b}$. In this section, we briefly describe how they can be efficiently computed using polynomial and rational Krylov methods.

A *polynomial Krylov subspace* associated to A and $\boldsymbol{b}$ is given by

$$\mathcal{P}_m(A,\boldsymbol{b}) = \mathrm{range}\left\{\boldsymbol{b}, A\boldsymbol{b}, \dots, A^{m-1}\boldsymbol{b}\right\} = \{p(A)\boldsymbol{b} : p \in \Pi_{m-1}\}.$$

More generally, given a sequence of poles $\{\xi_j\}_{j\geq 1} \subset (\mathbb{C}\cup\{\infty\})\setminus\sigma(A)\cup\{0\}$, we can define a *rational Krylov subspace* as follows,

$$\mathcal{Q}_m(A,\boldsymbol{b}) = q_{m-1}(A)^{-1}\mathcal{P}_m(A,\boldsymbol{b}) = \left\{r(A)\boldsymbol{b} : r(z) = \frac{p_{m-1}(z)}{q_{m-1}(z)}, \text{with } p_{m-1} \in \Pi_{m-1}\right\}, \tag{5.18}$$

where $q_{m-1}(z) = \prod_{j=1}^{m-1}(1 - z/\xi_j)$. If all poles are equal to ∞, we have $q_{m-1}(z) \equiv 1$ and $\mathcal{Q}_m(A,\boldsymbol{b})$ coincides with the polynomial Krylov subspace $\mathcal{P}_m(A,\boldsymbol{b})$, so $\mathcal{P}_m(A,\boldsymbol{b})$ can be considered as a special case of $\mathcal{Q}_m(A,\boldsymbol{b})$. Note that this definition of q_{m-1} does not allow us to have poles $\xi_j = 0$; this can be fixed by changing the definition of q_{m-1} but it is not required in our case, since we are only going to use real negative poles and poles at ∞.

Let us denote by $V_m = [\boldsymbol{v}_1 \dots \boldsymbol{v}_m]$ a matrix with orthonormal columns that spans the Krylov subspace $\mathcal{Q}_m(A,\boldsymbol{b})$, and by $A_m = V_m^T A V_m$ the projection of A onto the subspace. We can then project the problem on $\mathcal{Q}_m(A,\boldsymbol{b})$ and approximate $\psi = \boldsymbol{b}^T f(A)\boldsymbol{b}$ in the following way,

$$\psi \approx \psi_m = \boldsymbol{b}^T V_m f(A_m) V_m^T \boldsymbol{b}.$$

If the basis V_m is constructed incrementally using the rational Arnoldi algorithm [92], we have $\boldsymbol{v}_1 = \boldsymbol{b}/\left\|\boldsymbol{b}\right\|_2$ and therefore

$$\psi_m = \left\|\boldsymbol{b}\right\|_2^2 \boldsymbol{e}_1^T f(A_m)\boldsymbol{e}_1. \tag{5.19}$$

Note that the approximation ψ_m is closely related to the rational Krylov approximation of $f(A)\boldsymbol{b}$, which is given by

$$f(A)\boldsymbol{b} \approx V_m f(A_m) V_m^T \boldsymbol{b} = \left\|\boldsymbol{b}\right\|_2 V_m f(A_m)\boldsymbol{e}_1, \tag{5.20}$$

and is known to converge with a rate determined by the quality of rational approximations of f, in view of the following result.

Proposition 5.6 (Corollary 3.4 in [58])**.** *Let A be symmetric with spectrum contained in $[\lambda_{\min}, \lambda_{\max}]$. Then the arnoldi approximation of $f(A)\boldsymbol{b}$ satisfies*

$$\left|f(A)\boldsymbol{b} - \left\|\boldsymbol{b}\right\|_2 V_m f(A_m)\boldsymbol{e}_1\right| \leq 2\left\|\boldsymbol{b}\right\|_2 \min_{p\in\Pi_{m-1}} \left\|f - q_{m-1}^{-1}p\right\|_{[\lambda_{\min},\lambda_{\max}]}.$$

Rational Krylov methods are often used instead of polynomial ones when the function f is not analytic or has a singularity close to the spectrum of A, due to the generally better convergence rate of rational approximations than that of the polynomial ones for such functions. This is the case for $f(x) = x \log x$ when the matrix argument is singular or has many eigenvalues close to 0. For such functions, the higher cost per iteration induced by the linear system solves is justified by the much smaller number of iterations required to attain a prescribed accuracy. We refer to [57, 58] for an extensive discussion of rational Krylov methods for the computation of matrix functions. The approximation (5.19) can be also interpreted in terms of rational Gauss quadrature rules; see, e.g., [2, 88].

Remark 5.4. The standard Arnoldi algorithm is inherently sequential since the computation of the new vector of the Krylov basis $\boldsymbol{v}_{m+1}$ requires the previous computation of $\boldsymbol{v}_m$. It is possible to parallelize it by solving several linear systems simultaneously and expanding the Krylov basis with blocks of vectors, with one of the strategies presented in [22], at the cost of lower numerical stability. Since in this work we are expected to compute several quadratic forms $\boldsymbol{b}^T f(A)\boldsymbol{b}$, we can easily achieve parallelization by assigning the quadratic forms to different processors and thus we can neglect parallelism inside the computation of a single quadratic form.

Remark 5.5. We mention that for symmetric A it is possible to construct the Krylov basis V_m using a method based on short recurrences such as rational Lanczos [84]. This has the advantage of reducing the orthogonalization costs, which can become significant if m is large, and also avoids the need to store the matrix V_m when approximating the quadratic form $\boldsymbol{b}^T f(A)\boldsymbol{b}$; see (5.19). However, the implementation in finite arithmetic of short recurrence methods can suffer from loss of orthogonality, which in turn can lead to a slower convergence. In order to avoid this potential problem, we use the rational Arnoldi method with full orthogonalization. Since we expect to attain convergence in a small number of iterations, the orthogonalization costs remain modest compared to the cost of operations with A.

5.3.1 Convergence

By Proposition 5.6, the accuracy of the approximation (5.20) for $f(A)\boldsymbol{b}$ is related to the quality of rational approximations to the function f of the form $r(z) = q_{m-1}(z)^{-1}p_{m-1}(z)$, where $p_{m-1} \in \Pi_{m-1}$ and q_{m-1} is determined by the poles of the rational Krylov subspace.

In the case of quadratic forms we can prove a faster convergence rate using the fact that $\psi_m = \psi$ for rational functions of degree up to $(2m-1, 2m-2)$. The following fact, as stated in [18, Lemma 4.3], was already present in the literature; see [58, Remark 3.2] and [84, Proposition 3.1]. Here we report the proof in [18] for completeness.

Lemma 5.7. *Assume that A is symmetric. Let $p_{2m-1} \in \Pi_{2m-1}$ and define the rational function $r(z) = q_{m-1}(z)^{-2}p_{2m-1}(z)$. Then we have*

$$\boldsymbol{b}^T r(A)\boldsymbol{b} = \boldsymbol{b}^T V_m r(A_m) V_m^T \boldsymbol{b}.$$

Proof. It is sufficient to prove this fact for $p_{2m-1}(z) = z^k$, for $k = 0, \dots, 2m-1$. Assuming

for the moment that $k = 2j + 1$ is odd, we have

$$\boldsymbol{b}^T r(A)\boldsymbol{b} = \boldsymbol{b}^T s(A)As(A)\boldsymbol{b}, \qquad \text{with} \quad s(z) = q_{m-1}(z)^{-1}z^j, \quad j < m.$$

Now using [58, Lemma 3.1], we obtain

$$\boldsymbol{b}^T r(A)\boldsymbol{b} = \big(\boldsymbol{b}^T V_m s(A_m)V_m^T\big)A\big(V_m s(A_m)V_m^T\boldsymbol{b}\big) = \boldsymbol{b}^T V_m^T r(A_m)V_m^T\boldsymbol{b}.$$

The case of $p_{m-1}(z) = z^k$ with k even can be proved in the same way, by writing

$$\boldsymbol{b}^T r(A)\boldsymbol{b} = \boldsymbol{b}^T s(A)^2\boldsymbol{b}, \quad \text{with} \quad s(z) = q_{m-1}(z)^{-1}z^j, \quad j < m.$$

□

Lemma 5.7 leads to the following convergence result for the approximation of quadratic forms, with the same proof as Proposition 5.6 in [58, Corollary 3.4].

Proposition 5.8. *Let A be symmetric with spectrum contained in $[\lambda_{min}, \lambda_{max}]$, $\psi = \boldsymbol{b}^T f(A)\boldsymbol{b}$ and denote by ψ_m the approximation* (5.19)*. We have*

$$|\psi - \psi_m| \leq 2\,\|\boldsymbol{b}\|_2^2 \min_{p\in\Pi_{2m-1}} \left\|f - q_{m-1}^{-2}p\right\|_{[\lambda_{\min},\lambda_{\max}]}.$$

By comparing Proposition 5.8 with Proposition 5.6, we can expect the convergence for quadratic forms to be roughly twice as fast as the one for matrix-vector products with $f(A)$.

5.3.2 Poles for the Rational Krylov Subspace

Recall that the function $f(z) = x\log x$ has the integral expression (5.6), which corresponds to a Cauchy-Stieltjes function multiplied by the polynomial $x(1-x)$. This implies that we can expect that a pole sequence that yields fast convergence for Cauchy-Stieltjes functions will be effective also in our case, especially if we add two poles at ∞ to account for the degree-two polynomial.

The authors of [79] consider the case of a positive definite matrix A with spectrum in $[a, b]$ and a Cauchy-Stieltjes function f, and relate the error for the computation of $f(A)\boldsymbol{b}$ with a rational Krylov method to the *third Zolotarev problem* in approximation theory. The solution to this problem is known explicitly and it can be used to find poles on $(-\infty, 0)$ that provide in some sense an optimal convergence rate for the rational Krylov method [79, Corollary 4]. However, the optimal Zolotarev poles are not nested, so they cannot be used to expand the Krylov subspace incrementally, and they are practical only if one knows in advance how many iterations to perform, for instance by relying on an a priori error bound. This drawback can be overcome by constructing a nested sequence of poles that is equidistributed according to the limit measure identified by the optimal poles, which can be done with the method of equidistributed sequences (EDS) described in [79, Section 3.5]. These poles have the same asymptotic convergence rate as the optimal Zolotarev poles and are usually better for practical purposes. To be computed, they require the knowledge of $[a, b]$ or a positive interval Σ such that $[a, b] \subset \Sigma$.

As an alternative, one can also use poles obtained from Leja-Bagby points [8, 58]. These points can be computed with a greedy algorithm and they have an asymptotic convergence rate that is close to the optimal one. See [58, Section 4] and the references therein for additional information.

Remark 5.6. For the function $f(x) = x \log x$, the first few iterations of a polynomial Krylov method have a fast convergence rate that is close to the convergence rate of rational Krylov methods, even if it becomes asymptotically much slower for ill conditioned matrices. The faster initial convergence can be explained by the algebraic factor in the bound (5.7) for polynomial approximations of $x \log x$. Since polynomial Krylov iterations are cheaper than rational Krylov iterations, this suggests the use of a mixed polynomial-rational method, that starts with a few polynomial Krylov steps and then switches to a rational Krylov method with, e.g., EDS poles to achieve a higher accuracy. These methods are compared numerically in Example 5.10, where we also test the performance of the a posteriori error bound that we prove in Section 5.3.3.

Remark 5.7. Note that in the context of the graph entropy the matrix A is a graph Laplacian, which is a singular matrix. Therefore in principle it is not possible to use the poles described in this section, since here we assume that A is positive definite. However, we can use one of the desingularization strategies described in [20] to remove the 0 eigenvalue of the graph Laplacian, obtaining a matrix with spectrum contained in $[\lambda_2, \lambda_n]$, where λ_2 is the second smallest eigenvalue of A and λ_n is the largest one. In our implementation we use the approach that is called implicit desingularization in [20], which consists in replacing the initial vector $\boldsymbol{b}$ for the Krylov subspace with $\boldsymbol{c} = \boldsymbol{b} - \frac{\mathbb{1}^T \boldsymbol{b}}{n} \mathbb{1}$, where $\mathbb{1}$ is the vector of all ones. Since $\boldsymbol{c}$ is orthogonal to the eigenvector $\mathbb{1}$ associated to the eigenvalue 0, it can be shown that the convergence of a Krylov subspace method with starting vector $\boldsymbol{c}$ is the same as for a matrix with spectrum in $[\lambda_2, \lambda_n]$. An approximation of $f(A)\boldsymbol{b}$ can be then cheaply recovered from $f(A)\boldsymbol{c}$ using the fact that $f(A)\mathbb{1} = f(0)\mathbb{1}$, and similarly for $\boldsymbol{b}^T f(A)\boldsymbol{b}$. See [20] for more details.

5.3.3 A Posteriori Error Bound

In this section we prove an a posteriori bound for the error in the computation of the quadratic form $\boldsymbol{b}^T f(A)\boldsymbol{b}$ with a rational Krylov method. This bound is a variant of the one described in [57, Section 6.6.2] for $f(A)\boldsymbol{b}$, modified in order to account for the faster convergence rate in the case of quadratic forms.

We recall that after m iterations the rational Arnoldi algorithm yields the rational Arnoldi decomposition [21, Definition 2.3]

$$AV_{m+1}\underline{K_m} = V_{m+1}\underline{H_m}, \tag{5.21}$$

where $\operatorname{range} V_{m+1} = \mathcal{Q}_{m+1}(A, \boldsymbol{b})$ and $\underline{K_m}$, $\underline{H_m}$ are $(m+1)\times m$ upper Hessenberg matrices with full rank. Let us consider the situation when $\xi_m = \infty$: in this case the last row of $\underline{K_m}$ is zero, and the decomposition simplifies to

$$AV_m K_m = V_m H_m + \boldsymbol{v}_{m+1}\boldsymbol{h}_{m+1}^T,$$

where H_m and K_m denote the $m \times m$ leading principal blocks of $\underline{H_m}$ and $\underline{K_m}$, respectively, and $\boldsymbol{h}_{m+1}^T = h_{m+1,m}\boldsymbol{e}_m^T$ denotes the last row of $\underline{H_m}$. Note that K_m is nonsingular since $\underline{K_m}$ has full rank, so we can rewrite the decomposition as

$$AV_m = V_m A_m + \boldsymbol{v}_{m+1}\boldsymbol{h}_{m+1}^T K_m^{-1}, \qquad \text{where} \quad A_m = V_m^T A V_m = H_m K_m^{-1}. \tag{5.22}$$

Remark 5.8. To derive the bound, we assume that $\xi_m = \infty$ because it simplifies the rational Arnoldi decomposition and hence the expression of the bound. Such an assumption is not restrictive, since the value of ξ_m does not have any impact on V_m and A_m, but only on $\boldsymbol{v}_{m+1}$ and the last column of $\underline{H_m}$ and $\underline{K_m}$. As we shall see later, we can use a technique described in [57, Section 6.1] to compute the bound for all m, without having to set the corresponding poles $\xi_m = \infty$. Note that if $\xi_m \neq \infty$, then (5.22) does not hold, and in particular $V_m^T A V_m \neq H_m K_m^{-1}$.

By using the Cauchy integral formula (2.3), we can obtain the following expression for the error [57, Section 6.2.2]:

$$f(A)\boldsymbol{b} - V_m f(A_m) V_m^T \boldsymbol{b} = \frac{1}{2\pi i}\int_\Gamma f(z)(zI - A)^{-1}\boldsymbol{r}_m(z)\mathrm{d}z, \tag{5.23}$$

where Γ is a contour contained in the region of analyticity of f that encloses the spectrum of A, and

$$\boldsymbol{r}_m(z) = \boldsymbol{b} - (zI - A)\boldsymbol{x}_m(z), \qquad \text{with} \quad \boldsymbol{x}_m(z) = V_m(zI - A_m)^{-1}V_m^T\boldsymbol{b},$$

which can be seen as a residual vector of the shifted linear system $(zI - A)\boldsymbol{x} = \boldsymbol{b}$. It turns out that [57, Section 6.2.2]

$$\boldsymbol{r}_m(z) = \|\boldsymbol{b}\|_2\, \varphi_m(z)\boldsymbol{v}_{m+1}, \qquad \text{with} \quad \varphi_m(z) = \boldsymbol{h}_{m+1}^T K_m^{-1}(zI - A_m)^{-1}\boldsymbol{e}_1.$$

Observe that we have

$$\begin{aligned}\boldsymbol{b}^T(zI - A)^{-1}\boldsymbol{r}_m(z) &= \boldsymbol{r}_m(z)^T(zI - A)^{-1}\boldsymbol{r}_m(z) + \boldsymbol{x}_m(z)^T\boldsymbol{r}_m(z)\\ &= \boldsymbol{r}_m(z)^T(zI - A)^{-1}\boldsymbol{r}_m(z),\end{aligned}$$

where we exploited the fact that $\boldsymbol{x}_m(z) \in \mathcal{Q}_m(A, \boldsymbol{b}) \perp \boldsymbol{r}_m(z)$.

By using this property in conjunction with (5.23), we can write the error for the quadratic form $\boldsymbol{b}^T f(A)\boldsymbol{b}$ as

$$\begin{aligned}\psi - \psi_m &= \frac{1}{2\pi i}\int_\Gamma f(z)\boldsymbol{r}_m(z)^T(zI - A)^{-1}\boldsymbol{r}_m(z)\mathrm{d}z\\ &= \frac{1}{2\pi i}\|\boldsymbol{b}\|_2^2\int_\Gamma f(z)\varphi_m(z)^2\boldsymbol{v}_{m+1}^T(zI - A)^{-1}\boldsymbol{v}_{m+1}\mathrm{d}z.\end{aligned} \tag{5.24}$$

We can now follow the same steps used in [57, Section 6.2.2] to bound the integral in (5.24). Assume that A_m has the spectral decomposition $A_m = U_m D_m U_m^T$, with U_m orthogonal and $D_m = \mathrm{diag}(\theta_1, \ldots, \theta_m)$, and define the vectors

$$[\alpha_1, \ldots, \alpha_m] = \boldsymbol{h}_{m+1}^T K_m^{-1} U_m \qquad \text{and} \qquad [\beta_1, \ldots, \beta_m]^T = U_m^T\boldsymbol{e}_1,$$

so that we have

$$\varphi_m(z) = \boldsymbol{h}_{m+1}^T K_m^{-1} U_m (zI - D_m)^{-1} U_m^T \boldsymbol{e}_1 = \sum_{j=1}^{m} \alpha_j \beta_j \frac{1}{z - \theta_j},$$

and

$$\varphi_m(z)^2 = \sum_{j=1}^{m} \alpha_j^2 \beta_j^2 \frac{1}{(z-\theta_j)^2} + 2 \sum_{j=1}^{m} \alpha_j \beta_j \gamma_j \frac{1}{z - \theta_j},$$

where we defined $\gamma_j = \sum_{\ell \,:\, \ell \neq j} \alpha_\ell \beta_\ell \frac{1}{\theta_j - \theta_\ell}$. By plugging this expression into (5.24) we get

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{b}\|_2^2}(\psi - \psi_m) = \frac{1}{2\pi i} \int_\Gamma f(z) \varphi_m(z)^2 \boldsymbol{v}_{m+1}^T (zI - A)^{-1} \boldsymbol{v}_{m+1} \mathrm{d}z \\
&= \sum_{j=1}^{m} \alpha_j^2 \beta_j^2 \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{(z-\theta_j)^2} \boldsymbol{v}_{m+1}^T (zI - A)^{-1} \boldsymbol{v}_{m+1} \mathrm{d}z \\
&\quad + 2 \sum_{j=1}^{m} \alpha_j \beta_j \gamma_j \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{z - \theta_j} \boldsymbol{v}_{m+1}^T (zI - A)^{-1} \boldsymbol{v}_{m+1} \mathrm{d}z \\
&= \sum_{j=1}^{m} \alpha_j^2 \beta_j^2 \boldsymbol{v}_{m+1}^T \Big((f(A) - f(\theta_j) I)(A - \theta_j I)^{-2} - f'(\theta_j)(A - \theta_j I)^{-1} \Big) \boldsymbol{v}_{m+1} \\
&\quad + 2 \sum_{j=1}^{m} \alpha_j \beta_j \gamma_j \boldsymbol{v}_{m+1}^T (f(A) - f(\theta_j) I)(A - \theta_j I)^{-1} \boldsymbol{v}_{m+1},
\end{aligned}$$

where for the last equality we used the residue theorem [66, Theorem 4.7a].

Let us define

$$g_m(z) = \sum_{j=1}^{m} \begin{cases} \alpha_j^2 \beta_j^2 \left(\dfrac{f(z) - f(\theta_j)}{(z-\theta_j)^2} - \dfrac{f'(\theta_j)}{z - \theta_j} \right) + 2\alpha_j \beta_j \gamma_j \dfrac{f(z) - f(\theta_j)}{z - \theta_j} & \text{if } z \neq \theta_j, \\ \dfrac{1}{2} \alpha_j^2 \beta_j^2 f''(\theta_j) + 2\alpha_j \beta_j \gamma_j f'(\theta_j) & \text{if } z = \theta_j, \end{cases} \tag{5.25}$$

where the expression for $z = \theta_j$ is obtained by taking the limit for $z \to \theta_j$ in the definition for $z \neq \theta_j$. The above computations immediately lead to the following a posteriori bound for the quadratic form error.

Theorem 5.9. *Let A be a symmetric matrix with spectrum $\sigma(A) \subset [\lambda_{min}, \lambda_{max}]$. Using the same notation as above, we have*

$$\|\boldsymbol{b}\|_2^2 \min_{z \in [\lambda_{min}, \lambda_{max}]} |g_m(z)| \leq |\psi - \psi_m| \leq \|\boldsymbol{b}\|_2^2 \max_{z \in [\lambda_{min}, \lambda_{max}]} |g_m(z)|. \tag{5.26}$$

Remark 5.9. We are mainly interested in the upper bound in (5.26) to have a reliable stopping criterion for the rational Krylov method, although the lower bound can also be of interest. We

also mention that other bounds and error estimates can be obtained, such as those described in [57, Section 6.6], but we found that the one derived in this section worked well enough for our purposes. Under certain assumptions, it is also possible to obtain upper and lower bounds for the quadratic form $\boldsymbol{b}^T f(A)\boldsymbol{b}$ using pairs of rational Gauss quadrature rules, such as Gauss and Gauss-Radau quadrature rules. We refer to [2] for more details.

Example 5.10. In this example we test the accuracy of the lower and upper bounds given in (5.26) for polynomial and rational Krylov methods. We consider a 2000×2000 matrix A with eigenvalues given by the Chebyshev points for the interval $\Sigma = [10^{-3}, 10^3]$, and compute $\boldsymbol{b}^T f(A)\boldsymbol{b}$ with a random vector $\boldsymbol{b}$ and $f(x) = x\log(x)$. The upper and lower bounds are computed numerically by evaluating g_m on a discretization of the interval $[\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$. In addition to the lower and upper bounds, we also consider a heuristic estimate of the error given by the geometric mean of the upper and lower bound in (5.26), i.e.

$$\mathtt{est}_m = \|\boldsymbol{b}\|_2^2 \sqrt{\min_{z\in\Sigma} |g_m(z)| \max_{z\in\Sigma} |g_m(z)|}. \tag{5.27}$$

The results are shown in Figure 5.3. On the left plot we show the convergence for the polynomial Krylov method and for a rational Krylov method with poles from an EDS for Cauchy-Stieltjes functions, and on the right plot a mixed polynomial-rational method that uses 10 poles at ∞ (which correspond to polynomial Krylov steps) followed by 10 EDS poles (see Remark 5.6). The upper and lower bounds match the shape of the convergence curve quite well, although they are less accurate when polynomial iterations are used. Rather surprisingly, the geometric mean of the bounds gives a very accurate estimate for the error, even in the case when the bounds themselves are less accurate.

Remark 5.11. We do not have a rigorous explanation for the accuracy of the estimate based on the geometric mean of the bounds in (5.26), but from further experiments it seems to be very accurate also for other functions. Unfortunately, if the spectral interval Σ is known only approximately, the bounds become less tight and the geometric mean estimate usually ends up underestimating the actual error by one or two orders of magnitude.

Computation of the Bound

Recall that the a posteriori bounds in (5.26) hold after the m-th iteration only if $\xi_m = \infty$. In a practical scenario, i.e. when using the bounds as a stopping criterion for a rational Krylov method, it is desirable to evaluate the bounds after each iteration, without being forced to set the corresponding pole to ∞. As anticipated in Remark 5.8, we provide here two approaches to evaluate the bounds in (5.26) even when $\xi_m \neq \infty$.

One way to avoid setting poles to ∞, proposed in [57, Section 6.1], is to use an auxiliary basis vector $\boldsymbol{v}_\infty$, which is initialized as $\boldsymbol{v}_\infty^{(1)} = A\boldsymbol{v}_1$ at the beginning of the rational Arnoldi algorithm, and maintained orthonormal to the basis vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_j\}$ at each iteration j, at the cost of only one additional orthogonalization per iteration. The basis $[V_j \ \boldsymbol{v}_\infty^{(j)}]$ is an orthonormal basis of the rational Krylov subspace $\mathcal{Q}_{j+1}(A, \boldsymbol{b})$ with poles $\{\xi_1, \dots, \xi_{j-1}, \infty\}$, and it is associated to the auxiliary Arnoldi decomposition

$$AV_j\widetilde{K}_j = [V_j \ \boldsymbol{v}_\infty^{(j)}]\underline{\widetilde{H}_j},$$

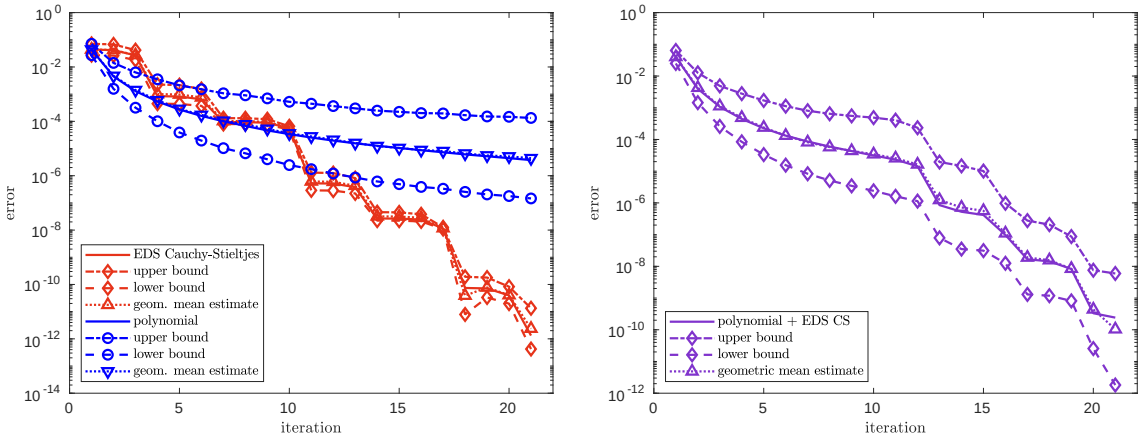

Figure 5.3: Accuracy of error bounds and estimates for the relative error in the computation of $\boldsymbol{b}^T f(A)\boldsymbol{b}$ with Krylov methods, where $\boldsymbol{b}$ is a random vector, $f(x) = x\log(x)$ and A is a 2000×2000 matrix with eigenvalues that are Chebyshev points in the interval $[10^{-3}, 10^3]$. Left: polynomial Krylov and rational Krylov with EDS poles for Cauchy-Stieltjes function. Right: 10 poles at ∞ and 10 EDS poles for Cauchy-Stieltjes functions.

where $\widetilde{K}_j \boldsymbol{e}_j = \boldsymbol{e}_1$ and the last column of $\underline{\widetilde{H}_j}$ contains the orthogonalization coefficients for $\boldsymbol{v}_\infty^{(j)}$. This decomposition can be used to compute the bound (5.26) since the last row of $\underline{\widetilde{K}_j}$ is zero by construction.

Remark 5.12. We point out that if $\xi_j = \infty$ for some j, then the approach described above will not work from iteration $j+1$ onward, since $A\boldsymbol{v}_1 \in \mathcal{Q}_{j+1}(A, \boldsymbol{b})$ and therefore $\boldsymbol{v}_\infty^{(j+1)} = \boldsymbol{0}$ after orthogonalization. This is easily fixed by switching to a different auxiliary vector at iteration $j+1$, such as $\boldsymbol{v}_\infty = A\boldsymbol{v}_{j+1}$, or by setting directly at the start $\boldsymbol{v}_\infty = A^{\ell+1}\boldsymbol{v}_1$, where ℓ is the number of poles at ∞ used to construct the rational Krylov subspace.

In our experience, the technique described above can be sometimes subject to instability due to a large condition number of the matrix $\widetilde{K}_j$. We therefore also propose another approach, which is inspired by the methods for moving the poles of a rational Krylov subspace presented in [21, Section 4]. The idea is to add a pole at ∞ at the beginning of the pole sequence, and reorder the poles at each iteration in order to always have the last pole equal to ∞.

First of all, we recall how to swap poles in a rational Arnoldi decomposition. This procedure is a special case of the algorithm described in [21], but we still describe it in some detail for completeness. Recall that the poles of a rational Krylov subspace are the ratios of the entries below the main diagonals of $\underline{H_j}$ and $\underline{K_j}$ [21, Definition 2.3], i.e. $\xi_j = h_{j+1,j}/k_{j+1,j}$. In other words, the poles $\{\xi_1, \ldots, \xi_j\}$ are the eigenvalues of the upper triangular pencil $(\widehat{H}_j, \widehat{K}_j)$, where we denote by $\widehat{H}_j$ the bottom $j \times j$ block of $\underline{H_j}$, and similarly for $\widehat{K}_j$. So we can obtain a transformation that swaps two adjacent poles in the same way as orthogonal transformations that reorder eigenvalues in a generalized Schur form [71]. Let U_j and W_j be $j \times j$ orthogonal matrices such that the pencil $U_j^T(\widehat{H}_j, \widehat{K}_j)W_j$ is still in upper triangular

form and has the last two eigenvalues in reversed order; the matrices U_j and W_j only involve 2×2 rotations, and they can be computed and applied cheaply as described in [71]. Defining $\widehat{U}_j = \mathrm{blkdiag}(1, U_j) \in \mathbb{R}^{(j+1)\times(j+1)}$, we have the new rational Arnoldi decomposition

$$A\widetilde{V}_{j+1}\underline{\widetilde{K}_j} = \widetilde{V}_{j+1}\underline{\widetilde{H}_j},$$

where

$$\widetilde{V}_{j+1} = V_{j+1}\widehat{U}_j, \qquad \underline{\widetilde{K}_j} = \widehat{U}_j^T \underline{K_j} W_j \qquad \text{and} \qquad \underline{\widetilde{H}_j} = \widehat{U}_j^T \underline{H_j} W_j.$$

This decomposition has the same poles as $AV_{j+1}\underline{K_j} = V_{j+1}\underline{H_j}$, with the difference that the last two poles ξ_{j-1} and ξ_j are now swapped. In particular, if $\xi_{j-1} = \infty$, the last pole of the new decomposition is now ∞, and hence the last row of $\underline{\widetilde{K}_j}$ is equal to zero.

Given the pole sequence $\{\xi_1, \xi_2, \dots\}$, let us consider the rational Krylov subspace associated to the modified pole sequence $\{\infty, \xi_1, \xi_2, \dots\}$. Clearly, after the first iteration both pole sequences identify the same subspace $\mathcal{Q}_1(A, \boldsymbol{b})$, but the last (and first) pole of the modified sequence is ∞, so the last row of $\underline{K_1}$ is zero and we can use the decomposition $AV_1K_1 = V_2\underline{H_1}$ to compute the bound (5.26). After the second iteration, if $\xi_1 \neq \infty$, we can swap the poles ξ_1 and ∞ with the procedure outlined above to obtain the decomposition $AV_2K_2 = V_3\underline{H_2}$ associated to the poles $\{\xi_1, \infty\}$, where again the last row of $\underline{K_2}$ is equal to zero (for simplicity we still use the notation K_j instead of $\widetilde{K}_j$, and similarly for V_j and H_j). If $\xi_1 = \infty$, there is no need to swap poles and we can proceed to the next iteration.

By repeating the same steps at each iteration, we can ensure that after j iterations we have a decomposition $AV_jK_j = V_{j+1}\underline{H_j}$, associated to the poles $\{\xi_1, \dots, \xi_{j-1}, \infty\}$ in this order, so that the last row of $\underline{K_j}$ is equal to zero and it can be used to compute the bound (5.26).

Remark 5.13. Note that V_j is a basis of $\mathcal{Q}_j(A, \boldsymbol{b})$, which is the same subspace that we would have obtained if we had run the rational Arnoldi algorithm with poles $\{\xi_1, \dots, \xi_{j-1}\}$; so the method described in this section actually computes the approximation ψ_j and the bound associated to the poles $\{\xi_1, \dots, \xi_{j-1}\}$, and not to the modified sequence $\{\infty, \xi_1, \dots, \xi_{j-2}\}$. The initial pole at ∞ is only added to enable the computation of the bound and it is never used in the actual approximation.

5.4 Implementation Aspects

In this section we outline the algorithm used to compute the entropy obtained by connecting the different components presented in the previous sections, and we briefly comment on some of the decisions that have to be taken in an implementation, especially concerning stopping criteria. Given a symmetric positive semidefinite matrix A and a target relative accuracy ϵ, the algorithm should output an estimate `trest` of $\mathrm{Tr}(f(A))$, where $f(x) = -x \log x$, such that

$$|\mathrm{Tr}(f(A)) - \texttt{trest}| \leq \epsilon \, \mathrm{Tr}(f(A)),$$

using either the deterministic probing approach of Section 4.2 or a stochastic trace estimator from Section 4.1. Observe that the entropy of an $n \times n$ density matrix is always bounded from above by $\log n$, but it may be in principle very small, so we prefer to aim for a certain relative accuracy rather than an absolute accuracy. Quadratic forms and matrix-vector products with $f(A)$ are computed using Krylov methods, specifically using a certain number of poles at ∞ followed by the EDS poles described in Section 5.3.2.

Remark 5.14. In the following, we are going to use $\hat{\epsilon}$ to denote an absolute error, to distinguish it from the target relative accuracy ϵ. Note that we can easily transform absolute inequalities for the error into relative inequalities if we know in advance an estimate or a lower bound for $\mathrm{Tr}(f(A))$. Recall that if A is an M-matrix, $\mathcal{T}_d(f(A))$ is actually a lower bound for $\mathrm{Tr}(f(A))$ (Proposition 5.3). In the general case, any rough approximation of the entropy can be used for this purpose, since the important point is determining the order of magnitude of $\mathrm{Tr}(f(A))$.

Remark 5.15. The error in the approximation of $S(A)$ can be divided into the error in the approximation of the trace using a probing method or a stochastic estimator, and the error in the approximation of the quadratic forms with $f(A)$ using a Krylov subspace method. For simplicity, in the following we impose that the relative error associated to each of these two components is smaller than $\epsilon/2$.

5.4.1 Probing Method Implementation

We begin by observing that it is not possible to cheaply estimate the error of a probing method a posteriori, since error estimates are usually based on approximations with different values of the distance d, which in general lead to completely different colorings that would require computing all quadratic forms from scratch.

For this reason, it is best to find a value of d that ensures a relative accuracy ϵ a priori when using a distance-d coloring. This can be done using one of the bounds in Corollary 5.2, but it can often lead to unnecessary additional work, since the bounds usually overestimate the error by a couple of orders of magnitude; see Figure 5.2. Therefore we also provide a heuristic criterion for choosing d that does not have the same theoretical guarantee as the bounds, but appears to work quite well in practice. In view of Corollary 5.2, we can expect the absolute error to behave as

$$|\mathrm{Tr}(f(A)) - \mathcal{T}_d(f(A))| \sim \frac{C}{d^k} q^d, \tag{5.28}$$

for $k = 2$ and some parameters $C > 0$ and $q \in (0, 1)$. However, we found that sometimes the actual error behavior is better described with a different value of k, such as $k = 3$, so we do not impose that $k = 2$. To estimate the values of the parameters, we compute $\mathcal{T}_d(f(A))$ for $d = 1, 2, 3$ and use the estimate

$$|\mathrm{Tr}(f(A) - \mathcal{T}_d(f(A))| \approx |\mathcal{T}_{d+1}(f(A)) - \mathcal{T}_d(f(A))|, \qquad d = 1, 2.$$

Assuming that (5.28) holds exactly and fixing the value of k, we can determine C and q by solving the equations

$$|\mathcal{T}_{d+1}(f(A)) - \mathcal{T}_d(f(A))| = \frac{C}{d^k} q^d, \qquad d = 1, 2.$$

We can check when the resulting estimate is below $\hat{\epsilon}$ to heuristically determine d, i.e., we select d as

$$d_\star = \min\left\{d \,:\, \frac{C}{d^k} q^d \leq \hat{\epsilon}\right\},$$

in order to have the approximate absolute error inequality

$$|\mathrm{Tr}(f(A)) - \mathcal{T}_{d_\star}(f(A))| \lesssim \hat{\epsilon}.$$

We found that the best results are obtained for $k = 2$ and $k = 3$, so in our implementation we use the maximum of the two corresponding estimates. Variants of this estimate include using other values of d to estimate the parameters instead of $d = 1, 2, 3$, and using four different values in order to also estimate the parameter k. However, they usually give results that are similar or sometimes worse than the estimate presented above, so they are often not worth the additional effort required to compute them. In particular, using four values of d raises the risk of misjudging the value of q, causing the estimate to be inaccurate for large values of d. The error estimate (5.28) is compared to the actual error and the theoretical bound (5.11) for two different graphs in Figure 5.4. The figure also includes a simple error estimate based on consecutive differences, which requires the computation of $\mathcal{T}_{d+1}(f(A))$ to estimate the error for $\mathcal{T}_d(f(A))$.

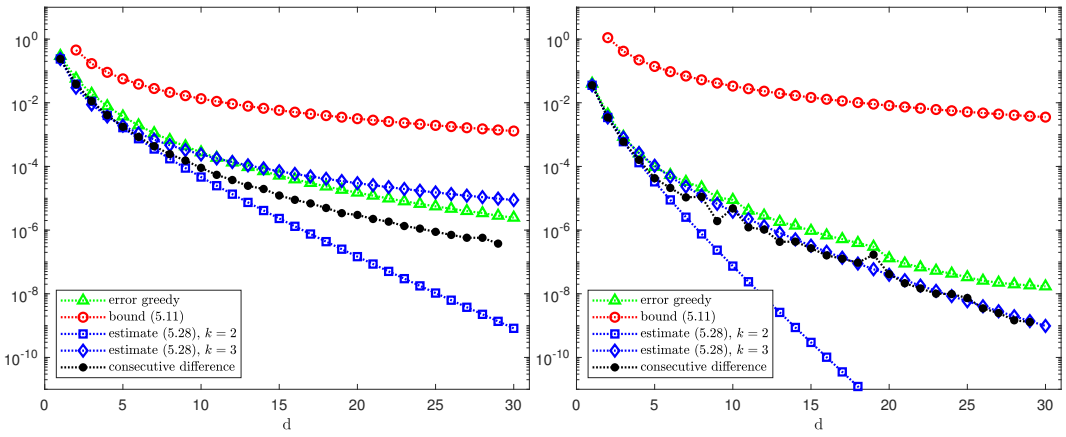

Figure 5.4: Absolute error of the probing method with greedy coloring (Algorithm 5) compared with the theoretical bound (5.11) using $a = \lambda_2$, the heuristic error estimate (5.28) and the simple error estimate $|\mathrm{Tr}(f(A) - \mathcal{T}_d(f(A))| \approx |\mathcal{T}_{d+1}(f(A)) - \mathcal{T}_d(f(A))|$. Left: Laplacian of the largest connected component of the graph `minnesota`, with 2640 nodes. Right: Laplacian of the largest connected component of the graph `eris1176`, with 1174 nodes.

Remark 5.16. The heuristic criterion for selecting d requires the computation of $\mathcal{T}_d(f(A))$ for $d = 1, 2, 3$, so it is more expensive to use than the theoretical bound (5.11). However, this cost is usually small compared to the cost of computing $\mathcal{T}_d(f(A))$ for the selected value of d, especially if the requested accuracy is small. Note also that the heuristic criterion always computes $\mathcal{T}_3(f(A))$, so it does more work than necessary when $d \leq 2$ would be sufficient.

Nevertheless, in such a situation the theoretical bound (5.11) may suggest to use an even higher value of d (see Figure 5.4).

After choosing d such that

$$|\mathrm{Tr}(f(A)) - \mathcal{T}_d(f(A))| \leq \hat{\epsilon}, \tag{5.29}$$

using either the a priori bound (5.11) or the estimate (5.28), a distance-d coloring can be computed with one of the coloring algorithms described in Section 4.2, depending on the properties of the graph. The greedy coloring [94, Algorithm 4.2] is usually a good choice for general graphs.

Let us now determine the accuracy required in the computation of the quadratic forms. Assume that we have

$$\mathcal{T}_d(f(A)) = \sum_{\ell=1}^{n_c} \boldsymbol{v}_\ell^T f(A) \boldsymbol{v}_\ell,$$

where $\{\boldsymbol{v}_\ell\}_{\ell=1}^{n_c}$ are the probing vectors used in the distance-d coloring. Let $\hat{\psi}_\ell$ denote the approximation of $\boldsymbol{v}_\ell^T f(A) \boldsymbol{v}_\ell$ obtained with a Krylov method. Recall that $\|\boldsymbol{v}_\ell\|_2 = |V_\ell|^{1/2}$, where V_ℓ denotes the set of the partition associated to the ℓ-th color. If we impose the conditions

$$\left| \boldsymbol{v}_\ell^T f(A) \boldsymbol{v}_\ell - \hat{\psi}_\ell \right| \leq \hat{\epsilon} \cdot \frac{|V_\ell|}{n} \qquad \ell = 1, \ldots, n_c, \tag{5.30}$$

where we normalized the accuracy requested for each quadratic form depending on $\|\boldsymbol{v}_\ell\|_2$, we obtain the desired absolute accuracy on the probing approximation

$$\left| \mathcal{T}_d(f(A)) - \sum_{\ell=1}^{n_c} \hat{\psi}_\ell \right| \leq \hat{\epsilon}. \tag{5.31}$$

If we are aiming for a relative accuracy ϵ, we should select $\hat{\epsilon} = \frac{1}{2}\epsilon\,\mathrm{Tr}(f(A))$ in (5.29) and (5.31). In practice, $\mathrm{Tr}(f(A))$ will be replaced by a rough approximation (see Remark 5.14).

The overall probing algorithm is summarized in Algorithm 6.

5.4.2 Adaptive Hutch++ Implementation

The probing approach requires that A is a sparse matrix and that the distance-d colorings are easily computable. In order to cover more general cases, we propose to use the Hutch++ estimator (cf., Section 4.1.2), and in particular its adaptive implementation. We use the Matlab code of [86, Algorithm 3] provided by the authors, modified to use Krylov methods for the computations with $f(A)$. This algorithm requires an absolute tolerance $\hat{\epsilon}$ and a failure probability δ, and outputs an approximation $\mathrm{Tr}_{\mathsf{adap}}(f(A))$ such that

$$\mathbb{P}\big[\,\big|\mathrm{Tr}(f(A)) - \mathrm{Tr}_{\mathsf{adap}}(f(A))\big| \geq \hat{\epsilon}\big] \leq \delta.$$

To obtain an approximation within a relative accuracy ϵ, we can use $\hat{\epsilon} \approx \epsilon\,\mathrm{Tr}(f(A))$, using a rough approximation of $\mathrm{Tr}(f(A))$. Similarly to the probing method, in order to have a final

Algorithm 6 Probing method for $S(A)$

Input: $A \in \mathbb{R}^{n\times n}$ symmetric positive semidefinite, ϵ relative error tolerance
Output: trest $\approx S(A)$ such that $|\mathrm{tr} - S(A)|\,/S(A) \lesssim \epsilon$

1: Select d such that $|\mathcal{T}_d(f(A)) - S(A)|\,/S(A) \lesssim \epsilon/2$, using either the bound (5.11) or the heuristic (5.28). The heuristic (5.28) requires the computation of $\mathcal{T}_d(f(A))$ for $d = 1, 2, 3$, which can be done by running steps 2 – 4.
2: Compute a distance-d coloring of $G(A)$ with, e.g., Algorithm 5 and the associated probing vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s\}$.
3: For $\ell = 1, \ldots, n_c$, compute $\hat{\psi}_\ell \approx \boldsymbol{v}_\ell^T f(A)\boldsymbol{v}_\ell$ such that (5.30) holds, using a rational Krylov method with either the upper bound (5.26) or the estimate (5.27) as stopping criterion.
4: **return** trest $= \sum_{\ell=1}^{n_c} \hat{\psi}_\ell$, satisfying $\Big|\sum_{\ell=1}^{n_c} \hat{\psi}_\ell - S(A)\Big|/S(A) \lesssim \epsilon$.

relative error bounded by ϵ, in our implementation we use a tolerance $\hat{\epsilon} \approx \frac{1}{2}\epsilon\,\mathrm{Tr}(f(A))$ for adaptive Hutch++, and we set the accuracy for the computation of matrix-vector products and quadratic forms in order to ensure that the total error due to the Krylov approximations remains below $\frac{1}{2}\epsilon\,\mathrm{Tr}(f(A))$. We omit the technical details to simplify the presentation.

5.4.3 Krylov Method Implementation

Quadratic forms with $f(A)$ are approximated using a Krylov method with some poles at ∞ followed by the EDS poles of Section 5.3.2, using as a stopping criterion either the a posteriori upper bound (5.26) or the estimate shown in Example 5.10.

The number of poles at ∞ is chosen in an adaptive way, switching to finite poles when the error reduction in the last few iterations of the polynomial Krylov method is "small". Specifically, we decide to switch to EDS poles after the k-th iteration if on average the last $\ell \geq 1$ iterations did not reduce the error bound or estimate `err_est` by at least a factor $c \in (0, 1)$, i.e. if

$$\frac{\texttt{err\_est}_k}{\texttt{err\_est}_{k-\ell-1}} \geq c^\ell.$$

In our implementation we use $\ell = 3$ and $c = 0.75$, usually leading to at most 10 polynomial Krylov iterations.

Since EDS poles are contained in $(-\infty, 0)$, each rational Krylov iteration involves the solution of a symmetric positive definite linear system, which can be computed either with a direct method using a sparse Cholesky factorization, or iteratively with the conjugate gradient method using a suitable preconditioner. Note that the same EDS poles can be used for all quadratic forms, so the number of different matrices that appear in the linear systems is usually small and independent of the total number of quadratic forms. Although this depends on the accuracy requested for the entropy, the number of EDS poles used is almost always bounded by 10, and often much smaller than that: see the numerical experiments in Section 5.5 for some examples. This is a great advantage for direct methods, especially when

the Cholesky factor remains sparse, since we can compute and store a Cholesky factorization for each pole and then reuse it for all quadratic forms. If the fill-in in the Cholesky factor is moderate, the cost of a rational iteration can become comparable to the cost of a polynomial one, leading to large savings when computing many quadratic forms. Of course, for large matrices with a general sparsity structure the computation of even a single Cholesky factor may be unfeasible, so the only option is to use a preconditioned iterative method. In such a situation, it is still possible to benefit from the small number of different matrices that appear in linear systems by storing and reusing preconditioners, but the gain is less evident compared to direct methods.

The matrix-vector products with $f(A)$ in the Hutch++ algorithm are approximated with the same Krylov subspace method, with the difference that we use the a posteriori upper and lower bounds from [57, eq. (6.15)]. A geometric mean estimate similar to the one used in Example 5.10 can also be used in this context. For the computation of the a posteriori bounds we use the pole swapping technique with an auxiliary pole at ∞ described in Section 5.3.3.

5.5 Numerical Experiments

The experiments were done in Matlab R2021b on a laptop with operating system Ubuntu 20.04, using a single core of an Intel i5-10300H CPU running at 2.5 GHz, with 32 GB of RAM. Since we are using Matlab, the execution times may not reflect the performance of a high performance implementation, but they are still a useful indicator when comparing different methods.

The target relative accuracy of the adaptive Hutch++ estimator will be 10^{-2} or 10^{-3}, since for higher precision the cost becomes prohibitive. Such accuracy is reasonable in the context of trace estimation for difficult problems; see [28, 28, 86]. When using probing, we are able to carry out the execution with a smaller input tolerance.

5.5.1 Test Matrices

We consider a number of symmetric test matrices from the SuiteSparse Matrix Collection [32]. All matrices are treated as binary matrices, i.e. all edge weights are set to one. For each matrix, we extract the graph Laplacian associated to the largest connected component and we normalize it so that it has unit trace. We report some information on the resulting matrices in Table 5.1. For the four smallest matrices, the eigenvalues were computed via diagonalization, while for the larger matrices the eigenvalues λ_2 and λ_n were approximated using `eigs`. The cost of solving a linear system with a direct method is highly dependent on the fill-in in the Cholesky factorization; the column labelled `fill-in` in Table 5.1 contains the ratios $\texttt{nnz}(R)/\texttt{nnz}(\rho)$, where ρ is the test matrix and R is the Cholesky factor of any shifted matrix $\rho + \alpha I$, for $\alpha > 0$ ($\texttt{nnz}(M)$ denotes the number of nonzeros of a matrix M). All matrices have been ordered using the approximate minimum degree reordering option available in Matlab before factorization.

Table 5.1: Information on the matrices used in the experiments.

test matrix	n	$\texttt{nnz}(\rho)$	`fill-in`	λ_2	λ_n	entropy
`yeast`	2224	15442	3.6	4.54e-06	4.96e-03	7.055
`minnesota`	2640	9244	1.3	1.28e-07	1.04e-03	7.607
`ca-HepTh`	8638	58250	7.5	4.92e-07	1.33e-03	8.540
`bcsstk29`	13830	618678	2.9	7.22e-08	1.25e-04	9.440
`cond-mat-2005`	36458	379926	21.7	5.63e-08	8.13e-04	9.958
`loc-Brightkite`	56739	482629	32.6	7.10e-08	2.67e-03	9.896
`ut2010`	115406	687472	1.2	2.72e-10	3.44e-04	11.361
`usroads`	126146	450046	1.4	2.39e-11	2.54e-05	11.478
`com-Amazon`	334863	2186607	105.9	6.69e-10	2.97e-04	12.400
`ny2010`	350167	2059711	1.8	4.67e-12	3.63e-05	12.541
`roadNet-PA`	1087562	4170590	1.6	5.54e-13	3.37e-06	13.628

5.5.2 Probing Bound vs. Estimate

In this experiment, we fix a relative error tolerance $\epsilon = 10^{-3}$ and we compare the choice of d given by the theoretical bound (5.11) with the one provided by the heuristic estimate (5.28). We report in Table 5.2 the error, the execution time, the value of d and the number of colors used in the two cases. When the theoretical bound is used, the selected value of d is significantly higher compared to the one chosen by the heuristic estimate, but in both cases the overall error remains below the tolerance ϵ. Moreover, for certain graphs using the theoretical bound leads to greedy colorings with a number of colors equal to the number of nodes in the graph, completely negating the advantage of using a probing method. Observe that the errors obtained with the larger value of d are not much smaller than the ones for the smaller value of d, because in both cases the quadratic forms are computed with target relative accuracy ϵ, so the probing error for the larger value of d is dominated by the error in the quadratic forms. In the case of the heuristic estimate, the execution time includes the time required to run the probing method for $d = 1, 2, 3$ in order to evaluate (5.28). The stopping criterion for the Krylov subspace method uses the upper bound (5.26). The execution time can be further reduced by using the estimate (5.27) for the Krylov subspace method, as the following experiment shows. The diagonalization time for the matrices used in this experiment can be found in the last column of Table 5.4. Note that for smaller matrices, diagonalization is often the fastest method, but the advantage of approximating the entropy with a probing method is already evident for matrices of size $n \approx 10000$.

5.5.3 Krylov Bound vs. Estimate

We fix an error tolerance $\epsilon = 10^{-5}$ and compare the performance of the geometric mean error estimate (5.27) with the theoretical upper bound (5.26) for the Krylov subspace method. The value of d for the probing method is selected using the heuristic estimate (5.28). The

Table 5.2: Comparison of the theoretical bound (5.11) against the heuristic estimate (5.28) for choosing d, using relative tolerance $\epsilon = 10^{-3}$ in the probing method. Top row: heuristic estimate. Bottom row: theoretical bound.

test matrix	n	error	d	colors	time (s)
`yeast`	2224	3.062e-04	3	222	0.952
		3.733e-05	25	2224	4.464
`minnesota`	2640	4.456e-04	5	24	0.196
		3.173e-05	18	255	0.779
`ca-HepTh`	8638	2.974e-04	3	252	2.273
		3.161e-05	27	8638	42.078
`bcsstk29`	13830	4.912e-05	3	176	2.292
		8.497e-05	12	2095	17.849

entropy error, execution time, and total number of polynomial and rational Krylov iterations are reported in Table 5.3. We can see that using the estimate instead of the theoretical bound moderately reduces the computational effort, while still attaining the requested accuracy ϵ on the entropy. In particular, observe that the number of rational Krylov iterations is significantly higher when using the upper bound (5.26). In this experiment, all linear systems are solved with direct methods and Cholesky factorizations are stored and reused.

Table 5.3: Comparison of the upper bound (5.26) against the geometric mean estimate (5.27) for Krylov methods used in the probing method, using relative tolerance $\epsilon = 10^{-5}$ for the probing method. Top row: geometric mean estimate. Bottom row: upper bound.

test matrix	n	error	poly iter	rat iter	time (s)
`yeast`	2224	4.405e-06	16140	748	7.095
		6.023e-06	16419	2237	8.231
`minnesota`	2640	5.728e-07	2983	289	1.535
		2.490e-06	2987	598	1.734
`ca-HepTh`	8638	8.195e-07	39481	2442	40.240
		2.395e-06	39747	6389	50.133
`bcsstk29`	13830	6.133e-06	4704	0	12.093
		7.545e-06	6363	44	16.047

5.5.4 Adaptive Hutch++

Here we test the performance and the accuracy of the adaptive variant of Hutch++ [86, Algorithm 3]. A relative accuracy ϵ is achieved by setting the absolute tolerance to $\epsilon S(\rho)$, where $S(\rho)$ is computed via diagonalization and considered as exact. The computational effort of Hutch++ is determined by the parameters N_r and N_H described in Section 4.1. In particular, the number of matrix-vector products is equal to N_r and the number of quadratic forms is

equal to $N_r + N_H$. We used $\epsilon = 10^{-2}, 10^{-3}$ as target tolerances and $\delta = 10^{-2}$ as failure probability. Matrix-vector products and quadratic forms are computed using the Krylov subspace method with the geometric mean estimate as a stopping criterion (5.27). In Table 5.4 we compare the results for the two tolerances, obtained as an average of 100 runs of the algorithm, including both the average and worst relative error. In the majority of cases, the worst error is below the input tolerance ϵ. We see that for $\epsilon = 10^{-2}$ the computation with Hutch++ is very fast for all test matrices; on the other hand, the cost becomes significantly higher for $\epsilon = 10^{-3}$, showing that the stochastic estimator quickly becomes inefficient as the required accuracy increases. Observe that for $\epsilon = 10^{-2}$ adaptive Hutch++ uses only 3 matvecs for all tests problems, which is the minimum amount that can be used by the implementation in [86]. This means that the internal criteria of the algorithm have determined that spending more matvecs in the low rank approximation is not beneficial, and hence the convergence of the method is roughly the same as for Hutchinson's estimator. A similar behavior can be observed in Table 5.7, and can be linked to the fact that for these test matrices ρ, the matrix function $-\rho \log \rho$ does not exhibit eigenvalue decay and hence cannot be well-approximated by low rank matrices. On the other hand, for problems where low rank approximation is more effective, stochastic estimators that exploit it, such as Hutch++, can have much faster convergence.

Table 5.4: Results for Hutch++ applied to some test matrices. For each matrix, the first and second row show the results for $\epsilon = 10^{-2}$ and $\epsilon = 10^{-3}$, respectively. The failure probability is $\delta = 10^{-2}$ in both cases. The last column contains the diagonalization times.

test matrix	avg error	worst error	N_r	$N_r + N_H$	time (s)	`eig` (s)
`yeast`	2.55e-03	9.94e-03	3	282	0.421	0.508
	3.56e-04	1.16e-03	1228	2160	19.735	
`minnesota`	3.46e-03	1.07e-02	3	154	0.111	0.819
	4.53e-04	1.26e-03	1854	2684	35.721	
`ca-HepTh`	2.83e-03	9.92e-03	3	81	0.205	22.170
	3.55e-04	9.14e-04	635	3968	66.350	
`bcsstk29`	1.84e-03	6.97e-03	3	38	0.171	86.696
	2.39e-04	8.24e-04	3	1883	13.276	

5.5.5 Larger Matrices

In this section we test the probing method and the adaptive Hutch++ algorithm on larger matrices, for which it would be extremely expensive to compute the exact entropy. In light of the results shown in Tables 5.2 and 5.3, we select the value of d for the probing method using the heuristic estimate (5.28) and we use the geometric mean estimate (5.27) for the Krylov subspace method. The results are reported in Tables 5.5 and 5.6 for the probing method, and in Table 5.7 for Hutch++. Figure 5.5 contains a more detailed breakdown of the execution time for the probing method used on the matrices of Table 5.5. We separate the time in

preprocessing, where we evaluate the heuristic (5.28) to select d, and the main run of the algorithm with the chosen value of d. The time for the main run is further divided in coloring, and polynomial and rational Krylov iterations. The time for the Cholesky factorizations refers to the whole process, since the factors are computed and stored when a certain pole for a rational Krylov iteration is encountered for the first time.

In Table 5.5, we consider matrices with a "large-world" sparsity structure, such as road networks, and we use a relative tolerance of $\epsilon = 10^{-4}$. For these matrices, for which the diameter and the average path length are relatively large, it is possible to quickly compute distance-d colorings with a relatively small number of colors. Moreover, Cholesky factorizations can be computed cheaply and have a small fill-in, so it is possible to rapidly solve linear systems using a direct method. On the other hand, in Table 5.6 we consider matrices with a "small-world" sparsity structure, more typical of social networks and scientific collaboration networks. These matrices require a much larger number of colors to construct distance-d colorings, even for small values of d. The cost of probing methods is thus significantly higher on this kind of problem. In Table 5.6, only polynomial Krylov iterations are used due to the low relative tolerance $\epsilon = 10^{-2}$, so it is never necessary to solve linear systems. Recall that these matrices also have a high fill-in in the Cholesky factorizations (see Table 5.1), so the conjugate gradient method with a suitable preconditioner is likely to be much more efficient than a direct method for solving a linear system.

Table 5.5: Results for the probing method applied to test matrices with large-world sparsity structure, using relative tolerance $\epsilon = 10^{-4}$.

test matrix	n	d	colors	poly iter	rat iter	time (s)
`ut2010`	115406	4	504	7070	919	79.60
`usroads`	126146	8	77	626	0	6.30
`ny2010`	350167	5	329	3914	15	111.39
`roadNet-PA`	1087562	8	106	827	0	84.46

Table 5.6: Results for the probing method applied to test matrices with small-world sparsity structure, using relative tolerance $\epsilon = 10^{-2}$.

test matrix	n	d	colors	poly iter	time (s)
`cond-mat-2005`	36458	3	1221	3883	13.809
`loc-Brightkite`	56739	3	3946	18765	90.564
`com-Amazon`	334863	3	625	1285	47.458

In Table 5.7 we show the results for the adaptive Hutch++ algorithm, using relative tolerance $\epsilon = 10^{-2}$. The results are obtained as an average of 100 runs of the algorithm. We can observe that the stochastic trace estimator works well for both large-world and small-world graphs, in contrast to the probing method.

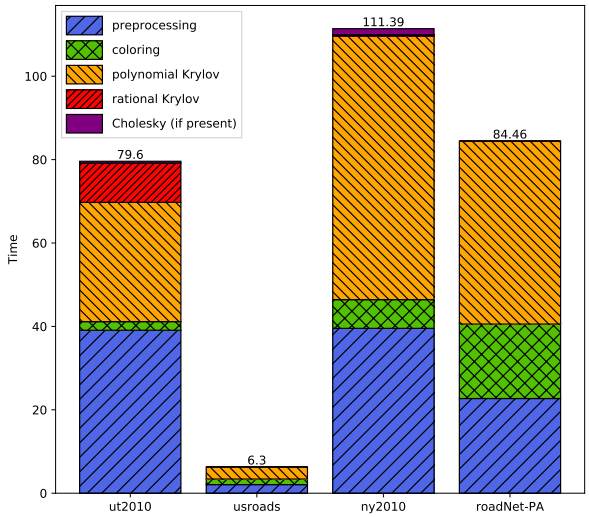

Figure 5.5: Breakdown of the execution time of the probing method for the test matrices in Table 5.5.

5.5.6 Algorithm Scaling

To investigate how the complexity of the algorithms scales with the matrix size, we compare the scaling of the probing method and the stochastic trace estimator on two different test problems with increasing dimension. The first one is the graph Laplacian of a 2D regular square grid, and the second one is the graph Laplacian of a Barabasi-Albert random graph, generated using the `pref` function of the CONTEST Matlab package [100]. For the probing method on the 2D grid, we use the optimal distance-d coloring with $\left\lceil \frac{1}{2}(d+1)^2 \right\rceil$ colors described in [41]. For both test problems, we use a relative tolerance $\epsilon = 10^{-4}$ for the probing method, and a relative tolerance $\epsilon = 10^{-2}$ and failure probability $\delta = 10^{-2}$ for adaptive Hutch++, averaging over 100 runs. The results are summarized in Figure 5.6, for graphs with a number of nodes from $n = 2^{10}$ to $n = 2^{20}$. As expected, the probing method is much more efficient in the case of the 2D grid, since the number of colors used in the distance-d colorings remains constant as n increases. On the other hand, for the Barabasi-Albert random graph, which has a small-world structure, the number of colors used in a distance-d coloring increases with the number of nodes, and hence the scaling for the probing method is significantly worse. The adaptive Hutch++ algorithm also has a better performance for the 2D grid, but the scaling in the problem size is good for both graph categories, since the number of vectors used in the trace approximation does not increase with the matrix dimension. However, the stochastic approach is only viable with a loose tolerance ϵ for this kind of problem since low rank approximation is not effective, as discussed in Section 5.5.4. The initial decrease in the execution time for Hutch++ as n increases is caused by the fact that

Table 5.7: Results for Hutch++ applied to large test matrices, with relative tolerance $\epsilon = 10^{-2}$ and failure probability $\delta = 10^{-2}$. The parameters N_r and N_H are defined in Section 4.1.

test matrix	n	N_r	$N_r + N_H$	time (s)
`ny2010`	350167	3	10	0.490
`usroads`	126146	3	14	0.190
`ny2010`	350167	3	10	0.487
`roadNet-PA`	1087562	3	8	1.126
`cond-mat-2005`	36458	3	34	0.448
`loc-Brightkite`	56739	3	42	4.576
`com-Amazon`	334863	3	11	1.171

the adaptive algorithm uses a larger number of vectors for the graphs with fewer nodes.

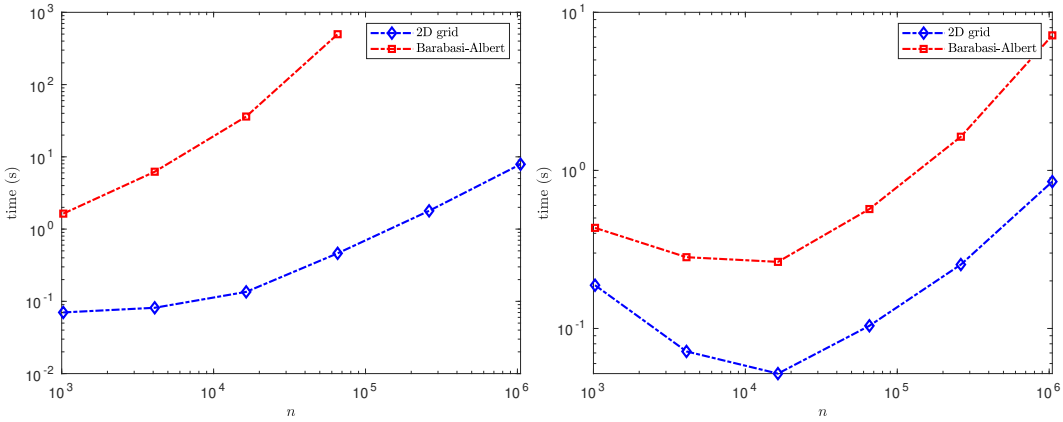

Figure 5.6: Execution times for the probing method (left, $\epsilon = 10^{-4}$) and the adaptive Hutch++ algorithm (right, $\epsilon = 10^{-2}$) on the graph Laplacian of a 2D regular grid and a Barabasi-Albert random graph, as a function of the number of nodes n.

5.5.7 Comparison with Stochastic Probing

Here we analyze the application of the stochastic probing technique for the computation of the von Neumann entropy.

In order to compare the accuracy and convergence of the stochastic probing approximation with the deterministic one, we repeat the experiment in Section 4.4.2. As a test matrix, we consider the density matrix $\rho = \mathcal{L}/\operatorname{Tr}(\mathcal{L})$, where $\mathcal{L}$ is the graph Laplacian of `minnesota`; see Table 5.1. We compare the accuracy of the deterministic probing estimator $\mathcal{T}_d(-\rho \log \rho)$, the stochastic probing estimator $\mathcal{T}_d^{\mathbb{1}}(f(A))$, for which one sample per color is used, and $\mathcal{T}_d^{\mathcal{N}}(f(A))$, for which $\mathcal{N} = \{N_1, \ldots, N_{n_c}\}$, and the N_i are chosen by using the criterion described in Remark 4.6. In figure 5.7, we show the absolute errors for different values of d,

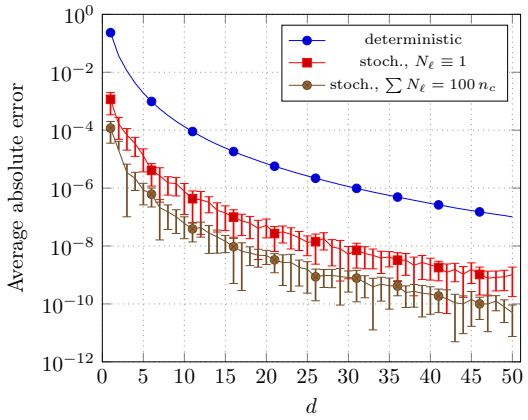

Figure 5.7: Absolute errors and asymptotic behavior for increasing d for the three methods in (4.43), applied to compute the entropy of the graph `minnesota`. For stochastic methods, it is shown the average error over 20 runs, together with a confidence interval.

ranging from 1 to 50. The distance-d colorings are computed via Algorithm 5. For stochastic methods, we consider the average absolute errors over 20 runs, together with a confidence interval obtained by adding and subtracting the empirical standard deviation.

We can see that the scaling of the error with d is the same for both deterministic and stochastic probing. However, due to the better scaling with the size, the error of the stochastic probing approximation has higher accuracy, which is further improved when using more than one probing vector per color, as discussed in Remark 4.6.

5.6 Conclusions and Further Developments

In this chapter, we have investigated two approaches for approximating the von Neumann entropy of a large, sparse, symmetric positive semidefinite matrix. The first method is a state-of-the-art randomized approach, while the second one is based on the idea of probing. Both methods require the computation of many quadratic forms involving the matrix function $f(A)$ with $f(x) = -x \log x$, an expensive task given the lack of smoothness of $f(x)$ at $x = 0$. We have examined the use of both polynomial and rational Krylov subspace methods, and combinations of the two. Pole selection strategy and several implementation aspects, such as heuristics and stopping criteria, have been investigated. Numerical experiments in which the entropy is computed for a variety of networks have been used to test the various approximation methods. Not surprisingly, the performance is affected by the structure of the underlying network, especially for the methods based on the probing idea. Our main conclusion is that the probing approach is better suited than the randomized one for graphs with a large-world structure, since they admit distance-d colorings with a relatively small number of colors. Such performance is further improved by using stochastic probing, as expected from the analysis in Chapter 4. Conversely, for complex networks with a small-world structure, the number of colors required for distance-d colorings is larger, so the probing approach becomes

more expensive. For this type of graphs, the randomized method is more competitive than the one based on probing, since it is less affected by the structure of the graph; however, for matrices in which low rank approximation cannot be exploited such as the graph Laplacians that we consider, randomized trace estimators are best suited for computing approximations with a relatively low accuracy, since their cost quickly grows as the requested accuracy is increased. The computation of the von Neumann entropy of density matrices expressed as matrix functions by using the same techniques can be subject of future research; some possible directions are outlined in Section 5.2.2, where we describe some situations in which the density matrix is itself expressed as a function of another (sparse) matrix.

Chapter 6

Conclusions

In this thesis we have investigated several mathematical and algorithmic topics arising in the computation of matrix functions of interest in physics and network analysis.

In the first part of thesis we analyzed the decay properties of the entries of certain matrix functions. We refined existing bounds on the entries of spectral projectors by expressing the projectors in terms of the sign function and exploiting an integral representation for the latter. We also provided bounds that take into account the eigenvalue distribution of the matrix argument. By removing the effect of isolated eigenvalues of maximum modulus from the spectral information, we were able to predict a superexponential decay in certain cases. The same techniques are also applied to other related examples, such as the Fermi-Dirac function and Cauchy-Stieltjes functions. While we are able to qualitatively describe the decay behavior of the entries in the projector, particularly in cases in which the spectral gap splits the spectrum into two symmetric intervals, further research is required in order to find quantitatively accurate bounds, especially for the case of nonsymmetric intervals.

The remainder of the thesis is devoted to the problem of trace estimation. First, we have analyzed, theoretically and experimentally, the accuracy of the stochastic probing estimator for approximating the trace of functions of sparse symmetric matrices. We have shown that the error scales, on average, at most with the square root of the size, improving on the linear scaling of the deterministic probing approach. We compared the performance with some important trace estimators from the literature and pointed out when stochastic probing can outperform the estimators based on low rank approximations. In particular, if the matrix function exhibits a fast decay of the eigenvalues, then we can compute the trace with high accuracy by using stochastic estimators based on low rank approximations, such as Hutch++ and XTrace, while, for slowly decaying eigenvalues, stochastic probing yields a better performance. Further developments on this topic might include a new and faster way to get the colorings.

Finally, we have considered the computation of the von Neumann entropy by using trace estimators and Krylov methods. We analyzed the accuracy of the deterministic probing approximation by developing a new bound for the best uniform polynomial approximation of $f(x) = -x \log x$ on an interval of nonngegative numbers. We obtained a posteriori error bounds and a truncation strategy for the approximation of quadratic forms $\boldsymbol{b}^T f(A)\boldsymbol{b}$ by mix-

ing polynomial and rational Krylov iterations. Our numerical experiments show the viability of the computation of the von Neumann entropy by combining trace estimators with Krylov methods. When using the adaptive Hutch++ algorithm, we rapidly achieve low accuracy with few operations indipendently of the matrix argument, while the task becomes prohibitive for a higher accuracy. On the other hand, if the sparsity pattern of the matrix argument is that of a large-world network, we are able to attain small errors with a moderate computational cost. This can be further enhanced by using stochastic probing, in view of the results of Chapter 4. In principle, the same techniques may be applied to the case of density matrices expressed as matrix functions, prominent examples of which are the Gibbs state and Fermi-Dirac function.

Bibliography

[1] J. Aarons and C. K. Skylaris. Electronic annealing Fermi operator expansion for DFT calculations on metallic systems. *J. Chem. Phys*, 148(7):074107, 2018.

[2] J. Alahmadi, M. Pranić, and L. Reichel. Rational Gauss quadrature rules for the approximation of matrix functionals involving Stieltjes functions. *Numer. Math.*, 151(2):443–473, 2022.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[4] G. B. Arfken and H. J. Weber. *Mathematical Methods for Physicists*. Harcourt/Academic Press, Burlington, MA, fifth edition, 2001.

[5] E. Aune, D. P. Simpson, and J. Eidsvik. Parameter estimation in high dimensional Gaussian distributions. *Stat. Comput.*, 24:247–263, 2014.

[6] H. Avron and S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM*, 58(2):1–34, 2011.

[7] R. Babich, R. C. Brower, M. A. Clark, G. T. Fleming, J. C. Osborn, C. Rebbi, and D. Schaich. Exploring strange nucleon form factors on the lattice. *Phys. Rev. D*, 85(5):054510, 2012.

[8] T. Bagby. On interpolation by rational functions. *Duke Math. J.*, 36:95–104, 1969.

[9] A. G. Baskakov. Estimates for the elements of inverse matrices, and the spectral analysis of linear operators. *Izv. Ross. Akad. Nauk Ser. Mat.*, 61(6):3–26, 1997.

[10] I. Bengtsson and K. Zyczkowski. *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge University Press, 2006.

[11] M. Benzi. Localization in matrix computations: theory and applications. In *Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications*, volume 2173 of *Lecture Notes in Math.*, pages 211–317. Springer, Cham, 2016.

[12] M. Benzi. Some uses of the field of values in numerical analysis. *Boll. Unione Mat. Ital.*, 14(1):159–177, 2021.

[13] M. Benzi and P. Boito. Matrix functions in network analysis. *GAMM-Mitt.*, 43(3):e202000012, 2020.

[14] M. Benzi, P. Boito, and N. Razouk. Decay properties of spectral projectors with applications to electronic structure. *SIAM Rev.*, 55(1):3–64, 2013.

[15] M. Benzi and G. H. Golub. Bounds for the entries of matrix functions with applications to preconditioning. *BIT*, 39(3):417–438, 1999.

[16] M. Benzi and N. Razouk. Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices. *Electron. Trans. Numer. Anal.*, 28:16–39, 2007/08.

[17] M. Benzi and M. Rinelli. Refined decay bounds on the entries of spectral projectors associated with sparse Hermitian matrices. *Linear Algebra Appl.*, 647:1–30, 2022.

[18] M. Benzi, M. Rinelli, and I. Simunec. Computation of the von Neumann entropy of large matrices via trace estimators and rational Krylov methods. *Numer. Math.*, 155(3-4):377–414, 2023.

[19] M. Benzi and V. Simoncini. Decay bounds for functions of Hermitian matrices with banded or Kronecker structure. *SIAM J. Matrix Anal. Appl.*, 36(3):1263–1282, 2015.

[20] M. Benzi and I. Simunec. Rational Krylov methods for fractional diffusion problems on graphs. *BIT*, 62(2):357–385, 2022.

[21] M. Berljafa and S. Güttel. Generalized rational Krylov decompositions with an application to rational approximation. *SIAM J. Matrix Anal. Appl.*, 36(2):894–916, 2015.

[22] M. Berljafa and S. Güttel. Parallelization of the rational Arnoldi algorithm. *SIAM J. Sci. Comput.*, 39(5):S197–S221, 2017.

[23] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*, volume 9 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. Revised reprint of the 1979 original.

[24] D. R. Bowler and T. Miyazaki. $O(N)$ methods in electronic structure calculations. *Reports on Progress in Physics*, 75(3):036503, Feb 2012.

[25] S. L. Braunstein, S. Ghosh, and S. Severini. The Laplacian of a graph as a density matrix: a basic combinatorial approach to separability of mixed states. *Ann. Comb.*, 10(3):291–317, 2006.

[26] T. Chen and E. Hallman. Krylov-aware stochastic trace estimation. *SIAM J. Matrix Anal. Appl.*, 44(3):1218–1244, 2023.

[27] H. Choi, J. He, H. Hu, and Y. Shi. Fast computation of von Neumann entropy for large-scale graphs via quadratic approximations. *Linear Algebra Appl.*, 585:127–146, 2020.

[28] A. Cortinovis and D. Kressner. On randomized trace estimates for indefinite matrices with an application to determinants. *Found. Comput. Math.*, 22(3):875–903, 2022.

[29] M. Cramer and J. Eisert. Correlations, spectral gap and entanglement in harmonic quantum systems on generic lattices. *New Journal of Physics*, 8(5):71–71, 2006.

[30] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *J. Funct. Anal.*, 244(2):668–690, 2007.

[31] M. Crouzeix and C. Palencia. The numerical range is a $(1 + \sqrt{2})$-spectral set. *SIAM J. Matrix Anal. Appl.*, 38(2):649–655, 2017.

[32] T. A. Davis and Y. Hu. The University of Florida sparse matrix collection. *ACM Trans. Math. Software*, 38(1):Art. 1, 25, 2011.

[33] M. De Domenico and J. Biamonte. Spectral entropies as information-theoretic tools for complex network comparison. *Phys. Rev. X*, 6:041062, 2016.

[34] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora. Structural reducibility of multilayer networks. *Nature Comm.*, 6(1), 2015.

[35] S. Demko, W. F. Moss, and P. W. Smith. Decay rates for inverses of band matrices. *Math. Comp.*, 43(168):491–499, 1984.

[36] R. Díaz Fuentes, M. Donatelli, C. Fenu, and G. Mantica. Estimating the trace of matrix functions with application to complex networks. *Numer. Algorithms*, 92(1):503–522, 2023.

[37] E. N. Epperly, J. A. Tropp, and R. J. Webber. XTrace: Making the most of every sample in stochastic trace estimation. *SIAM J. Matrix Anal. Appl.*, 45(1):1–23, 2024.

[38] A. Eremenko and P. Yuditskii. Uniform approximation of $\operatorname{sgn} x$ by polynomials and entire functions. *J. Anal. Math.*, 101:313–324, 2007.

[39] A. Eremenko and P. Yuditskii. Polynomials of the best uniform approximation to $\operatorname{sgn}(x)$ on two intervals. *J. Anal. Math.*, 114:285–315, 2011.

[40] E. Estrada and D. J. Higham. Network properties revealed through matrix functions. *SIAM Rev.*, 52(4):696–714, 2010.

[41] G. Fertin, E. Godard, and A. Raspaud. Acyclic and k-distance coloring of the grid. *Inform. Process. Lett.*, 87(1):51–58, 2003.

[42] M. Fiedler and H. Schneider. Analytic functions of M-matrices and generalizations. *Linear and Multilinear Algebra*, 13(3):185–201, 1983.

[43] A. Frommer, M. N. Khalil, and G. Ramirez-Hidalgo. A multilevel approach to variance reduction in the stochastic estimation of the trace of a matrix. *SIAM J. Sci. Comput.*, 44(4):A2536–A2556, 2022.

[44] A. Frommer, M. Rinelli, and M. Schweitzer. Analysis of stochastic probing methods for estimating the trace of functions of sparse symmetric matrices, arXiv preprint arXiv:2308.07722 [math.NA], 2023.

[45] 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. *Numer. Linear Algebra Appl.*, 25(4):e2131, 17, 2018.

[46] A. Frommer, C. Schimmel, and M. Schweitzer. Non-Toeplitz decay bounds for inverses of Hermitian positive definite tridiagonal matrices. *Electron. Trans. Numer. Anal.*, 48:362–372, 2018.

[47] A. Frommer, C. Schimmel, and M. Schweitzer. Analysis of probing techniques for sparse approximation and trace estimation of decaying matrix functions. *SIAM J. Matrix Anal. Appl.*, 42(3):1290–1318, 2021.

[48] A. Frommer and V. Simoncini. Matrix functions. In *Model Order Reduction: Theory, Research Aspects and Applications*, volume 13 of *Math. Ind.*, pages 275–303. Springer, Berlin, 2008.

[49] W. H. J. Fuchs. On Chebyshev approximation on sets with several components. In *Aspects of Contemporary Complex Analysis (Proc. NATO Adv. Study Inst., Univ. Durham, Durham, 1979)*, pages 399–408. Academic Press, London-New York, 1980.

[50] A. S. Gambhir, A. Stathopoulos, and K. Orginos. Deflation as a method of variance reduction for estimating the trace of a matrix inverse. *SIAM J. Sci. Comput.*, 39(2):A532–A558, 2017.

[51] A. Ghavasieh and M. De Domenico. Statistical physics of network structure and information dynamics. *J. Phys. Complex.*, 3(1):011001, 2022.

[52] A. Ghavasieh and M. De Domenico. Generalized network density matrices for analysis of multiscale functional diversity. *Phys. Rev. E*, 107(4), 2023.

[53] M. B. Giles. Multilevel Monte Carlo methods. *Acta Numer.*, 24:259–328, 2015.

[54] A. Girard. A fast "Monte-Carlo cross-validation" procedure for large least squares problems with noisy data. *Numer. Math.*, 56:1–23, 1989.

[55] A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17:Paper No. 117, 65, 2016.

[56] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[57] S. Güttel. *Rational Krylov Methods for Operator Functions*. PhD thesis, Technische Universität Bergakademie Freiberg, Germany, 2010. Dissertation available as MIMS Eprint 2017.39.

[58] S. Güttel. Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. *GAMM-Mitt.*, 36(1):8–31, 2013.

[59] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[60] N. Hale, N. J. Higham, and L. N. Trefethen. Computing $\mathbf{A}^\alpha$, $\log(\mathbf{A})$, and related matrix functions by contour integrals. *SIAM J. Numer. Anal.*, 46(5):2505–2523, 2008.

[61] E. Hallman and D. Troester. A multilevel approach to stochastic trace estimation. *Linear Algebra Appl.*, 638:125–149, 2022.

[62] P. R. Halmos. The theory of unbiased estimation. *Ann. Math. Statistics*, 17:34–43, 1946.

[63] I. Han, D. Malioutov, and J. Shin. Large-scale log-determinant computation through stochastic Chebyshev expansions. In *32nd International Conference on Machine Learning*, volume 2, pages 908–917. PMLR, 2015.

[64] L. Han, F. Escolano, E. R. Hancock, and R. C. Wilson. Graph characterizations from von Neumann entropy. *Pattern Recognit. Lett.*, 33(15):1958–1967, 2012.

[65] M. Hasson. The degree of approximation by polynomials on some disjoint intervals in the complex plane. *J. Approx. Theory*, 144(1):119–132, 2007.

[66] P. Henrici. *Applied and Computational Complex Analysis. Vol. 1.* Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988. Reprint of the 1974 original, A Wiley-Interscience Publication.

[67] N. J. Higham. *Functions of Matrices. Theory and Computation*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[68] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.*, 18(3):1059–1076, 1989.

[69] A. Iserles. How large is the exponential of a banded matrix? *New Zealand J. Math.*, 29(2):177–192, 2000.

[70] W. Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.*, 76:3168–3171, Apr 1996.

[71] D. Kressner. Block algorithms for reordering standard and generalized Schur forms. *ACM Trans. Math. Software*, 32(4):521–532, 2006.

[72] M. Kubale. *Graph colorings*, volume 352 of *Contemp. Math.* American Mathematical Society, Providence, RI, 2004.

[73] J. Laeuchli and A. Stathopoulos. Extending hierarchical probing for computing the trace of matrix inverses. *SIAM J. Sci. Comput.*, 42(3):A1459–A1485, 2020.

[74] L. D. Landau and E. M. Lifshitz. *Statistical Physics*. Pergamon Press, London, 1958.

[75] X. Li, Y. Shi, and I. Gutman. *Graph Energy*. Springer, New York, 2012.

[76] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[77] G. Mantica. Quantum dynamical entropy and an algorithm by Gene Golub. *Electr. Trans. Numer. Anal.*, 28:190–205, 2008.

[78] P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: foundations and algorithms. *Acta Numer.*, 29:403–572, 2020.

[79] S. Massei and L. Robol. Rational Krylov for Stieltjes matrix functions: convergence and pole selection. *BIT*, 61(1):237–273, 2021.

[80] G. Meinardus. *Approximation of Functions: Theory and Numerical Methods*. Expanded translation of the German edition. Translated by Larry L. Schumaker. Springer Tracts in Natural Philosophy, Vol. 13. Springer-Verlag New York, Inc., New York, 1967.

[81] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM, 2021.

[82] C. Morningstar, J. Bulava, J. Foley, K. J. Juge, D. Lenkner, M. Peardon, and C. H. Wong. Improved stochastic estimation of quark propagation with Laplacian Heaviside smearing in lattice QCD. *Phys. Rev. D*, 83(11):114505, 2011.

[83] A. M. N. Niklasson. Density matrix methods in linear scaling electronic structure theory. In R. Zalesny, M. G. Papadopoulos, P. G. Mezey, and J. Leszczynski, editors, *Linear-Scaling Techniques in Computational Chemistry and Physics: Methods and Applications*, pages 439–473. Springer, New York, 2011.

[84] D. Palitta, S. Pozza, and V. Simoncini. The short-term rational Lanczos method and applications. *SIAM J. Sci. Comput.*, 44(4):A2843–A2870, 2022.

[85] M. Penrose. *Random Geometric Graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

[86] D. Persson, A. Cortinovis, and D. Kressner. Improved variants of the Hutch++ algorithm for trace estimation. *SIAM J. Matrix Anal. Appl.*, 43(3):1162–1185, 2022.

[87] S. Pozza and V. Simoncini. Inexact Arnoldi residual estimates and decay properties for functions of non-Hermitian matrices. *BIT*, 59(4):969–986, 2019.

[88] M. S. Pranić and L. Reichel. Rational Gauss quadrature. *SIAM J. Numer. Anal.*, 52(2):832–851, 2014.

[89] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

[90] F. Roosta-Khorasani and U. Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Found. Comput. Math.*, 15(5):1187–1212, 2015.

[91] W. Rudin. *Real and Complex Analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[92] A. Ruhe. Rational Krylov algorithms for nonsymmetric eigenvalue problems. In *Recent Advances in Iterative Methods*, volume 60 of *IMA Vol. Math. Appl.*, pages 149–164. Springer, New York, 1994.

[93] A. K. Saibaba, A. Alexanderian, and I. C. F. Ipsen. Randomized matrix-free trace and log-determinant estimators. *Numer. Math.*, 137(2):353–395, 2017.

[94] C. Schimmel. *Bounds for the Decay in Matrix Functions and its Exploitation in Matrix Computations*. PhD thesis, Bergische Universität Wuppertal, Wuppertal, Germany, 2019.

[95] M. Schweitzer. Decay bounds for Bernstein functions of Hermitian matrices with applications to the fractional graph Laplacian. *Electron. Trans. Numer. Anal.*, 55:438–454, 2022.

[96] J. Sexton and D. Weingarten. Systematic expansion for full QCD based on the valence approximation, arXiv preprint arXiv:hep-lat/9411029, 1995.

[97] R. Shankar. *Principles of Quantum Mechanics*. Springer New York, NY, 2nd edition edition, 1994.

[98] A. Stathopoulos, J. Laeuchli, and K. Orginos. Hierarchical probing for estimating the trace of the matrix inverse on toroidal lattices. *SIAM J. Sci. Comput.*, 35(5):S299–S322, 2013.

[99] J. M. Tang and Y. Saad. A probing method for computing the diagonal of a matrix inverse. *Numer. Linear Algebra Appl.*, 19(3):485–501, 2012.

[100] A. Taylor and D. J. Higham. CONTEST: A controllable test matrix toolbox for MATLAB. *ACM Trans. Math. Softw.*, 35(4), 2009.

[101] C. Thron, S. Dong, K. Liu, and H. Ying. Padé-Z_2 estimator of determinants. *Phys. Rev. D*, 57(3):1642, 1998.

[102] S. Ubaru and Y. Saad. Applications of trace estimation techniques. In *High Performance Computing in Science and Engineering: Third International Conference, HPCSE 2017, Karolinka, Czech Republic, May 22–25, 2017, Revised Selected Papers*, pages 19–33. Springer, 2018.

[103] J. von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton, 1955. Translated by Robert T. Beyer.

[104] A. Wehrl. General properties of entropy. *Rev. Mod. Phys.*, 50:221–260, 1978.

[105] T. Whyte, A. Stathopoulos, E. Romero, and K. Orginos. Optimizing shift selection in multilevel monte carlo for disconnected diagrams in lattice QCD. *Comput. Phys. Commun.*, 294:108928, 2024.

[106] D. V. Widder. *The Laplace Transform*. Princeton Mathematical Series, vol. 6. Princeton University Press, Princeton, N. J., 1941.

[107] T. P. Wihler, B. Bessire, and A. Stefanov. Computing the entropy of a large matrix. *J. Phys. A*, 47(24):245201, 15, 2014.

[108] K. Wimmer, Y. Wu, and P. Zhang. Optimal query complexity for estimating the trace of a matrix. In *Automata, Languages, and Programming: 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I 41*, pages 1051–1062. Springer, 2014.