Bergische Universität Wuppertal

Fakultät für Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational
Mathematics (IMACM)

Preprint BUW-IMACM 20/34

A. Frommer, C.Schimmel and M. Schweitzer

Analysis of probing techniques for sparse approximation and trace estimation of decaying matrix functions

September 7, 2020

http://www.math.uni-wuppertal.de

ANALYSIS OF PROBING TECHNIQUES FOR SPARSE APPROXIMATION AND TRACE ESTIMATION OF DECAYING MATRIX FUNCTIONS[*]

ANDREAS FROMMER[†], CLAUDIA SCHIMMEL[†], AND MARCEL SCHWEITZER[‡]

Abstract. The computation of matrix functions $f(A)$, or related quantities like their trace, is an important but challenging task, in particular for large and sparse matrices A. In recent years, probing methods have become an often considered tool in this context, as they allow to replace the computation of $f(A)$ or $\mathrm{tr}(f(A))$ by the evaluation of (a small number of) quantities of the form $f(A)v$ or $v^T f(A)v$, respectively. These tasks can then efficiently be solved by standard techniques like, e.g., Krylov subspace methods. It is well-known that probing methods are particularly efficient when $f(A)$ is *approximately sparse*, e.g., when the entries of $f(A)$ show a strong off-diagonal decay, but a rigorous error analysis is lacking so far. In this paper we develop new theoretical results on the existence of sparse approximations for $f(A)$ and error bounds for probing methods based on graph colorings. As a by-product, by carefully inspecting the proofs of these error bounds, we also gain new insights into when to stop the Krylov iteration used for approximating $f(A)v$ or $v^T f(A)v$, thus allowing for a practically efficient implementation of the probing methods.

Key words. matrix functions, sparse approximation, trace, decay bounds, graph coloring, probing method, Krylov subspace method

AMS subject classifications. 05C12, 05C15, 15A16, 65F50, 65F60

1. Introduction.

Matrix functions $f(A)$, where $f : \mathbb{C} \to \mathbb{C}$ is a scalar function and $A \in \mathbb{C}^{n\times n}$ is a square matrix, play an essential role in many areas of science and engineering. The inverse A^{-1} is the most prominent example, another important case is the matrix exponential $f(A) = \exp(A)$, which is used for the numerical solution of time-dependent differential equations or the analysis of dynamical systems [23]. For the computation of communicability measures in network analysis, the matrix exponential and the resolvent, generated by the scalar function $f(z) = (\alpha - z)^{-1}$ with $\alpha \in \mathbb{C}$ are widely used [17–19]. The matrix sign function $f(A) = \mathrm{sign}(A)$ has applications in control theory [23, 46] and lattice quantum chromodynamics [7, 16, 39]. Inverse fractional powers $f(A) = A^{-\alpha}$ with $\alpha \in (0, 1)$ are strongly related to the matrix sign function and arise in generalized eigenvalue problems [40, Section 15.10], fractional differential equations [9] or sampling from multivariate Gaussian distributions [42].

For many of these applications, the explicit computation of $f(A)$ is not feasible as the matrix A is typically large and sparse, while $f(A)$ is a dense matrix. Therefore, one has to resort to approximation techniques when $f(A)$ or a related quantity like $f(A)b$, $b \in \mathbb{C}^n$, the diagonal $\mathrm{diag}(f(A))$ or the trace $\mathrm{tr}(f(A))$ is required. This work focuses on sparse approximations for the whole matrix $f(A)$ on the one hand and on approximating $\mathrm{tr}(f(A))$ on the other hand. Computing the trace $\mathrm{tr}(f(A))$ is a relevant task. For example, the trace of the inverse is required in the study of fractals [48], generalized cross-validation and its applications [26, 29], or when computing disconnected fermion loop contributions in lattice quantum chromodynamics (QCD) [14, 50]. In network analysis, the Estrada index—a total centrality measure for networks—is

[*] This work was partially supported by Deutsche Forschungsgemeinschaft through the Collaborative Research Centre SFB-TRR55 “Hadron Physics from Lattice QCD”.

[†] Department of Mathematics, Bergische Universität Wuppertal, 42097 Wuppertal, Germany, {frommer,schimmel}@math.uni-wuppertal.de

[‡] Mathematisch-Naturwissenschaftliche Fakultät, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany. E-mail: marcel.schweitzer@hhu.de.

defined as the trace of the exponential of the adjacency matrix of a graph [18,25] and an analogous measure is given by the trace of the resolvent [17, Section 8.1]. For Hermitian positive definite matrices A, one can compute the log-determinant $\log(\det(A))$ as the trace of the logarithm of A. Amongst others, the log-determinant is needed in machine learning and related fields [44,47]. Plenty of further applications can be found in [55,56], e.g.

In recent years, probing methods [8,35,52,54] have emerged as an important tool in this context. They obtain approximations by evaluating a small number of matrix-vector products or bilinear forms involving $f(A)$, which can be done by standard techniques (e.g., Krylov subspace methods). We briefly summarize the main idea of these methods in the following.

1.1. Probing methods.

Recall that the (directed) graph $G(A) = (V, E)$ of a sparse matrix $A \in \mathbb{C}^{n\times n}$ is given by the vertices $V = \{1, \dots, n\}$ and egdes $E = \{(i, j) : a_{ij} \neq 0, i \neq j\}$. By $\mathrm{d}(i, j)$ we denote the geodesic distance, i.e., the length of the shortest path, from node i to node j in $G(A)$ and by $\bar{\mathrm{d}}(i, j)$ the geodesic distance in the corresponding undirected graph $|G(A)|$ which results from $G(A)$ by removing the direction of the edges.

Given a partitioning of the nodes V of $G(A)$, i.e.,

$$V = V_1 \cup \ldots \cup V_m,\ \ V_\ell \neq \emptyset \text{ for } \ell = 1, \dots, m \text{ and } V_\ell \cap V_p = \emptyset \text{ for } \ell \neq p, \tag{1.1}$$

the corresponding probing vectors are defined as

$$v_\ell := \sum_{i \in V_\ell} e_i,\ \ell \in \{1, \dots, m\} \quad (e_i \text{ is the } i\text{th canonical unit vector}). \tag{1.2}$$

The vectors v_ℓ can be used to, e.g., estimate $\mathrm{tr}(f(A))$ via

$$\mathrm{tr}(f(A)) \approx \mathcal{T}(f(A)) := \sum_{\ell=1}^{m} v_\ell^H f(A) v_\ell, \tag{1.3}$$

or even construct a sparse approximation to $f(A)$ itself via

$$[f(A)^{[d]}]_{ij} := \begin{cases} [f(A)v_\ell]_i \text{ for } j \in V_\ell & \text{if } \bar{\mathrm{d}}(i, j) \leq d, \\ 0 & \text{if } \bar{\mathrm{d}}(i, j) > d, \end{cases} \tag{1.4}$$

where d is a prescribed distance threshold. We refer to, e.g., [8,35,52,54] for detailed discussions of such probing approaches and just expose the main motiviation: In the (unrealistic) situation that $f(A)$ is a sparse matrix with $[f(A)]_{ij} = 0$ for $\mathrm{d}(i, j) > d$, if the sets V_ℓ are chosen such that $[f(A)]_{ij} = 0$ for $i, j \in V_\ell, i \neq j$, both approximations (1.3) and (1.4) are actually exact. Therefore, if $f(A)$ is *approximately sparse*, and the V_ℓ are built such that $[f(A)]_{ij}$ is small for $i, j \in V_\ell, i \neq j$, we can expect probing methods to yield accurate approximations.

1.2. Exponential decay in matrix functions.

To make the notion of $f(A)$ precise, recall that $f(A)$ is defined if for all eigenvalues λ of A all derivatives of f at λ up to order $\nu(\lambda) - 1$ exist, where $\nu(\lambda)$ is the multiplicity of the elementary factor $(z - \lambda)$ in the minimal polynomial of A, see [33], e.g. We tacitly assume that this is always fulfilled whenever we consider $f(A)$. Note that $f(A)$ is then given as the polynomial in A which interpolates f on the spectrum of A in the Hermite sense.

One special form of approximate sparsity in $f(A)$ that is frequently encountered in practice is *exponential decay* of the entries of $f(A)$ away from the sparsity pattern of A.

Definition 1.1. *Let $A \in \mathbb{C}^{n\times n}$ and let f be defined on the spectrum of A. The matrix $f(A)$ has exponential decay (away from the sparsity pattern of A) if*

$$|[f(A)]_{ij}| \leq Cq^{\mathrm{d}(i,j)} \text{ for } i,j \in \{1,\ldots,n\}, \tag{1.5}$$

where d *is the geodesic distance in $G(A)$.*

Of course, for given A the relation (1.5) can always be satisfied if we choose C and $q \in (0,1)$ large enough. To be meaningful, the concept of exponential decay therefore implicitly assumes that C and, in particular, q are not too large or that (1.5) holds uniformly for a whole, possibly infinite, family of matrices.

Decay in matrix functions has been studied extensively, starting with [13], where accurate exponential decay bounds were presented for inverses of banded (Hermitian positive definite) matrices. Lots of other results and decay bounds for different types of functions and matrices can be found, e.g., in [4–6, 15, 21, 22, 38, 43]. Many of them are based on properties of polynomial approximations to f. Indeed, if i and j have distance $\mathrm{d}(i,j)$ in the graph $G(A)$, then for every polynomial p_s of degree at most $s = \mathrm{d}(i,j) - 1$ we have $[p_s(A)]_{ij} = 0$, see [5], e.g., which implies

$$|[f(A)]_{ij}| = |[f(A)]_{ij} - [p_s(A)]_{ij}| \leq \|f(A) - p_s(A)\|_2.$$

Herein, $\|f(A) - p_s(A)\|_2$ can be bounded further due to the following important approximation result which uses the numerical range $\mathcal{W}(A) = \{x^H Ax : \|x\| = 1\}$.

Theorem 1.2. *Let $A \in \mathbb{C}^{n\times n}$ and let $g : \mathbb{C} \to \mathbb{C}$ be defined on the spectrum of A. Then*

$$\|g(A)\|_2 \leq K \max_{z\in W(A)} |g(z)|, \tag{1.6}$$

where $K = 1$ if A is normal and $K = 1 + \sqrt{2}$ otherwise.

Note that this result is almost a triviality for A Hermitian, while the general case is much more involved, see, e.g., [11]. Applying Theorem 1.2 to $g = f - p_s$ immediately gives the following result which relates the accuracy of polynomial approximation to exponential decay in the matrix function. We will use it several times in this paper.

Theorem 1.3. *Assume that*

$$\min_{p_s\in\Pi_s} \max_{z\in W} |f(z) - p_s(z)| \leq Cq^s, \tag{1.7}$$

where Π_s is the set of all polynomials with degree $\leq s$. Then, if $\mathcal{W}(A) \subseteq W$ we have

$$|[f(A)]_{ij}| \leq \|f(A) - p_s(A)\|_2 \leq KCq^s \text{ whenever } \mathrm{d}(i,j) > s.$$

Thus, uniform exponential decay bounds for a family of matrices can be obtained if there is a common superset W of their numerical ranges for which (1.7) holds, as it is, e.g., the case for the results in [5, 13, 21].

In our error analysis to come we will sometimes distinguish between general exponential decay bounds for $f(A)$ and bounds which are explicitly based on (1.7).

1.3. Outline of the paper. The main goal of this paper is to obtain guidelines for choosing the sets V_ℓ in (1.1) and using this information to derive rigorous error bounds for the resulting approximations (1.3) and (1.4) in case that $f(A)$ exhibits an exponential decay property. In addition, our analysis also sheds light onto when to stop Krylov subspace iterations used for approximating $f(A)v_\ell$ or $v_\ell^H f(A)v_\ell$, respectively, in order to reach an implementation that is as efficient as possible without sacrificing accuracy.

This paper is organized as follows. In Section 2 we discuss the distance-d graph coloring problem as it forms the basis of the discussed probing methods. In Section 3 we first give some new theoretical results on the existence of sparse approximations of matrix functions and then prove new error bounds for the approximation (1.4) for $f(A)$. Section 4 covers error bounds for the approximation (1.3) of $\operatorname{tr}(f(A))$, while Section 5 uses the insights from Section 3 to develop stopping criteria for the Krylov subspace approximation inside the probing method. We illustrate the quality of the derived bounds in numerical experiments reported in Section 6. Concluding remarks are given in Section 7.

2. Distance-d coloring.

The quality of the approximations (1.3) and (1.4) crucially depends on the partitioning (1.1). If $f(A)$ has exponential decay with respect to $G(A)$, good partitionings can be obtained via graph colorings.

Definition 2.1. *A distance-d coloring of a graph* $G = (V, E)$ *is a mapping* $\operatorname{col} : V \to \{1, \dots, m\}$ *such that* $\operatorname{col}(i) \neq \operatorname{col}(j)$ *if* $\operatorname{d}(i, j) \leq d$. *A distance-d coloring is optimal if the number m of colors is minimal among all distance-d colorings of G.*

For $d = 1$, the computation of an optimal distance-d coloring corresponds to the classical graph coloring problem, which is known to be NP complete for general graphs [34]. In our setting, we are mainly interested in low-cost methods for computing a distance-d coloring with a sufficiently small number of colors. Efficient ways for computing such colorings of graphs are usually based on greedy strategies, see, e.g., [34]. For example, a distance-d coloring of a graph G with $V = \{w_1, \dots, w_n\}$ can be obtained via $\operatorname{col}(w_1) = 1$ and $\operatorname{col}(w_i) = \min\{k > 0 : k \neq \operatorname{col}(w) \text{ for all } w \in W_i\}$ for $i = 2, \dots, n$ where

$$W_i := \{w \in \{w_1, \dots, w_{i-1}\} : \operatorname{d}(w_i, w) \leq d\}. \tag{2.1}$$

This coloring uses at most $\Delta(G)^d + 1$ colors and can be implemented with cost $\mathcal{O}(\Delta(G)^d n)$, where $\Delta(G)$ is the maximal degree of G [49, Proposition 4.2]. In the next two sections, we discuss special classes of graphs where a (not necessarily optimal) distance-d coloring can be obtained with cost $\mathcal{O}(n)$.

2.1. Distance-d colorings for graphs of banded matrices.

Let A be a banded matrix with semi-bandwidth β, i.e. $[A]_{ij} = 0$ whenever $|i - j| > \beta$. Then it is easy to verify that a distance-d coloring for $G(A)$ with $m = d\beta + 1$ colors is given by

$$\operatorname{col}(i) = (i - 1) \bmod (d\beta + 1) + 1, \quad i = 1, \dots, n, \tag{2.2}$$

and this coloring is optimal if all entries within the band of A are non-zero. If A is sparse but not banded with small β, one can first determine an ordering of the nodes which aims at obtaining a (relatively) small bandwidth for the correspondingly permuted matrix and then define the coloring via (2.2) on the permuted nodes. Finding a permutation resulting in a small bandwidth is an important topic in the context of direct solvers for linear systems, and lots of low-cost methods have been proposed

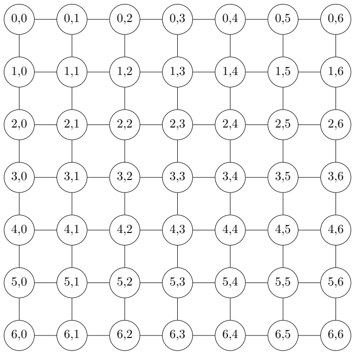

Fig. 2.1: Two-dimesional 7×7 lattice, where each node is defined by two coordinates $0 \leq w_1, w_2 \leq 6$.

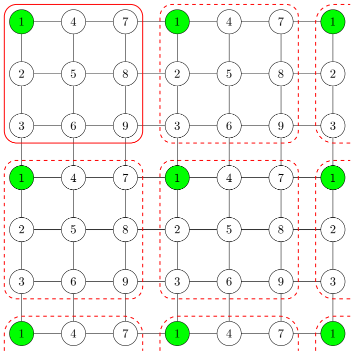

Fig. 2.2: Distance-2 coloring produced by Theorem 2.2.

over the years; see, e.g., [10,12,24,36,45,51]. A heuristic based on level sets is at the basis of the classical Cuthill-McKee algorithm [12] with cost $\mathcal{O}(|V|+|E|)$, and we refer to [30] for an overview and comparison of various other recent low-cost heuristics. The cost for computing the coloring is dominated by the cost for the computation of the permutation of the nodes.

2.2. Distance-d colorings for regular lattices.

As another special case, assume that the graph $G=(V,E)$ is a regular D-dimensional lattice for $D>1$. For $D=1$, the adjacency matrix is tridiagonal, a situation already covered by the banded case discussed before.

First, note that the greedy coloring approach can be made more explicit when applied to regular lattices: Each node w in a regular D-dimensional lattice can be identified with its coordinates $w=(w^{[1]},\dots,w^{[D]}) \in \mathbb{Z}^D$, see Figure 2.1(left) for an illustration. Using this representation, we have $d(v,w)=\|v-w\|_1 = |v^{[1]}-w^{[1]}|+\cdots+|v^{[D]}-w^{[D]}|$ and thus W_i from (2.1) is given as

$$W_i = \{w \in \{w_1,\dots,w_{i-1}\} : \|w-w_i\|_1 \leq d\}.$$

For an infinite lattice it is known [1, Theorem 2.7] that the cardinality $\ell_D(d)$ of the set $\{z \in \mathbb{Z}^D : \|z\|_1 \leq d\}$ is given as

$$\ell_D(d) = \sum_{k=0}^{D} \binom{D}{k}\binom{d+D-k}{D} \qquad \left(\text{where } \tbinom{d+D-k}{k} = 0 \text{ if } d<k\right). \tag{2.3}$$

So, in a greedy algorithm, W_i can be obtained by examining at most $\ell_D(d)-1$ nodes and check whether they have already been colored. Alternatively, a distance-d coloring for regular D-dimensional lattices can also be obtained directly, due to the following result which we prove in Appendix A.

Theorem 2.2. *Let $G=(V,E)$ be a D-dimensional $N_1 \times N_2 \cdots \times N_D$ lattice. Let any node $w \in V$ be defined by its coordinates $w=(w^{[1]},\dots,w^{[D]})$, with $0 \leq w^{[i]} \leq N_i - 1$, $i \in 1,\dots,D$. Then a distance-d coloring with $(d+1)^D$ colors is given by*

$$\mathrm{col}(w) = \left(\sum_{k=0}^{D-1} \widetilde{w^{[k]}}(d+1)^k\right)+1, \quad \textit{where} \quad \widetilde{w^{[k]}} = w^{[k]} \bmod (d+1). \tag{2.4}$$

Let us note that for $D = 2$, an optimal distance-d coloring is explicitly known with $\left\lceil \frac{1}{2}(d+1)^2 \right\rceil$ colors; see [20], while the coloring given in Theorem 2.2 needs approximately twice as many colors.

Two characteristics of the coloring of Theorem 2.2 for general D will further be exploited in the error analysis presented in Sections 3 and 4: Firstly, the construction is based on the fact that we color all nodes w in the cube $\{w : 0 \leq w^{[k]} \leq d \text{ for } k = 1, \dots, D\}$ with $(d+1)^D$ colors as illustrated in Figure 2.2 (red, solid square). This coloring is then repeated by shifting this initial cube through the entire lattice (red, dashed squares). Secondly, with this coloring every color class can be interpreted as representing a coarse grid, where the distances between the nodes in one color class are multiples of $d+1$. This is illustrated in Figure 2.2 where the green nodes represent one color class.

Remark 2.3. For D-dimensional lattices with an equal number of nodes in each dimension, a recursively computable *hierarchical distance-d coloring* was introduced in [52] for distances $d = 2^i$, $i = 0, 1, \dots,$ using $2^{Di+1} = 2d^D$ colors. This approach was recently extended to the case of lattices with an uneven number of nodes per dimension and even more general graphs in [35]. Note that for D small and d not too small, we have $(d+1)^D < 2d^D$. For example, $(d+1)^2 < 2d^2$ as soon as $d > 2$ and $(d+1)^3 < 2d^3$ as soon as $d > 3$. For the analysis in Section 3 and 4, the colorings discussed in the present paper are more appropriate and the analysis of the hierarchical probing approach is beyond the scope of this work. $\diamond$

We end the discussion of regular lattices with the following result which bounds the number of nodes that have exact distance d from a given node. The rather technical, combinatorial proof is presented in Appendix A.

Lemma 2.4. *Let* $\ell_D^{=}(d) := |\{z \in \mathbb{Z}^D : \|z\|_1 = d\}|$, *then* $\ell_D^{=}(d) \leq 2Dd^{D-1}$.

3. Sparse approximation of matrix functions.

In this section we analyze the error of the approximation (1.4) when one of the colorings from Section 2 is used. Before doing so, we first discuss some general results on the existence and quality of sparse approximations to reveal what is achievable at all.

3.1. General results on sparse approximations.

We place ourselves in a slightly broader context, as it was also done in [3], and formulate sparse approximation results in terms of a matrix $B \in \mathbb{C}^{n \times n}$ (instead of $f(A)$) with a decay property with respect to a general graph $G = (V, E)$ with $V = \{1, \dots, n\}$ (instead of $G(A)$)). The following essential result from [3] forms the basis for sparse approximations of matrices with exponential decay.

Theorem 3.1. *Let* $\{B_s\}_{s \in \mathcal{S}}$ *be a family of* $n_s \times n_s$ *matrices having exponential decay with respect to a family of corresponding graphs* $\{G_s\}_{s \in \mathcal{S}}$ *with geodesic distances* d_s,

$$|[B_s]_{ij}| \leq C q^{\mathrm{d}_s(i,j)}, \ \ i, j = 1 \dots, n,$$

with $C > 0, q \in (0,1)$ *independent of* s. *Assume that the graphs have bounded maximal degree* $\Delta(G_s) \leq c$ *for all* s. *Then for every* $\varepsilon > 0$, B_s *contains at most* $\mathcal{O}(n_s)$ *entries greater than* ε *in magnitude.*

Furthermore, the following result for matrices with exponential *off-diagonal* decay was also given in [3].

Theorem 3.2. *Let* $\{B_s\}_{s \in \mathcal{S}}$ *be a family of* $n_s \times n_s$ *matrices with*

$$|[B_s]_{ij}| \leq C q^{|i-j|}, \ \ i, j = 1 \dots, n,$$

with $C > 0, q \in (0,1)$ independent of s. Then for $\varepsilon > 0$ there exists $\widetilde{m}$ independent of s such that

$$\|B_s - B_s^{(m)}\|_1 < \varepsilon \;\; \text{for } m > \widetilde{m},$$

where $B_s^{(m)} \in \mathbb{C}^{n_s \times n_s}$ is the banded matrix with $[B_s^{(m)}]_{ij} = [B_s]_{ij}$ for $|i-j| \le m$ and $[B_s^{(m)}]_{ij} = 0$ for $|i-j| > m$. In particular, for any fixed $m > \widetilde{m}$ the matrices $B_s^{(m)}$ contain $\mathcal{O}(n_s)$ nonzeros.

To obtain a generalization for matrices with general exponential (not necessarily off-diagonal) decay, we define the level sets of a node $j \in V$ in a graph $G = (V, E)$ with $|V| = n$ as

$$\begin{aligned} L^{(\delta)}(j) &:= \{i \in V, \, \mathrm{d}(i,j) = \delta\}, \quad \delta = 0, \dots, n-1, \\ L^{(\infty)}(j) &:= \{i \in V, \, \mathrm{d}(i,j) = \infty\}. \end{aligned}$$

Note that for any node j we have

$$V \; = \cup_{\delta=0}^{n-1} L^{(\delta)}(j) \cup L^{(\infty)}(j),$$

since every node i has either distance smaller than n from j or cannot be reached from j, in which case $i \in L^{(\infty)}(j)$. With these notations we can give the following generalization of Theorem 3.2.

Theorem 3.3. *Let $\{B_s\}_{s \in \mathcal{S}}$ be a family of $n_s \times n_s$ matrices having exponential decay*

$$|[B_s]_{ij}| \le C q^{\mathrm{d}_s(i,j)}, \;\; i, j = 1, \dots, n$$

with respect to the distances d_s in a family of graphs $\{G_s\}_{s \in \mathcal{S}}$, where $C > 0, q \in (0,1)$ are independent of s. Furthermore, assume that for all nodes j all level sets $L_s^{(\delta)}(j)$ of all graphs G_s are polynomially bounded, i.e., we have

$$|L_s^{(\delta)}(j)| \le K \, \delta^{\alpha} \tag{3.1}$$

with $K > 0$ and $\alpha > 0$, both independent of s and j. For $m > 0$ define the matrix $B_s^{(m)}$ via

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

Then for $\varepsilon > 0$ there exists $\widetilde{m}$ independent of s such that $\|B_s - B_s^{(m)}\|_1 < \varepsilon$ for all $m > \widetilde{m}$. Moreover, for any fixed $m > \widetilde{m}$ the matrices $B_s^{(m)}$ contain $\mathcal{O}(n_s)$ nonzeros.

Proof. Let $m_1 := m_1(q, \alpha)$ be such that $\delta^{\alpha} q^{\frac{\delta}{2}} < 1$ holds for $\delta > m_1$. Then for $m > m_1$ we obtain

$$\begin{aligned} \|B_s - B_s^{(m)}\|_1 &= \max_{j=1}^{n_s} \sum_{i=1}^{n_s} \left| [B_s]_{ij} - [B_s^{(m)}]_{ij} \right| \; = \; \max_{j=1}^{n_s} \sum_{\substack{i \\ \mathrm{d}_s(i,j) > m}} |[B_s]_{ij}| \\ &\le \max_{j=1}^{n_s} \sum_{\substack{i \\ \mathrm{d}_s(i,j) > m}} C q^{\mathrm{d}_s(i,j)} \; = \; \max_{j=1}^{n_s} C \sum_{\delta=m+1}^{n_s - 1} |L_s^{(\delta)}(j)| q^{\delta} \; \le \; CK \sum_{\delta=m+1}^{n_s-1} \delta^{\alpha} q^{\delta} \\ &= CK \sum_{\delta=m+1}^{\infty} \delta^{\alpha} q^{\frac{\delta}{2}} q^{\frac{\delta}{2}} \; \le \; CK \sum_{\delta=m+1}^{\infty} q^{\frac{\delta}{2}} \; \le \; CK \, \frac{\sqrt{q}^{m+1}}{1 - \sqrt{q}} \, . \end{aligned}$$

Let $m_2 := m_2(q,\varepsilon)$ be such that

$$CK\,\frac{\sqrt{q}^{m+1}}{1-\sqrt{q}} < \varepsilon$$

holds for $m > m_2$. Then for $m > \widetilde{m} := \max\{m_1, m_2\}$ we have $\|B_s - B_s^{(m)}\|_1 < \varepsilon$, and the number of non-zero elements in $B_s^{(m)}$ is at most

$$\sum\nolimits_{j=1}^{n_s} \sum\nolimits_{\delta=0}^{m} |L_s^{(\delta)}(j)| = n_s \left(1 + \sum\nolimits_{\delta=1}^{m} K\delta^{\alpha}\right) = \mathcal{O}(n_s),$$

from which the assertion of the theorem follows. □

Note that off-diagonal decay is equivalent to decay with respect to a chain graph and thus Theorem 3.2 is covered by this theorem: For a chain the level sets $L^{\delta}(j)$ contain at most two elements, i.e., we have $\alpha = 0$. For general $\alpha > 0$ we now have a similar result for other important cases, e.g., when the graphs G_s are regular D-dimensional lattices.

Theorem 3.3 was formulated in [3] with the assumption (3.1) on polynomially bounded level sets replaced by the less restrictive assumption that the family of graphs $\{G_s\}$ has bounded maximal degree. This turns out to have been too optimistic, as the following example shows.

Example 3.4. Let $0 < q < 1$, let $t \in \mathbb{N}$ be such that $tq > 1$ holds, and let G_p be the full t-ary tree with height p, which has $n_p = 1 + t + \cdots + t^p = (t^{p+1}-1)/(t-1)$ nodes. Then the maximal degree of the graphs G_p is bounded, $\Delta(G_p) = t+1$. Let j be the root of this tree so that the level set $L_p^{(\delta)}(j)$ is formed exactly by all nodes at depth δ in the tree, implying

$$|L_p^{(\delta)}(j)| = t^{\delta}, \delta = 0, \ldots, p, \;\; L_p^{(\infty)}(j) = \emptyset.$$

Let B_p be the $n_p \times n_p$ matrix with $[B_p]_{ij} = q^{\mathrm{d}_p(i,j)}$, where d_p is the distance in G_p. Then B_p has exponential decay with respect to G_p, and for all m we have

$$\|B_p - B_p^{(m)}\|_1 \geq \sum_{\mathrm{d}_p(i,j)>m} |[B_p^{(m)}]_{ij}| = \sum_{\delta=m+1}^{p} |L_p^{(\delta)}(j)| q^{\delta} = \sum_{\delta=m+1}^{p} t^{\delta} q^{\delta} \geq (p-m)(tq)^{m+1},$$

where the last inequality holds because of $tq > 1$. Thus, the first m for which $\|B_p - B_p^{(m)}\|_1 < 1$ holds is $m = p = \Omega(\log n)$, in which case we have $B_p^{(m)} = B_p$. ⋄

In this example the exponential decay in B_p is not enough to compensate the exponential growth of the level sets. This motivated condition (3.1) in Theorem 3.3.

3.2. Analysis of probing for sparse approximation of $f(A)$.

We now turn back to the specific situation where $B = f(A)$ and $G = G(A)$. The existence results of the previous section do not reveal how a sparse approximation is obtained in practice without computing $f(A)$. We now investigate the probing approximation (1.4) for obtaining such an approximation.

The following result gives an entrywise bound for the probing approximation $f(A)^{[d]}$ from (1.4) provided the probing vectors are obtained from a distance-$2d$ coloring of $|G(A)|$.

Proposition 3.5. *Let $f(A)$ have exponential decay* (1.5)*, let the sets V_ℓ be the color classes of a distance-$2d$ coloring of $|G(A)|$ and v_ℓ the corresponding probing vectors*

(1.2). *Let $f(A)^{[d]}$ be the approximation defined by* (1.4). *Then with $\varepsilon = Cq^d$ the following entrywise error bound holds for all $i, j \in \{1, \ldots, n\}$*

$$|[f(A)]_{ij} - [f(A)^{[d]}]_{ij}| \leq \begin{cases} (|V_\ell| - 1)\varepsilon \text{ for } j \in V_\ell & \text{if } \bar{\mathrm{d}}(i,j) \leq d, \\ \varepsilon & \text{if } \bar{\mathrm{d}}(i,j) > d. \end{cases}$$

Proof. The assertion is trivial for $\bar{\mathrm{d}}(i,j) > d$, since then $\mathrm{d}(i,j) \geq \bar{\mathrm{d}}(i,j) > d$ and $[f(A)^{[d]}]_{ij} = 0$ by (1.4). For i, j with $\bar{\mathrm{d}}(i,j) \leq d$ we have

$$[f(A)^{[d]}]_{ij} = [f(A)v_\ell]_i = \sum_{k \in V_\ell} [f(A)]_{ik}, \text{ where } j \in V_\ell.$$

Thus,

$$[f(A)^{[d]}]_{ij} - [f(A)]_{ij} = \sum_{\substack{k \in V_\ell \\ k \neq j}} [f(A)]_{ik}. \tag{3.2}$$

If we had $\bar{\mathrm{d}}(i,k) \leq d$ for some $k \in V_\ell$ with $k \neq j$, then

$$\bar{\mathrm{d}}(j,k) \leq \bar{\mathrm{d}}(j,i) + \bar{\mathrm{d}}(i,k) = \bar{\mathrm{d}}(i,j) + \bar{\mathrm{d}}(i,k) \leq 2d, \tag{3.3}$$

which is a contradiction to $j, k \in V_\ell$. Thus $\mathrm{d}(i,k) \geq \bar{\mathrm{d}}(i,k) > d$, and therefore we have

$$|[f(A)]_{ij} - [f(A)^{[d]}]_{ij}| \leq \sum_{\substack{k \in V_\ell \\ k \neq j}} \varepsilon = (|V_\ell| - 1)\varepsilon,$$

which concludes the proof. □

Note that $\bar{\mathrm{d}}(i,j) = \bar{\mathrm{d}}(j,i)$ is crucial in (3.3), and that we do not necessarily have that $\mathrm{d}(j,k) \leq \mathrm{d}(i,j) + \mathrm{d}(i,k)$ for the distances in the *directed* graph. This is why for a structurally non-symmetric matrix the proposition has to rely on a coloring of the undirected graph rather than the directed one.

Proposition 3.5 immediately implies bounds for the 1-, 2- and Frobenius norms.

Corollary 3.6. *Let the assumptions of Proposition 3.5 hold and let $\gamma = \max_\ell |V_\ell|$. Then with $\varepsilon = Cq^d$ we have*

$$\|f(A) - f(A)^{[d]}\|_2 \leq \|f(A) - f(A)^{[d]}\|_F \leq n(\gamma - 1)\varepsilon \tag{3.4}$$

and

$$\|f(A) - f(A)^{[d]}\|_1 \leq n(\gamma - 1)\varepsilon. \tag{3.5}$$

For a family of matrices $\{A_s\}_{s \in \mathcal{S}}, A_s \in \mathbb{C}^{n_s \times n_s}$ with uniform exponential decay (1.5), the bounds in (3.4) and (3.5) are at least of order $\mathcal{O}(n_s \varepsilon)$. If, similarly to Theorem 3.3, we assume that the level sets are polynomially bounded, the bound for the 1-norm can be made independent of n_s.

Theorem 3.7. *Let $\{A_s\}_{s \in \mathcal{S}}$, be a family of $n_s \times n_s$ matrices such that $f(A_s)$ has uniform exponential decay* (1.5). *Assume that the sizes of the level sets $L_s^{(\delta)}(j)$ of the undirected graphs $|G(A_s)|$ satisfy*

$$|L_s^{(\delta)}(j)| \leq K\,\delta^\alpha \text{ for all nodes } j = 1, \ldots, n_s$$

and let $f(A_s)^{[d]}$ *be the approximation defined by* (1.4) *with probing vectors resulting from a distance* $2d$*-coloring of* $|G(A_s)|$*. Then with* $\varepsilon = Cq^d$ *there exists* $\widetilde{d}$ *independent of* s *such that for* $d \geq \widetilde{d}$ *we have*

$$\|f(A_s) - f(A_s)^{[d]}\|_1 \leq \varepsilon \textit{ for all } s \in \mathcal{S}\ .$$

Proof. For every $d > 0$ we have

$$\begin{aligned}
\|f(A_s) - f(A_s)^{[d]}\|_1 &= \max_{j=1}^{n_s} \sum_{i=1}^{n_s} |[f(A_s)]_{ij} - [f(A_s)^{[d]}]_{ij}| \\
&\leq \max_{j=1}^{n_s} \Big(\sum_{\substack{i \\ \bar{\mathrm{d}}_s(i,j) > d}} |[f(A)]_{ij}| + \sum_{\substack{i \\ \bar{\mathrm{d}}_s(i,j) \leq d}} \sum_{\substack{k \in V_{\ell(j)} \\ k \neq j}} |[f(A)]_{ik}| \Big) \qquad (j \in V_{\ell(j)}) \\
&\leq \max_{j=1}^{n} \Big(\sum_{\delta=d+1}^{n_s-1} |L_s^{(\delta)}(j)| Cq^\delta + \sum_{\substack{i \\ \bar{\mathrm{d}}_s(i,j) \leq d}} \sum_{\delta=d+1}^{n_s-1} |L_s^{(\delta)}(i)| Cq^\delta \Big) \\
&\leq \max_{j=1}^{n} \Big(\sum_{\delta=d+1}^{n_s-1} K\delta^\alpha Cq^\delta + \sum_{\substack{i \\ \bar{\mathrm{d}}_s(i,j) \leq d}} \sum_{\delta=d+1}^{n_s-1} K\delta^\alpha Cq^\delta \Big) \\
&= \max_{j=1}^{n} \Big(\sum_{\delta=d+1}^{n_s-1} K\delta^\alpha Cq^\delta + \sum_{\rho=0}^{d} |L_s^{(\rho)}(j)| \cdot \sum_{\delta=d+1}^{n_s-1} K\delta^\alpha Cq^\delta \Big) \qquad (3.6) \\
&\leq \sum_{\delta=d+1}^{\infty} K\delta^\alpha Cq^\delta + \big(1 + \sum_{\rho=1}^{d} K\rho^\alpha\big) \sum_{\delta=d+1}^{\infty} K\delta^\alpha Cq^\delta,
\end{aligned}$$

where we used (3.2) for the second line and that $\bar{\mathrm{d}}_s(i,k) > d$ for $k \in V_{\ell(j)}, k \neq j$, see (3.3), for the third. As shown in the proof of Theorem 3.3 there exists d_1 such that

$$\sum_{\delta=d+1}^{\infty} K\delta^\alpha Cq^\delta \leq \widehat{C} q^{\frac{d+1}{2}} \quad \text{for } d > d_1,$$

with $\widehat{C} = \frac{CK}{1-\sqrt{q}}$. Hence, we obtain

$$\|f(A_s) - f(A_s)^{[d]}\|_1 \leq \left(2 + \sum_{\rho=1}^{d} K\rho^\alpha\right) \widehat{C} q^{\frac{d+1}{2}} \quad \text{for } d > d_1.$$

Since $\sum\limits_{\rho=1}^{d} \rho^\alpha < d^{\alpha+1}$, we can find d_2 such that for $d > d_2$ we have

$$\Big(2 + \sum_{\rho=1}^{d} K\rho^\alpha\Big) \widehat{C} q^{\frac{d+1}{2}} < \varepsilon.$$

The assertion thus holds for $\widetilde{d} = \max\{d_1, d_2\}$. □

While the formulation of Theorem 3.7 is focused on the uniform approximation property, we can also directly use (3.6) to obtain error bounds for a single matrix A. We illustrate this for β-banded matrices, where—as opposed to the result formulated in Corollary 3.6—we now obtain a bound for the 1-norm that does not depend on n.

COROLLARY 3.8. *Let $A \in \mathbb{C}^{n\times n}$ be a β-banded matrix and let $f(A)$ have exponential decay* (1.5)*. Let $f(A)^{[d]}$ be the approximation defined by* (1.4) *with probing vectors resulting from the coloring* (2.2)*. Then*

$$\|f(A) - f(A)^{[d]}\|_1 \le 2\beta q \frac{2+2d\beta}{1-q}\varepsilon, \quad \text{where } \varepsilon = Cq^d.$$

Proof. For all nodes j and levels δ we have $|L^{(\delta)}(j)| \le 2\beta$. Using this and (3.6) the approximation error of $f(A)^{[d]}$ can be bounded as

$$\begin{aligned}\|f(A) - f(A)^{[d]}\|_1 &\le \sum_{\delta=d+1}^{n-1} 2\beta C q^\delta + \sum_{\rho=0}^{d} |L^\rho(j)| \sum_{\delta=d+1}^{n-1} 2\beta C q^\delta \\ &\le 2\beta C(1+1+2d\beta)\sum_{\delta=d+1}^{n-1} q^\delta \le 2\beta C(2+2d\beta)\sum_{\delta=d+1}^{\infty} q^\delta,\end{aligned}$$

which concludes the proof. □

Another situation in which it is possible to improve upon the result of Corollary 3.6, now for the Frobenius norm, is when the decay bounds that we have available have their origin in a polynomial approximation property (1.7).

THEOREM 3.9. *Let $A \in \mathbb{C}^{n\times n}$ and assume that the function f fulfills* (1.7)*. Let $f(A)^{[d]}$ be the approximation defined by* (1.4) *with probing vectors v_ℓ resulting from a distance $2d$-coloring of $|G(A)|$ with color classes V_ℓ. Then*

$$\|f(A) - f(A)^{[d]}\|_F \le 2K\sqrt{n}\varepsilon, \quad \text{with } \varepsilon = Cq^d,$$

where $K = 1$ when A is normal and $K = 1+\sqrt{2}$ otherwise.

Proof. Let p_d be a polynomial of degree d such that $|f(z) - p_d(z)| \le Cq^d$ for all $z \in \mathcal{W}(A)$, which exists since f satisfies (1.7). We now estimate the two terms in the triangle inequality

$$\|f(A) - f(A)^{[d]}\|_F \le \|f(A) - p_d(A)\|_F + \|f(A)^{[d]} - p_d(A)\|_F \tag{3.7}$$

individually. For the first term, note that for $i = 1,\ldots,n$ we have $\|f(A)e_i - p_d(A)e_i\|_2 \le \|f(A) - p_d(A)\|_2 \le K\varepsilon$ due to Theorem 1.2. This directly implies

$$\|f(A) - p_d(A)\|_F \le K\sqrt{n}\varepsilon. \tag{3.8}$$

Similarly, we also have $\|f(A)v_\ell - p_d(A)v_\ell\|_2 \le K\varepsilon\|v_\ell\|_2 = K\varepsilon\sqrt{|V_\ell|}$. For the degree d polynomial p_d the sparse approximation $p_d(A)^{[d]}$ is exact so that

$$\begin{aligned}\|f(A)^{[d]} - p_d(A)\|_F^2 \;=\; \|f(A)^{[d]} - p_d(A)^{[d]}\|_F^2 &= \sum\nolimits_{\ell=1}^{m} \|f(A)v_\ell - p_d(A)v_\ell\|_2^2 \\ &\le \sum\nolimits_{\ell=1}^{m} K^2\varepsilon^2|V_\ell| \;=\; K^2\varepsilon^2 n,\end{aligned}$$

which gives the estimate

$$\|f(A)^{[d]} - p_d(A)\|_F = \|f(A)^{[d]} - p_d(A)^{[d]}\|_F \le K\sqrt{n}\varepsilon. \tag{3.9}$$

Inserting (3.8) and (3.9) into (3.7) gives the desired result. □

4. Approximation of the trace of matrix functions. We now turn to investigating the accuracy of the probing method (1.3) for approximating the trace $\operatorname{tr}(f(A))$. As we will see, instead of using distance-$2d$ colorings for the undirected graph $|G(A)|$ we can now work with distance-d colorings in the directed graph $G(A)$. This reflects the fact that the trace, being the sum of bilinears $e_i^H f(A) e_i$ appears as a "quadratic" quantity. For the probing vectors defined in (1.2) we have

$$v_\ell^H f(A) v_\ell = \sum_{i \in V_\ell} [f(A)]_{ii} + \sum_{\substack{i,j \in V_\ell \\ i \neq j}} [f(A)]_{ij},$$

from which we immediately obtain

$$\operatorname{tr}(f(A)) = \sum_{i=1}^{n} [f(A)]_{ii} = \sum_{\ell=1}^{m} v_\ell^H f(A) v_\ell - \sum_{\ell=1}^{m} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} [f(A)]_{ij}.$$

Thus, the error of the approximation $\mathcal{T}(f(A))$ from in (1.3) is given by

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| = \left| \sum_{\ell=1}^{m} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} [f(A)]_{ij} \right|. \tag{4.1}$$

To obtain bounds for the error (4.1) when $f(A)$ has exponential decay (1.5), consider a distance-d coloring of $G(A)$ with color classes $V_\ell, \ell = 1, \ldots, m$. Then, with the corresponding probing vectors (1.2) and with $\varepsilon = Cq^d$, an immediate error bound is given by

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{m} \sum_{\substack{i,j \in V_\ell \\ i \neq j}} \varepsilon = \sum_{\ell=1}^{m} |V_\ell|(|V_\ell| - 1)\varepsilon. \tag{4.2}$$

If we assume that the size of the color classes is asymptotically given by $\mathcal{O}(\frac{n}{m})$, i.e., if the nodes are distributed uniformly among the color classes, and if the number of colors m is independent of n, then the error bound (4.2) is of order $\mathcal{O}(n^2)\varepsilon$. In the following we discuss cases in which better error bounds than (4.2) can be obtained. Similar to the case of the sparse approximation discussed in Section 3.2, we can give $\mathcal{O}(n)\varepsilon$ error bounds by exploiting knowledge about the specific coloring of $G(A)$. E.g., for banded matrices A, using the coloring (2.2), we obtain the following improved error bound. Note that the result also holds for matrices A for which a permutation P^TAP is banded if we permute the probing vectors accordingly.

THEOREM 4.1. *Assume that $A \in \mathbb{C}^{n \times n}$ is β-banded and that $f(A)$ has exponential decay* (1.5). *Let $\mathcal{T}(f(A))$ be the approximation* (1.3) *to the trace, where the vectors v_ℓ are computed with respect to the coloring* (2.2) *for a given distance d and put $\varepsilon = Cq^d$. Then*

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| \leq \varepsilon \frac{2n}{1 - q^d}.$$

Proof. The color classes of the coloring (2.2) are given as

$$V_\ell = \left\{ \ell + k(d\beta + 1), k = 0, \ldots, \left\lfloor \frac{n - \ell}{d\beta + 1} \right\rfloor \right\}, \quad \ell = 1, \ldots, d\beta + 1 =: m.$$

By inserting the decay bounds into (4.1), we obtain

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{m} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} |[f(A)]_{ij}| \leq \sum_{\ell=1}^{m} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} Cq^{\mathrm{d}(i,j)}. \tag{4.3}$$

Now, for any color ℓ, if $i,j \in V_\ell, i = \ell + rm, j = \ell + sm$, then $\mathrm{d}(i,j) = |r-s|d$. Thus, for all ℓ we have

$$\begin{aligned}\sum_{\substack{i,j\in V_\ell \\ i\neq j}} q^{\mathrm{d}(i,j)} &= \sum_{r=0}^{\lfloor \frac{n-\ell}{m} \rfloor} \sum_{\substack{s=0 \\ s\neq r}}^{\lfloor \frac{n-\ell}{m} \rfloor} q^{|s-r|d} \leq 2 \sum_{r=0}^{\lfloor \frac{n-\ell}{m} \rfloor} \sum_{k=1}^{\lfloor \frac{n-\ell}{m} \rfloor} q^{kd} \\ &\leq 2 \sum_{r=0}^{\lfloor \frac{n-\ell}{m} \rfloor} \frac{q^d}{1-q^d} = 2|V_\ell| \frac{q^d}{1-q^d}.\end{aligned}$$

Inserting this relation into (4.3) gives

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{m} |V_\ell| 2C \frac{q^d}{1-q^d} = 2nC \frac{q^d}{1-q^d} = \varepsilon \frac{2n}{1-q^d}.$$

□

A similar $\mathcal{O}(nq^d)$ bound can be formulated if $G(A)$ is a regular D-dimensional lattice and the coloring of Theorem 2.2 is used. We state this results using the polylogarithm $\mathrm{Li}_s(z) = \sum_{i=1}^{\infty} \frac{z^i}{i^s}$.

THEOREM 4.2. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix for which $G(A)$ is a regular D-dimensional lattice. Let $f(A)$ have exponential decay* (1.5). *Let $\mathcal{T}(f(A))$ be defined by* (1.3)*, where the vectors v_ℓ are computed with respect to the distance-d coloring of Theorem 2.2. Then*

$$|\,\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq 2CDn\,\mathrm{Li}_{1-D}(q^d).$$

Proof. Again,

$$|\mathrm{tr}(f(A)) - \mathcal{T}(f(A))| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} |[f(A)]_{ij}| \leq \sum_{\ell=1}^{k} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} Cq^{\mathrm{d}(i,j)},$$

with the color classes V_ℓ from Theorem 2.2. For this coloring, as illustrated in Figure 2.2, the distances between nodes from the same color class are multiples of d and these nodes actually form again a regular D-dimensional lattice. Lemma 2.4 shows that for each node the number of nodes with distance δ in this lattice, i.e., with distance δd in the original lattice, is bounded by $2D\,\delta^{D-1}$. Thus

$$\begin{aligned}\sum_{\ell=1}^{k} \sum_{\substack{i,j\in V_\ell \\ i\neq j}} Cq^{\mathrm{d}(i,j)} &\leq \sum_{\ell=1}^{k} |V_\ell| \sum_{\delta=1}^{\infty} 2D\,\delta^{D-1}\,Cq^{\delta d} \\ &\leq 2CDn \sum_{\delta=1}^{\infty} \delta^{D-1} q^{\delta d} \\ &= 2CDn\,\mathrm{Li}_{1-D}(q^d).\end{aligned}$$

□

REMARK 4.3. For a given value of D, the bound from Theorem 4.2 can be cast into a more explicit form by noting that all polylogarithms of negative integer order are rational functions of the form $\mathrm{Li}_{-s}(z) = \frac{p_s(z)}{(1-z)^{s+1}}$ where p_s is a polynomial of degree s such that $p_s(0) = 0$. An explicit representation can be found in terms of Eulerian numbers; see, e.g., [37]. In particular, the first few polylogarithms of negative integer order are given by

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

Using these relations we, e.g., find the following bound for $D = 4$

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| \le 8Cn\frac{q^d + 4q^{2d} + q^{3d}}{(1-q^d)^4},$$

which for large d behaves like $8Cnq^d$.

Let us further note that for the case $D = 1$, i.e., a tridiagonal matrix A, both Theorem 4.1 and Theorem 4.2 are applicable. Since $\mathrm{Li}_0(z) = \frac{z}{1-z}$, both theorems actually agree in this case. ◇

As in the situation where we looked at the approximation quality for the matrix function as a whole, we can again derive improved error bounds when we have a polynomial approximation property (1.7) available.

THEOREM 4.4. *Let $A \in \mathbb{C}^{n\times n}$ and assume that f fulfills condition* (1.7). *Let the approximation $\mathcal{T}(f(A))$ in* (1.3) *be obtained using a distance-d coloring of $G(A)$. Then, with $\varepsilon = Cq^d$ we have*

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| \le 2Kn\varepsilon, \tag{4.4}$$

where $K = 1$ when A is normal and $K = 1 + \sqrt{2}$ otherwise.

Proof. We proceed as in the proof of Theorem 3.9. Let p_d be a polynomial of degree d such that $|f(z) - p_d(z)| \le Cq^d \le \varepsilon$ for all $z \in \mathcal{W}(A)$, which exists since f fulfills (1.7). Then

$$\|f(A) - p_d(A)\|_2 \le K\varepsilon. \tag{4.5}$$

We write

$$|\operatorname{tr}(f(A)) - \mathcal{T}(f(A))| \le |\operatorname{tr}(f(A)) - \operatorname{tr}(p_d(A))| + |\mathcal{T}(f(A)) - \operatorname{tr}(p_d(A))|. \tag{4.6}$$

For the first term, we get, using the linearity of the trace, (4.5) and the Cauchy-Schwarz inequality

$$|\operatorname{tr}(f(A)) - \operatorname{tr}(p_d(A))| \le \sum_{\ell=1}^{n} \left|e_\ell^T \left(f(A) - p_d(A)\right) e_\ell\right| \le \sum_{\ell=1}^{n} K\varepsilon = Kn\varepsilon. \tag{4.7}$$

For the second term, note that the probing approximation $\mathcal{T}(p_d(A))$ for the trace is exact, $\operatorname{tr}(p_d(A)) = \mathcal{T}(p_d(A))$. Therefore, in a similar manner as for the first term, we obtain

$$\begin{aligned} |\mathcal{T}(f(A)) - \operatorname{tr}(p_d(A))| = |\mathcal{T}(f(A)) - \mathcal{T}(p_d(A))| &\le \sum_{\ell=1}^{m} \left|v_\ell^T \left(f(A) - p_d(A)\right) v_\ell\right| \\ &\le \sum_{\ell=1}^{m} K|V_\ell|\varepsilon = Kn\varepsilon. \end{aligned} \tag{4.8}$$

Inserting (4.7) and (4.8) into (4.6) concludes the proof. □
The numerical examples in Section 6 illustrate that the error of the probing-based approximations scales indeed linearly with the dimension n of the matrix. In this sense, $\mathcal{O}(n)\varepsilon$ error bounds are the best we can achieve.

5. Using Krylov subspace methods in the probing approach. Probing methods require the computation of matrix-vector products $f(A)v_\ell$ or bilinear forms $v_\ell^T f(A)v_\ell$. Both are standard tasks in numerical linear algebra, for which a plethora of different methods has been developed. Widely used methods for both tasks are Krylov subspace methods. As with any iterative method, an important question arising in this context is how to find a good stopping criterion in order to keep the computational cost as small as possible while at the same time guaranteeing that the desired overall accuracy is reached in the approximation of $f(A)$ or $\mathrm{tr}(f(A))$.
We now answer this question for the situation that the decay bounds we have available stem from a polynomial approximation property of the form (1.7). We begin by very shortly reviewing a few important facts about Arnoldi's method, the prototype Krylov subspace method; see, e.g., [23, Section 3.5] or [33, Section 13.2] for details. The approximation for $f(A)b$ from s steps of Arnoldi's method is given by

$$f_s = \|b\|_2 W_s f(H_s) e_1, \tag{5.1}$$

where the columns of W_s are the orthonormal Arnoldi basis vectors and $H_s = W_s^H A W_s \in \mathbb{C}^{s\times s}$ is the upper Hessenberg matrix containing the orthogonalization coefficients. We have that

$$f_s = \|b\|_2 W_s \widetilde{p}_{s-1}(H_s) e_1 = \widetilde{p}_{s-1}(A) b,$$

where $\widetilde{p}_{s-1}$ is the polynomial of degree $s-1$ that interpolates f on the eigenvalues of H_s in the Hermite sense. Theorem 1.2 together with (1.7) shows that there exists a polynomial p_{s-1} of degree $s-1$ such that

$$\|f(A)b - p_{s-1}(A)b\|_2 \leq \|f(A) - p_{s-1}(A)\|_2 \cdot \|b\|_2 \leq \|b\|_2 K C q^{s-1},$$

where $K = 1$ if A is normal and $K = 1 + \sqrt{2}$ otherwise. Due to the near-optimality property of the Arnoldi polynomial [31, Section 4.2.2], the error of the Arnoldi approximation is within a factor 2 of the best possible polynomial approximation, which implies

$$\|f(A)b - f_s\|_2 \leq 2\|b\|_2 K C q^{s-1}. \tag{5.2}$$

5.1. Sparse approximation. Let $f_s^{(\ell)}$ denote the Arnoldi approximation (5.1) for $f(A)v_\ell$. By replacing $f(A)v_\ell$ by $f_s^{(\ell)}$ in (1.4), we obtain the approximation

$$[\widetilde{f(A)}^{[d]}]_{ij} := \begin{cases} [f_s^{(\ell)}]_i \text{ for } j \in V_\ell & \text{if } \bar{\mathrm{d}}(i,j) \leq d, \\ 0 & \text{if } \bar{\mathrm{d}}(i,j) > d, \end{cases} \tag{5.3}$$

where the color classes V_ℓ come from a distance-$2d$ coloring of $|G(A)|$. In the triangle inequality

$$\|f(A) - \widetilde{f(A)}^{[d]}\| \leq \|f(A) - f(A)^{[d]}\| + \|f(A)^{[d]} - \widetilde{f(A)}^{[d]}\|, \tag{5.4}$$

Theorem 3.9 shows that for the Frobenius norm the first term in (5.4) can be bounded as

$$\|f(A) - \widetilde{f(A)}^{[d]}\|_F \le 2KCq^d\sqrt{n}. \tag{5.5}$$

For the second term, note that

$$\|f(A)^{[d]} - \widetilde{f(A)}^{[d]}\|_F^2 = \sum\nolimits_{\ell=1}^{m} \|f(A)v_\ell - f_s^{(\ell)}\|_2^2,$$

so that inserting (5.2), we obtain

$$\begin{aligned}\|f(A)^{[d]} - \widetilde{f(A)}^{[d]}\|_F &\le \Big(\sum\nolimits_{\ell=1}^{m} 4\|v_\ell\|_2^2 K^2C^2q^{2(s-1)}\Big)^{1/2}\\ &= \Big(4K^2C^2q^{2(s-1)}\sum\nolimits_{\ell=1}^{m}|V_\ell|\Big)^{1/2}\\ &= 2KCq^{(s-1)}\sqrt{n}. \end{aligned}\tag{5.6}$$

Inequalities (5.5) and (5.6), on the one hand, give us the final estimate

$$\|f(A) - \widetilde{f(A)}^{[d]}\|_F \le 2KC\sqrt{n}(q^d + q^{s-1}). \tag{5.7}$$

On the other hand, they also that show after $d+1$ Arnoldi steps we can expect the Krylov approximation error to have the same magnitude as the probing error. If we perform more than $s = d+1$ Arnoldi steps, the overall error is likely to be dominated by the probing error, so that further Arnoldi iterations will have no or little effect on the overall error. Choosing $s = d+1$ the overall bound simplifies to

$$\|f(A) - \widetilde{f(A)}^{[d]}\|_F \le 4K\sqrt{n}\varepsilon.$$

As we will illustrate in the numerical experiments in Section 6, performing more than $d+1$ Arnoldi steps does typically indeed not lead to any further reduction of the overall error. Heuristically this can be further motivated as follows: The entries of the vector $f(A)v_\ell$ that we approximate by the Arnoldi iterates does not contain the exact entries of $f(A)$, but perturbed entries due to the "mixing" of contributions from nodes of the same color. Until the $d+1$st iteration of the Arnoldi method, this mixing does not occur in the basis vectors; see Figure 5.1 for an illustration. For $s > d+1$, then, the additional accuracy with which we approximate $f(A)v_\ell$ is spoiled by the loss of accuracy in the approximation of $f(A)$ due to increased mixing.

5.2. Approximating the trace.

The s-step Arnoldi approximation for a bilinear form $v_\ell^H f(A)v_\ell$ is given by

$$v_\ell^H f(A)v_\ell \approx \alpha_s^{(\ell)} := \|v_\ell\|_2^2 e_1^H f(H_s)e_1. \tag{5.8}$$

Using the relation between Krylov subspace methods, Gaussian quadrature and moment matching, it has been shown in [27, 28, 53], e.g., that (5.8) is exact if f is a polynomial up to degree $2s-1$ when A is Hermitian and up to degree s when A is non-Hermitian. This leads to the following result.

Theorem 5.1. *The error of the approximation* (5.8) *satisfies*

$$|v_\ell^H f(A)v_\ell - \alpha_s^{(\ell)}| \;\le\; 2\|v_\ell\|_2^2 KC \cdot \begin{cases} q^{2s-1} & \textit{when } A \textit{ is Hermitian},\\ q^s & \textit{otherwise.}\end{cases}$$

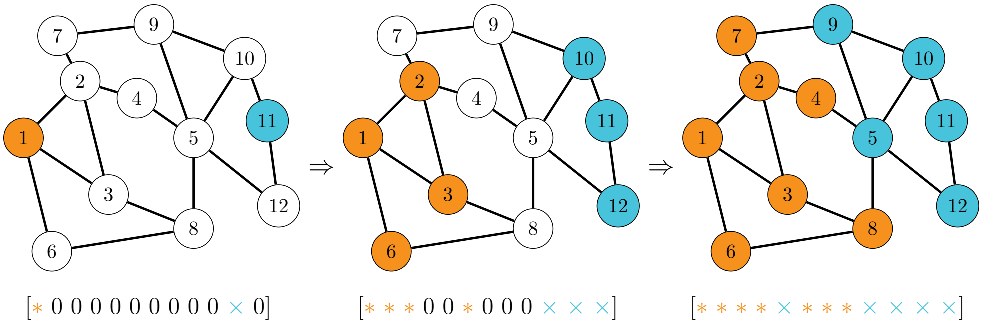

Fig. 5.1: Spreading of the nonzero entries in the first three Arnoldi basis vectors, starting with $v_\ell = e_1 + e_{11}$. Entries to which only the iteration corresponding to node 1 contributed are marked by an orange asterisk (∗) while entries to which only the iteration corresponding to node 11 contributed are marked by a blue cross (×). Because the nodes have a distance of 5, no mixing occurs in the first 3 basis vectors.

Proof. We consider just the Hermitian case; the non-Hermitian case follows analogously. Let $p^*_{2s-1}(z) \in \Pi_{2s-1}$ be such that

$$\max_{z\in\mathcal{W}(A)} |f(z) - p^*_{2s-1}(z)| \le Cq^{2s-1}. \tag{5.9}$$

As the approximation (5.8) is exact for $v_\ell^H p^*_{2s-1}(A)v_\ell$, we have

$$|v_\ell^H f(A)v_\ell - \alpha_s^{(\ell)}| \;=\; |v_\ell^H(f(A) - p^*_{2s-1}(A))v_\ell - \|v_\ell\|_2^2 e_1^H(f(H_s) - p^*_{2s-1}(H_s))e_1|.$$

From this, using the triangular inequality and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
&|v_\ell^H f(A)v_\ell - \|v_\ell\|_2^2 e_1^H f(H_s)e_1| \\
&\qquad \le \|v_\ell\|_2 \, \|f(A)v_\ell - p^*_{2s-1}(A)v_\ell\|_2 + \|v_\ell\|_2^2 \, \|f(H_s)e_1 - p^*_{2s-1}(H_s)e_1\|_2 \\
&\qquad \le \|v_\ell\|_2^2 \, \|f(A) - p^*_{2s-1}(A)\|_2 + \|v_\ell\|_2^2 \, \|f(H_s) - p^*_{2s-1}(H_s)\|_2
\end{aligned}$$

Now, $\|f(A) - p^*_{2s-1}(A)\|_2 \le Cq^{2s-1}$ due to (5.9) and (1.6), and the same bound applies to $\|f(H_s) - p^*_{2s-1}(H_s)\|_2$, since $\mathcal{W}(H_s) \subseteq \mathcal{W}(A)$ due to $H_s = W_s^H A W_s$ with W_s having orthonormal columns. □

Thus, choosing $s = \left\lceil \frac{d+1}{2} \right\rceil$ or $s = d$, we obtain the bound

$$|\mathcal{T}(f(A)) - \sum_{\ell=1}^{m} \alpha_s^\ell| \le KCq^d \sum_{\ell=1}^{m} \|v_\ell\|_2^2 = KCq^d \sum_{\ell=1}^{m} |V_l| = 2KCnq^d.$$

We are therefore in the order of magnitude of the bound for the probing error given in Theorem 4.4 for probing vectors coming from a distance-d coloring of $G(A)$ after d (or $\approx d/2$ if A is Hermitian) Arnoldi steps.

6. Numerical experiments.

In this section, we perform various numerical experiments both on model problems and on matrices coming from applications to investigate the quality of our error bounds, with particular emphasis on their scaling behavior with respect to growing matrix dimension n and increasing probing distance

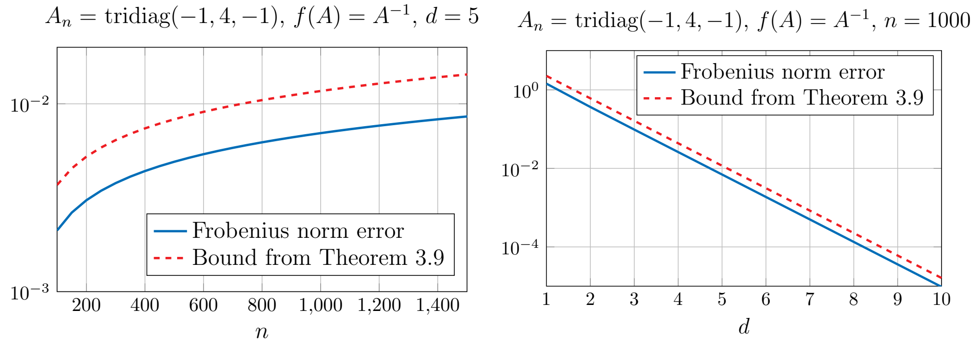

Fig. 6.1: Actual Frobenius norm error and error bound for the sparse approximation (1.4) corresponding to the coloring (2.2) for the matrix $A_n = \text{tridiag}(-1, 4, -1) \in \mathbb{C}^{n\times n}$ and $f(z) = 1/z$ for varying n (left) and d (right).

d. Unless explicitly stated otherwise, we compute the exact quantities $f(A)v_\ell$ and $v_\ell^T f(A)v_\ell$ used to obtain the exact error of our approximations to machine precision, using a factorization of A.

6.1. Tridiagonal model problem. As a first, simple test example, following [6], we consider the family of tridiagonal matrices $A_n = \text{tridiag}(-1, 4 -1) \in \mathbb{C}^{n\times n}$. The spectra of these matrices satisfy $\text{spec}(A_n) \subset [2, 6]$ independently of n. We consider the two functions $f(z) = 1/z$ and $f(z) = z^{-1/2}$ in the following and we always use the banded matrix coloring (2.2) with $\beta = 1$. In a first experiment, we compute sparse approximations of A_n^{-1} for varying dimension n while $d = 5$ is fixed and for varying d while $n = 1000$ is fixed. From [13, Theorem 2.4], the entries of A_n^{-1} exhibit an exponential decay with $C = \frac{1}{2}$ and $q = \frac{\sqrt{3}-1}{\sqrt{3}+1}$. The actual error norms together with our error bounds from Theorem 3.9 are depicted in Figure 6.1. In both cases, the bounds are quite tight and closely follow the actual error curve. We repeat the experiment for the inverse square root $A_n^{-1/2}$. The entries of this matrix function again decay exponentially, with $C = \sqrt{2}$ and $q = \frac{\sqrt{3}-1}{\sqrt{3}+1}$, see [21, Theorem 4]. This time, we compare the actual error to the 1-norm error bound of Corollary 3.8, because the decay bound from [21, Theorem 4] is not based on a polynomial approximation property of the form (1.7). The results of this experiment are shown in Figure 6.2. Again we see a good agreement between the actual error and the error bound, although it is not quite as sharp as before, overestimating the error by between one and two orders of magnitude. Still, the qualitative behavior is captured quite accurately. In particular, the 1-norm error is independent of n, as predicted by our theoretical results.

We also use this example to illustrate the influence of the number of Arnoldi steps used for approximating $f(A)v_\ell$ in the approximation (5.3), see Figure 6.3. We fix $n = 1000$ and $d = 5$ and compute the approximation error resulting when s Arnoldi steps per vector are performed, for $s = 1, \dots, 2d$ and compare it to the bound (5.7). We observe that the bound is in very good agreement with the actual error, and further, that the approximation error stagnates after $s = d + 1$, confirming our intuition explained in Section 5.1 that from this point on, the increased accuracy of the Krylov approximation is counteracted by the increased mixing between contributions of nodes from the same color class, so that no further decrease of the overall approximation

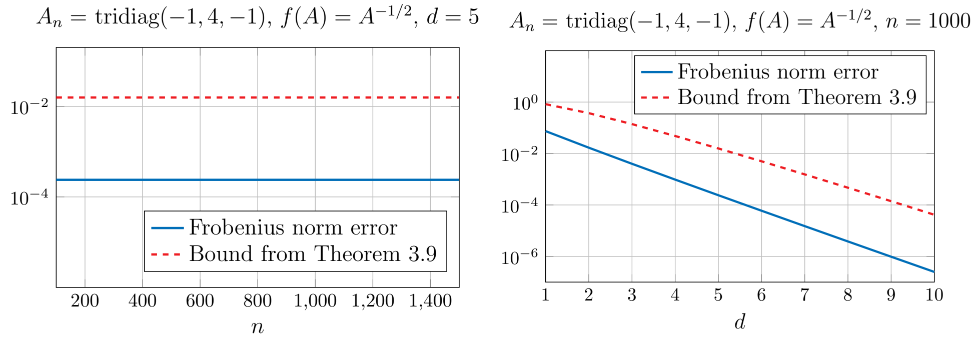

Fig. 6.2: Actual 1-norm error and error bound for the sparse approximation (1.4) corresponding to the coloring (2.2) for the matrix $A_n = \text{tridiag}(-1, 4, -1) \in \mathbb{C}^{n \times n}$ and $f(z) = z^{-1/2}$ for varying n (left) and d (right).

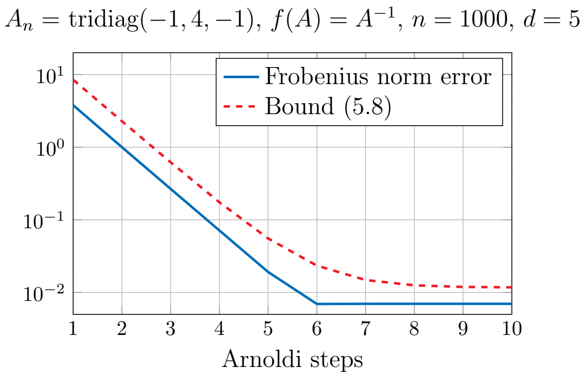

Fig. 6.3: Actual Frobenius norm error and error bound (in dependence of the number of Arnoldi steps) for the sparse approximation (5.3) corresponding to the coloring (2.2) for the matrix $A_n = \text{tridiag}(-1, 4, -1) \in \mathbb{C}^{n \times n}$ and $f(z) = 1/z$ with $n = 1000$, and $d = 5$.

error can be expected.
We conclude our first experiment by computing an estimate of the trace for both matrix functions, using exactly the same experimental parameters as before and compare the actual error to the bound (4.4) from Theorem 4.4. Note that we could alternatively use the bound from Theorem 4.1 which is tailored to banded matrices. Both bounds almost agree here, the latter one being slightly less sharp, by a factor $\frac{1}{1-q^d}$. Figure 6.4 shows that, as expected, the results are very similar to what can be observed in the context of computing a sparse approximation and we again observe a very good qualitative and quantitative agreement between the bounds and the actual error.

6.2. Shifted two-dimensional Laplace operator. As a second model problem, we consider the family of matrices $A_N \in \mathbb{C}^{N^2 \times N^2}$ arising from discretization of the Laplace equation with homogeneous Dirichlet boundary conditions on a regular

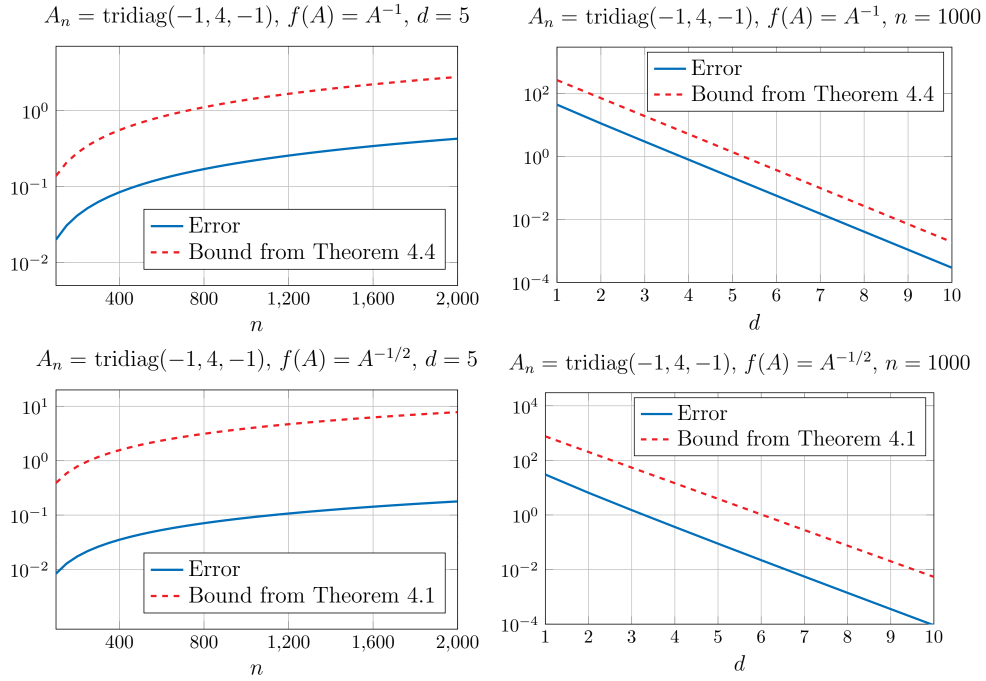

Fig. 6.4: Actual absolute error and error bound for the trace estimate (1.3) corresponding to the coloring (2.2) for the matrix A_n = tridiag$(-1, 4, -1) \in \mathbb{C}^{n\times n}$ and $f(z) = 1/z$ (top row) and $f(z) = z^{-1/2}$ (bottom row). Results for varying n are shown in the left column while results for varying d are shown in the right column.

square grid. We apply a shift of 4 to the diagonal of the matrix in order to obtain an N-independent decay in $f(A_N)$, giving

$$A_N = I_N \otimes M_N + M_N \otimes I_N \in C^{N^2\times N^2},$$

where M_N = tridiag$(-1, 4, -1) \in \mathbb{C}^{N\times N}$ is the tridiagonal matrix from the previous experiment. Applying this shift to the Laplacian matrix is common practice for obtaining good model problems for exponentially decaying matrices; see, e.g, [6, 52], were the same (or similar) families of matrices were considered.

We have spec$(A_N) \subset [4, 12]$ independent of N so that [13, Theorem 2.4] guarantees an exponential decay of the entries of A_N^{-1} with $C = \frac{1}{4}$ and $q = \frac{\sqrt{3}-1}{\sqrt{3}+1}$. We determine the color classes according to the optimal coloring for two-dimensional lattices from [20]. We again begin by approximating A_N^{-1} for increasing values of N while keeping $d = 5$ fixed and compare the actual error norm to the bound from Theorem 3.9. The results of this experiment are presented on the left-hand side of Figure 6.5 and we observe that the approximation error scales linearly with $N = \sqrt{n}$, as predicted by our theory. The magnitude of the error is overestimated by about one order of magnitude. On the right-hand side of Figure 6.5 the results for an experiment with varying d and fixed $N = 32$ are given. Again, we observe good qualitative and quantitative agreement between the error bound and actual error norm.

Next, in Figure 6.6, we also approximate tr(A_N^{-1}), using the same experimental setup

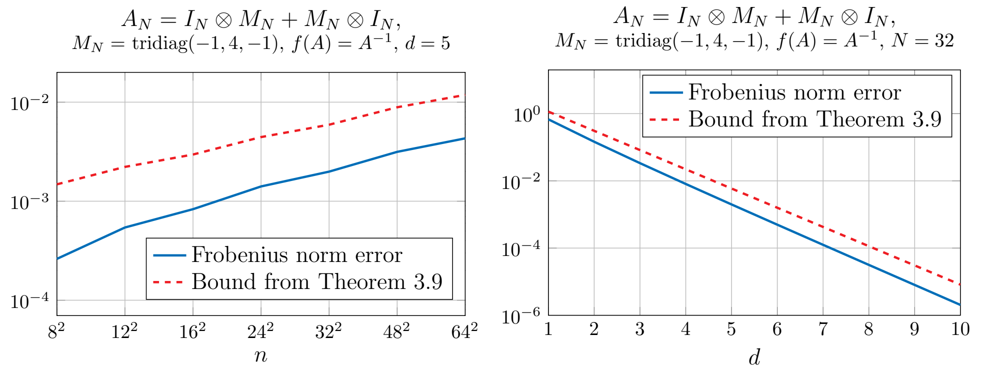

Fig. 6.5: Actual Frobenius norm error and error bound for the sparse approximation (1.4) corresponding to the coloring (2.4) for the matrix $A_N = I_N \otimes M_N + M_N \otimes I_N \in C^{N^2\times N^2}$, where $M_N = \mathrm{tridiag}(-1,4,-1) \in \mathbb{C}^{N\times N}$ and $f(z) = 1/z$ for varying N (left) and d (right).

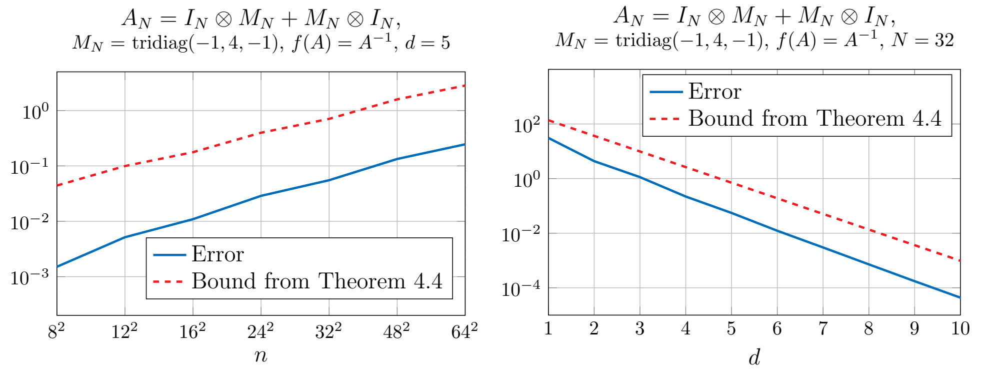

Fig. 6.6: Actual absolute error and error bound for the trace estimate (1.3) corresponding to the coloring (2.4) for the matrix $A_N = I_N \otimes M_N + M_N \otimes I_N \in C^{N^2\times N^2}$, where $M_N = \mathrm{tridiag}(-1,4,-1) \in \mathbb{C}^{N\times N}$ and $f(z) = 1/z$ for varying N (left) and d (right).

as for the sparse approximation and compare to the bound from Theorem 4.4. Again, we could also have used the lattice-specific bound from Theorem 4.2, which differs from that of Theorem 4.4 by a factor $\frac{2}{(1-q^d)^2} \approx 2$ in this case. The results of this experiment are shown in Figure 6.6. The scaling behavior for growing N and d is again captured very accurately, although we overestimate the actual error norm by quite a large margin.

6.3. Thresholded covariance matrix. For a next experiment, we consider the problem of computing a sparse approximation of an inverse covariance matrix, a task frequently occurring in uncertainty quanitification; see [2]. We use the example matrix from [54]: Let $A_{N^2} = \mathrm{cov}(N,\alpha,\beta) \in \mathbb{C}^{N^2\times N^2}$ be a covariance matrix corresponding to integer points (x_i, y_i) arranged as a regular $N \times N$ grid with respect to a decaying,

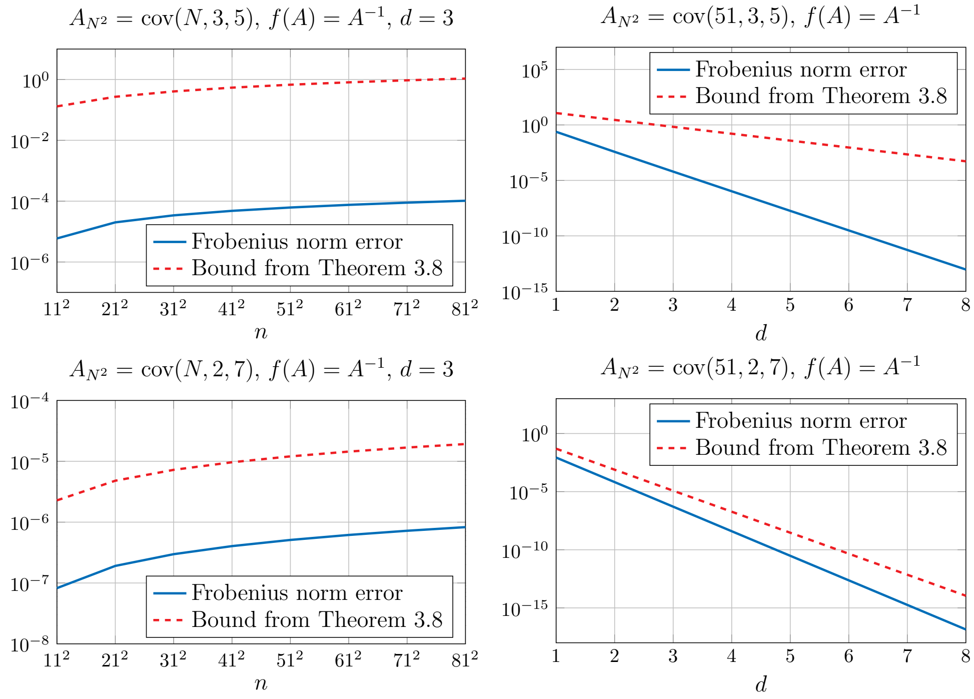

Fig. 6.7: Actual Frobenius norm error and error bound for the sparse approximation (1.4) corresponding to a greedy coloring for the matrix $A_{N^2} = \mathrm{cov}(N, \alpha, \beta) \in \mathbb{C}^{N^2 \times N^2}$ and $f(z) = 1/z$ for the parameter sets $\alpha = 3, \beta = 5$ (top) and $\alpha = 2, \beta = 7$ (bottom). Results for varying n are shown in the left column while results for varying d are shown in the right column.

thresholded covariance function. More precisely,

$$[A]_{ij} = \begin{cases} \left(1 - \frac{\|(x_i,y_i)-(x_j,y_j)\|_2}{\alpha}\right)^{\beta} & \text{if } \|(x_i, y_i) - (x_j, y_j)\|_2 \le \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

We use the two parameter sets $\alpha = 3, \beta = 5$ and $\alpha = 2, \beta = 7$ and compute a sparse approximation for A^{-1}. Again, we perform one experiment in which we vary $n = N^2$ while $d = 3$ is fixed and one experiment in which we vary d while $N = 51^2$ is fixed. The resulting Frobenius norms of the error together with our bounds are given in Figure 6.7. For the first parameter set, $\alpha = 3, \beta = 5$, we observe that while the qualitative behavior for growing n is accurately reproduced by our bound, we overestimate the actual error by several orders of magnitude. Thus, the bounds do give a valuable insight into the scaling behavior of the method but are not useful for judging whether the computed approximation is accurate enough for the application at hand. For growing d, we also observe that the slope of the error curve is much steeper than predicted by our bound, showing that also the qualitative behavior of the actual error is not accurately captured here. For the second parameter set, $\alpha = 2, \beta = 7$, our bounds look much better. For varying n, we still get an accurate impression of the qualitative scaling behavior while overestimating the error norm only by about one

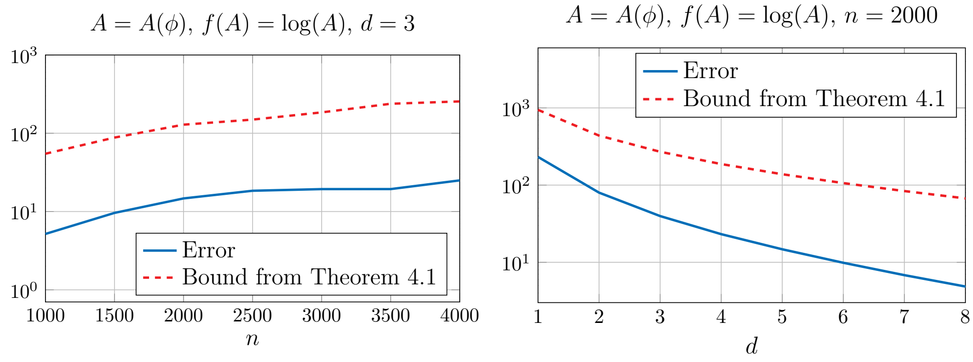

Fig. 6.8: Actual absolute error and error bound for the trace estimate (1.3) corresponding to the coloring (2.2) for the precision matrix $A(\phi)$ of a GMRF and $f(z) = \log(z)$ for varying n (left) and d (right)

order of magnitude. For varying d, we still do not get a completely accurate reflection of the slope of the error curve, but the slopes agree much better than before.

6.4. Maximum likelihood estimation for Gaussian Markov Random Fields. In a last experiment, we consider the problem of maximum likelihood estimation for Gaussian Markov Random Fields (GMRFs). A GMRF is a multivariate joint Gaussian distribution defined with respect to some underlying graph, where each random variable corresponds to a node of the graph. The GMRF can be described by the positive definite and sparse precision matrix $A \in \mathbb{R}^{n\times n}$ (which is the inverse of the covariance matrix Σ of the Gaussian distribution). Often, the precision matrix is parameterized by some unknown parameter ϕ, i.e., $A = A(\phi)$ which can be estimated by a maximum likelihood estimator. Let $x \in \mathbb{R}^n$ be a sample from the Gaussian distribution. The log-likelihood of this sample is then given by the functional

$$\log p(x \mid \phi) = \log\det(A(\phi)) - x^T A(\phi)x + G, \tag{6.1}$$

where G is a constant independent of ϕ; see, e.g., [32]. The computationally demanding part in the evaluation of (6.1) is the evaluation of the log-determinant. Due to the relation

$$\log\det(A(\phi)) = \operatorname{tr}\log(A(\phi)),$$

the log-determinant can be estimated by the probing approximation (1.3) applied to the matrix logarithm.
We consider the GMRF model from [41]. Given a set of n points $s_i \in [0,1]$, we define a Gaussian random variable $x_i, i = 1,\dots,n$ at each point. The entries of the precision matrix are

$$[A(\phi)]_{ij} = \begin{cases} 1 + \phi \sum_{k=1,k\neq i}^{n} \chi_{ij}^{\delta} & \text{if } i = j, \\ -\phi\chi_{ij}^{\delta} & \text{otherwise,} \end{cases} \tag{6.2}$$

where χ^δ is given by

$$\chi_{ij}^{\delta} = \begin{cases} 1 & \text{if } \|s_i - s_j\|_2 < \delta, \\ 0 & \text{otherwise,} \end{cases}$$

where δ is a distance threshold which determines which points s_i are connected in the graph underlying the GMRF. The resulting matrix is unstructured and sparse, but can be reordered to a matrix with rather narrow bandwidth by the Cuthill-McKee reordering, so that the coloring (2.2) can be used.
In our experiment, reported in Figure 6.8, we fix $\phi = 20$ and use $\delta = 0.02$ when $n = 1000$. For other values of n, we scale δ accordingly so that the average number of non-zeros per row and the bandwidth stay approximately constant for all values of n. In contrast to the previous experiments, we now mimick a situation that one typically faces in practice, namely that no explicit expressions for C and q in (1.5) are known, e.g., because the extremal eigenvalues of $A(\phi)$ are not known. In this case, one can obtain heuristic decay estimates by computing one (or a few) columns of $\log(A(\phi))$ (e.g., by a Krylov subspace method) and then estimating C and q from the observed decay pattern.
First, we vary n between 1000 and 5000 while keeping $d = 3$ fixed. Then, we fix $n = 2000$ and vary d between 1 and 10. We compare the actual error of the probing approximation for the trace of the logarithm to the bound from Theorem 4.1 for banded matrices, using the estimated values of C and q computed from a single column of $\log(A(\phi))$. In both cases, we can observe a good qualitative agreement between our bound and the actual error.

7. Conclusions.

We have presented a detailed a priori error analysis of probing methods for the computation of sparse approximations and trace estimates of matrix functions, with a special emphasis on graph coloring based probing and matrix functions that exhibit an exponential decay. As illustrated in several numerical experiments, our error bounds accurately predict the scaling behavior of the error with respect to the matrix dimension n or the coloring distance d. A particularly interesting observation in this context is that the error of the trace estimates decreases with exponent d, while the error of sparse approximations decreases only with exponent $\frac{d}{2}$. In addition to these error bounds for practical algorithms, we have also proven a new result on the existence of sparse approximations of matrix functions, improving on known results from the literature. While our results typically give a good idea of the qualitative behavior of the actual error, they sometimes severely overestimate the actual error. Possible directions for future research include developing further ideas to improve the quality of the error bounds and looking at new approaches for efficient distance-d coloring algorithms for appropriate classes of graphs.

Appendix A. Proof of Theorem 2.2 and Lemma 2.4.

A.1. Proof of Theorem 2.2.

Since for every node $w = (w^{[1]}, \ldots, w^{[D]})$ we have $\widetilde{w^{[k]}} \in \{0, \ldots, d\}$ for $k = 1, \ldots, D$, we know that the coloring

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

produces at most $(d+1)^D$ colors. Now assume $\operatorname{col}(w) = \operatorname{col}(v)$ for nodes $w \neq v$. We want to show that $\mathrm{d}(v, w) = \|v - w\|_1 > d$. Because of

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

we have $w^{[k]} = (d+1)a + \widetilde{w^{[k]}}$ and $v^{[k]} = (d+1)b + \widetilde{v^{[k]}}$ for some integers $a, b \geq 0$, and since $\operatorname{col}(w) = \operatorname{col}(v)$ we have $\widetilde{w^{[k]}} = \widetilde{v^{[k]}}$ for all $k = 1, \ldots, D$. Since $w \neq v$ there exists at least one k such that $w^{[k]} = (d+1)a + \widetilde{w^{[k]}} \neq (d+1)b + \widetilde{v^{[k]}} = v^{[k]}$ which is equivalent to $a \neq b$ for $d \geq 0$. By fixing such a k we obtain

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

which proves the assertion. □

A.2. Proof of Lemma 2.4. From (2.3) we obtain

$$\ell_D^{=}(d) = \ell_D(d) - \ell_D(d-1) = \sum_{k=0}^{D} \binom{D}{k}\binom{d+D-k-1}{D-1},$$

where we used $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$.
We will now use a proof technique called *double counting* to prove that

$$\sum_{k=0}^{D} \binom{D}{k}\binom{d+D-k-1}{D-1} \tag{A.1}$$

is equal to

$$\sum_{k=0}^{D-1} \binom{D}{k}\binom{d-1}{D-1-k} 2^{D-k}. \tag{A.2}$$

For this, we first give a combinatorial interpretation of (A.1), then formulate an equivalent statement which at last results in (A.2).
Let $X = \{X_1, \ldots, X_D\}$ be a set with D elements and let $Y = \{Y_1, \ldots, Y_{d-1}\}$ be a set with $d-1$ elements with $X \cap Y = \emptyset$. Then (A.1) counts the number of ways for choosing subsets $A \subseteq X$ and $B \subseteq X \cup Y$ with $|B| = D-1$ and $A \cap B = \emptyset$. This can be seen as follows: If $0 \leq k \leq D$ is the number of elements in A, then $\binom{D}{k}$ counts the number of ways for choosing A. Since $A \cap B = \emptyset$ there are $D + (d-1) - k$ elements left for the set B. Thus, the number of ways for choosing B with $|B| = D-1$ is given by $\binom{d+D-k-1}{D-1}$. The sum over the number of elements in A gives (A.1).
Now, choosing such a $B \subseteq X \cup Y$ with $|B| = D-1$ and $A \cap B = \emptyset$ is equivalent to choosing subsets $N \subseteq X$ and $M \subseteq Y$ such that $|M|+|N| = D-1$ and $(N \cup M) \cap A = \emptyset$. Hence, we now count the number of ways for choosing subsets $A \subseteq X$, $N \subseteq X$ and $M \subseteq Y$ with $|M| + |N| = D-1$ and $(N \cup M) \cap A = \emptyset$. If $1 \leq k \leq D-1$ is the number of elements in M, then there are $\binom{D}{k}$ ways for choosing M. The number of ways for choosing the left $D-1-k$ elements of N out of Y is given by $\binom{d-1}{D-1-k}$. Since $(N \cup M) \cap A = \emptyset$ there are $D-k$ elements left for A, i.e., there are 2^{D-k} ways for choosing A. The sum over the number of elements in M gives (A.2).
As a last step, we need to bound (A.2), where we use $\binom{n}{k} = \frac{n}{n-k}\binom{n-1}{k}$, $\binom{n}{k} \leq \frac{n^k}{k!}$ and

$2^n \leq (n+1)!$. We then have

$$\begin{aligned}
\sum_{k=0}^{D-1} \binom{D}{k}\binom{d-1}{D-1-k} 2^{D-k} &= \sum_{k=0}^{D-1} \frac{D}{D-k}\binom{D-1}{k}\binom{d-1}{D-1-k} 2^{D-k} \\
&\leq D \sum_{k=0}^{D-1} \frac{1}{D-k}\binom{D-1}{k} \frac{(d-1)^{D-1-k}}{(D-1-k)!} 2^{D-k} \\
&= 2D \sum_{k=0}^{D-1} \binom{D-1}{k} \frac{(d-1)^{D-1-k}}{(D-k)!} 2^{D-k-1} \\
&\leq 2D \sum_{k=0}^{D-1} \binom{D-1}{k} (d-1)^{D-k-1} \\
&= 2Dd^{D-1},
\end{aligned}$$

where the last equality comes from the binomial formula for $((d-1)+1)^{D-1}$. □

We want to remark that the estimate from Lemma 2.4 tends to severely overestimate the actual size of the level sets for larger values of D. This is exclusively due to the constant factor $2D$, while the factor d^{D-1} is actually the sharpest one possible. Asymptotically, for fixed D and growing d, we have

$$\ell_d^{=} \sim \frac{2^D}{(D-1)!} d^{D-1}.$$

This can be seen by carefully examining the proof above, noting that the asymptotic behavior is governed by the term corresponding to $k = 0$, i.e., $\binom{d-1}{D-1} 2^D$ and that asymptotically, for growing d we have $\binom{d-1}{D-1} \sim \frac{d^{D-1}}{(D-1)!}$.

REFERENCES

[1] M. Beck and S. Robins, *Computing the Continuous Discretely: Integer-Point Enumeration in Polyhedra*, Undergrad. Texts Math., Springer New York, 2015.

[2] C. Bekas, A. Curioni, and I. Fedulova, *Low cost high performance uncertainty quantification*, in Proceedings of the 2nd Workshop on High Performance Computational Finance, 2009, pp. 1–8.

[3] M. Benzi, P. Boito, and N. Razouk, *Decay Properties of Spectral Projectors with Applications to Electronic Structure*, SIAM Rev., 55 (2013), pp. 3–64.

[4] M. Benzi and G. H. Golub, *Bounds for the Entries of Matrix Functions with Applications to Preconditioning*, BIT, 39 (1999), pp. 417–438.

[5] M. Benzi and N. Razouk, *Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices*, Electron. Trans. Numer. Anal., 28 (2007), pp. 16–39.

[6] M. Benzi and V. Simoncini, *Decay bounds for functions of Hermitian matrices with banded or Kronecker structure*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1263–1282.

[7] J. Bloch, A. Frommer, B. Lang, and T. Wettig, *An iterative method to compute the sign function of a non-Hermitian matrix and its application to the overlap Dirac operator at nonzero chemical potential*, Computer Physics Communications, 177 (2007), pp. 933–943.

[8] M. Bollhöfer, A. Eftekhari, S. Scheidegger, and O. Schenk, *Large-scale Sparse Inverse Covariance Matrix Estimation*, SIAM J. Sci. Comput., 41 (2019), pp. A380–A401.

[9] K. Burrage, N. Hale, and D. Kay, *An Efficient Implicit FEM Scheme for Fractional-in-Space Reaction-Diffusion Equations*, SIAM J. Sci. Comput., 34 (2012).

[10] K. Y. Cheng, *Minimizing the bandwidth of sparse symmetric matrices*, Computing, 11 (1973), pp. 103–110.

[11] M. Crouzeix and C. Palencia, *The numerical range is a $(1+\sqrt{2})$-spectral set*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 649–655.

[12] E. Cuthill and J. McKee, *Reducing the Bandwidth of Sparse Symmetric Matrices*, in Proceedings of the 1969 24th National Conference, ACM '69, New York, NY, USA, 1969, ACM, pp. 157–172.

[13] S. Demko, W. F. Moss, and W. Smith, *Decay rates for inverses of banded matrices*, Math. Comp., 43 (1984), pp. 491–499.

[14] S. Dong and K. Liu, *Stochastic estimation with z_2 noise*, Phys. Lett. B, 328 (1994), pp. 130–136.

[15] V. Eijkhout and B. Polman, *Decay rates of inverses of banded M-matrices that are near to Toeplitz matrices*, Linear Algebra Appl., 109 (1988), pp. 247–277.

[16] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. A. van der Vorst, *Numerical methods for the QCD overlap operator, I. Sign-function and error bounds*, Comput. Phys. Commun., 146 (2002), pp. 203–224.

[17] E. Estrada, *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, Inc., New York, 2011.

[18] E. Estrada and D. Higham, *Network Properties Revealed Through Matrix Functions*, SIAM Rev., 52 (2010), pp. 696–714.

[19] E. Estrada and J. A. Rodríguez-Velázquez, *Subgraph Centrality in Complex Networks*, Phys. Rev. E, 71 (2005), p. 056103.

[20] G. Fertin, E. Godard, and A. Raspaud, *Acyclic and k-distance coloring of the grid*, Inform. Process. Lett., 87 (2003), pp. 51 – 58.

[21] 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 (2018), p. e2131.

[22] A. Frommer, C. Schimmel, and M. Schweitzer, *Non-Toeplitz decay bounds for inverses of Hermitian positive definite tridiagonal matrices*, Electron. Trans. Numer. Anal., 48 (2018), pp. 362–372.

[23] A. Frommer and V. Simoncini, *Matrix functions*, in Model Order Reduction: Theory, Research Aspects and Applications, W. H. A. Schilders, H. A. van der Vorst, and J. Rommes, eds., Springer, Berlin Heidelberg, 2008, pp. 275–303.

[24] N. E. Gibbs, W. G. Poole, and P. K. Stockmeyer, *An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix*, SIAM J. Numer. Anal., 13 (1976), pp. 236–250.

[25] Y. Ginosar, I. Gutman, T. Mansour, and M. Schork, *Estrada index and Chebyshev polynomials*, Chem. Phys. Lett., 454 (2008), pp. 145–147.

[26] G. H. Golub, M. Heath, and G. Wahba, *Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter*, Technometrics, 21 (1979), pp. 215–223.

[27] G. H. Golub and G. Meurant, *Matrices, Moments and Quadrature*, in Numerical Analysis 1993, D. F. Griffiths and G. A. Watson, eds., Essex, 1994, Longman Scientific & Technical, pp. 105–156.

[28] ———, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton and Oxford, 2010.

[29] G. H. Golub and U. von Matt, *Generalized Cross-Validation for Large Scale Problems*, J. Comput. Graph. Statist., 6 (1995), pp. 1–34.

[30] S. L. Gonzaga de Oliveira, J. A. B. Bernardes, and G. O. Chagas, *An evaluation of low-cost heuristics for matrix bandwidth and profile reductions*, Comput. Appl. Math., 37 (2018), pp. 1412–1471.

[31] S. Güttel, *Rational Krylov Methods for Operator Functions*, PhD thesis, Fakultät für Mathematik und Informatik der Technischen Universität Bergakademie Freiberg, 2010.

[32] I. Han, D. Malioutov, and J. Shin, *Large-scale log-determinant computation through stochastic Chebyshev expansions*, in International Conference on Machine Learning, 2015, pp. 908–917.

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

[34] M. Kubale, *Graph Colorings*, vol. 352 of Contemporary mathematics, American Mathematical Soc., 2004.

[35] J. Laeuchli and A. Stathopoulos, *Extending hierarchical probing for computing the trace of matrix inverses*, SIAM J. Sci. Comput., 42 (2020), pp. A1459–A1485.

[36] A. Lim, B. Rodrigues, and F. Xiao, *A fast algorithm for bandwidth minimization*, Int. J. Artif. Intell. T., 16 (2007), pp. 537–544.

[37] S. J. Miller, *An identity for sums of polylogarithm functions*, Integers, 8 (2008), p. A15.

[38] R. Nabben, *Two-sided bounds on the inverses of diagonally dominant tridiagonal matrices*, Linear Algebra Appl., 287 (1999), pp. 289–305.

[39] H. Neuberger, *The Overlap Dirac Operator*, in Numerical Challenges in Lattice Quantum

Chromodynamics. Lecture Notes in Computational Science and Engineering, F. A., L. T., M. B., and S. K., eds., vol. 15, Springer, Berlin, Heidelberg, 2000.

[40] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall Series in Computational Mathematics, Pearson Education Canada, 1980.

[41] A. N. Pettitt, I. S. Weir, and A. G. Hart, *A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data*, Statistics and Computing, 12 (2002), pp. 353–367.

[42] G. Pleiss, M. Jankowiak, D. Eriksson, A. Damle, and J. R. Gardner, *Fast matrix square roots with applications to Gaussian processes and Bayesian optimization*, tech. rep., 2020. arXiv:2006.11267.

[43] S. Pozza and V. Simoncini, *Inexact Arnoldi residual estimates and decay properties for functions of non-Hermitian matrices*, BIT, 59 (2019), pp. 969–986.

[44] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

[45] J. Reid and J. Scott, *Reducing the total bandwidth of a sparse unsymmetric matrix*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 805–821.

[46] J. D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Int. J. Control, 32 (1980), pp. 677–687.

[47] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory And Applications*, CRC press, 2005.

[48] B. Sapoval, T. Gobron, and A. Margolina, *Vibrations of fractal drums*, Phys. Rev. Lett., 67 (1991), pp. 2974–2977.

[49] C. Schimmel, *Bounds for the decay in matrix functions and its exploitation in matrix computations*, PhD thesis, Bergische Universität Wuppertal, 2019.

[50] J. Sexton and D. Weingarten, *Systematic expansion for full QCD based on the valence approximation*, tech. rep., Watson Research Center, 1994. arXiv:hep-lat/9411029.

[51] W. Smyth, *Algorithms for the reduction of matrix bandwidth and profile*, J. Comput. Appl. Math., 12–13 (1985), pp. 551–561.

[52] 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 (2013), pp. 299–322.

[53] Z. Strakoš, *Model reduction using the Vorobyev moment problem*, Numer. Algorithms, 51 (2009), pp. 363–379.

[54] J. M. Tang and Y. Saad, *A probing method for computing the diagonal of a matrix inverse*, Numer. Linear Algebra Appl., 19 (2012), pp. 485–501.

[55] S. Ubaru, J. Chen, and Y. Saad, *Fast estimation of* $\mathrm{tr}(f(A))$ *via stochastic lanczos quadrature*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1075–1099.

[56] S. Ubaru and Y. Saad, *Applications of trace estimation techniques*, in International Conference on High Performance Computing in Science and Engineering, Springer, 2017, pp. 19–33.