Title

Spectral Problems and Probabilistic Techniques

ISBN

9798189271106

Author

Detherage, Isabel

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Spectral Problems and Probabilistic Techniques

by

Isabel Detherage

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nikhil Srivastava, Chair
Professor Svetlana Jitomirskaya
Professor James Demmel

Spring 2026

Spectral Problems and Probabilistic Techniques

Abstract

Spectral Problems and Probabilistic Techniques

by

Isabel Detherage

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Nikhil Srivastava, Chair

This thesis investigates several problems unified by the central role of eigenvalues. Leveraging probabilistic techniques, we establish the following results:

1. We analyze a general class of matrix factorization algorithms, capable of computing standard factorizations such as QR, SVD, eigendecomposition, and Cholesky decomposition. Under a uniformly random pivoting strategy, any algorithm of this class is shown to exhibit the same linear rate of convergence, irrespective of the specific factorization computed.

2. A general framework for computing spectral sums is introduced, leveraging rational approximations of functions and stochastic trace estimation. We apply this framework to two applications: spectral density estimation and Schatten-1 norm approximation. We give an algorithm for spectral density estimation with local guarantees and an algorithm for Schatten-1 norm approximation improving upon the best known runtime.

3. We study the infimum of the spectrum, or ground state energy, of a discrete Schrödinger operator on $\theta\mathbb{Z}^d$ parameterized by a potential $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and a frequency parameter $\theta \in (0, 1)$. We relate this ground state energy to that of a corresponding continuous semiclassical Schrödinger operator on $\mathbb{R}^d$ with parameter θ, arising from the same choice of potential. We prove two-sided, non-asymptotic bounds relating these quantities.

4. Lastly, we prove both lower and upper bounds on the Ollivier-Ricci curvature of the basis exchange walk on a matroid. We give several examples of non-negatively curved basis exchange walks and negatively curved basis exchange walks.

Though these problems are distinct in nature, together they demonstrate the versatility of probabilistic methods in spectral problems.

Contents

Acknowledgments

First and foremost, thank you to Nikhil for being an incredible advisor. He gave advice that I have shared with dozens of other students, introduced me to areas of math that I loved, and showed great patience while I grew into a mathematician and researcher. Additionally, I would like to thank James Demmel for helpful feedback on this dissertation, Svetlana Jitomirskaya for serving on my dissertation committee, Shirshendu Ganguly for providing me with a strong foundation in probability theory, and Stefan Steinerberger for interesting conversations around research.

I undoubtedly would not have pursued a graduate degree without the encouragement and support of my undergraduate advisor, Wilhelm Schlag. His teaching sparked my initial desire to go to graduate school, which only grew while working with him further. I thank him for his unwavering belief in me (particularly at times when he believed in me far more than I believed in myself).

In a slightly unconventional turn, I must thank Dr. Mickey Collins; without his dedication to researching and treating concussions, I certainly would not have finished this degree. Thank you to Dr. Noah Bach Bar-Tura for her support during my first three years of graduate school, especially while I was recovering from my concussions.

Thank you to the fellow graduate students I collaborated with over the past years: Rikhav Shah, Zack Stier, Apoorv Vikram Singh, and Nicholas West. The work in this dissertation was far more enjoyable (not to mention possible) thanks to them.

I am also incredibly grateful to the many friends from graduate school that went through this experience alongside me: Mitsuki, for always being at the ready to gossip and / or commiserate (and for also being my roommate); Lewis, for his (sometimes brutal) honesty and the many hours we've spent bickering and gossiping; Robin, for being the officemate of my dreams and my go-to confidante; and Will, for his support and camaraderie during the tough days that frequented the first half of graduate school.

I am indebted to the many friends not in graduate school who served as a much needed reminder that there is more to the world than math. Thank you especially to Carly and Joyce for serendipitously finding themselves in San Francisco exactly when I found myself in Berkeley. Thank you to Katherine for many shared miles on the Ohlone Greenway, to Kathy for always being there when I need her most, and to Laura, Nick, Nico, and Rose for letting me become part of their community.

Thank you to my parents and my sisters for their constant support and for learning the words "theoretical computer science" and reciting them with diligence anytime they answer questions about what I do.

Finally, I must thank my parents for their unwavering support in my academic endeavors and for their dedication to fostering my love of learning from a young age. I thank them for advocating for me before I could do so for myself, particularly in a school system that was not prepared to support me as a student. They went above and beyond to make sure that I was always learning, especially in math. It seems that their efforts were not in vain.

Chapter 1

Introduction

We explore several questions with two common features: eigenvalues (or some spectral component) and probabilistic methods.

Eigenvalues are fundamental across numerous areas of mathematics and science. We name just a few prominent examples: in data analysis, eigenvalues of certain matrices are used for dimensionality reduction techniques (such as principal component analysis); in graph theory, many prominent algorithms leverage spectral information for complex structural tasks such as graph partitioning or community detection; and in probability theory, analysis of eigenvalues provides mixing time bounds for Markov chains. Given the importance of eigenvalues, significant resources have been devoted to developing computationally feasible methods for quickly finding eigenvalues (or rather, approximating eigenvalues). It is perhaps unsurprising that spectral problems, either focused on efficient computation of spectral information or applications to specific settings, are pervasive in mathematics.

We will consider several spectral problems in this thesis. Across these problems, the role of eigenvalues varies. In some instances, the focus is on finding estimates for eigenvalues or other spectral quantities, as in Chapter 3 and Chapter 4. In others, we use eigenvalues to measure the progress of a stochastic process, as in Chapter 2 and Chapter 3; here, the goal is not to estimate eigenvalues themselves, but to understand how they evolve with the stochastic process. Finally, in Chapter 5 and Chapter 6, we provide provide bounds on the smallest eigenvalue or the spectral gap of certain operators.

Probabilistic methods play a central role across all of these problems as well. In Chapter 2, Chapter 3, and Chapter 6, the role of probability is inherent to the problem: we are analyzing stochastic processes or probabilistic quantities directly. In Chapter 4 and Chapter 5, however, randomness is not inherent to the problem; rather we leverage it as a tool to design algorithms and to prove deterministic bounds on spectra.

We first introduce some necessary notation, then give an overview of the various problems explored in this thesis, presenting the main results and key ideas used to prove them.

Notation. We define notation for several spectral quantities:

- For an $n \times n$ real, symmetric matrix A, let $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A)$ denote its eigenvalues.
- For an $n \times d$ matrix A with $n > d$, let $\sigma_1(A) \geq \ldots \geq \sigma_d(A)$ denote its singular values.
- For an $n \times d$ matrix A with linearly independent columns, let $\kappa(A) = \sigma_1(A)/\sigma_d(A)$. This is known as the *condition number* of A; this is equivalent to $\kappa(A) = \|A\|_2 \cdot \left\|A^\dagger\right\|_2$ where $A^\dagger = (A^T A)^{-1} A^T$, the Moore-Penrose pseudo-inverse, and $\|A\|_2 = \sigma_1(A)$.

Additionally, for vectors $u, v \in \mathbb{R}^n$ we let $u \cdot v = \sum_{i=1}^n u_i v_i$ be the standard dot product.

Within this chapter, we only consider real matrices A and real vectors v; in later chapters, we will comment on when we extend to the complex case.

Bibliographic note. The content of Chapter 2 and Chapter 3 is joint work with Rikhav Shah; Chapter 2 appeared in [DS26b] and Chapter 3 has been submitted for publication (preprint: [DS25]). The content of Chapter 4 is part of ongoing joint work with Apoorv Vikram Singh and Nicholas West. The content of Chapter 5 is joint work with Nikhil Srivastava and Zack Stier, appearing in the publication [DSS24]. Parts of Chapter 6 appeared in the preprint [Det25] and have been submitted for publication. Parts of this chapter introducing each problem are adapted from the corresponding paper or preprint.

1.1 Chapter 2: Stochastic orthogonalization

Suppose we have some set of linearly independent unit vectors

$$\{v_1, v_2, \ldots, v_k\},$$

and we would like to generate an orthonormal basis for the span of these vectors. Recall that the well-known Gram-Schmidt algorithm orthogonalizes a set of vectors by replacing v_i with the projection of v_i onto the orthogonal complement of $\text{span}\{v_1, \ldots, v_{i-1}\}$. We can view the Gram-Schmidt process as an iterative process on pairs of vectors: choose two vectors v_i, v_j and replace v_j with $v_j - \frac{v_j \cdot v_i}{v_i \cdot v_i} v_i$, renormalized. In order for the Gram-Schmidt process to return an orthogonal set of vectors, we must follow a very careful ordering of pairs of vectors – specifically we apply this process to the following pairs in the following order: $(1,2), (2,3), (1,3), (3,4), (2,4), (1,4), \ldots$

Now suppose instead we applied this process to *randomly* chosen pairs of vectors. We ask two questions: does this process converge to an approximately orthogonal set? If so, at what rate?[1]

[1]There is a lingering question of *why* we would want to do this. For the reader familiar with numerical linear algebra, we remark that our original inspiration was to apply this in conjunction with a Kaczmarz solver. However, we found that this random strategy has more meaning when we connect this process to matrix factorizations; this is explored fully in Chapter 3.

To answer the first question, we must define more precisely what we mean for a set to converge to an approximately orthogonal set, both in terms of how we define approximately orthogonal and what sense of convergence. We consider convergence in probability; informally, we show the probability our vectors remain far from orthogonal goes to zero as the process runs. To measure how close our vectors are to orthogonal, we will use eigenvalues.

Potential function. Let $A = \begin{bmatrix} v_1 & v_2 & \dots & v_k \end{bmatrix}$, a matrix with unit vectors v_i as the columns. Note $\{v_1, \dots, v_k\}$ are orthogonal if and only if $A^T A = I$, i.e., all of the eigenvalues of $A^T A$ are 1. We define the potential function

$$\Phi(A) = -\frac{1}{2} \sum_{i=1}^{n} \log \lambda_i(A^T A).$$

We see this is zero if and only if the vectors are orthogonal, and we say our vectors are "close to orthogonal" if $\Phi(A)$ is close to 0. We can also define "close to orthogonal" in the more traditional sense, using the condition number $\kappa(A)$; more specifically, we use an optimization problem to relate $\Phi(A)$ and $\kappa(A)$.

Contributions. Let $\left\{ v_1^{(t)}, \dots v_k^{(t)} \right\}$ denote the set of unit vectors after t orthogonalization steps, where in each step a random pair of vectors is orthogonalized and the subsequent vectors are normalized. Let $A^{(t)}$ denote a matrix with $v_i^{(t)}$ as its columns. The key feature of our potential function $\Phi(A^{(t)})$ is that it is a *supermartingale*:

Definition 1.1.1 (Supermartingale)**.** A sequence of random variables $X_0, X_1, \dots$ defined over some probability space Ω is a *supermartingale* if for every n we have

$$\mathbb{E}(X_n | X_{n-1}, \dots, X_0) \leq X_{n-1}$$

where $\mathbb{E}(X_n | X_{n-1}, \dots, X_0)$ is the conditional expectation of X_n with respect to the sigma-algebra induced by $X_{n-1}, \dots, X_0$.

Informally, this means that on average, conditioned on all of the previous steps, $\Phi(A^{(t)})$ is decreasing. Leveraging properties of supermartingales, we show that this process converges to an approximately orthogonal set and give a preliminary bound on the rate of convergence. More specifically, we show:

Theorem 1.1.1 (Tail bound on $\Phi(A^{(t)})$)**.** *Fix any A with linearly independent columns of unit length and $\beta < 1$. For any $\delta > 0$,*

$$t = \Theta\left(n^2 \log\left(\frac{1}{|\det(A)\beta|} \right) \log\left(\frac{1}{\delta} \right) \right)$$

implies

$$\mathbb{P}\big(\Phi(A^{(t)}) \geq \beta \big) \leqslant \delta.$$

In Chapter 3, we both improve the rate of convergence (achieving the optimal rate, as conjectured in [Ste25]) and connect this process to a broad class of random algorithms for matrix factorizations.

1.2 Chapter 3: Matrix factorizations

We revisit the above orthogonalization procedure, connecting it to Jacobi's eigenvalue algorithm and showing how both can be viewed as specific instances in a class of iterative algorithms for matrix factorizations.

Connection to Jacobi's eigenvalue algorithm. Given a matrix A with linearly independent columns $\{v_1, \ldots, v_k\}$ of unit length, we view the above iterative orthogonalization process through the following lens: form A^TA, choose an off-diagonal entry (i, j) with $i < j$ uniformly at random, then form $R_{ij}^T A^T A R_{ij}$ where R is an upper triangular matrix with

$$R_{ij} = \begin{array}{c} \\ i \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} i & & j & & \end{array} \\ \begin{bmatrix} I & & & & \\ & 1 & & \frac{v_i \cdot v_j}{1-(v_i \cdot v_j)^2} & \\ & & I & & \\ & 0 & & \frac{1}{1-(v_i \cdot v_j)^2} & \\ & & & & I \end{bmatrix} \end{array}. \tag{1.1}$$

The matrix $R_{ij}^T A^T A R_{ij}$ now has zero in the (i, j) and (j, i) entries, and the rest of the matrix is unchanged except for the j^{th} row and column.

With this framing, we are reminded of Jacobi's eigenvalue algorithm for diagonalization, an iterative algorithm for diagonalizing a symmetric matrix B. At each step, we choose an off-diagonal entry (i, j) uniformly at random, then perform conjugation with an *orthogonal* matrix to zero out the (i, j) entry. More specifically, we choose Q_{ij} to be a Givens rotation,

$$Q_{ij} = \begin{array}{c} \\ i \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} i & & j & & \end{array} \\ \begin{bmatrix} I & & & & \\ & \cos\theta & & -\sin\theta & \\ & & I & & \\ & \sin\theta & & \cos\theta & \\ & & & & I \end{bmatrix} \end{array}, \tag{1.2}$$

where I is the appropriate size identity matrix and θ is chosen so that $Q_{ij}^T B Q_{ij}$ will be zero in the (i, j) and (j, i) entries. Note by the form of Q_{ij}, $Q_{ij}^T B Q_{ij}$ will only differ from B in the i, j rows and columns.

General algorithm for matrix factorizations. Both processes can be described under the following general framework. We start with some symmetric matrix B, either given or formed by taking $B = A^TA$ of a given matrix A. Initialize $B^{(0)} = B$, then at each iteration:

- Choose some off-diagonal index (i, j) uniformly at random (called a *pivot*).
- Choose S from some allowed set of matrices. This allowed set is some subset of matrices S such that $S^T B^{(t)} S$ has zero in the (i, j) and (j, i) entries and $S^T B^{(t)} S$ differs from $B^{(t)}$ only in the i, j rows and columns.

- Update $B^{(t+1)} \leftarrow S^T B^{(t)} S$.

Repeat until $B^{(t)}$ is diagonal, or sufficiently close to diagonal.

By tracking the S matrices used at each step, this process can be viewed as an iterative algorithm for computing matrix factorizations under a uniformly random pivoting strategy. We note that several major matrix factorizations can be computed with this framework: if we iterate with matrices of the form (1.1) and track the accumulations of upper triangular matrices R used at each step, this can be used to find the QR factorization of A, or the Cholesky decomposition of $A^T A$; if we iterate with matrices of the form (1.2) and track the accumulation of orthogonal matrices Q, this will give us the eigendecomposition of B; if we assume $B = A^T A$, this can be used to find the singular value decomposition of A.

Contributions. We show this process always converges to a diagonal matrix. Notably, this gives a unified proof of convergence for several significant (and previously unconnected) matrix factorizations.

We again use eigenvalues to track progress of an algorithm of this form. However, depending on the desired factorization, we have very different goals in how we want the eigenvalues to behave. If we are trying to compute the QR factorization, we want to apply upper triangular transformations to A until it has orthonormal columns, or $A^T A = I$. If we are trying to compute the SVD, we apply orthogonal transformations until $A^T A$ is diagonal; these transformations will preserve the eigenvalues. In one case, our goal is to change all of the eigenvalues to 1, while in the other, we will maintain them. In light of this, our previous potential function is insufficient for a unified analysis of all of these algorithms, as it is dependent on our goal being to transform all the eigenvalues to 1.

To define our new potential function, we first define $D_B = B \odot I$, where $\odot$ is the Hadamard product (i.e., $(A \odot B)_{ij} = A_{ij} B_{ij}$) and $\hat{B} = D_B^{-1/2} B D_B^{-1/2}$, the diagonal-normalization of B. Our new potential function is

$$\Gamma(B) = \operatorname{tr}(\hat{B}^{-1}) - n = \sum_{i=1}^{n} \left(\frac{1}{\lambda_i(\hat{B})} - 1 \right).$$

We will see that $\Gamma(B) = 0$ if and only if B is diagonal. For positive semi-definite B, we can make this relationship quantitative; first we define a notion of distance to diagonal, then we solve an optimization problem to bound this distance for a fixed value of $\Gamma(B)$.

Again, we find that $\Gamma(B^{(t)})$ is a supermartingale. Leveraging this, we show $\Gamma(B^{(t)})$ converges to zero in expectation.

Theorem 1.2.1. *Let $0 < \delta < 1$. For any positive definite B, $t \geq n^2 \log(n\kappa(\hat{B})/\delta)$ implies*

$$\mathbb{E}(\Gamma(B^{(t)})) \leq \delta.$$

Using this, we prove convergence of these iterative matrix factorization algorithms. We connect $\Gamma(B)$ to other more traditional quantities related to convergence of various matrix factorizations (for example, $\kappa(B)$ for QR factorization, $\|B - B \odot I\|_F$ for diagonalization) through optimization problems.

1.3 Chapter 4: Spectral sums and rational functions

In the above section, we studied a (randomized) algorithm to compute a full spectral decomposition. Computing such a decomposition, even with optimized algorithms, is generally expensive, and in many applications, we require spectral information more efficiently. More concretely, we consider the problem of computing a spectral sum, i.e., for a symmetric matrix $A \in \mathbb{R}^{n\times n}$, computing

$$\operatorname{tr}(f(A)) = \sum_{i=1}^{n} f(\lambda_i(A)),$$

without computing a full eigendecomposition. We provide a general framework for computing spectral sums, leveraging rational approximations of functions and stochastic trace estimation, such as that of [MMMW21].

For a given error tolerance, rational function approximations generally have lower degree compared to polynomial approximations for non-smooth functions, such as an indicator function or an absolute value function. However, evaluating rational matrix functions requires solving linear systems, which can be a bottleneck to computation time. Thus there is a tradeoff between the degree of the approximation and the time it takes to evaluate the approximation. Historically, polynomial approximations have been preferred: although one needs a higher degree for a desired error rate, evaluating $p(A)y$ for a polynomial p only requires matrix vector products with A.

In some instances though, when fast solvers are available (for example in certain structured matrices [ST14]), this tradeoff favors rational approximations. With advances in linear system solvers, such as [DS26a], rational approximations can be favorable for general matrices; this is largely problem dependent, depending on what improvement these system solvers give.

Applications. We apply this framework for two applications: spectral density estimation, via eigenvalue counts for intervals, and Schatten-1 norm estimation. Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the spectral density measure is

$$s_A(x) = \sum_{i=1}^{n} \frac{1}{n}\delta(x - \lambda_i(A)) \tag{1.3}$$

where $\delta(x)$ is a Dirac mass at 0. Approximating the spectral density has long been an important problem in chemistry and physics [DC70], [Tur88], [DS93], and has also found more recent applications in data analysis and machine learning (see for example [YGKM20] or [RL18]). Given local estimations of eigenvalue counts in intervals, one can construct a spectral density estimation; in this way, a related problem is to approximate the number of eigenvalues in a given interval I. Letting $\mathbf{1}_I$ be the indicator of the interval I, we can write

$$\text{number of eigenvalues in } I = \sum_{i=1}^{n} \mathbf{1}_I(\lambda_i(A)) =: \operatorname{tr} \mathbf{1}_I(A).$$

The *Schatten-1 norm* of a matrix $A \in \mathbb{R}^{n\times d}$, also known as the *nuclear norm* or the *trace norm*, is

$$\|A\|_* := \sum_{i=1}^{d} \sigma_i(A) = \sum_{i=1}^{d} \sqrt{\lambda_i(A^TA)} = \mathrm{tr}(\sqrt{A^TA}). \tag{1.4}$$

This is a quantity of interest in many fields: in matrix completion algorithms, it is used as a substitute for matrix rank [JNS13], [CR12], and it is used for applications in theoretical chemistry and differential privacy [Gut01], [HLM12].

Contributions. We provide explicitly computable rational approximations of a step function which avoid the complexity of elliptic integrals of the well-known *Zolotarev functions*. While Zolotarev functions provide the optimal-degree approximation with respect to error parameters, they are computationally cumbersome to implement. Our approach, which builds on a rational approximation of [PP11], yields functions that are simpler but inherit the desirable properties of Zolotarev functions: logarithmic degree with respect to error parameters, purely imaginary poles, and smallest pole bounded away from zero. Using this step function, we build rational approximations of several important non-smooth functions. We apply these rational approximations (using some choice of linear system solvers), then use stochastic trace estimators to approximate a desired spectral sum.

We give an algorithm for spectral density estimation with local guarantees as well as a Wasserstein distance guarantee, the first of its kind. In proving this algorithm, we show a novel connection between optimal degree rational approximations and optimal degree polynomial approximations of a step function.

Theorem 1.3.1 (Informally)**.** *Fix $\varepsilon > 0$. Let A be a symmetric matrix normalized as above. There is an algorithm that with probability at least $99/100$ returns a partitioning of $[-1, 1]$ into intervals I_t of width less than or equal to ε and estimates $\tilde{k}_t$ such that*

$$(1-\varepsilon)k_t - \varepsilon^2(k_{t-1} + k_t + k_{t+1}) \le \tilde{k}_t \le (1+\varepsilon)k_t + \varepsilon(k_{t-1} + k_t + k_{t+1})$$

where k_t is the true number of eigenvalues of A in I_t. Additionally, we can define the probability measure $\tilde{s}_A$ with point masses at the midpoints of each I_t with weight proportional to $\tilde{k}_t$. Then $\mathcal{W}(s_A, \tilde{s}_A) \le \varepsilon$ where $\mathcal{W}(\cdot,\cdot)$ denotes the Wasserstein distance. The runtime of this algorithm is $\tilde{O}(n^2/\varepsilon^4)$.

For the Schatten-1 norm, we give an algorithm that can be modified depending on the sparsity of the matrix A to optimize the runtime. We state our result for the dense case here.

Theorem 1.3.2. *Assuming $d = n$, $nnz(A) = n^2$, Algorithm 3 yields a $(1 \pm \varepsilon)$ approximation of the Schatten-1 norm with probability $99/100$ in time $\tilde{O}(n^2/\varepsilon^{1.5})$.*

This is an improvement in ε dependency over the previous best known runtime from [DS26a], the previous best runtime. Interestingly, their algorithm relies on first computing a spectral density estimation well-suited for this spectral sum, also via rational approximations. Our algorithm directly approximates the Schatten-1 norm.

1.4 Chapter 5: Ground state energies of Schrödinger operators

In this chapter, we consider the *ground state energy*, or infimum of the spectrum, of discrete and semi-classical *Schrödinger* operators. Informally, these are operators that contain a Laplacian term, measuring how quickly a function changes, and a potential term, measuring how much mass a function places at various points. These operators will be over infinite dimensional vector spaces, so we briefly comment on the spectrum of such operators.

Operators over infinite-dimensional vector spaces. For a linear operator T over finite-dimensional vector spaces, we define eigenvalues as scalars λ such that there exists a non-zero vector x satisfying $T(x) = \lambda x$. Equivalently, we can define eigenvalues as the set of λ such that $T - \lambda I$ is not invertible; the spectrum is the set of all such values. When we move to linear operators over infinite dimensional vector spaces, this second view is the appropriate definition to extend: we define the spectrum $\sigma(T)$ to be all λ such that $T - \lambda I$ is not invertible. Note, however, that $\lambda \in \sigma(T)$ need not imply that $T(x) = \lambda x$ for some x in this setting.

For a linear operator T, we define the ground state energy as $\mu(T) := \inf \sigma(T)$. Although there might not be an eigenfunction associated with $\mu(T)$, for self-adjoint T that are sufficiently "well-behaved," there is a variational characterization of $\mu(T)$, similar to the finite-dimensional case. More concretely, for an operator T acting on functions f from some Hilbert space, we can define

$$\mu(T) := \inf_{f:\|f\|^2=1} \langle f, Tf \rangle$$

where $\|f\|^2 = \langle f, f \rangle$. The operators we consider all admit such a variational characterization.

Schrödinger operators. More specifically, let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a smooth $\mathbb{Z}^d$-periodic potential and let $\theta \in (0, 1)$ be a parameter. We consider the family of discrete Schrödinger operators $A_\theta : \ell_2(\theta\mathbb{Z}^d) \to \ell_2(\theta\mathbb{Z}^d)$ defined by

$$A_\theta f(k\theta) := \left(2df(k\theta) - \sum_{j \sim k} f(j\theta) \right) + V(k\theta)f(k\theta) = (L + V)f(k\theta), \qquad \text{for } k \in \mathbb{Z}^d \quad (1.5)$$

where $j \sim k$ indicates that $j, k \in \mathbb{Z}^d$ differ in exactly one coordinate and L is the discrete Laplacian of the corresponding graph on $\theta\mathbb{Z}^d$.

We can consider the analogous semiclassical (or continuous) d-dimensional Schrödinger operator with potential V and semiclassical parameter θ, $B_\theta : H^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$, defined as

$$B_\theta f(x) := -\theta^2 \sum_{i=1}^{d} \frac{d^2}{dx_i^2} f(x) + V(x)f(x) = (-\theta^2\Delta + V)f(x). \quad (1.6)$$

B_θ is an unbounded self-adjoint nonnegative operator.

Motivated by questions in geometric group theory, there has been interest [BVZ97; BZ05; BDHMW17a; BDHMW17b; Oza22] in proving good upper and lower bounds on $\mu(A_\theta), \theta \in \left[0, \frac{1}{2}\right]$ for certain special cases of V — roughly speaking, this quantity arises naturally in the representation theory of the Heisenberg group and controls the spectral gap of the random walk on that group. More recently, a line of work initiated by Ozawa [Oza16; KNO19; KKN21] introduced a method for producing semidefinite programming proofs of Kazhdan's Property (T) for various groups including $SL_n(\mathbb{Z})$, in which effective bounds on certain $\mu(A_\theta)$ play a role. The best known bounds on $\mu(A_\theta)$ for the potential functions of interest due to [BZ05] rely on delicate trigonometric calculations, and as far as we are aware there is no systematic method for proving such bounds in general.

In contrast, in many cases $\mu(B_\theta)$ can be estimated using well-established tools of spectral theory. Establishing a relationship between $\mu(A_\theta)$ and $\mu(B_\theta)$ allows bounds in one setting to be transferred to the other.

Contributions. We prove two-sided bounds relating $\mu(A_\theta)$ and $\mu(B_\theta)$, the first of their kind. These bounds apply for a wide class of potentials and provide concrete bounds for fixed θ, as is required in the aforementioned applications.

Theorem 1.4.1 (Upper bound)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be continuous and $\mathbb{Z}^d$-periodic. For θ irrational,*

$$\mu(A_\theta) \leqslant \mu(B_\theta).$$

Theorem 1.4.2 (Lower bound)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, not necessarily periodic with $\|V\|_\infty \leqslant 1$, and let $\theta \in (0, 1)$. Suppose that there exists $a = a(\theta) > 0$ such that for all $y \in \theta\mathbb{Z}^d$,*

$$\frac{1}{\theta^d} \int_{\overline{C}(y)} V(x)dx \leqslant a \cdot \frac{1}{2^d} \sum_{k \in C(y)} V(k). \tag{1.7}$$

Then there exists a constant c_d dependent on a, d such that

$$c_d \mu(A_\theta) \geq \mu(B_\theta).$$

Methods. The proofs rely on the variational characterization of the ground state energy and use low-energy functions of one operator to construct low-energy functions of the other. The constructions for both directions can be viewed probabilistically.

For Theorem 1.4.1, we start with $g \in H^2(\mathbb{R}^d)$ and construct a test function $f \in \ell_2(\theta\mathbb{Z}^d)$. The most natural construction is to simply sample g on the lattice $\theta\mathbb{Z}^d$; however, this only takes into account the behavior of g on a null set, so we cannot hope that this sampling will yield a low energy test function for the discrete operator. In order to consider the behavior of g on its full domain, we instead consider g sampled on a randomly shifted lattice $\eta + \theta\mathbb{Z}^d$, where η is drawn uniformly at random from $[0, \theta]^d$. By considering the expected energy of this randomly sampled function, we show there must exist a shift that yields a low

energy function. Some complications arise from taking a shifted lattice, which necessitates the condition that θ is irrational.

For Theorem 1.4.2, we start with $f \in \ell_2(\theta\mathbb{Z}^d)$ and construct $g \in H^2(\mathbb{R}^d)$. Again, we use the natural construction, the multilinear extension of f. We can also view this as a probabilistic construction; consider the function

$$g(x_1, \dots, x_d) := \mathbb{E}_{x_1,\dots,x_d}\left[f(k_{x_1}\theta, \dots, k_{x_d}\theta)\right]$$

where each k_{x_i} is independent with the following distribution for $k_i\theta \leqslant x_i \leqslant (k_i+1)\theta$:

$$\mathbb{P}(k_{x_i} = k_i) = \frac{(k_i+1)\theta - x_i}{\theta},$$
$$\mathbb{P}(k_{x_i} = k_i + 1) = \frac{x_i - k_i\theta}{\theta} = 1 - \mathbb{P}(k_{x_i} = k_i).$$

Note this is precisely the multilinear extension of f. Through this lens, $g(x)$ can be interpreted as the expected value of a random walk on the continuous hypercube starting at x, where the vertices act as absorbing states with values defined by $f(k\theta)$.

Understanding the energy of this construction is significantly more difficult than in the previous direction. The function f only considers the behavior of the potential function on a discrete set of points, making it difficult to reason about the behavior of the potential term in B_θ for the constructed function g. We also offer an alternative lower bound with different assumptions on V; more specifically we require that we have some control over the behavior of V near its minima. This requires more technical definitions to formalize; we explore it fully in Chapter 5.

1.5 Chapter 6: Curvature of basis exchange walks

In the final chapter, we will not consider the spectrum of any operators directly; instead we will be concerned with the *discrete curvature* of a particular class of Markov chains. Considering the connection between curvature and spectral independence, we still consider this a spectral question; we define both notions and explore this connection below.

Discrete Ollivier-Ricci curvature. Considerable effort has been made over the past two decades to develop a discrete analogue of the notion of curvature in geometry. The key feature of curvature is a local-to-global phenomenon: curvature is a strictly local notion but can have global implications. One particularly fruitful definition of discrete curvature has been *Ollivier-Ricci curvature*. To define Ollivier-Ricci curvature, we first recall the definition of Wasserstein distance:

Definition 1.5.1 (Wasserstein distance)**.** Let μ, ν be two probability distributions over some metric space (X, d). The *Wasserstein distance* between μ and ν is defined as

$$\mathcal{W}(\mu, \nu) = \min_{X\sim\mu, Y\sim\nu} \mathbb{E}d(X, Y),$$

where the minimum is over all possible couplings of μ and ν.

Given a reversible Markov chain on a finite state space, we can consider the distance metric induced by the Markov chain, where states x, y have distance one if the probability of moving to x to y is greater than 0. We define the Ollivier-Ricci curvature as follows, where d is this distance metric:

Definition 1.5.2 (Ollivier–Ricci curvature)**.** For a Markov chain P over a metric space (X, d), the *Ollivier–Ricci curvature* of P is the minimum κ such that for all $S, T \in X$ with $S \sim T$

$$\mathcal{W}(P(S, \cdot), P(T, \cdot)) \leq (1 - \kappa).$$

Informally, a Markov chain has non-negative Ollivier-Ricci curvature if random walks started from any two states get closer on average after one step in the Markov chain.

Ollivier-Ricci curvature is a local property, based only on the two-neighborhoods of adjacent states S, T, but a lower bound on the Ollivier-Ricci curvature can imply global properties, such as a lower bound on the spectral gap of the Markov chain [Oll09]. In general, a Markov chain with non-negative curvature can be considered "well-behaved" in many respects; we see one example of this below and review several more in Chapter 6.

Spectral independence and curvature: We now define *spectral independence*, a local property based on the spectrum of certain matrices associated to a measure that implies global properties of an associated Markov chain. The following is a brief overview of spectral independence; the curious reader should see [Liu23] for a comprehensive exploration of spectral independence.

Given a measure π over subsets of a uniform size, $\pi : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, we define the following:

Definition 1.5.3 (Influence matrix)**.** The influence matrix $I_\pi \in \mathbb{R}^{n \times n}$ is defined by

$$I_\pi(i, j) := \mathbb{P}_{S \sim \pi}[j \in S \mid i \in S] - \mathbb{P}_{S \sim \pi}[j \in S].$$

Definition 1.5.4 (Spectral independence)**.** For $\eta \geq 0$, we say π is η-spectrally independent if $\lambda_{max}(I_\pi) \leq 1 + \eta$.

We then consider the *down-up walk* on the support of π. Informally, starting at a set S, one step of the down-up walk drops an element of S uniformly at random (the down step), then adds an element to form a new set T with probability proportional to $\pi(T)$ (the up step). Note this walk is a Markov chain. We say this Markov chain is spectrally independent if the underlying measure π is spectrally independent.

In [Liu21], non-negative Ollivier-Ricci curvature was shown to imply spectral independence. From this result, as well as that of [Oll09] connecting Ollivier-Ricci and the spectral gap of a Markov chain, we see that Ollivier-Ricci curvature is connected to several spectral properties of a Markov chain.

We study the converse of the aforementioned result – does spectral independence imply non-negative Ollivier Ricci curvature?

Basis exchange walks. We explore this question for down-up walks generated by the uniform distribution over bases in a *matroid*.

Definition 1.5.5. A *matroid* $\mathcal{M} = (E, \mathcal{B})$ is a set E and a non-empty collection of its subsets $\mathcal{B}$, called *bases*, satisfying the following:

- no basis is a proper subset of another basis, and
- for any $B_1, B_2 \in \mathcal{B}$, if $b_1 \in B_1 \setminus B_2$, there exists a $b_2 \in B_2 \setminus B_1$ such that $B_1 - b_1 + b_2 \in \mathcal{B}$.

The down-up walk on bases of a matroid (also known as a *basis-exchange walk*) is known to exhibit strong spectral independence [ALGV19].

Contributions. We explore whether basis exchange walks are non-negatively curved, given they exhibit many of the properties of non-negatively curved Markov chains, including spectral independence. Somewhat surprisingly, we find there exist examples of basis exchange walks with negative curvature.

Theorem 1.5.1. *There exist basis exchange walks, coming from both linear matroids and graphic matroids, that are negatively curved.*

This result reinforces the idea that non-negative curvature is a somewhat delicate phenomenon, and shows limitations of the connection of between curvature and spectral independence, both local-to-global phenomena.

Chapter 2

Stochastic orthogonalization

Our first problem concerns the convergence of a stochastic process for approximately orthogonalizing a collection of n linearly independent unit vectors. If we think of these vectors as columns of a matrix A, we can monitor progress of this process through the distribution of $\sigma(A^*A)$: the columns of A are orthogonal if and only $\lambda_i(A^*A) = 1$ for all i. If the columns are not orthogonal then we will have $\mathbb{E}\lambda_i(A^*A) = 1$ (since we take the columns to be unit vectors), but the distribution will not be concentrated exactly at 1. We will use a certain potential function $\Phi(\lambda_1, \ldots, \lambda_n)$ to make this idea quantitative. In particular, in this chapter the potential function we use is $-\log\det(|A|)$ where $|A| = (A^*A)^{1/2}$, i.e.,

$$\Phi(\lambda_1, \ldots, \lambda_n) = -\frac{1}{2}\sum \log \lambda_i.$$

Using this, we show that this stochastic process converges to a set of orthogonal vectors and give a preliminary bound on the rate of convergence. This work can be thought of as a warm up for Chapter 3. As we will see, this orthogonalization scheme can be thought of in a much more general context and related to an iterative algorithm for computing matrix factorizations. We will be able to sharpen the rate of convergence with a different choice of potential function, but we save this for Chapter 3.

2.1 Introduction

We consider a simple procedure for approximately orthogonalizing a collection of n linearly independent unit vectors, thought of as columns of a matrix $A \in \mathbb{C}^{d\times n}, d \geq n$. One iteration of this procedure is as follows: sample two columns, and replace one with its component perpendicular to the other, renormalized to be a unit vector. The only fixed points of this operation are matrices with orthonormal columns. We ask the following questions: *does this procedure converge to such matrices? If so, at what rate?*

We give a positive answer to the first question and a bound on the rate. Our approach is to consider the evolution of the n-volume of the parallelepiped generated by the vectors as

the orthogonalizations are performed. This volume is simply $\det(|A|)$ where $|A| = (A^*A)^{1/2}$. We show the n-volume is monotone even for adversarially chosen pairs of columns at each iteration, and logarithm of the n-volume increases non-trivially in expectation when the pair of columns is selected randomly.

This procedure bears some resemblance to the Kaczmarz method for solving linear systems. If one is solving $A^*x = 0$ using Kaczmarz, a row of A^* is sampled and x is projected to be orthogonal to the sampled row; this is then repeated many times. In our case, x is itself one of the columns of A, say $x = a_k$ and one applies one step of Kaczmarz to the system $(A^*x)_j = 0$ for $j \neq k$ (and then renormalizes).

We have two main results concerning $\det(|A_t|)$ where A_t is the matrix obtained by applying t iterations to A. Theorem 2.1.1 provides a bound on $\mathbb{E}\log\det(|A_t|)$, and Theorem 2.1.2 gives a tail bound for values of $\det(|A_t|)$ close to 1. The upshot is that $t = \Theta(n^2\log(1/(\det(|A|)\varepsilon)))$ iterations suffice to produce $\det(|A_t|) \geq 1-\varepsilon$ with constant probability. This in turn implies that $\kappa(A_t) \leq 1+\sqrt{\varepsilon}$, and further that there exists an orthonormal basis $B = \begin{bmatrix} b_1 & \cdots & b_n \end{bmatrix}$ for the span of the columns of A, denoted $\mathrm{Col}(A)$, with $\|A_t - B\|_F \leqslant \sqrt{\varepsilon}$ (see Corollary 2.1.3, Corollary 2.1.4).

Orthogonalization procedure

We first precisely define the procedure we analyze. Throughout this paper, let $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{C}^{d\times n}, d \geq n$ be the decomposition of A into columns. We assume the columns are unit length and linearly independent. We define the operation

$$\begin{aligned} &\texttt{orth} : \mathbb{C}^{d\times n} \times \left\{(i,j) \in [n]^2 : i \neq j\right\} \to \mathbb{C}^{d\times n} \\ &k\text{th column of } \texttt{orth}(A,i,j) = \begin{cases} a_k & k \neq i \\ \dfrac{a_i - \langle a_i, a_j\rangle a_j}{\|a_i - \langle a_i, a_j\rangle a_j\|} & k = i. \end{cases} \end{aligned} \tag{2.1}$$

At each timestep t, our procedure samples (i_t, j_t) uniformly at random from ordered pairs of distinct indices and updates $A_{t+1} \leftarrow \texttt{orth}(A_t, i_t, j_t)$. We study the evolution of the nonnegative potential function

$$\Phi(A) = -\log\det(|A|) \quad \text{where} \quad |A| = (A^*A)^{1/2}. \tag{2.2}$$

Note that since A has unit-length columns, $\Phi(A) = 0$ if and only if the columns of A are orthogonal. We show $\Phi(A_t)$ is monotonically decreasing no matter the selection of (i_t, j_t) (see Lemma 2.2.1). Over a random selection of (i_t, j_t), it strictly decreases in expectation unless $\Phi(A_t) = 0$ already (see Lemma 2.2.3). Our bound roughly shows that when $\Phi(A_t) \geq 1$, one should expect $\Phi(A_{t+1}) \leqslant \Phi(A_t) - O(1/n^2)$, i.e. steady additive progress is made, and when $\Phi(A_t) \leqslant 1$, one should expect $\Phi(A_{t+1}) \leqslant (1-1/n^2)\Phi(A_t)$, i.e. steady multiplicative progress is made.

Main results

Our main results are the following theorems about the expectation of $\Phi(A_t)$ and a tail bound on $\Phi(A_t)$. The first theorem shows $\Phi(A_t)$ converges to 0 almost surely, and the second gives a tail bound for values of $\det(|A_t|)$ close to 1.

Theorem 2.1.1 (Expectation bound on $\Phi(A_t)$)**.**

$$\mathbb{E}\Phi(A_t) \leqslant \exp\left(-\frac{t}{(n-1)^2(2\Phi(A)+1)}\right)\Phi(A).$$

Theorem 2.1.2 (Tail bound on $\Phi(A_t)$)**.** *Let $\varepsilon, \delta \in (0,1)$ and fix any A. If*

$$t \geq \Theta\left(n^2 \log\left(\frac{1}{(\det(|A|)\varepsilon)}\right)\log\left(\frac{1}{\delta}\right)\right)$$

then

$$\Pr(\Phi(A_t) \geq \varepsilon) \leqslant \delta.$$

We have the following corollaries, which restate Theorem 2.1.2 in terms of more traditional measures of orthogonality.

Corollary 2.1.3 (Convergence of condition number)**.** *Let $\varepsilon, \delta \in (0,1)$. Fix any A. For t as in Theorem 2.1.2,*

$$\Pr\big(\kappa(A_t) \geq 1 + \sqrt{\varepsilon}\big) \leqslant \delta.$$

Proof. This follows from Theorem 2.1.2 and Fact 2.3.1. □

Corollary 2.1.4 (Distance to unitary)**.** *Let $\varepsilon, \delta \in (0,1)$, and fix any A. For t as in Theorem Theorem 2.1.2, with probability at least $1-\delta$ there exists an orthonormal basis $B = [b_1 \cdots b_n]$ for* $\mathrm{Col}(A)$ *such that*

$$\|A_t - B\|_F \leqslant \sqrt{\varepsilon}.$$

Proof. This follows from Theorem 2.1.2 and Fact 2.3.2. □

The remainder of this chapter is organized as follows: we finish this section by introducing some notation and preliminaries; in Section 2.2, we consider the effect of a single step of this orthogonalization procedure; in Section 2.3 we then consider the effect of many steps, allowing us to prove Theorem 2.1.1 and Theorem 2.1.2; in Section 2.3, we connect $\Phi(\cdot)$ to other measures of orthogonality; and in Section 2.4, we discuss algorithmic applications of this procedure.

Notation and preliminaries

Throughout this chapter, $A, A', A_t \in \mathbb{C}^{d\times n}$ all denote matrices with unit length and linearly independent columns. We will consider several optimization problems throughout the analysis, and the following lemma will be convenient to have:

Lemma 2.1.5. *If the program*

$$x_j \geq 0, \qquad \sum_{j\in[n]} x_j = S, \qquad \prod_{j\in[n]} x_j = P$$

is feasible, then the minimizer of the objective

$$\sum_{j\in[n]} x_j^2$$

occurs for

$$x_{\pi(1)} \leqslant x_{\pi(2)} = \cdots = x_{\pi(n)}$$

for some permutation π.

Proof. Let $y_1, \ldots, y_n$ be a minimizer of the program and renumber the variables so that $y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n$ without loss of generality. If one fixes $x_j = y_j$ for $j = 3, \ldots, n-1$, then the triplet (y_1, y_2, y_n) should minimize the restricted program

$$\min x_1^2 + x_2^2 + x_n^2$$

subject to

$$x_1, x_2, x_n \geq 0, \qquad x_1 + x_2 + x_n = y_1 + y_2 + y_n =: S', \qquad x_1x_2x_n = y_1y_2y_n =: P'.$$

Since the constraints and objective are homogeneous, we may assume $S' = 3$ without loss of generality. Using Lagrange multipliers, one can see that the critical values occur when the set $\{x_1, x_2, x_n\}$ has only two distinct values. Let a and b be the value of the majority and minority element respectively, so that the constraints read

$$2a + b = 3, \qquad a^2b = P$$

and the objective is

$$\min g(a, b) = 2a^2 + b^2.$$

Combining the constraints shows that the only feasible values of a are the positive roots of

$$f(a) = 2a^3 - 3a^2 + P' = 2(a-1)^3 + 3(a-1)^2 + (P'-1).$$

Then one has

$$g(a, 3-2a) = 3 + 6(a-1)^2 = 3 + 2\cdot\left(-(P'-1) - 2(a-1)^3\right)$$

at those roots. Since $g(a, 3-2a)$ is monotonically decreasing in a, the minimizer occurs when a is the larger root of f. Since the minimum of $f(a)$ for $a \geq 0$ occurs at $a = 1$, this forces $b \leqslant 1 \leqslant a$. In particular, we must have $b = y_1 \leqslant y_2 = \cdots = y_n = a$ as required. □

2.2 Analysis of a single iteration

We first claim $\det(|A|)$ is deterministically monotone in applications of `orth`, irrespective of which columns are orthogonalized.

Lemma 2.2.1. *For any A and distinct indices i, j, one has*

$$\det(|A'|) = \frac{\det(|A|)}{\sqrt{1 - |\langle a_i, a_j \rangle|^2}}$$

where $A' = \texttt{orth}(A, i, j)$.

Proof. A' can be expressed as the composition of two elementary column operations: first, E_1 replaces a_i with the component $a_i - \langle a_i, a_j \rangle a_j$ so $\det(E_1) = 1$; second, E_2 renormalizes the ith column to be a unit vector, so

$$\det(E_2) = \frac{1}{\|a_i - \langle a_i, a_j \rangle a_j\|} = \frac{1}{\sqrt{1 - |\langle a_i, a_j \rangle|^2}}.$$

Then $A' = AE_1E_2$ so

$$\det(|A'|) = \det(E_2^* E_1^* A^* A E_1 E_2)^{1/2} = |\det(E_2)| \det(|A|)$$

as required. □

From the above lemma, we see that the change in the determinant depends only on $|\langle a_i, a_j \rangle|^2$. The next lemma relates the average value of this quantity (over a random selection of distinct i, j) to the determinant.

Lemma 2.2.2. *For any A,*

$$\frac{1}{n(n-1)} \sum_{i \neq j} |\langle a_i, a_j \rangle|^2 \geq \frac{1}{(n-1)^2}(1 - \det(|A|)^2).$$

Proof. Note that the left hand side is simply $\frac{1}{n(n-1)} \|A^*A - I\|_F^2$. Also note that $\|A\|_F^2 = n$ since A has unit length columns. Consider the optimization program $\min \|A^*A - I\|_F^2$ subject to $\|A\|_F^2 = n$ and $\det(|A|)^2 = P$. Using the $\|A\|_F^2 = n$ constraint, the objective can be rewritten as

$$\|A^*A - I\|_F^2 = \operatorname{tr}(A^*AA^*A) - 2\operatorname{tr}(A^*A) + n = \|A^*A\|_F^2 - n.$$

Both the constraints and objective function of this program depend only on the singular values of A, so we may rewrite the program in terms of them. Let $x_1, \ldots, x_n$ be the squares of the singular values of A so that the program becomes

$$\min \sum_{j \in [n]} x_j^2 - n$$

subject to

$$x_j \geq 0, \qquad \sum_{j\in[n]} x_j = n, \qquad \prod_{j\in[n]} x_j = P.$$

By Lemma 2.1.5, we may set $a = x_2 = \cdots = x_n \geq 1$ and $b = x_1 \leqslant 1$. The resulting program of a, b is

$$\min g(a,b) = (n-1)a^2 + b^2 - n$$

subject to

$$0 \leqslant b \leqslant 1 \leqslant a, \qquad (n-1)a + b = n, \qquad a^{n-1}b = P. \tag{2.3}$$

The equality constraints together imply that the feasible values of a must be solutions to

$$f(a) = a^n - \frac{n}{n-1}a^{n-1} = -\frac{P}{n-1}.$$

For every $P \in [0,1]$, there is only one solution to the right of 1, denote it $a(P)$. Since $f(\cdot)$ is monotone to the right of 1, $a(\cdot)$ is monotone as well with $a(1) = 1$ and $a(0) = \frac{n}{n-1}$. Let

$$\widetilde{f}(a) = (n-1)(a-1)^2 - \frac{1}{n-1}$$

and denote the largest solution to $\widetilde{f}(a) = -\frac{P}{n-1}$ by $\widetilde{a}(P)$. Observe that $f(a) \leqslant \widetilde{f}(a)$ on the interval $[1, n/(n-1)]$, so $\widetilde{a}(P) \leqslant a(P)$. The value of the objective function for $(n-1)a+b = n$ is

$$\begin{aligned} g(a,b) &= -n + (n-1)a^2 + b^2 \\ &= -n + (n-1)a^2 + (n-(n-1)a)^2 \\ &= n(n-1)(a-1)^2. \end{aligned} \tag{2.4}$$

In particular, this expression is monotone for $a \geq 1$. Thus the minimum value of g subject to the constraints (3.12) is

$$\begin{aligned} \min g(a,b) &= g(a(P), n-(n-1)a(P)) \\ &= n(n-1)(a(P)-1)^2 \\ &\geq n(n-1)(\widetilde{a}(P)-1)^2 \\ &\geq \frac{n}{n-1}(1-P) \end{aligned} \tag{2.5}$$

as required, where the last step used that $\widetilde{f}(\widetilde{a}(P)) = -P/(n-1)$. □

With these two lemmas in place, we can compute the expected value of

$$\Phi(\cdot) = -\log\det(|\cdot|)$$

in terms of itself after one application of `orth` for random indices. Define the iterative map

$$f(x) = x - \frac{1 - \exp(-2x)}{2(n-1)^2}. \tag{2.6}$$

When x is small, one should think of f as roughly $f(x) \approx (1 - O(1/n^2))x$. For larger x, one should think of f as roughly $f(x) \approx x - O(1/n^2)$.

Lemma 2.2.3 (One step estimate)**.** *Let (i, j) be a uniform sample from $\{(i, j) \in [n]^2 : i \neq j\}$. Then*

$$\Phi(A') \leqslant \Phi(A) \text{ pointwise} \qquad \text{and} \qquad \mathbb{E}\Phi(A') \leqslant f(\Phi(A))$$

where $A' = \texttt{orth}(A, i, j)$.

Proof. Lemma 2.2.1 states

$$\Phi(A') = \Phi(A) + \frac{1}{2}\log\bigl(1 - |\langle a_i, a_j\rangle|^2\bigr). \tag{2.7}$$

Since $|\langle a_i, a_j\rangle|^2 \geq 0$, the first claim follows immediately. By Jensen's inequality, when i, j are picked randomly one has

$$\begin{aligned}
\mathbb{E}\Phi(A') &\leqslant \Phi(A) + \frac{1}{2}\log\bigl(1 - \mathbb{E}\,|\langle a_i, a_j\rangle|^2\bigr) \\
&\leqslant \Phi(A) - \frac{1}{2}\mathbb{E}\,|\langle a_i, a_j\rangle|^2 \\
&= \Phi(A) - \frac{\|A^*A - I\|_F^2}{2n(n-1)}.
\end{aligned} \tag{2.8}$$

Now apply Lemma 2.2.2,

$$\begin{aligned}
\mathbb{E}\Phi(A') &\leqslant \Phi(A) - \frac{1 - \det(|A|)^2}{2(n-1)^2} \\
&= \Phi(A) - \frac{1 - \exp(-2\Phi(A))}{2(n-1)^2} \\
&= f(\Phi(A)).
\end{aligned} \tag{2.9}$$

□

Note Lemma 2.2.3 holds if (i, j) is instead a uniform sample from either $\{(i, j) \in [n]^2 : i > j\}$ or $\{(i, j) \in [n]^2 : i < j\}$. This is significant if one wishes to use this procedure to compute the QR decomposition of a matrix A. See Section 3.2 for more details.

2.3 Analysis of many iterations

The previous section describes the effect of one application of `orth` to the collection of vectors. This section considers the effect of several iterations. In particular, we're interested in two dual questions: what does this collection of vectors look like after a given number of iterations? How many iterations are required to achieve a particular state?

For this section, define the family of stochastic processes parameterized by the starting matrix A,

$$A_0 = A, \qquad A_{t+1} = \texttt{orth}(A_t, i_t, j_t) \tag{2.10}$$

where i_t, j_t are sampled uniformly at random (subject to $i_t \neq j_t$) at each step. We can define the corresponding hitting time of achieving a small potential value when starting with a matrix A,

$$\tau_\beta^{(A)} = \inf\{t \in \mathbb{Z}_{\geq 0} : \Phi(A_t) < \beta\}.$$

Further define the largest expected hitting time among starting matrices with potential bounded by α,

$$\mu_{\alpha\to\beta} = \sup_{X:\Phi(X)\leqslant\alpha} \mathbb{E}\tau_\beta^{(X)}. \tag{2.11}$$

This is the most expected amount of time this process takes to transform a matrix with potential at most α to a matrix with potential less than β.

Because applying f does not commute with taking expectations, obtaining a bound for several applications of `orth` by iteratively applying Lemma 2.2.3 is not immediate. Fortunately, f can be bounded by a linear function an adequate slope on bounded domains; this observation results in the following expectation bound.

Theorem 2.1.1 (Expectation bound on $\Phi(A_t)$)**.**

$$\mathbb{E}\Phi(A_t) \leqslant \exp\left(-\frac{t}{(n-1)^2(2\Phi(A)+1)}\right)\Phi(A).$$

Proof. Denote $\phi_t = \Phi(A_t)$. The second part of Lemma 2.2.3 implies

$$\mathbb{E}\phi_{t+1} \leqslant \mathbb{E}f(\phi_t)$$

The first part of Lemma 2.2.3 implies $\phi_t \leqslant \phi_0$ for all t with probability 1. Since f is convex, we have for $\phi \leqslant \phi_0$ that

$$f(\phi) \leqslant \frac{\phi}{\phi_0} \cdot f(\phi_0).$$

In particular, this inductively implies

$$\mathbb{E}\phi_{t+1} \leqslant \frac{f(\phi_0)}{\phi_0}\mathbb{E}\phi_t \leqslant \left(\frac{f(\phi_0)}{\phi_0}\right)^{t+1}\phi_0. \tag{2.12}$$

Finally, one may upper bound f via

$$\begin{aligned} f(\phi_0) &= \phi_0 - \frac{1 - \exp(2\phi_0)^{-1}}{2(n-1)^2} \\ &\leqslant \phi_0 - \frac{1 - (1+2\phi_0)^{-1}}{2(n-1)^2} \\ &= \phi_0 \left(1 - \frac{1}{(n-1)^2(2\phi_0+1)}\right) \\ &\leqslant \phi_0 \exp\left(-\frac{1}{(n-1)^2(2\phi_0+1)}\right). \end{aligned} \tag{2.13}$$

Combined with (2.12), this is exactly the desired result. □

This establishes control of the expectation of $\Phi(A_t)$. The next lemma establishes a tail bound in terms of the $\mu_{\alpha\to\beta}$ defined in (2.11).

Lemma 2.2.4 (Tail bound in terms of $\mu_{\alpha\to\beta}$)**.** *Fix any $\alpha > \beta > 0$. For every A with $\Phi(A) = \alpha$,*

$$\Pr(\Phi(A_t) \geq \beta) = \Pr(\tau_\beta^{(A)} > t) \leqslant \exp\left(-\left\lfloor \frac{t}{e\mu_{\alpha\to\beta}} \right\rfloor\right).$$

Proof. Conditioned on the event $\tau_\beta^{(A)} > k$, we have by definition,

$$\tau_\beta^{(A)} = k + \tau_\beta^{(A_k)}.$$

Since $\Phi(A_t)$ is pointwise monotone, we have that $\Phi(A_k) \leqslant \Phi(A) = \alpha$. This yields

$$\begin{aligned} \mathbb{E}(\tau_\beta^{(A)} \,|\, \tau_\beta^{(A)} > k) &= \mathbb{E}(\mathbb{E}(k + \tau_\beta^{(A_k)} \,|\, \tau_\beta^{(A)} > k, A_k)) \\ &= k + \mathbb{E}(\mathbb{E}(\tau_\beta^{(A_k)} \,|\, \tau_\beta^{(A)} > k, A_k)) \\ &\leqslant k + \sup_{X:\Phi(X)\leqslant\alpha} \mathbb{E}(\tau_\beta^{(A_k)} \,|\, \tau_\beta^{(A)} > k, A_k = X) \\ &\leqslant k + \sup_{X:\Phi(X)\leqslant\alpha} \mathbb{E}(\tau_\beta^{(X)}) \\ &= k + \mu_{\alpha\to\beta}. \end{aligned} \tag{2.14}$$

This allows us to compute a bound on the conditional tails of $\tau_\beta^{(A)}$ using Markov's inequality. For c which we specify later,

$$\begin{aligned} \Pr\left(\tau_\beta^{(A)} > k + c \,|\, \tau_\beta^{(A)} > k\right) &= \Pr\left(\tau_\beta^{(A)} - k > c \,|\, \tau_\beta^{(A)} > k\right) \\ &\leqslant \frac{\mathbb{E}(\tau_\beta^{(A)} - k \,|\, \tau_\beta^{(A)} > k)}{c} \\ &\leqslant \frac{\mu_{\alpha\to\beta}}{c}. \end{aligned} \tag{2.15}$$

Set $J = \lfloor t/c \rfloor$ and apply (2.15) for $k = 0, c, 2c, \dots (J-1)c$ to obtain

$$\Pr\Big(\tau_\beta^{(A)} > t\Big) \leqslant \Pr\Big(\tau_\beta^{(A)} > Jc\Big) = \prod_{j=1}^{J} \Pr\Big(\tau_\beta^{(A)} > jc \,|\, \tau_\beta^{(A)} > (j-1)c\Big) \leqslant (\mu_{\alpha\to\beta}/c)^J. \quad (2.16)$$

Pick $c = e\mu_{\alpha\to\beta}$ for the final result. □

To turn this into a concrete tail bound, we establish an estimate for $\mu_{\alpha\to\beta}$.

Lemma 2.2.5 (Expectation bound on $\mu_{\alpha\to\beta}$)**.** *For $\alpha > 1 > \beta > 0$,*

$$\mu_{\alpha\to\beta} \leqslant \frac{2(n-1)^2}{1-e^{-2}}(\alpha + e\lceil \log(1/\beta)\rceil).$$

For $1 > \alpha > \beta > 0$,

$$\mu_{\alpha\to\beta} \leqslant \frac{2e(n-1)^2}{1-e^{-2}}\lceil \log(\alpha/\beta)\rceil.$$

Proof. We show this in two steps: first we establish a crude upper bound, then we exploit the sub-additive nature of μ to boost this to a stronger bound. For the first step, we claim that for $\alpha > \beta > 0$, we have

$$\mu_{\alpha\to\beta} \leqslant \frac{2(n-1)^2\alpha}{1-\exp(-2\beta)}. \quad (2.17)$$

To see this, fix any A with $\Phi(A) \leqslant \alpha$. Denote $\phi_t = \Phi(A_t), \tau_\beta = \tau_\beta^{(A)}$ for brevity. Note that $f(x) - x$ is negative and monotonically decreasing, so

$$f(x) - x \leqslant \begin{cases} f(\beta) - \beta & x \geq \beta \\ 0 & x < \beta \end{cases}. \quad (2.18)$$

Using Lemma 2.2.3, this gives for each t the bound on the increments,

$$\begin{aligned} \mathbb{E}(\phi_{t+1} - \phi_t) \leqslant \mathbb{E}(f(\phi_t) - \phi_t) &\leqslant (f(\beta) - \beta) \cdot \Pr(\phi_t \geq \beta) \\ &= (f(\beta) - \beta) \cdot \Pr(\tau_\beta > t). \end{aligned} \quad (2.19)$$

This becomes a telescoping sum, so we obtain for all t that

$$\mathbb{E}(\phi_t) - \phi_0 \leqslant (f(\beta) - \beta) \sum_{j=0}^{t-1} \Pr(\tau_\beta > j).$$

In the limit as $t \to \infty$, the right hand side is simply $(f(\beta) - \beta)\mathbb{E}(\tau_\beta)$, and by Theorem 2.1.1 the left hand side is $-\phi_0$. By construction, $\phi_0 \leqslant \alpha$ so rearranging gives $\mathbb{E}(\tau_\beta) \leqslant \alpha/(\beta - f(\beta))$. Taking the supremum over all A with $\Phi(A) \leqslant \alpha$ gives (2.17).

Now we use (2.17) to prove the stated bounds. Consider any sequence $\alpha = \alpha_0 > \alpha_1 > \cdots > \alpha_\ell$ ending at $\alpha_\ell \leqslant \beta$. Set $k_0 = 0$, $k_j = \tau_{\alpha_j}^{(A_{k_{j-1}})}$. Then by definition,

$$\tau_\beta^{(A)} \leqslant \sum_{j=1}^{\ell} k_j.$$

Note also by definition that $\Phi(A_{k_{j-1}}) < \alpha_{j-1}$. Thus taking expectations and applying (2.17) gives

$$\mathbb{E}\tau_\beta^{(A)} \leqslant \sum_{j=1}^{\ell} \mathbb{E}k_j \leqslant \sum_{j=1}^{\ell} \mu_{\alpha_{j-1}\to\alpha_j} \leqslant \sum_{j=1}^{\ell} \frac{2(n-1)^2\alpha_{j-1}}{1-\exp(-2\alpha_j)}.$$

When $\alpha > 1 > \beta$, pick $\alpha_j = e^{1-j}$ for $j \geq 1$. Further use the approximation $1 - \exp(-2x) \geq (1-e^{-2})x$ when $x \leqslant 1$. Then the sum becomes

$$\begin{aligned}\frac{2(n-1)^2\alpha}{1-e^{-2}} + 2(n-1)^2 \sum_{j=2}^{\ell} \frac{e^{2-j}}{1-\exp(-2\cdot e^{1-j})} &\leqslant \frac{2(n-1)^2\alpha}{1-e^{-2}} + \frac{2(n-1)^2}{1-e^{-2}} \sum_{j=2}^{\ell} \frac{e^{2-j}}{e^{1-j}} \\ &= \frac{2(n-1)^2}{1-e^{-2}}(\alpha + e(\ell-1)).\end{aligned} \tag{2.20}$$

It suffices to take $\ell = \lceil \log(1/\beta) \rceil + 1$ giving the first result. If $1 \geq \alpha > \beta$, pick $\alpha_j = \alpha_0 e^{-j}$. Then the sum becomes

$$\sum_{j=1}^{\ell} \frac{2(n-1)^2\alpha_0 e^{1-j}}{1-\exp(-2\alpha_0 e^{-j})} \leqslant \sum_{j=1}^{\ell} \frac{2(n-1)^2\alpha_0 e^{1-j}}{(1-e^{-2})\alpha_0 e^{-j}} = \frac{2e(n-1)^2}{(1-e^{-2})}\ell.$$

It suffices to take $\ell = \lceil \log(\alpha/\beta) \rceil$. □

Theorem 2.1.2 (Tail bound on $\Phi(A_t)$)**.** *Fix any A and $\beta < 1$. Then*

$$\Pr(\Phi(A_t) \geq \beta) \leqslant \exp\left(-\left\lfloor \frac{1-e^{-2}}{2e} \frac{t}{(n-1)^2(\Phi(A) + e\lceil \log(1/\beta) \rceil)} \right\rfloor\right).$$

In particular for any $\delta > 0$,

$$t \geq \frac{2e^3}{e^2-1}(n-1)^2(\Phi(A) + e\lceil 2\log(4/\beta) \rceil)\lceil \log(1/\delta) \rceil = \Theta\left(n^2 \log\left(\frac{1}{|\det(A)\beta|}\right) \log\left(\frac{1}{\delta}\right)\right)$$

implies

$$\Pr(\Phi(A_t) \geq \beta) \leqslant \delta.$$

Proof. This is an immediate consequence of Lemma 2.2.4 and Lemma 2.2.5. □

2.3 Other measures of orthogonality

The previous section produced results concerning $\Phi(A_t) = -\log\det(|A_t|)$. Here we show convergence of $\Phi(\cdot)$ implies convergence of a couple other natural measures of orthogonality.

Fact 2.3.1 (Condition number)**.**

$$\Phi(A) \leqslant \beta \implies \log\kappa(A) \leqslant \min\left(\log(2)+\beta, \sqrt{2\beta}+\frac{\beta^{1.5}}{4}\right).$$

Proof. By applying Lagrange multipliers to the optimization problem

$$\max\kappa(A) = \sigma_1/\sigma_n$$

subject to

$$\|A\|_F^2 = \sigma_1^2+\cdots+\sigma_n^2 = n \qquad \text{and} \qquad \sigma_1\cdots\sigma_n = \det(|A|),$$

one finds that the optimum is achieved for

$$\sigma_1^2 = 1+\sqrt{1-\det(|A|)^2}, \qquad \sigma_2 = \cdots = \sigma_{n-1} = 1, \qquad \sigma_n^2 = 1-\sqrt{1-\det(|A|)^2}.$$

This gives a maximum value of

$$\max\kappa(A) = \sqrt{\frac{1+\sqrt{1-\det(|A|)^2}}{1-\sqrt{1-\det(|A|)^2}}} \leqslant \sqrt{\frac{1+\sqrt{1-\exp(-2\beta)}}{1-\sqrt{1-\exp(-2\beta)}}}$$

the log of which is upper bounded both by $\log(2)+\beta$ and $\sqrt{2\beta}+\beta^{1.5}/4$. □

Fact 2.3.2 (Distance to unitary)**.**

$$\Phi(A) \leqslant \beta \implies \inf\{\|A-B\|_F : B^*B = I, \mathrm{Col}(A) = \mathrm{Col}(B)\} \leqslant \sqrt{2\beta}. \tag{2.21}$$

Proof. By considering the Gram-Schmidt basis, one may assume without loss of generality that A is upper triangular with positive real diagonal entries. Let $1 = \lambda_1, \cdots, \lambda_n$ be those entries. Note by the Cauchy-Binet formula that $\det(A^*A) = (\lambda_1\cdots\lambda_n)^2$. Let $r_j \in \mathbb{C}^{j-1}$ be the portion of the jth column of A strictly above the diagonal. Take $B = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix}$. Then

$$\|A-B\|_F^2 = \sum_{j=2}^{n}\big((\lambda_j-1)^2+\|r_j\|^2\big) = \sum_{j=2}^{n}\big(1-2\lambda_j+\lambda_j^2+\|r_j\|^2\big). \tag{2.22}$$

Note $\lambda_j^2+\|r_j\|^2 = 1$ since the columns of A are unit length. Thus

$$\|A-B\|_F^2 = 2\sum_{j=1}^{n}(1-\lambda_j) \leqslant 2\sum_{j=1}^{n}\log(1/\lambda_j) = -2\log(\lambda_1\cdots\lambda_n) = 2\Phi(A). \tag{2.23}$$

□

2.4 Algorithmic considerations

The QR decomposition

The update $A_{t+1} = \texttt{orth}(A_t, i_t, j_t)$ can be expressed as

$$A_{t+1} = A_t C_t$$

where C_t is the appropriate column operation, taking just $O(d)$ time to perform. The corresponding decomposition of A as

$$A = A_\tau C \quad \text{where} \quad C = C_{\tau-1}^{-1} C_{\tau-2}^{-1} \cdots C_0^{-1}$$

is analogous to an approximate QR decomposition: the columns of A_τ form a well-conditioned basis for $\mathrm{Col}(A)$ when τ is sufficiently large. By Theorem 2.1.2, $\tau = O(n^2 \log(1/\det(|A|)))$ iterations suffice to achieve a constant condition number with constant probability; when $\det(|A|) = \mathrm{poly}(n)$, this gives a total arithmetic cost of $\widetilde{O}(dn^2)$, comparable to the arithmetic cost of Gram-Schmidt.

If one requires the decomposition $A = A_\tau C$ have C be upper triangular, one can simply sample the update pairs (i_t, j_t) subject to $i_t > j_t$. This does not change any of the analysis as $\mathbb{E}\,|\langle a_i, a_j\rangle|^2$ is the same. In this case, $A_\tau C$ would in fact be converging to the actual $A = QR$ decomposition.

There are a couple possible advantages to this method over Gram-Schmidt. First, one can end the process early when A_τ has a "reasonable" condition number depending on one's requirements; in particular, it could make a good preconditioner. The memory access pattern is very similar to that of the Kaczmarz method for solving the system $A^* x = b$; one could imagine updating the rows of A^* (and corresponding entries of b) to improve their conditioning at the same time the Kaczmarz solver runs. The idea is that the condition number gets smaller over time, so perhaps the increase in the rate of convergence of Kaczmarz justifies the additional computational expense. For a characterization of the rate of convergence of Kaczmarz in terms of the condition number, see [SV09]. A second possible advantage is that it's highly parallelizable. Below we suggest three concrete ways to perform up to $O(n)$ applications of `orth` in parallel.

Extension to allow parallelism

The process is currently described in terms of modifying a single column at a time. However, we can batch several of these updates together, computing the changes to multiple columns in parallel. Each batch of updates corresponds to a directed, rooted forest on the vertex set $[n]$, where edges (i, j) are oriented so that i is further from the root. Then for each edge (i, j), set the new value of a_i to be

$$\frac{a_i - \langle a_i, a_j\rangle\, a_j}{\|a_i - \langle a_i, a_j\rangle\, a_j\|}.$$

The orientation of the edges ensures that this is well defined. By listing the edges of the tree in order of furthest to the roots to the closest, say $(i_0, j_0), \cdots, (i_{m-1}, j_{m-1})$, one can express the effect of a batch update as A_m where

$$A = A_0, \qquad A_{t+1} = \texttt{orth}(A_t, i_t, j_t) \quad \forall t \in [m-1].$$

Note that once a column is changed, it is not used or changed again. Thus all of the updates to each A_t can be expressed in terms of the columns of the inital A_0. In particular, Lemma 2.2.1 simply implies

$$\Phi(A_t) = \Phi(A) + \frac{1}{2}\sum_{\ell=1}^{t} \log(1 - |\langle a_{i_\ell}, a_{j_\ell}\rangle|^2) \tag{2.24}$$

analogously to (2.7). If the forest is sampled so that the marginal distribution of each edge pair (i_t, j_t) is uniform, then following the remainder of the proof of Lemma 2.2.3 gives the generalization

$$\Phi(A_t) \leqslant \Phi(A) \text{ pointwise} \qquad \text{and} \qquad \mathbb{E}\Phi(A_t) \leqslant f_t(\Phi(A)) \tag{2.25}$$

where

$$f_t(x) = x - t \cdot \frac{1 - \exp(-2x)}{2n(n-1)}.$$

One can then define the process by sampling a sequence of independent forests and concatenating the lists of the edges to form the sequence $\{(i_t, j_t) : t \in \mathbb{Z}_{\geq 0}\}$. Replacing the result of Lemma 2.2.3 with (2.25) allows the subsequent proofs to go through.

There are many choices of how to sample the required forests. We end this section by suggesting four natural distributions.

1. Sample a uniformly random forest consisting of a single edge.
2. Sample a uniformly random perfect matching of the complete graph and flip a coin to orient each edge.
3. Sample a uniformly random permutation $\pi \colon [n] \to [n]$ and consider the path with edges $(i_t, j_t) = (\pi(t), \pi(t+1))$ for $t \in [n-1]$.
4. Sample a uniformly random vertex and consider the star graph rooted and centered at that selected vertex.

Chapter 3

Matrix factorizations

In this chapter, we revisit our stochastic orthogonalization scheme with one important difference: at each step, we now allow *any method* to orthogonalize v_1 and v_2 to be applied, so long as it preserves $\text{span}(v_1, v_2)$. In particular, we do not require that our vectors be renormalized as unit vectors.

Immediately, we notice that the analysis in the previous chapter no longer applies: from Lemma 2.2.1, we see the driving force of the convergence of $\Phi(A)$ was precisely the renormalization of the orthogonalized vectors. In fact, our original approach of considering the eigenvalues of A^TA is fundamentally flawed: since we are no longer guaranteed unit vectors, the goal is not to show A^TA is the identity, but rather that it is *diagonal*.

This reveals two things to us: first, when an appropriate procedure for the orthogonalization step is chosen, this orthogonalization scheme can be viewed as a *diagonalization* scheme. We discover that by choosing different orthogonalization rules, we can use this procedure to find different matrix factorizations. The "Gram-Schmidt" orthogonalization rule of the previous section gives us an iterative algorithm to find the QR factorization of A (or similarly the Cholesky decomposition of A^TA). Using Givens rotations to orthogonalize two columns of A, we can find the eigendecomposition of A^TA (or similarly the singular value decomposition of A). We recognize the second procedure as the well-known *Jacobi eigenvalue algorithm*. Second, we need a potential function that can measure proximity to a diagonal matrix. This potential function needs to be general enough to capture progress for any orthogonalization scheme, and also needs to be well-behaved under any orthogonalization scheme. Section 3.3 is devoted to defining and exploring our new potential function. We note that this potential function also allows us to achieve an improved rate of convergence compared to Chapter 2.

Overall, our contribution is a unified analysis of a class of iterative algorithms to compute many matrix factorizations under a randomized pivoting strategy (this corresponds to choosing two vectors at random to orthogonalize). Many of the same components we saw in the last chapter will reappear here: analysis of the potential function after one iteration, then after many iterations using martingale analysis, and the use of Lagrange multipliers to connect our potential function to other metrics.

3.1 Introduction

At first glance, the Jacobi eigenvalue algorithm for computing the eigendecomposition of Hermitian matrices and the Gram-Schmidt procedure for computing the QR decomposition of tall matrices are quite distinct: they take different inputs, compute different factorizations, have different provable asymptotic complexities, etc. However, there is an important commonality: both methods are iterative and consider only pairs of columns of their inputs at a time. In this work, this connection is taken further, and both algorithms are realized as special cases of a more general matrix factorization algorithm.

Our primary result, Theorem 3.4.4, is a unified analysis of all special cases of this general procedure when using a particular randomized pivoting rule; we show that these algorithms are all linearly convergent in expectation, and further that the rate of convergence is unaffected by which particular factorization is being computed.

In finite arithmetic, we provide a framework for showing that instances of this general procedure are numerically stable. This framework is applied to several orthogonalization methods (Theorem 3.5.7), as well as the Jacobi eigenvalue algorithm (Theorem 3.5.8, Theorem 3.5.9, Theorem 3.5.10), building on the work of [DV92] in the latter case.

We first give a brief overview of the history and context of these algorithms. As these algorithms have been extensively studied, we cannot give a comprehensive overview. We instead highlight their origins, practical relevance today, and interactions between them.

Eigendecomposition / Singular value decomposition: The classical Jacobi eigenvalue algorithm (which we henceforth refer to as the two-sided Jacobi eigenvalue algorithm[1]) is an iterative algorithm for computing an approximate eigendecomposition of a given Hermitian[2] matrix that dates back to 1846 [Jac46]. It was rediscovered and popularized in the 1940s and 50s when the advent of computers made large numerical computations feasible (we refer the reader to [GvdV00] for a detailed history). Today, it is included in popular textbooks [Dem97; Par98] and software packages [And+99]. Although it is often less preferred due to a slower runtime [Dem97], it has several strengths when compared with other popular methods for diagonalization. First, if the given matrix B is already close to diagonal, the arithmetic cost is greatly reduced [Sch64]. Second, the computation is highly parallelizable, and each thread only needs a small amount of memory (see, for example, [BS89]). Third, numerical experiments reveal classes of matrices for which all the eigenvalues appear to be computed with small relative error [DV92].

The corresponding algorithm for computing the singular value decomposition of a given matrix $A \in \mathbb{C}^{d\times n}$, today called the one-sided Jacobi eigenvalue algorithm, was introduced as an independent algorithm by Hestenes [Hes58], then rediscovered and connected to the

[1]The literature also uses "two-sided" to refer to a version of the Jacobi eigenvalue algorithm due to [Kog55] for computing the singular value decomposition; that method is different from the version of the Jacobi eigenvalue for computing the singular value decomposition that we consider here.

[2]One can replace $\mathbb{C}$ with $\mathbb{R}$ throughout this chapter. Hermitian becomes symmetric, (Special) unitary becomes (special) orthogonal, and the Hermitian adjoint $(\cdot)^*$ becomes the transpose $(\cdot)^\top$.

classical two-sided Jacobi eigenvalue algorithm by Chartres [Cha62]. It can be viewed as implicitly running the two-sided Jacobi eigenvalue algorithm on the matrix $B = A^*A$, without ever forming B. It enjoys many of the same advantages as the two-sided Jacobi eigenvalue algorithm [Bis89; DV92].

The numerical stability of both the one- and two-sided variants was investigated by Demmel and Veselić in a highly influential paper [DV92]. They primarily focus on the accuracy with which the small eigenvalues of positive definite matrices are computed. They find "modulo an assumption based on extensive numerical tests" that the Jacobi eigenvalue algorithm essentially produces the best approximation of the small eigenvalues one could hope for in the floating point arithmetic model. A family of matrices with more problematic behavior that those tested by [DV92] was published by [Mas94], though they do not result in instability of the method overall.

Cholesky decomposition / QR decomposition: Gaussian elimination is the oldest method for factoring a given matrix B as $B = LU$ where L is lower triangular and U is upper triangular [Grc11]. When B is positive definite, one is guaranteed that $U = L^*$ resulting in the Cholesky decomposition $B = LL^*$. Gaussian elimination with various pivoting rules is the primary algorithm used in practice for the Cholesky decomposition [TB97; Hig09; And$^+$99].

There are many methods for the QR decomposition. Both classical and modified Gram-Schmidt applied to $A \in \mathbb{C}^{d\times n}$ are mathematically equivalent to Gaussian elimination applied to the matrix $B = A^*A$ [Hig90]. Some other algorithms actively exploit the connection between the QR and Cholesky decompositions: [GV13] proposes computing the Cholesky decomposition $A^*A = R^*R$ and then computing $Q = AR^{-1}$, and the algorithm(s) of [YNYF16] and [FKNYY20] explore this idea further, with modifications to improve stability.

There is a notable way to use the Cholesky decomposition and one-sided Jacobi eigenvalue in conjunction for the purpose of a symmetric eigendecomposition as well. The idea is to compute the Cholesky factorization $B = LL^*$ of an input, and compute the SVD of $L^* = U\Sigma V^*$, which then implies the eigendecomposition $B = V\Sigma^2 V^*$. The proposal of [VH89] is to compute this SVD by applying the one-sided Jacobi eigenvalue algorithm to L (not to L^*). This technique avoids the aforementioned assumption needed by [DV92] in their theoretical analysis of the numerical stability of the Jacobi eigenvalue algorithm [Drm94]. However, while it ensures accurate eigenvalues, it does not ensure accurate eigenvectors [Mat95]. Nevertheless, it is observed to be faster [DV92]. The version of this technique which applies one-sided Jacobi to L^* is also faster, though it lacks a theoretical or practical advantage in terms of numerical stability [DV92]. Several additional variants of this idea were studied by [DV08].

Several works take inspiration from algorithms for one task for the other. For instance, [Luk86] proposes a Jacobi-like method for the QR decomposition, applying Givens rotations to transform a matrix A into an upper triangular one R. The accumulation of all the Givens rotations then forms Q (which is akin to the method for the QR decomposition using Householder reflections [TB97]). This method for QR now commonly appears in numerical linear algebra textbooks, e.g. [Dem97].

The central focus of this work is the striking similarity between the operations performed by the one-sided Jacobi eigenvalue algorithm and the modified Gram-Schmidt procedure. This similarity has been previously noted, for instance [Drm97] observes that the operations are actually the same under a certain kind of infinite matrix limit. We show that this similarity can be exploited to obtain a unified analysis of these algorithms and more.

Overview

In Section 3.2, we present the Jacobi eigenvalue algorithm and modified Gram-Schmidt procedure (MGS) in a way that highlights their similarity. In Section 3.2 we formally describe an abstract algorithm which encompasses the Jacobi eigenvalue algorithm and MGS as special cases.

Our results about this generalized algorithm are presented in the remaining sections. In Section 3.3 we introduce our main tool for evaluating progress in the algorithm, a certain potential function, and we relate this potential function to other standard metrics. Section 3.4 focuses on exact-arithmetic implementations of the algorithm, in the real RAM model. For inputs that are not too ill-conditioned, we show that $O(n^2 \log(n/\delta))$ iterations of this procedure suffice to produce δ-approximate factorizations (stated precisely in Theorem 3.4.4). For many reasonable special cases of this generalized algorithm, each iteration costs only $O(n)$ arithmetic operations, resulting in many $\widetilde{O}(n^3)$-time factorization algorithms. Additionally, this section confirms a recent conjecture of Steinerberger about the convergence of particular method for orthogonalization [Ste25] (see Remark 6).

Section 3.5 studies the general algorithm in finite arithmetic. Our focus is on proving statements roughly of the form "when machine precision is at most n^{-c} (for an explicit, small c), then the results of the algorithm have a guaranteed level of accuracy." Our analysis does not optimize c, and we do not advertise bounds that are particularly attractive for large matrices in single and double floating point arithmetic. Nevertheless, the dependence on n on the bits of precision required is shown to be logarithmic, and we prove that quad precision can handle a substantial range of inputs.

We have two primary results about finite arithmetic: the first is that a large class of orthogonalization algorithms are numerically stable (including an algorithm for computing the QR decomposition specifically). The second addresses the "assumption based on extensive numerical tests" needed by Demmel and Veselić in [DV92] to show stability of the Jacobi eigenvalue algorithm. This assumption is that a certain condition number remains bounded as iterations are performed. Their experiments showed that it doesn't seem to rise more than a constant amount above its initial value, but their proofs only bound it by certain exponentially growing quantities (exponential either in the size of the input or number of iterations performed). We show that with uniformly random pivoting, it is bounded by a polynomial in the input size and number of iterations performed (see Theorem 3.5.4).

Notation	Meaning
$\|\cdot\|$	ℓ_2 vector or operator norm
$\|\cdot\|_F$	Frobenius norm
$\lambda_j(A)$	j^{th} largest eigenvalue of A
$\kappa(A) = \|A\| \cdot \|A^{-1}\|$	Condition number of A
$D_A = \text{diag}(A_{11}, A_{22}, \ldots, A_{nn})$	Diagonal part of A
$\hat{A} = D_A^{-1/2} A D_A^{-1/2}$	Diagonal-normalization of A
$\hat{\kappa}(\cdot)$	Diagonal normalized condition number, $\hat{\kappa}(A) := \kappa(\hat{A})$
$A \odot B$	Hadamard or entrywise product, $(A \odot B)_{ij} = a_{ij} b_{ij}$
$O(\cdot), \widetilde{O}(\cdot)$	$O(\cdot)$ hides constants, $\widetilde{O}(\cdot)$ hides polylogarithmic factors.
$[n] = \{1, \ldots, n\}$	Set of indices 1 through n

3.2 Connecting SVD and QR

The core connection

Here we describe a simple connection between the SVD and QR decomposition (and the corresponding connection between the eigendecomposition and Cholesky decomposition) and algorithms for computing them. To see the connection, we first express the QR decomposition and a factorization that is almost the SVD in a common form: given A with full column rank, write

$$A = QT^{-1} \tag{3.1}$$

where Q has orthogonal columns and T is nonsingular. When T is unitary, by further decomposing $Q = U\sqrt{Q^*Q}$, one obtains the singular value decomposition. When T is upper triangular and $Q^*Q = I$, this is exactly the QR decomposition. Analogously, the eigendecomposition and Cholesky decomposition of $B = A^*A$ can both be expressed as $B = (T^{-1})^* D T$ where $D = Q^*Q$. The four cases are summarized in the following table.

	$B = (T^{-1})^* D T^{-1}$	$A = QT^{-1}$
T is unitary and $D = Q^*Q$ is diagonal	Eigendecomposition	Singular value decomposition
T is upper triangular and $D = Q^*Q = I$	Cholesky decomposition	QR decomposition

Notably, both Modified Gram-Schmidt (MGS) and one-sided Jacobi can be interpreted as methods for producing an orthogonal basis $Q = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix}$ for a given collection of vectors, expressed as columns of $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$ (along with the corresponding column operation sending A to Q). The algorithms themselves closely resemble each other too: they both iteratively pick pairs of indices (i, j) and replace the columns a_i, a_j with a different orthogonal basis for their span. By repeating this many times with appropriate selection of (i, j) each time, the collection of vectors will approach an orthogonal set. More specifically let $A^{(t)} =$

$\begin{bmatrix} a_1^{(t)} & \cdots & a_n^{(t)} \end{bmatrix}$ be the state after this has been done for t pairs (i, j). For each t, the Jacobi eigenvalue algorithm selects a pair (i, j) and then updates

$$\begin{aligned} a_i^{(t+1)} &= \cos(\theta) a_i^{(t)} + \sin(\theta) a_j^{(t)} \\ a_j^{(t+1)} &= -\sin(\theta) a_i^{(t)} + \cos(\theta) a_j^{(t)} \\ a_\ell^{(t+1)} &= a_\ell^{(t)} \quad \forall \ell \notin \{i, j\} \end{aligned} \tag{3.2}$$

where θ is such that the resulting columns $a_i^{(t+1)}$ and $a_j^{(t+1)}$ are orthogonal. MGS selects $i < j$ then updates

$$\begin{aligned} a_j^{(t+1)} &= \frac{a_j^{(t)} - \left\langle a_j^{(t)}, a_i^{(t)} \right\rangle a_i^{(t)}}{\left\| a_j^{(t)} - \left\langle a_j^{(t)}, a_i^{(t)} \right\rangle a_i^{(t)} \right\|} \\ a_\ell^{(t+1)} &= a_\ell^{(t)} \quad \forall \ell \neq j. \end{aligned} \tag{3.3}$$

The update equation (3.2) is a Givens rotation of $A^{(t)}$ (also called a plane rotation); importantly it is unitary. The product of these rotations forms T in the decomposition (3.1). The update equation (3.3) is an elementary matrix operation followed by a scaling; importantly, it is upper triangular. The product of these elementary operations forms T in the decomposition (3.1).

In both cases, the result of this update is that the ith and jth columns of $A^{(t+1)}$ will be orthogonal. Despite their similarity, the existing analysis of these algorithms is quite different, as is their empirical performance measured in terms of computational and communication costs and numerical stability.

Pivoting rules

In both cases, the strategy for picking (i, j) at each time t is determined by a "pivoting rule". Multiple pivoting rules have been proposed; we group them into three categories to discuss: fixed, adaptive, and randomized. Before doing so, we must clarify some terminology. There is some subtlety in what the literature refers to as "one iteration" of either procedure. For the Jacobi eigenvalue algorithm, one iteration is sometimes taken to mean one "sweep" (which we describe in a moment) of all $n(n-1)/2$ pairs (i, j). For MGS (more specifically, left-looking MGS), one iteration typically refers to the grouping of updates

$$(1, i), (2, i), \ldots, (i-1, i) \tag{3.4}$$

for each i. For Gaussian elimination (and right-looking MGS), it typically refers to the grouping of updates

$$(i, i+1), (i, i+2), \ldots, (i, n) \tag{3.5}$$

for each i. Because of this, the word "pivot" in those contexts typically refer to the single column index i. In this paper, we refer to each selection of (i, j) and update of the corresponding column(s) as one iteration, i.e. $A^{(t)}$ defined by either (3.2), (3.3) is the state after t iterations. One pivot then corresponds to a pair of column indices (i, j).

Fixed pivot rules: These are rules that fix a sequence of (i, j) ahead of time. The ones we highlight for the Jacobi eigenvalue algorithm are called cyclic or sweep rules, due to [Gre53], see also [FH60]. The row-cyclic version uses the sequence

$$(1,2),(1,3),(1,4),(1,5),\ldots,(1,n),(2,3),(2,4),(2,5),\ldots,(2,n),\ldots(n-1,n)$$

and repeats it over and over again. The column-cyclic version uses the sequence

$$(1,2),(1,3),(2,3),(1,4),(2,4),(3,4),(1,5),(2,5),(3,5),(4,5),\ldots,(n-1,n)$$

and repeats it over and over again. Each run through the sequence of $n(n-1)/2$ pairs is referred to as one "sweep." In the context of Gaussian elimination and MGS, those sweeps correspond to applying the groupings (3.4) or (3.5) for $i = 1, \ldots, n-1$. Running MGS for just one or two sweeps are both popular proposals.

For appropriately chosen rotations, [FH60] show that row-cyclic and column-cyclic Jacobi converges, and [Sch64] improve this to (eventual) quadratic convergence. However, there is no rigorous theory to predict the number of sweeps required to achieve a desired distance to diagonal [GV13]. In practice, it is observed to be a constant number of sweeps [Dem97]. Better theoretical results are available for the next category of pivoting rules.

Adaptive pivot rules: These are pivoting rules that use the current value of $A^{(t)}$ to decide the next pivots. The original rule proposed by Jacobi [Jac46], which we call the greedy pivoting rule, was to pick

$$(i,j) = \arg\max_{i,j:i\neq j} |\langle a_i, a_j\rangle| . \tag{3.6}$$

This is somewhat similar to column-pivoting rule of Gram-Schmidt (implicitly equivalent to complete pivoting rule of Gaussian elimination). This rule for Gram-Schmidt maintains a list of indices $\mathcal{I}$ initialized to $[n]$. Then

$$i = \arg\max_{i\in\mathcal{I}} \|a_i\| \tag{3.7}$$

is computed, i is removed from $\mathcal{I}$, and the updates $\{(i,j) : j \in \mathcal{I}\}$ are performed (in any order). This is repeated until $\mathcal{I}$ is empty. Strictly speaking, these rules will not necessarily compute the QR decomposition of A, since we no longer are assured that the update pairs (i, j) satisfy $i < j$. However, if one lets σ be the permutation which records the order in which indices are removed from $\mathcal{I}$, then we are assured $\sigma(i) < \sigma(j)$ and therefore the method computes QR decomposition of AP_σ where P_σ is the permutation matrix associated with σ.

The same pivoting strategy has been considered for one-sided Jacobi [dRij89]. Numerical experiments showed that this pivoting strategy can improve the rate of convergence in several (but not all) cases (see [dRij89] for more details).

Randomized pivot rules: There seems to be much less attention given to randomized pivot rules. A recent work of Chen et al. [CETW24] considers the partial Cholesky and QR decompositions for computing low-rank approximations. They study a randomized variant of the complete pivoting rule (3.7) where $i \in \mathcal{I}$ is sampled with probability proportional to $\|a_i\|^2$. This and similar ideas had been considered before as well [FKV04; DV06; DRVW06].

One exceedingly simple randomized rule is just to sample (i, j) uniformly at random among all pairs of indices. We were unable to find a reference that explicitly describes this rule (or any randomized rule) for the Jacobi eigenvalue algorithm; one reason that this is surprising is that the proof of global convergence of the greedy pivoting rule (3.6) essentially applies the probabilistic method to the hypothetical algorithm which uses this uniformly random pivot (see, for example, Section 5 of [GMvN59]).

In the context of orthogonalization, we note an important distinction between sampling (i, j) uniformly at random among all pairs of indices and just pairs of indices with $i < j$ (or $\sigma(i) < \sigma(j)$ for some permutation σ). In the latter case, we are assured the computed decomposition (if the method converges) is the QR decomposition $AP_\sigma = QR$. However, in the former case, we are given no assurances about R being upper triangular. That rule (without the $i < j$ constraint) was proposed by Steinerberger [Ste25], and both rules (both with and without the $i < j$ constraint) were concurrently proposed by the present author and Shah in [DS26b]. Both those works were motivated by a designing a preconditioner for the Kaczmarz linear system solver, and made analogies of the method to the Kaczmarz solver itself; neither work identified the connection to the Jacobi eigenvalue algorithm. Steinerberger conjectured that the rate of convergence of this procedure is $\kappa(A^{(t)}) \approx \exp(-O(t/n^2))\kappa(A)$ based on numerical experiments and a heuristic. The present author and Shah proved that $(-\log\det|A^{(t)}|) \leqslant \exp(-O(t/n^2))(-\log\det|A|)$.

Block variants

Both Gram-Schmidt and Jacobi admit block variants, which can be abstractly described as follows: rather than pick a *pair* of indices (i, j) at each step and updating columns i and j of $A^{(t)}$, one can instead pick any subset of indices $J \subset [n]$ and update all the corresponding columns. We call J the pivot set. In the non-block case, pivoting on (i, j) ensures that

$$\left\langle a_i^{(t+1)}, a_j^{(t+1)} \right\rangle = 0.$$

In the block variant, pivoting on J ensures that

$$\left\langle a_i^{(t+1)}, a_j^{(t+1)} \right\rangle = 0 \quad \forall i, j \in J, i \neq j.$$

These methods are often motivated by considering the practical performance of these algorithms on modern hardware. For example, one may be able to take advantage of large cache sizes to perform one iteration for $|J| > 2$ very quickly.

The block Jacobi eigenvalue algorithm of Drmač [Drm10] is one such method. It fixes a partition $I_1 \sqcup \cdots \sqcup I_m = [n]$ and divides the given matrix B into blocks corresponding to

those index sets. Then, the pivot set J is selected to be $I_\ell \cup I_{\ell'}$, where the pair (ℓ, ℓ') can be selected analogously to how (i, j) were selected in the non-block version. Both [BH24] and [HB17] show convergence of block Jacobi under various pivoting strategies. There are many block versions of Gram-Schmidt along similar lines, see [CLRT22] for a recent survey.

A generalized factorization

In describing the connection between Gram-Schmidt and the one-sided Jacobi eigenvalue algorithm, we've already hinted at what the generalized algorithm is. Once a pivot set J has been selected, we may abstract away the process by which the corresponding columns are updated: all we assume is that it is an invertible column operation which results in orthogonal columns. We will use the letter S for the $|J| \times |J|$ matrix performing this column operation. The Jacobi eigenvalue algorithm corresponds to the special case where S is unitary, and Gram-Schmidt to the case where S is upper triangular. The general pseudo-code for this and the two-sided version is below. We first clarify some notation. $\mathrm{GL}_k(\mathbb{C})$ is the set of all invertible $k \times k$ matrices over $\mathbb{C}$. $\binom{[n]}{k}$ is the set of all subsets of $[n]$ of size k. For an index set $J = \{i_1, \ldots, i_k\} \in \binom{[n]}{k}, i_1 < \cdots < i_k$, denote the submatrices

$$A_J = \begin{bmatrix} a_{1,i_1} & \cdots & a_{1,i_k} \\ \vdots & & \vdots \\ a_{d,i_1} & \cdots & a_{d,i_k} \end{bmatrix} \in \mathbb{C}^{d \times k} \quad \& \quad B_{JJ} = \begin{bmatrix} b_{i_1 i_1} & \cdots & b_{i_1 i_k} \\ \vdots & & \vdots \\ b_{i_k i_1} & \cdots & b_{i_k i_k} \end{bmatrix} \in \mathbb{C}^{k \times k}.$$

For a permutation σ on $[n]$ let P_σ be the associated permutation matrix, i.e. $P_\sigma e_j = e_{\sigma(j)}$. We present both versions of the algorithms side-by-side to highlight their mathematical equivalence: one can start with $B = A^*A$ and maintain $A^{(t)*}A^{(t)} = B^{(t)}$ by selecting the same T in line 7 of either algorithm. Because of this correspondence, we will refer to various lines of the algorithm without needing specify which version we are talking about. We also highlight that T need not be formed and inverted explicitly.

There are many pivot rules one can use for the selection in line 5. Here we use a randomized pivoting rule where J is sampled uniformly at random from $\binom{[n]}{k}$. We refer to this as "randomized size k pivoting".

We obtain a unified analysis of all special cases of Algorithm 1 with randomized size k pivoting. The mechanism by which S is selected in line 6 and the additional structure one wishes to impose upon S do not affect our analysis in any way—in expectation, the rate of convergence depends only on k and n. For concreteness, Table 3.1 lists several decompositions one can obtain by imposing different constraints on S and $(A_J^{(t)}S)^*(A_J^{(t)}S) = S^*B_{JJ}^{(t)}S$.

Algorithm 1 This is pseudo-code for the one-sided (left) and two-sided (right) versions of the generalized factorization algorithm.

Require: $A \in \mathbb{C}^{d\times n}$ has full column rank, pivot size $k \geq 2$.
Ensure: $A = QT_{\text{acc}}$
1: $A^{(0)} \leftarrow A$.
2: $T_{\text{acc}}^{(0)} \leftarrow I$.
3: $t \leftarrow 0$
4: **while** $A^{(t)*}A^{(t)}$ is not close to diagonal **do**
5: Select a pivot set $J = \{i_1, \ldots, i_k\}$ with $i_1 < \cdots < i_k$.
6: Construct $S \in \mathrm{GL}_k(\mathbb{C})$ such that $A_J^{(t)}S$ has orthogonal columns.
7: Define $T \in \mathbb{C}^{n\times n}$ as $T = P_\sigma^* \begin{bmatrix} S^{-1} & \\ & I \end{bmatrix} P_\sigma$ where σ is a permutation with $\sigma(i_j) = j$.
8: $A^{(t+1)} \leftarrow A^{(t)}T^{-1}$
9: $T_{\text{acc}}^{(t+1)} \leftarrow T_{\text{acc}}^{(t)}T$
10: $t \leftarrow t + 1$
11: **end while**
12: **return** $Q \leftarrow A^{(t)}$, $T_{\text{acc}} \leftarrow T_{\text{acc}}^{(t)}$

Require: $B \in \mathbb{C}^{n\times n}$ is positive definite, pivot size $k \geq 2$.
Ensure: $B = T_{\text{acc}}^* D T_{\text{acc}}$
1: $B^{(0)} \leftarrow B$.
2: $T_{\text{acc}}^{(0)} \leftarrow I$.
3: $t \leftarrow 0$
4: **while** $B^{(t)}$ is not close to diagonal **do**
5: Select a pivot $J = \{i_1, \ldots, i_k\}, i_1 < \cdots < i_k$.
6: Construct $S \in \mathrm{GL}_k(\mathbb{C})$ such that $S^*B_J^{(t)}S$ is diagonal.
7: Define $T \in \mathbb{C}^{n\times n}$ as $T = P_\sigma^* \begin{bmatrix} S^{-1} & \\ & I \end{bmatrix} P_\sigma$ where σ is a permutation with $\sigma(i_j) = j$.
8: $B^{(t+1)} \leftarrow (T^{-1})^*B^{(t)}T^{-1}$
9: $T_{\text{acc}}^{(t+1)} \leftarrow T_{\text{acc}}^{(t)}T$
10: $t \leftarrow t + 1$
11: **end while**
12: **return** $D \leftarrow B^{(t)}$, $T_{\text{acc}} \leftarrow T_{\text{acc}}^{(t)}$

Decomposition name	S	$(A_J^{(t)}S)^*(A_J^{(t)}S) = S^*B_{JJ}^{(t)}S$
Eigendecomposition / SVD	Unitary	Diagonal
Cholesky / QR	Upper triangular	Identity
LDL / unnamed	Unit upper triangular	Diagonal
unnamed / QL	Lower triangular	Identity
Gram matrix / Orthogonalization	Nonsingular	Identity

Table 3.1: Two-sided and one-sided decomposition of the form $B = T^*DT$ and $A = QT$ that one obtains by different selections of S in line 6 of Algorithm 1.

Remark 1 (Using recursion to compute S)**.** Let's revisit line 6 of either algorithm. Computing S is itself a matrix decomposition problem. In fact, the decomposition of B_{JJ} or A_J one must find is exactly the kind of decomposition one seeks to compute of B or A. This points to the natural choice of using the algorithm recursively. Alternatively, one may use any expensive factorization algorithm since the problem instance is smaller.

Remark 2 (Parallelizability)**.** When two pivot sets are disjoint, the corresponding updates may occur in parallel. The results here can be extended to the case where multiple disjoint

pivots $J_1, \ldots, J_\ell \in \binom{[n]}{k}$ are sampled at a time where the marginal distribution of each J_j is uniform, and then the corresponding pivots are performed in any order.

Remark 3 (Fixed pivot size)**.** There's no reason to keep the pivot size $|J| = k$ constant, other than it simplifies the analysis. J may be sampled from any set such that the probability $\{i, j\} \subset J$ is equal for all distinct pairs i, j.

Remark 4 (Meaning of "convergence")**.** The algorithm halts when $A^{(t)*}A^{(t)} = B^{(t)}$ is close to diagonal, in whatever sense the user demands. When we say that the method converges, we mean that the criteria that ends the while-loop in the algorithm is met. Other sources may impose a stronger condition on convergence: it means that the version of the algorithm which continues iterating indefinitely, never exiting the while-loop, must produce states such that the limits $\lim_{t\to\infty} A^{(t)}$ (or $\lim_{t\to\infty} B^{(t)}$) and $\lim_{t\to\infty} T^{(t)}_{\text{acc}}$ exist. In particular, if $A(t)$ has reached an orthonormal basis, but that basis changes each iteration, we will still say it has converged.

3.3 Main tool for analysis

In Section 3.3 we review the standard Jacobi analysis, from which we draw inspiration. This analysis rests on a certain potential function, used to measure progress of convergence to a diagonal matrix. In Section 3.3 we introduce our potential function, $\Gamma(\cdot)$, and relate $\Gamma(\cdot)$ to several other metrics.

Standard Jacobi analysis

Our analysis of Algorithm 1 follows a similar strategy as the classical analysis of the Jacobi eigenvalue algorithm. To highlight the connection, we review the classical analysis here. Say that $B = B^{(0)}, B^{(1)}, B^{(2)}, \ldots$ is the sequence of matrices produced by the iterations of the Jacobi eigenvalue algorithm. The proof strategy is to identify a potential function that behaves monotonically, i.e. some quantity which strictly decreases as iterations are performed. In the classical analysis, this potential is taken to be the squared distance of the current iterate $B^{(t)}$ to the closest diagonal matrix in Frobenius norm, denoted $\text{off}(\cdot)$. This quantity can be compactly expressed as

$$\text{off}(B) := \inf_{D \text{ diagonal}} \|B - D\|_F^2 = \|B - D_B\|_F^2 = \sum_{i,j\in[n], i\neq j} |b_{ij}|^2$$

where $D_B = \text{diag}(b_{11}, \ldots, b_{nn})$. A straightforward calculation reveals that that

$$\text{off}(B^{(t+1)}) = \text{off}(B^{(t)}) - 2|b^{(t)}_{ij}|^2 \tag{3.8}$$

where (i, j) is the pivot selected at time t. If (i, j) is picked according to the greedy pivot rule (i.e. $|b^{(t)}_{ij}|^2 = \arg\max_{i,j:i\neq j} |b^{(t)}_{ij}|^2$), then $|b^{(t)}_{ij}|^2$ will certainly be greater than the average

so one has

$$|b_{ij}^{(t)}|^2 \geq \frac{1}{n(n-1)} \operatorname{off}(B^{(t)})$$

which ensures global convergence of $\operatorname{off}(B^{(t)}) \to 0$ at a linear rate. If instead one uses a randomized size 2 pivoting strategy, one has equality in expected value,

$$\mathbb{E} \operatorname{off}(B^{(t+1)}) = \left(1 - \frac{2}{n(n-1)}\right) \operatorname{off}(B^{(t)}), \tag{3.9}$$

which again ensures global convergence in expectation at a linear rate. Although both cases reduce to reasoning about the expectation, it does not seem that a random pivoting strategy has been considered in the literature.

In the more general setting where $B^{(t)}$ is the sequence produced by non-unitary transformations, we no longer have (3.8) (and therefore not (3.9) either). Instead, we obtain an expression similar to (3.9) but with a different function in place of $\operatorname{off}(\cdot)$.

A new potential function

Our potential function admits a few equivalent expressions. Letting $\hat{B} = D_B^{-1/2} B D_B^{-1/2}$, i.e., the diagonal-normalization of B, we define $\Gamma(B)$ as

$$\Gamma(B) := \operatorname{tr}(B \odot B^{-1} - I) = \operatorname{tr}(\hat{B}^{-1}) - n = -\sum_{i,j \in [n], i \neq j} b_{ij} \cdot (B^{-1})_{ji}. \tag{3.10}$$

As with $\operatorname{off}(\cdot)$, it is clear from the definition that $\Gamma(B) = 0$ when B is diagonal. Less obvious, now, is the converse statement and quantitative versions. Notably, the converse fails among indefinite or general matrices[3]. Fortunately, among positive definite matrices, $\Gamma(B)$ is proportional to the *Jensen gap* of a randomly sampled eigenvalue of $\hat{B}$ for the function $\lambda \mapsto \lambda^{-1}$:

$$\frac{1}{n}\Gamma(B) = \mathbb{E}(\lambda^{-1}) - (\mathbb{E}(\lambda))^{-1} = \frac{1}{n} \operatorname{tr}\left(\hat{B}^{-1}\right) - \left(\frac{1}{n} \operatorname{tr}(\hat{B})\right)^{-1}. \tag{3.11}$$

The expressions (3.10), (3.11) are equivalent since $\hat{B}$ has all 1s along the diagonal, so $\operatorname{tr}(\hat{B}) = n$. Jensen's inequality directly implies $\Gamma(B) \geq 0$, with equality if and only if all the eigenvalues of $\hat{B}$ are equal, i.e. $\hat{B} = I$, which is equivalent to $B = D_B$ being diagonal. Unfortunately, the statement "$\Gamma(B) = 0$ implies B is diagonal" is not enough to conclude that if for a sequence $B^{(t)}$ with $\Gamma(B^{(t)}) \to 0$ that $\operatorname{off}(B^{(t)}) \to 0$. Consider, for example,

$$B^{(t)} = \begin{bmatrix} t^2 & t \\ t & t^2 \end{bmatrix}, \quad \hat{B}^{(t)} = \begin{bmatrix} 1 & t^{-1} \\ t^{-1} & 1 \end{bmatrix}.$$

[3] A counter example is $\Gamma\left(\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right) = 0$.

In this example, $\Gamma(B^{(t)}) = 2(t^2-1)^{-1} \to 0$ and yet $\text{off}(B^{(t)}) = 2t^2 \to \infty$. This example is capturing the difference between becoming closer to diagonal in an absolute versus relative sense. Indeed $B^{(t)}$ is moving further from the set of diagonal matrices in an absolute sense, but in a relative sense we should instead consider $\text{off}(\hat{B}^{(t)}) = 2t^{-2} \to 0$. It is this notion of distance-to-diagonal that is captured by $\Gamma(\cdot)$. Specifically, as shown in Theorem 3.3.1, $\Gamma(B^{(t)}) \to 0$ implies $\text{off}(\hat{B}^{(t)}) \to 0$ at the same rate. We justify considering only this relative sense of distance-to-diagonal by noting that it is used as a standard stopping criterion for the Jacobi eigenvalue algorithm [DV92]; therefore, convergence of $\Gamma(B^{(t)}) \to 0$ is enough to show convergence of Algorithm 1 for stopping criteria that are satisfied when $\text{off}(\hat{B})$ is reduced below some threshold.

This relationship between $\Gamma(\cdot)$ and $\text{off}(\cdot)$ is intuitively explained by a general fact that the Jensen gap is larger for distributions that are "spread out" and smaller for distributions that are highly concentrated. In particular, $\text{off}(\hat{B})$ is directly proportional to the variance of a randomly selected eigenvalue:

$$\mathbb{E}(\lambda^2) - \mathbb{E}(\lambda)^2 = \frac{1}{n}\operatorname{tr}(\hat{B}^2) - 1 = \frac{1}{n}\|\hat{B} - I\|_F^2 = \frac{1}{n}\operatorname{off}(\hat{B}).$$

This same mechanism can be used to control the ratio between the largest and smallest eigenvalues of $\hat{B}$: if this ratio is close to 1, the eigenvalues are clustered, and if it is large, they are spread out. This ratio is simply the condition number $\kappa(\hat{B})$. This quantity plays an important role in the numerical stability analysis of the Jacobi eigenvalue algorithm appearing in [DV92]; it is known as the diagonal-normalized condition number, which we denote as $\hat{\kappa}(B) := \kappa(\hat{B})$.

This condition number can be arbitrarily smaller than the usual condition number $\kappa(\cdot)$ (consider, for example, any diagonal matrix), but cannot be much larger (specifically, as shown in [Van69], we have $\hat{\kappa}(B) \leqslant n\kappa(B)$.). The relationship between $\Gamma(\cdot)$ and $\hat{\kappa}(\cdot)$ presented in Theorem 3.3.2 is what allows us to prove polynomial control over the numerical stability of Algorithm 1.

These relationships between $\Gamma(B)$ and $\kappa(\hat{B})$ and $\text{off}(\hat{B})$ are derived by considering an optimization program of the eigenvalues λ_j of $\hat{B}$. The constraints come from the invariant $\operatorname{tr}(\hat{B}) = n$ and the definition $\Gamma(B) = \Gamma(\hat{B}) = \operatorname{tr}(\hat{B}^{-1}) - n$. Converting these to statements about the eigenvalues themselves results in the constraints

$$\lambda_1, \ldots, \lambda_n > 0, \quad \lambda_1 + \cdots + \lambda_n = n, \quad \lambda_1^{-1} + \cdots + \lambda_n^{-1} = n + \Gamma(\hat{B}). \tag{3.12}$$

Strictly speaking, one may also wish to impose the ordering of the eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_n$. Due to the symmetry of the other constraints, this turns out not to be important. Observe that the constraint $1/\lambda_1 + \cdots + 1/\lambda_n = n + \Gamma(\hat{B})$ implies all feasible points avoid the boundary, $\lambda_j = 0$ for some j. The set of feasible points is also compact, so the maximizing point for any continuous objective function will be achieved for a point in the interior of the positive orthant. This allows us to find optima of smooth functions by the method of Lagrange multipliers. This method uses the following fact: if left-hand sides of the equalities in (3.12)

are $F_1(\lambda_1, \ldots, \lambda_n) = \lambda_1 + \ldots + \lambda_n$ and $F_2(\lambda_1, \ldots, \lambda_n) = \lambda_1^{-1} + \cdots + \lambda_n^{-1}$, and the objective is Φ, then the optimal assignment satisfies

$$\operatorname{rank}\left(\begin{bmatrix} \nabla\Phi(\lambda_1, \ldots, \lambda_n) \\ \nabla F_1(\lambda_1, \ldots, \lambda_n) \\ \nabla F_2(\lambda_1, \ldots, \lambda_n) \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} & \nabla\Phi(\lambda_1, \ldots, \lambda_n) & \\ 1 & \cdots & 1 \\ -1/\lambda_1^2 & \cdots & -1/\lambda_n^2 \end{bmatrix}\right) \leqslant 2. \quad (3.13)$$

In particular, the determinant of every 3×3 submatrix must vanish.

Lemma 3.3.1.

$$\operatorname{off}(\hat{B}) \leqslant \Gamma(B) \cdot \left(\sqrt{1 + \Gamma(B)/4} + \sqrt{\Gamma(B)/4}\right)^2$$

Proof. Put $g = \Gamma(\hat{B}) = \Gamma(B)$. We have

$$\|\hat{B} - I\|_F^2 = -n + \|\hat{B}\|_F^2 = -n + \lambda_1^2 + \cdots + \lambda_n^2. \quad (3.14)$$

We consider maximizing (3.14) subject to (3.12). This is a smooth objective so by (3.13),

$$\operatorname{rank}\left(\begin{bmatrix} 2\lambda_1 & \cdots & 2\lambda_n \\ 1 & \cdots & 1 \\ -1/\lambda_1^2 & \cdots & -1/\lambda_n^2 \end{bmatrix}\right) \leqslant 2 \quad (3.15)$$

By considering a vector $\begin{bmatrix} 1 & a & b \end{bmatrix}$ in the left-nullspace, we see that each eigenvalue must be the root of a common cubic polynomial $z^3 + az^2 - b = 0$. By Descartes' rule of signs, this polynomial can have at most two positive roots. Without loss of generality, say

$$\lambda_1 = \cdots = \lambda_k \geq \lambda_{k+1} = \cdots = \lambda_n$$

for some index k. Plugging these into these constraints gives

$$k\lambda_1 + (n-k)\lambda_n = n, \quad \frac{k}{\lambda_1} + \frac{n-k}{\lambda_n} = n + g.$$

Substituting for λ_n and clearing the denominators shows that λ_1 must be the larger root of

$$z^2 - \frac{n\,(2k+g)}{k\,(n+g)} z + \frac{n}{n+g} = 0.$$

By applying the quadratic formula, we see that the objective

$$k\lambda_1^2 + (n-k)\lambda_n^2$$

is monotonically increasing in k, so we should take $k = n-1$. By rearranging the expression obtained from the quadratic formula, this results in

$$\lambda_1 = \cdots = \lambda_{n-1} = \frac{1 + \frac{g}{2(n-1)} - \sqrt{\frac{g}{(n-1)n} + \frac{g^2}{4(n-1)^2}}}{1 + g/n} \qquad \& \qquad \lambda_n = n - (n-1)\lambda_1.$$

Then the maximum of the objective function can be expressed as

$$\begin{aligned} -n + (n-1)\lambda_1^2 + \lambda_n^2 &= -n + (n-1)\lambda_1^2 + (n - (n-1)\lambda_1)^2 \\ &= n(n-1)(\lambda_1 - 1)^2. \end{aligned} \tag{3.16}$$

The resulting expression is monotonically increasing in n, so one may upper bound this quantity by taking the limit as $n \to \infty$. This yields

$$\operatorname{off}(\hat{B}) \leqslant \left(g/2 + \sqrt{g + g^2/4}\right)^2.$$

□

Lemma 3.3.2.

$$1 + \Gamma(B)/n \leqslant \kappa(\hat{B}) \leqslant (1 + \sqrt{\Gamma(B)/2} + \Gamma(B)/2)^2.$$

Proof. For the lower bound on $\kappa(\hat{B})$, note that

$$n\Gamma(\hat{B}) = n \operatorname{tr}(\hat{B}^{-1}) - n^2 = \operatorname{tr}(\hat{B}) \cdot \operatorname{tr}(\hat{B}^{-1}) - n^2 \leqslant n\|\hat{B}\| \cdot n\|\hat{B}^{-1}\| - n^2 = n^2\kappa(\hat{B}) - n^2.$$

For the upper bound, put $g = \Gamma(B) = \Gamma(\hat{B})$. We have

$$\kappa(\hat{B}) = \max(\lambda_1, \ldots, \lambda_n) / \min(\lambda_1, \ldots, \lambda_n). \tag{3.17}$$

We consider maximizing (3.17) subject to (3.12). Since the constraints are all symmetric in the variables, we may permute the assignment to the variables which maximizes $\kappa(\hat{B})$ so that λ_1 and λ_n are the largest and smallest variables respectively. In particular, this means we may simply maximize the smooth objective λ_1/λ_n. By (3.13),

$$\operatorname{rank}\left(\begin{bmatrix} 1/\lambda_n & 0 & \cdots & 0 & -\lambda_1/\lambda_n^2 \\ 1 & 1 & \cdots & 1 & 1 \\ -1/\lambda_1^2 & -1/\lambda_2^2 & \cdots & -1/\lambda_{n-1}^2 & -1/\lambda_n^2 \end{bmatrix}\right) \leqslant 2. \tag{3.18}$$

The determinant of each 3×3 submatrix must be 0. Take the columns i, j, n for $1 < i, j < n$. This determinant is

$$\frac{\lambda_1}{\lambda_n^2}\left(\frac{1}{\lambda_j^2} - \frac{1}{\lambda_i^2}\right) = 0$$

which implies $\lambda_i = \lambda_j$. The rank condition (3.18) thus simplifies dramatically since the middle $n - 2$ columns coincide,

$$\det\left(\begin{bmatrix} 1/\lambda_n & 0 & -\lambda_1/\lambda_n^2 \\ 1 & 1 & 1 \\ -1/\lambda_1^2 & -1/\lambda_2^2 & -1/\lambda_n^2 \end{bmatrix}\right) = 0.$$

Many terms cancel in this determinant, and the constraint simplifies to just $\lambda_2^2 = \lambda_1\lambda_n$. Divide the constraint $\lambda_1+(n-2)\lambda_2+\lambda_n = n$ by λ_2 and multiply the constraint $\frac{1}{\lambda_1}+\frac{n-2}{\lambda_2}+\frac{1}{\lambda_n} = n+g$ by λ_2 to obtain

$$\frac{\lambda_1}{\lambda_2} + (n-2) + \frac{\lambda_n}{\lambda_2} = \frac{n}{\lambda_2}, \quad \frac{\lambda_2}{\lambda_1} + (n-2) + \frac{\lambda_2}{\lambda_n} = \lambda_2(n+g).$$

The lefthand sides of each of these constraints are both equal to $\sqrt{\lambda_1/\lambda_n}+\sqrt{\lambda_n/\lambda_1}+(n-2)$ since $\lambda_2 = \sqrt{\lambda_1\lambda_n}$. Consequently,

$$\lambda_2 = \cdots = \lambda_{n-1} = \sqrt{\frac{n}{n+g}}.$$

Observe that we have deduced the sum $\lambda_1 + \lambda_n = n - (n-2)\lambda_2$ and the product $\lambda_1\lambda_n = \lambda_2^2$ so by Viete's formulas λ_1 and λ_n are the larger and smaller roots of

$$z^2 - (n - (n-2)\lambda_2)z + \lambda_2^2 = 0$$

respectively. By applying the quadratic formula and rearranging, we obtain

$$\lambda_1/\lambda_n = 1 + \frac{1}{2}\Big(x(x+4) + (x+2)\sqrt{x(x+4)}\Big) \quad \text{where} \quad x = \frac{g}{\sqrt{1+g/n}+1}.$$

Notice that the expression for λ_1/λ_n is increasing in x, and that $x \leqslant g/2$. Plugging this in gives,

$$\lambda_1/\lambda_n \leqslant 1 + \sqrt{2g + g^2(g^2+16g+80)/64} + g + g^2/8$$

By treating this as a function of $\sqrt{g}$ and applying Taylor's theorem, this gives

$$\begin{aligned} \kappa(\hat{B}) &\leqslant 1 + \sqrt{2g} + g + \frac{5}{8\sqrt{2}}g^{1.5} + \frac{g^2}{4} \\ &\leqslant (1 + \sqrt{g/2} + g/2)^2 \end{aligned} \tag{3.19}$$

as required. □

Lemma 3.3.3.

$$\max\Big(\|\hat{B}\|, \|\hat{B}^{-1}\|\Big) \leqslant 1 + \frac{\Gamma(B)}{2} + \sqrt{\left(1 + \frac{\Gamma(B)}{2}\right)^2 - 1}.$$

Proof. Put $g = \Gamma(\hat{B})$. Here we are maximizing either λ_1 or $1/\lambda_n$ subject to (3.12). Since the constraints are symmetric, we may equivalently think of the maximum values of λ_1 and $1/\lambda_1$ instead, which correspond to just the maximum and minimum values of λ_1. Then by (3.12), we have

$$\det \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ -1/\lambda_1^2 & -1/\lambda_i^2 & -1/\lambda_j^2 \end{bmatrix} = 0,$$

i.e. $\lambda_i = \lambda_j$ for all $i, j > 1$. By plugging this into the constraints and solving for λ_1, we see that the extremal values of λ_1 are the roots of

$$(1 + g/n)z^2 - (2 + g)z + 1 = 0.$$

Notice that the larger root is increasing in n and the smaller root is decreasing in n, so we obtain correct bounds by taking the limit as n goes to infinity. The consequence is that the extremal values of λ_1 lie between the roots of $z^2 - (2 + g)z + 1$. Since the product of the roots is 1, the reciprocal of the smaller root is exactly the larger root, so we finally apply the quadratic formula to obtain

$$\max(\lambda_1, 1/\lambda_1) \leqslant 1 + \frac{\Gamma(B)}{2} - \sqrt{\left(1 + \frac{\Gamma(B)}{2}\right)^2 - 1}$$

as desired. □

3.4 Exact arithmetic

In this section we prove $\Gamma(\cdot)$ converges to zero as Algorithm 1 progresses, first by analyzing the change in $\Gamma(\cdot)$ during a single step of the algorithm and then iterating. As a corollary, we show that one- and two-sided versions of Algorithm 1 quickly converge to the factorizations $A = QT_{\text{acc}}$ and $B = T_{\text{acc}}^* D T_{\text{acc}}$ where Q is nearly orthogonal and D is nearly diagonal. See Theorem 3.4.4 for a precise statement on this convergence.

One-step analysis

As promised, we derive an analogous formula to (3.8) for how for $\Gamma(\cdot)$ changes upon a single iteration of Algorithm 1. Interestingly, although we will only apply this formula when B is positive semi-definite, it holds for general matrices and so we state it as such.

Lemma 3.4.1 (Deterministic update formula)**.** *Fix any $B \in \mathbb{C}^{n \times n}$. Let $J \subset [n]$ be any subset of indices and $S \in \mathbb{C}^{|J| \times |J|}$ be any nonsingular matrix such that $S^* B_{J,J} S$ is diagonal. Denote $T = P_\sigma \begin{bmatrix} S^{-1} & \\ & I \end{bmatrix} P_\sigma^*$ where $\sigma(J) = [k]$. Then*

$$\Gamma((T^{-1})^* B T^{-1}) = \Gamma(B) + \sum_{i,j \in J, i \neq j} b_{ij} \cdot (B^{-1})_{ji}.$$

Proof. Put $P = P_\sigma$. Note that $\Gamma(\cdot)$ is invariant under conjugation by permutation matrices, so

$$\Gamma(T^{-*} B T^{-1}) = \Gamma((P^* T^{-*} P)(P^* B P)(P^* T^{-1} P)).$$

We can express P^*BP and P^*TP as block matrices

$$P^*BP = \begin{bmatrix} B_{JJ} & B_{J,J^c} \\ B_{J^c,J} & B_{J^c,J^c} \end{bmatrix} \quad \& \quad P^*TP = \begin{bmatrix} S^{-1} & \\ & I \end{bmatrix}.$$

For the remainder of the proof, we replace P^*BP with B and P^*TP with T. $\odot$ has lower precedence than regular matrix multiplication, so $M_1M_2 \odot M_3M_4 = (M_1M_2) \odot (M_3M_4)$. Then

$$\begin{aligned} \Gamma(T^{-*}BT^{-1}) - \Gamma(B) &= \operatorname{tr}\big(T^{-*}BT^{-1} \odot TB^{-1}T^*\big) - \operatorname{tr}\big(B \odot B^{-1}\big) \\ &= \operatorname{tr}\left(\begin{bmatrix} S^* & \\ & I \end{bmatrix} B \begin{bmatrix} S & \\ & I \end{bmatrix} \odot \begin{bmatrix} S^{-1} & \\ & I \end{bmatrix} B^{-1} \begin{bmatrix} S^{-*} & \\ & I \end{bmatrix}\right) - \operatorname{tr}\big(B \odot B^{-1}\big) \\ &= \operatorname{tr}\big(S^*B_{JJ}S \odot S^{-1}(B^{-1})_{JJ}S^{-*}\big) - \operatorname{tr}(B_{JJ} \odot (B^{-1})_{JJ}) \end{aligned} \tag{3.20}$$

By assumption, $S^*B_{J,J}S = D$ is diagonal. In particular, we have

$$\operatorname{tr}\big(S^*B_{JJ}S \odot S^{-1}(B^{-1})_{JJ}S^{-*}\big) = \operatorname{tr}\big(S^*B_{JJ}S \cdot S^{-1}(B^{-1})_{JJ}S^{-*}\big) = \operatorname{tr}\big(B_{J,J} \cdot (B^{-1})_{J,J}\big).$$

This gives

$$\begin{aligned} \Gamma(T^*BT) - \Gamma(B) &= \operatorname{tr}\big(B_{JJ} \cdot (B^{-1})_{JJ}\big) - \operatorname{tr}\big(B_{JJ} \odot (B^{-1})_{JJ}\big) \\ &= \sum_{i,j \in J} b_{ij} \cdot (B^{-1})_{ji} - \sum_{j \in J} b_{jj} \cdot (B^{-1})_{jj} \\ &= \sum_{i,j \in J, i \neq j} b_{ij} \cdot (B^{-1})_{ji} \end{aligned} \tag{3.21}$$

as required. □

Completing the analogy of (3.8) to (3.9), we now compute the expected value of the bound from Theorem 3.4.1 under the randomized size k pivot rule.

Lemma 3.4.2 (Expected update)**.** *Fix any $B \in \mathbb{C}^{n \times n}$. Let T be as in Theorem 3.4.1 for a uniformly randomly sampled size k pivot set $J \in \binom{[n]}{k}$. Then*

$$\mathbb{E}\Gamma((T^{-1})^*BT^{-1}) = \left(1 - \frac{k(k-1)}{n(n-1)}\right)\Gamma(B).$$

Proof.

$$\mathbb{E}\left(\sum_{i,j \in J, i \neq j} b_{ij} \cdot (B^{-1})_{ji}\right) = \binom{n}{k}^{-1} \sum_{J \in \binom{[n]}{k}} \sum_{i,j \in J, i \neq j} b_{ij} \cdot (B^{-1})_{ji}.$$

Note for each ordered pair (i,j) that the term $b_{ij}\cdot(B^{-1})_{ji}$ appears exactly $\binom{n-2}{k-2}$ times. Thus

$$\begin{aligned}\mathbb{E}\left(\sum_{i,j\in J, i\neq j} b_{ij}\cdot(B^{-1})_{ji}\right) &= \binom{n}{k}^{-1}\binom{n-2}{k-2}\sum_{i,j\in[n],i\neq j} b_{ij}\cdot(B^{-1})_{ji} \\ &= -\frac{k(k-1)}{n(n-1)}\Gamma(B).\end{aligned} \tag{3.22}$$

□

Iterated analysis

Theorem 3.4.3 (Global linear convergence of Algorithm 1)**.** *Let $A^{(t)}, B^{(t)}$ be the sequence of matrices produced by any instantiation of the the one- and two-sided versions Algorithm 1 respectively with randomized size k pivoting. Then*

$$\begin{aligned}\mathbb{E}\Gamma(B^{(t)}) &= \left(1-\frac{k(k-1)}{n(n-1)}\right)^t\Gamma(B^{(0)}),\\ \mathbb{E}\Gamma(A^{(t)*}A^{(t)}) &= \left(1-\frac{k(k-1)}{n(n-1)}\right)^t\Gamma(A^{(0)*}A^{(0)}).\end{aligned}$$

Proof. Theorem 3.4.2 implies

$$\mathbb{E}\big(\Gamma(B^{(t)})\,|\,B^{(t-1)}\big) = \left(1-\frac{k(k-1)}{n(n-1)}\right)\Gamma(B^{(t-1)}). \tag{3.23}$$

By using induction with the law of iterated expectation, we obtain

$$\begin{aligned}\mathbb{E}\Gamma(B^{(t)}) &= \mathbb{E}\big(\mathbb{E}\big(\Gamma(B^{(t)})\,|\,B^{(t-1)}\big)\big)\\ &= \left(1-\frac{k(k-1)}{n(n-1)}\right)\mathbb{E}\Gamma(B^{(t-1)})\\ &= \left(1-\frac{k(k-1)}{n(n-1)}\right)^t\Gamma(B^{(0)}).\end{aligned} \tag{3.24}$$

□

Remark 5 (Monotonicity and determinism)**.** Unlike $\mathrm{off}(\cdot)$, the potential function $\Gamma(\cdot)$ is *not* strictly monotone, it is only monotone in expectation. If one seeks a deterministic pivoting strategy with monotone convergence, one could adopt a greedy strategy which is analogous to (3.6). Specifically, pick

$$(i,j) = \arg\min_{i,j:i\neq j} \Re\big(b_{ij}(B^{-1})_{ji}\big).$$

With this pivoting rule, the bounds of Theorem 3.4.3 would hold with an inequality $\leqslant$ instead of equality in expectation. We speculate that this could be done efficiently in the setting where one is initially given both B and B^{-1}, and updates both $B^{(t)}$ and $(B^{(t)})^{-1}$ in each iteration.

Remark 6 (Kaczmarz-Kac walk)**.** Consider the special case of the one-sided Algorithm 1, where $k = 2$ and S is either upper or lower triangular. This is precisely the procedure proposed in [Ste25] (where it is dubbed the "Kaczmarz-Kac walk") and concurrently in [DS26b]. With Theorem 3.3.2 to relate $\Gamma(\cdot)$ to $\kappa(\cdot)$, Theorem 3.4.3 confirms the rate of convergence of this method conjectured by Steinerberger [Ste25].

Our main result is a corollary of Theorem 3.4.3, showing the convergence of Algorithm 1.

Corollary 3.4.4. *Let $0 < \delta < 1$. Given positive definite $B \in \mathbb{C}^{n\times n}$, run Algorithm 1 with randomized size k pivoting for*

$$t \geq \frac{n(n-1)}{k(k-1)} \log\left(4n\hat{\kappa}(B)/\delta^2\right)$$

iterations. Let $T_{\text{acc}} = T_{\text{acc}}^{(t)}$, $D = B^{(t)}$ be the values returned by the algorithm. Let $D' = D \odot I$ be just the diagonal entries of D. Then

$$\|B - T_{\text{acc}}^* D' T_{\text{acc}}\| \leqslant \|B\| \frac{\delta}{\rho - \delta}$$

with probability at least $1 - \rho$. Additionally, $B = T_{\text{acc}}^* D T_{\text{acc}}$ exactly where D is close to diagonal in the sense that

$$\mathbb{E}\sqrt{\sum_{i\neq j} \frac{|d_{ij}|^2}{d_{ii}d_{jj}}} \leqslant \delta$$

where d_{ij} is the ij^{th} entry of D. Equivalently, if $B = A^*A$, then the one-sided version returns $Q = A^{(t)}$ forming the exact factorization $A = QT_{\text{acc}}$ where Q is nearly orthogonal in the sense that

$$\mathbb{E}\sqrt{\sum_{i\neq j} \frac{|\langle q_i, q_j\rangle|^2}{\|q_i\|^2 \|q_j\|^2}} \leqslant \delta$$

where q_i is the i^{th} column of Q. Finally, it is possible to execute lines 6-8 so that T_{acc} is ensured to be (special) unitary[4] or (unit) upper/lower triangular, giving any of the factorizations listed in Table 3.1.

Proof. First note that as outlined in Table 3.1, we can apply constraints to S to produce a specific factorization. Note that each listed property of S implies the corresponding property of T in line 7. That is, if S is unitary, T will be unitary, if S is upper triangular, T will be upper triangular, etc. These properties also are closed under multiplication, so the returned transformation T_{acc} will have the corresponding property as well. Picking S with the desired constraint is also always possible: these are just the corresponding decompositions of B_{JJ} and A_J, which always exist.

[4]Special unitary is any unitary matrix U with $\det(U) = 1$.

We now turn to the quantitative statements in the two-sided case. Theorem 3.4.2 implies

$$\mathbb{E}\big(\Gamma(B^{(t)}) \,|\, B^{(t-1)}\big) = \left(1 - \frac{k(k-1)}{n(n-1)}\right)\Gamma(B^{(t-1)})$$

By applying the same equality for smaller values of t and using the law of iterated expectation, we have

$$\begin{aligned}\mathbb{E}\Gamma(B^{(t)}) &= \mathbb{E}\big(\mathbb{E}\big(\Gamma(B^{(t)}) \,|\, B^{(t-1)}\big)\big) \\ &= \left(1 - \frac{k(k-1)}{n(n-1)}\right)\mathbb{E}\Gamma(B^{(t-1)}) \\ &= \left(1 - \frac{k(k-1)}{n(n-1)}\right)^t \Gamma(B^{(0)}).\end{aligned} \tag{3.25}$$

By Theorem 3.4.3, our choice of t, and Theorem 3.3.2 we have

$$\mathbb{E}\Gamma(B^{(t)}) = \left(1 - \frac{k(k-1)}{n(n-1)}\right)^t \Gamma(B^{(0)}) \leq \delta^2/4.$$

By Theorem 3.3.1 and Jensen's inequality, we have

$$\mathbb{E}\sqrt{\sum_{i\neq j} \frac{|d_{ij}|^2}{d_{ii}d_{jj}}} \leq \delta.$$

The statements for the one-sided case follow identically. The first statement in the corollary follows by considering the event

$$\sqrt{\sum_{i\neq j} \frac{|d_{ij}|^2}{d_{ii}d_{jj}}} \leq n\delta,$$

which occurs with probability $1-\rho$ by Markov's inequality. This is equivalent to

$$\left\|D'^{-1/2}T_{\text{acc}}^{-*}BT_{\text{acc}}^{-1}D'^{-1/2} - I\right\|_F \leqslant \delta/\rho$$

which by the triangle inequality and submultiplicativity implies

$$\begin{aligned}\|B - T_{\text{acc}}^* D' T_{\text{acc}}\|_F &\leqslant \|D'^{1/2}T_{\text{acc}}\|^2\delta/\rho \\ &= \|T_{\text{acc}}^* D' T_{\text{acc}}\|\delta/\rho \\ &\leqslant (\|B\| + \|B - T_{\text{acc}}^* D' T_{\text{acc}}\|)\delta/\rho.\end{aligned} \tag{3.26}$$

The desired result follows by rearranging. □

3.5 Finite arithmetic analysis

The previous section showed that every instantiation of Algorithm 1 (i.e. every process for selecting S in line 6) with randomized size k pivoting has the same expected behavior in terms of $\Gamma(\cdot)$. To obtain a convergence result in the presence of round-off error, one needs to assume the updates line 8 and 9 are performed with small mixed forward-backward error (see the diagram (3.27)).

In this setting, $\Gamma(\cdot)$ will play two roles: the first role, as before, is that the convergence of $\Gamma(\cdot) \to 0$ tracks the convergence of the algorithm. The second role is in controlling the accumulation of the error acquired during each iteration.

We let $\widetilde{A}^{(t)}$, $\widetilde{B}^{(t)}$ denote the state of the one- and two-sided algorithm respectively after t iterations are performed in finite precision with randomized size k pivoting (to contrast with $A^{(t)}$, $B^{(t)}$ without tildes computed exactly). The stability of each iteration can be defined in terms of the following commutative diagram, corresponding to a mixed forward-backward guarantee.

$$\begin{array}{ccc} \widetilde{A}^{(t)} & \xrightarrow{\text{floating}} & \widetilde{A}^{(t+1)} \\ \uparrow & & \uparrow \\ \approx & & \approx \\ \downarrow & & \downarrow \\ \widetilde{A}^{(t+1/3)} & \xrightarrow{\text{exact}} & \widetilde{A}^{(t+2/3)} \end{array} \qquad \begin{array}{ccc} \widetilde{B}^{(t)} & \xrightarrow{\text{floating}} & \widetilde{B}^{(t+1)} \\ \uparrow & & \uparrow \\ \approx & & \approx \\ \downarrow & & \downarrow \\ \widetilde{B}^{(t+1/3)} & \xrightarrow{\text{exact}} & \widetilde{B}^{(t+2/3)} \end{array} \tag{3.27}$$

Later in this section, we will need to refer to the state at various iterations and "one-third" iterations. To avoid ambiguity, we always take t to be an integral index (i.e. refers to the state after a complete number of iterations) and s an index in $\frac{1}{3}\mathbb{N}$ (i.e. can refer to an in-between state). The diagram commutes in the following sense: if $\widetilde{A}^{(t+1)}$ is the result of performing one iteration on $\widetilde{A}^{(t)}$ in floating point arithmetic, then we are guaranteed the existence of "intermediate states" $\widetilde{A}^{(t+1/3)}$ which is close to $\widetilde{A}^{(t)}$ and $\widetilde{A}^{(t+2/3)}$ which is close to $\widetilde{A}^{(t+1)}$ such that $\widetilde{A}^{(t+2/3)}$ is the result of applying one iteration to $\widetilde{A}^{(t+1/3)}$ in exact arithmetic. Analogous states exist for the two-sided variant. We define "closeness" by

$$A \approx A' \iff \operatorname{dist}_1(A, A') \leqslant \varepsilon \qquad \& \qquad B \approx B' \iff \operatorname{dist}_2(B, B') \leqslant \varepsilon$$

for pseudo-metrics

$$\operatorname{dist}_1(A, A') = \sup_j \left\| \frac{a_j}{\|a_j\|} - \frac{a'_j}{\|a'_j\|} \right\| \qquad \& \qquad \operatorname{dist}_2(B, B') = \sup_{ij} \left| \frac{b_{ij}}{\sqrt{b_{ii} b_{jj}}} - \frac{b'_{ij}}{\sqrt{b'_{ii} b'_{jj}}} \right|$$

where a_j, a'_j is the jth column of A, A' respectively. Note that we are implicitly assuming here that the diagonal entries of B and B' are positive. Importantly, $\operatorname{dist}_1(A, A') \leqslant \varepsilon$ implies that $\operatorname{dist}_2(A^T A, A'^T A') \leqslant 2\varepsilon + \varepsilon^2 = O(\varepsilon)$ so it suffices to consider only the hypothesis of the two-sided assumption of (3.27).

Control on $\Gamma(\widetilde{B}^{(t)})$

We begin by proving a finite-arithmetic analog of Theorem 3.4.3, i.e., we wish to show $\Gamma(\widetilde{B}^{(t)})$ exhibits linear convergence. In the last section, specifically, (3.23), we showed that the stochastic process $C_{n,k}^{-t}\Gamma(B^{(t)})$ was a martingale where $C_{n,k} = \left(1 - \frac{k(k-1)}{n(n-1)}\right)$. Thus $C^t \to 0$ roughly implied $\Gamma(B^{(t)}) \to 0$ at the same rate. In this section, we define a new martingale and show $C_{n,k}^{-t}\Gamma(\widetilde{B}^{(t)})$ is (nearly) bounded by it with high probability. To define the martingale, we introduce the following random variables.

$$X_t = \frac{\Gamma(\widetilde{B}^{(t+2/3)})}{\Gamma(\widetilde{B}^{(t+1/3)})} \qquad \& \qquad Y_{t_1t_2} = \prod_{t=t_1}^{t_2} X_t$$

where we use the convention $Y_{t_1t_2} = 1$ if $t_2 < t_1$.

Lemma 3.5.1 (Martingale behavior)**.** *For any index t_1, the process*

$$\left\{C_{n,k}^{t_1-1-t} Y_{t_1,t}\right\}_{t=t_1}^{\infty}$$

is a nonnegative martingale.

Proof. Theorem 3.4.2 states that

$$\mathbb{E}\Big(X_t \,|\, \widetilde{B}^{(t+1/3)}\Big) = \left(1 - \frac{k(k-1)}{n(n-1)}\right).$$

Theorem 3.4.1 shows that X_t is a deterministic function of $\widetilde{B}^{(t+1/3)}$ and the randomly chosen pivot $J \in \binom{[n]}{k}$. Since the pivot chosen at each step is independent of previously chosen pivots, one has

$$\mathbb{E}(X_t \,|\, X_{t-1}, \ldots, X_0) = C_{n,k} \implies \mathbb{E}(Y_{t_1,t} \,|\, Y_{t_1,t-1}, \ldots, Y_{t_1,t_1}) = C_{n,k} Y_{t_1,t-1}.$$

□

Our next lemma allows us to control the change in $\Gamma(\cdot)$ between the states $\widetilde{B}^{(t\pm1/3)}$ and the state $\widetilde{B}^{(t)}$.

Lemma 3.5.2 (Perturbation theory for $\Gamma(\cdot)$)**.** *Let B and P be square matrices with positive diagonal entries. Suppose that B is positive definite and that $\operatorname{dist}_2(B, P) \leqslant \varepsilon$ for $\varepsilon < 1/(n\Gamma(B) + n^2)$ then P is also positive definite and*

$$|\Gamma(B) - \Gamma(P)| \leqslant n\varepsilon \cdot \frac{(\Gamma(B) + n)^2}{1 - n\varepsilon(\Gamma(B) + n)}.$$

Proof. Let $X = D_B^{-1/2} B D_B^{-1/2}$ and $Y = D_P^{-1/2} P D_P^{-1/2}$. By definition of $\Gamma(\cdot)$, we have

$$\Gamma(B) = \Gamma(X) = \operatorname{tr}(X^{-1}) - n \quad \& \quad \Gamma(P) = \Gamma(Y) = \operatorname{tr}(Y^{-1}) - n.$$

Consequently, we can express their difference as

$$\begin{aligned} |\Gamma(X) - \Gamma(Y)| = \left|\operatorname{tr}(X^{-1} - Y^{-1})\right| &= \left|\operatorname{tr}\big(X^{-1}(Y - X)Y^{-1}\big)\right| \\ &= \left|\operatorname{tr}\big(Y^{-1}X^{-1}(Y - X)\big)\right|. \end{aligned} \tag{3.28}$$

Since the entries of $Y - X$ are bounded by ε, we can apply Hölder's inequality and Cauchy-Schwarz to obtain

$$\left|\operatorname{tr}\big(Y^{-1}X^{-1}(Y - X)\big)\right| \leqslant \varepsilon \sum_{i,j} \left|(Y^{-1}X^{-1})_{ij}\right| \leqslant \varepsilon \left\|Y^{-1}\right\|_F \left\|X^{-1}\right\|_F.$$

Then by rearranging the triangle inequality

$$\begin{aligned} \left\|Y^{-1}\right\|_F &\leqslant \left\|X^{-1}\right\|_F + \left\|X^{-1} - Y^{-1}\right\|_F \\ &\leqslant \left\|X^{-1}\right\|_F + \left\|X^{-1}\right\|_F \left\|Y^{-1}\right\|_F \|X - Y\|_F \end{aligned} \tag{3.29}$$

one obtains

$$\left\|Y^{-1}\right\|_F \leqslant \frac{\|X^{-1}\|_F}{1 - \|Y - X\|_F \|X^{-1}\|_F} \leqslant \frac{\|X^{-1}\|_F}{1 - n\varepsilon \|X^{-1}\|_F}.$$

Finally, note that the above quantity is monotone in $\|X^{-1}\|_F$ and apply

$$\left\|X^{-1}\right\|_F \leqslant \operatorname{tr}(X^{-1}) = \Gamma(X) + n.$$

Note that this argument applies if Y is replaced with any matrix in the line segment $Y_s = sY + (1 - s)X$. In particular, $\|Y_s^{-1}\|_F$ is finite for each choice of $s \in [0, 1]$. Since Y_t is a continuous path among Hermitian invertible matrices with one end point, X, that is positive definite, we conclude that the other end point, Y, is positive definite as well. □

To stitch these two lemmas together, we essentially argue that by selecting ε sufficiently small, one can ensure that the additive error one obtains from Theorem 3.5.2 between states $\widetilde{B}^{(t-1/3)}$ to $\widetilde{B}^{(t+1/3)}$ does not accumulate too much.

Theorem 3.5.3. *For any time step τ and parameters $\rho, \delta > 0$, if*

$$\varepsilon \leqslant \frac{\delta\rho}{2n(\Gamma(B)\tau\rho^{-1} + \delta + n)^2 \cdot \tau^2}$$

then

$$\Pr\left(\Gamma(\widetilde{B}^{(t)}) \leqslant \Gamma(B)\left(1 - \frac{k(k-1)}{n(n-1)}\right)^t \rho^{-1} + \delta \quad \forall t < \tau\right) \geq 1 - 2\rho.$$

Proof. Set $C = 1 - \frac{n(n-1)}{k(k-1)}$. Let $y = \tau\rho^{-1}$. Let Ω be the event that

$$\Omega \equiv \sup_{0 \leqslant t_1, t_2 < \tau} C^{t_1 - 1 - t_2} Y_{t_1 t_2} \leqslant y$$

and E_t the event that

$$E_t \equiv \Gamma(\widetilde{B}^{(t)}) \leqslant \Gamma(B) Y_{0,t-1} + \delta.$$

We claim that $\Omega \subset E_0 \cap \cdots \cap E_{\tau-1}$. We argue via induction on t. The base case of $\Omega \subset E_0$ follows since E_0 is guaranteed. It thus suffices to argue that $\Omega \cap E_0 \cap \ldots \cap E_t \subset E_{t+1}$. Condition on the events $\Omega, E_0, \ldots, E_t$. Define

$$d_t = \Gamma(\widetilde{B}^{(t+1/3)}) - \Gamma(\widetilde{B}^{(t-1/3)}) \qquad \& \qquad d'_t = \Gamma(\widetilde{B}^{(t)}) - \Gamma(\widetilde{B}^{(t-1/3)})$$

where we adopt the convention $\widetilde{B}^{(-1/3)} = \widetilde{B}^{(0)}$ Then

$$\begin{aligned} \Gamma(\widetilde{B}^{(t+1)}) &= \Gamma(B) Y_{0t} + \sum_{j=0}^{t} d_j Y_{jt} + d'_{t+1} \\ &\leqslant \Gamma(B) Y_{0t} + (t+1) y \sup_{0 \leqslant j \leqslant t} d_j + d'_{t+1}. \end{aligned} \tag{3.30}$$

Theorem 3.5.2 shows how to bound d_i, d'_i in terms of $\Gamma(\widetilde{B}^{(t)})$. Specifically,

$$d_j, d'_j \leqslant 2n\varepsilon \cdot \frac{(\Gamma(\widetilde{B}^{(j)}) + n)^2}{1 - n\varepsilon(\Gamma(\widetilde{B}^{(j)}) + n)} \leqslant 2n\varepsilon \cdot \frac{(\Gamma(B) C^j y + \delta + n)^2}{1 - n\varepsilon(\Gamma(B) C^j y + \delta + n)} \tag{3.31}$$

where we use Ω, E_j in the second inequality. The selection of ε is such that (3.30) with (3.31) implies E_{t+1} occurs, completing the induction step. The consequence is

$$\Pr(E_{\tau-1} \cap \cdots \cap E_0) \geq \Pr(\Omega).$$

Now consider the event

$$\Omega' \equiv \sup_{t < \tau} C^{-1-t} Y_{0t} \leqslant \rho^{-1}.$$

Then

$$\Omega' \cap E_t \implies \Gamma(\widetilde{B}^{(t)}) \leqslant \Gamma(B) C^t \rho^{-1} + \delta.$$

Therefore by union bound,

$$\begin{aligned} \Pr\Big(\Gamma(\widetilde{B}^{(t)}) \leqslant \Gamma(B) C^t \rho^{-1} + \delta \quad \forall t < \tau\Big) &\geq \Pr(E_{\tau-1} \cap \cdots \cap E_0 \cap \Omega') \\ &\geq \Pr(E_{\tau-1} \cap \cdots \cap E_0) - \Pr(\neg\Omega') \\ &\geq \Pr(\Omega) - \Pr(\neg\Omega'). \end{aligned} \tag{3.32}$$

By Theorem 3.5.1, events Ω, Ω' concern the concentration of a nonnegative supermartingale. For each t_1, one has by Ville's inequality that

$$\Pr\left(\sup_{t_2 \geq t_1} C^{t_1-1-t_2} Y_{t_1,t_2} \geq a\right) \leqslant a^{-1}. \tag{3.33}$$

Apply (3.33) for t_1 such that $0 \leq t_1 < \tau$ and $a = y$ to obtain by union bound

$$\Pr(\neg\Omega) \leqslant y^{-1}\tau = \rho.$$

Apply (3.33) for $t_1 = 0$ and $a = \rho^{-1}$ to obtain

$$\Pr(\neg\Omega') \leqslant \rho.$$

Plugging these inequalities into (3.32) gives the final result. □

As we stated at the outset, $\Gamma(\cdot)$ plays two roles: it must converge to 0 to show the algorithm halts, and it also must never see a large intermediate value to ensure the numerical error does not blow up. To these ends, we have two consequences of Theorem 3.5.3. In both, we set

$$\hat{B}^{(t)} = \operatorname{diag}\left(\widetilde{b}_{11}, \ldots, \widetilde{b}_{nn}\right)^{-1/2} \widetilde{B}^{(t)} \operatorname{diag}\left(\widetilde{b}_{11}, \ldots, \widetilde{b}_{nn}\right)^{-1/2}$$

to be the diagonal-normalization of a matrix.

Corollary 3.5.4 (Boundedness of $\|(\hat{B}^{(t)})^{-1}\|$)**.** *For any time step τ and parameters $\rho > 0$, if*

$$\varepsilon \leqslant \frac{\rho^3}{3n^3\|\hat{B}^{-1}\|^2\tau^4}$$

then

$$\Pr\left(\sup_{t<\tau} \|(\hat{B}^{(t)})^{-1}\| \leqslant \|\hat{B}^{-1}\| \cdot n\rho^{-1} + 3\right) \geq 1 - 2\rho.$$

Proof. Apply Theorem 3.3.3 to Theorem 3.5.3, set $\delta = 1$, and use $\left(1 - \frac{k(k-1)}{n(n-1)}\right)^t \leqslant 1$. □

Corollary 3.5.5 (Approximate convergence)**.** *For any $\rho \in (0, 1/2), \delta \in (0, 1/4)$, let*

$$t \geq \frac{n(n-1)}{k(k-1)} \log\left(n\|\hat{B}^{-1}\|/\rho\delta\right).$$

If

$$\varepsilon \leqslant \frac{\delta\rho^3}{3n^3\|\hat{B}^{-1}\|^2\tau^4}$$

then

$$\max\left(\|\hat{B}^{(t)}\|, \|(\hat{B}^{(t)})^{-1}\|\right) \leqslant 1 + 2\sqrt{\delta}$$

with probability $1 - 2\rho$.

Proof. Apply Theorem 3.3.3 to Theorem 3.5.3. □

Orthogonalization is stable

The most general result is that if each iteration of the one-sided algorithm satisfies (3.27), then the overall algorithm stably computes an orthonormal basis of the column-space of $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{C}^{d\times n}$. This can be extended to a basis of the flag $\operatorname{span}(a_1), \operatorname{span}(a_1, a_2), \operatorname{span}(a_1, a_2, a_3), \cdots$. We give a handful of examples of implementations of the algorithm which satisfy (3.27).

We first need to specify a metric on the Grassmanian $\mathrm{Gr}_k(\mathbb{C}^n)$ so that we may quantify the error of the basis output by the algorithm. For subspaces V and W of the same dimension k, let Q_V and Q_W denote orthonormal basis for those spaces respectively. Then we use the following standard metric, sometimes referred to as the *gap-metric*, which has a few equivalent expressions [QZL05],

$$\operatorname{dist}(V, W) = \|Q_V Q_V^* - Q_W Q_W^*\| = \inf_R \|Q_V - Q_W R\| = \sqrt{1 - \sigma_k(Q_V^* Q_W)^2} = \sin(\theta(V, W)),$$

where θ is the largest principle angle between V and W.

Lemma 3.5.6 (Perturbation theory for subspaces)**.** *Let $\hat{A}, \hat{A}' \in \mathbb{C}^{d\times n}$ have unit length columns. Then*

$$\operatorname{dist}(\operatorname{Col}(\hat{A}), \operatorname{Col}(\hat{A}')) \leqslant \frac{\sqrt{n}\operatorname{dist}_1(\hat{A}, \hat{A}')}{\max(\sigma_n(\hat{A}), \sigma_n(\hat{A}'))}.$$

Proof. Let $\hat{A} = QR$ be the QR decomposition of $\hat{A}$. Then

$$\operatorname{dist}(\operatorname{Col}(A), \operatorname{Col}(A')) \leqslant \left\| Q - \hat{A}' R^{-1} \right\| = \left\| (\hat{A} - \hat{A}') R^{-1} \right\| \leqslant \|\hat{A} - \hat{A}'\| \left\| R^{-1} \right\| = \frac{\|\hat{A} - \hat{A}'\|}{\sigma_n(\hat{A})}. \tag{3.34}$$

Now,

$$\|\hat{A} - \hat{A}'\| = \sup_y \frac{\|(\hat{A} - \hat{A}')y\|}{\|x\| \|y\|} \leqslant \sqrt{n} \sup_y \frac{\|(\hat{A} - \hat{A}')y\|}{\|x\| \|y\|_1} = \sqrt{n}\operatorname{dist}_1(\hat{A}, \hat{A}').$$

Finally, the metric is symmetric so we may swap the roles of A and A'. □

Theorem 3.5.7. *Say one is given $A \in \mathbb{C}^{d\times n}$. Let $\hat{A} = A \operatorname{diag}(\|a_1\|, \ldots, \|a_n\|)^{-1}$ normalize the columns of A. Let $Q = A^{(\tau)}$ be the result of running the one-sided Algorithm 1 with randomized size k pivoting on $A \in \mathbb{C}^{d\times n}$ with full column rank for*

$$\tau = O\left(n^2 k^{-2} \log\left(\frac{n}{\sigma_n(\hat{A})\delta} \right) \right)$$

iterations. Assume each iteration is computed stably in the sense of (3.27) *for*

$$\varepsilon \leqslant \frac{\delta \sigma_n(\hat{A})^4}{n^6 \tau^4}.$$

Then with probability $1 - 2/n$,

$$\|Q^* Q - I\| \leqslant 3\sqrt{\delta} \qquad \& \qquad \operatorname{dist}(\operatorname{Col}(Q), \operatorname{Col}(A)) \leqslant \delta.$$

Proof. Theorem 3.5.5 immediately implies $\|Q^*Q - I\| \leqslant \delta$. By the triangle inequality,

$$\begin{aligned}\operatorname{dist}(\operatorname{Col}(A), \operatorname{Col}(\widetilde{A}^{(\tau)})) \leqslant \sum_{t=1}^{\tau} \Big(& \operatorname{dist}(\operatorname{Col}(\widetilde{A}^{(t-1)}), \operatorname{Col}(\widetilde{A}^{(t-2/3)})) \\ & + \operatorname{dist}(\operatorname{Col}(\widetilde{A}^{(t-2/3)}), \operatorname{Col}(\widetilde{A}^{(t-1/3)})) \\ & + \operatorname{dist}(\operatorname{Col}(\widetilde{A}^{(t-1/3)}), \operatorname{Col}(\widetilde{A}^{(t)})) \Big).\end{aligned} \tag{3.35}$$

Since $\widetilde{A}^{(t-1/3)}$ is exactly a column operation applied to $\widetilde{A}^{(t-2/3)}$, the middle term in the summand vanishes. Note that one may replace $\widetilde{A}^{(s)}$ with its column-normalization $\hat{A}^{(s)}$. The first and last term are bounded by Theorem 3.5.6:

$$\operatorname{dist}(\operatorname{Col}(A), \operatorname{Col}(\widetilde{A}^{(\tau)})) \leqslant \tau \sup_{t \in [\tau]} \left(\frac{\sqrt{n}\, \operatorname{dist}_1(\hat{A}^{(t-1)}, \hat{A}^{(t-2/3)})}{\max(\sigma_n(\hat{A}^{(t-1)}), \sigma_n(\hat{A}^{(t-2/3)}))} + \frac{\sqrt{n}\, \operatorname{dist}_1(\hat{A}^{(t-1/3)}, \hat{A}^{(t)})}{\max(\sigma_n(\hat{A}^{(t-2/3)}), \sigma_n(\hat{A}^{(t)}))} \right).$$

But the distances $\operatorname{dist}_1(\cdot, \cdot)$ in the numerator are bounded by ε by assumption. Thus we have

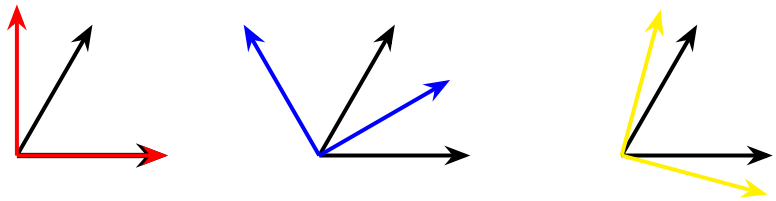
$$\operatorname{dist}(\operatorname{Col}(A), \operatorname{Col}(\widetilde{A}^{(\tau)})) \leqslant \frac{2\tau\sqrt{n}}{\inf_{s \leqslant \tau} \sigma_n(\hat{A}^{(s)})} \cdot \varepsilon.$$

Theorem 3.5.4 shows that

$$\left(\inf_{s \leqslant \tau} \sigma_n(\hat{A}^{(s)}) \right)^{-2} \leqslant \sigma_n(\hat{A})^{-2} n^2 + 3$$

with probability $1 - 2/n$. The selection of ε gives the desired result. □

Remark 7. The following update rules orthogonalize a pair of unit vectors and can be performed in floating point arithmetic with unit round-off $\mathbf{u}$ to satisfy (3.27) with $\varepsilon = O(n)\mathbf{u}$. We describe them for $J = \{1, 2\}$ with $\alpha = \langle a_1, a_2 \rangle$.



The rule depicted the left (red) corresonds to Gram-Schmidt:

$$\begin{bmatrix} a_1 & a_2 \end{bmatrix} \leftarrow \begin{bmatrix} a_1 & \frac{a_2 - \alpha a_1}{\sqrt{1-\alpha^2}} \end{bmatrix}. \tag{3.36}$$

The rule depicted in the middle (blue) is a normalized version of taking the SVD:

$$\begin{bmatrix} a_1 & a_2 \end{bmatrix} \leftarrow \begin{bmatrix} \frac{a_1 + a_2}{\sqrt{2+2\alpha}} & \frac{a_1 - a_2}{\sqrt{2-2\alpha}} \end{bmatrix}. \tag{3.37}$$

The rule depicted on the right (yellow) is a rule that treats the inputs symmetrically, and is equivalent to applying the rule depicted in the middle twice in a row:

$$\begin{bmatrix} a_1 & a_2 \end{bmatrix} \leftarrow \begin{bmatrix} \frac{(\sqrt{2-2\alpha}+\sqrt{2+2\alpha})a_1+(\sqrt{2-2\alpha}-\sqrt{2+2\alpha})a_2}{\left\|(\sqrt{2-2\alpha}+\sqrt{2+2\alpha})a_1+(\sqrt{2-2\alpha}-\sqrt{2+2\alpha})a_2\right\|} & \frac{(\sqrt{2-2\alpha}-\sqrt{2+2\alpha})a_1+(\sqrt{2-2\alpha}+\sqrt{2+2\alpha})a_2}{\left\|(\sqrt{2-2\alpha}-\sqrt{2+2\alpha})a_1+(\sqrt{2-2\alpha}+\sqrt{2+2\alpha})a_2\right\|}. \end{bmatrix}. \tag{3.38}$$

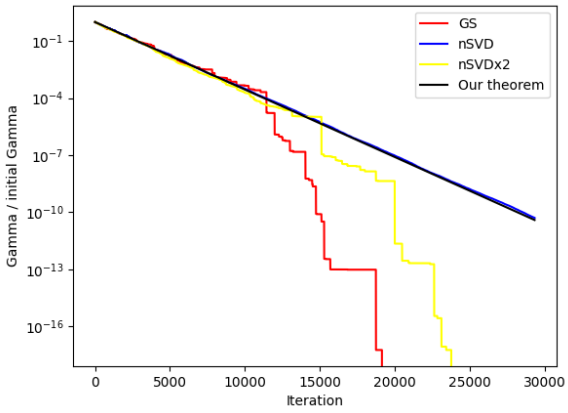


Figure 3.1: Demonstration of evolution of $\frac{\Gamma((A^{(t)})^* A^{(t)})}{\Gamma((A^{(0)})^* A^{(0)})}$ as iterations are performed according to each of the three rules depicted in Remark 7. Rule (3.36) is labeled GS in red. Rule (3.37) is labeled nSVD in blue. Rule (3.38) is labeled nSVDx2 in yellow. The plot for is a single trial applied to a 50x50 matrix with independent Haar unit vector columns. An interesting future line of inquery is understanding why nSVD seems to tightly hug the expected value, whereas GS and nSVD seem to have larger variance or thicker tails leading to a large probability that the observed behavior is better than the expectation.

Remark 8 (The QR decomposition)**.** One can strengthen the result of Theorem 3.5.7 if one implements the update $A_J \leftarrow A_J S$ by performing modified Gram-Schmidt to A_J. The returned basis $Q = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix}$ approximates the flag:

$$\operatorname{dist}(\operatorname{span}(q_1, \ldots, q_m), \operatorname{span}(a_1, \ldots, a_m)) \leqslant \delta$$

for all $m \leqslant n$. To see this, note that when using MGS in this way, the ordering of the pivot indices $i_1 < \cdots < i_k$ in line 5 ensures that each column is completely unaffected by columns to their right. Because of this, the algorithm using MGS run on input $\begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$ can be seen as simulating the same algorithm on all the inputs

$$\begin{bmatrix} a_1 \end{bmatrix} \quad \& \quad \begin{bmatrix} a_1 & a_2 \end{bmatrix} \quad \& \quad \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix}$$

with a certain number of no-ops when the pivot J (partially) falls outside the index set. However, these no-ops exactly counteract the reduced number of iterations required by the

smaller input size for convergence: just replace k with the random variable $|J \cap \{1, \cdots, m\}|$. The result would then follow by Theorem 3.5.7.

Diagonalization and SVD are stable

Demmel and Veselić study the stability of the Jacobi eigenvalue algorithm with an arbitrary pivoting rule. Their results make two assumptions. The first is that the pivoting rule is such that the method converges after some specified number of steps. The second is that the diagonal-normalized condition number of any iterate remains bounded. Under our randomized pivoting rule, both properties are satisfied. We restate the results of [DV92] concerning eigenvalue and singular value error in terms of our notation.

Theorem 3.5.8 ([DV92, Corollary 3.2])**.** *Given positive definite $B \in \mathbb{C}^{n\times n}$, let $\widetilde{B}^{(t)}$ be the t^{th} iterate of the two-sided Algorithm 1 using a Givens rotation in step 6 (i.e. use $k = 2$ and follow the Jacobi eigenvalue algorithm) with unit round-off $\mathbf{u}$. Then for each τ,*

$$\sup_{j\in[n]} \frac{\left|\lambda_j(B) - \widetilde{\lambda}_j^{(\tau)}\right|}{\lambda_j(B)} \leqslant O\Big(\sqrt{n}\tau\mathbf{u} + n\,\mathrm{dist}_2(\widetilde{B}^{(\tau)}, I)\Big) \sup_{t\leqslant\tau} \hat{\kappa}(\widetilde{B}^{(t)})$$

where $\widetilde{\lambda}_j^{(\tau)}$ is the jth largest diagonal entry of $\widetilde{B}^{(\tau)}$.

Theorem 3.5.9 ([DV92, Corollary 4.2])**.** *Given positive definite $A \in \mathbb{C}^{n\times n}$, let $\widetilde{A}^{(t)}$ be the t^{th} iterate of the one-sided Algorithm 1 using a Givens rotation in step 6 (i.e. use $k = 2$ and follow the Jacobi eigenvalue algorithm) with unit round-off $\mathbf{u}$. Then for each τ,*

$$\sup_{j\in[n]} \frac{\left|\sigma_j(B) - \widetilde{\sigma}_j^{(\tau)}\right|}{\sigma_j(B)} \leqslant O\big(\tau\mathbf{u} + n^2\mathbf{u} + n\,\mathrm{dist}_2(A^{(\tau)*}A^{(\tau)}, I)\big) \sup_{t\leqslant\tau} \hat{\kappa}(\widetilde{A}^{(t)*}\widetilde{A}^{(t)})$$

where $\widetilde{\sigma}_j^{(\tau)}$ is the length of the jth longest column of $\widetilde{A}^{(\tau)}$, the infimum is taken over orthogonal matrices, and $\widetilde{B}^{(t)} = \widetilde{A}^{(t)*}\widetilde{A}^{(t)}$.

Our results allow us to provide a concrete upper bound for randomized size 2 pivoting.

Corollary 3.5.10. *When using a randomized size 2 pivoting rule in the above theorems with ε as in Theorem 3.5.5, taking*

$$\tau = O(n^2 \log(n\hat{\kappa}(B)/\delta))$$

iterations ensures

$$\mathrm{dist}_2(\widetilde{B}^{(t)}, I) \leqslant \sqrt{\delta} \qquad \textit{and} \qquad \sup_{t\leqslant\tau} \hat{\kappa}(\widetilde{B}^{(t)}) \leqslant \kappa(\widetilde{B})n^3 + 3n.$$

with probability $1 - O(1/n)$

Proof. The bound on $\operatorname{dist}_2(\widetilde{B}^{(t)}, I)$ follows from Theorem 3.5.5. The bound on $\sup \hat{\kappa}(\cdot)$ follows from Theorem 3.5.4 along with the observation that $\|\hat{B}\|^{(t)} \leqslant n$. □

Similar bounds for the eigenvectors and singular vectors can also be found in [DV92] (see their Theorems 3.3 and 4.3). We omit repeating the precise statements, but note that the application of Theorem 3.5.10 provides concrete bounds analogous to the bounds for eigenvalues and singular values.

Chapter 4

Spectral sums and rational approximations

While the algorithms of Chapter 3 for computing matrix factorizations were theoretically interesting, they are not practical due to the inefficiencies of random pivoting. Here we shift to exploring efficient algorithms for computing spectral sums.

4.1 Introduction

In large-scale data analysis and numerical linear algebra, extracting spectral information without computing a full eigendecomposition is a fundamental challenge. More concretely, we consider the problem of computing a spectral sum, i.e., for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, computing

$$\operatorname{tr}(f(\boldsymbol{A})) = \sum_{i=1}^{n} f(\lambda_i(\boldsymbol{A})),$$

without computing a full eigendecomposition.

We provide a general framework for computing spectral sums, leveraging rational approximations of functions and stochastic trace estimation. Rational function approximations generally have better convergence properties for non-smooth functions, such as an indicator function or an absolute value function. However, evaluating rational matrix functions requires solving linear systems, which can be a bottleneck to computation time. Thus there is a tradeoff between the degree of the approximation and the time it takes to evaluate the approximation. Historically, polynomial approximations have been preferred: although one needs a higher degree for a desired error rate, evaluating $p(\boldsymbol{A})\boldsymbol{y}$ for a polynomial p only requires matrix vector products with $\boldsymbol{A}$.

In some instances though, when fast solvers are available, this tradeoff favors rational approximations. For example, for structured matrices such as HODLR (hierarchical off-diagonal low rank) or Toeplitz matrices, we always have fast solvers available, so rational approximations are favorable. With advances in linear system solvers, such as [DS26a], we

should sometimes favor rational approximations in general; this is largely problem dependent, depending on how quickly we can solve the desired systems. In other instances, by starting with a rational approximation and then using polynomials to approximate inverses, one can derive a near-optimal degree polynomial approximation.

We explore this framework for two applications: spectral density estimation, via eigenvalue counts for intervals, and Schatten-1 norm estimation. Given a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the spectral density measure is

$$s_{\boldsymbol{A}}(x) = \sum_{i=1}^{n} \frac{1}{n} \delta(x - \lambda_i(\boldsymbol{A})) \tag{4.1}$$

where $\delta(x)$ is a Dirac mass at 0. Approximating the spectral density has long been an important problem in chemistry and physics [DC70], [Tur88], [DS93], and has also found more recent applications in data analysis and machine learning (see for example [YGKM20] or [RL18]). Given local estimations of eigenvalues in intervals, one can construct a spectral density estimation; in this way, a related problem is to approximate the number of eigenvalues in a given interval I. Letting $\mathbf{1}_I$ be the indicator of the interval I, we can write

$$\text{number of eigenvalues in } I = \sum_{i=1}^{n} \mathbf{1}_I(\lambda_i(A)) = \operatorname{tr} \mathbf{1}_I(\boldsymbol{A}).$$

The *Schatten-1 norm* of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, also known as the *nuclear norm* or the *trace norm*, is

$$\|\boldsymbol{A}\|_* := \sum_{i=1}^{d} \sigma_i(\boldsymbol{A}). \tag{4.2}$$

This is a quantity of interest in many fields: in matrix completion algorithms, it is used as a substitute for matrix rank [JNS13], [CR12], and it is used for applications in theoretical chemistry and differential privacy [Gut01], [HLM12].

Throughout, we assume we have taken the preprocessing step of normalizing the spectral radius of $\mathbf{A}$, so $\|\mathbf{A}\| \leq 1$. We note this is standard among spectral density estimation algorithms, and this can be done in $\tilde{O}(nnz(\boldsymbol{A}))$ time using power methods or Lanczos methods [MM15], [KW92]. We will also assume $\|\boldsymbol{A}\| \geq 1/2$, which holds by the same arguments.

Our contributions

On the approximation theory side, we give an explicitly computable rational approximation of a step function, not involving any elliptic integrals, unlike the well-known Zolotarev functions. The starting point is a rational approximation from [PP11]; we then analyze many different properties of this function. In particular, we give a self-contained analysis of the error introduced by numerically approximating the poles. We then give an explicit expression for the partial fraction decomposition, which lends itself well to parallelizability and simplifies our later error analysis.

We apply this approximation to two spectral problems: spectral density estimation and Schatten-1 norm approximation. For the former, we show how to derive a near-optimal degree polynomial approximation of a step function from our rational approximation, and leverage this for a spectral density estimation algorithm with local guarantees, the first of its kind. For the Schatten-1 norm approximation, we use our approximation of a step function to give a rational approximation of the square root function. With the use of state of the art fast solvers, this yields an algorithm for Schatten-1 norm approximation with the best known dependency on n, with only a small trade off in the dependency on ε.

Main results

Spectral density estimation. For a symmetric matrix $\boldsymbol{A}$, we give a spectral density estimation satisfying both a Wasserstein distance guarantee and local "histogram" guarantees for the number of eigenvalues in a stochastically chosen partitioning of the spectral windows (see Theorem 4.4.7 for details). We also gives an algorithm to estimate the number of eigenvalues in a fixed target interval (see Theorem 4.4.6).

Theorem 4.1.1 (Informally)**.** *Fix $\varepsilon > 0$. Let $\boldsymbol{A}$ be a symmetric matrix normalized as above. With $\tilde{O}(1/\varepsilon^4)$ matrix vector products with $\boldsymbol{A}$, with probability at least 99/100 Algorithm 2 returns a partitioning of $[-1, 1]$ into intervals I_t of width less than or equal to ε and estimates $\tilde{k}_t$ such that*

$$(1-\varepsilon)k_t - \varepsilon^2(k_{t-1} + k_t + k_{t+1}) \leq \tilde{k}_t \leq (1+\varepsilon)k_t + \varepsilon(k_{t-1} + k_t + k_{t+1})$$

where k_t is the true number of eigenvalues of $\boldsymbol{A}$ in I_t. Additionally, we can define the probability measure $\tilde{s}_{\boldsymbol{A}}$ with point masses at the midpoints of each I_t with weight proportional to $\tilde{k}_t$. Then $\mathcal{W}(s_{\boldsymbol{A}}, \tilde{s}_{\boldsymbol{A}}) \leq \varepsilon$.

For concreteness, we give an example of a measure for which a Wasserstein distance guarantee does not imply histogram guarantees. Since we assume ε is a small constant and each eigenvalue corresponds to an atom with mass $1/n$, these examples are not hard to find.

More specifically, we give two measures with Wasserstein distance less than ε^4 where one has mass within 3ε of 0, and the other does not. The choice of ε^4 is due to the fact that if we run the well-known stochastic Lanczos quadrature (SLQ) algorithm for spectral density estimation with $\tilde{O}(1/\varepsilon^4)$ matrix vector products, we get a measure with a Wasserstein distance bound of ε^4 to the true spectral density. This example is not meant to compare this algorithm with SLQ, but rather to compare the *guarantees* we can give on these algorithms.

Example. Fix $\varepsilon < 1/10$ and suppose $n > 30/\varepsilon^3$. Consider a measure $s_{\boldsymbol{A}}$ corresponding to 10 eigenvalues at 0 and the rest in $(-1, -3\varepsilon) \cup (3\varepsilon, 1)$. Now consider a measure μ corresponding to same eigenvalues in $(-1, -3\varepsilon) \cup (3\varepsilon, 1)$ and the 10 eigenvalues at 0 now moved to 10ε (see below for a visualization).
By our choice of n, $\mathcal{W}(s_{\boldsymbol{A}}, \mu) \leq \frac{10}{n} \cdot 3\varepsilon < \varepsilon^4$, but μ will clearly not satisfy a local guarantee near 0. In contrast, our algorithm will return $\tilde{k} > 8$ for the interval containing 0.

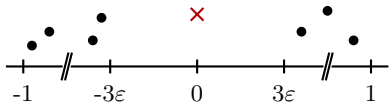


Figure 4.1: Measure $s_{\boldsymbol{A}}$ with mass at 0.

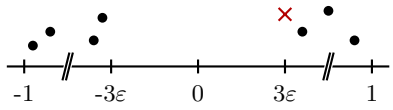


Figure 4.2: Measure μ with no mass at 0.

Schatten-1 norm approximation. We give an algorithm for Schatten-1 norm estimation. We state our result only for the dense case here; see Table 4.1 for optimized runtimes for various sparsity cases.

Theorem 4.1.2. *Assuming $d = n$, $nnz(\boldsymbol{A}) = n^2$, Algorithm 3 yields a $(1 \pm \varepsilon)$ approximation of the Schatten-1 norm with probability 99/100 in time $\tilde{O}(n^2/\varepsilon^{1.5})$.*

We compare this result to several other algorithms for Schatten-1 norm approximation of a dense matrix, specifically those of [DS26a] with runtime $O(n^2/\varepsilon^5)$ [MNSUW17] with runtime of $\tilde{O}(n^{2.18}/\varepsilon^3)$, and [BJMMR25] with runtime of $\tilde{O}(n^{2.5}/\varepsilon)$. In the case of [BJMMR25], we note that they restrict only to matrix-vector products. All of these algorithms first approximate a spectral density estimation in some form, then use this to approximate the Schatten-1 norm, while we directly approximate the Schatten-1 norm. Note these are all probabilistic algorithms (ours included).

Technical overview

Approximation of a step function: We first approximate the step function $\mathbf{1}_{[0,\infty)}$ on some symmetric interval $[-\beta, \beta]$. Our starting point is the explicit construction of a rational approximation of a step function, $s(x)$, given by [PP11]. We will want to take a partial fraction decomposition of $s(x)$, so we need explicit expressions for the poles of $s(x)$. We cannot explicitly factor the denominator of $s(x)$, but we can compute highly accurate numerically approximations of the poles very efficiently (see Theorem 4.3.5). Once we have explicit approximate poles, we use these to build an approximation $\hat{s}(x)$ of $s(x)$. We bound the error introduced by using approximate poles in Theorem 4.3.6.

Structure of applications: We use this explicit, rational approximation of a step function for the aforementioned applications. In each application, we follow a simple set of steps:

- Use step function approximation to approximate some function of interest.
- Use some choice of linear system solver to make rational approximation tractable.
- Perform stochastic approximation of the trace.

For spectral density estimation, the function of interest is an indicator of an interval, and we use Chebychev semi-iteration to solve the linear systems and Hutch++ (see Theorem 4.2.4

for definition) to approximate the trace. For Schatten-1 norm, the function of interest is the square root, we use the fast solvers of [DS26a] to solve the linear systems, and we use variance-reduced trace estimation (similar to Hutch++).

Remark 9. We comment on the different choices for each application. On the linear system solvers: the runtime for solvers using Chebychev semi-iteration is dependent on the condition number, but the runtime of [DS26a] depends on *average condition number* (see Theorem 4.5.1 for details). In the Schatten-1 norm setting, we can provide a much better bound on the average condition number than on the condition number. For spectral density estimation though, we cannot reasonably improve the average condition number without prior knowledge about the positioning of eigenvalues (which is what we are trying to approximate). Additionally, our linear systems across different intervals will *all* have the same Krylov subspace (for a fixed vector), so using polynomial methods will actually be quite beneficial for our spectral density algorithm.

On trace estimation, both methods roughly follow the same strategy. In both cases, part of the method is to construct a low rank approximation of our target matrix. In the Schatten-1 norm setting, we can build a low-rank approximation using Nystrom approximations (see Section 4.5 for details). For spectral density estimation, this approach for low-rank approximation is not an option due to certain properties of the sign function (in particular, the sign function is not operator monotone; see [PMM25] for details).

For each step, we are trying to choose the optimal strategy available to us; for these applications they happen to be different.

Prior work

Rational functions in numerical linear algebra: There has been considerable interest in using rational functions for tasks in numerical linear algebra, considering the superior approximation properties of rational functions over polynomials. Most of these have been centered around Zolotarev functions, optimal degree rational approximations of the sign function [Zol77]. For an extensive overview of rational approximations in linear algebra, see [NT25]. One key issue with rational approximations is the need to solve linear systems to compute a rational approximation. We mention several results and how they address this issue.

Zolotarev approximations were used in [JS19] to perform spectral filtering for principal component projection. Their work is similar to the ideas advanced here; they use asymmetric SVRG to approximately solve systems. They also use Zolotarev functions to construct a rational approximation of $\sqrt{x}$ and apply this to a matrix. Zolotarev functions were also used in [NF16] as part of an algorithm to find the polar decomposition of a matrix $\boldsymbol{A}$. They use a QR factorization to avoid solving a linear system (see Lemma 4.1 of [NF16] for details). The FEAST method for solving eigenproblems also uses Zolotarev approximations (see [TP14] and [GPTV15] for details). Other works implicitly use rational approximations through

quadrature of contour integrals. This method was proposed in [HHT08] to compute various matrix functions.

Spectral density and Schatten-1 norm estimation: For spectral density algorithms, polynomial methods have been historically more preferred (see [LSY16] for an overview of the most popular methods). Perhaps the most prominent spectral density estimation algorithm is stochastic Lanczos quadrature (SLQ), which is rooted in polynomial methods. With $O(1/\varepsilon)$ matrix-vector products, the spectral density estimation $\tilde{s}_{\boldsymbol{A}}$ from SLQ satisfies

$$\mathcal{W}(s_{\boldsymbol{A}}, \tilde{s}_{\boldsymbol{A}}) \leq \varepsilon$$

with high probability for sufficiently large n, where $s_{\boldsymbol{A}}$ is the true spectral density [CTU21]. This bound was tightened for $\boldsymbol{A}$ with decaying singular values in [BJMMR25]. Current results only give guarantees on the Wasserstein distance, but we expect SLQ can satisfy much stronger guarantees.

The core idea of our spectral density algorithm is to do some spectral filtering for specific regions of $[-\|\boldsymbol{A}\|, \|\boldsymbol{A}\|]$ where $\|\boldsymbol{A}\|$ is the spectral radius. This idea is explicitly suggested in [DPS16], where they discuss both polynomial and rational approximations of bump functions to perform spectral filtering. In [MNSUW17], Musco et al. implement this idea to construct a spectral density estimation satisfying *histogram* guarantees, meaning for certain intervals, they approximate the mass within that interval with relative guarantees. Although the measure they construct does not satisfy a Wasserstein distance bound, it can be used to approximate spectral sums, in particular the Schatten-1 norm. Interestingly, their approach can be seen as using a rational approximation of a step function, but with only one real pole. They use stochastic gradient descent to solve the arising systems.

We also mention that the spectral density algorithm in [BJMMR25] can be used to approximate the Schatten-1 norm; they directly estimate all of the singular values and use this to approximate the Schatten-1 norm.

4.2 Preliminaries

Notation

We denote matrices and vectors in bold, $\boldsymbol{A}$ and $\boldsymbol{v}$. For a matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}\|_F$ denote the spectral norm and the Frobenius norm, respectively. For a vector $\boldsymbol{v}$ and a positive definite matrix $\boldsymbol{M}$, $\|x\|_M = \boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x}$, i.e., the norm induced by $\boldsymbol{M}$. For a symmetric matrix $\boldsymbol{A}$, we let $\lambda_1(\boldsymbol{A}) \geq \ldots \geq \lambda_n(\boldsymbol{A})$ denote the eigenvalues of $\boldsymbol{A}$ and let $\sigma(\boldsymbol{A})$ denote the set of all eigenvalues. For any matrix $\boldsymbol{A}$, let $\sigma_i(\boldsymbol{A})$ be the singular values of $\boldsymbol{A}$. As defined above, we let $\|\boldsymbol{A}\|_* = \sum_{i=1}^{d} \sigma_i(\boldsymbol{A})$, i.e., the Schatten-1 norm or the nuclear norm, and let $\kappa(\boldsymbol{A}) = \sigma_1(\boldsymbol{A})/\sigma_d(\boldsymbol{A})$. $O(\cdot)$ hides constants, and $\tilde{O}(\cdot)$ hide constants and polylog factors. For an interval I, we let $\mathbf{1}_I(x)$ be the indicator function of that interval.

Solving linear systems

Since we are using rational approximations, we will need ways to solve linear systems. If our matrix $\boldsymbol{A}$ has some structure that makes solving the arising linear systems easy, we can take advantage of that. If not, we offer two alternatives: polynomial methods with Chebychev semi-iteration or near-optimal methods from [**DS26**].

Chebychev semi-iteration

For a positive definite matrix $\boldsymbol{M}$, the following proposition gives an explicit iterative scheme for approximating the action of $\boldsymbol{M}^{-1}\boldsymbol{y}$ by a degree-k polynomial $q_k(\boldsymbol{M})\boldsymbol{y}$. A key point is that the induced polynomial depends only on k and the *spectral interval*, and it does *not* depend of the input vector $\boldsymbol{y}$. Note the vector-independence means we replace each linear system appearing in $f(\boldsymbol{A})$ for a rational function f with a polynomial approximation, essentially rendering a polynomial approximation of f. We highlight this for two reasons: (1) this effectively becomes a polynomial approximation method, and should be compared to the (many) known polynomial methods, and (2) this makes the trace estimation analysis easier. Stochastic trace estimation methods are analyzed as querying a single fixed matrix across many matrix-vector products. Since this method uses a fixed polynomial independent of $\boldsymbol{y}$, we fit this model

Methods like conjugate gradient may be preferable in practice for solving the system, but its approximation of $\boldsymbol{M}^{-1}\boldsymbol{y}$ depends on the vector $\boldsymbol{y}$. We do not consider this approach here.

Theorem 4.2.1 (Chebyshev semi-iteration)**.** *Let $\boldsymbol{M}$ be a positive definite matrix with condition number κ. Given $\tau > 0$, there exists an explicit polynomial q_τ with degree at most $O(\sqrt{\kappa}\log(1/\tau))$ such that*

$$(1-\tau)\boldsymbol{M}^{-1} \preceq q_\tau(\boldsymbol{M}) \preceq (1+\tau)\boldsymbol{M}^{-1}.$$

Alternative methods: near-optimal linear system solvers

One can also opt for more "high-tech" linear system solvers, for example [DS26a]. This methods combines preconditioning, matrix sketching, low rank update formulas to get a near-optimal linear system solver. We note that one advantage of these methods is that they are not dependent on the condition number $\kappa(\boldsymbol{A}) = \frac{\sigma_d(\boldsymbol{A})}{\sigma_1(\boldsymbol{A})}$, but rather on the *average* condition number of a regularized matrix (defined below). We state the specific result that we will use in Section 4.5.

Theorem 4.2.2 (Modified [**DS26**])**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\boldsymbol{v} \in \mathbb{R}^d$, and $\lambda > 0$. For any $\varepsilon_L > 0$, there exists a randomized algorithm that returns a vector $\tilde{\boldsymbol{x}}$ satisfying*

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|^2_{\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I}} \le \varepsilon_L \|\boldsymbol{x}^*\|^2_{\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I}}, \tag{4.3}$$

with high probability, where $\boldsymbol{x}^* = (\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{v}$. *The algorithm runs in time*

$$\tilde{O}\big((nnz(\boldsymbol{A}) + d^2 \bar{\kappa}_{0,1}(\boldsymbol{A}_\lambda)) \log(1/\varepsilon_L)\big),$$

where $\boldsymbol{A}_\lambda \in \mathbb{R}^{(n+d)\times d}$ *is the augmented matrix formed by appending* $\sqrt{\lambda}\boldsymbol{I} \in \mathbb{R}^{d\times d}$ *to the rows of* $\boldsymbol{A}$, *and for any* $\boldsymbol{M} \in \mathbb{R}^{n\times d}$, $\bar{\kappa}_{0,1}(\boldsymbol{M}) := \frac{1}{d}\sum_{i=1}^{d} \frac{\sigma_i(\boldsymbol{M})}{\sigma_d(\boldsymbol{M})}$.

Krylov methods

As we will see in Section 4.4, many of the linear systems that we solve will have a shared right-hand side. We argue in this section that if one intends to approximate the solution of all of these linear systems using polynomial methods, such as Chebyshev semi-iteration, then one may save time by using a shared polynomial Krylov subspace. Notably, polynomial Krylov subspace methods have become a standard tool in numerical linear algebra for large-scale eigenproblems and linear systems [Saa03; Saa11]. We summarize our main claim in the following proposition.

Proposition 4.2.3. *Let* $p_1, p_2, \cdots, p_k$ *be polynomials of degree at most* $m-1$. *Then for a fixed vector* $\boldsymbol{v}$, *one may compute* $p_j(\boldsymbol{A})\boldsymbol{v}$ *for each* $j = 1, 2, \cdots, k$ *using* $O(m)$ *matrix-vector products with* $\boldsymbol{A}$ *and an additional* $O(nmk)$ *run-time.*

Proof. Form the matrix $\boldsymbol{K} = [\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \cdots, \boldsymbol{A}^{m-1}\boldsymbol{v}]$. Let $p(x) = c_0 + c_1 x + \cdots + c_{m-1}x^{m-1}$ be any polynomial of degree $m-1$ and let $\boldsymbol{c} = [c_0, c_1, \cdots, c_{m-1}]^T$ be the column vector of its coefficients. Then $p(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{K}\boldsymbol{c}$. Thus, to compute all of the $p_j(\boldsymbol{A})\boldsymbol{v}$, we can first use $O(m)$ matrix-vector products with $\boldsymbol{A}$ to form $\boldsymbol{K} \in \mathbb{R}^{n\times m}$ and second use $O(k)$ rectangular matrix-vector products of the form $\boldsymbol{c} \mapsto \boldsymbol{K}\boldsymbol{c}$ to calculate the $p_j(\boldsymbol{A})\boldsymbol{v}$, each of which requires $O(nm)$ time. Overall, we require $O(m)$ matrix-vector products with $\boldsymbol{A}$ plus $O(nmk)$ additional time. □

Remark 10. The monomial basis used in $\boldsymbol{K}$ is known to be poorly behaved numerically; in practice, one should use the output of the Lanczos procedure [Saa11, Chapter 6]. See also [Che24]. Stability is discussed also in [MMS18].

Trace estimation

To approximate the trace of a matrix function $f(\boldsymbol{A})$ we will use variance-reduced Hutchinson's, known as Hutch++ [MMMW21].

Lemma 4.2.4 (Hutch++)**.** *Let* $\boldsymbol{M}$ *be a symmetric matrix, then with* $O(\sqrt{\log(1/\delta)}/\epsilon + \log(1/\delta))$ *matrix vector products with* $\boldsymbol{M}$, *Hutch++ algorithm satisfies the following with probability* $\geq 1-\delta$:

$$|\mathit{Hutch{+}{+}}(\boldsymbol{M}) - \operatorname{tr}(\boldsymbol{M})| \leq \varepsilon \|\boldsymbol{M}\|_* .$$

At a high level, Hutch++ approximates the trace in two steps: first it finds a low rank approximation $\boldsymbol{M}_{(k)}$ of $\boldsymbol{M}$ and directly computes the trace, then it approximates the trace of $\boldsymbol{M} - \boldsymbol{M}_{(k)}$ using classical stochastic trace estimators, via quadratic forms with random Rademacher vectors.

We note that Hutch++ builds its low rank approximation using matrix-vector products computed with $\boldsymbol{M}$. In Section 4.5, computing these matrix-vector products will be expensive, and we will use an alternative method to construct a low-rank approximation of our matrix. Although we do not directly apply Hutch++, the same ideas will still apply: we approximate our target matrix with a low rank approximation and directly compute the trace, then use Hutchinson's to approximate the remaining portion.

4.3 Rational approximation of step function

Our first goal will be to find an explicit, computable rational approximation of the indicator function $\mathbf{1}_{[0,\infty)}(x)$. The well-known Zolotarev functions are the standard choice for a rational approximation of the step function, given their optimal rate trade off between error and degree. However, we opt for a different rational approximation, given by [PP11] (see Theorem 4.3.1 for an explicit definition).

The main advantage of this approximation as opposed to Zolotarev functions is the ability to avoid elliptic integrals. We will want to take a partial fraction decomposition of whatever rational approximation we take, so we will need to find the poles. For Zolotarev functions, this means computing elliptic integrals. For our construction, we will have to approximate the solutions to some system of equations, but as we will see in Section 4.3, this can be done very accurately very quickly. Additionally, using this rational approximation we will be able to explicitly write down the partial fraction decomposition in terms of the poles, and we can bound the error introduced by substituting approximate poles.

Overall, this choice of rational approximation is simpler but still enjoys many of the same benefits as Zolotarev functions: fast rate of convergence to the step function (as seen in Theorem 4.3.1), purely imaginary poles (as seen in Theorem 4.3.2), and smallest pole bounded away from zero (as seen in Theorem 4.3.3). We defer almost all of the proofs in this section to the appendix – they largely rely on standard tools from calculus and in our opinion are not conceptually interesting.

We remark that everything contained in this section is problem independent in a sense. The function we construct will only depend on our error parameters, but not on $\mathbf{A}$.

Theoretical approximation of step function

Our starting point is the following lemma from [PP11]:

Lemma 4.3.1 ([PP11])**.** *Let $\alpha, \beta, \gamma > 0$ and $\alpha/2 \le \beta$. Then there exists a rational function s such that*

$$\begin{aligned} |s(x)| &\le \gamma, & x &\in [-\beta, -\alpha/2] \\ |1 - s(x)| &\le \gamma, & x &\in [\alpha/2, \beta], \\ 0 \le s(x) &\le 1, & x &\in (-\infty, \infty) \end{aligned}$$

and letting deg *be the max degree of the numerator and denominator we have*

$$\deg s \le B \ln\left(e + \frac{2\beta}{\alpha}\right) \ln\left(e + \frac{1}{\gamma}\right)$$

where $B > 1$ is an absolute constant less than 5.

We have the following explicit construction of $s(x)$ from [PP11]. Let $\rho = \frac{1}{e+2\beta/\alpha}$ and $m = O(\ln(\frac{2\beta}{\alpha}) \ln(\frac{1}{\gamma}))$ and define

$$h(z) = \frac{p(-z)}{p(z)}, \qquad p(z) = \prod_{j=1}^{m} (z + \rho^{j/m}), \qquad s_1(z) = \frac{1}{h(z)^2 + 1}.$$

Then the rational approximation described above is $s(z) = s_1(\beta z)$. Note that the degree m is of the same order as the optimal degree rational approximation of this form, Zolotarev functions.

As we will see, since we assume our matrix $\boldsymbol{A}$ has $\|\boldsymbol{A}\| = 1$, we can take β to be a constant (in particular, $\beta = 3$ will suffice). We will continue to include β as a parameter in what follows, but we treat it as a constant. If we assume $\alpha < 1$, we can take $\rho = \frac{\alpha}{10}$, so $\rho \asymp \alpha$. As α is the more meaningful parameter in terms of our construction, we will always replace ρ with α in bounds. Note the dependency of the function on γ is through m.

We will find the partial fraction decomposition of this function. To start, we need to know where the poles are; below we show the poles are purely imaginary and give an explicit description.

Lemma 4.3.2. *The poles of $s(z)$ are $\left\{\pm i\frac{y_0}{\beta}, \ldots, \pm i\frac{y_{m-1}}{\beta}\right\}$ where y_k satisfies:*

$$\sum_{j=1}^{m} \arctan(y_k \rho^{-j/m}) = \frac{\pi}{4} + k\frac{\pi}{2}.$$

The pole closest to the real axis will be of particular interest to us, mostly in regards to conditioning of certain linear systems. We lower bound the magnitude of this pole.

Lemma 4.3.3. *Let $i\frac{y}{\beta}$ be a pole of $s(z)$. Then if $\rho < 1/2$,*

$$|y| \ge \frac{\pi\rho(1 - \rho^{1/m})}{4(1 - \rho)} \ge \frac{\pi}{2} \cdot \frac{\rho}{m} \ge O\Big(\frac{\rho}{m}\Big) = \tilde{O}(\alpha).$$

For computational reasons, it will be beneficial to express $s(z)$ as its partial fraction decomposition. Fortunately, we can compute this in terms of the poles of $s(z)$.

Lemma 4.3.4. *Denote the poles of $s(z)$ as $\{\pm iy_0/\beta, \dots, \pm iy_{m-1}/\beta\}$. The partial fraction decomposition of $s(z)$ is given by:*

$$s(z) = \frac{1}{2} + \sum_{k=0}^{m-1} \left(4 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + y_k^2} \right)^{-1} \cdot \left(\frac{1}{\beta z - i y_k} + \frac{1}{\beta z + i y_k} \right).$$

Note this also gives us the partial fraction decomposition over the reals:

$$s(x) = \frac{1}{2} + \sum_{k=0}^{m-1} \left(4 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + y_k^2} \right)^{-1} \left(\frac{2\beta x}{\beta^2 x^2 + y_k^2} \right). \tag{4.4}$$

This will be useful to avoid complex arithmetic. To simplify notation, let $C(y) = 4 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + y^2}$ so that

$$s(x) = \frac{1}{2} + \sum_{k=0}^{m-1} C(y_k)^{-1} \cdot \left(\frac{2\beta x}{\beta^2 x^2 + y_k^2} \right).$$

Remark 11. We see that to compute $\hat{s}(\boldsymbol{A})$ (or $\hat{s}(\boldsymbol{A})\boldsymbol{y}$ for some vector $\boldsymbol{y}$) we will need to compute the inverses:

$$(\beta^2 \boldsymbol{A}^2 + y_k^2 \boldsymbol{I})^{-1}.$$

Here we can begin to appreciate the advantage of having purely imaginary poles – no matter the spectrum of $\boldsymbol{A}$, our system still has eigenvalues bounded away from zero due to the y_k^2 term. In other words, having the imaginary component ensures the systems we are solving are (at least somewhat) well-conditioned.

Remark 12. We work with the partial fraction decomposition for a few reasons. First, it allows for parallel computations of each desired inverse when we run Hutch++. It also will be easier to analyze the error from approximately solving the inverses using the partial fraction decomposition. Finally, it will be convenient to work with the partial fraction decomposition when we analyze the error from approximating the poles in the next section.

Numerical approximation of step function

There is one key difficulty with the above approximation: we cannot explicitly compute the poles, so we cannot implement this approximation. We will first need to numerically approximate the poles (using a simple numerical method, such as bisection method or Newton's method), and then see how much this affects our approximation.

Lemma 4.3.5. *We can numerically approximate the poles to within ξ in $O(m^2 \log(m/\xi)) = \tilde{O}(\log(1/\xi))$ time.*

Let y_k^* be the approximate poles returned from Theorem 4.3.5. We consider the rational function $\hat{s}(z)$ constructed with the approximate poles of $s(x)$:

$$\hat{s}(x) = \frac{1}{2} + \sum_{k=0}^{m-1} C(y_k^*)^{-1} \cdot \left(\frac{2\beta x}{\beta^2 x^2 + (y_k^*)^2} \right). \tag{4.5}$$

We show the error from using the approximate poles is small.

Lemma 4.3.6. *Let $\hat{s}(z)$ be as above and assume $|y_k - y_k^*| \leq \xi \ll y_0$ for all m. Then for all real x*

$$|\hat{s}(x) - s(x)| \leq O(m\xi/y_0).$$

Then with the appropriate choice of parameters, we can approximate a step function:

Lemma 4.3.7. *In $\tilde{O}(1)$ time, we can find $\hat{s}(x)$ such that*

$$\begin{aligned} |\hat{s}(x)| &\leq \gamma, & x &\in [-\beta, -\alpha/2] \\ |1 - \hat{s}(x)| &\leq \gamma, & x &\in [\alpha/2, \beta], \\ -\gamma \leq \hat{s}(x) &\leq 1 + \gamma, & x &\in (-\infty, \infty). \end{aligned}$$

Proof. We will make an explicit choice of ξ, use Theorem 4.3.6 to bound the difference between $|\hat{s}(x) - s(x)|$, then use Theorem 4.3.1 to bound the difference to the step function.

Apply Theorem 4.3.6 with

$$\xi = \frac{\gamma \rho^2}{m^3} \leq \frac{\gamma y_0^2}{m}.$$

Then $O(d\xi/y_0) \leq O(\gamma)$ so we have $|\hat{s}(x) - s(x)| \leq O(\gamma)$. We note that the constant hidden here is less than 10, provided $\xi \leq y_0/2$. Adjusting our choice of γ by a constant factor, we can assume

$$|\hat{s}(x) - s(x)| \leq \gamma.$$

Applying triangle inequality and Theorem 4.3.1 gives us the desired error bounds to the step function.

Combining the runtime from Theorem 4.3.5 with our choice of ξ, the runtime is

$$O\left(m^2 \ln\left(\frac{m^4}{\gamma \rho^2} \right) \right) = \tilde{O}(1).$$

□

We have now achieved the goal of this section: an explicit, computable rational approximation of the step function, given to us in the form of $\hat{s}(x)$. Now we see what we can do with this approximation.

4.4 Spectral density estimation

We now leverage our estimate $\hat{s}(x)$ to approximate the number of eigenvalues in a given interval and to give a spectral density estimation with histogram guarantees. First, we construct an approximate bump function using the following observation:

$$\mathbf{1}_{[a_1,a_2]}(x) = \mathbf{1}_{[a_1,\infty)}(x) - \mathbf{1}_{[a_2,\infty)}(x) = \mathbf{1}_{[0,\infty)}(x - a_1) - \mathbf{1}_{[0,\infty)}(x - a_2). \tag{4.6}$$

Once we construct our approximate bump function, in order to apply it to our matrix $\boldsymbol{A}$, we need to solve the linear systems appearing in our expression. We will do this using Chebychev semi-iteration. We note that this will in fact give us a *polynomial* approximation of a bump function (and a step function). As we will see in Theorem 4.4.3, this polynomial achieves the best degree possible up to log factors. Once this is in place, we apply Hutch++ to give approximations for the number of eigenvalues in an interval.

Note our approximation of a bump function is smooth from the rational approximation and will have some non-zero boundary region of width $\alpha/2$ on each side. We offer different framings of how to interpret our approximation with respect to this boundary region. If there is a known gap in our spectrum, we can take advantage of this (Theorem 4.4.5); if not, we can use what we call a "sandwich" approximation (Theorem 4.4.6).

For a full spectral density estimation with histogram guarantees, we will apply the single interval framework many times. We will choose the partitioning of $[-1, 1]$ stochastically to avoid having a large number of eigenvalues in our boundary regions (this is inspired by the same random choice of intervals used in [MNSUW17]). The end result is a spectral density estimation with global and local guarantees (see Theorem 4.4.7).

Rational approximation: indicator function

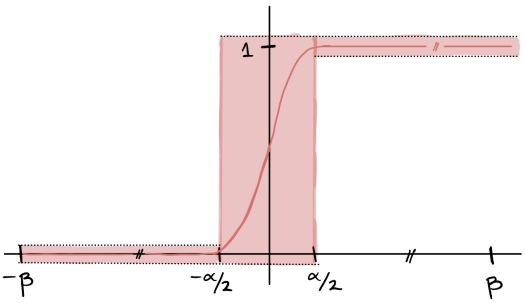
We first use $\hat{s}(x)$ to define an approximate indicator function of some interval.

Theorem 4.4.1. *Let $-\beta/2 + \alpha < a_1 < a_2 < \beta/2 - \alpha$. In $\tilde{O}(1)$ time, we can compute $\hat{b}_{[a_1,a_2]}(x)$ such that*

$$\left|\hat{b}_{[a_1,a_2]}(x) - \mathbf{1}_{[a_1,a_2]}(x)\right| \leq \begin{cases} \gamma & \text{for } x \in [-\beta/2, a_1 - \alpha] \cup [a_1, a_2] \cup [a_2 + \alpha, \beta/2] \\ 1 + \gamma & \text{otherwise.} \end{cases}$$

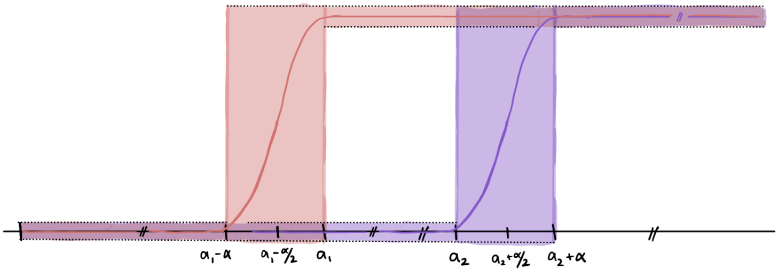
We also note for $x \in [a_1 - \alpha, a_1] \cup [a_2, a_2 + \alpha]$ (our boundary regions), $-\gamma \leq \hat{b}_{[a_1,a_2]}(x) \leq 1 + \gamma$.

Proof. We will construct our indicator as a difference of shifted step functions, as alluded to in (4.6). We need to have some care with our choice of shifts due to our boundary regions. Recall the shape of $\hat{s}(x)$:



where the shaded regions are the bounds we have on $\hat{s}(x)$ (we have drawn a representative curve for concreteness, but note we only know the curve falls in the shaded region).

Let $-\beta/2 + \alpha < a_1 < a_2 < \beta/2 - \alpha$. We want to take shifted copies of $\hat{s}(x)$:



We see the appropriate shifts to make sure we are close to 1 for all of $[a_1, a_2]$ is $a_1 - \alpha/2$ for one step function and $a_2 + \alpha/2$ for the other step function. We let

$$\hat{b}_{[a_1,a_2]}(x) = \hat{s}(x - (a_1 - \alpha/2)) - \hat{s}(x - (a_2 + \alpha/2)).$$

This gives us an approximate indicator of $[a_1, a_2]$ with boundary regions of width α on either side. Using Theorem 4.3.7 and the triangle inequality, we have

$$\left|\hat{b}_{[a_1,a_2]}(x) - \mathbf{1}_{[a_1,a_2]}(x)\right| \leq \begin{cases} 2\gamma & \text{for } x \in [-\beta/2, a_1 - \alpha] \cup [a_1, a_2] \cup [a_2 + \alpha, \beta/2] \\ 1 + 2\gamma & \text{otherwise} \end{cases} \tag{4.7}$$

Note for the first case we are leveraging our choice of a_1, a_2: for any $x \in [-\beta/2, a_1 - \alpha] \cup [a_2 + \alpha, \beta/2]$ we have $x - (a_1 - \alpha/2), x - (a_2 + \alpha/2) \in [-\beta, \beta]$, the region where we have more control over $\hat{s}(x)$.

Adjusting γ by a constant factor gives us the desired error bounds. The bounds for the boundary region come from applying the triangle inequality more carefully to our difference of step functions (or from our illustration above).

The runtime analysis is identical to the runtime for $\hat{s}(x)$. □

Remark 13 (Choice of β)**.** Recall that in our initial construction, we chose $\beta = 3$. Assuming $\alpha < 1/2$, we see that we can construct a bump function for any interval with $-1 \leq a_1 < a_2 \leq 1$, which is precisely our spectral window. We took β to be slightly larger than our spectral radius so that we would be able to do these shifts to get bump functions.

Recalling the partial fraction decomposition of $\hat{s}(x)$ from Theorem 4.3.4, we have:

$$\begin{aligned}\hat{b}_{[a_1,a_2]}(\mathbf{A}) = \sum_{k=0}^{m-1} C(y_k^*)^{-1} 2(\mathbf{A} - (a_1 - \alpha/2)\boldsymbol{I})(\beta^2(\mathbf{A} - (a_1 - \alpha/2)\boldsymbol{I})^2 + (y_k^*)^2\boldsymbol{I})^{-1} \\ - \sum_{k=0}^{m-1} C(y_k^*)^{-1} 2(\mathbf{A} - (a_2 + \alpha/2)\boldsymbol{I})(\beta^2(\mathbf{A} - (a_2 + \alpha/2)\boldsymbol{I})^2 + (y_k^*)^2\boldsymbol{I})^{-1}.\end{aligned} \tag{4.8}$$

Now we turn to solving the linear systems appearing in this expression.

Linear system solver: Chebychev semi-iteration

For ease of notation, let $\hat{b} := \hat{b}_{[a_1,a_2]}$ throughout.

If we have a linear system solver, then we are happy to work directly with $\hat{b}(x)$. If not, we can use Chebyshev semi-iteration to approximate the inverses with polynomials. As stated in Theorem 4.2.1, given a positive definite matrix $\boldsymbol{M}$ with condition number κ and error tolerance $\tau > 0$, we can find a polynomial q_τ with degree at most $O(\sqrt{\kappa}\log(1/\tau))$ such that

$$(1 - \tau)\boldsymbol{M}^{-1} \preceq q_k(\boldsymbol{M}) \preceq (1 + \tau)\boldsymbol{M}^{-1}.$$

We will first bound the condition number of the systems we are applying this to and bound the degree of the polynomial. Then we analyze the error introduced by substituting in these polynomial approximations. Since our bump function is just constructed as a difference of two step functions, we will work with step functions primarily (for ease of notation), and then apply the triangle inequality to control the error in $\hat{b}$.

Condition number bound

Theorem 4.4.2. *For each inverse appearing in $\hat{s}(\mathbf{A})$ or $\hat{b}(\boldsymbol{A})$, we can find a $(1 \pm \tau)$ polynomial approximate inverse with degree at most*

$$O\left(\frac{m\ln(\tau^{-1})}{\alpha}\right) = \tilde{O}\left(\frac{1}{\alpha}\right).$$

Note this means we need at most $\tilde{O}\left(\frac{1}{\alpha}\right)$ matrix-vector products to find the approximate inverse.

Proof. We want bound the condition numbers of matrices of the form

$$(\beta^2(\boldsymbol{A} - t\boldsymbol{I})^2 + (y_k^*)^2\boldsymbol{I})$$

for some shift t (which could be 0 in the case of $\hat{s}(\boldsymbol{A})$). The smallest eigenvalue is at least $(y_k^*)^2$ because of the second term. The largest is at most $O(1) + (y_k^*)^2$ since $\beta = 3$ and $\|\boldsymbol{A} - t\boldsymbol{I}\| \leq 3$ since $|t| \leq 1 + \alpha/2 \leq 2$. The condition number is always bounded by

$$\kappa \leq 1 + O\left(\frac{1}{(y_k^*)^2}\right) \leq O\left(\frac{1}{(y_0^*)^2}\right).$$

Using Theorem 4.3.3 and the fact that $y_0^* = O(y_0)$ by our choice of parameters in Theorem 4.3.7, we have

$$\sqrt{\kappa} \leq O\left(\frac{m}{\rho}\right) = \tilde{O}\left(\frac{1}{\alpha}\right).$$

Combining this bound with Theorem 4.2.1 we have that for each inverse appearing in $\hat{s}(\mathbf{A})$, we can find a $(1 \pm \tau)$-approximation using $O\left(\frac{m \ln(\tau^{-1})}{\alpha}\right)$ matrix vector products. □

Error analysis from inverse approximation

We now need to understand the error introduced into $\hat{s}(\boldsymbol{A})$ from plugging in these approximate inverses. We start with analyzing the error introduced to the scalar function $\hat{s}(x)$.

Lemma 4.4.3. *Recall β should be treated as a constant. Let $q_{\tau,k}(x)$ denote the polynomial from Theorem 4.2.1 satisfying for $x \in [(y_k^*)^2, \beta^4 + (y_k^*)^2]$,*

$$(1 - \tau)x^{-1} \leq q_{\tau,k}(x) \leq (1 + \tau)x^{-1}. \tag{4.9}$$

Let $\tilde{s}(x)$ denote the fully polynomial approximation of the step function

$$\tilde{s}(x) = \frac{1}{2} + \sum_{k=0}^{m-1} C(y_k^*)^{-1} \cdot 2\beta x \cdot q_{\tau,k}(\beta^2 x^2 + (y_k^*)^2). \tag{4.10}$$

Note by Theorem 4.4.2, $\tilde{s}(x)$ has degree at most $O\left(\frac{\ln(1/\alpha)\ln(1/\gamma)^2}{\alpha}\right) = \tilde{O}\left(\frac{1}{\alpha}\right)$. Then we have

$$\left|\tilde{s}(x) - \mathbf{1}_{[0,\infty)}(x)\right| \leq \gamma \quad \text{for } x \in [-\beta, -\alpha/2] \cup [\alpha/2, \beta]. \tag{4.11}$$

For $x \in [-\alpha/2, \alpha/2]$, we can guarantee $-\gamma \leq \tilde{s}(x) \leq 1 + \gamma$.

Proof. We start by bounding the difference between $\tilde{s}(x)$ and $\hat{s}(x)$. We will leverage the fact that for any $x \in [-\beta, \beta]$, by our choice of $q_{\tau,k}$ we have

$$\left|q_{\tau,k}(\beta^2 x^2 + (y_k^*)^2) - (\beta^2 x^2 + (y_k^*)^2)^{-1}\right| \leq \tau(\beta^2 x^2 + (y_k^*)^2)^{-1}.$$

By definitions of $\hat{s}(x)$ and $\tilde{s}(x)$, and the fact that $C(y_k^*) > 0$, we have

$$\begin{aligned}
|\hat{s}(x) - \tilde{s}(x)| &= \left| \sum_{k=0}^{m-1} C(y_k^*)^{-1} 2\beta x \cdot \left(q_\tau(\beta^2 x^2 + (y_k^*)^2) - \left(\beta^2 x^2 + (y_k^*)^2\right)^{-1} \right) \right| \\
&\leq \sum_{k=0}^{m-1} C(y_k^*)^{-1} \left| 2\beta x \right| \cdot \left| \left(q_\tau(\beta^2 x^2 + (y_k^*)^2) - \left(\beta^2 x^2 + (y_k^*)^2\right)^{-1} \right) \right| \\
&\leq \tau \sum_{k=0}^{m-1} C(y_k^*)^{-1} \left| 2\beta x \cdot (\beta^2 x^2 + (y_k^*)^2)^{-1} \right|.
\end{aligned}$$

Now notice that all terms in the sum are positive if $x > 0$ or negative if $x < 0$, so we can move the absolute value outside the sum again without any loss. Taking the absolute value of the sum, we see

$$|\hat{s}(x) - \tilde{s}(x)| \leq \tau \left| \sum_{k=0}^{m-1} C(y_k^*)^{-1} 2\beta x \cdot (\beta^2 x^2 + (y_k^*)^2)^{-1} \right| = \tau \left| \hat{s}(x) - 1/2 \right|.$$

Since $|\hat{s}(x)| \leq 2$, we have $|\hat{s}(x) - s(x)| \leq O(\tau)$ for $x \in [-\beta, \beta]$. We can adjust τ by a constant factor and incur no change to the order of the degree, so we assume without loss of generality that $|\hat{s}(x) - s(x)| \leq \tau$.

Now by the triangle inequality and Theorem 4.3.7, we have

$$\left| \hat{s}(x) - \mathbf{1}_{[0,\infty)}(x) \right| \leq \gamma + \tau \quad \text{for } x \in [-\beta, -\alpha/2] \cup [\alpha/2, \beta]$$

Within the boundary interval, we can bound the value of our approximation:

$$-(\gamma + \tau) \leq \hat{s}(x) \leq 1 + \gamma + \tau \quad \text{for } x \in [-\alpha/2, \alpha/2]$$

Letting $\tau = \gamma$ and scaling both by a constant factor we get a γ approximation to $\mathbf{1}_{[0,\infty)}(x)$ for $x \in [-\beta, -\alpha/2] \cup [\alpha/2, \beta]$.

Finally, we have the degree guaranteed by applying Theorem 4.4.2 and our choice of $\tau = \gamma$. □

Remark 14. By [EY06], this achieves the best polynomial approximation of a step function with respect to parameters α, γ up to log factors. Such a polynomial was used in [JS19] to perform principal component projections; they explicitly construct a polynomial approximation of $\text{sign}(x)$ with optimal degree with respect to analogous parameters. They give an efficient and stable construction using Chebyshev polynomials by first approximating a square root function. The connection between rational and polynomial approximations of optimal degree (up to log factors) has not been noted previously to our knowledge.

Our results for spectral density estimation with histogram guarantees can thus also be obtained by directly using such an optimal degree polynomial. However, we believe the

connection between rational and polynomial approximations of optimal degree (up to log factors) is theoretically interesting. We also note that given a structured matrix $\boldsymbol{A}$ where fast system solvers are available, it is possible to do better than the runtime achieved with a polynomial approximation. In this case, one would work directly with $\hat{b}(\boldsymbol{A})$ instead of $\tilde{b}(\boldsymbol{A})$ (defined below).

We can now define $\tilde{b}(x)$ as

$$\tilde{b}(x) = \tilde{s}(x - a_1 + \alpha/2) - \tilde{s}(x - a_2 - \alpha/2). \tag{4.12}$$

This is a fully polynomial approximation to the indicator function of $[a_1, a_2]$. As an immediate corollary of the above and Theorem 4.4.1 we have that $\tilde{b}$ is a good approximation to a bump function.

Corollary 4.4.4. *For $-\beta/2 - \alpha < a_1 < a_2 < \beta/2 + \alpha$, letting $\tilde{b}(x)$ be as defined in* (4.12)*, then*

$$\left|\tilde{b}_{[a_1,a_2]}(x) - \mathbf{1}_{[a_1,a_2]}(x)\right| \leq \begin{cases} \gamma & \text{for } x \in [-\beta/2, a_1 - \alpha] \cup [a_1, a_2] \cup [a_2 + \alpha, \beta/2] \\ 1 + \gamma & \text{otherwise.} \end{cases} \tag{4.13}$$

For $x \in [a_1 - \alpha, a_1] \cup [a_2, a_2 + \alpha]$ we have $-\gamma \leq \tilde{b}_{[a_1,a_2]}(x) \leq 1 + \gamma$. The degree of $\tilde{b}$ is at most $\tilde{O}(1/\alpha)$.

We note by the choice of our $q_{k,\tau}$ and the conditioning of our systems, this immediately extends to the matrix case.

Trace estimation: Hutch++

Our overall goal is to give an approximation for the number of eigenvalues in some target interval $[a_1, a_2]$. Let k_{eig} denote the number of eigenvalues in $[a_1, a_2]$ and note that $\mathrm{tr}(\mathbf{1}_{[a_1,a_2]}(\mathbf{A})) = k_{eig}$. Let k_b denote the number of eigenvalues in $[a_1 - \alpha/2, a_1] \cup [a_2, a_2 + \alpha/2]$, which we will refer to as the "boundary regions". Now that we have $\tilde{b}(x)$, an approximation of $\mathbf{1}_{[a_1,a_2]}(x)$, we can approximate k_{eig} by looking at $\mathrm{tr}(\tilde{b}(\mathbf{A}))$. From Theorem 4.4.1 and Theorem 4.4.4 respectively we have

$$-n\gamma \leq \mathrm{tr}(\tilde{b}(\mathbf{A})) - \mathrm{tr}(\mathbf{1}_{[a_1,a_2]}(\mathbf{A})) \leq k_b + n\gamma, \tag{4.14}$$

where we are using the fact that by our choice of $\beta = 3$ and normalization, $\sigma(\boldsymbol{A}) \subset [-\beta/2, \beta/2]$.

Now we turn to estimating $\mathrm{tr}(\tilde{b}(\boldsymbol{A}))$. To compute $\mathrm{tr}(\tilde{b}(\boldsymbol{A}))$ to ε relative accuracy, we will need to compute $O(1/\varepsilon)$ matrix-vector products with $\tilde{b}(\boldsymbol{A})$ by Theorem 4.2.4. By Theorem 4.4.2, $\tilde{b}(\boldsymbol{A})$ is a polynomial of degree at most $\tilde{O}(1/\alpha)$, so for each matrix-vector product

with $\tilde{b}(\boldsymbol{A})$, we need $\tilde{O}(1/\alpha)$ matrix-vector products with $\boldsymbol{A}$. We see that with $\tilde{O}(\frac{1}{\alpha\varepsilon})$ matrix-vector products with $\boldsymbol{A}$, we can find

$$\left|\text{Hutch++}(\tilde{b}(\boldsymbol{A})) - \text{tr}(\tilde{b}(\boldsymbol{A}))\right| \leq \varepsilon \left\|\tilde{b}(\mathbf{A})\right\|_* \tag{4.15}$$

From Theorem 4.4.4, we can bound $\left\|\tilde{b}(\boldsymbol{A})\right\|_* = \sum \left|\lambda_i(\tilde{b}(\boldsymbol{A}))\right| \leq k_{eig} + k_b + n\gamma$. Combining this with (4.14) and using the triangle inequality we see

$$(1-\varepsilon)k_{eig} - \varepsilon k_b - n\gamma \leq \text{Hutch++}(\tilde{b}(\boldsymbol{A})) \leq (1+\varepsilon)k_{eig} + (1+\varepsilon)k_b + n\gamma \tag{4.16}$$

where we have let γ absorb a constant factor. Note the runtime is inversely proportional to the width of the boundary region.

Single intervals

We apply the above to single intervals.

Isolated interval

Corollary 4.4.5. *Suppose for some target interval $[a_1, a_2]$ there is a choice of α such that $k_b = 0$ (i.e., there is a known gap of width α on either side of the target interval). In $\tilde{O}(\frac{n^2}{\alpha\varepsilon})$ time we can find $\tilde{k}$ such that*

$$\left|\tilde{k} - k_{eig}\right| \leq \varepsilon k_{eig}.$$

Proof. Applying the above with the given α, with $\tilde{O}(\frac{1}{\alpha\varepsilon})$ matvecs, we can find $\text{Hutch++}(\tilde{b}(\boldsymbol{A}))$ such that

$$\left|\text{Hutch++}(\tilde{b}(\boldsymbol{A})) - k_{eig}\right| \leq \frac{\varepsilon}{2}k_{eig} + n\gamma,$$

where we have used that $k_b = 0$ and let γ absorb the constant factor in (4.16). Letting $\gamma = \varepsilon/2n^2$, note $n\gamma \leq \frac{\varepsilon}{2n}$. Then if we return $\tilde{k} = \text{Hutch++}(\tilde{b}(\boldsymbol{A}))$, we have the desired bound since

$$\left|\text{Hutch++}(\tilde{b}(\boldsymbol{A})) - k_{eig}\right| \leq \frac{\varepsilon}{2}k_{eig} + n\gamma \leq \varepsilon k_{eig}.$$

We now just need to deal with the case when $k_{eig} = 0$. If $k_{eig} = 0$, then we have

$$\left|\text{Hutch++}(\tilde{b}(\boldsymbol{A}))\right| \leq \frac{\varepsilon}{2n}. \tag{4.17}$$

Additionally, provided $\varepsilon < 1$ and $n > 1$ we will only ever have (4.17) if $k_{eig} = 0$, so in this case we can safely round down and return $\tilde{k} = 0$. □

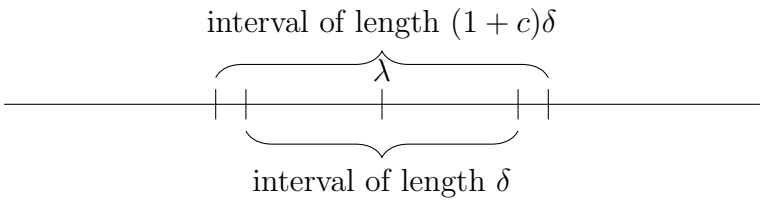


Figure 4.3: Intervals used in a "sandwich" approximation.

"Sandwich" approximations

Corollary 4.4.6. *Consider some target interval of length δ centered at some point λ: $I_1 = [\lambda - \delta/2, \lambda + \delta/2]$, and the "sandwiching" interval of length $(1+c)\delta$ for some constant c: $I_2 = [\lambda - (1+c)\delta/2, \lambda + (1+c)\delta/2]$ (see Figure 4.3). Let k_i be the number of eigenvalues of A in I_i. Then in $\tilde{O}(\frac{n^2}{\delta\varepsilon^2})$ time, we can find $\tilde{k}$ such that*

$$(1-\varepsilon)k_1 \leq \tilde{k} \leq (1+\varepsilon)k_2.$$

Proof. Instead of applying Hutch++ to estimate the trace, we will apply classic Hutchinson's. This requires $O(1/\varepsilon^2)$ matrix-vector products to give a $(1 \pm \frac{\varepsilon}{2})$ approximate of the trace; in other words we have

$$\left(1 - \frac{\varepsilon}{2}\right) \operatorname{tr}(\tilde{b}(\boldsymbol{A})) \leq \operatorname{Hutch}(\tilde{b}(\boldsymbol{A})) \leq \left(1 + \frac{\varepsilon}{2}\right) \operatorname{tr}(\tilde{b}(\boldsymbol{A})).$$

Using the bounds from (4.14), this gives

$$\left(1 - \frac{\varepsilon}{2}\right)(k_1 - n\gamma) \leq \operatorname{Hutch}(\tilde{b}(\boldsymbol{A})) \leq \left(1 + \frac{\varepsilon}{2}\right)(k_2 + n\gamma).$$

Assume $\varepsilon < 1/2$ and $n \geq 2$ and set $\gamma = \varepsilon/2n^2$. If $\operatorname{Hutch}(\tilde{b}(\boldsymbol{A})) \leq 0$, then return $\tilde{k} = 0$; otherwise return $\tilde{k} = \operatorname{Hutch}(\tilde{b}(\boldsymbol{A}))$. Using a similar argument to above, $\tilde{k}$ satisfies the desired bounds. □

Spectral density estimation

We use the machinery developed in the past few sections to give a spectral density estimation with histogram guarantees – essentially to approximate the number of eigenvalues in various intervals.

We consider a histogram approximation with intervals of uniform widths. We will choose our the endpoints of our intervals *randomly* so that with some constant probability, very few of the eigenvalues fall in the boundary regions (drawing inspiration from [MNSUW17]). Note that if we choose our intervals deterministically, we cannot possibly hope to make this guarantee.

We outline the procedure with the following algorithm:

Algorithm 2 Approximate spectral histogram with uniform interval widths

Require: Symmetric $\boldsymbol{A}$ with $\|\boldsymbol{A}\| \leq 1$, desired width w, and error parameters ε, ν
Ensure: Set of intervals I_t of width w with counts $\tilde{k}_t$ where $\tilde{k}_t$ approximates the number of eigenvalues of $\boldsymbol{A}$ in the interval I_t

▷ Choose partitioning and vectors for trace approximations.

Set $\alpha = O(w^2\nu)$.
Choose $x \in [-w/2 + \alpha/2, w/2 - \alpha/2]$ uniformly at random.
Set $I_t = [x + tw, x + (t+1)w]$ for $-\lfloor 1/w \rfloor - 1 \leq t \leq \lceil 1/w \rceil + 1$.
Set $S = \lfloor 1/\varepsilon \rfloor$. choose $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S \in \{-1, 1\}^n$ independently, uniformly at random.

▷ Find Krylov subspaces.

for each $\boldsymbol{y}_i$ **do**
    Form Krylov subspace $\mathcal{K}_{\boldsymbol{A}}(\boldsymbol{y}_i) = \left\{\boldsymbol{y}_i, \boldsymbol{A}\boldsymbol{y}_i, \ldots, \boldsymbol{A}^k\boldsymbol{y}_i\right\}$ where $k = \tilde{O}(1/\alpha)$.
end for

▷ Perform trace approximation.

for each I_t **do**
    Run Hutch++ [MMMW21] on $\tilde{b}_{I_t}(\boldsymbol{A})$, using $\mathcal{K}_{\boldsymbol{A}}(\boldsymbol{y}_i)$ to efficiently compute $\tilde{b}_{I_t}(\boldsymbol{A})\boldsymbol{y}_i$.
    Let $\tilde{k}_t$ = approximation returned by Hutch++
    If $\tilde{k}_t \leq \varepsilon\nu$, set $\tilde{k}_t = 0$.
end for

return I_t and $\tilde{k}_t$ for all t. ▷ Return intervals and estimates.

Theorem 4.4.7. *With probability at least* 99/100*, Algorithm 2 returns intervals* I_t *of width* w *and values* $\tilde{k}_t$ *such that*

$$(1-\varepsilon)k_t - \varepsilon\nu(k_{t-1} + k_t + k_{t+1}) \leq \tilde{k}_t \leq (1+\varepsilon)k_t + \nu(k_{t-1} + k_t + k_{t+1}).$$

The runtime for the algorithm is $\tilde{O}\left(\frac{n^2}{\varepsilon w^2 \nu}\right)$.

Proof. We break the proof up into several sections for clarity: (1) construction of intervals, (2) probabilistic argument, where we argue that by the construction of our intervals, there should be few eigenvalues lying in the boundary regions, (3) trace estimation, where we unravel the error from applying Hutch++, and (4) runtime analysis of all the steps of our algorithm.

Construction of intervals: We construct the intervals as outlined in Algorithm 2. Let w denote some fixed length and choose some $x \in [-w/2 + \alpha/2, w/2 - \alpha/2]$ uniformly at random. Consider the lattice points $\{x + tw\}$ with $-\lfloor 1/w \rfloor - 1 \leq t \leq \lceil 1/w \rceil + 1$. Let

$$I_t = [x + tw, x + (t+1)w], \qquad \overline{I}_t = [x + tw - \alpha/2, x + (t+1)w + \alpha/2].$$

Now define $M_t = [tw - w/2, (t+1)w + w/2]$. Notice that by our choice of x, we deterministically have

$$\overline{I}_t \subset M_t \subset I_{t-1} \cup I_t \cup I_{t+1}.$$

Probabilistic argument: Let k_t denote the number of eigenvalues in the interval I_t. Recall that

$$\operatorname{tr}(\tilde{b}_{I_t}(\boldsymbol{A})) \leq k_t + \left|\sigma(\boldsymbol{A}) \cap (\overline{I}_t \backslash I_t)\right| + n\gamma. \tag{4.18}$$

Let $B = \left\{i : \lambda_i(\boldsymbol{A}) \in \overline{I}_t \backslash I_t \text{ for some } t\right\}$, i.e., the indices of eigenvalues that fall in some boundary. Note that B is a random set, depending on the choice of x. We have that $\mathbb{P}(i \in B) \leq \frac{\alpha}{w-\alpha}$ for any i. (This is easiest to see by considering the probability that the point 0 is in a boundary region – this only occurs if $x \in [-\alpha/2, \alpha/2]$ so that 0 is within $\alpha/2$ of an endpoint. This occurs with probability $\alpha/(w-\alpha)$.)

Using this, we bound the expectation of $\left|\sigma(\boldsymbol{A}) \cap (\overline{I}_t \backslash I_t)\right|$:

$$\begin{aligned}
\mathbb{E}\left|\sigma(\boldsymbol{A}) \cap (\overline{I}_t \backslash I_t)\right| &= \mathbb{E} \sum_{\lambda_i(\boldsymbol{A}) \in \overline{I}_t \backslash I_t} \mathbf{1}(i \in B) && (4.19)\\
&\leq \mathbb{E} \sum_{\lambda_i(\boldsymbol{A}) \in M_t} \mathbf{1}(i \in B) && (4.20)\\
&\leq \sum_{\lambda_i(\boldsymbol{A}) \in M_t} \mathbb{P}(i \in B) && (4.21)\\
&\leq \frac{\alpha}{w-\alpha} \cdot (k_{t-1} + k_t + k_{t+1}). && (4.22)
\end{aligned}$$

By Markov, with probability $1 - \frac{w}{200}$:

$$\left|\sigma(\boldsymbol{A}) \cap (\overline{I}_t \backslash I_t)\right| \leq \frac{200\alpha}{w(w-\alpha)}(k_{t-1} + k_t + k_{t+1}) \leq \frac{400\alpha}{w^2}(k_{t-1} + k_t + k_{t+1})$$

where in the last step we have assumed $\alpha \leq w/2$. By a union bound, this holds for all intervals simultaneously with probably 99/100 (since we have $2/w$ intervals).

Taking $\alpha = O(w^2 \cdot \nu)$ we have for all I_t with probability 99/100

$$\operatorname{tr}(\tilde{b}_{I_t}(\boldsymbol{A})) \leq k_t + \nu(k_{t-1} + k_t + k_{t+1}) + n\gamma.$$

Trace estimation: Now we consider the Hutch++ step with $1/\varepsilon$ matvecs with $\tilde{b}_{I_t}(\boldsymbol{A})$. Letting $\tilde{k}_t = \text{Hutch++}(\tilde{b}_{I_t}(\boldsymbol{A}))$, by (4.16) we have

$$\left|\tilde{k}_t - k_t\right| \leq \varepsilon k_t + (1+\varepsilon)k_b + n\gamma.$$

We are conditioning on the event that the second term in this bound is small – in particular, with probability 99/100, we have

$$\left|\tilde{k}_t - k_t\right| \leq \varepsilon k_t + \nu(1+\varepsilon)(k_{t-1} + k_t + k_{t+1}) + n\gamma.$$

By adjusting our choice of ν, the $(1+\varepsilon)$ factor can be absorbed into ν (as we will see, this will not affect the runtime by more than a constant factor). We then have

$$\tilde{k}_t \leq (1+\varepsilon)k_t + \nu(k_{t-1} + k_t + k_{t+1}) + n\gamma.$$

Note we can set $\gamma = \varepsilon\nu/n$, which will again not affect the asymptotic runtime except for log factors. If k_{t-1}, k_t, or k_{t+1} are non-zero, the $n\gamma$ term can be absorbed into the second term. If $k_{t-1} = k_t = k_{t+1} = 0$, then the algorithm will return a value $\leq \varepsilon\nu$, in which case we round down to 0. In either case, the following upper bound holds:

$$\tilde{k}_t \leq (1+\varepsilon)k_t + \nu(k_{t-1} + k_t + k_{t+1}).$$

For the lower bound, recall that

$$tr(\tilde{b}(\boldsymbol{A})) \geq k_t - n\gamma.$$

Combining this with Hutch++, we have with probability 99/100:

$$\tilde{k}_t \geq (1-\varepsilon)k_t - \varepsilon\nu(k_{t-1} + k_t + k_{t+1}),$$

where we have absorbed the $n\gamma$ term into the second term by the above argument.

Run time analysis: Now we work out the runtime. In order to leverage Theorem 4.2.3, we fix one random set of vectors $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S\}$ where $S = \lfloor 1/\varepsilon \rfloor$ to be used when we apply Hutch++ for any $b_{I_t}(\boldsymbol{A})$. Recalling that each $\tilde{b}_{I_t}(\boldsymbol{A})$, we have from Theorem 4.2.3 that to compute $b_{I_t}(\boldsymbol{A})\boldsymbol{y}_i$ for a fixed $\boldsymbol{y}_i$ for all t will take $\tilde{O}(1/\alpha)$ matrix-vector products with $\boldsymbol{A}$ and an additional $\tilde{O}\big(\frac{n}{\alpha w}\big)$. We do this for $\lfloor 1/\varepsilon \rfloor$ vectors, so recalling $\alpha = O(w^2\nu)$, in total our runtime is

$$\tilde{O}\bigg(\frac{n^2}{\varepsilon w^2 \nu}\bigg) + \tilde{O}\bigg(\frac{n}{\varepsilon w^3 \nu}\bigg).$$

Note that the dominant time is computing the Krylov subspaces. □

Corollary 4.4.8. *Fix $1 > \varepsilon > 0$. With probability at least 99/100 in time $\tilde{O}(n^2/\varepsilon^4)$, Algorithm 2 returns I_t of width less than or equal to ε and $\tilde{k}_t$ such that*

$$(1-\varepsilon)k_t - \varepsilon^2(k_{t-1} + k_t + k_{t+1}) \leq \tilde{k}_t \leq (1+\varepsilon)k_t + \varepsilon(k_{t+1} + k_t + k_{t-1}).$$

Additionally, we can define the probability measure $\tilde{s}_{\boldsymbol{A}}$ with point masses at the midpoints of each I_t with weight proportional to $\tilde{k}_t$. Then $\mathcal{W}(s_{\boldsymbol{A}}, \tilde{s}_{\boldsymbol{A}}) \leq O(\varepsilon)$.

Proof. For the histogram guarantee, we simply apply Theorem 4.4.7 with $w = \nu = \varepsilon$.

We consider the Wasserstein distance guarantee. First consider an intermediary measure μ defined as follows: given the intervals $I_t = [a_t, b_t]$ and k_t, the number of eigenvalues in I_t, let

$$\mu\bigg(\frac{a_t + b_t}{2}\bigg) = \frac{k_t}{n}.$$

Note that each point mass in the original spectral distribution $s_{\boldsymbol{A}}$ could have moved by at most $\varepsilon/2$ since I_t have width $\leq \varepsilon$, so we have

$$\mathcal{W}(\mu, s_{\boldsymbol{A}}) \leq \frac{\varepsilon}{2}. \tag{4.23}$$

Now we turn to $\tilde{s}_{\boldsymbol{A}}$. Letting I_t be as above and $\tilde{k}_t$ be the approximate number of eigenvalues returned by the algorithm, we let

$$\tilde{s}_{\boldsymbol{A}}\left(\frac{a_t + b_t}{2}\right) = \frac{\tilde{k}_t}{\sum_i \tilde{k}_i}.$$

Note the sum of the approximate number of eigenvalues in each interval will not necessarily sum to n, hence the different normalization. Using the guarantees on each $\tilde{k}_t$, we have:

$$n(1 - \varepsilon - 3\varepsilon^2) \leq \sum_t \tilde{k}_t \leq n(1 + 4\varepsilon).$$

We write $\sum_t \tilde{k}_t = n(1 + \Delta)$ where $-4\varepsilon \leq \Delta \leq 4\varepsilon$, where we assume $\varepsilon < 1$ and $\gamma < \varepsilon$.

We will bound $\mathcal{W}_1(\mu, \tilde{s}_{\boldsymbol{A}})$ since the only difference is the mass they place at each midpoint, then use the triangle inequality to compare $\tilde{s}_{\boldsymbol{A}}$ to $s_{\boldsymbol{A}}$. To bound the Wasserstein distance, we have from [GS02]

$$\mathcal{W}(\mu, \tilde{s}_{\boldsymbol{A}}) \leq 2 \left\| \mu - \tilde{s}_{\boldsymbol{A}} \right\|_{TV}$$

where the 2 is coming from the fact that the whole interval has length 2. Let $\chi = \left\{ \frac{a_t + b_t}{2} : I_t = [a_t, b_t] \right\}$, i.e., the locations of all the point masses. By definition since both μ and $\tilde{s}_{\boldsymbol{A}}$ are defined on the same discrete set, we have

$$\begin{aligned}
\left\| \mu - \tilde{s}_{\boldsymbol{A}} \right\|_{TV} &= \frac{1}{2} \sum_{x \in \chi} |\mu(x) - \tilde{s}_{\boldsymbol{A}}(x)| = \frac{1}{2} \sum_t \left| \frac{k_t}{n} - \frac{\tilde{k}_t}{\sum \tilde{k}_t} \right| \\
&= \frac{1}{2} \sum_t \left| \frac{k_t(1 + \Delta) - \tilde{k}_t}{n(1 + \Delta)} \right| \\
&= \frac{1}{2} \sum_t \left(\frac{1}{1 + \Delta} \right) \cdot \left| \frac{\Delta k_t}{n} + \frac{k_t - \tilde{k}_t}{n} \right| \\
&\leq \frac{1}{2(1 + \Delta)} \cdot (|\Delta| + 4\varepsilon) \leq \frac{4\varepsilon}{1 - 4\varepsilon}
\end{aligned}$$

We have

$$\mathcal{W}(\mu, \tilde{s}_{\boldsymbol{A}}) \leq \frac{4\varepsilon}{1 - 4\varepsilon}.$$

Combined with (4.23), assuming $\varepsilon < 1/5$ we have

$$\mathcal{W}(s_{\boldsymbol{A}}, \tilde{s}_{\boldsymbol{A}}) \leq O(\varepsilon).$$

□

Remark 15. The individual interval count guarantees in Theorem 4.4.8 are reminiscent of the histogram approximation of [MNSUW17] but are ultimately incomparable. The spectral density estimation given from the histogram counts in [MNSUW17] does not achieve a Wasserstein distance guarantee, but it can be used to give an approximation of spectral sums. Our spectral density estimation is the opposite: while we satisfy a Wasserstein distance guarantee, the rounding we do is in an absolute sense (not a relative sense), so we cannot use this to give relative approximations of spectral sums. The key difference is the "boundary" regions of the two methods: in [MNSUW17], these boundary regions have size relative to the size of the endpoints of an interval; here, the boundary regions are always of a constant size.

The method of [MNSUW17] can also be viewed as using a rational approximation of a bump function, but they restrict to only one real pole. The advantage is then they only have one system to solve (repeatedly), so they can find a good preconditioner for that one system to then use repeatedly; the disadvantage is that the approximation is not as good.

4.5 Schatten-1 norm approximation

We look at the problem of approximating the *Schatten*-1 *norm* (or a nuclear norm) of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, again assuming we have normalized so $\|\boldsymbol{A}\| \in [1/2, 1]$. Recall from (4.2) the Schatten-1 norm of $\boldsymbol{A}$, denoted $\|A\|_*$, is defined as follows

$$\|\boldsymbol{A}\|_* := \sum_{i=1}^{d} \sigma_i(\boldsymbol{A}),$$

where $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_d(\boldsymbol{A})$ are the singular values of the matrix $\boldsymbol{A}$. Given an error parameter $\varepsilon > 0$, our goal is to compute an estimator Z such that $(1-\varepsilon)\|\boldsymbol{A}\|_* \leq Z \leq (1+\varepsilon)\|\boldsymbol{A}\|_*$, with probability 99/100.

We first observe that $\|A\|_* = \operatorname{tr}(\sqrt{\boldsymbol{A}^\top \boldsymbol{A}})$. We proceed by constructing a rational approximation to $\sqrt{x}$ using the previously defined step function $\hat{s}(x)$. We again assume we have taken a pre-processing step to normalize so that $\left\|\boldsymbol{A}^T \boldsymbol{A}\right\| \in [1/2, 1]$. Note this implies $\|\boldsymbol{A}\|_* \geq 1/2$.

Rational approximation: $\sqrt{x}$

For a non-negative $x \in \mathbb{R}$, let $x = t^2$, then $\sqrt{x} = |t| = t \cdot \operatorname{sign}(t)$. We form a rational approximation to the sign function $\hat{g}(t)$ using the step function $\hat{s}(t)$ from Theorem 4.3.7 as $\hat{g}(t) := 2\hat{s}(t) - 1$. We can also write this as

$$\hat{g}(t) = \sum_{k=0}^{m-1} \left(8 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + (y_k^*)^2} \right)^{-1} \cdot \left(\frac{2\beta t}{\beta^2 t^2 + (y_k^*)^2} \right).$$

Then, we have that a rational approximation to $\sqrt{x}$ can be represented as $t\hat{g}(t)$. Substituting $x = t^2$ we get the following expression

$$\hat{r}(x) = \sum_{k=0}^{m-1} \left(16 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + (y_k^*)^2} \right)^{-1} \cdot \left(\frac{\beta x}{\beta^2 x + (y_k^*)^2} \right).$$

where again we will assume $\beta = 3$. This gives us the following piecewise bound for $\hat{g}(x)$

$$\begin{aligned} -O(\gamma\sqrt{x}) \le \hat{g}(x) \le \sqrt{x} \pm O(\gamma\sqrt{x}), \qquad & x \in [0, \alpha^2/4] \\ \sqrt{x} - O(\gamma\sqrt{x}) \le \hat{g}(x) = \sqrt{x} + O(\gamma\sqrt{x}), \qquad & x \in [\alpha^2/4, \beta^2] \end{aligned}$$

where the degree, m, of $\hat{g}(x)$ is $O\Big(\ln\big(\frac{2\beta}{\alpha}\big) \ln\Big(\frac{1}{\gamma}\Big)\Big)$.

Setting $\alpha/2 \le \frac{\varepsilon \|\boldsymbol{A}\|_*}{d}$, we get that

$$(1 - \varepsilon - O(\gamma)) \|\boldsymbol{A}\|_* \le \operatorname{tr}(\hat{r}(\boldsymbol{A}^\top \boldsymbol{A})) \le (1 + O(\gamma)) \|\boldsymbol{A}\|_* . \tag{4.24}$$

We will use stochastic trace estimator to estimate the trace above, which requires getting an estimate of

$$\hat{r}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{v} = \sum_{k=0}^{m-1} \left(16 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + (y_k^*)^2} \right)^{-1} \cdot \beta \boldsymbol{A}^\top \boldsymbol{A} \big(\beta^2 \boldsymbol{A}^\top \boldsymbol{A} + (y_k^*)^2\big)^{-1} \boldsymbol{v},$$

for any vector $\boldsymbol{v}$. To get an estimate of $\hat{r}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{v}$, we will invoke the linear system solver of **DS26**, and the main work goes into estimating $\big(\beta^2 \boldsymbol{A}^\top \boldsymbol{A} + (y_k^*)^2\big)^{-1} \boldsymbol{v}$.

Linear system solver: [DS26a]

Theorem 4.5.1 (Modified [**DS26**])**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, $\boldsymbol{v} \in \mathbb{R}^d$, and $\lambda > 0$. For any $\varepsilon_L > 0$, there exists a randomized algorithm that returns a vector $\tilde{\boldsymbol{x}}$ satisfying*

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|^2_{\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}} \le \varepsilon_L \|\boldsymbol{x}^*\|^2_{\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}}, \tag{4.25}$$

with high probability, where $\boldsymbol{x}^* = (\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{v}$. The algorithm runs in time

$$\tilde{O}\big((nnz(\boldsymbol{A}) + d^2 \bar{\kappa}_{0,1}(\boldsymbol{A}_\lambda)) \log(1/\varepsilon_L)\big),$$

where $\boldsymbol{A}_\lambda \in \mathbb{R}^{(n+d) \times d}$ is the augmented matrix formed by appending $\sqrt{\lambda}\boldsymbol{I} \in \mathbb{R}^{d \times d}$ to the rows of $\boldsymbol{A}$, and for any $\boldsymbol{M} \in \mathbb{R}^{n \times d}$, $\bar{\kappa}_{0,1}(\boldsymbol{M}) := \frac{1}{d} \sum_{i=1}^{d} \frac{\sigma_i(\boldsymbol{M})}{\sigma_d(\boldsymbol{M})}$.

To estimate $\operatorname{tr}(\hat{r}(\boldsymbol{A}^\top \boldsymbol{A}))$, we will consider quadratic forms like $\boldsymbol{v}^\top \Big(\beta \boldsymbol{A}^\top \boldsymbol{A} \big(\beta^2 \boldsymbol{A}^\top \boldsymbol{A} + (y_i^*)^2\big)^{-1}\Big) \boldsymbol{v}$. To estimate this we will use the Theorem 4.5.1. The following fact simplifies the expression, and in the lemma after that we bound the error in the quadratic form produced by our procedure, versus the actual quadratic form.

Fact 4.5.2. *Let $\boldsymbol{M}$ be a positive semidefinite matrix and let $c > 0$. Then $\boldsymbol{M}(\boldsymbol{M} + c\boldsymbol{I})^{-1} = \boldsymbol{I} - c(\boldsymbol{M} + c\boldsymbol{I})^{-1}$.*

Lemma 4.5.3. *Let $\boldsymbol{v} \in \mathbb{R}^d$ be a Rademacher vector. Suppose that $\tilde{\boldsymbol{x}}$ is the vector returned by Theorem 4.5.1 solver with $\boldsymbol{x}^* = (\boldsymbol{A}^\top \boldsymbol{A} + ((y_k^*)^2/\beta^2)\boldsymbol{I})^{-1}\boldsymbol{v}$. Then, with high probability, the error E_i between the quadratic forms can be bounded as*

$$E_i := \left| \boldsymbol{v}^\top \frac{1}{\beta} \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{x}^* - \boldsymbol{v}^\top \frac{1}{\beta} \boldsymbol{A}^\top \boldsymbol{A} \tilde{\boldsymbol{x}} \right| \leq \frac{d\sqrt{\varepsilon_L}}{\beta}.$$

Proof. Let $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{A}$, and let $\lambda_i = (y_k^*/\beta)^2$. Then, by Theorem 4.5.2, the expression

$$\beta \boldsymbol{A}^\top \boldsymbol{A} \big(\beta^2 \boldsymbol{A}^\top \boldsymbol{A} + (y_i^*)^2\big)^{-1} = \frac{1}{\beta}(\boldsymbol{I} - \lambda_i(\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1}).$$

Multiplying a Rademacher vector $\boldsymbol{v}$ on both sides, we get that

$$\boldsymbol{v}^\top \frac{1}{\beta}(\boldsymbol{I} - \lambda_i(\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1})\boldsymbol{v} = \frac{1}{\beta}(d - \lambda_i \boldsymbol{v}^\top (\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1})\boldsymbol{v}.$$

Recall that $\tilde{\boldsymbol{x}}$ is the vector returned by Theorem 4.5.1 solver with $\boldsymbol{x}^* = (\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1}\boldsymbol{v}$. Then our procedure outputs $\frac{1}{\beta}(d - \lambda_i \boldsymbol{v}^\top \tilde{\boldsymbol{x}})$. We now bound the error E_i between the quadratic forms:

$$E_i = \left| \frac{1}{\beta}(d - \lambda_i \boldsymbol{v}^\top \tilde{\boldsymbol{x}}) - \frac{1}{\beta}(d - \lambda_i \boldsymbol{v}^\top \boldsymbol{x}^*) \right| = \frac{\lambda_i}{\beta} \left| \boldsymbol{v}^\top (\tilde{\boldsymbol{x}} - \boldsymbol{x}^*) \right|$$

By definition $\boldsymbol{x}^* = (\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1}\boldsymbol{v}$, so $\boldsymbol{v} = (\boldsymbol{M} + \lambda_i \boldsymbol{I})\boldsymbol{x}^*$, and we get that

$$E_i = \frac{\lambda_i}{\beta} \left| ((\boldsymbol{M} + \lambda_i \boldsymbol{I})\boldsymbol{x}^*)^\top (\tilde{\boldsymbol{x}} - \boldsymbol{x}^*) \right| \leq \frac{\lambda_i}{\beta} \left\| \boldsymbol{x}^* \right\|_{\boldsymbol{M}+\lambda_i \boldsymbol{I}} \left\| \tilde{\boldsymbol{x}} - \boldsymbol{x}^* \right\|_{\boldsymbol{M}+\lambda_i \boldsymbol{I}}. \tag{4.26}$$

Now, note from (4.25) that $\left\| \tilde{\boldsymbol{x}} - \boldsymbol{x}^* \right\|_{\boldsymbol{M}+\lambda_i \boldsymbol{I}} \leq \sqrt{\varepsilon_L} \left\| \boldsymbol{x}^* \right\|_{\boldsymbol{M}+\lambda_i \boldsymbol{I}}$, which implies that

$$E_i \leq \frac{\lambda_i}{\beta} \sqrt{\varepsilon_L} \left\| \boldsymbol{x}^* \right\|_{\boldsymbol{M}+\lambda_i \boldsymbol{I}}^2. \tag{4.27}$$

Furthermore, we can simplify

$$\begin{aligned} \sqrt{\varepsilon_L} \left\| \boldsymbol{x}^* \right\|_{\boldsymbol{M}+\lambda_i \boldsymbol{I}}^2 = \sqrt{\varepsilon_L} \boldsymbol{x}^{*\top} (\boldsymbol{M} + \lambda_i \boldsymbol{I}) \boldsymbol{x}^* &= \sqrt{\varepsilon_L} \big((\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1} \boldsymbol{v}\big)^\top (\boldsymbol{M} + \lambda_i \boldsymbol{I}) \big((\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1} \boldsymbol{v}\big) \\ &= \sqrt{\varepsilon_L} \boldsymbol{v}^\top (\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1} \boldsymbol{v} = \sqrt{\varepsilon_L} \boldsymbol{v}^\top \boldsymbol{x}^*. \end{aligned}$$

Plugging this into (4.27), we get that

$$E_i \leq \frac{\lambda_i}{\beta} \sqrt{\varepsilon_L} (\boldsymbol{v}^\top \boldsymbol{x}^*). \tag{4.28}$$

We can also bound $\boldsymbol{v}^\top \boldsymbol{x}^*$ by recalling that $\boldsymbol{v}^\top \boldsymbol{x}^* = \boldsymbol{v}^\top (\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1} \boldsymbol{v} \leq \frac{d}{\lambda_i}$, since the eigenvalues of $(\boldsymbol{M} + \lambda_i \boldsymbol{I})^{-1} \leq \lambda_i^{-1}$ and $\|\boldsymbol{v}\|_2^2 = d$. Moreover, $\boldsymbol{v}^\top \boldsymbol{x}^* \geq 0$, which from (4.28) implies that $E_i \leq \frac{d\sqrt{\varepsilon_L}}{\beta}$. □

Lemma 4.5.4. *For a Rademacher $\boldsymbol{v}$, let DS_Alg($\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}$) be our algorithm to estimate $\boldsymbol{v}^\top\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}$. Then*

$$\left|\textit{DS\_Alg}(\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}) - \boldsymbol{v}^\top\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}\right| \leq O(dm\sqrt{\varepsilon_L}).$$

Moreover, for a Rademacher vector $\boldsymbol{v}$, each call of DS_Alg($\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}$) takes time

$$\tilde{O}\left(\left(nnz(\boldsymbol{A}) + \frac{d^2}{\varepsilon}\right)\log(1/\varepsilon_L)\right).$$

Proof. From Theorem 4.5.3, we get that

$$\begin{aligned}\left|\text{DS\_Alg}(\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}) - \hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}\right| &\leq \left|\sum_{k=0}^{m-1}\left(16\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+(y_k^*)^2}\right)^{-1}\right|\cdot\frac{d\sqrt{\varepsilon_L}}{\beta}.\\ &\leq O(dm\sqrt{\varepsilon_L}),\end{aligned}$$

where the last inequality follows from Theorem 4.6.3, and the fact that β is a constant.

To show the running time, we first collect all the facts we need. Recall that to get a relative approximation to Schatten-1 norm ((4.24)) we set the parameter $\alpha/2 = \varepsilon\left\|\boldsymbol{A}\right\|_* /d$ and we assumed that the spectral radius, and hence the Schatten-1 norm, is at least a constant. Therefore, from Theorem 4.3.1, we get that

$$m := \deg(\hat{r}(x)) = O(\log(d/\varepsilon)\log(1/\gamma)). \tag{4.29}$$

Next, we compute the smallest y_k^* used in the expression for $\hat{a}$, because that would determine the condition number of our system. From Theorem 4.3.3, we know that the smallest $y_k^* = O((\varepsilon\left\|\boldsymbol{A}\right\|_* /d)\log(\gamma^{-1}))$.

We now compute the number $\bar{\kappa}_{0,1}$, and let y_0^* denote the smallest magnitude pole of $\hat{r}(x)$. Then the matrix $\boldsymbol{A}_\lambda$ in Theorem 4.5.1 is the matrix $\boldsymbol{A}$ with $y_0^*\boldsymbol{I}$ appended at the bottom. Calculating $\bar{\kappa}_{0,1}$, we get that

$$\bar{\kappa}_{0,1} \leq \frac{1}{d}\sum_{i=1}^{d}\frac{\sigma_i(\boldsymbol{A})+y_0^*}{y_0^*} \leq \frac{1}{d}\frac{\left\|\boldsymbol{A}\right\|_*}{\varepsilon\left\|\boldsymbol{A}\right\|_* /d} = \frac{1}{\varepsilon}.$$

We have to use Theorem 4.5.1 for each of the m poles, and therefore, we get that the running time of computing DS_Alg($\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}$) is bounded by

$$\tilde{O}\left(m\left(\text{nnz}(\boldsymbol{A}) + \frac{d^2}{\varepsilon}\right)\log(1/\varepsilon_L)\right) = \tilde{O}\left(\left(\text{nnz}(\boldsymbol{A}) + \frac{d^2}{\varepsilon}\right)\log(1/\varepsilon_L)\right),$$

where the equality follows from (4.29). □

Trace estimation: variance reduced Huchinson's

Instead of directly applying Hutchinson's trace estimator, we follow the variance reduced Hutchinson's approach [MMMW21] to estimate $\mathrm{tr}(\hat{r}(\boldsymbol{A}^\top \boldsymbol{A}))$. Towards that, we first get a low rank approximation to $\sqrt{(\boldsymbol{A}^\top \boldsymbol{A})}$. However, matrix-vector products with $\sqrt{(\boldsymbol{A}^\top \boldsymbol{A})}$ (or an approximation of it) are expensive, so we appeal to different method for constructing a low rank approximation. This exact problem has been addressed by Persson et al. [PMM25], where they show how that for operator monotone functions f, $\mathrm{tr}(f(\boldsymbol{A}^\top \boldsymbol{A}))$ can easily be approximated using low rank approximation to $\boldsymbol{A}^\top \boldsymbol{A}$. We show this in the following lemma.

Lemma 4.5.5. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, and let $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{A}$. For any $k \in [d]$, and constant $\eta > 0$ there exists a randomized procedure that computes a rank-k PSD matrix $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ such that, with high probability:*

1. *$\mathrm{tr}(\boldsymbol{H})$ can be computed in $O(k)$ time.*
2. *The matrix $\boldsymbol{H}$ satisfies*

$$\left\|\sqrt{\boldsymbol{M}} - \boldsymbol{H}\right\|_F^2 \leq (1+\eta)\left\|\sqrt{\boldsymbol{M}} - (\sqrt{\boldsymbol{M}})_{(k)}\right\|_F^2.$$

The total running time of this procedure is $\tilde{O}(nnz(\boldsymbol{A})k + dk^2 + k^3)$.

Proof. First, run $p = O(\log d)$ iterations of the randomized Block Krylov subspace method on $\boldsymbol{M}$ to obtain an orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{d\times k}$. From Musco and Musco [MM15, Theorem 11] we get that

$$\left\|\boldsymbol{M} - (\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{M})_{(k)}\right\|_F^2 \leq (1+\eta)\left\|\boldsymbol{M} - (\boldsymbol{M})_{(k)}\right\|_F^2.$$

Next, we explicitly construct the rank-k Nystrom approximation $\hat{\boldsymbol{M}} = \boldsymbol{M}\boldsymbol{Q}(\boldsymbol{Q}^\top \boldsymbol{M}\boldsymbol{Q})^\dagger \boldsymbol{Q}^\top \boldsymbol{M}$. Then, using Persson et al. [PMM25, Theorem 3.2] we get that

$$\|\boldsymbol{M}\|_F^2 - \left\|\hat{\boldsymbol{M}}\right\|_F^2 \leq (1+\eta)\left\|\boldsymbol{M} - (\boldsymbol{M})_{(k)}\right\|_F^2$$

and $\boldsymbol{M} \succeq \hat{\boldsymbol{M}} \succeq 0$. Finally, we apply the operator monotone function $f(x) = \sqrt{x}$ to the Nystrom approximation $\hat{\boldsymbol{M}}$, setting $\boldsymbol{H} = (\sqrt{\hat{\boldsymbol{M}}})_{(k)}$. From Persson et al. [PMM25, Theorem 2.5], we have the guarantee that

$$\left\|\sqrt{\boldsymbol{M}} - \boldsymbol{H}\right\|_F^2 \leq (1+\eta)\left\|\sqrt{\boldsymbol{M}} - (\sqrt{\boldsymbol{M}})_{(k)}\right\|_F^2.$$

Running Time: The block Krylov method takes $\tilde{O}(\mathrm{nnz}(\boldsymbol{A})k + dk^2)$ time, and the Nystrom method takes $\tilde{O}(\mathrm{nnz}(\boldsymbol{A})k + dk^2 + k^3)$ time, which leads to a total running time of $\tilde{O}(\mathrm{nnz}(\boldsymbol{A})k + dk^2 + k^3)$. □

Lemma 4.5.6. *Let $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{A}$, and let $\boldsymbol{H}$ be the matrix from the previous lemma, and let $q \in \mathbb{Z}$ be a positive integer. Let $\hat{r}(x)$ be the rational approximation to $\sqrt{x}$, of degree m. Then, there exists an algorithm that returns an estimate $\tilde{Z}$ that satisfies*

$$\left|\tilde{Z} - \|\boldsymbol{A}\|_*\right| \leq O(\varepsilon)\|\boldsymbol{A}\|_*, \text{ with probability at least } 1 - \frac{10}{q}\left(\frac{1}{d} + \frac{1}{k\varepsilon^2}\right).$$

Moreover, the time taken to construct such an estimator is

$$\tilde{O}\left(q\left(nnz(\boldsymbol{A}) + \frac{d^2}{\varepsilon}\right)\right).$$

Proof. Let q be the number of matrix-vector product queries to $\hat{r}(\boldsymbol{M})$. If we had exact access to them, then our ideal variance-reduced Hutchinson's estimator would be $Z = \operatorname{tr}(\boldsymbol{H}) + \frac{1}{q}\sum_{i=1}^{q} \boldsymbol{v}_i^\top(\hat{r}(\boldsymbol{M}) - \boldsymbol{H})\boldsymbol{v}_i$, where $\boldsymbol{v}_i$'s are random Rademacher vectors. However, since we are using the solvers of [**DS26**] to estimate $v_i^\top \hat{r}(\boldsymbol{M})\boldsymbol{v}_i$, our estimator (Theorem 4.5.4) is

$$\tilde{Z} := \operatorname{tr}(\boldsymbol{H}) + \frac{1}{q}\sum_{i=1}^{q}\big(\text{DS\_Alg}(\hat{r}(\boldsymbol{M})\boldsymbol{v}_i) - \boldsymbol{v}_i^\top \boldsymbol{H}\boldsymbol{v}_i\big).$$

Using triangle inequality, we get that the total error is

$$\left|\tilde{Z} - \|\boldsymbol{A}\|_*\right| \leq \underbrace{\left|\tilde{Z} - Z\right|}_{\text{Solver Error}} + \underbrace{|Z - \mathbb{E}Z|}_{\text{Statistical Error}} + \underbrace{|\mathbb{E}Z - \|\boldsymbol{A}\|_*|}_{\text{Approximation Error}}.$$

We bound each of those individually.

Approximation error: By definition $\mathbb{E}Z = \operatorname{tr}(\hat{r}(\boldsymbol{M}))$. In our setting with $\alpha/2 = \varepsilon\|\boldsymbol{A}\|_*/d$, and $\gamma \leq \varepsilon/(4d)$, we get that

$$|\mathbb{E}Z - \|\boldsymbol{A}\|_*| = \left|Tr(\hat{r}(\boldsymbol{M})) - \operatorname{tr}(\sqrt{\boldsymbol{M}})\right| \leq O(\varepsilon)\|\boldsymbol{A}\|_*.$$

Solver error: By Theorem 4.5.4, we have that

$$\left|\tilde{Z} - Z\right| \leq \frac{1}{q}\sum_{i=1}^{q}\left|\text{DS\_Alg}(\hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}_i) - \boldsymbol{v}_i^\top \hat{r}(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{v}_i\right| \leq O(dm\sqrt{\varepsilon_L}).$$

To ensure that this is at most $\varepsilon\|\boldsymbol{A}\|_*$, we set $\varepsilon_L = O(\varepsilon^2\|\boldsymbol{A}\|_*/(d^2m^2))$, and since we assumed that $\|\boldsymbol{A}\|_*$ is at least a constant we have that $\varepsilon_L = O(\varepsilon^2/(d^2m^2))$.

Statistical error: To bound $|Z - \mathbb{E}Z|$, we will use the Chebyshev concentration inequality, which requires computing its variance. Recall that $Z = \operatorname{tr}(\boldsymbol{H}) + \frac{1}{q}\sum_{i=1}^{q} \boldsymbol{v}_i^\top(\hat{r}(\boldsymbol{M}) - \boldsymbol{H})\boldsymbol{v}_i$, where $\boldsymbol{v}_i$'s are random Rademacher vectors. Then,

$$\operatorname{Var}(Z) = \frac{\operatorname{Var}(\boldsymbol{v}_1^\top(\hat{r}(\boldsymbol{M}) - \boldsymbol{H})\boldsymbol{v}_1)}{q} \leq \frac{2\left\|\hat{r}(\boldsymbol{M}) - \boldsymbol{H}\right\|_F^2}{q}.$$

We now bound $2\left\|\hat{r}(\boldsymbol{M}) - \boldsymbol{H}\right\|_F^2$. We decompose this as

$$2\left\|\hat{r}(\boldsymbol{M}) - \boldsymbol{H}\right\|_F^2 \leq 4\left\|\hat{r}(\boldsymbol{M}) - \sqrt{\boldsymbol{M}}\right\|_F^2 + 4\left\|\sqrt{\boldsymbol{M}} - \boldsymbol{H}\right\|_F^2.$$

The second term, $4\left\|\sqrt{\boldsymbol{M}} - \boldsymbol{H}\right\|_F^2$, can be bounded using the previous lemma as follows

$$\begin{aligned}
4\left\|\sqrt{\boldsymbol{M}} - \boldsymbol{H}\right\|_F^2 \leq 4(1+\eta)\left\|\sqrt{\boldsymbol{M}} - (\sqrt{\boldsymbol{M}})_{(k)}\right\|_F^2 &= 4(1+\eta)\sum_{i=k+1}^{d}\sigma_i^2(\boldsymbol{A}) \\
&\leq 4(1+\eta)\sigma_{k+1}(\boldsymbol{A})\sum_{i=k+1}^{d}\sigma_i(\boldsymbol{A}) \\
&\leq 4(1+\eta)\frac{\|\boldsymbol{A}\|_*^2}{k},
\end{aligned}$$

where the last inequality uses the fact that the singular values are sorted in descending order, which implies that $\sigma_{k+1}(\boldsymbol{A}) \leq \|\boldsymbol{A}\|_* / k$. Next, we bound $4\left\|\hat{r}(\boldsymbol{M}) - \sqrt{\boldsymbol{M}}\right\|_F^2$. Since $\alpha/2 = \varepsilon\|\boldsymbol{A}\|_*/d$, we get that

$$4\left\|\hat{r}(\boldsymbol{M}) - \sqrt{\boldsymbol{M}}\right\|_F^2 \leq 4\varepsilon^2\|\boldsymbol{A}\|_*^2/d + 4\gamma^2\|A\|_F^2 \leq 5\varepsilon^2\|A\|_*^2/d,$$

for $\gamma \leq \frac{\varepsilon}{4d}$. Therefore, we get that the variance of the Hutchinsons estimator for the tail is

$$\operatorname{Var}(Z) \leq \frac{5\|A\|_*^2(\varepsilon^2/d + (1+\eta)/k)}{q}.$$

Using the Chebyshev concentration inequality, we get

$$\Pr(|Z - \mathbb{E}[Z]| \geq \varepsilon\|\boldsymbol{A}\|_*) \leq \frac{5\|A\|_*^2(\varepsilon^2/d + (1+\eta)/k)}{q\varepsilon^2\|A\|_*^2} = \frac{1}{q}\left(\frac{5}{d} + \frac{5(1+\eta)}{k\varepsilon^2}\right).$$

Therefore, combining all three bounds above, we get that with probability $1 - \frac{1}{q}\left(\frac{5}{d} + \frac{5(1+\eta)}{k\varepsilon^2}\right)$, we have

$$\left|\tilde{Z} - \|A\|_*\right| \leq O(\varepsilon)\|A\|_*.$$

Running time analysis: We make a total of q matrix-vector queries to $\hat{r}(\boldsymbol{M})$. From Theorem 4.5.4, and with our parameters $\alpha/2 = \varepsilon \|A\|_* / d$, $\gamma = \varepsilon/(4d)$, $\varepsilon_L = O(\varepsilon^2/(d^2 m^2))$, where $m = \log(d/\varepsilon)\log(1/\gamma)$ ((4.29)), we get that the total running time of estimating $\tilde{Z}$ is

$$\tilde{O}\left(q\left(\operatorname{nnz}(\boldsymbol{A}) + \frac{d^2}{\varepsilon}\right)\log(1/\varepsilon_L)\right) = \tilde{O}\left(q\left(\operatorname{nnz}(\boldsymbol{A}) + \frac{d^2}{\varepsilon}\right)\right).$$

We note that the solver of **DS26** in Theorem 4.5.1 returns the answer with high probability, so our probability bound holds. □

We consolidate the above steps for the full algorithm, given in Algorithm 3.

Algorithm 3 Approximate Schatten-1 norm

Require: $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with $\|\boldsymbol{A}\| \le 1$
Ensure: $(1 \pm O(\varepsilon))$ relative approximation of $\sum_{i=1}^{d} \sigma_i(\boldsymbol{A})$

▷ Parameter Initialization

Set $\alpha = \varepsilon/d$, $\beta = 3$, $\gamma = \varepsilon/(4d)$
Set the rational degree $m = \Theta(\log(d/\varepsilon)\log(1/\gamma))$, and solver tolerance $\varepsilon_L = \Theta(\varepsilon^2/(d^2m^2))$.
Set k and q based on sparsity of $\boldsymbol{A}$, and let $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{A}$.

▷ Head: Low Rank Approximation

Run Block Krylov method on $\boldsymbol{M}$ to obtain basis $\boldsymbol{Q} \in \mathbb{R}^{d\times k}$.
Form the Nystrom approximation $\hat{\boldsymbol{M}} = \boldsymbol{M}\boldsymbol{Q}(\boldsymbol{Q}^\top \boldsymbol{M}\boldsymbol{Q})^\dagger \boldsymbol{Q}^\top \boldsymbol{M}$.
Compute rank-k square root $\boldsymbol{H} = \left(\sqrt{\hat{\boldsymbol{M}}}\right)_{(k)}$.
Calculate the exact trace of the head: $T_{\text{head}} := \operatorname{tr}(\boldsymbol{H})$.

▷ Tail: Hutchinson's Stochastic Trace Estimator

Initialize $S = 0$ to accumulate trace of the tail.
for $i = 1$ **to** q **do**
  Draw random Rademacher vector $\boldsymbol{v}_i \in \{-1, 1\}^d$, and let $Q_i = 0$.
  for $k = 0$ **to** $m - 1$ **do**
    Let $\lambda_k = (y_k^*)^2/\beta^2$.
    Use DS_Alg to find $\tilde{\boldsymbol{x}}$ satisfying $(\boldsymbol{M} + \lambda_k \boldsymbol{I})\boldsymbol{x} = \boldsymbol{v}_i$ to precision ε_L. ▷ [DS26a]
    Let $c_k = \left(16 \sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + (y_k^*)^2}\right)^{-1}$. ▷ Rational Coefficients
    $Q_i \leftarrow Q_i + c_k \cdot \frac{1}{\beta}\left(d - \lambda_k \boldsymbol{v}_i^\top \tilde{\boldsymbol{x}}\right)$.
  end for
  $S \leftarrow S + Q_i - \boldsymbol{v}_i^\top \boldsymbol{H} \boldsymbol{v}_i$.
end for

return $Z = T_{\text{head}} + \frac{S}{q}$ ▷ Combine the Outputs

Theorem 4.5.7. *Assuming $d = n$, $nnz(\boldsymbol{A}) = n^2$, Algorithm 3 yields a $(1 \pm \varepsilon)$ approximation of the Schatten-1 norm with probability $99/100$ in time $\tilde{O}(n^2/\varepsilon^{1.5})$.*

Proof. Combining the running time of two lemmas, we get that we can estimate Schatten-1 norm with probability $1 - \frac{1}{q}\left(\frac{5}{d} + \frac{5(1+\eta)}{k\varepsilon^2}\right)$ and with a total running time of

$$\tilde{O}(\mathrm{nnz}(\boldsymbol{A})k + dk^2 + k^3) + \tilde{O}\left(\mathrm{nnz}(\boldsymbol{A})q + \frac{d^2}{\varepsilon}q\right).$$

We optimize k and q under various settings of $\mathrm{nnz}(\boldsymbol{A})$ v/s d, and note down their running time in the following table: □

Matrix Condition	**Optimal Rank (k)**	**Optimal Queries (q)**	**Total Running Time**
$\mathrm{nnz}(\boldsymbol{A}) \gg d^2/\varepsilon$	$\Theta(1/\varepsilon)$	$\Theta(1/\varepsilon)$	$\tilde{O}\left(\frac{\mathrm{nnz}(\boldsymbol{A})}{\varepsilon} + \frac{d^2}{\varepsilon^2}\right)$
$\mathrm{nnz}(\boldsymbol{A}) = \Theta(d^2)$	$\Theta(1/\varepsilon^{1.5})$	$\Theta(1/\varepsilon^{0.5})$	$\tilde{O}\left(\frac{d^2}{\varepsilon^{1.5}}\right)$
$d^{4/3}/\varepsilon \leqslant \mathrm{nnz}(\boldsymbol{A}) \ll d^2$	$\Theta\left(\frac{d}{\sqrt{\mathrm{nnz}(\boldsymbol{A})}\varepsilon^{1.5}}\right)$	$\Theta\left(\frac{\sqrt{\mathrm{nnz}(\boldsymbol{A})}}{d\varepsilon^{0.5}}\right)$	$\tilde{O}\left(\frac{d\sqrt{\mathrm{nnz}(\boldsymbol{A})}}{\varepsilon^{1.5}}\right)$
$\mathrm{nnz}(\boldsymbol{A}) = \Theta(d)$	$\Theta(d^{1/3}/\varepsilon)$	$\Theta(1/(d^{1/3}\varepsilon))$	$\tilde{O}\left(\frac{d^{5/3}}{\varepsilon^2}\right)$

Table 4.1: Optimized runtime of Algorithm 3 for various sparsities.

4.6 Appendix

Proofs for theoretical approximation of step function

Results on poles:

Proof of Theorem 4.3.2. We show the poles of $s(z)$ must be purely imaginary and give explicit equations defining them.

By definition:

$$s(z) = s_1(\beta z) = \frac{1}{h^2(\beta z) + 1}$$

so the poles are precisely when $h^2(\beta z) + 1 = 0 \implies h(\beta z) = \pm i$. Now consider $\beta z = x + iy$:

$$|h(\beta z)| = \left|\frac{p(-\beta z)}{p(\beta z)}\right| = \prod_{i=1}^{m}\left|\frac{-x - iy + \rho^{j/m}}{x + iy + \rho^{j/m}}\right| = \prod_{i=1}^{m}\frac{(\rho^{j/m} - x)^2 + y^2}{(\rho^{j/m} + x)^2 + y^2}.$$

Note that $\rho^{j/m}$ is a real, positive number. If $x > 0$, all of the factors will be less than 1, so $|h(\beta z)| < 1$. If $x < 0$, all of the factors will be greater than 1, so $|h(\beta z)| > 1$. We must have $x = 0$, so all the poles are strictly imaginary. We also note that since the poles are the roots of $p^2(\beta z) + p^2(-\beta z)$, a real-coefficient polynomial, they must be closed under complex conjugation.

Suppose we have a pole at $z = i\frac{y}{\beta}$. Then by above, we must have $\arg(h(iy)) = \pm\frac{\pi}{2} + m\pi$ for some m. Using the definition of $h(z)$, we have

$$\arg(h(iy)) = \arg\left(\prod_{j=1}^{m} \frac{\rho^{j/m} - iy}{\rho^{j/m} + iy}\right) = \sum_{j=1}^{m} \arg\left(\frac{\rho^{j/m} - iy}{\rho^{j/m} + iy}\right) \tag{4.30}$$

$$= \sum_{j=1}^{m} \arg(\rho^{j/m} - iy) - \arg(\rho^{j/m} + iy) \tag{4.31}$$

$$= -2\sum_{j=1}^{m} \arctan(y/\rho^{j/m}), \tag{4.32}$$

so for some m, y must satisfy

$$\sum_{j=1}^{m} \arctan(y/\rho^{j/m}) = \pm\frac{\pi}{4} + k\frac{\pi}{2}. \tag{4.33}$$

First consider equations of the form

$$\sum_{j=1}^{m} \arctan(y/\rho^{j/m}) = \frac{\pi}{4} + k\frac{\pi}{2}.$$

Since $|\arctan(z)| \le \pi/2$, we see the above only has a solution for $0 \le m \le d-1$. Let y_k be the solution to

$$\sum_{j=1}^{m} \arctan(y/\rho^{j/m}) = \frac{\pi}{4} + k\frac{\pi}{2} \qquad \text{for } 0 \le m \le d-1.$$

This gives us d poles, $\left\{i\frac{y_0}{\beta}, i\frac{y_1}{\beta}, \ldots, i\frac{y_{m-1}}{\beta}\right\}$. Including the complex conjugates, we have exactly $2d$ poles. □

Proof of Theorem 4.3.3. Our goal is to lower bound the minimum pole.

We note that $\sum_{j=1}^{m} \arctan(y\rho^{-j/m})$ is strictly increasing in y, so the smallest pole will be when

$$\sum_{j=1}^{m} \arctan\left(\frac{y}{\rho^{j/m}}\right) = \frac{\pi}{4},$$

or equivalently when the right hand side is $-\pi/4$. Since $z \geq \arctan(z)$, we must have

$$\sum_{j=1}^{m} \frac{y}{\rho^{j/m}} \geq \frac{\pi}{4}.$$

We analyze the geometric series appearing in the sum

$$\sum_{j=1}^{m} \rho^{-j/m} = \rho^{-1/m} \cdot \frac{\rho^{-1} - 1}{\rho^{-1/m} - 1} = \frac{1-\rho}{\rho(1-\rho^{1/m})}.$$

Plugging this back into the above, we get the lower bound

$$y \geq \frac{\pi\rho(1-\rho^{1/m})}{4(1-\rho)}.$$

To get the second bound, we recall that $m = O(\ln(1/\rho))$, so we have $\rho^{1/m} = e^{\frac{\ln\rho}{m}} \lesssim e^{-1} \leq 1 - \frac{1}{m}$ for $m > 1$. Rearranging this we have

$$\rho(1-\rho^{1/m}) \gtrsim \frac{\rho}{m}.$$

□

Results on partial fraction decomposition: To find the partial fraction deomposition, we will use the residue method. We state and prove it now for the sake of completeness.

Theorem 4.6.1 (Residue method)**.** *Let $s(z)$ be a rational function and let f, g be any analytic functions such that $s(z) = \frac{f(z)}{g(z)}$. Suppose $s(z)$ has a pole at λ_k. Then in the partial fraction decomposition, the coefficient on the term $\frac{1}{z-\lambda_k}$ is given by*

$$\frac{f(\lambda_k)}{g'(\lambda_k)}.$$

Proof. Suppose we have a rational function $s(z)$ with poles λ_i and we want to find the partial fraction decomposition, i.e., we want to find r_i such that:

$$s(z) = \sum_i \frac{r_i}{z-\lambda_i}.$$

Suppose also we have some expression $s(z) = \frac{f(z)}{g(z)}$ where f, g are analytic functions (but not necessarily polynomials). Suppose we want to find r_k. Multiplying both sides by $z - \lambda_k$ we have

$$(z-\lambda_k)s(z) = r_k + \sum_{i \neq k} \frac{(z-\lambda_k)r_i}{z-\lambda_i} \implies r_k = \lim_{z\to\lambda_k} (z-\lambda_k)s(z).$$

Now we use the expression we had for $s(z)$:

$$\begin{aligned} s(z) = \lim_{z\to\lambda_k}(z-\lambda_k)s(z) &= \lim_{z\to\lambda_k}\frac{(z-\lambda_k)f(z)}{g(z)} \\ &= \lim_{z\to\lambda_k}\frac{f(z)+(z-\lambda_k)f'(z)}{g'(z)} \quad \text{by L'Hopital's} \\ &= \frac{f(\lambda_k)}{g'(\lambda_k)} \end{aligned}$$

where from the first to the second line we used that $(\lambda_k-\lambda_k)f(\lambda_k)=0$ and $g(\lambda_k)=0$ since λ_k is a pole. □

Proof of Theorem 4.3.4. We want to write $s(z)$ as

$$s(z) = q(z) + \sum_{k=0}^{m-1}\frac{A_m}{z-i\frac{y_k}{\beta}} + \frac{B_m}{z+i\frac{y_k}{\beta}} = q(z) + \sum_{k=0}^{m-1}\frac{\beta A_m}{\beta z - iy_k} + \frac{\beta B_m}{\beta z + iy_k}$$

where A_m and B_m are some (possibly complex-valued) constants.

Recall that

$$s(z) = \frac{p(\beta z)^2}{p(\beta z)^2 + p(-\beta z)^2} = \frac{1}{h(\beta z)^2+1}, \qquad h(\beta z) = \frac{p(-\beta z)}{p(\beta z)}.$$

Since the degrees of the numerator and denominator are equal, $q(z)$ must be a constant; in particular, we must have $\lim_{z\to\infty} s(z) = q(z)$. Since $p(z)$ is a degree d polynomial, we see $\lim_{z\to\infty} h(\beta z) = (-1)^m$, so $\lim_{z\to\infty} h(\beta z)^2 = 1$. Using this we have

$$q(z) = \lim_{z\to\infty} s(z) = \lim_{z\to\infty}\frac{1}{h(\beta z)^2+1} = \frac{1}{2}.$$

To find A_m and B_m, we invoke the residue method, as stated in Theorem 4.6.1. To Applying this to the above expression of $s(z)$, we have:

$$A_m = \frac{1}{\frac{d}{dz}(h(\beta z)^2+1)|_{z=iy_k/\beta}} = \frac{1}{2\beta h(iy_k)h'(iy_k)}.$$

We use the logarithmic derivative to differentiate $h(z)$:

$$\frac{d}{dz}\log(h(z)) = \frac{h'(z)}{h(z)} \implies h'(z)h(z) = h(z)^2\cdot\frac{d}{dz}\log(h(z)).$$

Using the definition of $h(z)$ and $p(z)$ we have:

$$\begin{aligned}\frac{d}{dz}\log(h(z)) &= \frac{d}{dz}\sum_{j=1}^{m}\log(-z+\rho^{j/m}) - \log(z+\rho^{j/m}) \\ &= \sum_{j=1}^{m}\frac{-1}{-z+\rho^{j/m}} - \frac{1}{z+\rho^{j/m}} \\ &= \sum_{j=1}^{m}\frac{-2\rho^{j/m}}{\rho^{2j/m}-z^2}.\end{aligned}$$

Plugging in iy_k we have

$$2\beta h(iy_k)h'(iy_k) = 2\beta h(iy_k)^2\sum_{j=1}^{m}\frac{-2\rho^{j/m}}{\rho^{2j/m}+y_k^2} = 4\beta\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y_k^2}$$

where we used $h(iy_k)^2 = -1$ by the choice of y_k. Plugging in $-iy_k$ we get the same expression for B_m:

$$B_m = \left(4\beta\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y_k^2}\right)^{-1}.$$

□

Bound on sum of coefficients: We first bound the largest pole.

Lemma 4.6.2. *Let y_{m-1} be the solution to*

$$\sum_{j=1}^{m}\arctan\Big(\frac{y}{\varepsilon^{j/m}}\Big) = \frac{\pi}{4} + (m-1)\frac{\pi}{2}.$$

Then $y_{m-1} \le \frac{4}{\pi}\frac{1}{1-\varepsilon^{1/m}} \le O(m)$

Proof. Using the identity $\arctan(x) = \frac{\pi}{2} - \frac{1}{x}$ for $x > 0$ we have

$$\begin{aligned}\sum_{j=1}^{m}\arctan\Big(\frac{y_{m-1}}{\varepsilon^{j/m}}\Big) = \frac{\pi}{4} + (m-1)\frac{\pi}{2} &\iff \sum_{j=1}^{m}\frac{\pi}{2} - \arctan\Big(\frac{y_{m-1}}{\varepsilon^{j/m}}\Big) = \frac{\pi}{4} \\ &\iff \sum_{j=1}^{m}\arctan\left(\frac{\varepsilon^{j/m}}{y_{m-1}}\right) = \frac{\pi}{4}.\end{aligned}$$

Now using that $\arctan(x) \le x$ we have that

$$\frac{\pi}{4} = \sum_{j=1}^{m}\arctan\left(\frac{\varepsilon^{j/m}}{y_{m-1}}\right) \le \sum_{j=1}^{m}\frac{\varepsilon^{j/m}}{y_{m-1}} \implies y_{m-1} \le \frac{4}{\pi}\sum_{j=1}^{m}\varepsilon^{j/m} \le \frac{4}{\pi}\frac{1}{1-\varepsilon^{1/m}}.$$

To arrive at the final expression, notice that

$$(1-\varepsilon) = (1-\varepsilon^{1/m})(1+\varepsilon^{1/m}+\varepsilon^{2/m}+\ldots\varepsilon^{m-1/m}) \leq (1-\varepsilon^{1/m})m.$$

Rearranging we have

$$\frac{1}{1-\varepsilon^{1/m}} \leq \frac{m}{1-\varepsilon} = O(m)$$

where we assume that $\varepsilon < 1/2$. □

Lemma 4.6.3. *Let y_k be the solutions to*

$$\sum_{j=1}^{m} \arctan\Big(\frac{y_k}{\varepsilon^{j/m}}\Big) = \frac{\pi}{4} + k\frac{\pi}{2}.$$

Then we have

$$\sum_{k=0}^{m-1}\left(\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y_k^2}\right)^{-1} \leq O(m).$$

Proof. We will show sum is dominated by the final term, i.e., the dominant term is

$$S_{m-1}^{-1} := \left(\sum_{j=1}^{m}\frac{\varepsilon^{j/m}}{\varepsilon^{2j/m}+y_{m-1}^2}\right)^{-1}.$$

First we upper bound this term and then we should the sum of the rest are the same order.

To upper bound S_{m-1}^{-1}, we want to lower bound S_{m-1}. We can rewrite S_{m-1} as:

$$\sum_{j=1}^{m}\frac{\varepsilon^{j/m}}{\varepsilon^{2j/m}+y_{m-1}^2} = \frac{1}{y_{m-1}}\sum_{j=1}^{m}\frac{\frac{\varepsilon^{j/m}}{y_{m-1}}}{(\frac{\varepsilon^{j/m}}{y_{m-1}})^2+1}.$$

From Theorem 4.6.2, we have an upper bound on y_{m-1}, so we have a lower bound on $\frac{1}{y_{m-1}}$. Our goal then is to lower bound the sum

$$\sum_{j=1}^{m}\frac{\frac{\varepsilon^{j/m}}{y_{m-1}}}{(\frac{\varepsilon^{j/m}}{y_{m-1}})^2+1}.$$

As shown in Theorem 4.6.2 we know $\sum_{j=1}^{m}\arctan\Big(\frac{\varepsilon^{j/m}}{y_{m-1}}\Big) = \frac{\pi}{4}$. We define $\theta_j = \arctan(\frac{\varepsilon^{j/m}}{y_{m-1}})$ then we have:

$$\sum_{j=1}^{m}\theta_j = \frac{\pi}{4}.$$

We can now think of this as the following optimization problem

$$\text{minimize } \sum_{j=1}^{m}\frac{\tan\theta_j}{\tan\theta_j^2+1} \quad \text{subject to } \sum_{j=1}^{m}\theta_j = \frac{\pi}{4} \text{ and } \theta_j \geq 0.$$

Simplifying the trigonometry, this is equivalent to

$$\text{minimize } \frac{1}{2}\sum_{j=1}^{m}\sin(2\theta_j) \quad \text{subject to } \sum_{j=1}^{m}\theta_j = \frac{\pi}{4} \text{ and } \theta_j \geq 0.$$

By the concavity of our objective function, we see that the minimum will occur at one of the vertices of the boundary, i.e., when one $\theta_k = \pi/4$ and the rest are 0. In this case we have

$$\frac{1}{2}\sum_{j=1}^{m}\sin(2\theta_j) = \frac{1}{2}\sin\Big(\frac{\pi}{2}\Big) = \frac{1}{2}.$$

Recalling our original expression, we have the upper bound $S_{m-1}^{-1} \leq 2y_{m-1}$.

Now we bound the rest of the sum. In particular, we will show

$$\sum_{i=0}^{m-2}\left(\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y_k^2}\right)^{-1} \leq \frac{2}{\pi}y_{m-1}.$$

First note that letting $f(y) = \sum_{j=1}^{m}\arctan(y/\rho^{j/m})$ we have

$$f'(y) = \sum_{j=1}^{m}\frac{1}{1+(y/\rho^{j/m})^2}\cdot\frac{1}{\rho^{j/m}} = \sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y^2}.$$

We can rewrite our target expression as $\sum_{i=0}^{m-2}\frac{1}{f'(y_k)}$. Now we want to invoke the inverse function theorem. Recall by choice of y_k we have $f(y_k) = \frac{\pi}{4} + i\frac{\pi}{2}$. Let $x_i = \frac{\pi}{4} + i\frac{\pi}{2}$, then by the inverse function theorem

$$\sum_{j=1}^{m-2}\frac{1}{f'(y_k)} = \sum_{j=1}^{m-2}(f^{-1}(x_i))'.$$

Now we will relate the right hand side as an integral. Note that $x_{i+1} - x_i = \frac{\pi}{2}$ for all i, so

$$\sum_{j=1}^{m-2}(f^{-1}(x_i))' = \frac{2}{\pi}\sum_{j=1}^{m-2}(x_{i+1}-x_i)(f^{-1}(x_i))'.$$

We interpret this as an integral. We claim that $(f^{-1}(x))'$ is an increasing function. Provided this fact:

$$\begin{aligned}
\frac{2}{\pi}\sum_{j=1}^{m-2}(x_{i+1}-x_i)(f^{-1}(x_i))' &\leq \frac{2}{\pi}\sum_{j=1}^{m-2}\int_{x_i}^{x_{i+1}}(f^{-1}(x))'dx \\
&= \frac{2}{\pi}\sum_{j=1}^{m-2}(f^{-1}(x_{i+1}) - f^{-1}(x_i)) \\
&= \frac{2}{\pi}(f^{-1}(x_{m-1}) - f^{-1}(x_1)) \\
&\leq \frac{2}{\pi}y_{m-1}.
\end{aligned}$$

The only outstanding thing to show is that $(f^{-1}(x))'$ is an increasing function. But recall that $(f^{-1}(x))' = \frac{1}{f'(y)}$ for $x = f(y)$. Since arctan is an increasing function and f is a sum of arctans, y increases as x increases. Thus, it suffices to show $\frac{1}{f'(y)}$ is increasing in y. We have an explicit form of $\frac{1}{f'(y)}$, and we see this is increasing in y.

Putting these bounds together with Theorem 4.6.2 we have

$$\sum_{k=0}^{m-1}\left(\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y_k^2}\right)^{-1} \leq O(y_{m-1}) \leq O(m).$$

□

Corollary 4.6.4. *Let y_k^* be the approximate poles as in Theorem 4.3.5, with $|y_k^* - y_k| \ll y_0$. Then*

$$\sum_{k=0}^{m-1}\left(\sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+(y_k^*)^2}\right)^{-1} \leq O(m).$$

Proof. We omit the details, but the core idea is that the perturbation is sufficiently small so that the relative change to the denominator is a constant factor. □

Proofs for numerical approximation of step function

Proof of Theorem 4.3.5. We want to find the zero of the function

$$f_m(y) = \sum_{j=1}^{m}\arctan(y/\rho^{j/m}) - (\pi/4 + m\pi/2).$$

Note that $f_m(0) < 0$, and for $y \geq 0$, $f_m'(y) = \sum_{j=1}^{m}\frac{\rho^{j/m}}{\rho^{2j/m}+y^2} > 0$. Moreover, let $y_\rho = \frac{4}{\pi}\sum_{j=1}^{m}\rho^{j/m}$. Since $m \leq d-1$, using the bound $\arctan(t) \geq \pi/2 - 1/t$, for $t > 0$, we obtain that

$$\sum_{j=1}^{m}\arctan\left(\frac{y_\rho}{\rho^{j/m}}\right) \geq d\frac{\pi}{2} - \frac{1}{y_\rho}\sum_{j=1}^{m}\rho^{j/m} = d\frac{\pi}{2} - \frac{\pi}{4} \implies f_m(y_\rho) \geq d\frac{\pi}{2} - \frac{\pi}{4} - \left(\frac{\pi}{4} + k\frac{\pi}{2}\right) \geq 0$$

We have established that f_m is a strictly increasing function with $f(0) < 0$ and $f(y_\rho) \geq 0$. Therefore, the root $y_k \in [0, y_\rho]$. By using the bisection algorithm, y_k can be approximated to error at most ρ in at most $O(\log(y_\rho/\xi))$ iterations, where each iteration cost at most $O(d)$ arithmetic operations. Moreover $y_\rho = \frac{4}{\pi}\sum_{j=1}^{m}\rho^{j/m} = O(d)$. Hence, all d poles can be approximated in time at most $O(d^2\log(d/\xi))$. □

Remark 16. One could alternatively use Newton's method to approximate the poles. This should give faster convergence, but since either method is independent of the size of our matrix $\mathbf{A}$ and gives polylog runtime with respect to the error parameters, we opted for the simpler one.

Proof of Theorem 4.3.6. To simplify notation, let $C(y) = 4\sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m}+y^2}$ so that:

$$s(x) = \frac{1}{2} + \sum_{k=0}^{m-1} C(y_k)^{-1} \cdot \left(\frac{2\beta x}{\beta^2 x^2 + y_k^2}\right), \qquad \hat{s}(x) = \frac{1}{2} + \sum_{k=0}^{m-1} C(y_k^*)^{-1} \cdot \left(\frac{2\beta x}{\beta^2 x^2 + (y_k^*)^2}\right).$$

Let $x \in \mathbb{R}$ and consider x/β. We have:

$$|\hat{s}(x/\beta) - s(x/\beta)| = \left|\sum_{k=0}^{m-1} \left(\frac{2x}{(x^2+(y_k^*)^2)C(y_k^*)}\right) - \left(\frac{2x}{(x^2+y_k^2)C(y_k)}\right)\right| \tag{4.34}$$

$$\leq \sum_{k=0}^{m-1} \left|\left(\frac{2x}{(x^2+(y_k^*)^2)C(y_k^*)}\right) - \left(\frac{2x}{(x^2+y_k^2)C(y_k)}\right)\right|. \tag{4.35}$$

We apply the mean value theorem to the function:

$$f(x,y) = \frac{2x}{(x^2+y^2)C(y)}.$$

Letting I_m be an interval with endpoints y_k and y_k^*, by the mean value theorem

$$|f(x,y_k^*) - f(x,y_k)| \leq |y_k - y_k^*| \cdot \sup_{y\in I_m} \left|\frac{\partial f}{\partial y}\right|.$$

Taking the derivative we have

$$\left|\frac{\partial f}{\partial y}\right| = \left|2x((x^2+y^2)C(y))^{-2} \cdot \big((x^2+y^2)C'(y) + 2yC(y)\big)\right| \tag{4.36}$$

$$= \left|f(x,y)\left(\frac{2y}{x^2+y^2} + \frac{C'(y)}{C(y)}\right)\right| \tag{4.37}$$

$$\leq |f(x,y)| \left(\left|\frac{2y}{x^2+y^2}\right| + \left|\frac{C'(y)}{C(y)}\right|\right) \tag{4.38}$$

$$\leq |f(x,y)| \left(\left|\frac{2}{y}\right| + \left|\frac{C'(y)}{C(y)}\right|\right). \tag{4.39}$$

Computing $C'(y)$, we have

$$\left|\frac{C'(y)}{C(y)}\right| = \left|\frac{\sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m}+y^2} \cdot \frac{2y}{\rho^{2j/m}+y^2}}{\sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m}+y^2}}\right| \leq \max_j \left|\frac{2y}{\rho^{2j/m}+y^2}\right| \leq \left|\frac{2}{y}\right|,$$

where to see the first inequality, we interpret this as a weighted average of $\frac{2y}{\rho^{2j/m}+y^2}$. Our last term to bound is $|f(x,y)|$. Using the fact that $(x-y)^2 \geq 0$, we have

$$|f(x,y)| = \left|\frac{2x}{x^2+y^2}\right| \left|\frac{1}{C(y)}\right| \leq \left|\frac{1}{y \cdot C(y)}\right|.$$

All together we have

$$\left|\frac{\partial f}{\partial y}\right| \leq |f(x,y)| \cdot \left|\frac{4}{y}\right|.$$

Applying this to our original expression we see

$$|\hat{s}(x/\beta) - s(x/\beta)| \leq \sum_{k=0}^{m-1} |y_k - y_k^*| \cdot \sup_{y \in I_m} \left|\frac{4}{y^2 \cdot C(y)}\right|. \tag{4.40}$$

Note this bound is independent of x. To bound the supremum, we note that $y^2 C(y)$ is a function increasing with y, so the supremum will be maximized for small values of y. We first find a lower bound for $C(y)$ for sufficiently small y.

Suppose $y \leq \rho$, then $y^2 \leq \rho^{2j/m}$ for all j so

$$C(y) = 4\sum_{j=1}^{m} \frac{\rho^{j/m}}{\rho^{2j/m} + y^2} \geq 4\sum_{j=1}^{m} \frac{\rho^{j/m}}{2\rho^{2j/m}} = 2\sum_{j=1}^{m} \frac{1}{\rho^{j/m}}.$$

Now recall that the smallest y we will consider is $y_0 - \rho$, corresponding to the smallest pole. We assume that $\rho \ll y_0$, so it suffices to bound the value of $y^2 C(y)$ for $y \geq y_0/2$. Recall that

$$\sum_{j=1}^{m} \arctan\left(\frac{y_0}{\rho^{j/m}}\right) = \frac{\pi}{4}.$$

Looking at the $j = d$ term and using the fact that $\arctan(x) \leq x$, we see

$$\frac{y_0}{\rho} \leq \frac{\pi}{4} \implies y_0 \leq \rho \implies y_0/2 \leq \rho.$$

This means we can apply our lower bound for $C(y)$ so we see

$$(y_0/2)^2 C(y_0/2) \geq \frac{y_0}{2}\sum_{j=1}^{m} \frac{y_0}{\rho^{j/m}} \geq \frac{y_0}{2}\sum_{j=1}^{m} \arctan(y_0 \rho^{-j/m}) = y_0 \frac{\pi}{8}.$$

Combining this with the above, we have

$$|\hat{s}(x) - s(x)| \leq O\left(\frac{m\xi}{(y_0 - \xi)}\right) = O\left(\frac{m\xi}{y_0}\right). \tag{4.41}$$

□

Chapter 5

Ground state energies of Schrödinger operators

In this chapter we turn to a question from mathematical physics on relating the ground state energies of analogous discrete and continuous Schrödinger operators. Though this is a question from mathematical physics, we approach it from a theoretical computer science perspective. We view the ground state energy as an optimization problem, using the variational characterization of the ground state energy, and construct (pseudo-)ground state functions for each operator from (pseudo-)ground state functions for the other.

5.1 Introduction

Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a smooth $\mathbb{Z}^d$-periodic potential and let $\theta \in (0, 1)$ be a parameter. We study the family of discrete Schrödinger operators $A_\theta : \ell_2(\theta\mathbb{Z}^d) \to \ell_2(\theta\mathbb{Z}^d)$ defined by

$$A_\theta f(k\theta) := \left(2df(k\theta) - \sum_{j\sim k} f(j\theta)\right) + V(k\theta)f(k\theta) = (L + V)f(k\theta), \qquad \text{for } k \in \mathbb{Z}^d \quad (5.1)$$

where $j \sim k$ indicates that $j, k \in \mathbb{Z}^d$ differ in exactly one coordinate by ± 1 and L is the discrete Laplacian of the corresponding graph on $\theta\mathbb{Z}^d$. This family includes several well-studied operators, such as the almost-Mathieu operator which corresponds to the special case

$$V(x) := \lambda(2 - 2\cos(2\pi x)), \qquad \text{for } \lambda > 0 \quad (5.2)$$

in $d = 1$. The operator A_θ is bounded and self-adjoint with $\sigma(A_\theta) \subseteq [0, 4d + \|V\|_\infty]$. We refer to the infimum of its spectrum as its *ground state energy*, denoted $\mu(A_\theta)$, which admits

the well-known variational characterization

$$\begin{aligned}\mu(A_\theta) &= \inf_{\substack{f\in\ell_2(\theta\mathbb{Z}^d)\\ \|f\|=1}} \langle f, A_\theta f\rangle \\ &= \inf_{\substack{f\in\ell_2(\theta\mathbb{Z}^d)\\ \|f\|=1}} \left(\sum_{j\sim k\in\mathbb{Z}^d} (f(k\theta)-f(j\theta))^2 + \sum_{k\in\mathbb{Z}^d} V(k\theta)f(k\theta)^2\right).\end{aligned} \tag{5.3}$$

We can consider the analogous semiclassical (or continuous) d-dimensional Schrödinger operator with potential V and semiclassical parameter θ, $B_\theta : H^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$, defined as

$$B_\theta f(x) := -\theta^2 \sum_{i=1}^d \frac{d^2}{dx_i^2} f(x) + V(x)f(x) = (-\theta^2\Delta + V)f(x). \tag{5.4}$$

B_θ is an unbounded selfadjoint nonnegative operator. Let $\mu(B_\theta)$ denote the ground state energy of B_θ. We then have the variational characterization (see e.g. [RS81, Chapter VIII])

$$\begin{aligned}\mu(B_\theta) &= \inf_{\substack{g\in H^2(\mathbb{R}^d)\\ \|g\|=1}} \langle g, B_\theta g\rangle \\ &= \inf_{\substack{g\in H^2(\mathbb{R}^d)\\ \|g\|=1}} \int_{\mathbb{R}^d} \theta^2 \|\nabla g(x)\|^2 + V(x)g(x)^2 dx.\end{aligned} \tag{5.5}$$

The purpose of this work is to establish a relationship between $\mu(A_\theta)$ and $\mu(B_\theta)$ for a wide class of V, that are non-asymptotic in θ.

Motivation

Motivated by questions in geometric group theory, there has been interest [BVZ97; BZ05; BDHMW17a; BDHMW17b; Oza22] in proving good upper and lower bounds on $\mu(A_\theta), \theta \in \left[0, \frac{1}{2}\right]$ for the special case (5.2) — roughly speaking, this quantity arises naturally in the representation theory of the Heisenberg group and controls the spectral gap of the random walk on that group. More recently, a line of work initiated by Ozawa [Oza16; KNO19; KKN21] introduced a method for producing semidefinite programming proofs of Kazhdan's Property (T) for various groups including $SL_n(\mathbb{Z})$, in which effective bounds on certain $\mu(A_\theta)$ play a role. The best known bounds on $\mu(A_\theta)$ for the almost-Mathieu operator due to [BZ05] rely on delicate trigonometric calculations, and as far as we are aware there is no systematic method for proving such bounds in general.

In contrast, in many cases $\mu(B_\theta)$ can be estimated using well-established tools of spectral theory — see e.g. [Kat52; GGS92] for examples of such bounds in 1 dimension, and Theorem 5.1.4 for coarse bounds in any dimension.

Establishing a relationship between $\mu(A_\theta)$ and $\mu(B_\theta)$ allows bounds in one setting to be transferred to the other. Our results are general, in that they apply for a wide class of potentials, and effective, in that they provide concrete bounds for fixed θ, as is required in applications to group theory mentioned above.

Statement of main results

We prove the following three results which articulate that while the spectra of A_θ and B_θ are clearly different, their ground state energies are in many cases comparable. With the exception of Theorem 5.1.1, we do not expect the bounds obtained here to be sharp in terms of their dependence on θ.

Upper bound on $\mu(A_\theta)$ for continuous periodic V, irrational θ. We refer to this direction as the "easy" direction. As the proofs will show, proving this relationship is much simpler and yields a clean bound, given below.

Theorem 5.1.1. *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be continuous and $\mathbb{Z}^d$-periodic. For θ irrational,*

$$\mu(A_\theta) \leqslant \mu(B_\theta).$$

The irrationality assumption above cannot be removed in general — see Section 5.3 for a concrete example and Remark 17 for a discussion. However the above result can be extended to all θ by instead considering the infimum of the *union spectrum* of A_θ (see Definition 5.3.1) instead of $\mu(A_\theta)$. The proof of Theorem 5.1.1 appears in Section 5.3.

For the reverse direction (which we refer to as the "hard" direction), we will require more conditions on V. We prove two different results, each requiring different conditions on V. The first is a general bound, for any V that satisfies the local condition given in (5.6). This condition states that the local discrete averages of V can be related to the local continuous averages of V, where local is over any "cube" defined by the lattice points $\theta\mathbb{Z}^d$. This result holds for any θ, provided the condition (5.6) (which is θ dependent) still holds.

The second result is for smooth, periodic V and roughly requires V to not grow too quickly near its minimums (a more precise definition is given in Definition 5.1.2). We note this result only holds for sufficiently small θ (in particular, we need sufficiently small θ to apply Lemma 5.4.9).

Lower bound on $\mu(A_\theta)$ for bounded V and any θ. For $y \in \theta\mathbb{Z}^d$, let

$$C(y) := y + \{0, \theta\}^d, \quad \& \quad \overline{C}(y) := y + [0, \theta]^d.$$

We then have the following comparison between $\mu(A_\theta)$ and $\mu(B_\theta)$ under the less restrictive assumptions that V is bounded, nonnegative, and its local averages on the cubes $C(y)$ and $\overline{C}(y)$ are comparable for all $y \in \theta\mathbb{Z}^d$.

Theorem 5.1.2. *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, not necessarily periodic with $\|V\|_\infty \leqslant 1$, and let $\theta \in (0, 1)$. Suppose that there exists $a = a(\theta) > 0$ such that for all $y \in \theta\mathbb{Z}^d$,*

$$\frac{1}{\theta^d} \int_{\overline{C}(y)} V(x)dx \leqslant a \cdot \frac{1}{2^d} \sum_{k \in C(y)} V(k). \tag{5.6}$$

Then

$$c_d \mu(A_\theta) \geq \mu(B_\theta)$$

where $c_d = \max\left\{\frac{3^d}{2^d} + 5 \cdot 3^d, 4 \cdot \frac{3^d}{2^d} \cdot a\right\}$.

Combined with Theorem 5.1.1 this establishes that for V satisfying the conditions of both theorems and θ irrational, $\mu(A_\theta)$ and $\mu(B_\theta)$ agree up to a multiplicative factor. The quantity c_d arises as the worst-case relationship between local Rayleigh quotients on individual cubes $C(y)$ and $\overline{C}(y)$. The requirement that $a > 0$ in particular rules out the possibility that $V(y) = 0$ for all $y \in \theta\mathbb{Z}^d$ while having $V \neq 0$ elsewhere, in which case we could have $0 = \mu(A_\theta) < \mu(B_\theta)$. The proof of Theorem 5.1.2 appears in Section 5.4.

Lower bound on $\mu(A_\theta)$ for smooth periodic V, small θ. Under stronger assumptions on V, we establish the reverse inequality $\mu(A_\theta) \geq \mu(B_\theta)(1 - o(1))$ as $\theta \to 0$. To state our result we introduce the following definitions.

Definition 5.1.1 (Critical point)**.** For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, let

$$D^\alpha := \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d}.$$

A point $m \in \mathbb{R}^d$ is a *critical point of order p* of V if $D^\alpha V(m) = 0$ for all α satisfying $\|\alpha\|_1 \leqslant p - 1$, and $D^\alpha V(m) \neq 0$ for some α satisfying $\|\alpha\|_1 = p$.

Definition 5.1.2 (Coercive)**.** A potential V is *P-coercive* with parameters (t_0, K, P), $K \geq 1$, if for every $t \leqslant t_0$ and $y \in \mathbb{R}^d$, $V(y) \leqslant t + \inf V$ implies that there exists a critical point m of order p with $2 \leqslant p \leqslant P$ such that $\|y - m\| \leqslant K t^{\frac{1}{p}}$. We say that V is *uniformly p-coercive* if it is p-coercive and all of its critical points have order exactly p.

Coercivity guarantees that if a function is close to its minimum value, then its input must be polynomially close to a critical point. Note that any real analytic periodic potential with finitely many local minima satisfies this property. The bound below depends on the parameters of coercivity as well as V's smoothness.

Theorem 5.1.3. *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be smooth, $\mathbb{Z}^d$-periodic, and P-coercive with bounded directional derivative in every sequence of P directions — i.e., for each $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leqslant P$, there is an upper bound on $D^\alpha V(x)$ independent of α and x. Then for θ sufficiently small,*

$$\left(1 + O\left(\theta^{\frac{1}{2(P+1)}}\right)\right)\mu(A_\theta) \geq \mu(B_\theta) - O\left(\theta^{2 - \frac{3P+1}{2P(P+1)}}\right).$$

The proof of Theorem 5.1.3 appears in Section 5.4, where a more precise version (Theorem 5.4.6) specifying the constants in the asymptotic notation is provided. The estimate above is nontrivial at least for V with isolated critical points: for such potentials, it can be

shown that the additive $O(\theta^{2-\frac{3P+1}{2P(P+1)}})$ term on the right hand side is of smaller order than $\mu(B_\theta)$ as $\theta \to 0$. In particular we prove the following estimates in Section 5.4 using simple arguments which we suspect are known but could not find in the literature.

Proposition 5.1.4. *Suppose $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a smooth $\mathbb{Z}^d$-periodic potential which is uniformly p-coercive for some $p \geq 2$ and has only isolated critical points. Then as $\theta \to 0$ we have*

(a) $\mu(B_\theta) = \Omega\left(\theta^{\frac{2p}{p+1}}\right)$ *if* $d = 1, 2$,

(b) $\mu(B_\theta) = \Omega\left(\theta^{\frac{2p}{p+2}}\right)$ *if* $d \geq 3$, *and*

(c) $\mu(B_\theta) = O\left(\theta^{\frac{2p}{p+2}}\right)$ *for all* $d \geq 1$,

where the implicit constants depend on V.

Example. As an example, consider the almost-Mathieu operator. The potential (5.2) is coercive with parameters $\left(4\Lambda, \frac{1}{2\pi\sqrt{\Lambda}}, 2\right)$, its Lipschitz constant is $4\pi\Lambda$, and its second derivative is bounded by $8\pi^2\Lambda$. Therefore we find from Theorem 5.4.6 that for all $0 < q < 1$

$$\mu(B_\theta) \leqslant \mu(A_\theta) + \frac{1}{1-\mu(A_\theta)}\left(8\pi\sqrt{\Lambda}\theta^{1+q/2} + \left(\mu(A_\theta) + \frac{8\pi\Lambda\theta^{1-q}}{1-4\pi\Lambda\theta^{1-q}}\right)\mu(A_\theta)\right).$$

This choice of V also obeys the conditions of Theorem 5.1.4, so from the choice $q = \frac{5}{6}$ and the constant in (5.33) being $\sqrt[4]{8}\pi\sqrt{\Lambda}$, we obtain the multiplicative upper bound

$$\frac{\mu(B_\theta)}{\mu(A_\theta)} \leqslant 1 + \frac{1}{1-\sqrt[4]{8}\pi\sqrt{\Lambda}\theta^2}\left(8\pi\sqrt{\Lambda}\theta^{\frac{1}{12}} + \sqrt[4]{8}\pi\sqrt{\Lambda}\theta^2 + \frac{8\pi\Lambda\theta^{\frac{1}{6}}}{1-4\pi\Lambda\theta^{\frac{1}{6}}}\right).$$

The lower bound of 1 follows from Theorem 5.1.1. We note that the large error bound of $O\left(\theta^{\frac{1}{12}}\right)$ (in comparison to [BWZ24]'s (5.7); see the following section for further discussion) is a lossy consequence of applying a Lipschitz bound for the potential on its entire domain.

In the setting of Theorem 5.1.2 and Theorem 5.4.2, when $\theta < \frac{1}{2}$, V has $a \leqslant \frac{2}{3}$; note that if $\theta = \frac{1}{2}$ such an a does not exist. Thus c_1 as in the Theorem statement is $\frac{33}{2}$ (we note that this can be improved to $c_1 = 9$ with modified choices of parameters in the proof which are preferable in $d = 1$).

Related work

The question of quantitative bounds on $\mu(A_\theta)$ in the case of the almost-Mathieu operator (5.2) was first investigated by Béguin, Valette, and Zuk [BVZ97] in the context of proving mixing time bounds for random walks on the Heisenberg group; they proved $\mu(A_\theta) = \Omega(1)$

for θ in a neighborhood of $\frac{1}{2}$ and conjectured a constant lower bound for $\theta \in \left[\frac{1}{4}, \frac{1}{2}\right]$. This was significantly improved by Boca and Zaharescu [BZ05] who proved nonasymptotic bounds in the full range $\theta \in \left[0, \frac{1}{2}\right]$ via an elementary trigonometric calculation (see also Ozawa [Oza22] for a simplified proof of a slightly weaker result). Bump, Diaconis, Hicks, Miclo, and Widom [BDHMW17a; BDHMW17b] gave alternative proofs of similar bounds based on an uncertainty principle for circulant matrices and by a comparison to the harmonic oscillator spectrum via a limiting argument.[1] Our results do not improve the best known bound of [BZ05], but provide an alternate and more conceptual route to proving such bounds.

There has been an extensive study of the fine structure of the spectrum of discrete Schrödinger operators in the mathematical physics community, but surprisingly few works giving quantitative bounds on the edge of the spectrum. Davies [Dav99] studied the numerical problem of estimating $\mu(A_\theta)$ and observed that it is related to subtle dynamical phenomena. Akkouche [Akk10] gave a necessary and sufficient condition for the ground state energy of a discrete Schrödinger operator $-L + \lambda V$ on $\mathbb{Z}^d$ to be positive for a large (not necessarily periodic or nonnegative) class of potentials V, for small λ: V merely needs to have mean bounded away from 0 on all sufficiently large intervals. This is related to the condition (5.6) but holds only on some fixed macroscopic scale, whereas $a(\theta)$ is posited to exist for a chosen scale θ. Milatovic [Mil13] generalized this result to other bounded degree graphs.

We will not attempt to survey the vast literature on the spectrum of Schrödinger operators on $\mathbb{R}^d$; most relevant for this paper are the works by Kato [Kat52], Gesztesy, Graf, and Simon [GGS92], and Arendt and Barry [AB96; AB97], who proved bounds on $\mu(B_\theta)$ in $d = 1$ for various classes of potentials. More broadly, there appear to be no results in the literature relating spectral bounds for discrete and continuous Schrödinger operators. Establishing such bounds, as is done here, could provide a fruitful bridge between the two areas.

Motivated by this work, Becker, Wittsten, and Zworski [BWZ24] used semiclassical methods to significantly improve some aspects of Theorem 5.1.3. They establish

$$\frac{\mu(A_\theta)}{\mu(B_\theta)} = 1 + O_d(\theta) \qquad \text{as } \theta \to 0 \tag{5.7}$$

for all periodic smooth confining potentials $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ where $O_d(\cdot)$ hides constants and d-dependent factors, with a full asymptotic expansion in θ when V satisfies further nondegeneracy assumptions. Their result is stronger than Theorem 5.1.3 in that it obtains a smaller ratio as a function of θ, but weaker in that it does not provide explicit estimates for fixed θ. The classes of potentials considered in the two theorems are incomparable: [BWZ24] requires V to be confining, whereas Theorem 5.1.3 requires V to be coercive with critical points of finite order. Theorem 5.1.2 proves a much weaker bound than (5.7), but requires fewer assumptions on V, and Theorem 5.1.1 proves a stronger upper bound of 1 for irrational θ.

[1] This latter argument only gave asymptotic bounds as $\theta \to 0$.

5.2 Preliminaries and overview of approach

Before giving an overview of our approach, we define the *Rayleigh quotient* and ε-pseudo-ground-states.

Rayleigh quotient and ε-pseudo-ground-states. We introduce the *Rayleigh quotient* for any operator T and function f in the domain of T, $D(T)$:

$$\mathcal{R}_T(f) = \frac{\langle f, Tf\rangle}{\langle f, f\rangle}.$$

For $f \in D(T)$, we say that f is a *ε-pseudo-ground-state* or *ε-p.g.s.* if $\mathcal{R}_T(f) \leqslant \mu(T) + \varepsilon$. When T is either A_θ or B_θ and f is an ε-p.g.s. of T, we may assume that f is non-negative by considering $|f|$, which is easily seen to have a Rayleigh quotient less than or equal to that of f. We will sometimes want to assume f is positive; in this case we add a small everywhere-positive perturbation from $D(T)$ which increases the Rayleigh quotient by an arbitrarily small amount.

Approach. Our approach is simple and inspired by methods in theoretical computer science. For the upper bound, given any $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, we exhibit $f : \theta\mathbb{Z}^d \to \mathbb{R}_{\geq 0}$ such that $\mathcal{R}_{A_\theta}(f) \leqslant \mathcal{R}_{B_\theta}(g)$. The naive construction is to construct f by sampling g on $\theta\mathbb{Z}^d$. However, we can never hope to construct a "good" test function f with this method since we are only sampling g on a null set. Instead, we will sample g many times across many *shifted* lattices: $\theta\mathbb{Z}^d + \eta$. We will show that at least one of these shifts give a discrete function f with low energy for θ irrational.

Similarly for the lower bound, given a ε-pseudo-ground-state function f for A_θ, we will try to construct a function g for B_θ such that $\mathcal{R}_{B_\theta}(g) \leqslant c\mathcal{R}_{A_\theta}(f)$ for some constant c (as we will see, we will be unable to have $c = 1$ as in the upper bound). The most natural way to construct g is through interpolation of f, which is defined on the lattice points $\theta\mathbb{Z}^d$. It will be more challenging to relate the energy these functions, so we will need to require certain conditions on V.

Preliminaries. One main technical tool we will use in the second method for the "hard direction" is discrete Fourier analysis (or analysis of Boolean functions). In particular, we will often consider functions restricted to $C(y)$ for some choice of $y \in \theta\mathbb{Z}^d$, and it will be useful to view these as functions defined on a hypercube.

To that end, we recall some facts from Fourier analysis on a hypercube: consider a function F defined on the d-dimensional hypercube $\{-1, 1\}^d$ with the standard edge set. The functions $\chi_S(x) := \prod_{i \in S} x_i$ for $S \subseteq [d]$ form an orthogonal basis for the space of functions on $\{-1, 1\}^d$ with respect to the inner product

$$\langle f, g\rangle := \sum_{x \in \{-1,1\}^d} f(x)g(x).$$

We can then write $F(x) = \sum_{S \subseteq [d]} \widehat{F}(S)\chi_S(x)$. From [ODo14] we have

$$\mathbb{E}\left[F(x)^2\right] = \sum_{S \subseteq [d]} \widehat{F}(S)^2, \tag{5.8}$$

$$\mathbb{E}\left[F(x) \cdot LF(x)\right] = 2 \sum_{S \subseteq [d]} |S| \, \widehat{F}(S)^2, \text{ and} \tag{5.9}$$

$$\langle F, LF \rangle = 2^{d+1} \sum_{S \subseteq [d]} |S| \, \widehat{F}(S)^2 \tag{5.10}$$

where the expectation is taken over x drawn uniformly at random from $\{-1, 1\}^d$ and L is the unnormalized graph Laplacian: $Lf(x) := 2df(x) - \sum_{y \sim x} f(y)$.

Notation. We use $\|\cdot\|$ to denote the ℓ_2 and L^2 norms of functions, and the operator norms of operators. We use $\mathbb{E}$ to denote the expectation of a random variable, and $\mathbb{E}_x\left[f(x, y)\right]$ to denote the partial expectation of an expression with respect to the random variable x only, treating the other variables as constants.

5.3 Easy direction

In this section we prove Theorem 5.1.1. Recall that this requires that θ be irrational. We start with an explicit example to show that in general, we cannot remove the irrationality assumption.

Motivating example. The core issue is that B_θ is invariant to shifts of the potential, but A_θ is not. Consider the following two potentials:

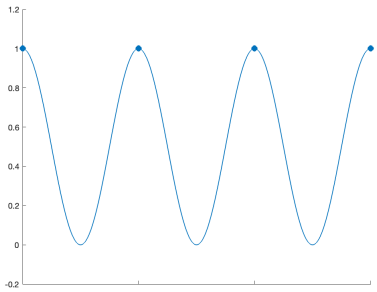


Figure 5.1: Potential V_1 over $\mathbb{R}$ and sampled points on $\theta\mathbb{Z}$ (given by dots).

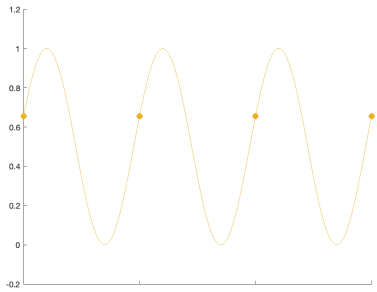


Figure 5.2: Potential V_2 over $\mathbb{R}$ and sampled points on $\theta\mathbb{Z}$ (given by dots).

Consider V_1 and V_2 given in Figure 5.1 and Figure 5.2. Note that V_2 is simply a shfited version of V_1, and since B_θ is shift invariant, letting $B_{\theta,i}$ correspond to the B_θ operator with potential V_i, $\mu(B_{\theta,1}) = \mu(B_{\theta,2})$. However, we can see that A_θ definitely not shift invariant, particularly with this choice of θ. $V_1|_{\theta\mathbb{Z}} = 1$ while $V_2|_{\theta\mathbb{Z}} \approx 0.65$. Since both of these are constants, we see $\mu(A_{\theta,1}) = 1$ and $\mu(A_{\theta,2}) \approx 0.65$, where $A_{\theta,i}$ corresponds to the A_θ operator with potential V_i. Since $\mu(A_\theta)$ depends so heavily on the shift of of the potential and $\mu(B_\theta)$ is completely invariant to it, it is unsurprising that in this setting we cannot relate $\mu(A_\theta)$ and $\mu(B_\theta)$.

Note that since V is assumed to be $\mathbb{Z}$-periodic, by forcing θ to be irrational, we can avoid this constant behavior of V on $\theta\mathbb{Z}$. In fact, requiring θ to be irrational will force $\mu(A_\theta)$ to similarly be invariant to shifts of the potential function (see Lemma 5.3.1 for details).

Constructing test functions.

As stated above, for any $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, we exhibit $f : \theta\mathbb{Z}^d \to \mathbb{R}_{\geq 0}$ such that $\mathcal{R}_{A_\theta}(f) \leqslant \mathcal{R}_{B_\theta}(g)$. The naive construction is to construct f by sampling g on $\theta\mathbb{Z}^d$. However, we can never hope to construct a "good" test function f with this method since we are only sampling g on a null set.

The key observation is that for θ irrational, ergodicity implies there is *a priori* no reason to prefer $\theta\mathbb{Z}^d \subset \mathbb{R}^d$ over any other shift $\theta\mathbb{Z}^d + \eta \subset \mathbb{R}^d$. We show by averaging over all such translations that there exists a translation giving rise to a satisfactory f.

Let $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. Let $\eta \in [0, \theta]^d$, drawn uniformly at random with each coordinate independent. We define the test function f_η as $f_\eta(k\theta) := g(k\theta + \eta)$.

Since we sample g at $\theta\mathbb{Z}^d + \eta$ to construct f_η, it is only natural to sample V on the same lattice. To this end define $A_{\theta,\eta}$, a discrete operator on functions defined on $\theta\mathbb{Z}^d$, via

$$A_{\theta,\eta} f(k\theta) := 2df(k\theta) - \sum_{\substack{j,k\in\mathbb{Z}^d \\ j\sim k}} f(j\theta) + \sum_{k\in\mathbb{Z}^d} V(k\theta + \eta) f(k\theta).$$

We aim to show that there exists an η so $\mathcal{R}_{A_{\theta,\eta}}(f_\eta) \leqslant \mathcal{R}_{B_\theta}(g)$. However, this will not be sufficient to relate $\mu(A_\theta)$ and $\mu(B_\theta)$ unless we can first relate $\mu(A_\theta)$ and $\mu(A_{\theta,\eta})$. This is precisely where we will need to use the irrationality condition on θ.

Relating $\mu(A_\theta)$ and $\mu(A_{\theta,\eta})$

We have the following lemma, relating $\mu(A_{\theta,\eta})$ and $\mu(A_\theta)$, which is known (appearing in e.g. [Jit21] for the almost-Mathieu operator), but we include a proof for completeness.

Lemma 5.3.1. *If $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is continuous and periodic and θ is irrational, then $\mu(A_{\theta,\eta})$ is independent of η.*

Proof. Observe that

$$\|A_{\theta,0} - A_{\theta,\eta}\| \leqslant \sup_{n\in\mathbb{Z}^d} |V(n\theta+\eta) - V(n\theta)| \leqslant \omega(\eta)$$

where $\omega(\cdot)$ is the modulus of continuity of V, so $\mu(A_{\theta,\eta})$ is continuous in η.

Let $\varepsilon > 0$ and $\eta \in \mathbb{R}^d$. Since θ is irrational, ergodicity in each coordinate of $n\theta+\eta$ implies that there is some $\eta' \in \mathbb{R}^d$ and $n_0 \in \mathbb{Z}^d$ such that $|\eta' - \eta| \leqslant \varepsilon$ and $n_0\theta \equiv \eta' \pmod 1$ (solve for an appropriate entry of n_0 separately for each component). Observe that $A_{\theta,\eta'} = A_{\theta,0}$ since the Laplacian part of A_θ is translation-invariant and we have $V(n_0\theta+n\theta) = V(n\theta+\eta')$. By continuity we now have $|\mu(A_0) - \mu(A_\eta)| \leqslant \varepsilon$. Letting $\varepsilon \to 0$ yields the result. □

Proof of Theorem 5.1.1

Proof of Theorem 5.1.1. Given $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, define f_η as above, i.e. $f_\eta(k\theta) := g(k\theta+\eta)$.

By above, for θ irrational, it is sufficient to relate $\mu(A_{\theta,\eta})$ and $\mu(B_\theta)$. We claim that it suffices to show

$$\frac{\mathbb{E}\left[\langle f_\eta, A_{\theta,\eta} f_\eta\rangle\right]}{\mathbb{E}\left[\|f_\eta\|^2\right]} \leq \frac{\left[\langle g, B_\theta g\rangle\right]}{\left[\|g\|^2\right]}. \tag{5.11}$$

Note that if X and Y are non-negative random variables with $\mathbb{E}[|X|] < \infty$ and $X \neq 0$ almost surely, then with nonzero probability

$$\frac{Y}{X} \leqslant \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}. \tag{5.12}$$

By Section 5.2, we can assume that $f > 0$, so $\|f_\eta\|^2 > 0$ almost surely, and we can apply (5.12). In particular, (5.11) implies that there exists an η such that

$$\frac{\left[\langle f_\eta, A_{\theta,\eta} f_\eta\rangle\right]}{\left[\|f_\eta\|^2\right]} \leq \frac{\left[\langle g, B_\theta g\rangle\right]}{\left[\|g\|^2\right]}.$$

Minimizing over all choices of g, we can then conclude there exists an η such that $\mu(A_{\theta,\eta}) \leq \mu(B_\theta)$. Finally by Lemma 5.3.1, $\mu(A_\theta) \leqslant \mu(B_\theta)$.

We now turn to proving (5.11). Expanding the numerator, we have

$$\frac{\mathbb{E}\left[\langle f_\eta, A_{\theta,\eta} f_\eta\rangle\right]}{\mathbb{E}\left[\|f_\eta\|^2\right]} = \frac{\mathbb{E}\left[\langle f_\eta, Lf_\eta\rangle\right] + \mathbb{E}\left[\langle f_\eta, V_\eta f_\eta\rangle\right]}{\mathbb{E}\left[\|f_\eta\|^2\right]}.$$

We will evaluate each quantity appearing in the right hand side individually.

First we bound the expected norm and potential term, both of which are straightforward.

Claim 5.13. The norms of f_η and g satisfy $\mathbb{E}_\eta\left[\|f_\eta\|^2\right] = \frac{1}{\theta^d}\|g\|^2$.

Proof. This follows from directly evaluating this expectation:

$$\begin{aligned}\mathbb{E}_\eta\left[\|f_\eta\|^2\right] &= \mathbb{E}_\eta\left[\sum_{k\in\mathbb{Z}^d} f_\eta(k\theta)^2\right] = \sum_{k\in\mathbb{Z}^d}\mathbb{E}_\eta\left[g(k\theta+\eta)^2\right] \\ &= \sum_{k\in\mathbb{Z}^d}\mathbb{E}_{\eta_1}\left[\cdots\mathbb{E}_{\eta_d}\left[g(k\theta+\eta)^2\right]\cdots\right] = \frac{1}{\theta^d}\|g\|^2.\end{aligned}$$

□

Claim 5.14. The quadratic forms of f_η and g with the respective potential operators satisfy

$$\mathbb{E}_\eta\left[\langle f_\eta, Vf_\eta\rangle\right] = \frac{1}{\theta^d}\langle g, Vg\rangle.$$

Proof. Again, we directly evaluate this expectation

$$\begin{aligned}\mathbb{E}\left[\langle f_\eta, Vf_\eta\rangle\right] &= \mathbb{E}_\eta\left[\sum_{k\in\mathbb{Z}^d} V(k\theta+\eta)f_\eta(k\theta)^2\right] \\ &= \sum_{k\in\mathbb{Z}^d}\mathbb{E}_\eta\left[V(k\theta+\eta)g(k\theta+\eta)^2\right] \\ &= \frac{1}{\theta^d}\langle g, Vg\rangle.\end{aligned}$$

□

In bounding the Laplacian term, the following segmentation will be useful in our analysis when we want to vary only one coordinate: for $k\in\mathbb{Z}^d, 1\leqslant i\leqslant d, \eta\in[0,\theta)^d$, let

$$I^\eta_{k,i} := \{x : x_j = (k\theta+\eta)_j\ \forall j\neq i,\ k_i\theta+\eta_i \leqslant x_i < (k_i+1)\theta+\eta_i\}.$$

Letting $\{e_i\}_{i=1}^d$ be the coordinate basis of $\mathbb{R}^d$, we can also express these sets as

$$I^\eta_{k,i} = (k\theta+\eta) + [0,\theta)e_i.$$

Given k and i, let $k_{\neg i} := (k_1,\ldots,k_{i-1},k_{i+1},\ldots,k_d)\in\mathbb{Z}^{d-1}$. For $k'\in\mathbb{Z}^{d-1}$ define

$$J^\eta_{k',i} := \bigcup_{\substack{k\in\mathbb{Z}^d\\ k_{\neg i}=k'}} I^\eta_{k,i}.$$

Immediately from this we have

$$\bigsqcup_{\substack{0\leqslant\eta_j<\theta\\ j\neq i}} J^\eta_{k',i} = \bigcup_{\substack{k\in\mathbb{Z}^d\\ k_{\neg i}=k'}}\overline{C}(\theta k), \qquad \text{therefore} \qquad \bigsqcup_{k'\in\mathbb{Z}^{d-1}}\bigsqcup_{\substack{0\leqslant\eta_j<\theta\\ j\neq i}} J^\eta_{k',i} = \mathbb{R}^d. \tag{5.15}$$

Note $J^\eta_{k',i}$ is independent of η_i but is dependent on each η_j, $j\neq i$. See Figure 5.3 for examples of $I^\eta_{k,i}$ and $J^\eta_{k',i}$ and Figure 5.4 for an illustration of (5.15).

Equipped with this, we bound the expectation of the Laplacian term.

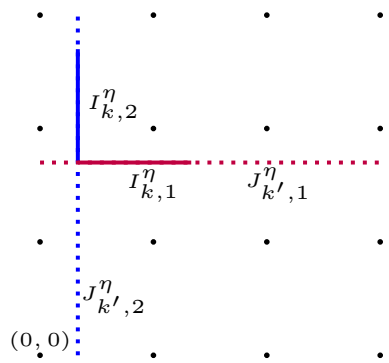


Figure 5.3: In $d = 2$, $I^\eta_{k,i}$ (solid) and $J^\eta_{k',i}$ (dotted) for $i = 1$ (in purple) and $i = 2$ (in blue), $k = (0, 1)$, and $\eta = \left(\frac{1}{3}\theta, \frac{7}{10}\theta\right)$.

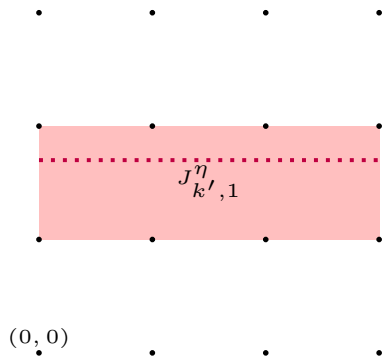


Figure 5.4: In $d = 2$, $J^\eta_{k',1}$ (purple, dotted) and $\bigsqcup_{0\leqslant\omega_2<\theta} J^\omega_{k',1}$ (pink) for $k' = k_2 = 1$. Note that this union is independent of $\eta = \left(\frac{1}{3}\theta, \frac{7}{10}\theta\right)$.

Claim 5.16. The quadratic forms of f_η and g with the respective Laplacian operators satisfy

$$\mathbb{E}_\eta\left[\langle f_\eta, Lf_\eta\rangle\right] \leqslant \frac{1}{\theta^{d-2}} \|\nabla g\|^2 .$$

Proof. First we evaluate $\langle f_\eta, Lf_\eta\rangle$. Using the definition of f_η and Cauchy-Schwarz, we see

$$\begin{aligned}
\langle f_\eta, Lf_\eta\rangle &= \sum_{k\in\mathbb{Z}^d}\sum_{i=1}^{d} \left(f_\eta((k+e_i)\theta) - f_\eta(k\theta)\right)^2 \\
&= \sum_{i=1}^{d}\sum_{k\in\mathbb{Z}^d} \left(g((k+e_i)\theta+\eta) - g(k\theta+\eta)\right)^2 \\
&= \sum_{i=1}^{d}\sum_{k\in\mathbb{Z}^d} \left(\int_{I^\eta_{k,i}} \frac{\partial}{\partial x_i} g(x)dx\right)^2 \\
&\leqslant \sum_{i=1}^{d}\sum_{k\in\mathbb{Z}^d} \theta \int_{I^\eta_{k,i}} \left(\frac{\partial}{\partial x_i} g(x)\right)^2 dx
\end{aligned}$$

Then by Tonelli's theorem, the invariance of $J^{\eta}_{k',i}$ with respect to η_i, and (5.15), we have

$$\begin{aligned}
\mathbb{E}_\eta\left[\langle f_\eta, Lf_\eta\rangle\right] &\leqslant \theta \sum_{i=1}^{d} \sum_{k'\in\mathbb{Z}^{d-1}} \mathbb{E}_\eta\left[\int_{J^{\eta}_{k',i}} \left(\frac{\partial}{\partial x_i} g(x)\right)^2 dx\right] \\
&= \theta \sum_{i=1}^{d} \mathbb{E}_\eta\left[\sum_{k'\in\mathbb{Z}^{d-1}} \int_{J^{\eta}_{k',i}} \left(\frac{\partial}{\partial x_i} g(x)\right)^2 dx\right] \\
&= \theta \sum_{i=1}^{d} \mathbb{E}_{(\eta_1,\ldots,\eta_{i-1},\eta_{i+1},\ldots,\eta_d)}\left[\sum_{k'\in\mathbb{Z}^{d-1}} \int_{J^{\eta}_{k',i}} \left(\frac{\partial}{\partial x_i} g(x)\right)^2 dx\right] \\
&= \frac{\theta}{\theta^{d-1}} \sum_{i=1}^{d} \int_{\mathbb{R}^d} \left(\frac{\partial}{\partial x_i} g(x)\right)^2 dx \\
&= \frac{1}{\theta^{d-2}} \left\|\nabla g\right\|^2.
\end{aligned}$$

□

Combining the above claims we have

$$\frac{\mathbb{E}_\eta\left[\langle f_\eta, A_{\theta,\eta} f_\eta\rangle\right]}{\mathbb{E}_\eta\left[\langle f_\eta, f_\eta\rangle\right]} \leqslant \left(\frac{1}{\theta^{d-2}}\left\|\nabla g\right\|^2 + \frac{1}{\theta^d}\langle g, Vg\rangle\right) \frac{\theta^d}{\left\|g\right\|^2} = \frac{\theta^2\left\|\nabla g\right\|^2 + \langle g, Vg\rangle}{\left\|g\right\|^2} = \mathcal{R}_{B_\theta}(g).$$

As we have shown (5.11), we are done. □

The above proof considers the infimum of the spectrum of the operator A_θ. However, we can instead consider its *union spectrum*, which is more natural in some contexts.

Definition 5.3.1 (Union Spectrum, cf. [MJ17])**.** The *union spectrum* of A_θ is

$$S_+(A_\theta) := \bigcup_{\eta} \sigma(A_{\theta,\eta}).$$

Note that if θ is irrational, then by Lemma 5.3.1, $\inf S_+(A_\theta) = \mu(A_\theta)$. By considering the infimum of the union spectrum, we can extend Theorem 5.1.1 to all θ:

Corollary 5.3.2. *For all $\theta > 0$,* $\inf S_+(A_\theta) \leqslant \mu(B_\theta)$.

Proof. By the above proof we know that there exists η such that

$$\mu(A_{\theta,\eta}) \leqslant \frac{\langle f_\eta, A_{\theta,\eta} f_\eta\rangle}{\langle f_\eta, f_\eta\rangle} \leqslant \frac{\langle g, B_\theta g\rangle}{\langle g, g\rangle}.$$

Minimizing over g and recalling the definition of $S_+(A_\theta)$, the claim is shown. □

For choices of V such that $\mu(A_\theta)$ is continuous with respect to θ, this result extends immediately to rational θ by continuity.

Corollary 5.3.3. *For V such that $\mu(A_\theta)$ is continuous with respect to θ, for all θ*

$$\mu(A_\theta) \leqslant \mu(B_\theta).$$

Remark 17. Choices of V satisfying Theorem 5.3.3 are not immediately apparent. A notable counterexample is the almost-Mathieu operator. Let $V_\eta(x) = 2 - 2\cos(2\pi x + \eta)$ and $V = V_0$. Then for $\theta = 1$ and any η, we have $V_\eta(k\theta) = V(\eta)$ for all k, so $\mu(A_{\theta,\eta}) = V(\eta)$. Now consider $\theta' = 1 - \varepsilon$ where ε is a small positive irrational (so θ' is irrational). Then for $\eta = 0$, we have $V(k\theta') = V(k(1-\varepsilon)) = V(-k\varepsilon) = V(k\varepsilon)$, so $\mu(A_{1-\varepsilon}) = \mu(A_\varepsilon)$. By Theorem 5.1.4 and Theorem 5.1.1, $\mu(A_\varepsilon) \leqslant O(\varepsilon)$, so as $\varepsilon \to 0$ along a sequence of irrationals, $\mu(A_{\theta'}) \to 0$. Since θ' is irrational, $\mu(A_{\theta',\eta}) \to 0$ for all η, so we see $\mu(A_{\theta,\eta})$ is not continuous in θ at $\theta = 1$ for any η such that $V(\eta) > 0$ (as $\mu(A_{1,\eta}) = V(\eta)$). By considering e.g. $2 - 2\cos(2\pi k x)$ for integers $k \geq 1$, one can also create discontinuities at $\theta = \frac{1}{k}$.

5.4 Hard direction

In this section we prove Theorem 5.1.3. Whereas the previous direction of the proof used the probabilistic method to find a pseudo-ground-state for A_θ from one for B_θ, here we deterministically construct a pseudo-ground-state for B_θ from one for A_θ. Our construction will simply be the multilinear interpolation of a pseudo-ground-state function f. In $d = 1$ this amounts just to "connecting" the function values by line segments. In higher dimensions, on each hypercube in $\theta\mathbb{Z}^d$, the interpolating function is the unique multilinear function in each variable which equals the given function on the corners of the hypercube. The techniques brought to bear are first Boolean Fourier-analytic, in terms of understanding both local and global behavior of the functions, and then a careful consideration of the potential's polynomial behavior near its minimizers.

It will be significantly more difficult to relate the energies of these functions than in the previous direction. We start with two motivating examples, to show where these difficulties can arise.

Motivating examples

Example 1: Hidden high potential regions. Consider the following discrete potential function and low-energy test function.

Consider the discrete potential function given in blue in Figure 5.5. This potential has a relatively large low potential region (or a "well") where the test function (in red) places a large amount of mass. Our discrete test function will not incur a large increase in energy by placing a large amount of mass in this region because the discretized potential is low here.

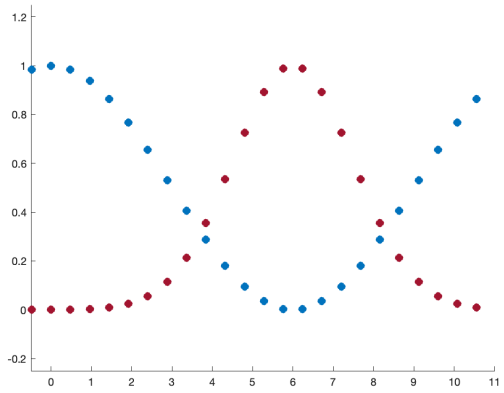


Figure 5.5: Discrete potential function (blue) and a low-energy test function (red).

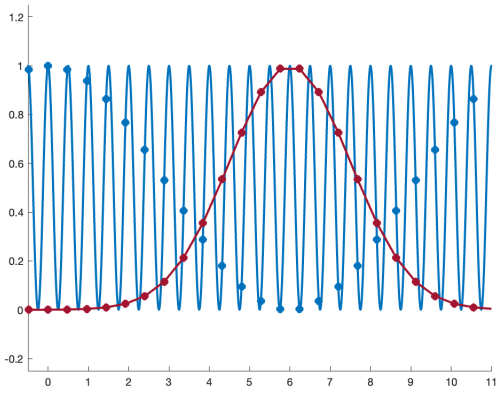


Figure 5.6: Continuous potential (blue) and interpolation of discrete function (red).

In contrast, consider the continuous potential function given in blue in Figure 5.6. We see that the discretized potential function was in fact quite misleading and obscured many high potential regions. If we consider the linear interpolation of the discrete test function, we see that we will be heavily penalized for placing a large amount of mass in these high potential regions.

The core issue here was that the discretized potential is not representative of the continuous potential. To combat this, we must impose some conditions on the potential. In our first method, we will require that we can relate the average of the endpoints of any interval to the continuous average over the whole interval, effectively forcing the discretized potential to be representative of the continuous potential. In the second method, we require that the potential cannot grow too quickly near a minimum. This means that if the discretized potential is low at some point, the continuous neighborhood around it must also have relatively low potential.

Example 2: Introducing non-zero terms. Recall that in the previous section, we directly related the potential term of the discrete test function (or at least the expectation over the potential term over all test functions) to the potential term of the continuous function. In the following section, we will want to do something similar. Letting g be the multilinear extension of f (which we can think of simply as the linear interpolation for now), we would like to relate $\langle g, Vg\rangle$ to $\langle f, Vf\rangle$. In particular, we would like to relate these for each interval.

However, we will be unable to do this. Consider the function given in Figure 5.7. Notice that at the left endpoint of the interval $f(x) = 0$, and at the right endpoint, the potential is 0. We see that overall then, $\langle f, Vf\rangle_I = 0$, where $\langle f, Vf\rangle_I$ denotes the potential term restricted to the interval I. However, we see that $\langle g, Vg\rangle_I$ is definitively not 0. This means

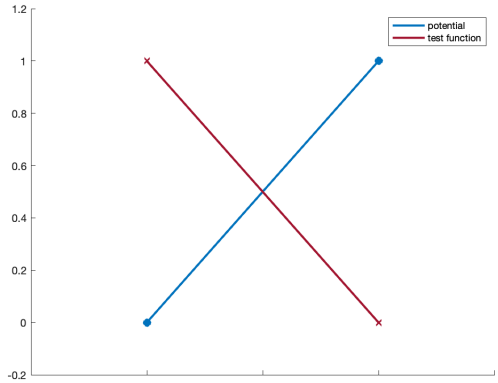


Figure 5.7: A sample discrete test function and its linear interpolation (blue), as well as a sample potential function (red) across one interval.

that we will be unable to upper bound $\langle g, Vg\rangle_I$ by any constant multiple of $\langle f, Vf\rangle_I$.

Instead, we will bound $\langle g, Vg\rangle$ by both $\langle f, Vf\rangle$ and $\langle f, Lf\rangle$. Roughly, if f is relatively constant across an interval, we will be able to relate $\langle g, Vg\rangle$ and $\langle f, Vf\rangle$; otherwise, we bound $\langle g, Vg\rangle$ by $\langle f, Lf\rangle$. See Lemma 5.4.5 for details.

Constructing the test function

In both methods, we will use the same construction: the multilinear extension of a given discrete function f. We outline that construction here.

First we set up some notation and basic facts. For $y \in \theta\mathbb{Z}^d$, $C(y)$ identifies naturally with the vertices of the hypercube $\{-1, 1\}^d$, and we let $E(C(y))$ be the edge set of the subgraph induced by the vertices in $C(y)$.

Let f be a positive bounded function on $\theta\mathbb{Z}^d$. Let g be the piecewise multilinear extension of f, defined on $\mathbb{R}^d$ as

$$g(x_1, \ldots, x_d) := \mathbb{E}_{x_1,\ldots,x_d}\left[f(k_{x_1}\theta, \ldots, k_{x_d}\theta)\right] \tag{5.17}$$

where each k_{x_i} is independent with the following distribution for $k_i\theta \leqslant x_i \leqslant (k_i+1)\theta$:

$$\mathbb{P}(k_{x_i} = k_i) = \frac{(k_i+1)\theta - x_i}{\theta},$$
$$\mathbb{P}(k_{x_i} = k_i+1) = \frac{x_i - k_i\theta}{\theta} = 1 - \mathbb{P}(k_{x_i} = k_i).$$

The function g obtained in this way is continuous on each closed hypercube, and is continuous on all of $\mathbb{R}^d$ since g is defined uniquely on the intersection of any two adjacent cubes: if $x_i = k_i\theta$ then $\mathbb{P}(k_{x_i} = k_i - 1) = \mathbb{P}(k_{x_i} = k_i + 1) = 0$ and so the expectation is taken over

the same set of lattice points, namely those forming the codimension 1 hypercube of the intersection.

Fourier analytic results

As mentioned in Section 5.2, we will use discrete Fourier analysis to help with the analysis of g. We discuss the necessary translations and dilations to relate a given $C(y)$ to the Boolean hypercube $\{\pm 1\}^d$, then collect several facts about the relationships between a function defined on $\{\pm 1\}^d$ and its multilinear extension defined on $[-1,1]^d$. These results will primarily be used in Section 5.4.

Single cube analysis

Let f be a function defined on $\theta\mathbb{Z}^d$ and g be its multilinear extension. Given $y \in \theta\mathbb{Z}^d$, let v_y be a vector such that $x \mapsto \frac{\theta}{2}x - v_y$ sends $C(y) \to \{-1,1\}^d$ and $y \mapsto -\mathbf{1}$. Define the following functions:

$$F(x) := f\left(\frac{\theta}{2}x\right), \text{ a dilation of } f,$$
$$F_y(x) := f\left(\frac{\theta}{2}x - v_y\right), \text{ a dilation and translation of } f.$$

Note that F is independent of y, and F_y is simply a shifted version of F. We will only ever consider $F_y(x)$ for $x \in \{-1,1\}^d$, so we restrict it to this domain. As introduced in Section 5.2 we can write F_y as

$$F_y(x) = \sum_{S \subseteq [d]} \widehat{F}_y(S)\chi_S(x).$$

Similarly, we define $G(x) := g\left(\frac{\theta}{2}x\right)$ and $G_y := g\left(\frac{\theta}{2}x - v_y\right)$. Since there is a unique multilinear function agreeing with F_y on $\{-1,1\}^d$, for $x \in [-1,1]^d$ we have the expansion

$$G_y(x) = \sum_{S \subseteq [d]} \widehat{F}_y(S)\chi_S(x).$$

It will be convenient to relate $\mathbb{E}\left[G_y(x)^2\right]$ (where the expectation is taken over x drawn uniformly at random from $[-1,1]^d$) and $\langle G, -\Delta G\rangle$ to the Fourier coefficients of F, denoted $\widehat{F}_y(S)$. This simply amounts to integrating the characters with respect to the Lebesgue

measure and differentiating. For any χ_S, we have

$$\begin{aligned}\int_{[-1,1]^d} (\chi_S(x))^2 dx &= \int_{[-1,1]^d} \prod_{i\in S} x_i^2 dx \\ &= 2^{d-|S|} \prod_{i\in S} \int_{-1}^{1} x_i^2 dx_i \\ &= 2^{d-|S|} \left(\frac{2}{3}\right)^{|S|} \\ &= 2^d \left(\frac{1}{3}\right)^{|S|}.\end{aligned}$$

We then have for every $y \in \theta\mathbb{Z}^d$:

$$\mathbb{E}\left[G_y(x)^2\right] = \sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F_y}(S)^2, \tag{5.18}$$

$$\mathbb{E}\left[\left(\frac{\partial}{\partial x_i} G_y(x)\right)^2\right] = \sum_{S:i\in S} \left(\frac{1}{3}\right)^{|S|-1} \widehat{F_y}(S)^2, \tag{5.19}$$

$$\langle G_y, -\Delta G_y\rangle = 2^d \sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|-1} |S|\, \widehat{F_y}(S)^2. \tag{5.20}$$

Full domain analysis

We want to find expressions for $\|f\|^2, \|g\|^2, \langle f, Lf\rangle$, and $\langle g, -\Delta g\rangle$ with respect to $\widehat{F}_y$ for $y \in \theta\mathbb{Z}^d$. Since we won't be looking at a single cube in these settings, we define

$$m_k = \sum_{S:|S|=k} \sum_{y\in\mathbb{Z}^d} \widehat{F}_y(S)^2.$$

Proposition 5.4.1. *For a function f defined on $\theta\mathbb{Z}^d$ and its multilinear extension g, we have the following expansions with respect to the Fourier coefficients*

$$\|f\|^2 = \sum_{k=0}^{d} m_k, \quad \& \quad \langle f, Lf\rangle = 4\sum_{k=0}^{d} k m_k, \tag{5.21}$$

$$\|g\|^2 = \theta^d \sum_{k=0}^{d} \left(\frac{1}{3}\right)^k m_k, \quad \& \quad \langle g, -\Delta g\rangle = 4\theta^{d-2} \sum_{k=0}^{d} \left(\frac{1}{3}\right)^{k-1} k m_k. \tag{5.22}$$

Proof. We prove the results for the norm and Laplacians side by side, as the steps are identical. Recall from (5.8) that for every $y \in \theta\mathbb{Z}^d$,

$$\mathbb{E}\left[F_y(x)^2\right] = \sum_{S\subseteq[d]} \widehat{F}_y(S)^2 \quad \& \quad \langle F_y, LF_y\rangle = 2^{d+1} \sum_{S\subseteq[d]} |S|\, \widehat{F}_y(S)^2.$$

We relate the norm and Laplacian terms on a single cube to those on the whole space:

$$\|f\|^2 = \frac{1}{2^d}\sum_y \|F_y\|^2 \quad \& \quad \langle f, Lf\rangle = \frac{1}{2^{d-1}}\sum_y \langle F_y, LF_y\rangle .$$

Combining these we have

$$\|f\|^2 = \sum_{k=0}^{d} m_k \quad \& \quad \langle f, Lf\rangle = 4\sum_{k=0}^{d} k m_k,$$

as desired. Recall from (5.18) and (5.20) that

$$\mathbb{E}\left[G_y(x)^2\right] = \sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F}_y(S)^2 \quad \& \quad \mathbb{E}\left[\left(\frac{\partial}{\partial x_i}G(x)\right)^2\right] = \sum_{S:i\in S} \left(\frac{1}{3}\right)^{|S|-1} \widehat{F}(S)^2.$$

Taking into account the dilation and the change of variables, we see that

$$\|g\|^2 = \left(\frac{\theta}{2}\right)^d \sum_y \|G_y\|^2 \quad \& \quad \langle g, -\Delta g\rangle = \frac{2^2}{\theta^2}\left(\frac{\theta}{2}\right)^d \langle G, -\Delta G\rangle .$$

Combining these we have

$$\|g\|^2 = \theta^d \sum_{k=0}^{d} \left(\frac{1}{3}\right)^k m_k \quad \& \quad \langle g, -\Delta g\rangle = 4\theta^{d-2}\sum_{k=0}^{d} \left(\frac{1}{3}\right)^{k-1} k m_k,$$

as desired. □

Proof of Theorem 5.1.2

In this section we prove an upper bound on $\mu(B_\theta)$ with respect to $\mu(A_\theta)$ with few constraints on our potential V. We start with this result as the approach is more closely aligned with the approach of Section 5.3. The result we prove is the following, a slightly sharper version of Theorem 5.1.2.

Theorem 5.4.2. *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, not necessarily periodic with $\|V\|_\infty \leqslant 1$, and let $\theta \in (0,1)$. Suppose that there exists $a \in \mathbb{R}$ such that for all $y \in \theta\mathbb{Z}^d$,*

$$\frac{1}{\theta^d}\int_{\overline{C}(y)} V(x)^2 dx \leqslant a \cdot \frac{1}{2^d}\sum_{k\in C(y)} V(k)^2.$$

Then,

$$c_d\mu(A_\theta) \geq \mu(B_\theta)$$

where $c_d = \max\left\{\frac{3^d}{2^d} + 5\cdot 3^d, \frac{3^d}{2^d}\frac{a}{r}\right\}$ and r is a d-dependent constant between $\frac{1}{4}$ and 1.

Throughout this section, we let f be a positive function on $\theta\mathbb{Z}^d$ and let g be the multilinear extension of f. Recall we can think of g as an expectation, as outlined in (5.17). We want to show

$$\frac{\langle g, B_\theta g\rangle}{\|g\|^2} = \frac{-\theta^2 \langle g, Lg\rangle + \langle g, Vg\rangle}{\|g\|^2} \leq c_d \frac{\langle f, A_\theta f\rangle}{\|f\|^2}$$

where c_d is as defined above. We will try to directly relate each term appearing on the left hand side to the corresponding term on the right hand side. As alluded to in Section 5.4, this will not be possible for the potential term. In this case, we will relate the potential term for continuous function to both the potential term and Laplacian term for the discrete function (see Lemma 5.4.5 for details). We give bounds for each term in the following lemmas, then combine them for the eventual proof of Theorem 5.4.2.

We start by relating the norms of f and g. The following bound achieves a better constant than Theorem 5.4.1.

Lemma 5.4.3. *The norms of f and g satisfy $\|g\|^2 \geq \left(\frac{2}{3}\right)^d \theta^d \|f\|^2$.*

Proof. We prove this via induction on d. For $d = 1$ we have

$$\begin{aligned}
\int_{\mathbb{R}} g(x)^2 dx &= \sum_{k\in\mathbb{Z}} \int_{k\theta}^{(k+1)\theta} \frac{1}{\theta^2} \left((x-k\theta)f(k\theta) + ((k+1)\theta - x)f((k+1)\theta)\right)^2 dx \\
&\geq \sum_{k\in\mathbb{Z}} \frac{1}{\theta^2} \int_{k\theta}^{(k+1)\theta} (x-k\theta)^2 f(k\theta)^2 + ((k+1)\theta - x)^2 f((k+1)\theta)^2 dx \\
&= \frac{\theta}{3} \sum_{k\in\mathbb{Z}} (f(k\theta)^2 + f((k+1)\theta)^2) \\
&= \frac{2\theta}{3} \|f\|^2 .
\end{aligned}$$

Assume that for any function f defined on $\theta\mathbb{Z}^d$, its multilinear extension g satisfies $\|g\|^2 \geq \left(\frac{2}{3}\right)^d \theta^d \|f\|^2$.

Consider f defined on $\theta\mathbb{Z}^{d+1}$ and its multilinear extension g. For any $x_1 \in \mathbb{R}$ with $k\theta \leqslant x_1 \leqslant (k+1)\theta$ and for any $y \in \theta\mathbb{Z}^d$, define

$$f_{x_1}(y) := \frac{x_1 - k\theta}{\theta} f(k\theta, y) + \frac{(k+1)\theta - x_1}{\theta} f((k+1)\theta, y).$$

Let g_{x_1} be the multilinear extension of f_{x_1} and observe that by definition $g_{x_1}(y) = g(x_1, y)$.

Now we evaluate the norm. Using the inductive hypothesis we have

$$\begin{aligned}\int_{\mathbb{R}^{d+1}} g(x)^2 dx &= \int_{\mathbb{R}}\int_{\mathbb{R}^d} g(x_1,y)^2 dy dx_1 = \int_{\mathbb{R}}\int_{\mathbb{R}^d} g_{x_1}(y)^2 dy dx_1 \\ &\geq \int_{\mathbb{R}} \left(\frac{2\theta}{3}\right)^d \|f_{x_1}\|^2 dx_1 \\ &= \left(\frac{2\theta}{3}\right)^d \int_{\mathbb{R}} \sum_{y\in\theta\mathbb{Z}^d} f_{x_1}(y)^2 dx_1\end{aligned} \tag{5.23}$$

Now we apply Tonelli's theorem to interchange the integral and the sum:

$$\begin{aligned}\int_{\mathbb{R}} \sum_{y\in\theta\mathbb{Z}^d} f_{x_1}(y)^2 dx_1 &= \sum_{y\in\theta\mathbb{Z}^d}\sum_{k\in\mathbb{Z}} \int_{k\theta}^{(k+1)\theta} \frac{1}{\theta^2}\left((x_1-k\theta)f(k\theta,y) + ((k+1)\theta - x_1)f((k+1)\theta,y)\right)^2 dx_1 \\ &\geq \sum_{y\in\theta\mathbb{Z}^d}\sum_{k\in\mathbb{Z}} \frac{1}{\theta^2}\int_{k\theta}^{(k+1)\theta} (x_1-k\theta)^2 f(k\theta,y)^2 + ((k+1)\theta-x_1)^2 f((k+1)\theta,y)^2 dx_1 \\ &= \frac{\theta}{3}\sum_{y\in\theta\mathbb{Z}^d}\sum_{k\in\mathbb{Z}} f(k\theta,y)^2 + f((k+1)\theta,y)^2 \\ &= \left(\frac{2\theta}{3}\right)\|f\|^2 .\end{aligned} \tag{5.24}$$

Combining (5.23) and (5.24), we have the desired result. □

Now we relate the Laplacian terms.

Lemma 5.4.4. *The quadratic forms of f and g with the respective Laplacian operators satisfy*

$$\langle g, -\Delta g\rangle \leqslant \theta^{d-2}\langle f, Lf\rangle .$$

Proof. From (5.21),

$$\begin{aligned}\langle f, Lf\rangle &= 4\sum_{k=0}^{d} k m_k \\ \langle g, -\Delta g\rangle &= 4\theta^{d-2}\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k-1} k m_k .\end{aligned}$$

Immediately we see

$$\langle g, -\Delta g\rangle \leqslant \theta^{d-2}\langle f, Lf\rangle .$$

□

Finally, we relate the potential term of g to the potential and Laplacian terms of f.

Lemma 5.4.5. *Let $V : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, not necessarily periodic with $\|V\|_\infty \leqslant 1$, and let $\theta \in (0,1)$. Suppose that there exists $a \in \mathbb{R}$ such that for all $y \in \theta\mathbb{Z}^d$,*

$$\frac{1}{\theta^d}\int_{\overline{C}(y)} V(x)^2 dx \leqslant a \cdot \frac{1}{2^d}\sum_{k\in C(y)} V(k)^2.$$

The quadratic forms of f and g with the respective potential operators satisfy

$$\langle g, Vg\rangle \leqslant \theta^d\left(10\cdot 2^{d-1}\langle f, Lf\rangle + \frac{a}{r}\langle f, Vf\rangle\right)$$

where r is a d-dependent constant between $\frac{1}{4}$ and 1.

Proof. We will consider cases based on the size of the local Laplacian Rayleigh quotients. Let $\alpha(d)$ be a function of d to be determined later. Let $y \in \theta\mathbb{Z}^d$ and let

$$\langle f, Lf\rangle_{C(y)} := \sum_{(u,v)\in E(C(y))} (f(u) - f(v))^2.$$

Case 1. $\dfrac{\langle f, Lf\rangle_{C(y)}}{\|f\|^2_{C(y)}} \geq \alpha(d)$.

We use the naive bound

$$\frac{\langle g, Vg\rangle_{\overline{C}(y)}}{\|g\|^2_{\overline{C}(y)}} \leqslant 1 \quad \Longrightarrow \quad \frac{\langle g, Vg\rangle_{\overline{C}(y)}}{\|g\|^2_{\overline{C}(y)}} \leqslant \frac{1}{\alpha(d)}\frac{\langle f, Lf\rangle_{C(y)}}{\|f\|^2_{C(y)}}.$$

Note that $\|g\|^2_{\overline{C}(y)} = \theta^d \sum_{S\subset[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F}_y(S)^2$ and $\|f\|^2_{C(y)} = 2^d \sum_{S\subset[d]} \widehat{F}_y(S)^2$, so we see

$$\frac{\|g\|^2_{\overline{C}(y)}}{\|f\|^2_{C(y)}} \leqslant \frac{\theta^d}{2^d}.$$

Using this and rearranging gives

$$\langle g, Vg\rangle_{\overline{C}(y)} \leqslant \frac{\|g\|^2_{\overline{C}(y)}}{\|f\|^2_{C(y)}}\frac{1}{\alpha(d)}\langle f, Lf\rangle_{C(y)} \leqslant \frac{\theta^d}{2^d}\frac{1}{\alpha(d)}\langle f, Lf\rangle_{C(y)}. \tag{5.25}$$

Case 2. $\dfrac{\langle f, Lf\rangle_{C(y)}}{\|f\|^2_{C(y)}} \leqslant \alpha(d)$.

Assume $\|f\|^2_{C(y)} = 1$. Let $\mu := \frac{\langle f, \mathbf{1}\rangle}{2^d}$. By the Poincaré inequality,

$$\|f - \mu\mathbf{1}\|^2_{C(y)} \leqslant \langle f, Lf\rangle \leqslant \alpha(d).$$

Let $\|f - \mu\mathbf{1}\|^2_{C(y),\infty}$ be the infinity norm of $f - \mu\mathbf{1}$ on $C(y)$. Then

$$\|f - \mu\mathbf{1}\|^2_{C(y),\infty} \leqslant \|f - \mu\mathbf{1}\|^2_{C(y)} \leqslant \alpha(d).$$

By its definition, $\mu = \sqrt{\mathbb{E}\left[f^2\right] - \mathrm{Var}(f)} \geq \sqrt{\frac{1}{2^d} - \frac{\alpha(d)}{2^d}}$. We see that for all $x \in C(y)$,

$$\sqrt{\frac{1}{2^d}\left(1 - \alpha(d)\right)} - \sqrt{\alpha(d)} \leqslant f(x) \leqslant \sqrt{\frac{1}{2^d}} + \sqrt{\alpha(d)}.$$

Now we relate the minimum and maximum values of f on $C(y)$. Define r by

$$\sqrt{r} = \frac{\sqrt{1 - \alpha(d)} - \sqrt{2^d\alpha(d)}}{1 + \sqrt{2^d\alpha(d)}}.$$

Then we have $\min\limits_{k\in C(y)} f(k)^2 \geq r \max\limits_{k\in C(y)} f(k)^2$.

Combining the above,

$$\begin{aligned}
\langle f, Vf\rangle_{C(y)} &\geq \min_{k\in C(y)} f(k)^2 \sum_{k\in C(y)} V(k) \\
&\geq \frac{2^d}{\theta^d}\frac{r}{a} \max_{k\in C(y)} f(k)^2 \int_{\overline{C}(y)} V(x)dx \qquad (5.26) \\
&\geq \frac{2^d}{\theta^d}\frac{r}{a} \langle g, Vg\rangle_{\overline{C}(y)}.
\end{aligned}$$

From (5.25) and (5.26), for any $y \in \theta\mathbb{Z}^d$

$$\langle g, Vg\rangle_{\overline{C}(y)} \leqslant \frac{\theta^d}{2^d}\frac{a}{r} \langle f, Vf\rangle_{C(y)} + \frac{\theta^d}{2^d}\frac{1}{\alpha(d)} \langle f, Lf\rangle_{C(y)}.$$

Now we sum over all $y \in \theta\mathbb{Z}^d$. Observe that in summing $\langle f, Vf\rangle_{C(y)}$ over y, each vertex is counted 2^d times; similarly, in summing $\langle f, Lf\rangle_{C(y)}$, each edge is counted 2^{d-1} times. Thus,

$$\begin{aligned}
\langle g, Vg\rangle = \sum_{y\in\theta\mathbb{Z}^d} \langle g, Vg\rangle_{\overline{C}(y)} &\leqslant \frac{\theta^d}{2^d}\frac{a}{r} \sum_{y\in\theta\mathbb{Z}^d} \langle f, Vf\rangle_{C(y)} + \frac{\theta^d}{2^d}\frac{1}{\alpha(d)} \sum_{y\in\theta\mathbb{Z}^d} \langle f, Lf\rangle_{C(y)} \\
&= \frac{\theta^d}{2^d}\frac{a}{r} 2^d \langle f, Vf\rangle + \frac{\theta^d}{2^d}\frac{1}{\alpha(d)} 2^{d-1} \langle f, Lf\rangle \\
&= \theta^d \left(\frac{a}{r} \langle f, Vf\rangle + \frac{1}{2\alpha(d)} \langle f, Lf\rangle\right).
\end{aligned}$$

Letting $\alpha(d) = \frac{1}{10\cdot 2^d}$ we have $r \geq \frac{1}{4}$ and we achieve the desired result. □

Proof of Theorem 5.4.2. From Lemma 5.4.3, Lemma 5.4.4, and Lemma 5.4.5 we have

$$\begin{aligned}\frac{\langle g, B_\theta g\rangle}{\|g\|_2^2} &\leqslant \frac{3^d}{2^d\theta^d\|f\|^2}\left(\theta^d\left(1+10\cdot 2^{d-1}\right)\langle f, Lf\rangle + \frac{a}{r}\theta^d\langle f, Vf\rangle\right)\\ &= \frac{1}{\|f\|^2}\left(\left(\frac{3^d}{2^d}+5\cdot 3^d\right)\langle f, Lf\rangle + \frac{3^d}{2^d}\frac{a}{r}\langle f, Vf\rangle\right).\end{aligned}$$

This proves our claim with

$$c_d = \max\left\{\frac{3^d}{2^d}+5\cdot 3^d, \frac{3^d}{2^d}\frac{a}{r}\right\}.$$

□

Proof of Theorem 5.1.3

We now prove our alternative lower bound, Theorem 5.1.3. To prove this bound, we will again consider the same multilinear extension g of a given discrete function f. However, we will not consider each term $\|g\|, -\theta^2\langle g, \Delta g\rangle, \langle g, Vg\rangle$ separately as done in previous sections. Instead we will compare the Rayleigh quotients of both the Laplacian term and the potential terms. In particular, for the Rayleigh quotient of the potential term, we will do a local analysis of the Rayleigh quotient at each cube, then view the full Rayleigh quotient as a weighted average of the local components.

The precise result that we prove, an explicit form of Theorem 5.1.3, is:

Theorem 5.4.6. *Let $V : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be $\mathbb{Z}^d$-periodic, sufficiently smooth, and (t_0, K, P) coercive. Let L be the Lipschitz constant of V and let $\frac{D_v^k D_u V}{k!} \leqslant M_P$ for any unit vectors u, v, and $k \leqslant P-1$ where D_v is the directional derivative with respect to v. For any $q \in (0,1)$ and and θ small enough that $\theta^q \leqslant t_0$,*

$$\left(1+\frac{1}{1-\mu(A_\theta)}\left(\mu(A_\theta)+\frac{2L\theta^{1-q}}{1-L\theta^{1-q}}\right)\right)\mu(A_\theta) \geq \mu(B_\theta) - \frac{1}{1-\mu(A_\theta)}\left(2M_P K\sqrt{d}\theta^{1+q(P-1)/P}\right).$$

To prove Theorem 5.4.6, we will analyze the contributions from the Laplacian terms and the potential terms to the Rayleigh quotient separately.

Laplacian term analysis

Given the above proposition, we now prove the following.

Lemma 5.4.7. *Let f be an ε-pseudo-ground-state for A_θ and let g be its multilinear extension. Then*

$$\frac{\mathcal{R}_{-\theta^2\Delta}(g)}{\mathcal{R}_L(f)} \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}.$$

Proof. In the following, keep in mind that m_k is dependent on θ.

Assume without loss of generality that $\|f\|^2 = \sum_{k=0}^{d} m_k = 1$. From (5.21) we have

$$\mathcal{R}_{-\theta^2\Delta}(g) = -\theta^2 \frac{\langle g, \Delta g\rangle}{\|g\|^2} = \frac{4\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k-1} k m_k}{\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k} m_k}$$

$$4\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k-1} k m_k \leqslant 4\sum_{k=0}^{d} k m_k = \langle f, Lf\rangle$$

so that

$$\mathcal{R}_{-\theta^2\Delta}(g) \leqslant \frac{\langle f, Lf\rangle}{\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k} m_k} = \frac{\mathcal{R}_L(f)}{\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k} m_k}.$$

We also note that if $d = 1$, this becomes an equality.

Now we want to lower bound $\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k} m_k$. It will suffice to just lower bound m_0. Since $\sum_{k=0}^{d} m_k = 1$, we can view $\langle f, Lf\rangle$ as a weighted average of elements of $[d]$ (times 4). Since f is an ε-p.g.s., we know $\mathcal{R}_{A_\theta}(f) \leqslant \mu(A_\theta) + \varepsilon$. From this and (5.21) we must have

$$m_0 \geq 1 - (\mu(A_\theta) + \varepsilon) = 1 - \mu(A_\theta) - \varepsilon.$$

From this we see

$$\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k} m_k = m_0 + \frac{1}{3}m_1 + \cdots \geq 1 - \mu(A_\theta) - \varepsilon.$$

Altogether we have

$$\frac{\mathcal{R}_{-\theta^2\Delta}(g)}{\mathcal{R}_L(f)} \leqslant \frac{1}{\sum_{k=0}^{d}\left(\frac{1}{3}\right)^{k} m_k} \leqslant \frac{1}{1 - \mu(A_\theta) - \varepsilon}$$

as desired. □

Potential term analysis

Now all that remains is to analyze the potential terms. We will analyze the Rayleigh quotient of the potential terms restricted to each $C(y)$ (or $\overline{C}(y)$ for the continuous case). We define the following quantities:

$$\|g\|^2_{\overline{C}(y)} := \int_{\overline{C}(y)} g(x)^2 dx, \qquad \langle g, Vg\rangle_{\overline{C}(y)} := \int_{\overline{C}(y)} V(x) g(x)^2 dx,$$

$$\|f\|^2_{C(y)} := \sum_{x\in C(y)} f(x)^2, \qquad \langle f, Vf\rangle_{C(y)} := \sum_{x\in C(y)} V(x) f(x)^2.$$

If we do a "cube by cube" analysis, we then need some way to relate this to the whole Rayleigh quotient. Observe that we can relate the "local" Rayleigh quotients to the "global" Rayleigh quotients via

$$\begin{aligned}\frac{\langle g, Vg\rangle}{\|g\|^2} &= \sum_{y\in\theta\mathbb{Z}^d} \frac{\|g\|^2_{\overline{C}(y)}}{\|g\|^2} \frac{\langle g, Vg\rangle_{\overline{C}(y)}}{\|g\|^2_{\overline{C}(y)}} \\ \frac{\langle f, Vf\rangle}{\|f\|^2} &= \sum_{y\in\theta\mathbb{Z}^d} \frac{\|f\|^2_{C(y)}}{2^d\|f\|^2} \frac{\langle f, Vf\rangle_{C(y)}}{\|f\|^2_{C(y)}}\end{aligned} \tag{5.27}$$

where the 2^d factor corrects for overcounting, since each vertex x is contained in 2^d cubes. It is natural to define the following probability measures on $\theta\mathbb{Z}^d$:

$$\begin{aligned}\nu_g(y) &:= \frac{\|g\|^2_{\overline{C}(y)}}{\|g\|^2} \\ \nu_f(y) &:= \frac{\|f\|^2_{C(y)}}{2^d\|f\|^2} = \frac{\mathbb{E}_{k\in C(y)} f(k)^2}{\|f\|^2}.\end{aligned} \tag{5.28}$$

Lemma 5.4.8. *For f and g as above, for all $y \in \theta\mathbb{Z}^d$,*

$$\nu_g(y) \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}\nu_f(y).$$

Proof. Assume $\|f\|^2 = \sum_{k=0}^d m_k = 1$. Recall from (5.8) and (5.21) that

$$\begin{aligned}\|g\|^2 &= \theta^d \sum_{k=0}^d \left(\frac{1}{3}\right)^k m_k \\ \|g\|^2_{\overline{C}(y)} &= \theta^d \sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F}_y(S)^2.\end{aligned}$$

Then we have

$$\nu_g(y) = \frac{\sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F}_y(S)^2}{\sum_{k=0}^d \left(\frac{1}{3}\right)^k m_k}.$$

Recall from the proof of Lemma 5.4.7 we must have $m_0 \geq 1-\mu(A_\theta)-\varepsilon$. We see $\sum_{k=0}^d \left(\frac{1}{3}\right)^k m_k \geq m_0 \geq 1-\mu(A_\theta)-\varepsilon$, so

$$\nu_g(y) \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}\left(\sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F}_y(S)^2\right).$$

From (5.8), we can write $\nu_f(y)$ as

$$\nu_f(y) = \sum_{S \subseteq [d]} \widehat{F}_y(S)^2.$$

Clearly $\sum_{S\subseteq[d]} \left(\frac{1}{3}\right)^{|S|} \widehat{F}_y(S)^2 \leqslant \sum_{S\subseteq[d]} \widehat{F}_y(S)^2$, so the claim is shown. □

Now we arrive at the main estimate.

Lemma 5.4.9. *Let $V : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be smooth, $\mathbb{Z}^d$-periodic and (t_0, K, P)-coercive. Let L be the Lipschitz constant of V and let $\frac{D_v^k D_u V}{k!} \leqslant M_P$ for any unit vectors u, v, and $k \leqslant P-1$. For any $q \in (0,1)$ and θ small enough that $\theta^q \leqslant t_0$,*

$$|\mathcal{R}_V(g) - \mathcal{R}_V(f)| \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}\left(2M_P K\sqrt{d}\theta^{1+q(P-1)/P} + \frac{2L\theta^{1-q}}{1-L\theta^{1-q}}\mathcal{R}_V(f)\right).$$

Proof. Suppose $y \in \theta\mathbb{Z}^d$. Let $q \in (0,1)$ and θ small enough that $\theta^q \leqslant t_0$.
Case 1. $V(y) \geq \theta^q$. Then, for $x \in \overline{C}(y)$,

$$1 - L\theta^{1-q} \leqslant \frac{V(x)}{V(y)} \leqslant 1 + L\theta^{1-q},$$

so that we have

$$\begin{aligned}
\mathcal{R}_{V,\overline{C}(y)}(g) &\leqslant V(y)\left(1+L\theta^{1-q}\right),\\
\mathcal{R}_{V,C(y)}(f) &\geq V(y)\left(1-L\theta^{1-q}\right),\\
\frac{\mathcal{R}_{V,\overline{C}(y)}(g)}{\mathcal{R}_{V,C(y)}(f)} &\leqslant \frac{1+L\theta^{1-q}}{1-L\theta^{1-q}},\\
\left|\mathcal{R}_{V,\overline{C}(y)}(g) - \mathcal{R}_{V,C(y)}(f)\right| &\leqslant \frac{2L\theta^{1-q}}{1-L\theta^{1-q}}\mathcal{R}_{V,C(y)}(f).
\end{aligned}$$

Case 2. $V(y) \leqslant \theta^q$.

Consider the difference $\left|\mathcal{R}_{V,\overline{C}(y)}(g) - \mathcal{R}_{V,C(y)}(f)\right|$. Being the difference of two random variables supported in $\overline{C}(y)$, it is at most

$$\sup \lim_{x\in\overline{C}(y)} V(x) - \inf_{x\in\overline{C}(y)} V(x) \leqslant 2 \sup_{x\in\overline{C}(y)} |V(x) - V(y)|.$$

By the fundamental theorem of calculus,

$$\sup_{x\in\overline{C}(y)} |V(x)-V(y)| \leqslant \sqrt{d}\theta \cdot \sup_{\substack{z\in\overline{C}(y)\\ u\in\mathbb{S}^{d-1}}} D_u V(z).$$

Let z and u maximize this expression. By coercivity, m is a critical point of order $p \leqslant P$. We know $D^k_{z-m} D_u V(m) = 0$ for any $k < p-1$. By Taylor's remainder theorem we can bound

$$D_u V(z) \leqslant \sup_{\|\xi - m\|_\infty < \sqrt{d}\theta} D^{p-1}_{z-m} D_u V(\xi) \cdot \frac{\|z-m\|_2^{p-1}}{(p-1)!}.$$

By coercivity, we know that $\|z-m\|_2^{p-1} \leqslant K\theta^{q(p-1)/p}$. Using the uniform bound M_P we have

$$\max_{x \in \overline{C}(y)} |V(x) - V(y)| \leqslant M_P K\sqrt{d}\theta \cdot \theta^{q(p-1)/p} = M_P K\sqrt{d}\theta^{1+q(p-1)/p} \leqslant M_P K\sqrt{d}\theta^{1+q(P-1)/P}. \tag{5.29}$$

To relate the global Rayleigh quotients we recall (5.27). By Theorem 5.4.8, $\nu_g \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}\nu_f$, so by (5.29) we have in both cases:

$$|\mathcal{R}_V(g) - \mathcal{R}_V(f)| \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}\left(2M_P K\sqrt{d}\theta^{1+q(P-1)/P} + \frac{2L\theta^{1-q}}{1-L\theta^{1-q}}\mathcal{R}_V(f)\right). \tag{5.30}$$

□

Proof of Theorem 5.4.6

Proof of Theorem 5.4.6. We first verify that the g constructed in (5.17) is a valid test function for (5.5). Note that g is continuous and equal to a polynomial of degree at most d on each cube $\overline{C}(y)$. Since $f \in L^\infty(\mathbb{R}^d)$, g has all derivatives $\partial_i g$ uniformly bounded on $\mathbb{R}^d = \bigcup_{y \in \theta\mathbb{Z}^d} \operatorname{int}(\overline{C}(y))$. Thus, g is Lipschitz. By e.g. [Eva22, pp. 279], we see that $g \in W^{1,\infty}(\mathbb{R}^d)$. By Jensen's inequality we further have $\|g\|^2_{\overline{C}(y)} \leqslant O\left(\|f\|^2_{C(y)}\right)$ for every y, so $g \in H^1(\mathbb{R}^d)$, as desired.

We now aim to bound the difference between

$$\mathcal{R}_{B_\theta}(g) = \mathcal{R}_{-\theta^2\Delta}(g) + \mathcal{R}_V(g) \quad \text{and} \quad \mathcal{R}_{A_\theta}(f) = \mathcal{R}_L(f) + \mathcal{R}_V(f).$$

From Lemma 5.4.7, multiplying through and rearranging yields the following relation on the Laplacian terms:

$$|\mathcal{R}_{-\theta^2\Delta}(g) - \mathcal{R}_L(f)| \leqslant \frac{\mu(A_\theta)+\varepsilon}{1-\mu(A_\theta)-\varepsilon}\mathcal{R}_L(f).$$

Combining this with (5.30) we have

$$|\mathcal{R}_{B_\theta}(g) - \mathcal{R}_{A_\theta}(f)| \leqslant \frac{1}{1-\mu(A_\theta)-\varepsilon}\left(2M_P K\sqrt{d}\theta^{1+q(P-1)/P} + \left(\mu(A_\theta)+\varepsilon + \frac{2L\theta^{1-q}}{1-L\theta^{1-q}}\right)\mathcal{R}_{A_\theta}(f)\right). \tag{5.31}$$

Minimizing over all f we have the following bound on $\mu(B_\theta)$:

$$\mu(B_\theta) \leqslant \mu(A_\theta) + \frac{1}{1-\mu(A_\theta)} \left(2M_P K \sqrt{d} \theta^{1+q(P-1)/P} + \left(\mu(A_\theta) + \frac{2L\theta^{1-q}}{1-L\theta^{1-q}} \right) \mu(A_\theta) \right). \tag{5.32}$$

□

Remark 18. It is clear from the proof that Theorem 5.1.3 holds for any periodic $V \in C^P$, not necessarily C^∞.

Lower bounds on $\mu(B_\theta)$

Lastly, to show Theorem 5.1.3 is non-trivial, we prove asymptotic lower bounds on $\mu(B_\theta)$ for uniformly p-coercive potentials with isolated critical points, as advertised in Section 5.1.

Proof of Theorem 5.1.4. Assume $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ with $\|f\|^2 = 1$ satisfies

$$\theta^2 \|\nabla f\|^2 + \int_{\mathbb{R}^d} V(x) f(x)^2 dx \leqslant \theta^r,$$

for r a constant depending on p and d to be chosen later. Let

$$W := \{x : V(x) \leqslant 2\theta^r\}$$

and notice by Markov's inequality (applied to the probability distribution defined by the mass of f) that

$$\sum_C \|f|_C\|^2 \frac{\|f|_{\overline{W}\cap C}\|^2}{\|f|_C\|^2} = \|f|_{\overline{W}}\|^2 \leqslant \frac{1}{2},$$

where the sum is over unit volume cubes in $\mathbb{R}^d$. Let S be the set of cubes C satisfying

$$\frac{\|f|_{\overline{W}\cap C}\|^2}{\|f|_C\|^2} \leqslant \frac{3}{4},$$

and notice by Markov's inequality that

$$\sum_{C\in S} \|f|_C\|^2 \geq c$$

for a constant c. Let $q > 2$. For each cube $C \in S$ we have $\|f|_{W\cap C}\|^2 \geq \frac{\|f|_C\|^2}{4}$ as well as

$$\frac{\|f|_{W\cap C}\|_q^2}{\|f|_{W\cap C}\|^2} \geq \mathrm{vol}(W \cap C)^{\frac{2}{q}-1},$$

which together yield

$$\|f\|_q^2 \geq \sum_{C\in S} \|f|_{W\cap C}\|_q^2 \geq \frac{1}{4}\sum_{C\in S} \mathrm{vol}(W\cap C)^{\frac{2}{q}-1}\,\|f|_C\|^2\,.$$

Since f has isolated critical points and is uniformly p-coercive, we have the uniform estimate

$$\mathrm{vol}(W\cap C) = O\left(\theta^{\frac{rd}{p}}\right)$$

for each cube C. Substituting, we have

$$\|f\|_q^2 \geq \Omega(1)\cdot\theta^{\frac{rd}{p}\left(\frac{2}{q}-1\right)}\sum_{C\in S}\|f_C\|^2 = \Omega(1)\cdot\theta^{\frac{rd}{p}\left(\frac{2}{q}-1\right)}.$$

(b) If $d\geq 3$ set $q=\frac{2d}{d-2}$ and apply the Sobolev inequality to obtain

$$\|\nabla f\|^2 = \Omega(1)\cdot\theta^{-\frac{2r}{p}},$$

whence $\theta^2\|\nabla f\|^2 = \Omega\left(\theta^{2-\frac{2r}{p}}\right)$. Choosing $r=2-\frac{2r}{p}=\frac{2p}{p+2}$ yields the claim.

(a) If $d=1,2$ we instead use the Sobolev inequalities

$$\|\nabla f\|^2 + \|f\|^2 \geq C_q\,\|f\|_q^2\,,$$

for $q=\infty$ and q large, respectively. Since $\|f\|^2=1$ and the right hand side is unbounded as $\theta\to 0$, this term can be ignored and we have $\|f\|_q^2\geq\Omega(1)\cdot\theta^{-\frac{r}{p}}$ for $d=1$ and $\|f\|_q^2\geq\Omega(1)\cdot\theta^{-\frac{2r}{p}+o(1)}$ for $d=2$. Choosing $r=2-\frac{r}{p}$ and $r=2-\frac{2r}{p}+o(1)$ yields $r=\frac{2p}{p+1}$ and $r=\frac{2p}{p+2}+o(1)$, as advertised.

Now, we set out to prove the upper bound (c) through a different approach. By boundedness, there is a constant H such that $V(x)\leqslant H\|x\|^p =: W(x)$ where we assume that $V(0)=0$ and that 0 is a critical point of order p. For all test functions g, $\mathcal{R}_{B_\theta}(g)\leqslant\mathcal{R}_{-\theta^2\Delta+W}(g)$. Let

$$g_\sigma(x) = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right).$$

Then,

$$\begin{aligned}\langle g_\sigma, W g_\sigma\rangle &= H\int_{\mathbb{R}^d}\|x\|^p\,g_\sigma(x)dx\\ &= HS_{d-1}\int_{\mathbb{R}_{\geq 0}} r^{p+d-1}\exp\left(-\frac{r^2}{2\sigma^2}\right)dr\\ &= HS_{d-1}2^{\frac{p+d-2}{2}}\left(\frac{p+d-2}{2}\right)!\sigma^{p+d}\end{aligned}$$

where $S_{d-1} = \frac{2\sqrt{\pi}^d}{\left(\frac{d-2}{2}\right)!}$ is the surface area of $\mathbb{S}^{d-1}$. We also compute directly that

$$\left\langle g_\sigma, \left(-\theta^2\Delta + H\right) g_\sigma \right\rangle = \frac{1}{2}\left(d\frac{\theta^2}{\sigma^2} + \operatorname{tr} H\sigma^2\right)$$
$$\|g_\sigma\|^2 = \left(\sqrt{\pi}\sigma\right)^d$$

so that all told,

$$\left\langle g_\sigma, \left(-\theta^2\Delta + W\right) g_\sigma \right\rangle = \frac{1}{2}d\frac{\theta^2}{\sigma^2} + 2^{\frac{p+d}{2}} H \frac{\left(\frac{p+d-2}{2}\right)!}{\left(\frac{d-2}{2}\right)!}\sigma^p.$$

By weighted AM–GM this is minimized by

$$\frac{2d}{p}\sqrt[\frac{p}{2}+1]{\frac{p}{d}2^{\frac{d-2}{2}} H \frac{\left(\frac{p+d-2}{2}\right)!}{\left(\frac{d-2}{2}\right)!}}\theta^{\frac{2p}{p+2}} \tag{5.33}$$

where the exponent of θ matches the claim. □

Chapter 6

Curvature of basis exchange walks

The work in this section largely follows that of [Det25].

In this chapter the question we explore has a slightly different flavor, focusing on the discrete curvature of a specific class of Markov chains. Although we don't directly consider the spectrum of any operator, we still consider it a spectrally motivated question due to the connection between curvature and the spectral gap of a Markov chain.

In particular, we ask about the *Ollivier-Ricci curvature* of *basis exchange walks*, a specific class of Markov chains. We spend some time discussing the properties of basis exchange walks and the properties of discrete Markov chains with non-negative Ollivier-Ricci curvature to properly motivate the following question:

Do all basis exchange walks have non-negative Ollivier-Ricci curvature?

We find that although basis exchange walks exhibit many of the hallmark properties of Markov chains with non-negative Ollivier-Ricci curvature, there exist examples with negative curvature. This result is somewhat surprising, and implies that local notions of expansion are not equivalent (see the next section for details).

6.1 Introduction

Considering the long history of translating ideas from geometry to spectral graph theory, considerable effort has been made in the last two decades to find a discrete analogue of the notion of curvature. One particularly fruitful definition has been *Ollivier-Ricci curvature*. Ollivier-Ricci curvature on metric spaces was first introduced in [Oll09], as an analog to Ricci curvature on manifolds. Notably, this is a *local* notion, depending only on the two neighborhood of adjacent states in a Markov chain, but can have *global* implications (similar to curvatures in differential geometry settings). We will see several examples of such implications for a Markov chain with non-negative Ollivier-Ricci curvature in Section 6.3.

Here we consider the Ollivier-Ricci curvature of basis exchange walks, and show that despite the many desirable properties uniformly exhibited by basis exchange walks, there

exist both negatively and non-negatively curved basis exchange walks. Furthermore, we show even among standard classes of matroids, such as linear matroids or graphic matroids, there exist both matroids with negatively and non-negatively curved basis exchange walks. This work reinforces the idea that non-negative curvature is a somewhat delicate phenomenon, which cannot be guaranteed by a strongly log-concave stationary distribution (a feature of all basis exchange walks) or the representability of the matroid.

We highlight in particular the implication this has for the connection between spectral independence and Ollivier-Ricci curvature, both local certificates of expansion. Concurrent works [BCC$^+$22] and [Liu21] prove non-negative Ollivier-Ricci curvature *implies* spectral independence (see either work for precise, quantified statements). The validity of the converse to this statement was not previously known, but this work disproves it: basis exchange walks are uniformly 0-spectrally independent, but we establish several examples with negative Ollivier-Ricci curvature.

6.2 Markov chains and basis exchange walks

We will think of basis exchange walks as a particularly type of "well-behaved" Markov chain. Before formally defining a basis exchange walk, we pause to discuss several different desirable properties of Markov chains. We will see that basis exchange walks have all of these properties.

Desirable properties of Markov chains

We define several desirable properties of Markov chains, noting that all of these ultimately serve the same purpose: to bound the mixing time of the Markov chain. For each notion, we show the connection to mixing time.

Let P be the transition matrix of a reversible, irreducible Markov chain over a finite state space χ. Let π denote the stationary distribution of P. We first recall the definition of the mixing time:

Definition 6.2.1 (Mixing time)**.** The mixing time of a Markov chain is

$$t_{min}(\epsilon) = \inf\left\{t : \left\|P^t(x,\cdot) - \pi\right\|_{TV} \leq \epsilon \ \forall x \in \chi\right\}$$

where $\|\mu - \nu\|_{TV} = \sum_{x\in\chi} |\mu(x) - \nu(x)|$, the total variation distance.

Spectral gap. The first property we look at is a spectral gap.

Definition 6.2.2 (Spectral gap)**.** Let $\lambda_* = \max\{|\lambda| : \lambda \in \sigma(P)\}$. The *absolute spectral gap* of P is defined as $\gamma_* = 1 - \lambda_*$.

Note for a lazy Markov chain the absolute spectral gap coincides with the spectral gap, that is $\gamma_* = \gamma = 1 - \lambda_2$.

It will be convenient to consider the inverse spectral gap, which we refer to as the relaxation time and denote $t_{rel} = \frac{1}{\gamma_*}$. The subsequent fact shows that a large spectral gap implies that the Markov chain mixes quickly (which, considering a spectral decomposition of P, should not be surprising).

Theorem 6.2.1 (Mixing time bound from spectral gap, [LP17]). *Let $\pi_{min} := \min_{x\in\chi} \pi(x)$. Then letting , we have*

$$t_{mix}(\epsilon) \leq t_{rel} \log\left(\frac{1}{\epsilon \pi_{min}}\right).$$

Modified log-Sobolev constant. Next we define another quantity we can associate to a Markov chain: the modified log-Sobolev constant. To define this, we first need to define the Dirichlet form of a reversible Markov chain and the normalized relative entropy.

Definition 6.2.3 (Dirichlet form). Given $f, g : \chi \to \mathbb{R}$, we define

$$\mathcal{E}_P(f, g) := \sum_{x,y\in\chi} \pi(x) f(x)[I - P](x, y)g(y)$$

where I denotes the identity matrix.

Definition 6.2.4 (Normalized relative entropy). Given $f : \chi \to \mathbb{R}_{\geq 0}$ we define

$$\mathrm{Ent}_\pi(f) := \mathbb{E}_\pi(f \log f) - \mathbb{E}_\pi f \log \mathbb{E}_\pi f$$

where we adopt the convention that $0 \log 0 = 0$.

Definition 6.2.5 (Modified log-Sobolev constant). The modified log-Sobolev constant of a Markov chain P is defined as

$$\rho := \inf\left\{ \frac{\mathcal{E}_P(f, \log f)}{\mathrm{Ent}_\pi(f)} \mid f : \chi \to \mathbb{R}_{\geq 0}, \mathrm{Ent}_\pi(f) \neq 0 \right\}.$$

A bound on the modified log-Sobolev constant allows us control over the decay of entropy as the Markov chain runs. More formally, we can bound the KL-divergence of $P^t(x, \cdot)$ and π in terms of ρ, which by Pinsker's inequality will give us control over the TV distance between $P^t(x, \cdot)$ and π. This, in turn, gives us a bound on the mixing time, which we state below.

Lemma 6.2.2 (from [BT06]). *Letting $\pi_{min} = \min_{x\in\chi} \pi(x)$, we have*

$$t_{mix}(\epsilon) \leq \frac{1}{\rho}\left(\log\log\frac{1}{\pi_{min}} + \log\frac{1}{2\epsilon^2}\right).$$

Spectral independence. Lastly we look at a property specific to certain Markov chains over product spaces known as *spectral independence*, first introduced in [ALG21]. Spectral independence can also be considered a local notion of expansion; to understand spectral independence in this way, one must go back to the definition of spectral independence and its motivation from high-dimensional expanders. The curious reader should see [Liu23] for a comprehensive exploration of spectral independence.

Spectral independence itself is a property of a *measure* π. We will first define spectral independence, then discuss a Markov chain associated with π. Finally we will see that spectral independence of π implies a spectral gap (and therefore a mixing time bound) for the associated Markov chain.

We start with a measure over subsets of a uniform size $\pi : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$. Spectral independence is a condition on the influence matrix defined by π, as seen below.

Definition 6.2.6 (Influence matrix)**.** The influence matrix $I_\pi \in \mathbb{R}^{n \times n}$ is defined by

$$I_\pi(i,j) := \mathbb{P}_{S \sim \pi}[j \in S \mid i \in S] - \mathbb{P}_{S \sim \pi}[j \in S].$$

Definition 6.2.7 (Spectral independence)**.** For $\eta \geq 0$, we say π is η-spectrally independent if $\lambda_{max}(I_\pi) \leq 1 + \eta$.

The associated Markov chain we will consider is the *down-up walk* defined by this chain, which we denote P_π. Informally, starting at a set S, one step of the down-up walk drops an element of S uniformly at random (the down step), then adds an element to form a new set T with probability proportional to $\pi(T)$ (the up step). See Definition 6.2.11 for a formal definition in the context of matroids.

The following connects spectral independence of μ and the spectral gap of P_μ. Note we can then translate this to a mixing time bound.

Lemma 6.2.3 (from [ALG21])**.** *Suppose there exists $\eta \geq 0$ such that π and all of its conditional measures are all η-spectrally independent. Then $\gamma(P_\pi) \geq \Omega(k^{-1+\eta})$.*

Basis exchange walks

We now define basis exchange walks, starting with the definition of a matroid.

Definition 6.2.8 (Matroid)**.** A *matroid* $\mathcal{M} = (E, \mathcal{B})$ is a set E and a non-empty collection of its subsets $\mathcal{B}$, called *bases*, satisfying the following:

- no proper subset of a basis is itself a basis, and
- for any $B_1, B_2 \in \mathcal{B}$, if $b_1 \in B_1 \backslash B_2$, there exists a $b_2 \in B_2 \backslash B_1$ such that $B_1 - b_1 + b_2 \in \mathcal{B}$.

The second condition is known as the *basis exchange property*.

Definition 6.2.9 (Rank)**.** The *rank* of a matroid $\mathcal{M} = (E, \mathcal{B})$ is the size of any basis $B \in \mathcal{B}$. Note that by the basis exchange property, all bases must be the same size.

A particularly interesting class of matroids are those induced by a graph G, known as graphic matroids. We consider several graphic matroids as examples, so we define them here.

Definition 6.2.10 (Graphic matroids)**.** A *graphic matroid* induced by a graph G is a matroid whose basis sets are the spanning forests of G. If G is connected, the basis sets are the spanning trees of G.

Definition 6.2.11 (Down-up walk)**.** The *down-up walk* or the *basis exchange walk* is a Markov chain on the bases of a matroid. Starting from a basis $S \in \mathcal{B}$, the random walk proceeds by dropping an element $u \in S$ uniformly at random, then moving to a basis containing $S - u$ uniformly at random.

We note that the above describes the basis exchange walk on a set of bases given by the uniform distribution (i.e., the uniform distribution over all bases will be the stationary distribution). Given a distribution π over $\mathcal{B}$, one can define the corresponding down-up walk, but in considering the question of curvature of basis exchange walks, we will only be concerned with the uniform case.

With the definitions of basis exchange walks in hand, we point out that basis exchange walks have all of the "desirable" properties of Markov chains mentioned above. One key tool used to prove these properties is the *generating polynomial* of the stationary distribution associated to a basis exchange walk. In particular, this generating polynomial is *strongly log-concave* (see [ALGV19] or [ALGVV21] for further details).

We have the following properties of basis exchange walks:

Theorem 6.2.4 (from [ALGV19])**.** *Let π be the uniform distribution over the bases of a matroid. Then π is 0-spectrally independent.*

Theorem 6.2.5 (from [ALGV19])**.** *The spectral gap γ of a basis exchange walk over a rank-k matroid satisfies $\gamma \geq 1/k$.*

Theorem 6.2.6 (from [CGM19])**.** *The modified log-Sobolev constant ρ of a basis exchange walk over a rank-k matroid satisfies $\rho \geq 1/k$.*

We highlight that the state spaces for these Markov chains could be exponential in the size of the ground set n, but the bounds for the spectral gap and modified log-Sobolev constant depend only on k, the rank of the matroid.

6.3 Ollivier-Ricci curvature

We now turn to a particular notion of discrete curvature, Ollivier-Ricci curvature. To define Ollivier-Ricci curvature, we first recall the definition of Wasserstein distance:

Definition 6.3.1 (Wasserstein distance)**.** Let μ, ν be two probability distributions over some metric space (X, d). The *Wasserstein distance* between μ and ν is defined as

$$\mathcal{W}(\mu, \nu) = \min_{X \sim \mu, Y \sim \nu} \mathbb{E} d(X, Y),$$

where the minimum is over all possible couplings of μ and ν.

Now we are ready to define Ollivier-Ricci curvature:

Definition 6.3.2 (Ollivier–Ricci curvature)**.** For a Markov chain P over a metric space (X, d), we define the *Ollivier–Ricci curvature* of P as the minimum κ such that for all $S, T \in X$

$$\mathcal{W}(P(S, \cdot), P(T, \cdot)) \leq (1 - \kappa) \cdot d(S, T).$$

Note if d is the metric induced by P, i.e., $d(S, T) := \inf\{t : P^t(S, T) > 0\}$, we can consider the minimum over only *adjacent* $S, T \in X$ in the above definition (see [Oll09] for further details). Throughout, we will take d to be the metric induced by P.

Properties of non-negatively curved Markov chains

We have the following results about Markov chains with non-negative curvature Ollivier-Ricci curvature: we see that non-negative curvature allows us to bound the spectral gap, gives us control over the entropy, and implies spectral independence.

Lemma 6.3.1 (from [Oll09])**.** *Suppose a Markov chain P has Ollivier-Ricci curvature $\kappa \geq 0$. Then $\gamma(P) \geq \kappa$.*

Non-negative Ollivier-Ricci curvature can imply a modified log-Sobolev constant, but additional conditions must also be satisfied (see [Mün23] for details). We can, however, get control on other aspects of the entropy if we have a non-negatively curved Markov chain. In particular, we can bound the the variance of the KL divergence (also known as the *varentropy*) of $P^t(x, \cdot)$ with respect to the stationary distribution. We refrain from stating any formal results here, but direct the reader to [Sal23] for more details.

Lemma 6.3.2 ([Liu21], informally)**.** *Non-negative Ollivier-Ricci curvature implies spectral independence.*

Note that both curvature and spectral independence are quantitative properties, and a Markov chain with stronger positive curvature will have stronger spectral independence. See [Liu21] for full details.

Curvature and degree

We briefly comment that non-negative curvature is only a useful notion for a family of Markov chains with high degree (where the degree is the number of neighboring states a given state can move to). We have the following fact about bounded degree expanders:

Lemma 6.3.3 (from [Sal21])**.** *Bounded degree expanders always have negative curvature.*

Since bounded degree expanders always have negative curvature, we see that if we have an infinite family of constant degree Markov chains, they cannot have non-negative curvature bounded below by a constant. If such a family did exist, then by Lemma 6.3.1, they should be expanders. But this is a contradiction with the above fact.

From this, we see that curvature is only a useful notion for families of Markov chains with superconstant degree (which basis exchange walks are).

6.4 A lower bound and positively curved examples

We now specifically consider the Ollivier-Ricci curvature of basis exchange walks.

From Definition 6.3.1 and Definition 6.3.2, we see that to show a Markov chain is positively curved, it suffices to find a coupling for each edge (S, T) so that in one step, a random walk started from S and a random walk started from T move closer in expectation. In the following section we define such a coupling. Intuitively, this coupling tries to keep the walks close by matching their down-steps and then matching their up-steps whenever possible.

Using this coupling, we define give a lower bound on the Ollivier-Ricci curvature of a basis exchange walk in terms of the rank of the matroid and size of the ground set. We also give a more precise lower bound in terms of more local measures. We then use this to give several examples of positively curved basis exchange walks.

The down-step coupling

We define a candidate coupling, which we call the *down-step coupling*, for the down-up walk starting from any adjacent bases S, T. The following will be useful to denote all possible "up-steps" of a given down-step:

Definition 6.4.1. If B is a basis and $u \in B$, let $N(B-u) = \{x \in E : B-u+x \text{ is a basis}\}$.

Definition 6.4.2 (Down-step coupling)**.** Let $S = \{s, u_1, \ldots, u_{k-1}\}$, $T = \{t, u_1, \ldots, u_{k-1}\}$ be bases of a rank-k matroid. Let P be the basis exchange walk, $X \sim P(S, \cdot)$, and $Y \sim P(T, \cdot)$. We couple X and Y by first matching the down-steps when possible, then matching the up-steps when possible. More concretely:

- If S drops s, then T drops t. Since $S - s = T - t$, we can always couple the up steps to be the same.

- If S drops u_i, then T drops u_i. Without loss of generality, suppose

$$\#N(S-u_i) \geq \#N(T-u_i).$$

  We couple the up steps in the following way:

  - If $S-u_i$ adds t, $T-u_i$ adds s (note if $t \in N(S-u_i)$, then $s \in N(T-u_i)$ since $S-u_i+t = T-u_i+s$).
  - If $S-u_i$ adds $v \in N(S-u_i) \cap N(T-u_i)$, $T-u_i$ adds v.
  - If $S-u_i$ adds $v \notin N(S-u_i) \cap N(T-u_i)$ with $v \neq t$, then $T-u_i$ adds a random element from $N(T-u_i)$ with any distribution preserving the desired marginal across $N(T-u_i)$.

The distribution mentioned in the last line is simply a distribution such that the up-step for $T-u_i$ is uniform across $N(T-u_i)$; such a distribution will always be possible given that $\#N(S-u_i) \geq \#N(T-u_i)$. Note that with this coupling, the walks are at distance at most 2 by the basis exchange property.

Remark 19. We remark that while this coupling might not be optimal, coupling different down steps for S and T can only decrease $\mathbb{E}d(X,Y)$ in a small number of cases. For example, consider $S = \{s,u,u'\}$ and $T = \{t,u,u'\}$. Suppose S drops u and adds a, while T drops u' and adds b, so S moves to $\{s,u,a\}$ and T moves to $\{t,b,u'\}$. We are only in a better position (i.e., at a distance less than 2) if: (1) $a = t$ and $b = s$, in which case the distance is 1, or (2) $a = u'$ and $b = u$, in which case the distance is also 1. We see that for most possible up-steps, it is more advantageous to couple based on the down-step. We make this more precise with an upper bound on curvature in Section 6.5.

Example. We give an example of this coupling on the graphic matroid induced by K_4. Let S and T be the following spanning trees, respectively:

The following table shows how walks would move from S and T in the down-step coupling, where s is the bottom edge and t is the diagonal edge from the upper left to bottom right. For clarity, we group outcomes under which element is dropped in the down step of the down-up walk.

Lower bound

Our first theorem provides a lower bound on the curvature of the basis exchange walk in terms of the rank and the size of the ground set.

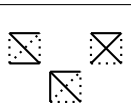
initial state	drop s/drop t			drop top edge			drop left edge			
										
										
$\mathbb{P}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{36}$ each pair
	add same edge			add same edge			add t /add s	add left edge	add diag.	

Table 6.1: The down-step coupling for two states in the graphic matroid induced by K_4.

Theorem 6.4.1. *The basis exchange walk of a rank-k matroid over a set of size $n > k+1$ satisfies*

$$\kappa \geq -1 + \frac{2}{k} + \frac{3(k-1)}{k(n-k+1)}.$$

If $n = k+1$, the basis exchange walk satisfies $\kappa \geq 1/k$.

Proof. We show that the above coupling gives us the desired bound.

First note if S drops s and T drops t (which occurs with probability $\frac{1}{k}$) we have

$$\mathbb{E}\left(d(X,Y) \mid s,t \text{ dropped in down step}\right) = 0.$$

Now we consider what happens when we drop u_i. Letting $U' = \{u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_{k-1}\}$, we have two cases:

Case 1: $U' + s + t \notin \mathcal{B}$.

We claim $N(S-u) = N(T-u)$ by the basis exchange property. Let $x \in N(S-u)$ and $y \in N(T-u)$, so

$$S - u + x = U' + s + x \in \mathcal{B}, \qquad T - u + y = U' + t + y \in \mathcal{B}.$$

Applying the basis exchange property to $U'+s+x, U'+t+y$, and x, we must have $U'+s+y \in \mathcal{B}$ since $U'+s+t \notin \mathcal{B}$, so $y \in N(S-u)$. Similarly, we can see that $U'+t+x \in \mathcal{B}$, so $x \in N(T-u)$.

Since $N(S-u) = N(T-u)$, we can couple these walks to have the same up-step and in this case

$$\mathbb{E}\left(d(X,Y) \mid u_i \text{ dropped}\right) = 1.$$

Case 2: $U' + s + t \in \mathcal{B}$

In this case, we can couple the walks to both move to $U'+s+t$ with some probability and the distance will be 0. Note that $N(S-u), N(T-u) \leq n-(k-1)$ since $(S-u) \cap N(S-u) = \emptyset$ (and similarly for T), so

$$\min\{\mathbb{P}(U' + s \to U' + s + t), \mathbb{P}(U' + t \to U' + s + t)\} \geq \frac{1}{n-k+1}$$

where $\mathbb{P}(U' + s \to U' + s + t)$ denotes the probability of adding t on the up-step, given u_i was dropped on the down-step (i.e., $1/N(S-u_i)$). Similarly, we can couple the walks to stay at S and T respectively, which also occurs with probability $\geq \frac{1}{n-k+1}$, and the distance is 1. Otherwise, the distance is always less than or equal to 2 since

$$U' + s + x \sim U' + s + t \sim U' + t + y.$$

All together then we have

$$\mathbb{E}\left(d(X,Y) \mid u_i \text{ dropped}\right) \leq 2 \cdot \left(1 - \frac{2}{n-k+1}\right) + \frac{1}{n-k+1} = 2 - \frac{3}{n-k+1}.$$

If $n > k+1$, in either case we have

$$\mathbb{E}\left(d(X,Y) \mid u_i \text{ dropped}\right) \leq 2 - \frac{3}{n-k+1}.$$

Then we have

$$\mathcal{W}(P(S,\cdot), P(T,\cdot)) \leq \frac{(k-1)}{k} \cdot \left(2 - \frac{3}{n-k+1}\right).$$

By definition of κ, we have:

$$\kappa \geq 1 - \frac{(k-1)}{k} \cdot \left(2 - \frac{3}{n-k+1}\right) \tag{6.1}$$

$$= 1 - \frac{2(k-1)}{k} + \frac{3(k-1)}{k(n-k+1)} \tag{6.2}$$

$$= -1 + \frac{2}{k} + \frac{3(k-1)}{k(n-k+1)}. \tag{6.3}$$

If $n = k+1$, then in either case $\mathbb{E}(d(X,Y) \mid u_i \text{ dropped}) \leq 1$, so we have $\kappa \geq 1/k$. □

By more carefully tracking which case we are in, we can sharpen this bound. We define the set of elements we can drop that would place us in case two from above:

Definition 6.4.3. Let $J = \{u : t \in N(S-u) \text{ and } u \neq s\}$.

Corollary 6.4.2. *The basis exchange walk over any matroid satisfies*

$$\kappa \geq \min_{S \sim T} \left\{ -\frac{\#J}{k} + \frac{1}{k} + \frac{1}{k} \sum_{u \in J} \frac{2 + \#(N(S-u) \cap N(T-u))}{\max\{\#N(S-u), \#N(T-u)\}} \right\}.$$

Proof. This follows immediately from the down-step coupling and definitions. □

While this expression is not particularly friendly, it makes explicit the idea that if there is a high overlap between $N(S-u)$ and $N(T-u)$ for all u, the basis exchange walk will be positively curved. In the following examples, we directly analyze the coupling, which is equivalent to applying this corollary (although we refrain from referring to this expression).

Positively curved examples

Non-negative curvature from smallness

We have the following corollaries to Theorem 6.4.1 for rank-2 matroids, matroids over small ground sets, and graphic matroids over small graphs:

Corollary 6.4.3. *The basis exchange walk on any rank-2 matroid is positively curved.*

Corollary 6.4.4. *The basis exchange walk on any matroid with $n \leq 7$ is non-negatively curved.*

Corollary 6.4.5. *The basis exchange walk on any graphic matroid induced by a simple graph $G = (V, E)$ with $\#V \leq 4$ is non-negatively curved.*

Uniform matroid

Lemma 6.4.6. *The basis exchange walk over the rank-k uniform matroid with size n ground set satisfies*

$$\kappa \geq 1 - \frac{(k-1)(n-k)}{k(n-k+1)}.$$

Proof. The down-step coupling gives

$$\mathbb{E}d(X,Y) = \frac{1}{k} \cdot 0 + \frac{k-1}{k} \left(\frac{1}{n-k+1} \cdot 0 + \frac{n-k}{n-k+1} \cdot 1 \right).$$

□

Vámos matroid

The *Vámos matroid* is a rank-4 matroid over a ground set of size 8 that cannot be represented as a matrix over any field. However, this matroid is close, in some sense, to a uniform matroid; it includes 65 of the possible 70 subsets of size 4 as bases. The excluded sets are the shaded parallelograms in Figure 6.1[1].

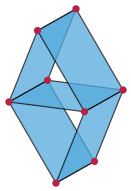


Figure 6.1: Dependent sets in the Vámos matroid.

The following lemma shows that the Vámos matroid is positively curved. We can interpret this as a consequence of how close the Vámos matroid is to the rank-4 uniform matroid over 8 elements.

Lemma 6.4.7. *The basis exchange walk over the Vámos matroid is positively curved.*

Proof. Again we show the down-step coupling demonstrates positive curvature.

Let $S = \{a, b, c, s\}$ and $T = \{a, b, c, t\}$ for some distinct elements a, b, c, s, t. For $u \neq s, t$ we consider the possibilities for $N(S-u)$ and $N(T-u)$. By definition of the Vámos matroid, for any $S \in \mathcal{B}$ and any $u \in S$, $\#N(S-u) = 5$ or 4 (i.e., it either contains all the elements not in $S-u$ or one less). We compute the expected distance in each case.

Suppose $\#N(S-u) = \#N(T-u) = 5$. Since $t \in N(S-u)$, the down-step coupling in this case gives

$$\mathbb{E}\left(d(X,Y) \mid u \text{ dropped}\right) = \frac{4}{5}.$$

Suppose $\#N(S-u) = 5$ and $\#N(T-u) = 4$. Again since $t \in N(S-u)$, in this case we have

$$\mathbb{E}\left(d(X,Y) \mid u \text{ dropped}\right) \leq \frac{1}{5} \cdot 0 + \frac{3}{5} \cdot 1 + \frac{1}{5} \cdot 2 = 1.$$

Suppose $\#N(S-u) = \#N(T-u) = 4$. As seen in the proof of Theorem 6.4.1, if $S - u + t \notin \mathcal{B}$, we can couple so the distance is always 1, so we assume $t \in S-u$ and $s \in T-u$. Then we must have $\{u, d\} \subseteq N(S-u) \cap N(T-u)$ for some d. Conditioned on being in this case, we have

$$\mathbb{E}\left(d(X,Y) \mid u \text{ dropped}\right) \leq \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot 1 + 2 \cdot \frac{1}{4} = 1.$$

In all cases with $u \neq s, t$ the expected distance is less than or equal to 1, so we see the basis exchange walk on this matroid has positive curvature. □

[1]Provided by David Eppstein, CC0.

Graphic matroids

Now we consider a specific class of graphic matroids that induce positively curved basis exchange walks. This shows that in addition to having non-negatively curved basis exchange walks on graphic matroids induced by sufficiently small graphs, one can have non-negatively curved basis exchange walks on graphic matroids induced by arbitrarily large graphs.

Lemma 6.4.8. *Basis exchange walks over graphic matroids induced by graphs with edge-disjoint cycles are positively curved.*

Proof. We express the graph as a union of cycles, C_i, and edges with effective resistance 1, G':

$$G = C_1 + \ldots + C_k + G'.$$

Any spanning tree S of the graph must be a union of spanning trees S_i of each cycle and G' so we have

$$S = S_1 + \ldots + S_k + G'.$$

Any adjacent spanning tree T must differ in only one spanning tree of one cycle. Without loss of generality we assume the different spanning tree is in the cycle labeled as C_1, so

$$T = T_1 + S_2 + \ldots + S_k + G'.$$

We must show for any $u \in S \cap T$, we can couple walks from S and T conditioned on dropping u so that the expected distance is not greater than 1. We consider different cases for u.

Suppose $u \in G'$. For any $B \in \mathcal{B}$ we have $N(B-u) = \{u\}$ (i.e, the only edge that can be added on the up step is u). For this case, we have

$$\mathbb{E}\left(d(X,Y) \mid u \text{ dropped}\right) = 1.$$

Suppose $u \in S_i$ with $i \neq 1$, then $N(S-u) = N(T-u) = \{u, u'\}$ where $\{u'\} = C_i \backslash S_i$. We can always add the same element to each of $S-u$ and $T-u$, so we again have

$$\mathbb{E}\left(d(X,Y) \mid u \text{ dropped}\right) = 1.$$

Suppose $u \in S_1$, then $N(S-u) = \{u, t\}$ and $N(T-u) = \{u, s\}$ where $t = C_1 \backslash S_1$ and $s = C_1 \backslash T_1$ (similar to previous notation, s is simply the edge included in S but not in T). Since $S_1 = C_1 - t$ and $T_1 = C_1 - s$, $S_1 - u + t = T_1 - u + s$. We couple by adding u to both S and T or adding t to S and s to T. In this case, we have

$$\mathbb{E}\left(d(X,Y) \mid u \text{ dropped}\right) = 1/2.$$

This covers all cases, so the basis exchange walk is positively curved. □

6.5 An upper bound and negatively curved examples

In this section we give an upper bound on the curvature of an edge (S, T) of a basis exchange walk in terms of the sets $N(T-u)$ and $N(S-u)$, given in Theorem 6.5.1. Using this, we then give explicit examples of negatively curved basis exchange walks.

Note that to prove an upper bound on the curvature of an edge, we will need to prove a lower bound on the expected distance of two walks starting from S, T for *all possible couplings*. We do so by exhaustively considering *all* events where the distance can drop to 0, and upper bounding the probability that the distance goes to 0 under any coupling. We then consider events where the distance *must* be at least 2 for any coupling and lower bound this probability.

An upper bound

We now give an upper bound on the curvature of the basis exchange walk (we note the main function of this bound will be to prove specific examples have negative curvature). In order to state this bound, we fix any $S \sim T$ and define the following notation.

Definition 6.5.1. Let $J = \{u : t \in N(S-u) \text{ and } u \neq s\}$.

Definition 6.5.2. For any $u \in J$, let $A_u = (N(S-u)-t)\backslash N(T-u)$, i.e., the elements that are not t and that can be added to $S-u$ but *cannot* be added to $T-u$.

Remark 20. These sets will be to identify neighbors of S that are all far from all but a small number of neighbors of T (see Proposition 6.5.2 for a more precise statement). If $t \notin N(S-u)$, then as seen in the proof of Theorem 6.4.1, $N(S-u) = N(T-u)$. In this case, every neighbor of S induced by dropping u has a unique neighbor of T induced by dropping u.

Theorem 6.5.1. *Consider any bases $S \sim T$ of a matroid $\mathcal{M}$. Then the basis exchange walk on $\mathcal{M}$ satisfies*

$$\kappa \leq \frac{1}{k} + \frac{1}{k} \cdot \sum_{u \in J} \left(\frac{1}{\#N(T-u)} - \frac{\#A_u}{\#N(S-u)} \right).$$

Proof. Let S and T be adjacent bases with $S = \{s, u_1, \ldots, u_{k-1}\}$ and $T = \{t, u_1, \ldots, u_{k-1}\}$. Let $X \sim P(S, \cdot)$ and $Y \sim P(T, \cdot)$. Note for any coupling we have

$$\begin{aligned} \mathbb{E}d(X,Y) &\geq 0 \cdot \mathbb{P}(d(X,Y)=0) + 1 \cdot \mathbb{P}(d(X,Y)=1) + 2 \cdot \mathbb{P}(d(X,Y) \geq 2) && (6.4) \\ &= (1 - \mathbb{P}(d(X,Y)=0) - \mathbb{P}(d(X,Y) \geq 2)) + 2 \cdot \mathbb{P}(d(X,Y) \geq 2) && (6.5) \\ &= 1 - \mathbb{P}(d(X,Y)=0) + \mathbb{P}(d(X,Y) \geq 2). && (6.6) \end{aligned}$$

Recalling the definition of curvature, we see then

$$\kappa \leq \mathbb{P}(d(X,Y)=0) - \mathbb{P}(d(X,Y) \geq 2).$$

We will upper bound this quantity for any coupling.

We start by considering events where $d(X,Y) \geq 2$. In particular, we want to identify neighbors of S that are from most neighbors of T. The following proposition shows that any neighbor of S of the form $S-u+a$ for some $u \in J$ and some $a \in A_u$ is far from almost all neighbors of T.

Proposition 6.5.2. *For any $u \in J$ and $a \in A_u$, $S-u+a$ is at distance at least 2 from any neighbor of T aside from $T-u+s$, $T-t+s$, or $T-t+a$ (if these sets are bases).*

Proof. Without loss of generality, let $u = u_1$. Then $S-u+a = \{s, a, u_2, \ldots, u_{k-1}\}$. We proceed by cases on which element T drops.

Case 1: T drops u_1 and adds some element b, so $T-u_1+b = \{t, b, u_2, \ldots, u_{k-1}\}$.
Recall by definition of A_{u_1}, $b \notin N(T-u_1)$ so $b \neq a$. Since $a \neq u_1, t$, these are at distance 2 unless $b = s$.

Case 2: T drops t and adds some element b, so $T-t+b = \{b, u_1, \ldots, u_{k-1}\}$.
Since $a \neq u_1$, these are at distance 2 unless $b = s$ or $b = a$.

Case 3: T drops $u_i \neq u_1$ and adds b.
Then $T-u_i+b$ contains t and u_1, neither of which are contained in $S-u_1+a$, so $T-u_i+b$ is at distance at least 2 from $S-u_1+a$. □

We define the following relevant events:

$$\mathcal{F}_{u,a} := \{S \text{ drops } u \text{ and adds } a \in A_u\} \quad \text{for any } u \in J, a \in A_u \tag{6.7}$$
$$\mathcal{B}_u := \{T \text{ drops } u \text{ and adds } s\} \quad \text{for any } u \in J, \tag{6.8}$$
$$\mathcal{E}_a := \{T \text{ drops } t \text{ and adds } a\} \quad \text{for any } a \in N(T-t). \tag{6.9}$$

The above proposition shows

$$\mathbb{P}(d(X,Y) \geq 2) \geq \sum_{u \in J} \sum_{a \in A_u} \mathbb{P}(\mathcal{F}_{u,a}) - \mathbb{P}(\mathcal{F}_{u,a} \cap \mathcal{B}_u) - \mathbb{P}(\mathcal{F}_{u,a} \cap \mathcal{E}_s) - \mathbb{P}(\mathcal{F}_{u,a} \cap \mathcal{E}_a). \tag{6.10}$$

Now we consider events where $d(X,Y) = 0$. The following table gives all possible pairs of exchanges by S and T so that $d(X,Y) = 0$.

Exchange by S	Exchange by T
S drops s, adds t	T drops u, adds u for any $u \in T$
S drops s, adds $a \neq t$	T drops t, adds a
S drops $u \neq s$, adds u	T drops t, adds s
S drops $u \in J$, adds t	T drops u, adds s

We define the additional relevant events, based on the exchange made by S:

$$\mathcal{C}_{1,a} := \{S \text{ drops } s, \text{ adds } a\}, \tag{6.11}$$
$$\mathcal{C}_{2,u} := \{S \text{ drops } u, \text{ adds } u\} \quad \text{for any } u \in S, u \neq s, \tag{6.12}$$
$$\mathcal{C}_{3,u} := \{S \text{ drops } u, \text{ adds } t\} \quad \text{for any } u \in J. \tag{6.13}$$

Then we have

$$\mathbb{P}(d(X,Y)=0) \leq \mathbb{P}(\mathcal{C}_{1,t}) + \sum_{a\neq t} \mathbb{P}(\mathcal{C}_{1,a} \cap \mathcal{E}_a) + \sum_{\substack{u\in S \\ u\neq s}} \mathbb{P}(\mathcal{C}_{2,u} \cap \mathcal{E}_s) + \sum_{u\in J} \mathbb{P}(\mathcal{C}_{3,u} \cap \mathcal{B}_u) \quad (6.14)$$

Now we want to combine these two expressions. First note that if we collect some of the terms in the difference between 6.14 and 6.10, we can simplify as follows:

$$\mathbb{P}(\mathcal{C}_{1,t}) + \sum_{a\neq t} \mathbb{P}(\mathcal{C}_{1,a} \cap \mathcal{E}_a) + \sum_{\substack{u\in S \\ u\neq s}} \mathbb{P}(\mathcal{C}_{2,u} \cap \mathcal{E}_s) + \sum_{u\in J} \sum_{a\in A_u} (\mathbb{P}(\mathcal{F}_{u,a} \cap \mathcal{E}_s) + \mathbb{P}(\mathcal{F}_{u,a} \cap \mathcal{E}_a)) \quad (6.15)$$

$$\leq \mathbb{P}(\mathcal{C}_{1,t}) + \sum_{\substack{a\in N(T-t) \\ a\neq t}} \mathbb{P}(\mathcal{E}_a) = \frac{1}{k} \cdot \left(\frac{1}{\#N(S-s)} + \frac{\#N(S-s)-1}{\#N(S-s)} \right) = \frac{1}{k}. \quad (6.16)$$

Passing to the full expression we have

$$\begin{aligned} \mathbb{P}(d(X,Y)=0) - \mathbb{P}(d(X,Y)\geq 2) &\leq \frac{1}{k} + \sum_{u\in J} \mathbb{P}(\mathcal{C}_{3,u} \cap \mathcal{B}_u) \\ &\quad + \sum_{u\in J} \sum_{a\in A_u} \mathbb{P}(\mathcal{F}_{u,a} \cap \mathcal{B}_u) - \sum_{u\in J} \sum_{u\in A_u} \mathbb{P}(\mathcal{F}_{u,a}) \\ &\leq \frac{1}{k} + \sum_{u\in J} \mathbb{P}(\mathcal{B}_u) - \sum_{u\in J} \sum_{u\in A_u} \mathbb{P}(\mathcal{F}_{u,a}). \end{aligned}$$

Explicitly computing these probabilities, we have

$$\kappa \leq \frac{1}{k} + \frac{1}{k} \sum_{u\in J} \left(\frac{1}{\#N(T-u)} - \frac{\#A_u}{\#N(S-u)} \right). \quad (6.17)$$

□

An immediate corollary of the above is a sufficient condition for negative curvature:

Corollary 6.5.3. *The basis exchange walk on $\mathcal{M}$ is negatively curved if there exist bases $S \sim T$ such that*

$$\sum_{u\in J} \left(\frac{\#A_u}{\#N(S-u)} - \frac{1}{\#N(T-u)} \right) > 1.$$

Negatively curved examples

By Corollary 6.4.3, all basis exchange walks over rank-2 matroids are positively curved. However, as soon as we move to rank-3 matroids, this is no longer true, as shown in the example below. One might then hope that certain classes of matroids, for example regular matroids, induce non-negatively curved basis exchange walks. We show this is not the case with an explicit example: we give a graphic matroid (thus also a regular matroid) with a negatively curved basis exchange walk.

Rank-3 matroid

The bound from Theorem 6.5.1 says that in order to ensure negative curvature, it suffices to have S and T so that $N(S-u)$ is very different from $N(T-u)$ for many u. We construct a matroid with this property.

Let $E = \{s, t, u, u', v_1, \ldots, v_5, w_1, \ldots, w_5\}$. We define a matroid over E by defining the bases:

- $S = \{s, u, u'\}$,
- $T = \{t, u, u'\}$,
- $\{s, t, u\}, \{s, t, u'\}$,
- $\{s, u, v_i\}, \{s, u', v_i\}$ for all i,
- $\{t, u, w_i\}, \{t, u', w_i\}$ for all i,
- $\{u, u', v_i\}, \{u, u', w_i\}$ for all i, and
- $\{u, v_i, w_j\}, \{u', v_i, w_j\}$ for all i, j.

This is a $\mathbb{R}$-linear matroid, with the elements identified with the following vectors in $\mathbb{R}^3$:

$$s = \begin{bmatrix}1\\0\\0\end{bmatrix}, \quad t = \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad u = \begin{bmatrix}0\\0\\1\end{bmatrix}, \quad u' = \begin{bmatrix}1\\1\\1\end{bmatrix}, \quad v_i = \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad w_i = \begin{bmatrix}1\\0\\0\end{bmatrix}.$$

We have $J = \{u, u'\}$, $\#A_j = 5$, and $N(S-j) = N(T-j) = 7$ for $j \in J$, so by Theorem 6.5.1

$$\kappa \leq \frac{1}{3} + \frac{1}{3} \cdot \left(\frac{2}{7} - \frac{2 \cdot 5}{7}\right) = -\frac{1}{21}. \tag{6.18}$$

Graphic matroid

We want to construct a graphic matroid with spanning trees S and T so that $N(S-u)$ and $N(T-u)$ are very different for many u. We look at the graphic matroid induced by K_6; in order to make the set J as large as possible, we choose S and T to be paths. We label the "outer" edges of K_6 as in Figure 6.2, and consider the spanning trees $S = \{s, 1, 2, 3, 4\}$ and $T = \{t, 1, 2, 3, 4\}$.

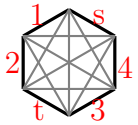


Figure 6.2: Labeling of outer edges of K_6.

We count the relevant quantities for Theorem 6.5.1. In the following table, let i be the element dropped in the "down" step of the down-up walk.

i	$\#N(S-i)$	$\#N(T-i)$	$\#A_i$
1	$2 \cdot 4$	$1 \cdot 5$	5
2	$1 \cdot 5$	$2 \cdot 4$	2
3	$1 \cdot 5$	$2 \cdot 4$	2
4	$2 \cdot 4$	$1 \cdot 5$	5

Table 6.2: Quantities needed for computing curvature upper bound

By Theorem 6.5.1 we have

$$\kappa \leq \frac{1}{5} + \frac{1}{5}\left(\frac{2}{5} + \frac{2}{8} - \frac{2 \cdot 5}{8} - \frac{2 \cdot 2}{5}\right) = -\frac{2}{25}. \tag{6.19}$$

Bibliography

[Akk10] Sofiane Akkouche. The Spectral Bounds of the Discrete Schrödinger Operator. *Journal of Functional Analysis*, 259(6):1443–1465, 2010 (cited on page 105).

[ALG21] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral Independence in High-Dimensional Expanders and Applications to the Hardcore Model. *SIAM Journal on Computing*, 53(6):FOCS20–1, 2021 (cited on page 134).

[ALGV19] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. Log-Concave Polynomials II: High-Dimensional Walks and an FPRAS for Counting Bases of a Matroid. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC '19, pages 1–12. ACM, June 2019. DOI: 10.1145/3313276.3316385. URL: http://dx.doi.org/10.1145/3313276.3316385 (cited on pages 12, 135).

[ALGVV21] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, Cynthia Vinzant, and Thuy-Duong Vuong. Log-Concave Polynomials IV: Approximate Exchange, Tight Mixing Times, and Near-Optimal Sampling of Forests. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC '21, pages 408–420. ACM, June 2021. DOI: 10.1145/3406325.3451091. URL: http://dx.doi.org/10.1145/3406325.3451091 (cited on page 135).

[And+99] E. Anderson et al. *LAPACK Users' Guide*. 3rd edition. SIAM. 3rd edition, 1999. URL: https://netlib.org/lapack/lug/ (cited on pages 28, 29).

[AB96] Wolfgang Arendt and Charles J. K. Batty. The Spectral Bound of Schrödinger Operators. *Potential Analysis*, 5:207–230, 1996 (cited on page 105).

[AB97] Wolfgang Arendt and Charles J. K. Batty. The Spectral Function and Principal Eigenvalues for Schrödinger Operators. *Potential Analysis*, 7(1):415–436, 1997 (cited on page 105).

[BWZ24] Simon Becker, Jens Wittsten, and Maciej Zworski. *Discrete vs. Continuous in the Semiclassical Limit*. 2024 (cited on pages 104, 105).

[BH24] Erna Begović and Vjeran Hari. Convergence of the Complex Block Jacobi Methods Under the Generalized Serial Pivot Strategies. *Linear Algebra and its Applications*, 699:421–458, October 2024. ISSN: 0024-3795. DOI: 10.1016/j.laa.2024.07.012. URL: http://dx.doi.org/10.1016/j.laa.2024.07.012 (cited on page 35).

[BVZ97] Cédric Béguin, Alain Valette, and Andrzej Zuk. On the Spectrum of a Random Walk on the Discrete Heisenberg Group and the Norm of Harper's Operator. *Journal of Geometry and Physics*, 21(4):337–356, 1997 (cited on pages 9, 101, 104).

[BS89] Michael Berry and Ahmed Sameh. An Overview of Parallel Algorithms for the Singular Value and Symmetric Eigenvalue Problems. *Journal of Computational and Applied Mathematics*, 27(1-2):191–213, September 1989. ISSN: 0377-0427. DOI: 10.1016/0377-0427(89)90366-x. URL: http://dx.doi.org/10.1016/0377-0427(89)90366-X (cited on page 28).

[BJMMR25] Rajarshi Bhattacharjee, Rajesh Jayaram, Cameron Musco, Christopher Musco, and Archan Ray. Improved Spectral Density Estimation via Explicit and Implicit Deflation. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2693–2754. SIAM, 2025 (cited on pages 61, 63).

[Bis89] Christian H. Bischof. Computing the Singular Value Decomposition on a Distributed System of Vector Processors. *Parallel Computing*, 11:171–186, August 1989. DOI: 10.1016/0167-8191(89)90027-6 (cited on page 29).

[BCC+22] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. On Mixing of Markov Chains: Coupling, Spectral Independence, and Entropy Factorization. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3670–3692. Society for Industrial and Applied Mathematics, January 2022. ISBN: 9781611977073. DOI: 10.1137/1.9781611977073.145. URL: http://dx.doi.org/10.1137/1.9781611977073.145 (cited on page 132).

[BT06] Sergey G. Bobkov and Prasad Tetali. Modified Logarithmic Sobolev Inequalities in Discrete Settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006 (cited on page 133).

[BZ05] Florin P. Boca and Alexandru Zaharescu. Norm Estimates of Almost Mathieu Operators. *Journal of Functional Analysis*, 220(1):76–96, 2005 (cited on pages 9, 101, 105).

[BDHMW17a] Daniel Bump, Persi Diaconis, Angela Hicks, Laurent Miclo, and Harold Widom. An Exercise (?) in Fourier Analysis on the Heisenberg Group. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 26, pages 263–288, 2017 (cited on pages 9, 101, 105).

[BDHMW17b] Daniel Bump, Persi Diaconis, Angela Hicks, Laurent Miclo, and Harold Widom. Useful Bounds on the Extreme Eigenvalues and Vectors of Matrices for Harper's Operators. *Large Truncated Toeplitz Matrices, Toeplitz Operators, and Related Topics: The Albrecht Böttcher Anniversary Volume*:235–265, 2017 (cited on pages 9, 101, 105).

[CR12] Emmanuel J. Candès and Benjamin Recht. Exact Matrix Completion via Convex Optimization. *Communications of the ACM*, 55(6):111–119, 2012 (cited on pages 7, 59).

[CLRT22] Erin Carson, Kathryn Lund, Miroslav Rozložník, and Stephen Thomas. Block Gram-Schmidt Algorithms and Their Stability Properties. *Linear Algebra and its Applications*, 638:150–195, 2022. ISSN: 0024-3795. DOI: 10.1016/j.laa.2021.12.017 (cited on page 35).

[Cha62] B. A. Chartres. Adaptation of the Jacobi Method for a Computer with Magnetic-Tape Backing Store. *The Computer Journal*, 5:51–60, January 1962. DOI: 10.1093/comjnl/5.1.51 (cited on page 29).

[Che24] Tyler Chen. The Lanczos Algorithm for Matrix Functions: A Handbook for Scientists. *arXiv preprint arXiv:2410.11090*, 2024 (cited on page 65).

[CTU21] Tyler Chen, Thomas Trogdon, and Shashanka Ubaru. Analysis of Stochastic Lanczos Quadrature for Spectrum Approximation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 1728–1739. PMLR, 2021 (cited on page 63).

[CETW24] Yifan Chen, Ethan N. Epperly, Joel A. Tropp, and Robert J. Webber. Randomly Pivoted Cholesky: Practical Approximation of a Kernel Matrix with Few Entry Evaluations. *Communications on Pure and Applied Mathematics*, 78(5):995–1041, December 2024. ISSN: 1097-0312. DOI: 10.1002/cpa.22234. URL: http://dx.doi.org/10.1002/cpa.22234 (cited on page 34).

[CGM19] Mary Cryan, Heng Guo, and Giorgos Mousa. Modified Log-Sobolev Inequalities for Strongly Log-Concave Distributions. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1358–1370. IEEE, November 2019. DOI: 10.1109/focs.2019.00083. URL: http://dx.doi.org/10.1109/FOCS.2019.00083 (cited on page 135).

[Dav99] E. Brian Davies. Ground State Energy of Almost Periodic Schrödinger Operators. *Ergodic Theory and Dynamical Systems*, 19(3):591–609, 1999 (cited on page 105).

[dRij89] P. P. M. de Rijk. A One-Sided Jacobi Algorithm for Computing the Singular Value Decomposition on a Vector Computer. *SIAM Journal on Scientific and Statistical Computing*, 10(2):359–371, March 1989. ISSN: 2168-3417. DOI: 10.1137/0910023. URL: http://dx.doi.org/10.1137/0910023 (cited on page 33).

[DV92] James Demmel and Krešimir Veselić. Jacobi's Method is More Accurate than QR. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1204–1245, October 1992. ISSN: 1095-7162. DOI: 10.1137/0613074. URL: http://dx.doi.org/10.1137/0613074 (cited on pages 28–30, 39, 56, 57).

[Dem97] James W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. DOI: 10.1137/1.9781611971446. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611971446 (cited on pages 28, 29, 33).

[DS26a] Michał Dereziński and Aaron Sidford. Approaching Optimality for Solving Dense Linear Systems with Low-Rank Structure. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 925–938. SIAM, 2026. DOI: 10.1137/1.9781611978971.37 (cited on pages 6, 7, 58, 61, 62, 64, 83, 89).

[DRVW06] Amit Deshpande, Luis Rademacher, Santosh S. Vempala, and Grant Wang. Adaptive Sampling and Fast Low-Rank Matrix Approximation. *Theory of Computing*, 2(1):225–247, 2006. ISSN: 1557-2862. DOI: 10.4086/toc.2006.v002a012. URL: http://dx.doi.org/10.4086/toc.2006.v002a012 (cited on page 34).

[DV06] Amit Deshpande and Santosh Vempala. *Adaptive Sampling and Fast Low-Rank Matrix Approximation*. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Springer Berlin Heidelberg, 2006, pages 292–303. ISBN: 9783540380450. DOI: 10.1007/11830924_28. URL: http://dx.doi.org/10.1007/11830924_28 (cited on page 34).

[Det25] Isabel Detherage. Curvature of Basis Exchange Walks. *arXiv preprint arXiv:2509.18510*, 2025 (cited on pages 2, 131).

[DS25] Isabel Detherage and Rikhav Shah. Matrix Factorizations with Uniformly Random Pivoting. *arXiv preprint arXiv:2505.02023*, 2025 (cited on page 2).

[DS26b] Isabel Detherage and Rikhav Shah. A Kaczmarz-Inspired Method for Orthogonalization. *Quarterly of Applied Mathematics*, 84(1):31–43, 2026 (cited on pages 2, 34, 46).

[DSS24] Isabel Detherage, Nikhil Srivastava, and Zachary Stier. On the Ground State Energies of Discrete and Semiclassical Schrödinger Operators. *Pure and Applied Analysis*, 6(4):955–976, 2024 (cited on page 2).

[DPS16] Edoardo Di Napoli, Eric Polizzi, and Yousef Saad. Efficient Estimation of Eigenvalue Counts in an Interval. *Numerical Linear Algebra with Applications*, 23(4):674–692, 2016 (cited on page 63).

[DS93] David A. Drabold and Otto F. Sankey. Maximum Entropy Approach for Linear Scaling in the Electronic Structure Problem. *Physical Review Letters*, 70(23):3631, 1993 (cited on pages 6, 59).

[Drm94] Zlatko Drmač. *The Generalized Singular Value Problem*. PhD thesis, FernUniversität in Hagen, Hagen, Germany, 1994 (cited on page 29).

[Drm97] Zlatko Drmač. Implementation of Jacobi Rotations for Accurate Singular Value Computation in Floating Point Arithmetic. *SIAM Journal on Scientific Computing*, 18(4):1200–1222, July 1997. ISSN: 1095-7197. DOI: 10.1137/s1064827594265095. URL: http://dx.doi.org/10.1137/S1064827594265095 (cited on page 30).

[Drm10] Zlatko Drmač. A Global Convergence Proof for Cyclic Jacobi Methods with Block Rotations. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1329–1350, 2010. DOI: 10.1137/090748548 (cited on page 34).

[DV08] Zlatko Drmač and Krešimir Veselić. New Fast and Accurate Jacobi SVD Algorithm. I. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1322–1342, January 2008. ISSN: 1095-7162. DOI: 10.1137/050639193. URL: http://dx.doi.org/10.1137/050639193 (cited on page 29).

[DC70] Francois Ducastelle and Françoise Cyrot-Lackmann. Moments Developments and Their Application to the Electronic Charge Distribution of d Bands. *Journal of Physics and Chemistry of Solids*, 31(6):1295–1306, 1970 (cited on pages 6, 59).

[EY06] Alexandre Eremenko and Peter Yuditskii. Uniform Approximation of $\operatorname{sgn}(x)$ by Polynomials and Entire Functions. *arXiv preprint math/0604324*, 2006 (cited on page 74).

[Eva22] Lawrence C. Evans. *Partial Differential Equations*, volume 19. American Mathematical Society, 2022 (cited on page 127).

[FH60] George E. Forsythe and Peter Henrici. The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix. *Transactions of the American Mathematical Society*, 94(1):1–23, 1960. ISSN: 1088-6850. DOI: 10.1090/s0002-9947-1960-0109825-2. URL: http://dx.doi.org/10.1090/S0002-9947-1960-0109825-2 (cited on page 33).

[FKV04] Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast Monte-Carlo Algorithms for Finding Low-Rank Approximations. *Journal of the ACM*, 51(6):1025–1041, November 2004. ISSN: 1557-735X. DOI: 10.1145/1039488.1039494. URL: http://dx.doi.org/10.1145/1039488.1039494 (cited on page 34).

[FKNYY20] Takeshi Fukaya, Ramaseshan Kannan, Yuji Nakatsukasa, Yusaku Yamamoto, and Yuka Yanagisawa. Shifted Cholesky QR for Computing the QR Factorization of Ill-Conditioned Matrices. *SIAM Journal on Scientific Computing*, 42(1):A477–A503, January 2020. ISSN: 1095-7197. DOI: 10.1137/18m1218212. URL: http://dx.doi.org/10.1137/18M1218212 (cited on page 29).

[GGS92] Fritz Gesztesy, Gian Michele Graf, and Barry Simon. The Ground State Energy of Schrödinger Operators. *Communications in Mathematical Physics*, 150:375–384, 1992 (cited on pages 101, 105).

[GS02] Alison L. Gibbs and Francis Edward Su. On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435, 2002 (cited on page 81).

[GMvN59] Herman H. Goldstine, Francis J. Murray, and John von Neumann. The Jacobi Method for Real Symmetric Matrices. *Journal of the ACM*, 6(1):59–96, January 1959. ISSN: 1557-735X. DOI: 10.1145/320954.320960. URL: http://dx.doi.org/10.1145/320954.320960 (cited on page 34).

[GvdV00] Gene H. Golub and Henk A. van der Vorst. Eigenvalue Computation in the 20th Century. *Journal of Computational and Applied Mathematics*, 123:35–65, November 2000. DOI: 10.1016/s0377-0427(00)00413-1 (cited on page 28).

[GV13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013 (cited on pages 29, 33).

[Grc11] Joseph F. Grcar. How Ordinary Elimination Became Gaussian Elimination. *Historia Mathematica*, 38(2):163–218, 2011. ISSN: 0315-0860. DOI: 10.1016/j.hm.2010.06.003 (cited on page 29).

[Gre53] Robert T. Gregory. Computing Eigenvalues and Eigenvectors of a Symmetric Matrix on the ILLIAC. *Mathematics of Computation*, 7(44):215–220, 1953. ISSN: 1088-6842. DOI: 10.1090/s0025-5718-1953-0057643-6. URL: http://dx.doi.org/10.1090/S0025-5718-1953-0057643-6 (cited on page 33).

[Gut01] Ivan Gutman. The Energy of a Graph: Old and New Results. In *Algebraic Combinatorics and Applications*, pages 196–211. Springer, 2001 (cited on pages 7, 59).

[GPTV15] Stefan Güttel, Eric Polizzi, Ping Tak Peter Tang, and Gautier Viaud. Zolotarev Quadrature Rules and Load Balancing for the FEAST Eigensolver. *SIAM Journal on Scientific Computing*, 37(4):A2100–A2122, 2015 (cited on page 62).

[HHT08] Nicholas Hale, Nicholas J. Higham, and Lloyd N. Trefethen. Computing A^α, $\log(A)$, and Related Matrix Functions by Contour Integrals. *SIAM Journal on Numerical Analysis*, 46(5):2505–2523, 2008 (cited on page 63).

[HLM12] Moritz Hardt, Katrina Ligett, and Frank McSherry. A Simple and Practical Algorithm for Differentially Private Data Release. *Advances in Neural Information Processing Systems (NIPS)*, 25, 2012 (cited on pages 7, 59).

[HB17] Vjeran Hari and Erna Begović. Convergence of the Cyclic and Quasi-Cyclic Block Jacobi Methods. *Electronic Transactions on Numerical Analysis*, 46:107–147, 2017. DOI: 10.48550/ARXIV.1604.05825 (cited on page 35).

[Hes58] Magnus R. Hestenes. Inversion of Matrices by Biorthogonalization and Related Results. *Journal of the Society for Industrial and Applied Mathematics*, 6:51–90, March 1958. DOI: 10.1137/0106005 (cited on page 28).

[Hig90] Nicholas J. Higham. *Analysis of the Cholesky Decomposition of a Semi-Definite Matrix*. In *Reliable Numerical Computation*. Oxford University Press, September 1990, pages 161–186. ISBN: 9781383025675. DOI: 10.1093/oso/9780198535645.003.0010. URL: http://dx.doi.org/10.1093/oso/9780198535645.003.0010 (cited on page 29).

[Hig09] Nicholas J. Higham. Cholesky Factorization. *WIREs Computational Statistics*, 1(2):251–254, June 2009. ISSN: 1939-0068. DOI: 10.1002/wics.18. URL: http://dx.doi.org/10.1002/wics.18 (cited on page 29).

[Jac46] C. G. J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *Journal für die reine und angewandte Mathematik*, 1846(30):51–94, 1846. DOI: 10.1515/crll.1846.30.51 (cited on pages 28, 33).

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-Rank Matrix Completion Using Alternating Minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing (STOC)*, pages 665–674. ACM, 2013 (cited on pages 7, 59).

[JS19] Yujia Jin and Aaron Sidford. Principal Component Projection and Regression in Nearly Linear Time Through Asymmetric SVRG. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019 (cited on pages 62, 74).

[Jit21] Svetlana Jitomirskaya. On Point Spectrum of Critical Almost Mathieu Operators. *Advances in Mathematics*, 2021 (cited on page 108).

[KKN21] Marek Kaluba, Dawid Kielak, and Piotr W. Nowak. On Property (T) for Aut(F_n) and SL_n(Z). *Annals of Mathematics*, 193(2):539–562, 2021 (cited on pages 9, 101).

[KNO19] Marek Kaluba, Piotr W. Nowak, and Narutaka Ozawa. Aut(F_5) Has Property (T). *Mathematische Annalen*, 375:1169–1191, 2019 (cited on pages 9, 101).

[Kat52] Tosio Kato. Note on the Least Eigenvalue of the Hill Equation. *Quarterly of Applied Mathematics*, 10(3):292–294, 1952 (cited on pages 101, 105).

[Kog55] E. G. Kogbetliantz. Solution of Linear Equations by Diagonalization of Coefficients Matrix. *Quarterly of Applied Mathematics*, 13:123–132, July 1955. DOI: 10.1090/qam/88795 (cited on page 28).

[KW92] Jacek Kuczyński and Henryk Woźniakowski. Estimating the Largest Eigenvalue by the Power and Lanczos Algorithms with a Random Start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992 (cited on page 59).

[LP17] David Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, October 2017. ISBN: 9781470442323. DOI: 10.1090/mbk/107. URL: http://dx.doi.org/10.1090/mbk/107 (cited on page 133).

[LSY16] Lin Lin, Yousef Saad, and Chao Yang. Approximating Spectral Densities of Large Matrices. *SIAM Review*, 58(1):34–65, 2016 (cited on page 63).

[Liu21] Kuikui Liu. From Coupling to Spectral Independence and Blackbox Comparison with the Down-Up Walk. In *APPROX-RANDOM 2021*, 2021. DOI: 10.4230/LIPIcs.APPROX/RANDOM.2021.32. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.APPROX/RANDOM.2021.32 (cited on pages 11, 132, 136).

[Liu23] Kuikui Liu. *Spectral Independence: A New Tool to Analyze Markov Chains*. University of Washington, 2023 (cited on pages 11, 134).

[Luk86] Franklin T. Luk. A Jacobi-like Algorithm for Computing the QR-Decomposition. *SIAM Journal on Scientific and Statistical Computing*, 7(2):452–459, 1986. DOI: 10.1137/0907031. Previously circulated as Cornell University Technical Report, 1984 (cited on page 29).

[MJ17] Christoph A. Marx and Svetlana Jitomirskaya. Dynamics and Spectral Theory of Quasi-Periodic Schrödinger-Type Operators. *Ergodic Theory and Dynamical Systems*, 37(8):2353–2393, 2017 (cited on page 112).

[Mas94] Walter F. Mascarenhas. A Note on Jacobi Being More Accurate Than *QR*. *SIAM Journal on Matrix Analysis and Applications*, 15(1):215–218, January 1994. ISSN: 1095-7162. DOI: 10.1137/s089547989222792x. URL: http://dx.doi.org/10.1137/S089547989222792X (cited on page 29).

[Mat95] Roy Mathias. Accurate Eigensystem Computations by Jacobi Methods. *SIAM Journal on Matrix Analysis and Applications*, 16(3):977–1003, July 1995. ISSN: 1095-7162. DOI: 10.1137/s089547989324820x. URL: http://dx.doi.org/10.1137/S089547989324820X (cited on page 29).

[MMMW21] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal Stochastic Trace Estimation. In *Proceedings of the 2021 Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM, 2021 (cited on pages 6, 65, 78, 86).

[Mil13] Ognjen Milatovic. A Spectral Property of Discrete Schrödinger Operators with Non-Negative Potentials. *Integral Equations and Operator Theory*, 76:285–300, 2013 (cited on page 105).

[Mün23] Florentin Münch. Ollivier Curvature, Isoperimetry, Concentration, and Log-Sobolev Inequalities. *arXiv preprint arXiv:2309.06493*, 2023 (cited on page 136).

[MM15] Cameron Musco and Christopher Musco. Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition. In *Advances in Neural Information Processing Systems (NIPS 2015)*, volume 28, pages 1396–1404. Curran Associates, Inc., 2015 (cited on pages 59, 86).

[MMS18] Cameron Musco, Christopher Musco, and Aaron Sidford. Stability of the Lanczos Method for Matrix Function Approximation. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1605–1624. SIAM, 2018 (cited on page 65).

[MNSUW17] Cameron Musco, Praneeth Netrapalli, Aaron Sidford, Shashanka Ubaru, and David P. Woodruff. Spectrum Approximation Beyond Fast Matrix Multiplication: Algorithms and Hardness. *arXiv preprint arXiv:1704.04163*, 2017 (cited on pages 61, 63, 70, 77, 82).

[NF16] Yuji Nakatsukasa and Roland W. Freund. Using Zolotarev's Rational Approximation for Computing the Polar, Symmetric Eigenvalue, and Singular Value Decompositions. *SIAM Review*, 58(3):463–493, 2016 (cited on page 62).

[NT25] Yuji Nakatsukasa and Lloyd N. Trefethen. Applications of AAA Rational Approximation, 2025. arXiv: `2510.16237 [math.NA]`. URL: `https://arxiv.org/abs/2510.16237` (cited on page 62).

[ODo14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014 (cited on page 107).

[Oll09] Yann Ollivier. Ricci Curvature of Markov Chains on Metric Spaces. *Journal of Functional Analysis*, 256(3):810–864, February 2009. ISSN: 0022-1236. DOI: `10.1016/j.jfa.2008.11.001`. URL: `http://dx.doi.org/10.1016/j.jfa.2008.11.001` (cited on pages 11, 131, 136).

[Oza16] Narutaka Ozawa. Noncommutative Real Algebraic Geometry of Kazhdan's Property (T). *Journal of the Institute of Mathematics of Jussieu*, 15:85–90, 2016 (cited on pages 9, 101).

[Oza22] Narutaka Ozawa. A Substitute for Kazhdan's Property (T) for Universal Non-Lattices. *arXiv preprint arXiv:2207.05272*, 2022 (cited on pages 9, 101, 105).

[Par98] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1998. DOI: `10.1137/1.9781611971163`. URL: `https://epubs.siam.org/doi/book/10.1137/1.9781611971163` (cited on page 28).

[PMM25] David Persson, Raphael A. Meyer, and Christopher Musco. Algorithm-Agnostic Low-Rank Approximation of Operator Monotone Matrix Functions. *SIAM Journal on Matrix Analysis and Applications*, 46(1):1–21, 2025. DOI: 10.1137/23M1619435 (cited on pages 62, 86).

[PP11] Penco Petrov Petrushev and Vasil Atanasov Popov. *Rational Approximation of Real Functions*, number 28 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2011 (cited on pages 7, 59, 61, 66, 67).

[QZL05] Li Qiu, Yanxia Zhang, and Chi-Kwong Li. Unitarily Invariant Metrics on the Grassmann Space. *SIAM Journal on Matrix Analysis and Applications*, 27(2):507–531, 2005. DOI: 10.1137/040607605 (cited on page 53).

[RL18] Aditya Ramesh and Yann LeCun. Backpropagation for Implicit Spectral Densities. *arXiv preprint arXiv:1806.00499*, 2018 (cited on pages 6, 59).

[RS81] Michael Reed and Barry Simon. *I: Functional Analysis*, volume 1. Academic Press, 1981 (cited on page 101).

[Saa03] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003 (cited on page 65).

[Saa11] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, 2011 (cited on page 65).

[Sal21] Justin Salez. Sparse Expanders Have Negative Curvature. *arXiv preprint arXiv:2101.08242*, 2021 (cited on page 137).

[Sal23] Justin Salez. Cutoff for Non-Negatively Curved Markov Chains. *Journal of the European Mathematical Society*, 26(11):4375–4392, 2023 (cited on page 136).

[Sch64] Arnold Schönhage. Zur quadratischen Konvergenz des Jacobi-Verfahrens. *Numerische Mathematik*, 6(1):410–412, December 1964. ISSN: 0945-3245. DOI: 10.1007/bf01386091. URL: http://dx.doi.org/10.1007/BF01386091 (cited on pages 28, 33).

[ST14] Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014 (cited on page 6).

[Ste25] Stefan Steinerberger. Kaczmarz Kac Walk, 2025. URL: https://api.semanticscholar.org/CorpusID:276384030 (cited on pages 3, 30, 34, 46).

[SV09] Thomas Strohmer and Roman Vershynin. A Randomized Kaczmarz Algorithm with Exponential Convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009. DOI: 10.1007/s00041-008-9030-4 (cited on page 25).

[TP14] Ping Tak Peter Tang and Eric Polizzi. FEAST as a Subspace Iteration Eigensolver Accelerated by Approximate Spectral Projection. *SIAM Journal on Matrix Analysis and Applications*, 35(2):354–390, 2014 (cited on page 62).

[TB97] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. DOI: 10.1137/1.9780898719574 (cited on page 29).

[Tur88] I. Turek. A Maximum-Entropy Approach to the Density of States Within the Recursion Method. *Journal of Physics C: Solid State Physics*, 21(17):3251–3260, 1988 (cited on pages 6, 59).

[Van69] Abraham Van der Sluis. Condition Numbers and Equilibration of Matrices. *Numerische Mathematik*, 14(1):14–23, 1969 (cited on page 39).

[VH89] Krešimir Veselić and Vjeran Hari. A Note on a One-Sided Jacobi Algorithm. *Numerische Mathematik*, 56(6):627–633, June 1989. ISSN: 0945-3245. DOI: 10.1007/bf01396349. URL: http://dx.doi.org/10.1007/BF01396349 (cited on page 29).

[YNYF16] Yusaku Yamamoto, Yuji Nakatsukasa, Yuka Yanagisawa, and Takeshi Fukaya. Roundoff Error Analysis of the CholeskyQR2 Algorithm in an Oblique Inner Product. *JSIAM Letters*, 8(0):5–8, 2016. ISSN: 1883-0617. DOI: 10.14495/jsiaml.8.5. URL: http://dx.doi.org/10.14495/jsiaml.8.5 (cited on page 29).

[YGKM20] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: Neural Networks Through the Lens of the Hessian. In *2020 IEEE International Conference on Big Data*, pages 581–590. IEEE, 2020 (cited on pages 6, 59).

[Zol77] Egor Ivanovich Zolotarev. Application of Elliptic Functions to Questions of Functions Deviating Least and Most from Zero. *Zap. Imp. Akad. Nauk. St. Petersburg*, 30(5):1–59, 1877 (cited on page 62).