Title

Random Matrix Theory in Numerical Linear Algebra

Author

Kulkarni, Archit

Publication Date

2020

Peer reviewed|Thesis/dissertation

Random Matrix Theory in Numerical Linear Algebra

by

Archit U. Kulkarni

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:

Assistant Professor Nikhil Srivastava, Chair
Associate Professor Prasad Raghavendra
Professor Marc Rieffel

Spring 2020

Random Matrix Theory in Numerical Linear Algebra

Abstract

Random Matrix Theory in Numerical Linear Algebra

by

Archit U. Kulkarni

Doctor of Philosophy in Mathematics

University of California, Berkeley

Assistant Professor Nikhil Srivastava, Chair

We use techniques from random matrix theory and high-dimensional probability to shed light on several problems in numerical linear algebra. We focus on two main topics: (1) the problem of approximately computing the eigenvalues and eigenvectors of a given non-Hermitian matrix, and (2) the problem of approximating the spectral distribution and the extreme eigenvalues of a Hermitian matrix via the Lanczos algorithm.

Diagonalization. We confirm a 2007 conjecture of Davies [37] that for each $\delta \in (0,1)$, every matrix $A \in \mathbb{C}^{n\times n}$ is at least $\delta\|A\|$-close to one whose eigenvectors have condition number at worst c_n/δ, for some c_n depending only on n. We further show that the dependence on δ cannot be improved to $1/\delta^p$ for any constant $p < 1$.

Our proof uses tools from random matrix theory to show that the pseudospectrum of A can be regularized with the addition of a complex Gaussian perturbation. Along the way, we explain how a variant of a theorem of Śniady implies a conjecture of Sankar, Spielman and Teng on the optimal constant for smoothed analysis of condition numbers.

Next, using this idea of adding a complex Gaussian perturbation as a preprocessing step, we exhibit a randomized algorithm which given a square matrix $A \in \mathbb{C}^{n\times n}$ with $\|A\| \le 1$ and $\delta > 0$, computes with high probability an invertible V and diagonal D such that

$$\|A - VDV^{-1}\| \le \delta$$

in $O(T_{\mathsf{MM}}(n)\log^2(n/\delta))$ arithmetic operations on a floating point machine with $O(\log^4(n/\delta)\log n)$ bits of precision. The computed similarity V additionally satisfies $\|V\|\|V^{-1}\| \le O(n^{2.5}/\delta)$. Here $T_{\mathsf{MM}}(n)$ is the number of arithmetic operations required to multiply two $n \times n$ complex matrices numerically stably, known to satisfy $T_{\mathsf{MM}}(n) = O(n^{\omega+\eta})$ for every $\eta > 0$ where ω is the exponent of matrix multiplication [48]. After the initial Gaussian perturbation, the remainder of the algorithm is a variant of the spectral bisection algorithm in numerical linear algebra [17]. Our running time is optimal up to polylogarithmic factors, in the sense that

verifying that a given similarity diagonalizes a matrix requires at least matrix multiplication time.

The Lanczos algorithm. We study the Lanczos algorithm where the initial vector is sampled uniformly from $\mathbb{S}^{n-1}$. Let A be an $n \times n$ Hermitian matrix. We show that when run for few iterations, the output of the algorithm on A is almost deterministic. More precisely, we show that for any $\varepsilon \in (0,1)$ there exists $c > 0$ depending only on ε and a certain global property of the spectrum of A (in particular, not depending on n) such that when Lanczos is run for at most $c \log n$ iterations, the Jacobi coefficients and the Ritz values deviate from their medians by t with probability at most $\exp(-n^{\varepsilon} t^2)$, for $t < \|A\|$. A similar result is derived for the Ritz vectors. The proof relies on the local Lévy lemma, a tool in high-dimensional probability regarding concentration of measure for functions that are Lipschitz on a large region of the sphere, as well as on classical connections between the Lanczos algorithm and orthogonal polynomials.

Furthermore, we show that the Lanczos algorithm fails with high probability to identify outliers of the spectrum when run for at most $c' \log n$ iterations, where again c' depends only on the same global property of the spectrum of A. Classical results imply that the bound $c' \log n$ is tight up to a constant factor.

Our techniques also yield asymptotic results: Suppose we have a sequence of Hermitian matrices $A_n \in M_n(\mathbb{C})$ whose spectral distributions converge in Kolmogorov distance with rate $O(n^{-\varepsilon})$ to a density, for some $\varepsilon > 0$. Then we show that for large enough n, and for $k = O(\sqrt{\log n})$, the Ritz values after k iterations concentrate around the roots of the kth orthogonal polynomial with respect to the limiting density.

Contents

Acknowledgments

It is a pleasure to thank my advisor Nikhil Srivastava for his generosity, enthusiasm and patience. It is an understatement to say this thesis could not have happened without his guidance. His mentorship will have a lasting influence on me, and it is an immense privilege to have been his student.

I thank my coauthors and fellow graduate students Jess Banks, Jorge Garza Vargas, and Satyaki Mukherjee, for a joyful and productive collaboration.

I would also like to thank Benson Au, Aurelien Gribinski, Jonathan Leake, Mohan Ravichandran, and Nick Ryder, for countless helpful mathematical conversations.

Part of the work in this thesis was done at the Institute of Pure and Applied Mathematics at UCLA, and at the Simons Institute for the Theory of Computing at UC Berkeley. I thank these institutions for providing an excellent setting to meet collaborators and do research. I am also grateful for the financial support of a James H. Simons fellowship.

Finally, I thank my family for their guidance and unwavering support.

Chapter 1

Introduction

Problems involving eigenvalues and eigenvectors of $n \times n$ matrices for large n are ubiquitous in science and engineering. By the Abel-Ruffini theorem, for $n \geq 5$ there is no formula for the eigenvalues and eigenvectors involving a finite number of arithmetic operations, powers and kth roots—that is, there is no algorithm to compute these quantities exactly in finite time. A significant part of the field of numerical linear algebra is concerned with constructing algorithms that provably and efficiently compute *approximate* eigenvalues, eigenvectors, matrix factorizations and so on.

The field of random matrix theory, on the other hand, does not traditionally deal with algorithmic questions. In random matrix theory, attention is mainly devoted to highly symmetric random matrix ensembles, and matrices from these ensembles tend to have very predictable properties [42, 55]. As an example, any large square matrix with independent, identically distributed complex entries of mean 0 and variance 1 will have eigenvalues approximately uniformly distributed within the unit disc in the complex plane [123]. Thus, a typical deterministic matrix of interest in scientific or numerical applications, whose eigenvalues carry important information about an underlying physical system, may not be accurately modeled by a random matrix from one of these ensembles.

Nevertheless, as we will see in this thesis, the tools used in random matrix theory can still be used to shed light on numerical applications. The popularity of exploiting randomness in numerical linear algebra has grown considerably in recent years; see [91] for a comprehensive survey. In this dissertation, we will focus on two topics: (1) approximate diagonalization of non-Hermitian matrices, and (2) approximation of extreme eigenvalues of Hermitian matrices via the Lanczos algorithm. In each of these areas, we will use previously unexploited techniques from random matrix theory and high-dimensional probability to substantially improve upon existing results.

A detailed overview of mathematical preliminaries will be given in Section 1.1, but first let us go over the most basic aspects of non-Hermitian and non-normal matrices. This will allow us to better motivate some of our main results.

Recall that a matrix M is called *Hermitian* if it is equal to its conjugate transpose M^*,

and *normal* if $MM^* = M^*M$. The eigenvectors of a normal matrix are orthogonal. However, this does not hold for non-normal matrices, and this leads to increased instability of the eigenvalues under small perturbations. We review one standard example: consider the $n \times n$ "Jordan block" matrix J_n, which contains 1s on the superdiagonal and 0s in every other entry. The spectrum of this matrix is simply $\{0\}$. However, upon adding ε to the lower-left entry, the characteristic polynomial becomes $\lambda^n - \varepsilon$, so the eigenvalues are evenly spaced on the circle of radius $\varepsilon^{1/n}$. Thus, perturbations of J_n that are exponentially small in n can still move its eigenvalues by a macroscopic amount. This stands in stark contrast to the case of normal matrices, whose eigenvalues are 1-Lipschitz with respect to the operator norm.

With regards to the stability of eigenvalues, there is a range of behavior between normal matrices and nondiagonalizable matrices such as the Jordan block. One metric that is useful for quantifying this behavior is the *eigenvector condition number* of a matrix, which ranges from 1 in the case of a normal matrix to ∞ in the case of a nondiagonalizable matrix. The definition appears in Section 1.1, along with several consequences.

We now give an outline of the thesis and a summary of our contributions. Theorem statements and detailed accounts of related work can be found in the chapters indicated below.

In Chapter 2, we show in a precise quantitative way that every matrix is close to a matrix that has a small eigenvector condition number. Our result confirms a 2007 conjecture of Davies [36], and leads to a numerically stable way of computing analytic functions of a matrix. We prove this using the probabilistic method: we add small independent complex Gaussian random variables to each entry of the matrix, and then show that the resulting matrix has a small eigenvector condition number on average. Several ingredients of our proof come from random matrix theory. These include lower tail bounds for the least singular value of Gaussian random matrices first computed by Edelman [52], as well as a coupling lemma of Śniady [116] originally used to relate random matrices to a concept called *Brown measure* used in operator theory and free probability. The novelty of our approach lies in the way we connect these results to the eigenvector condition number by bounding the expected *area* of the pseudospectrum. This connection has not been exploited before in numerical linear algebra to the best of our knowledge.

In Chapter 3, we revisit the idea of regularization by a complex Gaussian perturbation, and use it to make substantial progress on the algorithmic problem of diagonalizing an arbitrary $n \times n$ non-Hermitian matrix in a numerically stable way on a machine with finite-precision arithmetic. We exhibit an algorithm running in nearly matrix-multiplication time, while the previously best known provable algorithm runs in time $O(n^9)$ [4]. Our algorithm is based on a variant of the well-known spectral bisection algorithm. Each step of this algorithm requires the computation of the so-called matrix sign function, which is approximated by repeating the Newton iteration $A \mapsto \frac{A+A^{-1}}{2}$. We show that the adding a Gaussian perturbation at the beginning of the algorithm regularizes the eigenvector condition number and the minimum eigenvalue gap of the input matrix. By analyzing the pseudospectrum and bounding the resolvent using the holomorphic functional calculus (see Section 1.1), we can then show that

these regularity properties are maintained throughout the entire Newton iteration—thus ensuring that the eigenvalues remain stable, and that the accumulated error from roundoff can be controlled. The tools from random matrix theory used in Chapter 2 make a reappearance here, alongside lower bounds on the *second-smallest* singular value, which are required to bound the minimum eigenvalue gap. The resolvent bounds rely on a new property which we call *pseudospectral shattering*, which ensures that eigenvalues stay far from relevant contours even in the presence of roundoff, and which is achieved by adding a Gaussian perturbation.

In Chapter 4 we analyze a different iterative algorithm, the Lanczos algorithm, from a rather different perspective. This algorithm is used for the related problem of approximating the extreme eigenvalues of high-dimensional Hermitian matrices, and is well-studied as it is used often in practice. Upper bounds are known on the number of iterations required to obtain a satisfactory approximation of the outlying eigenvalues, but below this threshold, the behavior of the algorithm is less understood. It is a randomized algorithm, taking as input a deterministic matrix and a single uniform random vector u from the unit sphere. As such, tools from high-dimensional probability can be brought to bear. We show that the random output of the Lanczos algorithm when run for few iterations is in fact tightly concentrated and can still be used to produce a useful approximation to the bulk distribution of the spectrum. To prove our concentration result, we first prove that the output of the Lanczos algorithm is locally Lipschitz in the input unit vector u. We then bring in a tool from high-dimensional probability known as the local Lévy lemma, which says that functions Lipschitz on a large region of the unit sphere are tightly concentrated about their medians. These techniques do not appear to have been used in previous work on the Lanczos algorithm.

1.0.1 Bibliographic Note

The contents of this thesis are the result of joint works in various stages of publication. Chapter 2 is based on joint work [11] with Jess Banks, Satyaki Mukherjee, and Nikhil Srivastava. Chapter 3 is based on joint work [13] with Jess Banks, Jorge Garza Vargas, and Nikhil Srivastava. Chapter 4 is based on joint work [127] with Jorge Garza Vargas. Parts of the introductory material are adapted from these works as well.

1.1 Preliminaries

In this section, we review the basic theory and definitions that will be used in the forthcoming chapters. All of it should be accessible to anyone with some background in linear algebra and complex analysis.

1.1.1 Eigenvalue perturbation theory

Suppose a matrix $M \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. One may then form the spectral decomposition

$$A = \sum_{i=1}^{n} \lambda_i v_i w_i^* = V D V^{-1},$$

where the w_i^* and the v_i are respectively the left and right eigenvectors of A, normalized so that $w_i^* v_i = 1$ for all i. In the case where A is normal (that is, A commutes with its conjugate transpose A^*), we have $v_i = w_i$ for all i.

We will often be interested in the eigenvalues of a small perturbation $A + tE$ for some other matrix E. Informally, one has that the eigenvalues are differentiable in t, and that the derivative of λ_i is equal to $w_i^* E v_i$. Thus for $\|E\| = 1$, the magnitude of the derivative is at most $\|w_i\|\|v_i\|$. Similar results can be obtained for higher derivatives and for the derivatives of eigenvectors; see [65] for a survey of first-order perturbation theory for eigenvalues and eigenvectors.

Motivated by this, in the spirit of numerical analysis one may define a condition number for an eigenvalue as follows, measuring the sensitivity of an eigenvalue to matrix perturbations of small norm:

Definition 1.1.1. For M, λ_i, w_i^* and v_i as above, the *eigenvalue condition number of* λ_i is defined as

$$\kappa(\lambda_i) := \|v_i\|\|w_i\|.$$

Note that for normal matrices, all eigenvalue condition numbers are equal to 1.

A related notion of spectral stability is the condition number of the matrix of eigenvectors, defined as follows:

Definition 1.1.2. For a diagonalizable matrix M, the *eigenvector condition number of* M is defined as

$$\kappa_V(M) := \inf_{V : M = V D V^{-1}} \|V\|\|V^{-1}\|. \tag{1.1}$$

The eigenvector condition number ranges between 1 and ∞ when A is normal and nondiagonalizable respectively, where $\|\cdot\|$ denotes the operator norm. Matrices with small κ_V enjoy many of the desirable properties of normal matrices, such as stability of the spectrum under small perturbations (this is the content of the Bauer-Fike theorem [16]).

We record a lemma relating the eigenvector and eigenvalue condition numbers. For related results, including an extension of this lemma to the more general context of block diagonalization, see the thesis of Demmel [45, Equation 3.6].

Lemma 1.1.3. Let M be an $n \times n$ matrix with distinct eigenvalues, and let V be the matrix whose columns are the eigenvectors of M normalized to have unit norm. Then

$$\kappa_V(M) \leq \kappa(V) \leq \sqrt{n \sum_{i=1}^{n} \kappa(\lambda_i)^2}.$$

Proof. Note that the left eigenvectors w_i are the rows of V^{-1}. Then $\|V\|_F^2 = n$ and $\|V^{-1}\|_F^2 = \sum_{i=1}^n \|w_i\|^2 = \sum_{i=1}^n \kappa(\lambda_i)^2$, so

$$\kappa(V) = \|V\|\|V^{-1}\| \leq \|V\|_F \|V^{-1}\|_F = \sqrt{n \sum_{i=1}^n \kappa(\lambda_i)^2}.$$

□

Eigenvalue condition numbers and eigenvector condition numbers do not tell the whole story, however. A richer object is the so-called ε-pseudospectrum, defined as follows:

Definition 1.1.4. For any matrix $M \in \mathbb{C}^{n\times n}$ and any $\varepsilon > 0$, the *ε-pseudospectrum* of M is defined as follows:

$$\begin{aligned} \Lambda_\epsilon(M) &:= \{z \in \mathbb{C} : z \in \Lambda(M + E) \text{ for some } \|E\| < \epsilon\} && (1.2)\\ &= \left\{z \in \mathbb{C} : \|(zI - M)^{-1}\| > 1/\epsilon\right\} && (1.3)\\ &= \left\{z \in \mathbb{C} : \sigma_n(zI - M) < \epsilon\right\}, && (1.4) \end{aligned}$$

where $\lambda(X)$ denotes the spectrum of any matrix X and $\sigma_n(X)$ denotes the least singular value of X. In other words, the ε-pseudospectrum is a level set of the norm of the resolvent matrix $(z - M)^{-1}$.

For a proof of the equality of these three sets and a comprehensive treatment of pseudospectra, see the beautiful book of Trefethen and Embree [125]. Note that for a normal matrix, we have

$$\Lambda_\epsilon(M) = \Lambda(M) + \bigcup_{i=1}^n D(\lambda_i, \epsilon),$$

whereas for a nonnormal matrix such as a Jordan block, Λ_ϵ can be much larger.

Note also that the eigenvector condition number and pseudospectrum are related as follows:

Lemma 1.1.5 ([125])**.** Let $D(z, r)$ denote the open disk of radius r centered at $z \in \mathbb{C}$. For every $M \in \mathbb{C}^{n\times n}$,

$$\bigcup_i D(\lambda_i, \varepsilon) \subset \Lambda_\varepsilon(X) \subset \bigcup_i D(\lambda_i, \varepsilon\kappa_V(M)). \tag{1.5}$$

We can relate the pseudospectra of a matrix and of a perturbation:

Proposition 1.1.6 ([125], Theorem 52.4)**.** For any $n \times n$ matrices M and E and any $\varepsilon > 0$, $\Lambda_{\varepsilon - \|E\|}(M) \subseteq \Lambda_\varepsilon(M + E)$.

It is also immediate that $\Lambda(M) \subset \Lambda_\epsilon(M)$, and in fact a stronger relationship holds as well:

Proposition 1.1.7 ([125], Theorem 4.3)**.** For any $n \times n$ matrix M, any bounded connected component of $\Lambda_\varepsilon(M)$ must contain an eigenvalue of M.

1.1.2 Random Matrix Theory

For an $n \times n$ matrix A with eigenvalues λ_i, we say that the *empirical spectral distribution* of A is the atomic probability measure $\frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i}$, where δ_x denotes the Dirac mass at x.

The following definition describes the type of matrix perturbation we will use in Chapters 2 and 3:

Definition 1.1.8. A *complex Ginibre matrix* is an $n \times n$ random matrix $G_n = (g_{ij})$ with i.i.d complex Gaussian entries $g_{ij} \sim N(0, 1_{\mathbb{C}}/n)$, by which we mean $\mathbb{E}g_{ij} = 0$ and $\mathbb{E}|g_{ij}|^2 = 1/n$. Equivalently, the real and imaginary parts of each g_{ij} are independent $N(0, 1/2n)$ random variables.

With this normalization in n, the empirical spectral distribution of G_n converges weakly almost surely to "circular law," the uniform measure on the unit disc in the complex plane, as $n \to \infty$. The same convergence was in fact proven to hold for any i.i.d. complex matrices with entries of zero mean and variance in [123], the culmination of a long line of work by many authors.

1.1.3 Functional Analysis

Let $M \in \mathbb{C}^{n\times n}$, with eigenvalues $\lambda_1, ..., \lambda_n$. We say that a matrix P is a *spectral projector* for M if $MP = PM$ and $P^2 = P$. For instance, each of the terms $v_i w_i^*$ appearing in the spectral expansion (3.7) is a spectral projector, as $Av_i w_i^* = \lambda_i v_i w_i^* = v_i w_i^* A$ and $w_i^* v_i = 1$. If Γ_i is a simple closed positively oriented rectifiable curve in the complex plane separating λ_i from the rest of the spectrum, then it is well-known that

$$v_i w_i^* = \frac{1}{2\pi i} \oint_{\Gamma_i} (z - M)^{-1} \mathrm{d}z,$$

by taking the Jordan normal form of the the *resolvent* $(z - M)^{-1}$ and applying Cauchy's integral formula.

Since every spectral projector P commutes with M, its range agrees exactly with an invariant subspace of M. We will often find it useful to choose some region of the complex plane bounded by a simple closed positively oriented rectifiable curve Γ, and compute the spectral projector onto the invariant subspace spanned by those eigenvectors whose eigenvalues lie inside Γ. Such a projector can be computed by a contour integral analogous to the above.

Recall that if f is any function, and M is diagonalizable, then we can meaningfully define $f(M) := Vf(D)V^{-1}$, where $f(D)$ is simply the result of applying f to each element of the diagonal matrix D. The *holomorphic functional calculus* gives an equivalent definition that extends to the case when M is non-diagonalizable and also applies to infinite-dimensional operators. As we will see in Chapter 3, it has the added benefit that bounds on the norm of the resolvent of M can be converted into bounds on the norm of $f(M)$.

Proposition 1.1.9 (Holomorphic Functional Calculus)**.** Let A be any matrix, $D \supset \Lambda(M)$ be an open neighborhood of its spectrum (not necessarily connected), and $\Gamma_1, ..., \Gamma_k$ be simple closed positively oriented rectifiable curves whose interiors together contain all of $\Lambda(M)$. Then if f is holomorphic on D, the definition

$$f(M) := \frac{1}{2\pi i} \sum_{j=1}^{k} \oint_{\Gamma_j} f(z)(z - M)^{-1} \mathrm{d}z$$

is an *algebra homomorphism* in the sense that $(fg)(M) = f(M)g(M)$ for any f and g holomorphic on D.

Finally, the *resolvent identity*

$$(z - M)^{-1} - (z - M')^{-1} = (z - M)^{-1}(M - M')(z - M')^{-1}$$

will frequently come in handy to analyze perturbations of contour integrals.

1.1.4 Orthogonal polynomials

Orthogonal polynomials have a rich connection to random matrix theory; see [41] for a survey. Here, we summarize some of the basic theory of orthogonal polynomials, with a view towards their application in the analysis of the Lanczos algorithm.

For now, let μ be a finite Borel measure on $\mathbb{R}$ and assume that its support, which we denote as $\mathrm{supp}(\mu)$, is compact and has infinitely many points. The set of square integrable functions $L^2(\mathbb{R}, d\mu)$ becomes a Hilbert space when endowed with the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x)g(x)d\mu(x).$$

The hypothesis that $|\mathrm{supp}(\mu)| = \infty$ implies that the monomials $\{1, x, x^2, \dots\}$ are linearly independent in $L^2(\mathbb{R}, d\mu)$. Hence, we can use the Gram-Schmidt procedure to obtain an infinite sequence of polynomials $p_k(x)$ with $\deg(p_k(x)) = k$ and

$$\int p_k(x)p_l(x)d\mu(x) = \delta_{kl}.$$

The leading coefficient of $p_k(x)$ is a quantity of interest in this chapter and will be denoted by γ_k. We will denote the monic orthogonal polynomials by $\pi_k(x)$. That is, $\pi_k(x) = \gamma_k^{-1} p_k(x)$ and clearly

$$\gamma_k = \left(\int_{\mathbb{R}} \pi_k^2(x) d\mu(x) \right)^{-\frac{1}{2}}. \tag{1.6}$$

Since $\pi_k(x)$ is orthogonal to all polynomials with degree less than k, the polynomial $x^k - \pi_k(x)$ is the orthogonal projection of x^k onto the span of $\{1, \dots, x^{k-1}\}$. Hence,

$$\int_{\mathbb{R}} \pi_k^2(x) d\mu(x) = \min_{q \in \Gamma_k} \int_{\mathbb{R}} q^2(x) d\mu(x)$$

where Γ_k denotes the space of monic polynomials of degree k.

Favard's theorem ensures that there is a sequence of real numbers α_k and a sequence of positive real numbers β_k such that the following *three-term recurrence* holds:

$$\begin{aligned} xp_k(x) &= \beta_{k-1} p_{k-1}(x) + \alpha_k p_k(x) + \beta_k p_{k+1}(x), \quad k \geq 1 \\ \text{and} \quad xp_0(x) &= \alpha_0 p_0(x) + \beta_0 p_1(x), \quad k = 0. \end{aligned}$$

It is clear from the three-term recurrence that the following identity holds:

$$\gamma_k = \left(\prod_{i=0}^{k-1} \beta_i \right)^{-1}. \tag{1.7}$$

These so-called *Jacobi coefficients* α_k and β_k encode all the information of the measure μ. In fact, since the Stieltjes transform of μ has a continued fraction expansion in terms of its Jacobi coefficients, knowing the few first elements in these sequences allows one to approximate the measure. See Chapter 4.3 in [41] for an example.

We denote by J_k the $k \times k$ Jacobi matrix of μ; that is, J_k is the tridiagonal symmetric matrix with $(J_k)_{ii} = \alpha_{i-1}$ and $(J_k)_{i+1,i} = (J_k)_{i,i+1} = \beta_{i-1}$. It is a standard fact that $\pi_k(x) = \det(xI - J_k)$ and that in particular, the roots of $p_k(x)$ are exactly the eigenvalues of J_k, which are real since J_k is symmetric.

Another object of importance in this theory is the Hankel matrix of a measure. We will denote M_k the $(k+1) \times (k+1)$ Hankel matrix of μ, in other words, if m_i denotes the ith moment of μ then $(M_k)_{ij} = m_{i+j-2}$ for every $1 \leq i, j \leq k+1$. From the elementary theory it is known (see [41], Section 3.1) that if we define $D_k = \det M_k$ then

$$\beta_k = \frac{\sqrt{D_{k-1} D_{k+1}}}{D_k} \quad \text{and} \quad \gamma_k = \sqrt{\frac{D_{k-1}}{D_k}}, \quad k \geq 0, \tag{1.8}$$

where we define $D_{-1} = 1$. Note that the second identity in (1.8) implies

$$D_k = \prod_{i=0}^{k} \gamma_i^{-2}. \tag{1.9}$$

Moreover, if $\tilde{M}_k(x)$ denotes the matrix obtained by replacing the last row of M_k by the row $(1 \; x \; x^2 \; \cdots \; x^k)$, we have the following useful identity

$$p_k(x) = \frac{\det \tilde{M}_k(x)}{\sqrt{D_{k-1} D_k}}. \tag{1.10}$$

Note that in the case in which $\mathrm{supp}(\mu)$ has n points, for n a positive integer, the set of monomials $\{1, x, x^2, \dots\}$ is not linearly independent in $L^2(\mathbb{R}, d\mu)$. Moreover, the Gram-Schmidt procedure stops after n iterations, and hence it only makes sense to talk about the orthogonal polynomials $p_k(x)$ for $k \leq n-1$. However, sometimes it is convenient to define the nth monic orthogonal polynomial as the unique monic polynomial of degree n whose roots are the elements of $\mathrm{supp}(\mu)$. In this case, the facts mentioned previously still hold for $k \leq n$.

1.1.5 The Lanczos algorithm

The Lanczos algorithm is a randomized iterative algorithm that takes three inputs: an $n \times n$ Hermitian matrix A, a random vector u distributed uniformly in $\mathbb{S}^{n-1}$ and an integer $1 \leq k \leq n$. The output is a $k \times k$ symmetric tridiagonal matrix J_k whose diagonal entries will be denoted by α_i, for $i = 0, \dots, k-1$, and whose subdiagonal and superdiagonal entries will be denoted by β_i, for $i = 0, \dots, k-2$. The eigenvalues of J_k are called the Ritz values and we will usually denote them as $r_1 \geq \cdots \geq r_k$. The eigenvectors of J_k give rise to the Ritz vectors, which after a change of basis yield the approximations for the eigenvectors of A. Algorithm 1 below describes how the procedure generates the Jacobi coefficients α_i and β_i.

input: A, k, u
initialize: $v_0 = u$
for $j = 0, \dots, k-1$ **do**
  $W_j = \mathrm{span}\{v_0, \dots, v_j\}$
  $\alpha_j = \langle Av_j, v_j \rangle$
  $\beta_j = \|\mathrm{Proj}_{W_j^\perp}(Av_j)\|_2$
  if $\beta_j = 0$ **then**
    stop
  else
    $v_{j+1} = \frac{\mathrm{Proj}_{W_j^\perp}(Av_j)}{\|\mathrm{Proj}_{W_j^\perp}(Av_j)\|_2}$
  end if
end for
return J_k

Algorithm 1: The Lanczos algorithm

This algorithm has a natural interpretation in terms of orthogonal polynomials. To every $u \in \mathbb{S}^{n-1}$ we can associate a measure supported on the spectrum of A as follows. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of A and $u_1, \dots, u_n$ be the coordinates of u when writen in the eigenbasis of A. We define the probability measure

$$\mu^u = \sum_{i=1}^{n} u_i^2 \delta_{\lambda_i}. \tag{1.11}$$

In the language of functional analysis, μ^u is the spectral measure of the operator A induced by the vector state u; that is, $\langle f(A)u, u\rangle = \int f(x)\, d\mu^u(x)$ for all (say) polynomials f. Note that the expectation of the random measure μ^u is just the empirical spectral distribution of A, namely

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i}.$$

It is not hard to see that if $p_j(x)$ are the orthogonal polynomials with respect to μ^u then $v_j = p_j(A)u$. Hence, the coefficients α_j and β_j output by the Lanczos algorithm are the Jacobi coefficients of the measure μ^u, and the Ritz values after k iterations are the roots of $p_k(x)$.

As a last remark, observe that the output of Algorithm 1 scales linearly with A. Hence, to simplify notation, in some of the proofs in Chapter 4 we will start by assuming that $\|A\| = 1$.

Chapter 2

Davies' Conjecture

2.1 Background

In this chapter we study the following question posed by E. B. Davies in [37]:

> *How well can an arbitrary matrix be approximated by one with a small eigenvector condition number?*

Our main theorem is as follows.

Theorem 2.1.1. Suppose $A \in \mathbb{C}^{n\times n}$ and $\delta \in (0,1)$. Then there is a matrix $E \in \mathbb{C}^{n\times n}$ such that $\|E\| \leq \delta\|A\|$ and

$$\kappa_V(A+E) \leq 4n^{3/2}\left(1+\frac{1}{\delta}\right).$$

In other words, every matrix is at most inverse polynomially close to a matrix whose eigenvectors have condition number at most polynomial in the dimension. The previously best known general bound in such a result was [37, Theorem 3.8]:

$$\kappa_V(A+E) \leq \left(\frac{n}{\delta}\right)^{(n-1)/2}, \tag{2.1}$$

so our theorem constitutes an exponential improvement in the dependence on both δ and n. We show in Proposition 2.4.1 that the $1/\delta$-dependence in Theorem 2.1.1 cannot be improved beyond $1/\delta^{1-1/n}$, so our bound is essentially optimal in δ for large n.

2.1.1 Davies' Conjecture

Theorem 2.1.1 implies a positive resolution to a conjecture of Davies [37].

Conjecture 2.1.2. For every positive integer n there is a constant c_n such that for every $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$ and $\epsilon \in (0,1)$:

$$\inf_{E\in\mathbb{C}^{n\times n}} (\kappa_V(A+E)\epsilon + \|E\|) \leq c_n\sqrt{\epsilon}. \tag{2.2}$$

Proof of Conjecture 2.1.2. Given $\epsilon > 0$, set $\delta = d_n\sqrt{\varepsilon}$ for some $d_n > 0$ and apply Theorem 2.1.1. This yields $c_n = 4n^{3/2} + 4n^{3/2}/d_n + d_n$. This is minimized at $d_n = 2n^{3/4}$, which yields $c_n = 4n^{3/2} + 4n^{3/4} \leq 8n^{3/2}$. □

The phrasing of Conjecture 2.1.2 is motivated by a particular application in numerical analysis. Suppose one wants to evaluate analytic functions $f(A)$ of a given matrix A, which may be nonnormal. If A is diagonalizable, one can use the formula $f(A) = Vf(D)V^{-1}$, where $f(D)$ means the function is applied to the scalar diagonal entries of D. However, this may be numerically infeasible if $\kappa_V(A)$ is very large: if all computations are carried to precision ϵ, the result may be off by an error of $\kappa_V(A)\epsilon$. Davies' idea was to replace A by a perturbation $A + E$ with a much smaller $\kappa_V(A + E)$, and compute $f(A + E)$ instead. In [37, Theorem 2.4], he showed that the net error incurred by this scheme for a given $\epsilon > 0$ and sufficiently regular f is controlled by:

$$\kappa_V(A + E)\epsilon + \|E\|,$$

which is the quantity appearing in (2.2). The key desirable feature of (2.2) is the dimension-independent fractional power of ϵ on the right-hand side, which shows that the total error scales slowly.

Davies proved his conjecture in the special case of upper triangular Toeplitz matrices, in dimension $n = 3$ with the constant $c_n = 2$, as well as in the general case with the weaker dimension-dependent and nonconstructive bound $(n + 1)\epsilon^{2/(n+1)}$. This last result corresponds to (2.1) above. He also speculated that a *random* regularizing perturbation E suffices to prove Conjecture 2.1.2, and presented empirical evidence to that effect. Our proof of Theorem 2.1.1 below indeed follows this strategy.

2.1.2 Gaussian Regularization

Theorem 2.1.1 follows from a probabilistic result concerning complex Gaussian perturbations of a given matrix A.

We show that adding a small Ginibre perturbation regularizes the eigenvalue condition numbers of any matrix in the following averaged sense.

Theorem 2.1.3. Suppose $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$ and $\delta \in (0, 1)$. Let G_n be a complex Ginibre matrix, and let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be the (random) eigenvalues of $A + \delta G_n$. Then for every measurable open set $B \subset \mathbb{C}$,

$$\mathbb{E}\sum_{\lambda_i \in B} \kappa(\lambda_i)^2 \leq \frac{n^2}{\pi\delta^2}\mathrm{Leb}(B).$$

Note that the $\kappa(\lambda_i)$ appearing above are well-defined because $A+\delta G_n$ has distinct eigenvalues with probability one.

2.1.3 Related Work

Random Matrix Theory. There have been numerous studies of the eigenvalue condition numbers $\kappa(\lambda_i)^2$, sometimes called eigenvector *overlaps* in the random matrix theory and mathematical physics literature, for non-Hermitian random matrix models of type $A + \delta G_n$. In the centered case $A = 0$ and $\delta = 1$ of a standard complex Ginibre matrix, the seminal work of Chalker and Mehlig [32] calculated the large-n limit of the conditional expectations

$$\mathbb{E}[\kappa(\lambda)^2 | \lambda = z] \underset{n\to\infty}{\sim} n(1 - |z|^2),$$

whenever $|z| < 1$. Recent works by Bourgade and Dubach [24] and Fyodorov [58] improved on this substantially by giving exact nonasymptotic formulas for the distribution of $\kappa(\lambda)^2$ conditional on the location of the eigenvalue λ, as well as concise descriptions of the scaling limits for these formulas. The paper [22] proved (in the more general setup of invariant ensembles) that the angles between the right eigenvectors $(v_i^* v_j)/\|v_i\|\|v_j\|$ have subgaussian tails, which has some bearing on κ_V (for instance, a small angle between unit eigenvectors causes $\|V^{-1}\|$ and therefore κ_V to blow up.)

In the non-centered case, Davies and Hager [38] showed that if A is a Jordan block and $\delta = n^{-\alpha}$ for some appropriate α, then almost all of the eigenvalues of $A + \delta G_n$ lie near a circle of radius $\delta^{1/n}$ with probability $1 - o_n(1)$. Basak, Paquette, and Zeitouni [15, 14] showed that for a sequence of banded Toeplitz matrices A_n with a finite symbol, the spectral measures of $A_n + n^{-\alpha} G_n$ converge weakly in probability, as $n \to \infty$, to a predictable density determined by the symbol. Both of the above results were recently and substantially improved by Sjöstrand and Vogel [112, 113] who proved that for any Toeplitz A, almost all of the eigenvalues of $A + n^{-\alpha} G_n$ are close to the symbol curve of A with exponentially good probability in n. Note that none of the results mentioned in this paragraph explicitly discuss the $\kappa(\lambda_i)$; however, they do deal qualitatively with related phenomena surrounding spectral instability of non-Hermitian matrices.

The idea of managing spectral instability by adding a random perturbation can be traced back to the influential papers of Haagerup and Larsen [69] and Śniady [116] (see also [68, 56]), who used it to study convergence of the eigenvalues of certain non-Hermitian random matrices to a limiting Brown measure, in the context of free probability theory.

There are three notable differences between Theorem 2.1.3 and the results mentioned above:

1. Our result is much coarser, and only guarantees an upper bound on the $\mathbb{E}\kappa(\lambda_i)^2$, rather than a precise description of any distribution, limiting or not.

2. It applies to any $A \in \mathbb{C}^{n\times n}$ and $\delta \in (0, 1)$.

3. It is completely nonasymptotic and does not require $n \to \infty$ or even sufficiently large n.

Numerical Analysis. In the numerical linear algebra literature, several works have analyzed the condition numbers of Gaussian matrices (notably the seminal results of Demmel [45] and

Edelman [52]) as well as perturbations of arbitrary matrices by Gaussian matrices (beginning with [109]) in the nonasymptotic regime. In contrast, we study the condition numbers of the *eigenvectors* of such matrices, rather than of the matrices themselves.

The idea of approximating matrix functions by adding a regularizing perturbation was introduced in [37] and has since appeared in several works regarding numerical computation of the matrix logarithm, sine, cosine, and related functions [94, 73, 95, 97, 40].

2.1.4 Techniques and Organization

We first collect some tools from random matrix theory in Section 2.2, along the way proving a conjecture of Sankar, Spielman, and Teng [109] regarding the optimal constant in their smoothed analysis of condition numbers of matrices under *real* Gaussian perturbations in Section 2.2.3. Section 2.3 contains the proofs of our main results, Theorems 2.1.1 and 2.1.3. In Section 2.4, we prove optimality of the $1/\delta$-dependence in Theorem 2.1.1 as discussed above, and show that Theorem 2.1.3 is sharp up to a small constant factor. We conclude with a discussion of some open problems in Section 2.5.

2.2 Tools from Random Matrix Theory

2.2.1 Nonasymptotic Extreme Singular Value Estimates

Let us record some standard non-asymptotic estimates for the extreme singular values of complex Ginibre matrices. The lower tail behavior of the smallest singular value of a Ginibre matrix was worked out by Edelman [52, Chapter 5], and with our normalization it translates to:

Theorem 2.2.1. For a complex Ginibre matrix G_n,

$$\mathbb{P}[\sigma_n(G_n) < \varepsilon] = 1 - e^{-\varepsilon^2 n^2} \leq \varepsilon^2 n^2.$$

We will also require a cruder tail estimate on the largest singular value. We believe the lemma holds with a constant 2 instead of $2\sqrt{2}$, but did not find a reference to a nonasymptotic result to this effect; since the difference is not very consequential in this context, we reduce to the real case.

Lemma 2.2.2. For a complex Ginibre matrix G_n,

$$\mathbb{P}[\sigma_1(G_n) > 2\sqrt{2} + t] \leq 2\exp(-nt^2).$$

Proof. We can write $G_n = \frac{1}{\sqrt{2}}(X + iY)$ where X and Y are independent with i.i.d. *real* $N(0, 1/n)$ entries. It is well-known (e.g. [35, Theorem II.11]) that:

$$\mathbb{E}\sigma_1(G_n) \leq \frac{2}{\sqrt{2}}\mathbb{E}\|X\| \leq 2\sqrt{2}.$$

Lipschitz concentration of functions of real Gaussian random variables yields the result. □

2.2.2 Śniady's Comparison Theorem

To bound the least singular value of noncentered Gaussian matrices, we will lean on a remarkable theorem of Śniady [116].

Theorem 2.2.3 (Śniady)**.** Let A_1 and A_2 be $n \times n$ complex matrices such that $\sigma_i(A_1) \leq \sigma_i(A_2)$ for all $1 \leq i \leq n$. Assume further that $\sigma_i(A_1) \neq \sigma_j(A_1)$ and $\sigma_i(A_2) \neq \sigma_j(A_2)$ for all $i \neq j$. Then for every $t \geq 0$, there exists a joint distribution on pairs of $n \times n$ complex matrices (G_1, G_2) such that

1. the marginals G_1 and G_2 are distributed as (normalized) complex Ginibre matrices G_n, and
2. almost surely $\sigma_i(A_1 + \sqrt{t}G_1) \leq \sigma_i(A_2 + \sqrt{t}G_2)$ for every i.

We will briefly sketch the proof of this theorem for the reader's benefit, since it is quite beautiful and we will need to perform a slight modification to prove the conjecture of Sankar-Spielman-Teng in the next subsection.

Sketch of proof. The key insight of the proof is that it is possible to couple the distributions of G_1 and G_2 through their singular values. To do so, one first derives a stochastic differential equation satisfied by the singular values $s_1, ..., s_n$ of a matrix Brownian motion (i.e., a matrix whose entries are independent complex Brownian motions):

$$ds_i = \frac{1}{\sqrt{2n}} dB_i + \frac{dt}{2s_i}\left(1 - \frac{1}{2n} + \sum_{j \neq i} \frac{s_i^2 + s_j^2}{n(s_i^2 - s_j^2)}\right), \tag{2.3}$$

where the B_i are independent standard real Brownian motions. Next, one uses a single n-tuple of real Brownian motions $B_1, ..., B_n$ to drive two processes $(s_1^{(1)}, \ldots, s_n^{(1)})$ and $(s_1^{(2)}, \ldots, s_n^{(2)})$ according to (2.3), with initial conditions $s_i^{(1)}(0) = \sigma_i(A_1)$ and $s_i^{(2)}(0) = \sigma_i(A_2)$ for all i. (To do this rigorously, one needs existence and uniqueness of strong solutions to the above SDE; this is shown in [79] under the hypothesis $s_i(0) \neq s_j(0)$ for all $i \neq j$.)

Things have been arranged so that the joint distribution of $(s_1^{(j)}, \ldots, s_n^{(j)})$ at time t matches the joint distribution of the singular values of $A_j + \sqrt{t}G_j$ for each $j = 1, 2$. One can then sample unitaries U_j and V_j from the distribution arising from the singular value decomposition $A_j + \sqrt{t}G_j = U_j D_j V_j^*$, conditioned on $D_j = \mathrm{diag}(s_1^{(j)}, \ldots, s_n^{(j)})$. Thus each G_j is separately Ginibre-distributed. However, $A_1 + \sqrt{t}G_1$ and $A_2 + \sqrt{t}G_2$ are coupled through the shared randomness driving the evolution of their singular values. In particular, since the same B_i were used for both processes, from (2.3) one can verify that the n differences $s_i^{(2)} - s_i^{(1)}$ are C^1 in t. By taking derivatives, one can then show the desired monotonicity property: if $s_i^{(2)} - s_i^{(1)} \geq 0$ holds for all i at $t = 0$, it must hold for all $t \geq 0$. □

2.2.3 Sankar-Spielman-Teng Conjecture

The proof technique of Śniady can be adapted to prove a counterpart of Theorem 3.3.2 for *real* Ginibre perturbations (by this we mean matrices with i.i.d. real $N(0, 1/n)$ entries). Because a rigorous proof requires some stochastic analysis, we defer the proof and discussion of the following theorem to Appendix A.1.

Theorem 2.2.4. Let A_1 and A_2 be $n \times n$ real matrices such that $\sigma_i(A_1) \leq \sigma_i(A_2)$ for all $1 \leq i \leq n$. Assume further that $\sigma_i(A_1) \neq \sigma_j(A_1)$ and $\sigma_i(A_2) \neq \sigma_j(A_2)$ for all $i \neq j$. Then for every $t \geq 0$, there exists a joint distribution on pairs of real $n \times n$ matrices (G_1, G_2) such that

1. the marginals G_1 and G_2 are distributed as real Ginibre matrices (with i.i.d. $N(0, 1/n)$ entries), and
2. almost surely $\sigma_i(A_1 + \sqrt{t}G_1) \leq \sigma_i(A_2 + \sqrt{t}G_2)$ for every i.

This resolves Conjecture 1 in [109], which we restate below as a proposition:

Proposition 2.2.5. Let G be an $n \times n$ matrix with i.i.d. real $N(0, 1)$ entries, and A be any $n \times n$ matrix with real entries. Then

$$\mathbb{P}[\sigma_n(A + G) < \varepsilon] \leq \varepsilon\sqrt{n}.$$

Proof. The case $A = 0$ is a result of Edelman [52]. The proposition for general A would then follow from Theorem 2.2.4 with $A_1 = 0$ and $A_2 = A$ if not for the hypothesis $\sigma_i(A_1) \neq \sigma_j(A_1)$ and $\sigma_i(A_2) \neq \sigma_j(A_2)$ for all $i \neq j$. So we approach 0 and A by matrices satisfying this hypothesis, apply Theorem 2.2.4, and take limits, using the continuous mapping theorem and continuity of $\sigma_n(\cdot)$. □

2.3 Proof of Theorems 2.1.1 and 2.1.3

Proof of Theorem 2.1.1 given Theorem 2.1.3. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the random matrix $A + \delta G_n$, and $t > 2\sqrt{2}$ and $s > 1$ be parameters to be optimized later. Davies' original bound (2.1) implies our bound for $n \leq 3$, so assume $n \geq 4$. Then Lemma 2.2.2 tells us that

$$\mathbb{P}[\|\delta G_n\| \geq t\delta] \leq 2e^{-4(t-2\sqrt{2})^2}. \tag{2.4}$$

Letting $B = D(0, \|A\| + t\delta)$, we have

$$\mathbb{P}\left[\sum_{\lambda_i \in B} \kappa(\lambda_i)^2 \neq \sum_{i \leq n} \kappa(\lambda_i)^2\right] \leq \mathbb{P}[\|\delta G_n\| \geq t\delta] \leq 2e^{-4(t-2\sqrt{2})^2} \tag{2.5}$$

since $\max_{i \leq n} |\lambda_i| \leq \|A\| + \|\delta G_n\|$. On the other hand, by Theorem 2.1.3 applied to B and Markov's inequality:

$$\mathbb{P}\left[\sum_{\lambda_i \in B} \kappa(\lambda_i)^2 \geq s\frac{n^2 \mathrm{Leb}(B)}{\delta^2 \pi}\right] \leq \frac{1}{s}. \tag{2.6}$$

By the union bound, if we choose s and t such that

$$2e^{-4(t-2\sqrt{2})^2} + \frac{1}{s} < 1 \tag{2.7}$$

then there exists a choice of G_n such that neither of the events (2.5), (2.6) occurs. Letting $E = \delta G_n$ for this choice, we have

$$\sum_{i=1}^{n} \kappa(\lambda_i)^2 = \sum_{i \in B} \kappa(\lambda_i)^2 \leq s \frac{n^2 \mathrm{Leb}(B)}{\pi \delta^2}.$$

Taking a square root and applying Lemma 1.1.3, we have

$$\kappa_V(A+E) \leq \frac{\sqrt{s} n^{3/2}}{\delta}(\|A\| + t\delta) \leq \frac{\sqrt{s} n^{3/2}\|A\|}{\delta} + t\sqrt{s} n^{3/2}.$$

Because $\|E\| \leq t\delta$ and not δ, replacing δ by δ/t yields the bound

$$\kappa_V(A+E) \leq \frac{t\sqrt{s} n^{3/2}\|A\|}{\delta} + t\sqrt{s} n^{3/2}.$$

To get the best bound, we must minimize $t\sqrt{s}$ subject to the constraints (2.7), $t > 2\sqrt{2}$ and $s > 1$. Solving for s this becomes a univariate optimization problem, and one can check numerically that the optimum is achieved at $t \approx 3.7487$ and $t\sqrt{s} \approx 3.8822 < 4$, as advertised. □

We begin the proof of Theorem 2.1.3 by relating the eigenvalue condition numbers of a matrix to the rate at which its pseudospectrum Λ_ϵ shrinks as a function of the parameter $\epsilon > 0$. The following proposition is not new; the proof essentially appears for example in Section 3.6 of [24], but we include it for completeness since it is critical to our argument.

Lemma 2.3.1 (Limiting Area of the Pseudospectrum)**.** Let M be an $n \times n$ matrix with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and let $B \subset \mathbb{C}$ be an open set whose boundary contains none of the λ_i. Then

$$\lim_{\varepsilon \to 0} \frac{\mathrm{Leb}(\Lambda_\varepsilon(M) \cap B)}{\varepsilon^2} = \pi \sum_{\lambda_i \in B}^{n} \kappa(\lambda_i)^2.$$

Proof. Write the spectral decomposition

$$(zI - M)^{-1} = \sum_{i=1}^{n} \frac{v_i w_i^*}{z - \lambda_i},$$

where the v_i and w_i^* are right and left eigenvectors of M, respectively. Since the λ_i are distinct, we may choose $\epsilon_0 > 0$ sufficiently small to guarantee that there exists a constant

$C > 0$ satisfying (1) the disks $D(\lambda_i, \epsilon_0)$ are disjoint; (2) for every $\lambda_i \in B$ the disk $D(\lambda_i, \epsilon_0)$ is contained in B; and (3) whenever $z \in D(\lambda_i, \epsilon_0)$ for some i,

$$\|(zI - M)^{-1}\| \geq \frac{\|v_i w_i^*\|}{|z - \lambda_i|} - C = \frac{\kappa(\lambda_i)}{|z - \lambda_i|} - C. \tag{2.8}$$

Using the definition of the ε-pseudospectrum in (1.3), we rearrange (2.8) to obtain

$$\Lambda_\varepsilon(M) \cap B \supset \left\{ z : |z - \lambda_i| \leq \min \left\{ \epsilon_0, \frac{\kappa(\lambda_i)\varepsilon}{1 + \varepsilon C} \right\}, \text{ for some } \lambda_i \in B \right\}.$$

Thus, taking ε small enough, we have

$$\liminf_{\varepsilon \to 0} \frac{\text{Leb}(\Lambda_\varepsilon(M) \cap B)}{\varepsilon^2} \geq \pi \sum_{\lambda_i \in B} \kappa(\lambda_i)^2.$$

For the opposite inequality, Theorem 52.1 of [125] states that the ε-pseudospectrum is contained in disks around the eigenvalues λ_i of radii $\varepsilon\kappa(\lambda_i) + O(\varepsilon^2)$. Choosing ϵ small enough so that for $\lambda_i \in B$ these disks are entirely contained in B:

$$\text{Leb}(\Lambda_\varepsilon \cap B) \leq \sum_{\lambda_i \in B} \pi(\varepsilon\kappa(\lambda_i) + O(\varepsilon^2))^2 \Rightarrow \limsup_{\varepsilon \to 0} \frac{\text{Leb}(\Lambda_\varepsilon \cap B)}{\varepsilon^2} \leq \sum_{\lambda_i \in B} \pi\kappa(\lambda_i)^2.$$

□

Next, we show that for fixed $\epsilon > 0$, any particular point $z \in \mathbb{C}$ is unlikely to be in $\Lambda_\epsilon(A + \delta G_n)$. This is based on the following singular value estimate, which generalizes Theorem 2.2.1.

Lemma 2.3.2 (Small Ball Estimate for σ_n)**.** Let M be an $n \times n$ matrix with complex entries, and G be drawn from the Ginibre ensemble. Then for all $\delta > 0$ and $\epsilon > 0$

$$\mathbb{P}\left[\sigma_n(M + \delta G_n) < \epsilon\right] \leq n^2 \frac{\epsilon^2}{\delta^2}.$$

Proof. Repeat the proof of Proposition 2.2.5 using instead Theorems 2.2.1 and 3.3.2. □

Remark 2.3.3. If one is willing to lose a small constant factor in the bound, Lemma 2.3.2 has an elementary geometric proof (which avoids stochastic calculus), essentially identical to the proof of Sankar-Spielman-Teng [109, Theorem 3.1] in the case of real Ginibre perturbations. Note however that it is crucial to use a *complex* Gaussian for our purposes. A real Gaussian would yield a small ball estimate of order ϵ (see Proposition 2.2.5) rather than ϵ^2, which is not good enough to take the limit below.

Proof of Theorem 2.1.3. For every $z \in \mathbb{C}$ we have the upper bound

$$\mathbb{P}[z \in \Lambda_\epsilon(A + \delta G_n)] = \mathbb{P}[\sigma_n(zI - (A + \delta G_n)) < \epsilon] \leq n^2 \frac{\varepsilon^2}{\delta^2}, \tag{2.9}$$

by applying Lemma 2.3.2 to $M = zI - A$ and noting that G and $-G$ have the same distribution.

Fix a measurable open set $B \subset \mathbb{C}$. Then

$$\begin{aligned}
\mathbb{E}\,\mathrm{Leb}(\Lambda_\epsilon(A + \delta G_n) \cap B) &= \mathbb{E} \int_B \mathbb{1}_{\{z \in \Lambda_\epsilon(A+\delta G_n)\}}\, dz \\
&= \int_B \mathbb{E}\{z \in \Lambda_\epsilon(A + \delta G_n)\}\, dz && \text{by Fubini} \\
&\leq \int_B n^2 \frac{\varepsilon^2}{\delta^2}\, dz && \text{by (3.12)} \\
&= n^2 \frac{\varepsilon^2}{\delta^2} \mathrm{Leb}(B) && (2.10)
\end{aligned}$$

where the integrals are with respect to Lebesgue measure on $\mathbb{C}$. Finally, taking a limit as $\epsilon \to 0$ yields the desired bound:

$$\begin{aligned}
\mathbb{E} \sum_{\lambda_i \in B} \kappa(\lambda_i^2) &= \mathbb{E} \liminf_{\epsilon \to 0} \frac{\mathrm{Leb}(\Lambda_\epsilon(A + \delta G_n) \cap B)}{\pi \epsilon^2} && \text{by Lemma 2.3.1} \\
&\leq \liminf_{\epsilon \to 0} \mathbb{E} \frac{\mathrm{Leb}(\Lambda_\epsilon(A + \delta G_n) \cap B)}{\pi \epsilon^2} && \text{by Fatou's Lemma} \\
&\leq \frac{n^2 \mathrm{Leb}(B)}{\pi \delta^2} && \text{by (2.10).}
\end{aligned}$$

□

2.4 Optimality of the Bounds

We first show that Theorem 2.1.1 has essentially the optimal dependence on δ for n large. The example which requires this dependence is simply a Jordan block J, for which Davies [37] established the upper bound $\kappa_V(J + \delta E) \leq 2/\delta^{1-1/n}$, for some E with $||E|| < 1$.

Proposition 2.4.1. Fix $n > 0$ and let $J \in \mathbb{C}^{n \times n}$ be the upper triangular Jordan block with ones on the superdiagonal and zeros everywhere else. Then there exist $c_n > 0$ and $\delta_n > 0$ such that for all $E \in \mathbb{C}^{n \times n}$ with $\|E\| \leq 1$ and all $\delta < \delta_n$, we have

$$\kappa_V(J + \delta E) \geq \frac{c_n}{\delta^{1-1/n}}.$$

Proof. As a warm-up, we'll need the following bound on the pseudospectrum of J. Let λ be an eigenvalue of $J + \delta E$, with v its associated right eigenvector; then $(J + \delta E)^n v = \lambda^n v$ and, accordingly, $|\lambda|^n \leq \|(J + \delta E)^n\|$. Expanding, using nilpotence of J, $\|J\| = 1$, and submultiplicativity of the operator norm, we get

$$|\lambda|^n \leq \|(J + \delta E)^n\| \leq (1 + \delta)^n - 1 = O(\delta) \tag{2.11}$$

where the big-O refers to the limit $\delta \to 0$ (recall n is fixed).

Writing $J + \delta E = V^{-1} D V$, we want to lower bound the condition number of V. As above, let λ be an eigenvalue of $J + \delta E$, now writing w^* and v for its left and right eigenvectors. We'll use the lower bound

$$\kappa(V) = \|V^{-1}\|\|V\| \geq \frac{\|w^*\|\|v\|}{|w^* v|} = \kappa(\lambda).$$

Since the formula above is agnostic to the scaling of the left and right eigenvectors, we'll assume that both have unit length and show that $|w^* v|$ is small.

Let $0 \leq k \leq n$. Then $\|(J + \delta E)^k v\| = |\lambda|^k$, and analogously to (2.11),

$$\|(J + \delta E)^k - J^k\| \leq (1 + \delta)^k - 1 = O(\delta).$$

Since J acts on the left as a left shift,

$$\begin{aligned}
\left(\sum_{i=k+1}^{n} |v_i|^2 \right)^{1/2} &= \|J^k v\| \\
&\leq \|(J + \delta E)^k v\| + \|(J^k - (J + \delta E)^k) v\| \\
&\leq |\lambda|^k + O(\delta) \\
&= O(\delta^{k/n}),
\end{aligned}$$

where the final line follows from (2.11). Similarly,

$$\left(\sum_{i=1}^{n-k} |w_i|^2 \right)^{1/2} = \|w^* J^k\| = O(\delta^{k/n}).$$

Finally, we have $\kappa(V)^{-1} = |w^* v| \leq \sum_{j=1}^{n} |w_j||v_j|$, which in turn is at most

$$\sum_{j=1}^{n} \left(\sum_{i=1}^{j} |w_i|^2 \right)^{1/2} \left(\sum_{i=j}^{n} |v_i|^2 \right)^{1/2} = O(\delta^{(n-j)/n} \delta^{(j-1)/n}) = O(\delta^{1-1/n}).$$

□

We end by showing that the dependence on n in Theorem 2.1.3 cannot be improved.

Proposition 2.4.2. There exists $c > 0$ such that for all n,

$$\mathbb{E} \sum_{i \in [n]} \kappa^2 \left(\lambda_i(G_n)\right) \geq cn^2.$$

Proof. Bourgade and Dubach [24, Theorem 1.1, Equation 1.8] show that eigenvalue condition numbers in the bulk of the spectrum of complex Ginibre matrices are of order $\sqrt{n}$. Precisely,

$$\lim_{n \to \infty} \frac{\mathbb{E}[\kappa(\lambda_i)^2 | \lambda_i = z]}{n} = 1 - |z|^2$$

uniformly for (say) $z \in D(0, r)$ for any $r < 1$. The classical *circular law* for the limiting spectral distribution of Ginibre matrices ensures that

$$\lim_{n \to \infty} \frac{\mathbb{E}|\Lambda(G_n) \cap D(0, r)|}{n} = \frac{\mathrm{vol}(D(0, r))}{\mathrm{vol}(D(0, 1))} = r^2.$$

Thus,

$$\liminf_{n \to \infty} \frac{\mathbb{E} \sum_{i \in [n]} \kappa \left(\lambda_i(G_n)\right)^2}{n^2} \geq r^2(1 - r^2) > 0.$$

□

2.5 Conclusion and Discussion

A key theme in our work is the interplay between the related notions of eigenvector condition number κ_V, eigenvalue condition number $\kappa(\lambda_i)$ and pseudospectrum Λ_ε. Equally important is the fact that global objects such as κ_V and Λ_ε can be controlled by local quantities, specifically the least singular values of shifts $\sigma_n(zI - M)$ for each $z \in \mathbb{C}$. The proof also heavily exploits the left and right unitary invariance of the Ginibre ensemble (via Theorem 3.3.2, due to Śniady) as well as anticoncentration of the complex Gaussian.

One natural question is whether similar results hold if one replaces Gaussian perturbations with a different class of random perturbations G'. To apply the approach in this chapter, the key difficulty would be obtaining suitable bounds for the least singular value of $z - A - \delta G'$. Davies [37] presents experimental evidence that Theorem 2.1.1 holds for random real rank-one perturbations and random real Gaussian perturbations. In fact, recent work shows that the theorem does not hold for perturbations of bounded rank [12], but a real version of the theorem can be proved, albeit with a weaker dependence on δ; see [12, 76]. See Remark 2.3.3 for a discussion of why the present proof does not extend to the case of real Gaussian perturbations.

One may also ask if Theorem 2.1.1 can be derandomized; that is, if the regularizing perturbation E can be chosen by a deterministic algorithm given A as input. One natural

choice would be to perturb in the direction of the nearest normal matrix in either operator or Frobenius norm, the latter of which can be written as a certain optimization problem over unitary matrices [106].

Proposition 2.4.1 shows that the upper bound in Theorem 2.1.1 is tight in the perturbation size δ. Now, let c_n be the smallest constant such that Theorem 2.1.1 holds with an upper bound of c_n/δ. Theorem 2.1.1 implies that $c_n \leq 8n^{3/2}$, and since $\kappa_V = \|V\|\|V^{-1}\| \geq 1$ for any matrix, we have $c_n \geq 1$. It would be interesting to determine the correct asymptotic behavior of c_n. Davidson, Herrero, and Salinas asked in 1989 [34] whether the statement of Theorem 2.1.1 is possible with $\kappa_V(A+E)$ depending only on δ and not on n. In the present context, we can ask the more refined question: does Theorem 2.1.1 hold with bounded c_n, or must c_n go to infinity with n?

Chapter 3

Approximate Diagonalization

3.1 Background

In this chapter, we study the algorithmic problem of approximately finding all of the eigenvalues and eigenvectors of a given arbitrary $n \times n$ complex matrix. While this problem is quite well-understood in the special case of Hermitian matrices (see, e.g., [103]), the general non-Hermitian case has remained mysterious from a theoretical standpoint even after several decades of research. In particular, the currently best known *provable* algorithms for this problem run in time $O(n^9/\delta^2)$ [4] or $O(n^c \log(1/\delta))$ [30] with $c \geq 12$ where $\delta > 0$ is an error parameter, depending on the model of computation and notion of approximation considered.[1] To be sure, the non-Hermitian case is well-motivated: coupled systems of differential equations, linear dynamical systems in control theory, transfer operators in mathematical physics, and the nonbacktracking matrix in spectral graph theory are but a few situations where finding the eigenvalues *and eigenvectors* of a non-Hermitian matrix is important.

The key difficulties in dealing with non-normal matrices are the interrelated phenomena of *non-orthogonal eigenvectors* and *spectral instability*, the latter referring to extreme sensitivity of the eigenvalues and invariant subspaces to perturbations of the matrix. Non-orthogonality slows down convergence of standard algorithms such as the power method, and spectral instability can force the use of very high precision arithmetic, also leading to slower algorithms. Both phenomena together make it difficult to reduce the eigenproblem to a subproblem by "removing" an eigenvector or invariant subspace, since this can only be done approximately and one must control the spectral stability of the subproblem.

In this chapter, we overcome these difficulties by identifying and leveraging a phenomenon we refer to as *pseudospectral shattering*: adding a small complex Gaussian perturbation to any matrix yields a matrix with well-conditioned eigenvectors and a large minimum gap between the eigenvalues, implying spectral stability. This result builds on the recent solution of Davies' conjecture [11], and is of independent interest in random matrix theory, where minimum eigenvalue gap bounds in the non-Hermitian case were previously only known for

[1] A detailed discussion of these and other related results appears in Section 3.1.3.

i.i.d. models [110, 60].

We complement the above by proving that a variant of the well-known spectral bisection algorithm in numerical linear algebra [17] is both fast and numerically stable (i.e., can be implemented using a polylogarithmic number of bits of precision) when run on a pseudospectrally shattered matrix. The key step in the bisection algorithm is computing the *sign function* of a matrix, a problem of independent interest in many areas such including control theory and approximation theory [78]. Our main algorithmic contribution is a rigorous analysis of the well-known Newton iteration method [105] for computing the sign function *in finite arithmetic*, showing that it converges quickly and numerically stably on matrices for which the sign function is well-conditioned, in particular on pseudospectrally shattered ones.

The end result is an algorithm which reduces the general diagonalization problem to a polylogarithmic (in the desired accuracy and dimension n) number of invocations of standard numerical linear algebra routines (multiplication, inversion, and QR factorization), each of which is reducible to matrix multiplication [47], yielding a nearly matrix multiplication runtime for the whole algorithm. This improves on the previously best known running time of $O(n^3 + n^2 \log(1/\delta))$ arithmetic operations even in the Hermitian case [103].

We now proceed to give precise mathematical formulations of the eigenproblem and computational model, followed by statements of our results and a detailed discussion of related work.

3.1.1 Problem Statement

An *eigenpair* of a matrix $A \in \mathbb{C}^{n \times n}$ is a tuple $(\lambda, v) \in \mathbb{C} \times \mathbb{C}^n$ such that

$$Av = \lambda v,$$

and v is normalized to be a unit vector. The *eigenproblem* is the problem of finding a maximal set of linearly independent eigenpairs (λ_i, v_i) of a given matrix A; note that an eigenvalue may appear more than once if it has geometric multiplicity greater than one. In the case when A is diagonalizable, the solution consists of exactly n eigenpairs, and if A has distinct eigenvalues then the solution is unique, up to the phases of the v_i.

3.1.1.1 Accuracy and Conditioning

As mentioned in the introduction, due to the Abel-Ruffini theorem, it is impossible to have a finite-time algorithm which solves the eigenproblem exactly using arithmetic operations and radicals. Thus, all we can hope for is *approximate* eigenvalues and eigenvectors, up to a desired accuracy $\delta > 0$. There are two standard notions of approximation. We assume $\|A\| \leq 1$ for normalization.

Forward Approximation. Compute pairs (λ_i', v_i') such that

$$|\lambda_i - \lambda_i'| \leq \delta \quad \text{and} \quad \|v_i - v_i'\| \leq \delta$$

for the true eigenpairs (λ_i, v_i), i.e., find a solution close to the exact solution. This makes sense in contexts where the exact solution is meaningful; e.g. the matrix is of theoretical/mathematical origin, and unstable (in the entries) quantities such as eigenvalue multiplicity can have a significant meaning.

Backward Approximation. Compute (λ'_i, v'_i) which are the exact eigenpairs of a matrix A' satisfying

$$\|A' - A\| \leq \delta,$$

i.e., find the exact solution to a nearby problem. This is the appropriate and standard notion in scientific computing, where the matrix is of physical or empirical origin and is not assumed to be known exactly (and even if it were, roundoff error would destroy this exactness). Note that since diagonalizable matrices are dense in $\mathbb{C}^{n\times n}$, one can hope to always find a complete set of eigenpairs for some nearby $A' = VDV^{-1}$, yielding an *approximate diagonalization* of A:

$$\|A - VDV^{-1}\| \leq \delta. \tag{3.1}$$

Note that the eigenproblem in either of the above formulations is *not* easily reducible to the problem of computing eigenvalues, since they can only be computed approximately and it is not clear how to obtain approximate eigenvectors from approximate eigenvalues. We now introduce a condition number for the eigenproblem, which measures the sensitivity of the eigenpairs of a matrix to perturbations and allows us to relate its forward and backward approximate solutions.

Condition Numbers.

We define the *condition number of the eigenproblem* to be[2]:

$$\kappa_{\text{eig}}(A) := \frac{\kappa_V(A)}{\text{gap}(A)} \in [0, \infty]. \tag{3.2}$$

It follows from the following proposition (whose proof appears in the preliminaries of this chapter) that a δ-backward approximate solution of the eigenproblem is a $\kappa_{\text{eig}}(A)^2\delta$-forward approximate solution[3].

Proposition 3.1.1. If $\|A\|, \|A'\| \leq 1$, $\|A - A'\| \leq \delta$, and $\{(v_i, \lambda_i)\}_{i\leq n}$, $\{(v'_i, \lambda'_i)\}_{i\leq n}$ are eigenpairs of A, A' with distinct eigenvalues, and $\delta < \frac{\text{gap}(A)}{4\kappa_V(A)}$, then

$$\|v'_i - v_i\| \leq 4\kappa_{\text{eig}}(A)^2\delta \text{ and } \|\lambda'_i - \lambda_i\| \leq 4\kappa_{\text{eig}}(A)^2\delta \quad \forall i = 1, \ldots, n, \tag{3.3}$$

after possibly multiplying the v_i by phases.

[2]This quantity is inspired by but not identical to the "distance to ill-posedness" for the eigenproblem considered by Demmel [46], to which it is polynomially related.

[3]In fact, it can be shown that $\kappa_{\text{eig}}(A)$ is polynomially related to the smallest constant for which (3.3) holds for all sufficiently small $\delta > 0$.

Note that $\kappa_{\text{eig}} = \infty$ if and only if A has a double eigenvalue; in this case, a relation like (3.3) is not possible since different infinitesimal changes to A can produce macroscopically different eigenpairs.

In this chapter we will present a backward approximation approximation for the eigenproblem with running time scaling polynomially in $\log(1/\delta)$, which by (3.3) yields a forward approximation algorithm with running time scaling polynomially in $\log(1/\kappa_{\text{eig}}\delta)$.

Remark 3.1.2 (Multiple Eigenvalues)**.** A backward approximation algorithm for the eigenproblem can be used to accurately find bases for the eigenspaces of matrices with multiple eigenvalues, but quantifying the forward error requires introducing condition numbers for invariant subspaces rather than eigenpairs. A standard treatment of this can be found in any numerical linear algebra textbook, e.g. [43], and we do not discuss it further in this work for simplicity of exposition.

3.1.1.2 Models of Computation

These questions may be studied in various computational models: exact *real arithmetic* (i.e., infinite precision), *variable precision rational arithmetic* (rationals are stored exactly as numerators and denominators), and *finite precision arithmetic* (real numbers are rounded to a fixed number of bits which may depend on the input size and accuracy). Only the last two models yield actual Boolean complexity bounds, but introduce a second source of error stemming from the fact that computers cannot exactly represent real numbers.

We study the third model in this chapter, axiomatized as follows.

Finite Precision Arithmetic. We use the standard axioms from [70]. Numbers are stored and manipulated approximately up to some machine precision $\mathbf{u} := \mathbf{u}(\delta, n) > 0$, which for us will depend on the instance size n and desired accuracy δ. This means every number $x \in \mathbb{C}$ is stored as $\mathsf{fl}(x) = (1+\Delta)x$ for some adversarially chosen $\Delta \in \mathbb{C}$ satisfying $|\Delta| \leq \mathbf{u}$, and each arithmetic operation $\circ \in \{+, -, \times, \div\}$ is guaranteed to yield an output satisfying

$$\mathsf{fl}(x \circ y) = (x \circ y)(1+\Delta) \quad |\Delta| \leq \mathbf{u}.$$

It is also standard and convenient to assume that we can evaluate $\sqrt{x}$ for any $x \in \mathbb{R}$, where again $\mathsf{fl}(\sqrt{x}) = \sqrt{x}(1+\Delta)$ for $|\Delta| \leq \mathbf{u}$.

Thus, the outcomes of all operations are adversarially noisy due to roundoff. The bit lengths of numbers stored in this form remain fixed at $\lg(1/\mathbf{u})$, where lg denotes the logarithm base 2. The *bit complexity* of an algorithm is therefore the number of arithmetic operations times $O^*(\log(1/\mathbf{u}))$, the running time of standard floating point arithmetic, where the $*$ suppresses $\log\log(1/\mathbf{u})$ factors. We will state all running times in terms of arithmetic operations accompanied by the required number of bits of precision, which thereby immediately imply bit complexity bounds.

Remark 3.1.3 (Overflow, Underflow, and Additive Error)**.** Using p bits for the exponent in the floating-point representation allows one to represent numbers with magnitude in the range $[2^{-2^p}, 2^{2^p}]$. It can be easily checked that all of the nonzero numbers, norms, and condition numbers appearing during the execution of our algorithms lie in the range $[2^{-\lg^c(n/\delta)}, 2^{\lg^c(n/\delta)}]$ for some small c, so overflow and underflow do not occur. In fact, we could have analyzed our algorithm in a computational model where every number is simply rounded to the nearest rational with denominator $2^{\lg^c(n/\delta)}$—corresponding to *additive* arithmetic errors. We have chosen to use the multiplicative error floating point model since it is the standard in numerical analysis, but our algorithms do not exploit any subtleties arising from the difference between the two models.

The advantages of the floating point model are that it is realistic (being and potentially yields very fast algorithms by using a small number of bits of precision (polylogarithmic in n and $1/\delta$), in contrast to rational arithmetic, where even a simple operation such as inverting an $n \times n$ integer matrix requires n extra bits of precision (see, e.g., Chapter 1 of [66]). An iterative algorithm that can be implemented in finite precision (typically, polylogarithmic in the input size and desired accuracy) is called *numerically stable*, and corresponds to a dynamical system whose trajectory to the approximate solution is robust to adversarial noise (see, e.g. [114]).

The disadvantage of the model is that it is only possible to compute forward approximations of quantities which are *well-conditioned* in the input—in particular, discontinuous quantities such as eigenvalue multiplicity cannot be computed in the floating point model, since it is not even assumed that the input is stored exactly.

3.1.2 Results and Techniques

Eigenvalue Gaps, κ_V, and Pseudospectral Shattering. The key probabilistic result of the paper is that a random *complex* Gaussian perturbation of any matrix yields a nearby matrix with large minimum eigenvalue gap and small κ_V.

Theorem 3.1.4 (Smoothed Analysis of gap and κ_V)**.** Suppose $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$, and $\gamma \in (0, 1/2)$. Let G_n be an $n \times n$ matrix with i.i.d. complex Gaussian $N(0, 1_{\mathbb{C}}/n)$ entries, and let $X := A + \gamma G_n$. Then

$$\kappa_V(X) \leq \frac{n^2}{\gamma}, \quad \text{gap}(X) \geq \frac{\gamma^4}{n^5}, \quad \text{and} \quad \|G_n\| \leq 4,$$

with probability at least $1 - 1/n - O(1/n^2)$ where the implicit constant is at most 600.

The proof of Theorem 3.1.4 appears in Section 3.3.1. The key idea is to first control $\kappa_V(X)$ using [11], and then observe that for a matrix with small κ_V, two eigenvalues of X near a complex number z imply a small *second-least* singular value of $z - X$, which we are able to control.

In Section 3.3.2 we develop the notion of *pseudospectral shattering*, which is implied by Theorem 3.1.4 and says roughly that the pseudospectrum consists of n components that lie in separate squares of an appropriately coarse grid in the complex plane. This is useful in the analysis of the spectral bisection algorithm in Section 3.5.

Matrix Sign Function. The sign function of a number $z \in \mathbb{C}$ with $\Re(z) \neq 0$ is defined as $+1$ if $\Re(z) > 0$ and -1 if $\Re(z) < 0$. The *matrix sign function* of a matrix A with Jordan normal form

$$A = V \begin{bmatrix} N & \\ & P \end{bmatrix} V^{-1},$$

where N (resp. P) has eigenvalues with strictly negative (resp. positive) real part, is defined as

$$\operatorname{sgn}(A) = V \begin{bmatrix} -I_N & \\ & I_P \end{bmatrix} V^{-1},$$

where I_P denotes the identity of the same size as P. The sign function is undefined for matrices with eigenvalues on the imaginary axis. Quantifying this discontinuity, Bai and Demmel [7] defined the following condition number for the sign function:

$$\kappa_{\text{sign}}(M) := \inf \left\{ 1/\epsilon^2 : \Lambda_\epsilon(M) \text{ does not intersect the imaginary axis} \right\}, \tag{3.4}$$

and gave perturbation bounds for $\operatorname{sgn}(M)$ depending on κ_{sign}.

Roberts [105] showed that the simple iteration

$$A_{k+1} = \frac{A_k + A_k^{-1}}{2} \tag{3.5}$$

converges globally and quadratically to $\operatorname{sgn}(A)$ in exact arithmetic, but his proof relies on the fact that all iterates of the algorithm are simultaneously diagonalizable, a property which is destroyed in finite arithmetic since inversions can only be done approximately.[4] In Section 3.4 we show that this iteration is indeed convergent when implemented in finite arithmetic for matrices with small κ_{sign}, given a numerically stable matrix inversion algorithm. This leads to the following result:

Theorem 3.1.5 (Sign Function Algorithm)**.** There is a deterministic algorithm SGN which on input an $n \times n$ matrix A with $\|A\| \leq 1$, a number K with $K \geq \kappa_{\text{sign}}(A)$, and a desired accuracy $\beta \in (0, 1/12)$, outputs an approximation $\mathsf{SGN}(A)$ with

$$\|\mathsf{SGN}(A) - \operatorname{sgn}(A)\| \leq \beta,$$

in

$$O((\log K + \log\log(1/\beta)) T_{\mathsf{INV}}(n)) \tag{3.6}$$

[4] Doing the inversions exactly in rational arithmetic could require numbers of bit length n^k for k iterations, which will typically not even be polynomial.

arithmetic operations on a floating point machine with

$$\lg(1/\mathbf{u}) = O(\log n \log^3 K(\log K + \log(1/\beta)))$$

bits of precision, where $T_{\mathsf{INV}}(n)$ denotes the number of arithmetic operations used by a numerically stable matrix inversion algorithm (satisfying Definition 3.2.3).

The main new idea in the proof of Theorem 3.1.5 is to control the evolution of the pseudospectra $\Lambda_{\epsilon_k}(A_k)$ of the iterates with appropriately decreasing (in k) parameters ϵ_k, using a sequence of carefully chosen shrinking contour integrals in the complex plane. The pseudospectrum provides a richer induction hypothesis than scalar quantities such as condition numbers, and allows one to control all quantities of interest using the holomorphic functional calculus. This technique is introduced in Sections 3.4.1 and 3.4.2, and carried out in finite arithmetic in Section 3.4.3, yielding Theorem 3.1.5.

Diagonalization by Spectral Bisection. Given an algorithm for computing the sign function, there is a natural and well-known approach to the eigenproblem pioneered in [17]. The matrices $(I \pm \mathrm{sgn}(A))/2$ are spectral projectors onto the invariant subspaces corresponding to the eigenvalues of A in the left and right open half planes, so if some shift of A or iA has roughly half its eigenvalues on either side of the imaginary axis, the problem can be reduced to smaller subproblems appropriate for recursion.

The two difficulties in carrying out the above approach are: (a) efficiently computing the sign function (b) finding a balanced splitting along an axis that is well-separated from the spectrum. These are nontrivial even in exact arithmetic, since the iteration (3.5) converges slowly if (b) is not satisfied, even without roundoff error. We use Theorem 3.1.4 to ensure that a good splitting always exists after a small Gaussian perturbation of order δ, and Theorem 3.1.5 to compute splittings efficiently in finite precision. Combining this with well-understood techniques such as rank-revealing QR factorization, we obtain the following theorem, whose proof appears in Section 3.5.1.

Theorem 3.1.6 (Backward Approximation Algorithm)**.** There is a randomized algorithm EIG which on input any matrix $A \in \mathbb{C}^{n\times n}$ with $\|A\| \le 1$ and a desired accuracy parameter $\delta > 0$ outputs a diagonal D and invertible V such that

$$\|A - VDV^{-1}\| \le \delta \quad \text{and} \quad \kappa(V) \le 32n^{2.5}/\delta$$

in

$$O\left(T_{\mathsf{MM}}(n) \log^2 \frac{n}{\delta}\right)$$

arithmetic operations on a floating point machine with

$$O(\log^4(n/\delta) \log n)$$

bits of precision, with probability at least $1 - 2/n - O(1/n^2)$, where the implied constant is at most 600. Here $T_{\mathsf{MM}}(n)$ refers to the running time of a numerically stable matrix multiplication algorithm (detailed in Section 3.2.3).

Considering (3.3), we have the following immediate corollary by invoking EIG with accuracy $\delta/\kappa_{\text{eig}}$.

Corollary 3.1.7 (Forward Approximation Algorithm)**.** There is a randomized algorithm which on input any matrix $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$, a desired accuracy parameter $\delta > 0$, and an estimate $K \geq \kappa_{\text{eig}}(A)$ outputs a $\delta-$forward approximate solution to the eigenproblem for A in

$$O\left(T_{\mathsf{MM}}(n)\log^2\frac{nK}{\delta}\right)$$

arithmetic operations on a floating point machine with

$$O(\log^4(nK/\delta)\log n)$$

bits of precision, with probability at least $1 - 2/n - O(1/n^2)$. Here $T_{\mathsf{MM}}(n)$ refers to the running time of a numerically stable matrix multiplication algorithm (detailed in Section 3.2.3).

Remark 3.1.8 (Accuracy vs. Precision)**.** The gold standard of "backward stability" in numerical analysis postulates that

$$\log(1/\mathbf{u}) = \log(1/\delta) + \log(n),$$

i.e., the number of bits of precision is linear in the number of bits of accuracy. The relaxed notion of "logarithmic stability" introduced in [48] requires

$$\log(1/\mathbf{u}) = \log(1/\delta) + O(\log^c(n)\log(\kappa))$$

for some constant c, where κ is an appropriate condition number. In comparison, Theorem 3.1.6 obtains the weaker relationship

$$\log(1/\mathbf{u}) = O(\log^4(1/\delta)\log(n) + \log^5(n)),$$

which is still polylogarithmic in n in the regime $\delta = 1/\text{poly}(n)$.

3.1.3 Related Work

Minimum Eigenvalue Gap. The minimum eigenvalue gap of random matrices has been studied in the case of Hermitian and unitary matrices, beginning with the work of Vinson [130], who proved an $\Omega(n^{-4/3})$ lower bound on this gap in the case of the Gaussian Unitary Ensemble (GUE) and the Circular Unitary Ensemble (CUE). Bourgade and Ben Arous [5] derived exact limiting formulas for the distributions of all the gaps for the same ensembles. Nguyen, Tao, and Vu [100] obtained non-asymptotic inverse polynomial bounds for a large class of non-integrable Hermitian models with i.i.d. entries (including Bernoulli matrices).

In a different direction, Aizenman et al. proved an inverse-polynomial bound [1] in the case of an arbitrary Hermitian matrix plus a GUE matrix or a Gaussian Orthogonal Ensemble

(GOE) matrix, which may be viewed as a smoothed analysis of the minimum gap. Theorem 3.3.6 may be viewed as a non-Hermitian analogue of the last result.

In the non-Hermitian case, Ge [60] obtained an inverse polynomial bound for i.i.d. matrices with real entries satisfying some mild moment conditions, and [110][5] proved an inverse polynomial lower bound for the complex Ginibre ensemble. Theorem 3.3.6 may be seen as a generalization of these results to non-centered complex Gaussian matrices.

Smoothed Analysis and Free Probability. The study of numerical algorithms on Gaussian random matrices (i.e., the case $A = 0$ of smoothed analysis) dates back to [131, 115, 44, 52]. The powerful idea of improving the conditioning of a numerical computation by adding a small amount of Gaussian noise was introduced by Spielman and Teng in [117], in the context of the simplex algorithm. Sankar, Spielman, and Teng [109] showed that adding real Gaussian noise to any matrix yields a matrix with polynomially-bounded condition number; [11] can be seen as an extension of this result to the condition number of the eigenvector matrix, where the proof crucially requires that the Gaussian perturbation is complex rather than real. The main difference between our results and most of the results on smoothed analysis (including [4]) is that our running time depends logarithmically rather than polynomially on the size of the perturbation.

The broad idea of regularizing the spectral instability of a nonnormal matrix by adding a random matrix can be traced back to the work of Śniady [116] and Haagerup and Larsen [69] in the context of Free Probability theory.

Matrix Sign Function. The matrix sign function was introduced by Zolotarev in 1877. It became a popular topic in numerical analysis following the work of Beavers and Denman [18, 17, 49] and Roberts [105], who used it first to solve the algebraic Ricatti and Lyapunov equations and then as an approach to the eigenproblem; see [78] for a broad survey of its early history. The numerical stability of Roberts' Newton iteration was investigated by Byers [27], who identified some cases where it is and isn't stable. Malyshev [90], Byers, He, and Mehrmann [28], Bai, Demmel, and Gu [8], and Bai and Demmel [7] studied the condition number of the matrix sign function, and showed that if the Newton iteration converges then it can be used to obtain a high-quality invariant subspace[6], but did not prove convergence in finite arithmetic and left this as an open question.[7] The key issue in analyzing the convergence of the iteration is to bound the condition numbers of the intermediate matrices that appear, as N. Higham remarks in his 2008 textbook:

> Of course, to obtain a complete picture, we also need to understand the effect of rounding errors on the iteration prior to convergence. This effect is surprisingly difficult to analyze. . . . Since errors will in general occur on each iteration, the

[5] At the time of writing, the work [110] is still an unpublished arXiv preprint.

[6] This is called an *a fortiriori bound* in numerical analysis.

[7] [28] states: "A priori backward and forward error bounds for evaluation of the matrix sign function remain elusive."

> overall error will be a complicated function of $\kappa_{sign}(X_k)$ and E_k for all k. . . . We are not aware of any published rounding error analysis for the computation of $sign(A)$ via the Newton iteration. –[71, Section 5.7]

This is precisely the problem solved by Theorem 3.1.5, which is as far as we know the first provable algorithm for computing the sign function of an arbitrary matrix which does not require computing the Jordan form.

In the special case of Hermitian matrices, Higham [72] established efficient reductions between the sign function and the polar decomposition. Byers and Xu [29] proved backward stability of a certain scaled version of the Newton iteration for Hermitian matrices, in the context of computing the polar decomposition. Higham and Nakatsukasa [99] (see also the improvement [98]) proved backward stability of a different iterative scheme for computing the polar decomposition, and used it to give backward stable spectral bisection algorithms for the Hermitian eigenproblem with $O(n^3)$-type complexity.

Non-Hermitian Eigenproblem. *Floating Point Arithmetic.* The eigenproblem has been thoroughly studied in the numerical analysis community, in the floating point model of computation. While there are provably fast and accurate algorithms in the Hermitian case (see the next subsection) and a large body of work for various structured matrices (see, e.g., [23]), the general case is not nearly as well-understood. As recently as 1998, J. Demmel remarked in his well-known textbook [43]: ". . . the problem of devising an algorithm [for the non-Hermitian eigenproblem] that is numerically stable and globally (and quickly!) convergent remains open."

Demmel's question remained entirely open until 2015, when it was answered in the following sense by Armentano, Beltrán, Bürgisser, Cucker, and Shub in the remarkable paper [4]. They exhibited an algorithm (see their Theorem 2.28) which given any $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$ and a desired accuracy $\delta > 0$ produces in $O(n^9/\delta^2)$ expected arithmetic operations the diagonalization of the nearby random perturbation $A + \delta G$ where G is a matrix with standard complex Gaussian entries. By setting δ sufficiently small, this may be viewed as a backward approximation algorithm for diagonalization, in that it solves a nearby problem essentially exactly.[8] Their algorithm is based on homotopy continuation methods, which they argue informally are numerically stable and can be implemented in finite precision arithmetic. Our algorithm is similar on a high level in that it adds a Gaussian perturbation to the input and then obtains a high accuracy forward approximate solution to the perturbed problem. The difference is that their overall running time depends polynomially rather than logarithmically on the accuracy δ desired with respect to the original unperturbed problem.

Other Models of Computation. If we relax the requirements further and ask for any provable algorithm in any model of Boolean computation, there is only one more positive result with a polynomial bound on the number of bit operations: Jin Yi Cai showed in 1994 [30] that given a rational $n \times n$ matrix A with integer entries of bit length a, one can find an

[8] The output of their algorithm is n vectors on each of which Newton's method converges quadratically to an eigenvector.

Result	Error	Arithmetic Ops	Boolean Ops	Restrictions
[103]	Backward	$n^3 + n^2 \log(1/\delta)$	$n^3 \log(n/\delta) + n^2 \log(1/\delta) \log(n/\delta)$	Hermitian
[3]	Backward	n^9/δ^2	$n^9/\delta^2 \cdot \mathrm{polylog}(n/\delta)$[a]	
[21]	Backward	$n^{\omega+1}\mathrm{polylog}(n)\log(1/\delta)$	$n^{\omega+1}\mathrm{polylog}(n)\log(1/\delta)$	Hermitian
Theorem 3.1.6[b]	Backward	$T_{\mathsf{MM}}(n)\log^2(n/\delta)$	$T_{\mathsf{MM}}(n)\log^6(n/\delta)\log(n)$	
Corollary 3.1.7	Forward	$T_{\mathsf{MM}}(n)\log^2(n\kappa_{\mathrm{eig}}/\delta)$	$T_{\mathsf{MM}}(n)\log^6(n\kappa_{\mathrm{eig}}/\delta)\log(n)$	

[a] Does not specify a particular bound on precision.
[b] $T_{\mathsf{MM}}(n) = O(n^{\omega+\eta})$ for every $\eta > 0$, see Definition 3.2.2 for details.

Table 3.1: Results for finite-precision floating-point arithmetic

Result	Model	Error	Arithmetic Ops	Boolean Ops	Restrictions
[30]	Rational	Forward[a]	$\mathrm{poly}(a, n, \log(1/\delta))$[b]	$\mathrm{poly}(a, n, \log(1/\delta))$	
[102]	Rational	Forward	$n^\omega + n \log\log(1/\delta)$	$n^{\omega+1}a + n^2 \log(1/\delta)\log\log(1/\delta)$	Eigenvalues only[c]
[89]	Finite[c]	Forward	$n^\omega \log(n)\log(1/\delta)$	$n^\omega \log^4(n)\log^2(n/\delta)$	Hermitian, λ_1 only

[a] Actually computes the Jordan Normal Form. The degree of the polynomial is not specified, but is at least 12 in n.
[b] In the bit operations, a denotes the bit length of the input entries.
[c] Uses a custom bit representation of intermediate quantities.

Table 3.2: Results for other models of arithmetic

δ-forward approximation to its Jordan Normal Form $A = VJV^{-1}$ in time $\mathrm{poly}(n, a, \log(1/\delta))$, where the degree of the polynomial is at least 12. This algorithm works in the rational arithmetic model of computation, so it does not quite answer Demmel's question since it is not a numerically stable algorithm. However, it enjoys the significant advantage of being able to compute forward approximations to discontinuous quantities such as the Jordan structure.

As far as we are aware, there are no other published provably polynomial-time algorithms for the general eigenproblem. The two standard references for diagonalization appearing most often in theoretical computer science papers do not meet this criterion. In particular, the widely cited work by Pan and Chen [102] proves that one can compute the *eigenvalues* of A in $O(n^\omega + n \log\log(1/\delta))$ (suppressing logarithmic factors) *arithmetic* operations by finding the roots of its characteristic polynomial, which becomes a bound of $O(n^{\omega+1}a + n^2 \log(1/\delta)\log\log(1/\delta))$ bit operations if the characteristic polynomial is computed exactly in rational arithmetic and the matrix has entries of bit length a. However that paper does not give any bound for the amount of time taken to find approximate eigenvectors from approximate eigenvalues, and states this as an open problem.[9]

Finally, the important work of Demmel, Dumitriu, and Holtz [47] (see also the followup [9]), which we rely on heavily, does not claim to provably solve the eigenproblem either—it

[9] "The remaining nontrivial problems are, of course, the estimation of the above output precision p [sufficient for finding an approximate eigenvector from an approximate eigenvalue], ... We leave these open problems as a challenge for the reader." – [102, Section 12].

bounds the running time of one iteration of a specific algorithm, and shows that such an iteration can be implemented numerically stably, without proving any bound on the number of iterations required in general.

Hermitian Eigenproblem. For comparison, the eigenproblem for Hermitian matrices is much better understood. We cannot give a complete bibliography of this huge area, but mention one relevant landmark result: the work of Wilkinson [132] and Hoffman-Parlett [74] in the 60's and 70's, which shows that the Hermitian eigenproblem can be solved with backward error δ in $O(n^3 + n^2 \log(1/\delta))$ arithmetic operations with $O(\log(n/\delta))$ bits of precision. There has also recently been renewed interest in this problem in the theoretical computer science community, with the goal of bringing the runtime close to $O(n^\omega)$: Louis and Vempala [89] show how to find a $\delta-$approximation of just the largest eigenvalue in $O(n^\omega \log^4(n) \log^2(1/\delta))$ bit operations, and Ben-Or and Eldar [21] give an $O(n^{\omega+1}\text{polylog}(n))$-bit-operation algorithm for finding a $1/\text{poly}(n)$-approximate diagonalization of an $n \times n$ Hermitian matrix normalized to have $\|A\| \leq 1$.

Remark 3.1.9 (Davies' Conjecture)**.** The beautiful paper [36] introduced the idea of approximating a matrix function $f(A)$ for nonnormal A by $f(A + E)$ for some well-chosen E regularizing the eigenvectors of A. This directly inspired our approach to solving the eigenproblem via regularization.

The existence of an approximate diagonalization (3.1) for every A with a *well-conditioned similarity* V (i.e, $\kappa(V)$ depending polynomially on δ and n) was precisely the content of Davies' conjecture [36], which was recently solved by some of the authors and Mukherjee in [11]. The existence of such a V is a prerequisite for proving that one can always efficiently find an approximate diagonalization in finite arithmetic, since if $\|V\|\|V^{-1}\|$ is very large it may require many bits of precision to represent. Thus, Theorem 3.1.6 can be viewed as an efficient algorithmic answer to Davies' question.

Reader Guide. This chapter contains a lot of parameters and constants. On first reading, it may be good to largely ignore the constants not appearing in exponents, and to keep in mind the typical setting $\delta = 1/\text{poly}(n)$ for the accuracy, in which case the important auxiliary parameters $\omega, 1-\alpha, \epsilon, \beta, \eta$ are all $1/\text{poly}(n)$, and the machine precision is $\log(1/\mathbf{u}) = \text{polylog}(n)$.

3.2 Preliminaries

Let $M \in \mathbb{C}^{n\times n}$ be a complex matrix, not necessarily normal. We will write matrices and vectors with uppercase and lowercase letters, respectively. Let us denote by $\Lambda(M)$ the spectrum of M and by $\lambda_i(M)$ its individual eigenvalues. In the same way we denote the singular values of M by $\sigma_i(M)$ and we adopt the convention $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_n(M)$. When M is clear from the context we will simplify notation and just write Λ, λ_i or σ_i respectively.

Recall that the *operator norm* of M is

$$\|M\| = \sigma_1(M) = \sup_{\|x\|=1} \|Mx\|.$$

As usual, we will say that M is *diagonalizable* if it can be written as $M = VDV^{-1}$ for some diagonal matrix D whose nonzero entries contain the eigenvalues of M. In this case we have the spectral expansion

$$M = \sum_{i=1}^{n} \lambda_i v_i w_i^*, \tag{3.7}$$

where the right and left eigenvectors v_i and w_j^* are the columns and rows of V and V^{-1} respectively, normalized so that $w_i^* v_i = 1$.

We now record the proof of Proposition 3.1.1, which follows from a simple contour integral argument appearing in Section 3.5.1.

Proof of Proposition 3.1.1. We repeat the proof of Lemma 3.5.8 with $\eta = \delta$, where instead of using a grid square we use a circular contour around λ_i with radius $\omega = \frac{\text{gap}}{2}$, so that by (1.5) we may set $\varepsilon = \frac{\text{gap}}{2\kappa_V}$. Note that $\eta < \varepsilon/2$ and $\text{gap}(A) \leq 2$ by hypothesis, which gives the desired bound. □

3.2.1 Finite-Precision Arithmetic

We briefly elaborate on the axioms for floating-point arithmetic given in Section 3.1.1. Similar guarantees to the ones appearing in that section for scalar-scalar operations also hold for operations such as matrix-matrix addition and matrix-scalar multiplication. In particular, if A is an $n \times n$ complex matrix,

$$\mathsf{fl}(A) = A + A \circ \Delta \qquad |\Delta_{i,j}| < \mathbf{u}.$$

It will be convenient for us to write such errors in additive, as opposed to multiplicative form. We can convert the above to additive error as follows. Recall that for any $n \times n$ matrix, the spectral norm (the $\ell^2 \to \ell^2$ operator norm) is at most $\sqrt{n}$ times the $\ell^2 \to \ell^1$ operator norm, i.e. the maximal norm of a column. Thus we have

$$\|A \circ \Delta\| \leq \sqrt{n} \max_i \|(A \circ \Delta)e_i\| \leq \sqrt{n} \max_{i,j} |\Delta_{i,j}| \max_i \|Ae_i\| \leq \mathbf{u}\sqrt{n}\|A\|. \tag{3.8}$$

For more complicated operations such as matrix-matrix multiplication and matrix inversion, we use existing error guarantees from the literature. This is the subject of Section 3.2.3.

We will also need to compute the trace of a matrix $A \in \mathbb{C}^{n \times n}$, and normalize a vector $x \in \mathbb{C}^n$. Error analysis of these is standard (see for instance the discussion in [70, Section 3.1-3.4, 4.1]) and the results in this chapter are highly insensitive to the details. For simplicity, calling $\hat{x} := x/\|x\|$, we will assume that

$$|\mathsf{fl}\,(\text{Tr}A) - \text{Tr}A| \leq n\|A\|\mathbf{u} \tag{3.9}$$

$$\|\mathsf{fl}(\hat{x}) - \hat{x}\| \leq n\mathbf{u}. \tag{3.10}$$

Each of these can be achieved by assuming that $\mathbf{u}n \leq \epsilon$ for some suitably chosen ϵ, independent of n, a requirement which will be depreciated shortly by several tighter assumptions on the machine precision.

Throughout this chapter, we will take the pedagogical perspective that our algorithms are games played between the practitioner and an adversary who may additively corrupt each operation. In particular, we will include explicit error terms (always denoted by $E_{(\cdot)}$) in each appropriate step of every algorithm. In many cases we will first analyze a routine in exact arithmetic—in which case the error terms will all be set to zero—and subsequently determine the machine precision $\mathbf{u}$ necessary to make the errors small enough to guarantee convergence.

3.2.2 Sampling Gaussians in Finite Precision

For various parts of the algorithm, we will need to sample from normal distributions. For our model of arithmetic, we assume that the complex normal distribution can be sampled up to machine precision in $O(1)$ arithmetic operations. To be precise, we assume the existence of the following sampler:

Definition 3.2.1 (Complex Gaussian Sampling)**.** A c_{N}-stable Gaussian sampler $\mathsf{N}(\sigma)$ takes as input $\sigma \in \mathbb{R}_{\geq 0}$ and outputs a sample of a random variable $\widetilde{G} = \mathsf{N}(\sigma)$ with the property that there exists $G \sim N_{\mathbb{C}}(0, \sigma^2)$ satisfying

$$|\widetilde{G} - G| \leq c_{\mathsf{N}} \sigma \cdot \mathbf{u}$$

with probability one, in at most T_{N} arithmetic operations for some universal constant $T_{\mathsf{N}} > 0$.

We will only sample $O(n^2)$ Gaussians during the algorithm, so this sampling will not contribute significantly to the runtime. Here as everywhere in the chapter, we will omit issues of underflow or overflow. To simplify some of our bounds, we will also assume that $c_{\mathsf{N}} \geq 1$.

3.2.3 Black-box Error Assumptions for Multiplication, Inversion, and QR

Our algorithm uses matrix-matrix multiplication, matrix inversion, and QR factorization as primitives. For our analysis, we must therefore assume some bounds on the error and runtime costs incurred by these subroutines. In this section, we first formally state the kind of error and runtime bounds we require, and then discuss some implementations known in the literature that satisfy each of our requirements with modest constants.

Our definitions are inspired by the definition of *logarithmic stability* introduced in [47]. Roughly speaking, they say that implementing the algorithm with floating point precision $\mathbf{u}$ yields an accuracy which is at most polynomially or quasipolynomially in n worse than $\mathbf{u}$ (possibly also depending on the condition number in the case of inversion). Their definition has the property that while a logarithmically stable algorithm is not strictly-speaking backward stable, it can attain the same forward error bound as a backward stable algorithm at the

cost of increasing the bit length by a polylogarithmic factor. See Section 3 of their paper for a precise definition and a more detailed discussion of how their definition relates to standard numerical stability notions.

Definition 3.2.2. A $\mu_{\mathsf{MM}}(n)$-*stable multiplication algorithm* takes as input $A, B \in \mathbb{C}^{n\times n}$ and a precision $\mathbf{u} > 0$ and outputs $C = \mathsf{MM}(A, B)$ satisfying

$$\|C - AB\| \leq \mu_{\mathsf{MM}}(n) \cdot \mathbf{u}\|A\|\|B\|,$$

on a floating point machine with precision $\mathbf{u}$, in $T_{\mathsf{MM}}(n)$ arithmetic operations.

Definition 3.2.3. A $(\mu_{\mathsf{INV}}(n), c_{\mathsf{INV}})-$*stable inversion algorithm* $\mathsf{INV}(\cdot)$ takes as input $A \in \mathbb{C}^{n\times n}$ and a precision $\mathbf{u}$ and outputs $C = \mathsf{INV}(A)$ satisfying

$$\|C - A^{-1}\| \leq \mu_{\mathsf{INV}}(n) \cdot \mathbf{u} \cdot \kappa(A)^{c_{\mathsf{INV}} \log n}\|A^{-1}\|,$$

on a floating point machine with precision $\mathbf{u}$, in $T_{\mathsf{INV}}(n)$ arithmetic operations.

Definition 3.2.4. A $\mu_{\mathsf{QR}}(n)$-*stable QR factorization algorithm* $\mathsf{QR}(\cdot)$ takes as input $A \in \mathbb{C}^{n\times n}$ QR and a precision $\mathbf{u}$, and outputs $[Q, R] = \mathsf{QR}(A)$ such that

1. R is exactly upper triangular.
2. There is a unitary Q' and a matrix A' such that

$$Q'A' = R, \tag{3.11}$$

and

$$\|Q' - Q\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}, \quad \text{and} \quad \|A' - A\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}\|A\|,$$

on a floating point machine with precision $\mathbf{u}$. Its running time is $T_{\mathsf{QR}}(n)$ arithmetic operations.

Remark 3.2.5. Throughout this chapter, to simplify some of our bounds, we will assume that

$$1 \leq \mu_{\mathsf{MM}}(n), \mu_{\mathsf{INV}}(n), \mu_{\mathsf{QR}}(n), c_{\mathsf{INV}} \log n.$$

The above definitions can be instantiated with traditional $O(n^3)$-complexity algorithms for which $\mu_{\mathsf{MM}}, \mu_{\mathsf{QR}}, \mu_{\mathsf{INV}}$ are all $O(n)$ and $c_{\mathsf{INV}} = 1$ [70]. This yields easily-implementable practical algorithms with running times depending cubically on n.

In order to achieve $O(n^\omega)$-type efficiency, we instantiate them with fast-matrix-multiplication-based algorithms and with $\mu(n)$ taken to be a low-degree polynomial [47]. Specifically, the following parameters are known to be achievable.

Theorem 3.2.6 (Fast and Stable Instantiations of $\mathsf{MM}, \mathsf{INV}, \mathsf{QR}$)**.**

1. If ω is the exponent of matrix multiplication, then for every $\eta > 0$ there is a $\mu_{\mathsf{MM}}(n)-$stable multiplication algorithm with $\mu_{\mathsf{MM}}(n) = n^{c_\eta}$ and $T_{\mathsf{MM}}(n) = O(n^{\omega+\eta})$, where c_η does not depend on n.

2. Given an algorithm for matrix multiplication satisfying (1), there is a $(\mu_{\mathsf{INV}}(n), c_{\mathsf{INV}})$-stable inversion algorithm with

$$\mu_{\mathsf{INV}}(n) \leq O(\mu_{\mathsf{MM}}(n)n^{\lg(10)}), \qquad c_{\mathsf{INV}} \leq 8,$$

   and $T_{\mathsf{INV}}(n) \leq T_{\mathsf{MM}}(3n) = O(T_{\mathsf{MM}}(n))$.

3. Given an algorithm for matrix multiplication satisfying (1), there is a $\mu_{\mathsf{QR}}(n)-$stable QR factorization algorithm with

$$\mu_{\mathsf{QR}}(n) = O(n^{c_{\mathsf{QR}}}\mu_{\mathsf{MM}}(n)),$$

   where c_{QR} is an absolute constant, and $T_{\mathsf{QR}}(n) = O(T_{\mathsf{MM}}(n))$.

In particular, all of the running times above are bounded by $T_{\mathsf{MM}}(n)$ for an $n \times n$ matrix.

Proof. (1) is Theorem 3.3 of [48]. (2) is Theorem 3.3 (see also equation (9) above its statement) of [47]. The final claim follows by noting that $T_{\mathsf{MM}}(3n) = O(T_{\mathsf{MM}}(n))$ by dividing a $3n \times 3n$ matrix into nine $n \times n$ blocks and proceeding blockwise, at the cost of a factor of 9 in $\mu_{\mathsf{INV}}(n)$. (3) appears in Section 4.1 of [47].

□

We remark that for specific existing fast matrix multiplication algorithms such as Strassen's algorithm, specific small values of $\mu_{\mathsf{MM}}(n)$ are known (see [48] and its references for details), so these may also be used as a black box, though we will not do this in this work.

3.3 Pseudospectral Shattering

This section is devoted to our central probabilistic result, Theorem 3.1.4, and the accompanying notion of *pseudospectral shattering* which will be used extensively in our analysis of the spectral bisection algorithm in Section 3.5.

3.3.1 Smoothed Analysis of Gap and Eigenvector Condition Number

As is customary in the literature, we will refer to an $n \times n$ random matrix G_n whose entries are independent complex Gaussians drawn from $\mathcal{N}(0, 1_{\mathbb{C}}/n)$ as a *normalized complex Ginibre random matrix*. To be absolutely clear, and because other choices of scaling are quite common, we mean that $\mathbb{E}G_{i,j} = 0$ and $\mathbb{E}|G_{i,j}|^2 = 1/n$.

In the course of proving Theorem 3.1.4, we will need to bound the probability that the second-smallest singular value of an arbitrary matrix with small Ginibre perturbation is atypically small. We begin with a well-known lower tail bound on the singular values of a Ginibre matrix alone.

Theorem 3.3.1 ([119, Theorem 1.2])**.** For G_n an $n \times n$ normalized complex Ginibre matrix and for any $\alpha \geq 0$ it holds that

$$\mathbb{P}\left[\sigma_j(G_n) < \frac{\alpha(n-j+1)}{n}\right] \leq \left(\sqrt{2e}\,\alpha\right)^{2(n-j+1)^2}.$$

As in Chapter 2, we can transfer this result to case of a Ginibre perturbation via the remarkable coupling result of Śniady, which we restate below:

Theorem 3.3.2 (Śniady [116])**.** Let A_1 and A_2 be $n \times n$ complex matrices such that $\sigma_i(A_1) \leq \sigma_i(A_2)$ for all $1 \leq i \leq n$. Assume further that $\sigma_i(A_1) \neq \sigma_j(A_1)$ and $\sigma_i(A_2) \neq \sigma_j(A_2)$ for all $i \neq j$. Then for every $t \geq 0$, there exists a joint distribution on pairs of $n \times n$ complex matrices (G_1, G_2) such that

1. the marginals G_1 and G_2 are distributed as normalized complex Ginibre matrices, and
2. almost surely $\sigma_i(A_1 + \sqrt{t}G_1) \leq \sigma_i(A_2 + \sqrt{t}G_2)$ for every i.

Corollary 3.3.3. For any fixed matrix M and parameters $\gamma, t > 0$

$$\mathbb{P}[\sigma_{n-1}(M + \gamma G_n) < t] \leq (e/2)^4 (tn/\gamma)^8 \leq 4(tn/\gamma)^8.$$

Proof. Applying Theorem 3.3.2 to $A_1 = 0$ and $A_2 = M$ shows that

$$\mathbb{P}[\sigma_{n-1}(M + \gamma G_n) < t] \leq \mathbb{P}[\sigma_{n-1}(\gamma G_n) < t] = \mathbb{P}[\sigma_{n-1}(G_n) < t/\gamma].$$

Invoking Theorem 3.3.1 with $j = n - 1$ and α replaced by $tn/2\gamma$ yields the claim. □

We will need as well the main theorem of Chapter 2, which shows that the addition of a small complex Ginibre to an arbitrary matrix tames its eigenvalue condition numbers. We restate this below:

Theorem 3.3.4 ([11, Theorem 1.5])**.** Suppose $A \in \mathbb{C}^{n \times n}$ with $\|A\| \leq 1$ and $\delta \in (0, 1)$. Let G_n be a complex Ginibre matrix, and let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be the (random) eigenvalues of $A + \delta G_n$. Then for every measurable open set $B \subset \mathbb{C}$,

$$\mathbb{E} \sum_{\lambda_i \in B} \kappa(\lambda_i)^2 \leq \frac{n^2}{\pi \delta^2} \mathrm{Leb}(B).$$

Our final lemma before embarking on the proof in earnest shows that bounds on the j-th smallest singular value and eigenvector condition number are sufficient to rule out the presence of j eigenvalues in a small region. For our particular application we will take $j = 2$. .

TODO: improve this and ... p r o p a g a t e ?

Lemma 3.3.5. Let $D(z_0, r) := \{z \in \mathbb{C} : |z - z_0| < r\}$. If $M \in \mathbb{C}^{n\times n}$ is a diagonalizable matrix with at least j eigenvalues in $D(z_0, r)$ then

$$\sigma_{n-j+1}(z_0 - M) \le r\kappa_V(M).$$

Proof. Write $M = VDV^{-1}$. By Courant-Fischer:

$$\begin{aligned}
\sigma_{n-j+1}(z_0 - M) &= \min_{S:\dim(S)=j} \max_{x\in S\setminus\{0\}} \frac{\|V(z_0 - D)V^{-1}x\|}{\|x\|} \\
&= \min_{S:\dim(S)=j} \max_{y\in V(S)\setminus\{0\}} \frac{\|V(z_0 - D)y\|}{\|Vy\|} && \text{setting } y = Vx \\
&= \min_{S:\dim(S)=j} \max_{y\in S\setminus\{0\}} \frac{\|V(z_0 - D)y\|}{\|Vy\|} && \text{since } V \text{ is invertible} \\
&\le \min_{S:\dim(S)=j} \max_{y\in S\setminus\{0\}} \frac{\|V\|\|(z_0 - D)y\|}{\sigma_n(V)\|y\|} \\
&\le \kappa_V(M)\sigma_{n-j+1}(z_0 - D).
\end{aligned}$$

Since $z_0 - D$ is diagonal its singular values are just $|z_0 - \lambda_i|$, so the j-th smallest is at most r, finishing the proof. □

We now present the main tail bound that we use to control the minimum gap and eigenvector condition number.

Theorem 3.3.6 (Multiparameter Tail Bound)**.** Let $A \in \mathbb{C}^{n\times n}$. Assume $\|A\| \le 1$ and $\gamma < 1/2$, and let $X := A + \gamma G_n$ where G_n is a complex Ginibre matrix. For every $t, r > 0$:

$$\mathbb{P}[\kappa_V(X) < t,\ \mathrm{gap}(X) > r,\ \|G_n\| < 4] \ge 1 - \left(\frac{144}{r^2} \cdot 4(trn/\gamma)^8 + (9n^2/\gamma^2 t^2) + 2e^{-2n}\right). \tag{3.12}$$

Proof. Write $\Lambda(X) := \{\lambda_1, \ldots, \lambda_n\}$ for the (random) eigenvalues of $X := A + \gamma G_n$, in increasing order of magnitude (there are no ties almost surely). Let $\mathcal{N} \subset \mathbb{C}$ be a minimal $r/2$-net of $B := D(0, 3)$, recalling the standard fact that one exists of size no more than $(3 \cdot 4/r)^2 = 144r^2$. The most useful feature of such a net is that, by the triangle inequality, for any $a, b \in D(0, 3)$ with distance at most r, there is a point $y \in \mathcal{N}$ with $|y - (a + b)/2| < r/2$ satisfying $a, b \in D(y, r)$. In particular, if $\mathrm{gap}(X) < r$, then there are two eigenvalues in the disk of radius r centered at some point $y \in \mathcal{N}$.

Therefore, consider the events

$$\begin{aligned}
E_{\mathrm{gap}} &:= \{\mathrm{gap}(X) < r\} \subset \{\exists y \in \mathcal{N} : |D(y, r) \cap \Lambda(X)| \ge 2\} \\
E_D &:= \{\Lambda(X) \not\subseteq D(0, 3)\} \subset \{\|G_n\| \ge 4\} := E_G \\
E_\kappa &:= \{\kappa_V(X) > t\} \\
E_y &:= \{\sigma_{n-1}(y - X) < rt\}, \qquad y \in \mathcal{N}.
\end{aligned}$$

Lemma 3.3.5 applied to each $y \in \mathcal{N}$ with $j = 2$ reveals that

$$E_{\text{gap}} \subseteq E_D \cup E_\kappa \cup \bigcup_{y \in \mathcal{N}} E_y,$$

whence

$$E_{\text{gap}} \cup E_\kappa \subseteq E_D \cup E_\kappa \cup \bigcup_{y \in \mathcal{N}} E_y.$$

By a union bound, we have

$$\mathbb{P}[E_{\text{gap}} \cup E_\kappa] \leq \mathbb{P}[E_D \cup E_\kappa] + |\mathcal{N}| \max_{y \in \mathcal{N}} \mathbb{P}[E_y]. \tag{3.13}$$

From the tail bound on the operator norm of a Ginibre matrix in [11, Lemma 2.2],

$$\mathbb{P}[E_D] \leq \mathbb{P}[E_G] \leq 2e^{-(4-2\sqrt{2})^2 n} \leq 2e^{-2n}. \tag{3.14}$$

Observe that by (1.1),

$$\left\{ \kappa_V(X) > \sqrt{n \sum_{\lambda_i \in D(0,3)} \kappa(\lambda_i)^2} \right\} \subset E_D,$$

which implies that

$$E_\kappa \subset E_D \cup \left\{ \sum_{\lambda_i \in D(0,3)} \kappa(\lambda_i)^2 > t^2/n \right\}.$$

Theorem 3.3.4 and Markov's inequality yields

$$\mathbb{P}\left[\sum_{\lambda_i \in D(0,3)} \kappa(\lambda_i)^2 > t^2/n \right] \leq \frac{9n^2}{\gamma^2} \frac{n}{t^2} = \frac{9n^3}{t^2 \gamma^2}.$$

Thus, we have

$$\mathbb{P}[E_\kappa \cup E_G] \leq \frac{9n^3}{t^2\gamma^2} + 2e^{-2n}.$$

Corollary 3.3.3 applied to $M = -y + A$ gives the bound

$$\mathbb{P}[E_y] \leq 4 \left(\frac{trn}{\gamma} \right)^8,$$

for each $y \in \mathcal{N}$, and plugging these estimates back into (3.13) we have

$$\mathbb{P}[E_{\text{gap}} \cup E_\kappa \cup E_G] \leq \frac{144}{r^2} \cdot 4 \left(\frac{trn}{\gamma} \right)^8 + \frac{9n^2}{\gamma^2 t^2} + 2e^{-2n},$$

as desired. □

A specific setting of parameters in Theorem 3.3.6 immediately yields Theorem 3.1.4.

Proof of Theorem 3.1.4. Applying Theorem 3.3.6 with parameters $t := \frac{n^2}{\gamma}$ and $r := \frac{\gamma^4}{n^5}$, we have

$$\mathbb{P}[\mathrm{gap}(X) > r,\ \kappa_V(X) < t,\ \Lambda(X) \subset D(0,3)] \geq 1 - 600\frac{n^{10}}{\gamma^8}\left(\frac{\gamma^2}{n^2}\right)^8 - \frac{9}{n^2} - 2e^{-2n} \geq 1 - O(n^{-2}), \tag{3.15}$$

as desired. □

Since it is of independent interest in random matrix theory, we record the best bound on the gap alone that is possible to extract from the theorem above.

Corollary 3.3.7 (Minimum Gap Bound)**.** For X as in Theorem 3.3.6,

$$\mathbb{P}[\mathrm{gap}(X) < r] \leq 2 \cdot 9^{4/5}(144 \cdot 4)^{1/5}(n/\gamma)^{2+6/5} r^{6/5} \leq 42(n/\gamma)^{3.2} r^{1.2} + 2e^{-2n}.$$

In particular, the probability is $o(1)$ if $r = o((\gamma/n)^{8/3})$.

Proof. Setting

$$t^{10} = \frac{9}{144 \cdot 4}(\gamma/nr)^6$$

in Theorem 3.3.6 balances the first two terms and yields the advertised bound. □

3.3.2 Shattering

Propositions 1.1.6 and 1.1.7 in the preliminaries together tell us that if the ϵ-pseudospectrum of an $n \times n$ matrix M has n connected components, then each eigenvalue of any size-ϵ perturbation $\widetilde{M}$ will lie in its own connected component of $\Lambda_\epsilon(M)$. The following key definitions make this phenomenon quantitative in a sense which is useful for our analysis of spectral bisection.

Definition 3.3.8 (Grid)**.** A *grid* in the complex plane consists of the boundaries of a lattice of squares with lower edges parallel to the real axis. We will write

$$\mathsf{grid}(z_0, \omega, s_1, s_2) \subset \mathbb{C}$$

to denote an $s_1 \times s_2$ grid of $\omega \times \omega$-sized squares and lower left corner at $z_0 \in \mathbb{C}$. Write $\mathrm{diam}(\mathsf{g}) := \omega\sqrt{s_1^2 + s_2^2}$for the diameter of the grid.

Definition 3.3.9 (Shattering)**.** A pseudospectrum $\Lambda_\epsilon(A)$ is *shattered* with respect to a grid g if:

1. Every square of g has at most one eigenvalue of A.
2. $\Lambda_\varepsilon(A) \cap \mathsf{g} = \emptyset$.

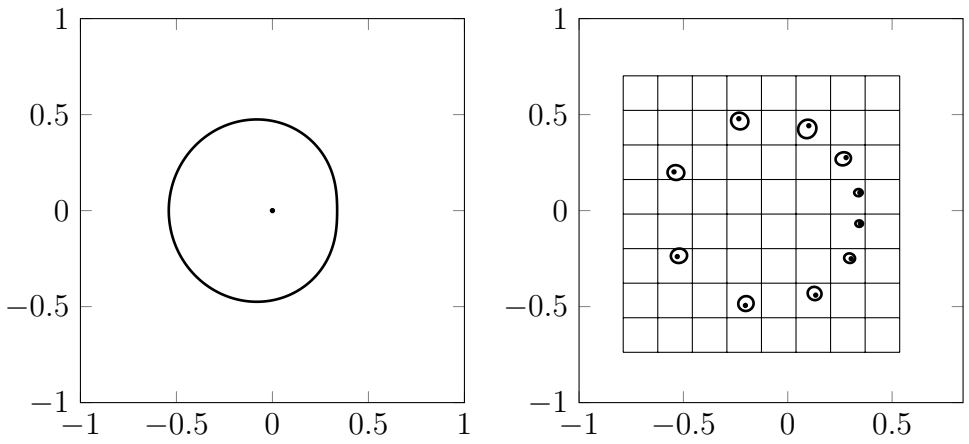

Figure 3.1: T is a sample of an upper triangular 10×10 Toeplitz matrix with zeros on the diagonal and an independent standard real Gaussian repeated along each diagonal above the main diagonal. G is a sample of a 10×10 complex Ginibre matrix with unit variance entries. Using the MATLAB package EigTool [133], the boundaries of the ε-pseudospectrum of T (left) and $T + 10^{-6}G$ (right) for $\varepsilon = 10^{-6}$ are plotted along with the spectra. The latter pseudospectrum is shattered with respect to the pictured grid.

Observation 3.3.10. As $\Lambda_\epsilon(A)$ contains a ball of radius ϵ about each eigenvalue of A, shattering of the ϵ-pseudospectrum with respect to a grid with side length ω implies $\epsilon \leq \omega/2$.

As a warm-up for more sophisticated arguments later on, we give here an easy consequence of the shattering property.

Lemma 3.3.11. If $\Lambda_\epsilon(M)$ is shattered with respect to a grid g with side length ω, then every eigenvalue condition number satisfies $\kappa_i(M) \leq \frac{2\omega}{\pi\epsilon}$.

Proof. Let v, w^* be a right/left eigenvector pair for some eigenvalue λ_i of M, normalized so that $w^*v = 1$. Letting Γ be the positively oriented boundary of the square of g containing λ_i, we can extract the projector vw^* by integrating, and pass norms inside the contour integral to obtain

$$\kappa_i(A) = \|vw^*\| = \left\| \frac{1}{2\pi i} \oint_\Gamma (z - M)^{-1} \mathrm{d}z \right\| \leq \frac{1}{2\pi} \oint_\Gamma \left\| (z - M)^{-1} \right\| \mathrm{d}z \leq \frac{2\omega}{\pi\epsilon}. \tag{3.16}$$

In the final step we have used the fact that, given the definition of pseudospectrum (1.3) above, $\Lambda_\epsilon(M) \cap \mathsf{g} = \emptyset$ means $\|(z - M)^{-1}\| \leq 1/\epsilon$ on g. □

The theorem below quantifies the extent to which perturbing by a Ginibre matrix results in a shattered pseudospectrum. See Figure 3.1 for an illustration in the case where the initial matrix is poorly conditioned. In general, not all eigenvalues need move so far upon such a perturbation, in particular if the respective κ_i are small.

Theorem 3.3.12 (Exact Arithmetic Shattering)**.** Let $A \in \mathbb{C}^{n\times n}$ and $X := A + \gamma G_n$ for G_n a complex Ginibre matrix. Assume $\|A\| \le 1$ and $0 < \gamma < 1/2$. Let $\mathsf{g} := \mathsf{grid}(z, \omega, \lceil 8/\omega \rceil, \lceil 8/\omega \rceil)$ with $\omega := \frac{\gamma^4}{4n^5}$, and z chosen uniformly at random from the square of side ω cornered at $-4-4i$. Then, $\kappa_V(X) \le n^2/\gamma$, $\|A - X\| \le 2\gamma$, and $\Lambda_\varepsilon(X)$ is shattered with respect to g for

$$\epsilon := \frac{\gamma^5}{16n^9},$$

with probability at least $1 - 1/n - O(1/n^2)$ where the implied constant is at most 600.

Proof. Condition on the event in Theorem 3.1.4, so that

$$\kappa_V(X) \le \frac{n^2}{\gamma}, \quad \|X - A\| \le 4\gamma, \quad \text{and } \operatorname{gap}(X) \ge \frac{\gamma^4}{n^5} = 4\omega.$$

Consider the random grid g. Since $D(0,3)$ is contained in the square of side length 8 centered at the origin, every eigenvalue of X is contained in one square of g with probability 1. Moreover, since $\operatorname{gap}(X) > 4\omega$, no square can contain two eigenvalues. Let

$$\operatorname{dist}_{\mathsf{g}}(z) := \min_{y\in\mathsf{g}} |z - y|.$$

Let $\lambda_i := \lambda_i(X)$. We now have for each λ_i and every $s < \frac{\omega}{2}$:

$$\mathbb{P}[\operatorname{dist}_{\mathsf{g}}(\lambda_i) > s] = \frac{(\omega - 2s)^2}{\omega^2} = 1 - \frac{4s}{\omega} + \frac{4s^2}{\omega^2} \ge 1 - \frac{4s}{\omega},$$

since the distribution of λ_i inside its square is uniform with respect to Lebesgue measure. Setting $s = \omega/4n^2$, this probability is at least $1 - 1/n^2$, so by a union bound

$$\mathbb{P}[\min_{i\le n} \operatorname{dist}_{\mathsf{g}}(\lambda_i) > \omega/4n^2] > 1 - 1/n, \tag{3.17}$$

i.e., every eigenvalue is well-separated from g with probability $1 - 1/n$.

We now recall from (1.5) that

$$\Lambda_\varepsilon(X) \subset \bigcup_{i\le n} D(\lambda_i, \kappa_V(X)\epsilon).$$

Thus, on the events (3.15) and (3.17), we see that $\Lambda_\varepsilon(X)$ is shattered with respect to g as long as

$$\kappa_V(X)\epsilon < \frac{\omega}{4n^2},$$

which is implied by

$$\epsilon < \frac{\gamma^4}{4n^5} \cdot \frac{1}{4n^2} \cdot \frac{\gamma}{n^2} = \frac{\gamma^5}{16n^9}.$$

Thus, the advertised claim holds with probability at least

$$1 - \frac{1}{n} - O(1/n^2),$$

as desired. □

Finally, we show that the shattering property is retained when the Gaussian perturbation is added in finite precision rather than exactly. This also serves as a pedagogical warmup for our presentation of more complicated algorithms later in this chapter: we use E to represent an adversarial roundoff error (as in step 2), and for simplicity neglect roundoff error completely in computations whose size does not grow with n (such as steps 3 and 4, which set scalar parameters).

SHATTER

Input: Matrix $A \in \mathbb{C}^{n \times n}$, Gaussian perturbation size $\gamma \in (0, 1/2)$.
Requires: $\|A\| \leq 1$.
Algorithm: $(X, \epsilon) = \mathsf{SHATTER}(A, \gamma)$

1. $G_{ij} \leftarrow \mathsf{N}(1/n)$ for $i, j = 1, \ldots, n$.
2. $X \leftarrow A + \gamma G + E$.
3. Let g be a random grid with $\omega = \frac{\gamma^4}{4n^5}$ and bottom left corner z chosen as in Theorem 3.3.12.
4. $\epsilon \leftarrow \frac{1}{2} \cdot \frac{\gamma^5}{16n^9}$

Output: Matrix $X \in \mathbb{C}^{n \times n}$, grid g, shattering parameter $\epsilon > 0$.
Ensures: $\|X - A\| \leq 4\gamma$, $\kappa_V(X) \leq n^2/\gamma$, and $\Lambda_\epsilon(X)$ is shattered with respect to g, with probability at least $1 - 1/n - O(1/n^2)$.

Theorem 3.3.13 (Finite Arithmetic Shattering)**.** Assume there is a c_{N}-stable Gaussian sampling algorithm N satisfying the requirements of Definition 3.2.1. Then SHATTER has the advertised guarantees as long as the machine precision satisfies

$$\mathbf{u} \leq \frac{1}{2} \frac{\gamma^5}{16n^9} \cdot \frac{1}{(3 + c_{\mathsf{N}})\sqrt{n}}, \tag{3.18}$$

and runs in

$$n^2 T_{\mathsf{N}} + n^2 = O(n^2)$$

arithmetic operations.

Proof. The two sources of error in SHATTER are:

1. An additive error of operator norm at most $n \cdot c_{\mathsf{N}} \cdot (1/\sqrt{n}) \cdot \mathbf{u} \leq \cdot c_{\mathsf{N}} \sqrt{n} \mathbf{u}$ from N, by Definition 3.2.1.
2. An additive error of norm at most $\sqrt{n} \cdot \|X\| \cdot \mathbf{u} \leq 3\sqrt{n}\mathbf{u}$, with probability at least $1 - 1/n$, from the roundoff E in step 2.

Thus, as long as the precision satisfies (3.18), we have

$$\|\mathsf{SHATTER}(A,\gamma) - \mathrm{shatter}(A,\gamma)\| \leq \frac{1}{2}\frac{\gamma^5}{16n^9},$$

where shatter$(\cdot)$ refers to the (exact arithmetic) outcome of Theorem 3.3.12. The correctness of SHATTER now follows from Proposition 1.1.6. Its running time is bounded by

$$n^2 T_{\mathsf{N}} + n^2$$

arithmetic operations, as advertised. □

3.4 Matrix Sign Function

The algortithmic centerpiece of this work is the analysis, in finite arithmetic, of a well-known iterative method for approximating to the matrix sign function. Recall from Section 3.1 that if A is a matrix whose spectrum avoids the imaginary axis, then

$$\mathrm{sgn}(A) = P_+ - P_-$$

where the latter two are the spectral projectors corresponding to eigenvalues in the open right and left half-planes respectively. The iterative algorithm we consider approximates the matrix sign function by repeated application to A of the function

$$g(z) := \frac{1}{2}(z + z^{-1}) \tag{3.19}$$

This is simply Newton's method to find a root of $z^2 - 1$, but one can verify that the function fixes the left and right halfplanes, and thus we should expect it to push eigenvalues in the former towards -1, and in the latter towards $+1$.

In Subsection 3.4.1 we briefly discuss the specific preliminaries that will be used throughout this section. In Subsection 3.4.2 we give a *pseudospectral* proof of the rapid global convergence of SGN when implemented in exact arithmetic. In Subsection 3.4.3 we show that the proof provided in Subsection 3.4.3 is robust enough to handle the finite arithmetic case; a formal statement of this main result is the content of Theorem 3.4.9.

3.4.1 Circles of Apollonius

It has been known since antiquity that a circle in the plane may be described as the set of points a fixed ratio of distances to two focal points. By choosing two such points and varying the ratio in question, we get a family of circles named for the Greek geometer Apollonius of Perga. We will exploit several interesting properties enjoyed by these *Circles of Apollonius* in the analysis below.

SGN

Input: Matrix $A \in \mathbb{C}^{n\times n}$, pseudospectral guarantee ϵ, circle parameter α, and desired accuracy δ
Requires: $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$.
Algorithm: $S = \mathsf{SGN}(A, \epsilon, \alpha, \delta)$

1. $N \leftarrow \lceil \lg(1/(1-\alpha)) + 3\lg\lg(1/(1-\alpha)) + \lg\lg(1/(\beta\varepsilon)) + 7.59 \rceil$
2. $A_0 \leftarrow A$
3. For $k = 1, ..., N$,
   a) $A_k \leftarrow \frac{1}{2}(A_{k-1} + A_{k-1}^{-1}) + E_k$
4. $S \leftarrow A_N$

Output: Approximate matrix sign function S
Ensures: $\|S - \mathrm{sgn}(A)\| \leq \delta$

More precisely, we analyze the Newton iteration map g in terms of the family of Apollonian circles whose foci are the points $\pm 1 \in \mathbb{C}$. For the remainder of this section we will write $m(z) = \frac{1-z}{1+z}$ for the Möbius transformation taking the right half-plane to the unit disk, and for each $\alpha \in (0,1)$ we denote by

$$\mathsf{C}_\alpha^+ = \{z \in \mathbb{C} : |m(z)| \leq \alpha\}, \quad \mathsf{C}_\alpha^- = \{z \in \mathbb{C} : |m(z)|^{-1} \leq \alpha\}$$

the closed region in the right (respectively left) half-plane bounded by such a circle. Write $\partial\mathsf{C}_\alpha^-$ for their boundaries, and $\mathsf{C}_\alpha = \mathsf{C}_\alpha^+ \cup \mathsf{C}_\alpha^-$ for their union.

The region C_α^+ is a disk centered at $\frac{1+\alpha^2}{1-\alpha^2} \in \mathbb{R}$, with radius $\frac{2\alpha}{1-\alpha^2}$, and whose intersection with the real line is the interval $(m(\alpha), m(\alpha)^{-1})$; C_α^- can be obtained by reflecting C_α^+ with respect to the imaginary axis. For $\alpha > \beta > 0$, we will write

$$\mathsf{A}_{\alpha,\beta}^+ = \mathsf{C}_\alpha^+ \setminus \mathsf{C}_\beta^+$$

for the *Apollonian annulus* lying inside C_α^+ and outside C_β^+; note that in our notation this set does not include $\partial\mathsf{C}_\beta^+$. In the same way define $\mathsf{A}_{\alpha,\beta}^-$ for the left half-plane and write $\mathsf{A}_{\alpha,\beta} = \mathsf{A}_{\alpha,\beta}^+ \cup \mathsf{A}_{\alpha,\beta}^-$.

Observation 3.4.1 ([105])**.** The Newton map g is a two-to-one map from C_α^+ to $\mathsf{C}_{\alpha^2}^+$, and a two-to-one map from C_α^- to $\mathsf{C}_{\alpha^2}^-$.

Proof. This follows from the fact that for each z in the right half-plane,

$$|m(g(z))| = \left|\frac{1 - \frac{1}{2}(z + 1/z)}{1 + \frac{1}{2}(z + 1/z)}\right| = \left|\frac{(1-z)^2}{(z+1)^2}\right| = |m(z)|^2$$

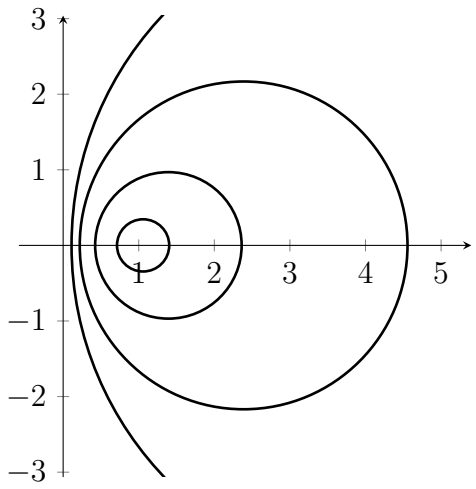

Figure 3.2: Apollonian circles appearing in the analysis of the Newton iteration. Depicted are $\mathsf{C}^+_{\alpha^{2^k}}$ for $\alpha = 0.8$ and $k = 0, 1, 2, 3$.

and similarly for the left half-plane. □

It follows from Observation 3.4.1 that under repeated application of the Newton map g, any point in the right or left half-plane converges to $+1$ or -1, respectively.

3.4.2 Exact Arithmetic

Here we denote by $A_0 = A$ and $A_{k+1} = g(A_k)$. In the case of exact arithmetic, Observation 3.4.1 implies global convergence of the Newton iteration when A is diagonalizable. For the convenience of the reader we provide this argument (due to [105]) below.

Proposition 3.4.2. Let A be a diagonalizable matrix and assume that $\Lambda(A) \subset \mathsf{C}_\alpha$ for some $\alpha \in (0, 1)$. Then for every $N \in \mathbb{N}$ we have the guarantee

$$\|A_N - \operatorname{sgn}(A)\| \leq \frac{4\alpha^{2^N}}{\alpha^{2^{N+1}} + 1} \cdot \kappa_V(A).$$

Moreover, when A does not have eigenvalues on the imaginary axis the minimum α for which $\Lambda(A) \subset \mathsf{C}_\alpha$ is given by

$$\alpha^2 = \max_i \left\{ 1 - \frac{4|\Re(\lambda_i(A))|}{|\lambda_i(A) - \operatorname{sgn}(A)|^2} \right\}$$

Proof. Consider the spectral decomposition $A = \sum_{i=1}^n \lambda_i v_i w_i^*$, and denote by $\lambda_i^{(N)}$ the eigenvalues of A_N.

By Observation 3.4.1 we have that $\Lambda(A_N) \subset \mathsf{C}_{\alpha^{2^N}}$ and $\operatorname{sgn}(\lambda_i) = \operatorname{sgn}(\lambda_i^{(N)})$. Moreover, A_N and $\operatorname{sgn}(A)$ have the same eigenvectors. Hence

$$\|A_N - \operatorname{sgn}(A)\| \leq \left\| \sum_{\Re(\lambda_i)>0}^{n} (\lambda_i^{(N)} - 1) v_i w_i^* \right\| + \left\| \sum_{\Re(\lambda_i)<0} (\lambda_i^{(N)} + 1) v_i w_i^* \right\|. \tag{3.20}$$

Now we will use that for any matrix X we have that $\|X\| \leq \kappa_V(X)\mathsf{spr}(X)$ where $\mathsf{spr}(X)$ denotes the spectral radius of X. Observe that the spectral radii of the two matrices appearing on the right hand side of (3.20) are bounded by $\max_i |\lambda_i - \operatorname{sgn}(\lambda_i)|$, which in turn is bounded by the radius of the circle $\mathsf{C}^+_{\alpha^{2^N}}$, namely $2\alpha^{2^N}/(\alpha^{2^{N+1}}+1)$. On the other hand, the eigenvector condition number of these matrices is bounded by $\kappa_V(A)$. This concludes the first part of the statement.

In order to compute α note that if $z = x + iy$ with $x > 0$, then

$$|m(z)|^2 = \frac{(1-x)^2 + y^2}{(1+x)^2 + y^2} = 1 - \frac{4x}{(1+x)^2 + y^2},$$

and analogously when $x < 0$ and we evaluate $|m(z)|^{-2}$. □

The above analysis becomes useless when trying to prove the same statement in the framework of finite arithmetic. This is due to the fact that at each step of the iteration the roundoff error can make the eigenvector condition numbers of the A_k grow. In fact, since $\kappa_V(A_k)$ is sensitive to infinitesimal perturbations whenever A_k has a multiple eigenvalue, it seems difficult to control it against adversarial perturbations as the iteration converges to $\operatorname{sgn}(A_k)$ (which has very high multiplicity eigenvalues). A different approach [105] yields a proof of convergence in exact arithmetic even when A is not diagonalizable. However, this proof relies heavily on the fact that $m(A_N)$ is an exact power of $m(A_0)$, or more precisely, it requires the sequence A_k to have the same generalized eigenvectors, which is again not the case in the finite arithmetic setting.

Therefore, a *robust* version, tolerant to perturbations, of the above proof is needed. To this end, instead of simultaneously keeping track of the eigenvector condition number and the spectrum of the matrices A_k, we will just show that the ϵ_k-pseudospectra of these matrices are contained in a certain shrinking region dependent on k. This invariant is inherently robust to perturbations smaller than ϵ_k, unaffected by clustering of eigenvalues due to convergence, and allows us to bound the accuracy and other quantities of interest via the functional calculus. The following lemma shows how to obtain a bound on $\|A_N - \operatorname{sgn}(A)\|$ solely using information from the pseudospectrum.

Lemma 3.4.3 (Pseudospectral Error Bound)**.** Assume that $\varepsilon_N > 0$ and $\alpha_N \in (0,1)$ satisfy $\Lambda_{\epsilon_N}(A_N) \subset \mathsf{C}_{\alpha_N}$. Then we have the guarantee

$$\|A_N - \operatorname{sgn}(A)\| \leq \frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\varepsilon_N}. \tag{3.21}$$

Proof. Note that $\operatorname{sgn}(A) = \operatorname{sgn}(A_N)$. Using the functional calculus we get

$$\begin{aligned}
\|A_N - \operatorname{sgn}(A_N)\| &= \left\| \frac{1}{2\pi i} \oint_{\partial \mathsf{C}^+_{\alpha_N}} z(z-A_n)^{-1} - (z-A_n)^{-1}\, dz \right. \\
&\qquad \left. + \frac{1}{2\pi i} \oint_{\partial \mathsf{C}^-_{\alpha_N}} z(z-A_n)^{-1} + (z-A_n)^{-1}\, dz \right\| \\
&\leq \frac{\ell(\partial \mathsf{C}^+_{\alpha_N})}{\pi} \sup\{|z-1| : z \in \mathsf{C}^+_{\alpha_N}\} \frac{1}{\varepsilon_N} \quad \text{by the triangle inequality and symmetry} \\
&= \frac{4\alpha_N}{1-\alpha_N^2} \left(\frac{1+\alpha_N}{1-\alpha_N} - 1 \right) \frac{1}{\varepsilon_N} \\
&= \frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\varepsilon_N}.
\end{aligned}$$

□

The key feature of (3.21) is that the accuracy depends on the *square* of the circle parameter α. In view of this bound it is now enough to find sequences α_k and ϵ_k such that

$$\Lambda_{\epsilon_k}(A_k) \subset \mathsf{C}^+_{\alpha_k}$$

and α_k^2/ϵ_k converges quadratically to zero, which will follow if ϵ_k shrinks at roughly the same rate as α_k. The lemma below is instrumental in determining such sequences.

Lemma 3.4.4 (Key Lemma)**.** If $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, then for every $\alpha' > \alpha^2$, we have $\Lambda_{\epsilon'}(g(A)) \subset \mathsf{C}_{\alpha'}$ where

$$\epsilon' := \epsilon \frac{(\alpha' - \alpha^2)(1-\alpha^2)}{8\alpha}.$$

Proof. From the definition of pseudospectrum, our hypothesis implies $\|(z-A)^{-1}\| < 1/\epsilon$ for every z outside of $\mathsf{C}^+_\alpha \cup \mathsf{C}^-_\alpha$. The proof will hinge on the observation that, for each $\alpha' \in (\alpha^2, \alpha)$, this resolvent bound allows us to bound the resolvent of $g(A)$ everywhere in the Appolonian annuli $\mathsf{A}_{\alpha,\alpha'}$.

Let $w \in \mathsf{A}_{\alpha,\alpha'}$. We must show that $w \notin \Lambda_{\varepsilon'}(g(A))$. Since $w \notin \mathsf{C}_{\alpha^2}$, Observation 3.4.1 ensures no $z \in \mathsf{C}_\alpha$ satisfies $g(z) = w$; in other words, the function $(w - g(z))^{-1}$ is holomorphic in z on C_α. As $\Lambda(A) \subset \Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, Observation 3.4.1 also guarantees that $\Lambda(g(A)) \subset \mathsf{C}_{\alpha^2}$. Thus for w in the union of the two Appolonian annuli in question, we can calculate the resolvent of $g(A)$ at w using the holomorphic functional calculus:

$$(w - g(A))^{-1} = \frac{1}{2\pi i} \oint_{\partial \mathsf{C}_\alpha} (w - g(z))^{-1} (z - A)^{-1} \mathrm{d}z,$$

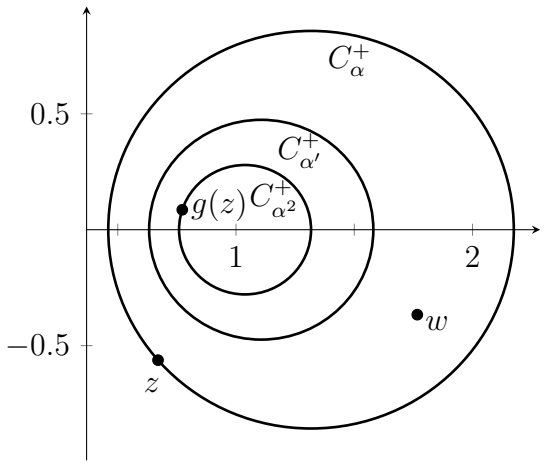

Figure 3.3: Illustration of the proof of Lemma 3.4.4

where by this we mean to sum the integrals over $\partial \mathsf{C}_\alpha^+$ and $\partial \mathsf{C}_\alpha^-$, both positively oriented. Taking norms, passing inside the integral, and applying Observation 3.4.1 one final time, we get:

$$\begin{aligned}\left\|(w - g(A))^{-1}\right\| &\leq \frac{1}{2\pi} \oint_{\partial \mathsf{C}_\alpha} |(w - g(z))^{-1}| \cdot \|(z - A)^{-1}\| \mathrm{d}z \\ &\leq \frac{\ell\left(\partial \mathsf{C}_\alpha^+\right) \sup_{y \in \mathsf{C}_{\alpha^2}^+} |(w-y)^{-1}| + \ell((\partial \mathsf{C}_\alpha^-) \sup_{y \in \mathsf{C}_{\alpha^2}^-} |(w-y)^{-1}|}{2\pi\epsilon} \\ &\leq \frac{1}{\epsilon} \frac{8\alpha}{(\alpha' - \alpha^2)(1 - \alpha^2)}.\end{aligned}$$

In the last step we also use the forthcoming Lemma 3.4.5. Thus, with ϵ' defined as in the theorem statement, $\mathsf{A}_{\alpha,\alpha'}$ contains none of the ϵ'-pseudospectrum of $g(A)$. Since $\Lambda(g(A)) \subset \mathsf{C}_{\alpha^2}$, Theorem 1.1.7 tells us that there can be no ϵ'-pseudospectrum in the remainder of $\mathbb{C} \setminus \mathsf{C}_{\alpha'}$, as such a connected component would need to contain an eigenvalue of $g(A)$. □

Lemma 3.4.5. Let $1 > \alpha, \beta > 0$ be given. Then for any $x \in \partial \mathsf{C}_\alpha$ and $y \in \partial \mathsf{C}_\beta$, we have $|x - y| \geq (\alpha - \beta)/2$.

Proof. Without loss of generality $x \in \partial \mathsf{C}_\alpha^+$ and $y \in \partial \mathsf{C}_\beta^+$. Then we have

$$|\alpha - \beta| = |m(x) - m(y)| = \frac{2|x - y|}{|1 + x||1 + y|} \leq 2|x - y|.$$

□

Lemma 3.4.4 will also be useful in bounding the condition numbers of the A_k, which is necessary for the finite arithmetic analysis.

Corollary 3.4.6 (Condition Number Bound)**.** Using the notation of Lemma 3.4.4, if $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, then

$$\|A^{-1}\| \leq \frac{1}{\epsilon} \quad \text{and} \quad \|A\| \leq \frac{4\alpha}{(1-\alpha)^2\varepsilon}.$$

Proof. The bound $\|A^{-1}\| \leq 1/\varepsilon$ follows from the fact that $0 \notin \mathsf{C}_\alpha \supset \Lambda_\varepsilon(A)$. In order to bound A we use the contour integral bound

$$\begin{aligned}\|A\| &= \left\| \frac{1}{2\pi i} \oint_{\partial \mathsf{C}_\alpha} z(z-A)^{-1}\, dz \right\| \\ &\leq \frac{\ell(\partial \mathsf{C}_\alpha)}{2\pi} \left(\sup_{z \in \partial \mathsf{C}_\alpha} |z| \right) \frac{1}{\varepsilon} \\ &= \frac{4\alpha}{1-\alpha^2} \frac{1+\alpha}{1-\alpha} \frac{1}{\varepsilon}.\end{aligned}$$

□

Another direct application of Lemma 3.4.4 yields the following.

Lemma 3.4.7. Let $\epsilon > 0$. If $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, and $1/\alpha > D > 1$ then for every N we have the guarantee

$$\Lambda_{\epsilon_N}(A_N) \subset \mathsf{C}_{\alpha_N},$$

for $\alpha_N = (D\alpha)^{2^N}/D$ and $\epsilon_N = \frac{\alpha_N \epsilon}{\alpha} \left(\frac{(D-1)(1-\alpha^2)}{8D} \right)^N$.

Proof. Define recursively $\alpha_0 = \alpha$, $\epsilon_0 = \epsilon$, $\alpha_{k+1} = D\alpha_k^2$ and $\epsilon_{k+1} = \frac{1}{8}\epsilon_k \alpha_k (D-1)(1-\alpha_0^2)$. It is easy to see by induction that this definition is consistent with the definition of α_N and ϵ_N given in the statement.

We will now show by induction that $\Lambda_{\epsilon_k}(A_k) \subset \mathsf{C}_{\alpha_k}$. Assume the statement is true for k, so from Lemma 3.4.4 we have that the statement is also true for A_{k+1} if we pick the pseudospectral parameter to be

$$\epsilon' = \epsilon_k \frac{(\alpha_{k+1} - \alpha_k^2)(1-\alpha_k^2)}{8\alpha_k} = \frac{1}{8}\epsilon_k \alpha_k (D-1)(1-\alpha_k^2).$$

On the other hand

$$\frac{1}{8}\epsilon_k \alpha_k (D-1)(1-\alpha_k^2) \geq \frac{1}{8}\epsilon_k \alpha_k (D-1)(1-\alpha_0^2) = \epsilon_{k+1},$$

which concludes the proof of the first statement. □

We are now ready to prove a pseudospectral version of Proposition 3.4.2.

Proposition 3.4.8. Let $A \in \mathbb{C}^{n\times n}$ be a diagonalizable matrix and assume that $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$ for some $\alpha \in (0,1)$. Then, for any $1 < D < \frac{1}{\alpha}$ for every N we have the guarantee

$$\|A_N - \operatorname{sgn}(A)\| \le (D\alpha)^{2^N} \cdot \frac{\pi\alpha(1-\alpha^2)^2}{8\epsilon} \cdot \left(\frac{8D}{(D-1)(1-\alpha^2)}\right)^{N+2}.$$

Proof. Using the choice of α_k and ϵ_k given in the proof of Lemma 3.4.7 and the bound (3.21), we get that

$$\begin{aligned}
\|A_N - \operatorname{sgn}(A)\| &\le \frac{8\pi\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\varepsilon_N} \\
&= \frac{8\pi\alpha_0\alpha_N}{\epsilon_0(1-\alpha_N)^2(1+\alpha_N)}\left(\frac{8D}{(D-1)(1-\alpha_0^2)}\right)^N \\
&= (D\alpha_0)^{2^N}\frac{8D^3\pi\alpha_0}{(D-(D\alpha_0)^{2^N})^2(D+(D\alpha_0)^{2^N})\epsilon_0}\left(\frac{8D}{(D-1)(1-\alpha_0^2)}\right)^N \\
&\le (D\alpha_0)^{2^N}\frac{8D^2\pi\alpha_0}{(D-1)^2\epsilon_0}\left(\frac{8D}{(D-1)(1-\alpha_0^2)}\right)^N \\
&= (D\alpha_0)^{2^N}\frac{\pi\alpha_0(1-\alpha_0^2)^2}{8\epsilon_0}\left(\frac{8D}{(D-1)(1-\alpha_0^2)}\right)^{N+2},
\end{aligned}$$

where the last inequality was taken solely to make the expression more intuitive, since not much is lost by doing so. □

3.4.3 Finite Arithmetic

Finally, we turn to the analysis of SGN in finite arithmetic. By making the machine precision small enough, we can bound the effect of roundoff to ensure that the parameters α_k, ε_k are not too far from what they would have been in the exact arithmetic analysis above. We will stop the iteration before any of the quantities involved will become exponentially small, so we will only need $\operatorname{polylog}(1-\alpha_0, \varepsilon_0, \beta)$ bits of precision, where β is the accuracy parameter.

In exact arithmetic, recall that the Newton iteration is given by $A_{k+1} = g(A_k) = \frac{1}{2}(A_k + A_k^{-1})$. Here we will consider the finite arithmetic version G of the Newton map g, defined as $\mathsf{G}(A) := g(A) + E_A$ where E_A is an adversarial perturbation coming from the round-off error. Hence, the sequence of interest is given by $\widetilde{A}_0 := A$ and $\widetilde{A}_{k+1} := \mathsf{G}(\widetilde{A}_k)$.

In this subsection we will prove the following theorem concerning the runtime and precision of SGN. Our assumptions on the size of the parameters α_0, β are in place only to simplify the analysis; these assumptions are not required for the execution of the algorithm.

Theorem 3.4.9 (Main guarantees for SGN)**.** Assume INV is a $(\mu_{\mathsf{INV}}(n), c_{\mathsf{INV}})$-stable matrix inversion algorithm satisfying Definition 3.2.3. Let $\varepsilon_0 \in (0,1), \beta \in (0, 1/12)$, and assume $A = \widetilde{A}_0$ has its ε_0-pseudospectrum contained in C_{α_0} where $0 < 1 - \alpha_0 < 1/100$. Run SGN with

$$N = \lceil \lg(1/(1-\alpha_0)) + 3\lg\lg(1/(1-\alpha_0)) + \lg\lg(1/(\beta\varepsilon_0)) + 7.59 \rceil$$

iterations, as in the statement of the algorithm. Then $\widetilde{A_N} = \mathsf{SGN}(A)$ satisfies the advertised accuracy guarantee

$$\|\widetilde{A_N} - \operatorname{sgn}(A)\| \leq \beta$$

when run with machine precision satisfying

$$\mathbf{u} \leq \frac{\alpha_0^{2^{N+1}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N},$$

corresponding to at most

$$\lg(1/\mathbf{u}) = O(\log n \log^3(1/(1-\alpha_0))(\log(1/\beta) + \log(1/\varepsilon_0)))$$

required bits of precision. The number of arithmetic operations is at most

$$N(4n^2 + T_{\mathsf{INV}}(n)).$$

Later on, we will need to call SGN on a matrix with shattered pseudospectrum; the lemma below calculates acceptable parameter settings for shattering so that the pseudospectrum is contained in the required pair of Appolonian circles, satisfying the hypothesis of Theorem 3.4.9.

Lemma 3.4.10. If A has ϵ-pseudospectrum shattered with respect to a grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$ that includes the imaginary axis as a grid line, then one has $\Lambda_{\varepsilon_0}(A) \subseteq \mathsf{C}_{\alpha_0}$ where $\epsilon_0 = \epsilon/2$ and

$$\alpha_0 = 1 - \frac{\epsilon}{\operatorname{diam}(\mathsf{g})^2}.$$

In particular, if ε is at least $1/\mathrm{poly}(n)$ and ωs_1 and ωs_2 are at most $\mathrm{poly}(n)$, then ε_0 and $1 - \alpha_0$ are also at least $1/\mathrm{poly}(n)$.

Proof. First, because it is shattered, the $\epsilon/2$-pseudospectrum of A is at least distance $\epsilon/2$ from g. Recycling the calculation from Proposition 3.4.2, it suffices to take

$$\alpha_0^2 = \max_{z \in \Lambda_{\epsilon/2}(A)} \left(1 - \frac{4|\Re z|}{|z - \operatorname{sgn}(z)|^2}\right).$$

From what we just observed about the pseudospectrum, we can take $|\Re z| \geq \epsilon/2$. To bound the denominator, we can crudely use the fact any two points inside the grid are at distance no more than $\operatorname{diam}(\mathsf{g})$. Finally, we use $\sqrt{1-x} \leq 1 - x/2$ for any $x \in (0,1)$. □

The proof of Theorem 3.4.9 will proceed as in the exact arithmetic case, with the modification that ε_k must be decreased by an additional factor after each iteration to account for roundoff. At each step, we set the machine precision $\mathbf{u}$ small enough so that the ε_k remain close to what they would be in exact arithmetic. For the analysis we will introduce an explicit auxiliary sequence e_k that lower bounds the ε_k, provided that $\mathbf{u}$ is small enough.

Lemma 3.4.11 (One-step additive error)**.** Assume the matrix inverse is computed by an algorithm INV satisfying the guarantee in Definition 3.2.3. Then $\mathsf{G}(A) = g(A) + E$ for some error matrix E with norm

$$\|E\| \leq \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}} \log n}\|A^{-1}\|\right) 4\sqrt{n}\mathbf{u}. \tag{3.22}$$

The proof of this lemma is deferred to Appendix A.2.

With the error bound for each step in hand, we now move to the analysis of the whole iteration. It will be convenient to define $s := 1 - \alpha_0$.. As in the exact arithmetic case, for $k \geq 1$, we will recursively define decreasing sequences α_k and ε_k maintaining the property

$$\Lambda_{\varepsilon_k}(\widetilde{A}_k) \subset \mathsf{C}_{\alpha_k} \qquad \text{for all } k \geq 0 \tag{3.23}$$

by induction as follows:

1. The base case $k = 0$ holds because by assumption, $\Lambda_{\varepsilon_0} \subset \mathsf{C}_{\alpha_0}$.

2. Here we recursively define α_{k+1}. Set

   $$\alpha_{k+1} := (1 + s/4)\alpha_k^2.$$

   In the notation of Subsection 3.4.2, this corresponds to setting $D = 1 + s/4$. This definition ensures that $\alpha_k^2 \leq \alpha_{k+1} \leq \alpha_k$ for all k, and also gives us the bound $(1 + s/4)\alpha_0 \leq 1 - s/2$. We also have the closed form

   $$\alpha_k = (1 + s/4)^{2^k - 1}\alpha_0^{2^k},$$

   which implies the useful bound

   $$\alpha_k \leq (1 - s/2)^{2^k}. \tag{3.24}$$

3. Here we recursively define ε_{k+1}. Combining Lemma 3.4.4, the recursive definition of α_{k+1}, and the fact that $1 - \alpha_k^2 \geq 1 - \alpha_0^2 \geq 1 - \alpha_0 = s$, we find that $\Lambda_{\varepsilon'}\left(g(\widetilde{A}_k)\right) \subset \mathsf{C}_{\alpha_{k+1}}$, where

   $$\varepsilon' = \varepsilon_k \frac{(\alpha_{k+1} - \alpha_k^2)\,(1 - \alpha_k^2)}{8\alpha_k} = \varepsilon_k \frac{s\alpha_k(1 - \alpha_k^2)}{32} \geq \varepsilon_k \frac{\alpha_k s^2}{32}.$$

   Thus in particular

   $$\Lambda_{\varepsilon_k \alpha_k s^2/32}\left(g(\widetilde{A}_k)\right) \subset \mathsf{C}_{\alpha_{k+1}}.$$

   Since $\widetilde{A}_{k+1} = \mathsf{G}(\widetilde{A}_k) = g(\widetilde{A}_k) + E_k$, for some error matrix E_k arising from roundoff, Proposition 1.1.6 ensures that if we set

   $$\varepsilon_{k+1} := \varepsilon_k \frac{s^2\alpha_k}{32} - \|E_k\| \tag{3.25}$$

   we will have $\Lambda_{\varepsilon_{k+1}}(\widetilde{A}_{k+1}) \subset \mathsf{C}_{\alpha_{k+1}}$, as desired.

We now need to show that the ε_k do not decrease too fast as k increases. In view of (3.25), it will be helpful to set the machine precision small enough to guarantee that $\|E_k\|$ is a small fraction of $\varepsilon_k \frac{\alpha_k s^2}{32}$.

First, we need to control the quantities $\|\widetilde{A}_k\|$, $\|\widetilde{A}_k^{-1}\|$, and $\kappa(\widetilde{A}_k) = \|\widetilde{A}_k\|\|\widetilde{A}_k^{-1}\|$ appearing in our upper bound (3.22) on $\|E_k\|$ from Lemma 3.4.11, as functions of ϵ_k. By Corollary 3.4.6, we have

$$\|\widetilde{A}_k^{-1}\| \leq \frac{1}{\epsilon_k} \quad \text{and} \quad \|\widetilde{A}_k\| \leq 4\frac{\alpha_k}{(1-\alpha_k)^2\varepsilon_k} \leq \frac{4}{s^2\varepsilon_k}.$$

Thus, we may write the coefficient of $\mathbf{u}$ in the bound (3.22) as

$$K_{\varepsilon_k} := \left[\frac{4}{s^2\varepsilon_k} + \frac{1}{\varepsilon_k} + \mu_{\mathsf{INV}}(n)\left(\frac{4}{s^2\varepsilon_k^2}\right)^{c_{\mathsf{INV}}\log n}\frac{1}{\varepsilon_k}\right]4\sqrt{n}$$

so that Lemma 3.4.11 reads

$$\|E_k\| \leq K_{\varepsilon_k}\mathbf{u}. \tag{3.26}$$

Plugging this into the definition (3.25) of ε_{k+1},we have

$$\varepsilon_{k+1} \geq \varepsilon_k\frac{s^2\alpha_k}{32} - K_{\varepsilon_k}\mathbf{u}. \tag{3.27}$$

Now suppose we take $\mathbf{u}$ small enough so that

$$K_{\varepsilon_k}\mathbf{u} \leq \frac{1}{3}\varepsilon_k\frac{s^2\alpha_k}{32}. \tag{3.28}$$

For such $\mathbf{u}$, we then have

$$\varepsilon_{k+1} \geq \frac{2}{3}\varepsilon_k\frac{s^2\alpha_k}{32}, \tag{3.29}$$

which implies

$$\|E_k\| \leq \frac{1}{2}\varepsilon_{k+1}; \tag{3.30}$$

this bound is loose but sufficient for our purposes. Inductively, we now have the following bound on ε_k in terms of α_k:

Lemma 3.4.12 (Lower bound on ε_k)**.** Let $k \geq 0$, and for all $0 \leq i \leq k-1$, assume $\mathbf{u}$ satisfies the requirement (3.28):

$$K_{\varepsilon_i}\mathbf{u} \leq \frac{1}{3}\varepsilon_i\frac{s^2\alpha_i}{32}.$$

Then we have

$$\varepsilon_k \geq e_k := \varepsilon_0\left(\frac{s^2}{50}\right)^k\alpha_k.$$

In fact, it suffices to assume the hypothesis only for $i = k-1$.

Proof. The last statement follows from the fact that K_{ε_i} is increasing and ε_i is decreasing in i.

Since (3.28) implies (3.29), we may apply (3.29) repeatedly to obtain

$$\begin{aligned}
\varepsilon_k &\geq \varepsilon_0 (s^2/48)^k \prod_{i=0}^{k-1} \alpha_i \\
&= \varepsilon_0 (s^2/48)^k (1+s/4)^{2^k-1-k} \alpha_0^{2^k-1} && \text{by the definition of } \alpha_i \\
&= \varepsilon_0 \left(\frac{s^2}{48(1+s/4)} \right)^k \frac{\alpha_k}{\alpha_0} \\
&\geq \varepsilon_0 \left(\frac{s^2}{50} \right)^k \alpha_k. && \alpha_0 \leq 1, s < 1/8
\end{aligned}$$

□

We now show that the conclusion of Lemma 3.4.12 still holds if we replace ε_i everywhere in the hypothesis by e_i, which is an explicit function of ε_0 and α_0 defined in Lemma 3.4.12. Note that we do not know $\varepsilon_i \geq e_i$ a priori, so to avoid circularity we must use a short inductive argument.

Corollary 3.4.13 (Lower bound on ε_k with predictable hypothesis)**.** Let $k \geq 0$, and for all $0 \leq i \leq k-1$, assume $\mathbf{u}$ satisfies

$$K_{e_i}\mathbf{u} \leq \frac{1}{3} e_i \frac{s^2 \alpha_i}{32} \tag{3.31}$$

where e_i is defined in Lemma 3.4.12. Then we have

$$\varepsilon_k \geq e_k.$$

In fact, it suffices to assume the hypothesis only for $i = k-1$.

Proof. The last statement follows from the fact that e_i is decreasing in i.

Assuming the full hypothesis of this lemma, we prove $\varepsilon_i \geq e_i$ for $0 \leq i \leq k$ by induction on i. For the base case, we have $\varepsilon_0 \geq e_0 = \varepsilon_0 \alpha_0$.

For the inductive step, assume $\varepsilon_i \geq e_i$. Then as long as $i \leq k-1$, the hypothesis of this lemma implies

$$K_{\varepsilon_i}\mathbf{u} \leq \frac{1}{3} \varepsilon_i \frac{s^2 \alpha_i}{32},$$

so we may apply Lemma 3.4.12 to obtain $\varepsilon_{i+1} \geq e_{i+1}$, as desired. □

Lemma 3.4.14. Suppose $\mathbf{u}$ satisfies the requirement (3.28) for all $0 \leq k \leq N$. Then

$$\|\widetilde{A}_N - \operatorname{sgn}(A)\| \leq \frac{8}{s} \sum_{k=0}^{N-1} \frac{\|E_k\|}{\varepsilon_{k+1}^2} + \frac{8 \cdot 50^N}{s^{2N+2} \varepsilon_0} (1 - s/2)^{2^N}. \tag{3.32}$$

Proof. Since $\operatorname{sgn} = \operatorname{sgn} \circ g$, for every k we have

$$\|\operatorname{sgn}(\widetilde{A_{k+1}}) - \operatorname{sgn}(\widetilde{A_k})\| = \|\operatorname{sgn}(\widetilde{A_{k+1}}) - \operatorname{sgn}(g(\widetilde{A_k}))\| = \|\operatorname{sgn}(\widetilde{A_{k+1}}) - \operatorname{sgn}(\widetilde{A_{k+1}} - E_k)\|.$$

From the holomorphic functional calculus we can rewrite $\|\operatorname{sgn}(\widetilde{A_{k+1}}) - \operatorname{sgn}(\widetilde{A_{k+1}} - E_k)\|$ as the norm of a certain contour integral, which in turn can be bounded as follows:

$$\begin{aligned}
&\frac{1}{2\pi}\Bigg\| \oint_{\partial \mathsf{C}^+_{\alpha_{k+1}}} [(z - \widetilde{A_{k+1}})^{-1} - (z - (\widetilde{A_{k+1}} - E_k))^{-1}]\, dz \\
&\qquad - \oint_{\partial \mathsf{C}^-_{\alpha_{k+1}}} [(z - \widetilde{A_{k+1}})^{-1} - (z - (\widetilde{A_{k+1}} - E_k))^{-1}]\, dz \Bigg\| \\
=&\frac{1}{2\pi}\Bigg\| \oint_{\partial \mathsf{C}^+_{\alpha_{k+1}}} [(z - (\widetilde{A_{k+1}} - E_k))^{-1} E_k (z - \widetilde{A_{k+1}})^{-1}]\, dz \\
&\qquad - \oint_{\partial \mathsf{C}^-_{\alpha_{k+1}}} [(z - (\widetilde{A_{k+1}} - E_k))^{-1} E_k (z - \widetilde{A_{k+1}})^{-1}]\, dz \Bigg\| \\
\leq&\frac{1}{\pi} \oint_{\partial \mathsf{C}^+_{\alpha_{k+1}}} \|(z - (\widetilde{A_{k+1}} - E_k))^{-1}\| \|E_k\| \|(z - \widetilde{A_{k+1}})^{-1}\|\, dz \\
\leq&\frac{1}{\pi} \ell(\partial \mathsf{C}^+_{\alpha_{k+1}}) \|E_k\| \frac{1}{\varepsilon_{k+1} - \|E_k\|} \frac{1}{\varepsilon_{k+1}} \\
=&\frac{4\alpha_{k+1}}{1 - \alpha_{k+1}^2} \|E_k\| \frac{1}{\varepsilon_{k+1} - \|E_k\|} \frac{1}{\varepsilon_{k+1}},
\end{aligned}$$

where we use the definition (1.3) of pseudospectrum and Proposition 1.1.6, together with the property (3.23). Ultimately, this chain of inequalities implies

$$\|\operatorname{sgn}(\widetilde{A_{k+1}}) - \operatorname{sgn}(\widetilde{A_{k+1}} - E_k)\| \leq \frac{4\alpha_{k+1}}{1 - \alpha_{k+1}^2} \|E_k\| \frac{1}{\varepsilon_{k+1} - \|E_k\|} \frac{1}{\varepsilon_{k+1}}.$$

Summing over all k and using the triangle inequality, we obtain

$$\begin{aligned}
\|\operatorname{sgn}(\widetilde{A_N}) - \operatorname{sgn}(\widetilde{A_0})\| &\leq \sum_{k=1}^{N-1} \frac{4\alpha_{k+1}}{1 - \alpha_{k+1}^2} \|E_k\| \frac{1}{\varepsilon_{k+1} - \|E_k\|} \frac{1}{\varepsilon_{k+1}} \\
&\leq \frac{8}{s} \sum_{k=0}^{N-1} \frac{\|E_k\|}{\varepsilon_{k+1}^2},
\end{aligned}$$

where in the last step we use $\alpha_k \leq 1$ and $1 - \alpha_{k+1}^2 \geq s$, as well as (3.30).

By Lemma 3.4.3, we have

$$\begin{aligned}\|\widetilde{A_N} - \operatorname{sgn}(\widetilde{A_N})\| &\leq \frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\varepsilon_N} \\ &\leq \frac{8}{s^2}\alpha_N\frac{\alpha_N}{\varepsilon_N} \\ &\leq \frac{8}{s^2}\alpha_N\frac{1}{\varepsilon_0}\left(\frac{50}{s^2}\right)^N \\ &\leq \frac{8}{s^2\varepsilon_0}(1-s/2)^{2^N}\left(\frac{50}{s^2}\right)^N \\ &\leq \frac{8\cdot 50^N}{s^{2N+2}\varepsilon_0}(1-s/2)^{2^N}.\end{aligned}$$

where we use $s < 1/2$ in the last step.

Combining the above with the triangle inequality, we obtain the desired bound.

□

We would like to apply Lemma 3.4.14 to ensure $\|\widetilde{A_N} - \operatorname{sgn}(A)\| < \beta$. The bound in Lemma 3.4.14 is the sum of two terms; we will make each term less than $\beta/2$. The bound for the second term will yield a sufficient condition on the number of iterations N. Given that, the bound on the first term will give a sufficient condition on the machine precision $\mathbf{u}$. This will be the content of Lemmas 3.4.16 and 3.4.17.

We start with the second term. The following preliminary lemma will be useful:

Lemma 3.4.15. Let $1/800 > t > 0$ and $1/2 > c > 0$ be given. Then for

$$j \geq \lg(1/t) + 2\lg\lg(1/t) + \lg\lg(1/c) + 1.62,$$

we have

$$\frac{(1-t)^{2^j}}{t^{2j}} < c.$$

The proof is deferred to Appendix A.2.

Lemma 3.4.16. Suppose we have

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\varepsilon_0)) + 4.$$

Then

$$\frac{8\cdot 50^N}{s^{2N+2}\varepsilon_0}(1-s/2)^{2^N} \leq \beta/2.$$

Proof. It is sufficient that

$$\frac{8\cdot 64^N}{s^{2N+2}\varepsilon_0}(1-s/8)^{2^N} \leq \beta/2.$$

The result now follows from applying Lemma 3.4.15 with $c = \beta s^2\varepsilon_0/16$ and $t = s/8$.

□

Now we move to the first term in the bound of Lemma 3.4.14.

Lemma 3.4.17. Suppose

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2 \varepsilon_0)) + 1.62,$$

and suppose the machine precision $\mathbf{u}$ satisfies

$$\mathbf{u} \leq \frac{(1-s)^{2^{N+1}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N}.$$

Then we have

$$\frac{8}{s}\sum_{k=0}^{N-1} \frac{\|E_k\|}{\varepsilon_{k+1}^2} \leq \beta/2.$$

Proof. It suffices to show that for all $0 \leq k \leq N-1$,

$$\|E_k\| \leq \frac{\beta \varepsilon_{k+1}^2 s}{16N}.$$

In view of (3.26), which says $\|E_k\| \leq K_{\varepsilon_k}\mathbf{u}$, it is sufficient to have for all $0 \leq k \leq N-1$

$$\mathbf{u} \leq \frac{1}{K_{\varepsilon_k}} \frac{\beta \varepsilon_{k+1}^2 s}{16N}. \tag{3.33}$$

For this, we claim it is sufficient to have for all $0 \leq k \leq N-1$

$$\mathbf{u} \leq \frac{1}{K_{e_k}} \frac{\beta e_{k+1}^2 s}{16N}. \tag{3.34}$$

Indeed, using the facts $\beta < 1/6$, and $e_{k+1} \leq e_k \leq \varepsilon_0 \alpha_k \leq \alpha_k$, the condition (3.34) is the hypothesis of Corollary 3.4.13, so we have the conclusion $e_k \leq \varepsilon_k$ for all $0 \leq k \leq N$, which implies the condition (3.33).

Finally, because $1/K_{e_k}$ and e_k are decreasing in k, it is sufficient to have the single condition

$$\mathbf{u} \leq \frac{1}{K_{e_N}} \frac{\beta e_N^2 s}{16N}.$$

We continue the chain of sufficient conditions on $\mathbf{u}$, where each line implies the line above:

$$\begin{aligned}
\mathbf{u} &\leq \frac{1}{K_{e_N}} \frac{\beta e_N^2 s}{16N} \\
\mathbf{u} &\leq \frac{1}{\left[\frac{4}{s^2 e_N} + \frac{1}{e_N} + \mu_{\mathsf{INV}}(n)\left(\frac{4}{s^2 e_N^2}\right)^{c_{\mathsf{INV}}\log n} \frac{1}{e_N}\right] 4\sqrt{n}} \frac{\beta e_N^2 s}{16N} \\
\mathbf{u} &\leq \frac{1}{6\mu_{\mathsf{INV}}(n)\left(\frac{4}{s^2 e_N}\right)^{c_{\mathsf{INV}}\log n + 1} 4\sqrt{n}} \frac{\beta e_N^2 s}{16N} \\
\mathbf{u} &\leq \frac{\beta}{6 \cdot 4 \cdot 16 \mu_{\mathsf{INV}}(n)\sqrt{n}N} \left(\frac{e_N s^2}{4}\right)^{c_{\mathsf{INV}}\log n + 3}.
\end{aligned}$$

where we use the bound $\frac{1}{e_N} \leq \frac{4}{s^2 e_N^2}$ without much loss, and we also assume $\mu_{\mathsf{INV}}(n) \geq 1$ and $c_{\mathsf{INV}} \log n \geq 1$ for simplicity.

Substituting the value of e_N as defined in Lemma 3.4.12, we get the sufficient condition

$$\mathbf{u} \leq \frac{\beta}{384\mu_{\mathsf{INV}}(n)\sqrt{n}N} \left(\frac{\varepsilon_0 (s^2/50)^N \alpha_N s^2}{4} \right)^{c_{\mathsf{INV}} \log n + 3}.$$

Replacing α_N by the smaller quantity $\alpha_0^{2^N} = (1-s)^{2^N}$ and cleaning up the constants yields the sufficient condition

$$\mathbf{u} \leq \frac{\beta}{400\mu_{\mathsf{INV}}(n)\sqrt{n}N} \left(\frac{\varepsilon_0 (s^2/50)^N (1-s)^{2^N} s^2}{4} \right)^{c_{\mathsf{INV}} \log n + 3}.$$

Now we finally use our hypothesis on the size of N to simplify this expression. Applying Lemma 3.4.16, we have

$$\varepsilon_0 (s^2/50)^N / 4 \geq \frac{4(1-s)^{2^N}}{\beta}.$$

Thus, our sufficient condition becomes

$$\mathbf{u} \leq \frac{\beta}{400\mu_{\mathsf{INV}}(n)\sqrt{n}N} \left(\frac{4(1-s)^{2^{N+1}}}{\beta} \right)^{c_{\mathsf{INV}} \log n + 3}.$$

To make the expression simpler, since $c_{\mathsf{INV}} \log n + 3 \geq 4$ we may pull out a factor of $4^4 > 200$ and remove the occurrences of β to yield the sufficient condition

$$\mathbf{u} \leq \frac{(1-s)^{2^{N+1}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N}.$$

□

Matching the statement of Theorem 3.4.9, we give a slightly cleaner sufficient condition on N, the proof of which is deferred to Appendix A.2.

Lemma 3.4.18 (Final condition on N)**.** If

$$N = \lceil \lg(1/s) + 3 \lg\lg(1/s) + \lg\lg(1/(\beta\varepsilon_0)) + 7.59 \rceil,$$

then

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2 \varepsilon_0)) + 1.62.$$

Lemma 3.4.19 (Bit length computation)**.** Suppose

$$N = \lceil \lg(1/s) + 3 \lg\lg(1/s) + \lg\lg(1/(\beta\varepsilon_0)) + 7.59 \rceil$$

and

$$\mathbf{u} \leq \frac{(1-s)^{2^{N+1}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N}.$$

Then

$$\log(1/\mathbf{u}) = O\big(\log n \log(1/s)^3(\log(1/\beta) + \log(1/\varepsilon_0))\big).$$

Proof. Immediately we have

$$\log(1/\mathbf{u}) = O\left(\log(1/\beta) + \log \mu_{\mathsf{INV}}(n) + \log n + \log N + (\log n)2^{N+1}\log(1/(1-s))\right).$$

We first focus on the term $2^{N+1}\log(1/(1-s))$. Note that $\log(1/(1-s)) = O(s)$. Thus,

$$2^{N+1}\log(1/(1-s)) = (1/s)\cdot 2^{3\lg\lg(1/s)+\lg\lg(1/(\beta\varepsilon_0))+9.59}\cdot O(s) = O(\log(1/s)^3(\log(1/\beta)+\log(1/\varepsilon_0))).$$

Using that $\mu_{\mathsf{INV}}(n) = \mathrm{poly}(n)$ and discarding subdominant terms, we obtain the desired bound.

□

This completes the proof of Theorem 3.4.9. Finally, we may prove the theorem advertised in the Introduction.

Proof of Theorem 3.1.5. Set $\varepsilon := \min\{\frac{1}{K}, 1\}$. Then $\Lambda_\varepsilon(A)$ does not intersect the imaginary axis, and furthermore $\Lambda_\varepsilon(A) \subseteq D(0,2)$ because $\|A\| \leq 1$. Thus, we may apply Lemma 3.4.10 with $\mathrm{diam}(\mathbf{g}) = 4\sqrt{2}$ to obtain parameters α_0, ε_0 with the property that $\log(1/(1-\alpha_0))$ and $\log(1/\varepsilon_0)$ are both $O(\log K)$. Theorem 3.4.9 now yields the desired conclusion. □

3.5 Spectral Bisection Algorithm

In this section we will prove Theorem 3.1.6. As discussed in Section 3.1, our algorithm is not new, and in its idealized form it reduces to the two following tasks:

Split: Given an $n \times n$ matrix A, find a partition of the spectrum into pieces of roughly equal size, and output spectral projectors $P_\pm$ onto each of these pieces.

Deflate: Given an $n \times n$ rank-k projector P, output an $n \times k$ matrix Q with orthogonal columns that span the range of P.

These routines in hand, on input A one can compute $P_\pm$ and the corresponding $Q_\pm$, and then find the eigenvectors and eigenvalues of $A_\pm := Q_\pm^* A Q_\pm$. The observation below verifies that this recursion is sound.

Observation 3.5.1. The spectrum of A is exactly $\Lambda(A_+) \sqcup \Lambda(A_-)$, and every eigenvector of A is of the form $Q_\pm v$ for some eigenvector v of one of $A_\pm$.

The difficulty, of course, is that neither of these routines can be executed exactly: we will never have access to true projectors $P_\pm$, nor to the actual orthogonal matrices $Q_\pm$ whose columns span their range, and must instead make do with approximations. Because our algorithm is recursive and our matrices nonnormal, we must take care that the errors in the sub-instances $A_\pm$ do not corrupt the eigenvectors and eigenvalues we are hoping to find. Additionally, the Newton iteration we will use to split the spectrum behaves poorly when an eigenvalue is close to the imaginary axis, and it is not clear how to find a splitting which is balanced.

Our tactic in resolving these issues will be to pass to our algorithms a matrix *and* a grid with respect to which its ϵ-pseudospectrum is shattered. To find an approximate eigenvalue, then, one can settle for locating the grid square it lies in; containment in a grid square is robust to perturbations of size smaller than ϵ. The shattering property is robust to small perturbations, inherited by the subproblems we pass to, and—because the spectrum is quantifiably far from the grid lines—allows us to run the Newton iteration in the first place.

Let us now sketch the implementations and state carefully the guarantees for SPLIT and DEFLATE; the analysis of these will be deferred to Appendices A.3 and A.4. Our splitting algorithm is presented a matrix A whose ϵ-pseudospectrum is shattered with respect to a grid g. For any vertical grid line with real part h, $\operatorname{Tr}\operatorname{sgn}(A-h)$ gives the difference between the number of eigenvalues lying to its left and right. As

$$|\operatorname{Tr}\mathsf{SGN}(A-h) - \operatorname{Tr}\operatorname{sgn}(A-h)| \le n\|\mathsf{SGN}(A-h) - \operatorname{sgn}(A-h)\|,$$

we can determine these eigenvalue counts *exactly* by running SGN to accuracy $O(1/n)$ and rounding $\operatorname{Tr}\mathsf{SGN}(A-h)$ to the nearest integer. We will show in Appendix A.3 that, by mounting a binary search over horizontal and vertical lines of g, we will always arrive at a partition of the eigenvalues into two parts with size at least $\min\{n/5, 1\}$. Having found it, we run SGN one final time at the desired precision to find the approximate spectral projectors.

Theorem 3.5.2 (Guarantees for SPLIT)**.** Assume INV is a $(\mu_{\mathsf{INV}}, c_{\mathsf{INV}})$-stable matrix inversion algorithm satisfying Definition 3.2.3. Let $\epsilon \le 0.5$, $\beta \le 0.05/n$, and $\|A\| \le 4$ and g have side lengths of at most 8, and define

$$N_{\mathsf{SPLIT}} := \lg\frac{256}{\epsilon} + 3\lg\lg\frac{256}{\epsilon} + \lg\lg\frac{4}{\beta\epsilon} + 7.59.$$

Then SPLIT has the advertised guarantees when run on a floating point machine with precision

$$\mathbf{u} \le \mathbf{u}_{\mathsf{SPLIT}} := \min\left\{\frac{\left(1-\frac{\epsilon}{256}\right)^{2^{N_{\mathsf{SPLIT}}+1}(c_{\mathsf{INV}}\log n+3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{SPLIT}}}, \frac{\epsilon}{100n}\right\},$$

Using at most

$$T_{\mathsf{SPLIT}}(n,\mathsf{g},\epsilon,\beta) \le 12\lg\frac{1}{\omega(\mathsf{g})}\cdot N_{\mathsf{SPLIT}}\cdot\left(T_{\mathsf{INV}}(n)+O(n^2)\right)$$

SPLIT

Input: Matrix $A \in \mathbb{C}^{n\times n}$, pseudospectral parameter ϵ, grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$, and desired accuracy β
Requires: $\Lambda_\epsilon(A)$ is shattered with respect to g, and $\beta \le 0.05/n$
Algorithm: $(\widetilde{P_\pm}, \mathsf{g}_\pm) = \mathsf{SPLIT}(A, \epsilon, \mathsf{g}, \beta)$

1. Execute a binary search over horizontal grid shifts h until

$$\operatorname{Tr} \mathsf{SGN}\left(A - h, \epsilon/4, 1 - \frac{\epsilon}{2\operatorname{diam}(\mathsf{g})^2}, \beta\right) \le 3n/5.$$

2. If this fails, set $A \leftarrow iA$ and repeat with vertical grid shifts

3. Once a shift is found,

$$\widetilde{P_\pm} \leftarrow \tfrac{1}{2}\left(\mathsf{SGN}\left(A - h, \epsilon/4, 1 - \frac{\epsilon}{2\operatorname{diam}(\mathsf{g})^2}, \beta\right) \pm 1\right),$$

and $\mathsf{g}_\pm$ are set to the two subgrids

Output: Two matrices $\widetilde{P_\pm} \in \mathbb{C}^{n\times n}$, two subgrids $\mathsf{g}_\pm$, and two numbers $n_\pm$
Ensures: Each subgrid $\mathsf{g}_\pm$ contains $n_\pm$ eigenvalues of A, $n_\pm \ge n/5$, and $\|\widetilde{P_\pm} - P_\pm\| \le \beta$, where $P_\pm$ are the true spectral projectors for the eigenvalues in the subgrids $\mathsf{g}_\pm$ respectively.

arithmetic operations. The number of bits required is

$$\lg 1/\mathbf{u}_{\mathsf{SPLIT}} = O\left(\log n \log^3 \frac{256}{\epsilon}\left(\log\frac{1}{\beta} + \log\frac{4}{\epsilon}\right)\right).$$

Deflation of the approximate projectors we obtain from SPLIT amounts to a standard rank-revealing QR factorization. This can be achieved deterministically in $O(n^3)$ time with the classic algorithm of Gu and Eisenstat [67], or probabilistically in matrix-multiplication time with a variant of the method of [47]; we will use the latter.

Theorem 3.5.3 (Guarantees for DEFLATE)**.** Assume MM and QR are matrix multiplication and QR factorization algorithms satisfying Definitions 3.2.2 and 3.2.4. Then DEFLATE has the advertised guarantees when run on a machine with precision:

$$\mathbf{u} \le \mathbf{u}_{\mathsf{DEFLATE}} := \min\left\{\frac{\beta}{4\|\widetilde{P}\|\max(\mu_{\mathsf{QR}}(n), \mu_{\mathsf{MM}}(n))}, \frac{\eta}{2\mu_{\mathsf{QR}}(n)}\right\}.$$

The number of arithmetic operations is at most:

$$T_{\mathsf{DEFLATE}}(n) = n^2 T_{\mathsf{N}} + 2T_{\mathsf{QR}}(n) + T_{\mathsf{MM}}(n).$$

DEFLATE

Input: Matrix $\widetilde{P} \in \mathbb{C}^{n\times n}$, desired rank k, input precision β, and desired accuracy η
Requires: $\|\widetilde{P} - P\| \leq \beta \leq \frac{1}{4}$ for some rank-k projector P.
Algorithm: $\widetilde{Q} = \mathsf{DEFLATE}(P, k, \beta, \eta)$

1. $H \leftarrow n \times n$ Haar unitary $+E_1$
2. $(U, R) \leftarrow \mathsf{QR}(PH^*)$
3. $\widetilde{Q} \leftarrow$ first k columns of U.

Output: A tall matrix $\widetilde{Q} \in \mathbb{C}^{n\times k}$
Ensures: There exists a matrix $Q \in \mathbb{C}^{n\times k}$ whose orthogonal columns span range(P), such that $\|\widetilde{Q} - Q\| \leq \eta$, with probability at least $1 - \frac{(20n)^3\sqrt{\beta}}{\eta^2}$.

Remark 3.5.4. The proof of the above theorem, which is deferred to Appendix A.4, closely follows and builds on the analysis of the randomized rank revealing factorization algorithm (RURV) introduced in [47] and further studied in [10]. The parameters in the theorem are optimized for the particular application of finding a basis for a deflating subspace given an approximate spectral projector.

The main difference with the analysis in [47] and [10] is that here, to make it applicable to complex matrices, we make use of Haar unitary random matrices instead of Haar orthogonal random matrices. In our analysis of the unitary case, we discovered a strikingly simple formula (Corollary A.4.6) for the density of the smallest singular value of an $r \times r$ sub-matrix of an $n \times n$ Haar unitary; this formula is leveraged to obtain guarantees that work for any n and r, and not only for when $n - r \geq 30$, as was the case in [10]. Finally, we explicitly account for finite arithmetic considerations in the Gaussian randomness used in the algorithm, where true Haar unitary matrices can never be produced.

We are ready now to state completely an algorithm EIG which accepts a shattered matrix and grid and outputs approximate eigenvectors and eigenvalues with a *forward-error* guarantee. Aside from the a priori un-motivated parameter settings in lines 2-3 and 9—which we promise to justify in the analysis to come—EIG implements an approximate version of the split and deflate framework that began this section.

Theorem 3.5.5 (EIG: Finite Arithmetic Guarantee)**.** Assume MM, QR, and INV are numerically stable algorithms for matrix multiplication, QR factorization, and inversion satisfying Definitions 3.2.2, 3.2.4, and 3.2.3. Let $\delta < 1$, $A \in \mathbb{C}^{n\times n}$ have $\|A\| \leq 3.5$ and, for some $\epsilon < 1$, have ϵ-pseudospectrum shattered with respect to a grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$ with side lengths at most 8 and $\omega \leq 1$. Define

$$N_{\mathsf{EIG}} := \lg \frac{256n}{\epsilon} + 3 \lg \lg \frac{256n}{\epsilon} + \lg \lg \frac{(5n)^{26}}{\theta^2 \delta^4 \epsilon^9}.$$

EIG

Input: Matrix $A \in \mathbb{C}^{m\times m}$, desired eigenvector accuracy δ, grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$, pseudospectral guarantee ϵ, acceptable failure probability θ, and global instance size n
Requires: $\Lambda_\epsilon(A)$ is shattered with respect to g, and $m \le n$.
Algorithm: $\mathsf{EIG}(A, \delta, \mathsf{g}, \epsilon, \theta, n)$

1. If A is 1×1, $(\widetilde{V}, \widetilde{D}) \leftarrow (1, A)$
2. $\eta \leftarrow \frac{\delta\epsilon^2}{200}$
3. $\beta \leftarrow \frac{\eta^4}{(20n)^6}\frac{\theta^2}{4n^8}$
4. $(\widetilde{P}_+, \widetilde{P}_-, \mathsf{g}_+, \mathsf{g}_-, n_+, n_-) \leftarrow \mathsf{SPLIT}(A, \epsilon, \mathsf{g}, \beta)$
5. $\widetilde{Q}_\pm \leftarrow \mathsf{DEFLATE}(\widetilde{P}_\pm, n_\pm, \beta, \eta)$
6. $\widetilde{A}_\pm \leftarrow \widetilde{Q}_\pm^* \widetilde{A} \widetilde{Q}_\pm + E_{6,\pm}$
7. $(\widetilde{V}_\pm, \widetilde{D}_\pm) \leftarrow \mathsf{EIG}(\widetilde{A}_\pm, 4\delta/5, \mathsf{g}_\pm, 4\epsilon/5, \theta, n)$.
8. $\widetilde{V} \leftarrow \begin{pmatrix} \widetilde{Q}_+\widetilde{V}_+ & \widetilde{Q}_-\widetilde{V}_- \end{pmatrix} + E_8$
9. $\widetilde{V} \leftarrow \mathrm{normalize}(\widetilde{V}) + E_9$
10. $\widetilde{D} \leftarrow \begin{pmatrix} \widetilde{D}_+ & \\ & \widetilde{D}_- \end{pmatrix}$

Output: Eigenvectors and eigenvalues $(\widetilde{V}, \widetilde{D})$
Ensures: With probability at least $1-\theta$, each entry $\widetilde{\lambda}_i = \widetilde{D}_{i,i}$ lies in the same square as exactly one eigenvalue $\lambda_i \in \Lambda(A)$, and each column $\widetilde{v}_i$ of $\widetilde{V}$ has norm $1 \pm n\mathbf{u}$, and satisfies $\|\widetilde{v}_i - v_i\| \le \delta$ for some exact unit right eigenvector $Av_i = \lambda_i v_i$.

Then EIG has the advertised guarantees when run on a floating point machine with precision satisfying:

$$\begin{aligned}\lg 1/\mathbf{u} &\ge \max\left\{\lg^3\frac{n}{\epsilon}\lg\left(\frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8}\right)2^{14.83}(c_{\mathsf{INV}}\log n + 3) + \lg N_{\mathsf{EIG}}, \lg\frac{(5n)^{30}}{\theta^2\delta^4\epsilon^8} + \lg\max\{\mu_{\mathsf{MM}}(n), \mu_{\mathsf{QR}}(n), n\}\right\} \\ &= O\left(\log^3\frac{n}{\epsilon}\log\frac{n}{\theta\delta\epsilon}\log n\right).\end{aligned}$$

The number of arithmetic operations is at most

$$\begin{aligned} T_{\mathsf{EIG}}(n, \delta, \mathsf{g}, \epsilon, \theta, n) &= 60 N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})} \left(T_{\mathsf{INV}}(n) + O(n^2) \right) + 10 T_{\mathsf{QR}}(n) + 25 T_{\mathsf{MM}}(n) \\ &= O\left(\log \frac{1}{\omega(\mathsf{g})} \left(\log \frac{n}{\epsilon} + \log \log \frac{1}{\theta \delta} \right) T_{\mathsf{MM}}(n) \right). \end{aligned}$$

Remark 3.5.6. We have not fully optimized the large constant $2^{14.83}$ appearing in the bit length above.

Theorem 3.5.5 easily implies Theorem 3.1.6 when combined with SHATTER.

Theorem 3.5.7 (Restatement of Theorem 3.1.6)**.** There is a randomized algorithm EIG which on input any matrix $A \in \mathbb{C}^{n \times n}$ with $\|A\| \leq 1$ and a desired accuracy parameter $\delta \in (0, 1)$ outputs a diagonal D and invertible V such that

$$\|A - VDV^{-1}\| \leq \delta \quad \text{and} \quad \kappa(V) \leq 32 n^{2.5} / \delta$$

in

$$O\left(T_{\mathsf{MM}}(n) \log^2 \frac{n}{\delta} \right)$$

arithmetic operations on a floating point machine with

$$O\left(\log^4 \frac{n}{\delta} \log n \right)$$

bits of precision, with probability at least $1 - 2/n - O(1/n^2)$, where the implied constant is at most 600. Here $T_{\mathsf{MM}}(n)$ refers to the running time of a numerically stable matrix multiplication algorithm (detailed in Section 3.2.3).

Proof. Given A and δ, consider the following two step algorithm:

1. $(X, \mathsf{g}, \epsilon) \leftarrow \mathsf{SHATTER}(A, \delta/8)$.
2. $(V, D) \leftarrow \mathsf{EIG}(X, \delta', \mathsf{g}, \epsilon, 1/n, n)$, where

$$\delta' := \frac{\delta^3}{n^{2.5} \cdot 6 \cdot 128 \cdot 2}.$$

We will show that this choice of δ' guarantees

$$\|X - VDV^{-1}\| \leq \delta/2.$$

Theorem 3.3.13 implies that $X = WCW^{-1}$ is diagonalizable with probability one, and moreover

$$\kappa(W) = \|W\| \|W\|^{-1} \leq 8 n^2 / \delta$$

when W is normalized to have unit columns, by (1.1) (where we are using the proof of Theorem 3.3.6), with probability at least $1 - 1/n - O(1/n^2)$.

Since $\|X\| \leq \|A\| + \|A - X\| \leq 1 + 4\gamma \leq 3$ from Theorem 3.3.13, the hypotheses of Theorem 3.5.5 are satisfied. Thus EIG succeeds with probability at least $1 - 1/n$. Taking a union bound with the success of SHATTER, we have $V = W + E$ for some $\|E\| \leq \delta'\sqrt{n}$, so

$$\|V - W\| \leq \delta'\sqrt{n},$$

as well as

$$\sigma_n(V) \geq \sigma_n(W) - \|E\| \geq \frac{\delta}{8n^2} - \delta'\sqrt{n} \geq \frac{\delta}{16n^2},$$

since our choice of δ' satisfies.

$$\delta' \leq \frac{\delta}{16n^{2.5}},$$

This implies that

$$\kappa(V) = \|V\|\|V^{-1}\| \leq 2\sqrt{n} \cdot \frac{16n^2}{\delta},$$

establishing the last item of the theorem.

We can control the perturbation of the inverse as:

$$\begin{aligned}\|V^{-1} - W^{-1}\| &= \|W^{-1}(W - V)V^{-1}\| \\ &\leq 2\left(\frac{8n^2}{\delta}\right)^2 \delta'\sqrt{n} \\ &\leq \frac{128n^{2.5}\delta'}{\delta^2}.\end{aligned}$$

Combining this with $\|D - C\| \leq \delta$ from Theorem 3.5.5, we have:

$$\begin{aligned}\|VDV^{-1} - WCW^{-1}\| &\leq \|(V - W)DV^{-1}\| + \|W(D - C)V^{-1}\| + \|WC(V^{-1} - W^{-1})\| \\ &\leq \delta'\sqrt{n} \cdot 5 \cdot \frac{16n^2}{\delta} + \sqrt{n}\delta'\frac{16n^2}{\delta} + \sqrt{n} \cdot 5 \cdot \frac{128n^{2.5}\delta'}{\delta^2} \\ &= \frac{\delta' n^{2.5}}{\delta}\left(5 \cdot 16 + 16 + \frac{5 \cdot 128}{\delta}\right) \\ &\leq \frac{\delta' n^{2.5}}{\delta^2} \cdot 6 \cdot 128\end{aligned}$$

which is at most $\delta/2$, for δ' chosen as above. We conclude that

$$\|A - VDV^{-1}\| \leq \|A - X\| + \|X - VDV^{-1}\| \leq \delta,$$

with probability $1 - 2/n - O(1/n^2)$ as desired.

To compute the running time and precision, we observe that SHATTER outputs a grid with parameters

$$\omega = \Omega\left(\frac{\delta^4}{n^5}\right), \quad \epsilon = \Omega\left(\frac{\delta^5}{n^9}\right).$$

Plugging this into the guarantees of EIG, we see that it takes

$$O\left(\log\frac{n}{\delta}\left(\log\frac{n}{\delta}+\log\log\frac{n}{\delta}\right)T_{\mathsf{MM}}(n)\right)=O(T_{\mathsf{MM}}(n)\log^2(n/\delta))$$

arithmetic operations, on a floating point machine with precision

$$O\left(\log^3\frac{n}{\delta}\log\frac{n}{\delta}\log n\right)=O(\log^4(n/\delta)\log(n))$$

bits, as advertised. □

3.5.1 Proof of Theorem 3.5.5

A key stepping-stone in our proof will be the following elementary result controlling the spectrum, pseudospectrum, and eigenvectors after perturbing a shattered matrix.

Lemma 3.5.8 (Eigenvector Perturbation for a Shattered Matrix)**.** Let $\Lambda_\epsilon(A)$ be shattered with respect to a grid whose squares have side length ω, and assume that $\|\widetilde{A}-A\|\le\eta<\epsilon$. Then, (i) each eigenvalue of $\widetilde{A}$ lies in the same grid square as exactly one eigenvalue of A, (ii) $\Lambda_{\epsilon-\eta}(\widetilde{A})$ is shattered with respect to the same grid, and (iii) for any right unit eigenvector $\widetilde{v}$ of $\widetilde{A}$, there exists a right unit eigenvector of A corresponding to the same grid square, and for which

$$\|\widetilde{v}-v\|\le\frac{\sqrt{8}\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}.$$

Proof. For (i), consider $A_t=A+t(\widetilde{A}-A)$ for $t\in[0,1]$. By continuity, the entire trajectory of each eigenvalue is contained in a unique connected component of $\Lambda_\eta(A)\subset\Lambda_\epsilon(A)$. For (ii), $\Lambda_{\epsilon-\eta}(\widetilde{A})\subset\Lambda_\epsilon(A)$, which is shattered by hypothesis. Finally, for (iii), let w^* and $\widetilde{w}^*$ be the corresponding left eigenvectors to v and $\widetilde{v}$ respectively, normalized so that $w^*v=\widetilde{w}^*\widetilde{v}=1$. From the contour integral definition of spectral projectors, if we call Γ the boundary of the grid square containing the eigenvalues associated to v and $\widetilde{v}$ respectively,

$$\begin{aligned}\|\widetilde{v}\widetilde{w}^*-vw^*\|&=\frac{1}{2\pi}\left\|\oint_\Gamma(z-A)^{-1}-(z-\widetilde{A})^{-1}\mathrm{d}z\right\|\\&=\frac{1}{2\pi}\left\|\oint_\Gamma(z-A)^{-1}(A-\widetilde{A})(z-\widetilde{A})^{-1}\mathrm{d}z\right\|\\&\le\frac{2\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}.\end{aligned}$$

Thus, using that $\|v\|=1$ and $w^*v=1$,

$$||\widetilde{v}\widetilde{w}^*-vw^*||\ge\|(\widetilde{v}\widetilde{w}^*-vw^*)v\|=\|(\widetilde{w}^*v)\widetilde{v}-v\|.$$

Now, since $(\widetilde{v}^*v)\widetilde{v}$ is the orthogonal projection of v onto the span of $\widetilde{v}$ we have that

$$\|(\widetilde{w}^*v)\widetilde{v} - v\| \geq \|(\widetilde{v}^*v)\widetilde{v} - v\| = \sqrt{1-(\widetilde{v}^*v)^2}.$$

Then, multiplying v by a phase we can assume without loss of generality that $\widetilde{v}^*v \geq 0$ which implies that

$$\sqrt{1-(\widetilde{v}^*v)^2} = \sqrt{(1-\widetilde{v}^*v)(1+\widetilde{v}^*v)} \geq \sqrt{1-\widetilde{v}^*v}.$$

The above discussion can now be summarized in the following chain of inequalities

$$\sqrt{1-\widetilde{v}^*v} \leq \sqrt{1-(\widetilde{v}^*v)^2} \leq \|(\widetilde{w}^*v)\widetilde{v} - v\| \leq \|\widetilde{v}\widetilde{w}^* - vw^*\| \leq \frac{2\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}.$$

Finally, note that $\|v-\widetilde{v}\| = \sqrt{2-2\widetilde{v}^*v} \leq \frac{\sqrt{8}\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}$ as we wanted to show. □

The algorithm EIG works by recursively reducing to subinstances of smaller size, but requires a pseudospectral guarantee to ensure speed and stability. We thus need to verify that the pseudospectrum does not deteriorate too subtantially when we pass to a sub-problem.

Lemma 3.5.9 (Compressing a Shattered Matrix)**.** Suppose P is a spectral projector of $A \in \mathbb{C}^{n\times n}$ of rank k and Q is an $n \times k$ matrix with $Q^*Q = I_k$ and $PQQ^* = QQ^*P$. Then for every $\epsilon > 0$,

$$\Lambda_\epsilon(Q^*AQ) \subset \Lambda_\epsilon(A).$$

Proof. Take $z \in \Lambda_\epsilon(Q^*AQ)$. Then, there exists $v \in \mathbb{C}^k$ satisfying $\|(z-Q^*AQ)v\| \leq \epsilon\|v\|$. Since $I_k = Q^*I_nQ$ we have

$$\|Q^*(z-A)Qv\| \leq \epsilon\|v\|.$$

Since Q^* is an isometry on range(Q) and $(z-A)Qv \in$ range(Q), we have $\|Q^*(z-A)Qv\| = \|(z-A)Qv\|$ and hence

$$\|(z-A)Qv\| \leq \epsilon\|v\| = \epsilon\|Qv\|,$$

showing that $z \in \Lambda_\epsilon(A)$. □

Observation 3.5.10. Since $\delta, \omega(\mathsf{g}), \epsilon \leq 1$, our assumption on η in Line 2 of the pseudocode of EIG implies the following bounds on η which we will use below:

$$\eta \leq \min\left\{0.02, \epsilon/75, \delta/100, \frac{\delta\epsilon^2}{200\omega(\mathsf{g})}\right\}.$$

Initial lemmas in hand, let us begin to analyze the algorithm. At several points we will make an assumption on the machine precision on the right hand side. These will be collected at the end of the proof, where we will verify that they follow from the precision hypothesis of Theorem 3.5.5.

Correctness.

Lemma 3.5.11 (Accuracy of $\widetilde{\lambda}_i$)**.** When DEFLATE succeeds, each eigenvalue of A shares a unique square of g with exactly one from $\widetilde{A_\pm}$, and $\Lambda_{4\epsilon/5}(\widetilde{A_\pm}) \subset \Lambda_\epsilon(A)$.

Proof. Let $P_\pm$ be the true projectors onto the two bisection regions found by $\mathsf{SPLIT}(A, \beta)$, $Q_\pm$ be the matrices whose orthogonal columns span their ranges, and $A_\pm := Q_\pm^* A Q_\pm$. From Theorem 3.5.3 the event that DEFLATE succeeds, the approximation $\widetilde{Q_\pm}$ that it outputs satisfies $\|\widetilde{Q_\pm} - Q_\pm\| \le \eta$, so in particular $\|\widetilde{Q_\pm}\| \le 2$ as $\eta \le 1$. The error $E_{6,\pm}$ from performing the matrix multiplications necessary to compute $\widetilde{A_\pm}$ admits the bound

$$\begin{aligned}
\|E_{6,\pm}\| &\le \mu_{\mathsf{MM}}(n)\|\widetilde{Q_\pm}\|\|A\widetilde{Q_\pm}\|\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\|\widetilde{Q_\pm}A\|\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\|\widetilde{Q_\pm}\|^2\|A\|\mathbf{u} \\
&\le 16\left(\mu_{\mathsf{MM}}(n)\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\mathbf{u}^2\right) && \|A\| \le 4 \text{ and } \|\widetilde{Q_\pm}\| \le 1 + \eta \le \\
&\le 3\eta && \mathbf{u} \le \frac{\eta}{10\mu_{\mathsf{MM}}(n)^2}.
\end{aligned}$$

Iterating the triangle inequality, we obtain

$$\begin{aligned}
\|\widetilde{A_\pm} - A_\pm\| &\le \|E_{6,\pm}\| + \|(\widetilde{Q_\pm} - Q_\pm)A\widetilde{Q_\pm}\| + \|Q_\pm A(\widetilde{Q_\pm} - Q_\pm\| \\
&\le 3\eta + 8\eta + 4\eta && \|\widetilde{Q_\pm} - Q_\pm\| \le \eta \\
&\le \epsilon/5 && \eta \le \epsilon/75.
\end{aligned}$$

We can now apply Lemma 3.5.8. □

Everything is now in place to show that, if every call to DEFLATE succeeds, EIG has the advertised accuracy guarantees. After we show this, we will lower bound this success probability and compute the running time.

When $A \in \mathbb{C}^{1\times 1}$, the algorithm works as promised. Assume inductively that EIG has the desired guarantees on instances of size strictly smaller than n. In particular, maintaining the notation from the above lemmas, we may assume that

$$(\widetilde{V_\pm}, \widetilde{D_\pm}) = \mathsf{EIG}(\widetilde{A_\pm}, 4\epsilon/5, \mathsf{g}_\pm, 4\delta/5, \theta, n)$$

satisfy (i) each eigenvalue of $\widetilde{D_\pm}$ shares a square of $\mathsf{g}_\pm$ with exactly one eigenvalue of $\widetilde{A_\pm}$, and (ii) each column of $\widetilde{V_\pm}$ is $4\delta/5$-close to a true eigenvector of $\widetilde{A_\pm}$. From Lemma 3.5.8, each eigenvalue of $\widetilde{A_\pm}$ shares a grid square with exactly one eigenvalue of A, and thus the output

$$\widetilde{D} = \begin{pmatrix} \widetilde{D_+} & \\ & \widetilde{D_-} \end{pmatrix}$$

satisfies the eigenvalue guarantee.

To verify that the computed eigenvectors are close to the true ones, let $\widetilde{\widetilde{v_\pm}}$ be some approximate right unit eigenvector of one of $\widetilde{A_\pm}$ output by EIG (with norm $1 \pm n\mathbf{u}$), $\widetilde{v_\pm}$ the

exact unit eigenvector of $\widetilde{A_\pm}$ that it approximates, and $v_\pm$ the corresponding exact unit eigenvector of $A_\pm$. Recursively, $\mathsf{EIG}(A, \epsilon, \mathsf{g}, \delta, \theta, n)$ will output an approximate unit eigenvector

$$\widetilde{v} := \frac{\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e}{\|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\|} + e',$$

whose proximity to the actual eigenvector $v := Qv_\pm$ we need now to quantify. The error terms here are e, a column of the error matrix E_8 whose norm we can crudely bound by

$$\|e\| \le \|E_8\| \le \mu_{\mathsf{MM}}(n)\|\widetilde{Q_\pm}\|\|\widetilde{V_\pm}\|\mathbf{u} \le 4\mu_{\mathsf{MM}}(n)\mathbf{u} \le \eta,$$

and e' is a column of E_9, the error incurred by performing the normalization in floating point; we assumed in (3.10) that $\|e'\| \le n\mathbf{u}$. The distance between $\widetilde{v}$ and $\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm}$ is just the difference in their norms—since they are parallel—so

$$\left\| \frac{\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e}{\|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\|} - \widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e \right\| \le \left| \|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\| - 1 \right| \le (1+\eta)(1+\mathbf{u}) + 4\mu_{\mathsf{MM}}\mathbf{u} - 1 \le 4\eta.$$

Inductively $\|\widetilde{\widetilde{v}_\pm} - \widetilde{\widetilde{v}_\pm}\| \le 4\delta/5$, and since $\|A_\pm - \widetilde{A_\pm}\| \le \epsilon/5$ and $A_\pm$ has shattered ϵ-pseudospectrum from Lemma 3.5.9, Lemma 3.5.8 ensures

$$\begin{aligned}
\|\widetilde{\widetilde{v}_\pm} - v_\pm\| &\le \frac{\sqrt{8}\omega(\mathsf{g}) \cdot 15\eta}{\pi \cdot \epsilon(\epsilon - 15\eta)} \\
&\le \frac{\sqrt{8}\omega(\mathsf{g}) \cdot 15\eta}{\pi \cdot 4\epsilon^2/5} && \eta \le \epsilon/75 \\
&\le \delta/10 && \eta \le \frac{\delta\epsilon^2}{200\omega(\mathsf{g})}.
\end{aligned}$$

Thus iterating the triangle identity and using $\|Q_\pm\| = 1$,

$$\begin{aligned}
\|\widetilde{v} - v\| &= \left\| \frac{\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e}{\|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\|} + e' - Q_\pm v_\pm \right\| \\
&\le \left\| \frac{\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e}{\|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\|} - \widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e \right\| + \|e'\| + \|e\| \\
&\quad + \|(\widetilde{Q_\pm} - Q_\pm)\widetilde{\widetilde{v}_\pm}\| + \|Q_\pm(\widetilde{\widetilde{v}_\pm} - \widetilde{v}_\pm)\| + \|Q_\pm(\widetilde{v}_\pm - v_\pm)\| \\
&\le 4\eta + n\mathbf{u} + \mu_{\mathsf{MM}}(n)\mathbf{u} + \eta(1 + n\mathbf{u}) + 4\delta/5 + \delta/10 \\
&\le 8\eta + 4\delta/5 + \delta/10 && n\mathbf{u}, \mu_{\mathsf{MM}}(n)\mathbf{u} \le \eta \\
&\le \delta && \eta \le \delta/200.
\end{aligned}$$

This concludes the proof of correctness of EIG.

Running Time and Failure Probability. Let's begin with a simple lemma bounding the depth of EIG's recursion tree.

Lemma 3.5.12 (Recursion Depth)**.** The recursion tree of EIG has depth at most $\log_{5/4} n$, and every branch ends with an instance of size 1×1.

Proof. By Theorem 3.5.2, SPLIT can always find a bisection of the spectrum into two regions containing $n_\pm$ eigenvalues respectively, with $n_+ + n_- = n$ and $n_\pm \geq 4n/5$, and when $n \leq 5$ can always peel off at least one eigenvalue. Thus the depth $d(n)$ satisfies

$$d(n) = \begin{cases} n & n \leq 5 \\ 1 + \max_{\theta \in [1/5, 4/5]} d(\theta n) & n > 5 \end{cases} \tag{3.35}$$

As $n \leq \log_{5/4} n$ for $n \leq 5$, the result is immediate from induction. □

We pause briefly and verify that the assumptions on $\delta < 1$, $\epsilon < 1/2$, and $\|A\| \leq 3.5$ in Theorem 3.5.5 ensure that every call to SPLIT throughout the algorithm satisfies the hypotheses in Theorems 3.5.2. Since δ, ϵ are non-increasing as we travel down the recursion tree of EIG, we need only verify for their initial settings. Theorem 3.5.2 needs $\epsilon < 1/2$, which is satisfied immediately, and we additionally have $\beta = \eta^4\theta^2/(20n)^6 \cdot 4n^8 \leq 1/20^6 n \leq 0.05/n$.

Finally, we need that every matrix passed to SPLIT throughout the course of the algorithm has norm at most 4. Lemma 3.5.11 shows that if $\|A\| \leq 4$ and has its ϵ-pseudospectrum shattered, then $\|\widetilde{A_\pm} - A_\pm\| \leq \epsilon/5$, and since $\|A_\pm\| = \|A\|$, this means $\|\widetilde{A_\pm}\| \leq \|A\| + \epsilon/5$. Thus each time we pass to a subproblem, the norm of the matrix we pass to EIG (and thus to SPLIT) increases by at most $\epsilon/5$. Since ϵ decreases by a factor of $4/5$ on each recursion, this means that by the end of the algorithm the norm of the matrix passed to EIG will be at most $\frac{1}{5 \cdot (1 - 4\epsilon/5)} \leq \epsilon \leq 1/2$. Thus we will be safe if our initial matrix has norm at most 3.5, as assumed.

Lemma 3.5.13 (Lower Bounds on the Parameters)**.** The input parameters given to every recursive call $\mathsf{EIG}(A', \delta', \mathsf{grid}', \epsilon', \theta, n)$ and $\mathsf{SPLIT}(A' - h', \epsilon', \mathsf{g}', \beta')$ satisfy

$$\delta' \geq \delta/n \qquad \epsilon' \geq \epsilon/n \qquad \eta \geq \frac{\delta\epsilon^2}{200n^3}4 \qquad \beta \geq \frac{\theta^2\delta^4\epsilon^8}{(5n)^{26}}.$$

Proof. Along each branch of the recursion tree, we replace $\epsilon \leftarrow 4\epsilon/5$ and $\delta \leftarrow 4\delta/5$ at most $\log_{5/4} n$ times, so each can only decrease by a factor of n from their initial settings. □

Lemma 3.5.14 (Failure Probability)**.** EIG fails with probability no more than θ.

Proof. Since each recursion splits into at most two subproblems, and the recursion tree has depth $\log_{5/4} n$, there are at most

$$2 \cdot 2^{\log_{5/4} n} = 2n^{\frac{\log 2}{\log 5/4}} \leq 2n^4$$

calls to DEFLATE. We have set every η and β so that the failure probability of each is $\theta/2n^4$, so a crude union bound finishes the proof. □

The arithmetic operations required for EIG satisfy the recursive relationship

$$\begin{aligned} T_{\mathsf{EIG}}(n,\delta,\mathsf{g},\epsilon,\theta,n) \leq\ & T_{\mathsf{SPLIT}}(n,\epsilon,\beta) + T_{\mathsf{DEFLATE}}(n,\beta,\eta) + 2T_{\mathsf{MM}}(n) \\ & + T_{\mathsf{EIG}}(n_+, 4\delta/5, \mathsf{g}_+, 4\epsilon/5, \theta, n) + T_{\mathsf{EIG}}(n_-, 4\delta/5, \mathsf{g}_-, 4\epsilon/5, \theta, n) \\ & + 2T_{\mathsf{MM}}(n) + O(n^2). \end{aligned}$$

Each of the $T_\circ$ terms is of the form $\mathrm{polylog}(n)\mathrm{poly}(n)$, where both polynomials have nonnegative coefficients, and the exponent on n is at least 2. Thus, when we split into problems of sizes $n_+ + n_- = n$ and $n_\pm \geq 4n/5$, by convexity $T_\circ(n_+,...) + T_\circ(n_-,...) \leq \frac{4^2+1^2}{5^2} T_\circ(n,...) = \frac{16}{25} T_\circ(n,...)$. Recursively then, if we were to keep all accuracy parameters fixed, the total cost of the operations we perform in each layer is at most 16/25 times the cost of the previous one. Using our parameter lower bounds from Lemma 3.5.13, and these geometrically decreasing bit operations, we then have

$$\begin{aligned} T_{\mathsf{EIG}}(n,\delta,\mathsf{g},\epsilon,\theta,n) &\leq \frac{25}{8}\left(T_{\mathsf{SPLIT}}\left(n,\epsilon/n,\mathsf{g},\frac{\delta^4\epsilon^8\theta^2}{(5n)^{26}}\right)\right. \\ &\qquad \left. + T_{\mathsf{DEFLATE}}\left(n,\beta/n,\epsilon/n,\frac{\delta^4\epsilon^8\theta^2}{(5n)^{26}}\right) + 4T_{\mathsf{MM}}(n) + O(n^2)\right) \\ &= \frac{25}{8}\left(12N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})}\left(T_{\mathsf{INV}}(n) + O(n^2)\right) + 2T_{\mathsf{QR}}(n)\right. \\ &\qquad \left. + 5T_{\mathsf{MM}}(n) + n^2 T_{\mathsf{N}} + O(n^2)\right) \\ &\leq 60N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})}\left(T_{\mathsf{INV}}(n) + O(n^2)\right) + 10T_{\mathsf{QR}}(n) + 25T_{\mathsf{MM}}(n) \end{aligned}$$

where

$$N_{\mathsf{EIG}} := \lg \frac{256n}{\epsilon} + 3\lg\lg\frac{256n}{\epsilon} + \lg\lg\frac{(5n)^{26}}{\theta^2\delta^4\epsilon^9}.$$

In the final expression for T_{EIG} we have used the fact that $T_{\mathsf{N}} = O(1)$. Thus we have

$$T_{\mathsf{EIG}}(n,\delta,\mathsf{g},\epsilon,\theta,n) = O\left(\log\frac{1}{\omega(\mathsf{g})}\left(\log\frac{n}{\epsilon} + \log\log\frac{1}{\theta\delta}\right) T_{\mathsf{MM}}(n,\mathbf{u})\right),$$

by Theorem 3.2.6.

Required Bits of Precision. We will need the following bound on the norms of all spectral projectors.

Lemma 3.5.15 (Sizes of Spectral Projectors)**.** Throughout the algorithm, every approximate spectral projector $\widetilde{P}$ given to $\mathsf{DEFLATE}$ satisfies $\|\widetilde{P}\| \leq 10n/\epsilon$.

Proof. Every such $\widetilde{P}$ is β-close to a true spectral projector P of a matrix whose ϵ/n-pseudosepctrum is shattered with respect to the initial 8×8 unit grid $\mathbf{g}$. Since we can generate P by a contour integral around the boundary of a rectangular subgrid, we have

$$\|\widetilde{P}\| \leq 2 + \|P\| \leq 2 + \frac{32}{2\pi}\frac{n}{\epsilon} \leq 10n/\epsilon,$$

with the last inequality following from $\epsilon < 1$. □

Collecting the machine precision requirements $\mathbf{u} \leq \mathbf{u}_{\mathsf{SPLIT}}, \mathbf{u}_{\mathsf{DEFLATE}}$ from Theorems 3.5.2 and 3.5.3, as well as those we used in the course of our proof so far, and substituting in the parameter lower bounds from Lemma 3.5.13, we need $\mathbf{u}$ to satisfy

$$\mathbf{u} \leq \min \left\{ \begin{array}{l} \dfrac{\left(1 - \frac{\epsilon}{256n}\right)^{2^{N_{\mathsf{EIG}}+1}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}}, \\ \dfrac{\epsilon}{100n^2}, \dfrac{\theta^2\delta^4\epsilon^8}{(5n)^{26}}\dfrac{1}{4\|\widetilde{P}\|\max\{\mu_{\mathsf{QR}}(n), \mu_{\mathsf{MM}}(n)\}}, \\ \dfrac{\delta\epsilon^2}{100n^3 \cdot 2\mu_{\mathsf{QR}}(n)}, \dfrac{\delta\epsilon^2}{100n^3\max\{4\mu_{\mathsf{MM}}(n), n, 2\mu_{\mathsf{QR}}(n)\}} \end{array} \right\}$$

From Lemma 3.5.15, $\|\widetilde{P}\| \leq 10n/\epsilon$, so the conditions in the second two lines are all satisfied if we make the crass upper bound

$$\mathbf{u} \leq \frac{\theta^2\delta^4\epsilon^8}{(5n)^{30}}\frac{1}{\max\{\mu_{\mathsf{QR}}(n), \mu_{MM}(n), n\}},$$

i.e. if $\lg 1/\mathbf{u} \geq O\left(\lg \frac{n}{\theta\delta\epsilon}\right)$. Unpacking the first requirement and using the definition of N_{EIG} and $1/2 \leq (1-x)^{2^{\lg x}}$ for $x \in (0,1)$, we have

$$\begin{aligned} \frac{\left(1 - \frac{\epsilon}{256n}\right)^{2^{N_{\mathsf{EIG}}+1}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}} &= \frac{\left(\left(1 - \frac{\epsilon}{256n}\right)^{\frac{256n}{\epsilon}}\right)^{\lg^3 \frac{256n}{\epsilon} \lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8} 2^{8.59}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}} \\ &\geq \frac{2^{-\lg^3 \frac{256n}{\epsilon} \lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8} 2^{8.59}(c_{\mathsf{INV}} \log n + 3)}}{2\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}}, \end{aligned}$$

so the final expression is a sufficient upper bound on $\mathbf{u}$. This gives

$$\begin{aligned} \lg 1/\mathbf{u} &\geq \lg^3 \frac{n}{\epsilon} \lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8} 2^{14.83}(c_{\mathsf{INV}} \log n + 3) + \lg N_{\mathsf{EIG}} \\ &= O\left(\log^3 \frac{n}{\epsilon} \log \frac{n}{\theta\delta\epsilon} \log n\right). \end{aligned}$$

This dominates the precision requirement above, and completes the proof of Theorem 3.5.5.

3.6 Conclusion and Open Questions

In this chapter, we reduced the approximate diagonalization problem to a polylogarithmic number of matrix multiplications, inversions, and QR factorizations on a floating point machine with precision depending only polylogarithmically on n and $1/\delta$. The key phenomena enabling this were: (a) every matrix is δ-close to a matrix with well-behaved pseudospectrum, and such a matrix can be found by a complex Gaussian perturbation. (b) The spectral bisection algorithm can be shown to converge rapidly to a forward approximate solution on such a well-behaved matrix, using a polylogarithmic in n and $1/\delta$ amount of precision and number of iterations. The combination of these facts yields a δ-backward approximate solution for the original problem.

Using fast matrix multiplication, we obtain algorithms with nearly optimal asymptotic computational complexity (as a function of n, compared to matrix multiplication), for general complex matrices with no assumptions. Using naive matrix multiplication, we get easily implementable algorithms with $O(n^3)$ type complexity and much better constants which are likely faster in practice. The constants in our bit complexity and precision estimates, while not huge, are likely suboptimal. The reasonable practical performance of spectral bisection based algorithms is witnessed by the many empirical papers (see e.g. [8]) which have studied it. The more recent of these works further show that such algorithms are communication-avoiding and have good parallelizability properties.

Remark 3.6.1 (Hermitian Matrices)**.** A curious feature of our algorithm is that even when the input matrix is Hermitian or real symmetric, it begins by adding a complex non-Hermitian perturbation to regularize the spectrum. If one is only interested in this special case, one can replace this first step by a Hermitian GUE or symmetric GOE perturbation and appeal to the result of [1] instead of Theorem 3.1.4, which also yields a polynomial lower bound on the minimum gap of the perturbed matrix. It is also possible to obtain a much stronger analysis of the Newton iteration in the Hermitian case, since the iterates are all Hermitian and $\kappa_V = 1$ for such matrices. By combining these observations, one can obtain a running time for Hermitian matrices which is significantly better (in logarithmic factors) than our main theorem. We do not pursue this further since our main goal was to address the more difficult non-Hermitian case.

We conclude by listing several directions for future research.

1. Devise a deterministic algorithm with similar guarantees. The main bottleneck to doing this is deterministically finding a regularizing perturbation, which seems quite mysterious. Another bottleneck is computing a rank-revealing QR factorization in near matrix multiplication time deterministically (all of the currently known algorithms require $\Omega(n^3)$ time).

2. Determine the correct exponent for smoothed analysis of the eigenvalue gap of $A + \gamma G$ where G is a complex Ginibre matrix. We currently obtain roughly $(\gamma/n)^{8/3}$ in Theorem

3.3.6. Is it possible to match the $n^{-4/3}$ type dependence [130] which is known for a pure Ginibre matrix?

3. Reduce the dependence of the running time and precision to a smaller power of $\log(1/\delta)$. The bottleneck in the current algorithm is the number of bits of precision required for stable convergence of the Newton iteration for computing the sign function. Other, "inverse-free" iterative schemes have been proposed for this, which conceivably require lower precision.

4. Study the convergence of "scaled Newton iteration" and other rational approximation methods (see [71, 98]) for computing the sign function on non-Hermitian matrices. Perhaps these have even faster convergence and better stability properties?

More broadly, we hope that the techniques introduced in this chapter—pseudospectral shattering and pseudospectral analysis of matrix iterations using contour integrals—are useful in attacking other problems in numerical linear algebra.

Chapter 4

The Lanczos Algorithm Under Few Iterations

4.1 Background

Eigenvalue problems are ubiquitous in science and engineering. However, most applications require analyzing matrices whose large dimension makes it impractical to exactly compute any important feature of their spectrum. It is for this reason that iterative randomized algorithms have proliferated in numerical linear algebra [107, 124].

In this context, iterative randomized algorithms provide an approximation of the spectrum of the matrix in question, where the accuracy of the approximation improves as the number of iterations increases. For any such algorithm, it is natural to ask the following questions:

(Q1) How much does the random output vary?

(Q2) How many iterations are necessary and sufficient to obtain a satisfactory approximation?

This chapter addresses the above questions for one of the most widely used algorithms for eigenvalue approximation, namely the Lanczos algorithm. Throughout, we assume exact arithmetic.

4.1.1 The Lanczos algorithm

Recall from Section 1.1 that when run for k iterations, the Lanczos algorithm outputs a $k \times k$ matrix, called the *Jacobi matrix*, then, the eigenvalues of the Jacobi matrix, namely the *Ritz values*, are used as an approximation for the spectrum of the matrix. In particular, when $k = n$, the Ritz values are exactly the eigenvalues of A, and hence the full spectrum is recovered. However, in practice it is usually too expensive to perform $\Theta(n)$ iterations.

The success of the Lanczos algorithm resides to some extent in its ability to find the *outliers* of the spectrum of the matrix A with very few iterations. By outliers, we mean

the eigenvalues distant from the region in which the majority of the spectrum accumulates (the *bulk*). Hence, the algorithm is of particular interest in most applications in science and engineering [107].

Lanczos-type methods can also be used to approximate the global spectral density of large matrices—for a survey of techniques see [88]. In applied mathematics, large matrices often arise as discretizations of infinite-dimensional operators such as the Laplacian. Computing the eigenvalues of the finite-dimensional operator then yields information about the infinite-dimensional operator and the underlying continuous system. For an example, see Section 7 of [126] for numerical experiments and bounds for the Lanczos algorithm applied to an explicit discretized Laplace operator.

In applications, sophisticated modifications of the Lanczos algorithm are used [63, 31, 87]. Since the goal of this chapter is to introduce proof techniques and theoretical tools that have not been exploited previously, we only deal with the simplest version of the Lanczos algorithm and do not strive to obtain optimal constants in our bounds and theorems when providing answers for questions (1) and (2).

4.1.2 Question (1): Our contributions

As far as we are aware, there is no previous work addressing this question for the Lanczos algorithm. In this chapter we show that there is a $c > 0$ depending on a global feature of the spectrum of the matrix, such that for n large enough, the output of the Lanczos procedure is almost deterministic when run for at most $c \log n$ iterations. More precisely, in Theorem 4.2.2 we show that for $\varepsilon \in (0, 1/2)$ deviations of the order $n^{-\varepsilon}$ occur with exponentially small probability.

From the point of view of random matrix theory, the problem treated in this chapter is atypical. In random matrix theory, most of the studied models have a rich probabilistic structure that can be exploited to obtain results about the eigenvalue distribution of the matrix. By contrast, in our case, the Jacobi matrix output by the Lanczos algorithm is a random matrix obtained by running a complicated deterministic dynamic over a minimal source of randomness—a single uniform random unit vector. Hence, in order to obtain results similar to the ones presented in this article, the structure of the algorithm needs to be exploited in an involved way. We use the ubiquitous concentration of measure phenomenon for Lipschitz functions in high dimension, together with a careful control of the variables appearing in the Lanczos algorithm and their Lipschitz constants as functions of the random input. Roughly speaking, the Lipschitz constant is exponential in the number of iterations, which yields concentration in the regime of at most $c \log n$ iterations for sufficiently small c. Throughout the analysis we use elementary results in the theory of orthogonal polynomials.

In view of the fact that the output of the Lanczos algorithm is sharply concentrated under few iterations, one may ask which values the output is concentrated around. Towards the end of this introduction we give an overview of our results in this direction.

4.1.3 Question (2): Previous work

For the Lanczos algorithm, theoretical answers to the sufficiency part of Question (2) posed above appeared decades ago. Most of them in essence give an *upper bound* on the number of iterations required to approximate the outliers of the spectrum of an n-dimensional matrix A with great accuracy–see [77, 101, 108]. Roughly speaking, previous literature provides inequalities that state that $k \geq C \log n$ iterations suffice for the output of the Lanczos algorithm approximates very well the true extreme eigenvalues of A, making the use of $O(\log n)$ iterations common in practice–see [81] or [126] for examples of inequalities that give this bound. The constant C in the results mentioned above is determined by features of the spectrum of A; typically, these features are the diameter of the spectrum and the gaps between the outliers and the bulk. In recent years, more refined arguments have yielded inequalities in which other features of the spectrum are considered, see [134] for an example or [20] for a survey.

Regarding the necessity part of Question (2), to the best of our knowledge, the only existing negative result regarding detection of outliers is the one given in the recent work [111]. There, a query complexity bound was proven for any algorithm that is allowed to make queries of matrix-vector products, which in particular applies to the Lanczos algorithm.

4.1.4 Question (2): Our contributions

In this chapter we study the Lanczos algorithm in the context of approximation of outliers, and answer the necessity part of Question (2). That is, we show that if run for at most $k \leq c \log n$ iterations, the Lanczos algorithm fails to approximate outliers with overwhelming probability. Thus, in essence we provide a *lower bound* on the number of iterations required for accuracy. The aforementioned c depends only on an easily computed global property of the spectrum which we call *equidistribution*.

To give some rough context, the result in [111] discussed above shows that if the empirical spectral distribution of a matrix is close to the semicircle distribution plus an outlying "spike," any algorithm in their class will fail to identify the spike with overwhelming probability, unless given at least $c \log n$ queries. In contrast, our result applies exclusively to the Lanczos algorithm, but shows that outliers are missed for a far more general class of measures than just the semicircle.

In order to analyze asymptotic behavior, we adopt a similar framework to that used in [83] and [19], in which a sequence of Hermitian matrices A_n with convergent spectra was considered. These papers studied the behavior of the Lanczos algorithm in the regime of $\Theta(n)$ iterations.

To show that the Lanczos algorithm misses outliers when run for at most $c \log n$ iterations, we use the elementary theory of orthogonal polynomials and standard techniques in high-dimensional probability. Roughly speaking, using a variational principle, we show that for small enough k, the roots of the kth orthogonal polynomial with respect to a certain random measure are contained in a small blow-up of the convex hull of the bulk of the true spectrum.

See Theorem 4.2.10 or Proposition 4.2.12 for a precise statement and Figure 4.1 for an illustration.

4.1.5 Our result on the locations of the Ritz values

One may ask if finer statements about the location of the Ritz values can be made. Previously, tools from potential theory have been used to answer this question in the regime of $k = \Theta(n)$ iterations [19, 83, 82]. Results in the regime of fixed k as $n \to \infty$ in the deterministic setting of orthogonal polynomials follow from [59, §4]. In the present work we use determinantal formulas for orthogonal polynomials and concentration of measure results to locate the Ritz values in the regime of $k = O(\sqrt{\log n})$ iterations. In particular, we prove that the Ritz values concentrate around the roots of the kth orthogonal polynomials for the limiting eigenvalue distribution. See Figure 4.2 for an illustration. Moreover, also when $k = O(\sqrt{\log n})$, we show that the Jacobi matrix obtained after k iterations is concentrated around the kth Jacobi matrix of the limiting measure.

These results may be of particular relevance in applications where an infinite dimensional operator is discretized with the goal of computing its density. In essence, Theorem 4.2.13 below states that in this situation the first iterations of the Lanczos algorithm are an accurate approximation of the true Jacobi coefficients of the spectral measure of the infinite dimensional operator, and hence the procedure is giving valuable information for recovering the limiting measure.

4.1.6 Organization

The chapter is organized as follows. In Section 4.2, we review the classical background of the Lanczos procedure and orthogonal polynomials, and formally state our main theorems. In Section 4.3, we develop machinery that in Section 4.4 will be used to prove concentration for the output of the Lanczos algorithm. In Section 4.5, we prove our complementary results about the location of the Ritz values and Jacobi coefficients. Finally, in Section 4.6 we discuss further research directions that may be of interest.

4.2 Preliminaries and statements of theorems

Throughout this chapter only elementary facts about orthogonal polynomials are used. For the reader's convenience in Section 4.2 we include a concise survey of the results that will be used in the sequel. Chapter 2 in [121] and Chapters 2 and 3 in [41] are introductory references containing these results.

In Section 1.1, we have described the Lanczos algorithm and its interpretation in terms of orthogonal polynomials. Some standard references for this matter are Chapter 6 in [124] and Chapter 6 in [107].

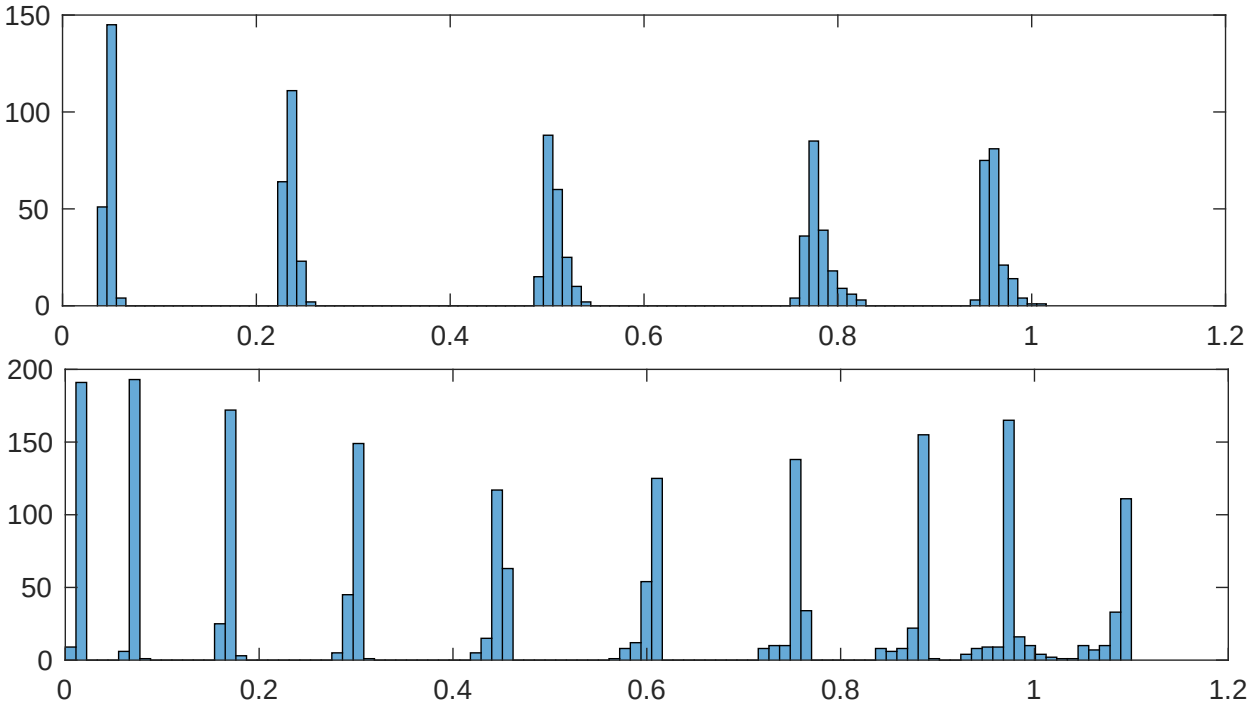

Figure 4.1: A is a 2000×2000 diagonal matrix with entries $\{0, 1/2000, 2/2000, \ldots, 1999/2000, 1.1\}$. This represents a discretization of Unif($[0, 1]$) plus an outlier at 1.1. Plotted is a histogram of the Ritz values output by Lanczos after $k = 5$ iterations (above) and after $k = 10$ iterations (below). To generate the histogram the procedure was run 200 times. Notice that to find the outlier with a decent probability, 10 iterations suffice (but 5 do not).

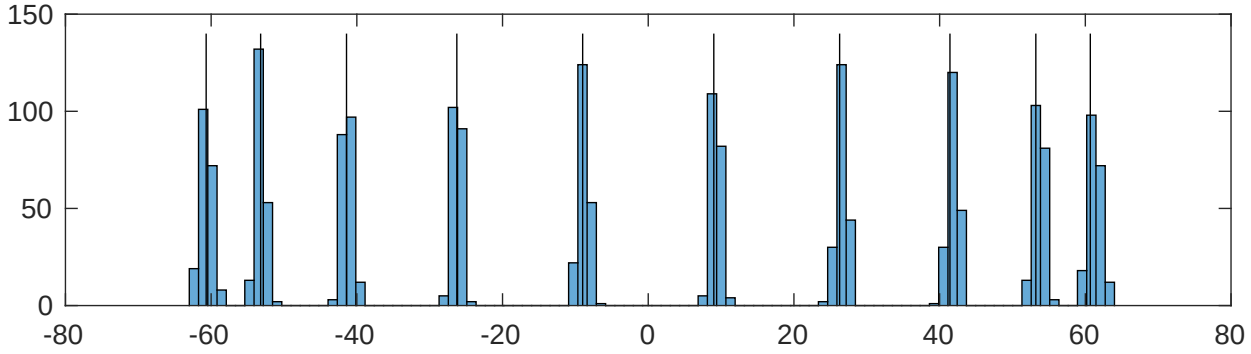

Figure 4.2: A is a fixed $n \times n$ matrix drawn from the Gaussian Orthogonal Ensemble (GOE) with $n = 2000$. Plotted is the histogram of the Ritz values after 200 repetitions of the Lanczos algorithm with $k = 10$ iterations. Also plotted are the roots of the 10th orthogonal polynomial with respect to the (suitably rescaled) semicircle law, which is the limit of the distribution of eigenvalues for GOE as $n \to \infty$.

In Section 2.3 we introduce the framework in which this chapter is developed and formally state the main contributions of our work.

4.2.1 Statement of results

Sections 4.3 and 4.4 are devoted to proving concentration results for the output of the Lanczos algorithm, Algorithm 1. In these sections, the input matrix will be fixed and denoted by A. We will use n to denote the dimension of A and usually the number of iterations of the procedure will be denoted by k. Note that the Jacobi coefficients α_i and β_i are assigned during the ith iteration of the algorithm and are unchanged during future iterations.

Since for our analysis it is necessary to compare outputs of the algorithm resulting from different input vectors $u \in \mathbb{S}^{n-1}$, we will stress this dependence by viewing the respective quantities as a function of u and denoting them by $\alpha_i(u), \beta_i(u), r_i(u), \gamma_i(u)$, $p_k^u(x)$, $v_i(u)$ and $J_k(u)$. Depending on the context, the aforementioned quantities will also be thought as random variables, random polynomials, random vectors and random matrices, respectively. One of the main steps in the proof of our concentration result, Theorem 4.2.2, consists of showing that these quantities are somehow stable under perturbations of the input vector.

For Theorem 4.2.2 a technical feature of the global behaviour of the spectrum is taken into account. Intuitively, we want to say that the spectrum is *equidistributed* if it is not grouped in a small number of small clusters (see Examples 4.2.4 and 4.2.5 below). As we show in Section 4.4, the family of well equidistributed point sets includes, but it is not limited to, those sets obtained by discretizing an absolutely continuous distribution. We use two parameters, δ and ω, to quantify how well-distributed the spectrum of a matrix is. We motivate and develop this notion in Section 4.4.

Definition 4.2.1 (Equidistribution)**.** Let Λ be any finite set of n real numbers. Let δ and ω be positive real numbers and let j be a natural number. We say that Λ is (δ, ω, j)-*equidistributed* if for any finite set T of at most j real numbers it holds that

$$\left| \left\{ \lambda \in \Lambda : \frac{1}{|T|} \sum_{t \in T} \log |\lambda - t| \geq \log \omega \right\} \right| \geq \delta n.$$

Theorem 4.2.2 (Concentration of Jacobi coefficients after i iterations)**.** Suppose the spectrum of A is (δ, ω, i)-equidistributed for some $\delta, \omega > 0$ and $i \in \mathbb{N}$. Let $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ denote the medians of the Jacobi coefficients $\alpha_i(u)$ and $\beta_i(u)$ respectively. Then for all $t > 0$, the quantities $\mathbb{P}[|\alpha_i(u) - \tilde{\alpha}_i| > t\|A\|]$ and $\mathbb{P}[|\beta_i(u) - \tilde{\beta}_i| > t\|A\|]]$ are both bounded above by

$$2 \exp\left\{ -\frac{\min\{\delta, 1/50\}^2}{32} n \right\} + 2 \exp\left\{ -\frac{1}{64} \left(\frac{\omega}{4\|A\|} \right)^{2i} \delta^2 t^2 n \right\}. \tag{4.1}$$

Remark 4.2.3. The constants δ, ω appearing in the above theorem are typically quite moderate in magnitude, and are easy to compute if one can obtain explicit bounds for certain

integrals with respect to the spectral distribution of A. Besides the examples provided below, in Section 4.4.1 we give more examples and a detailed discussion on how to compute these quantities.

Example 4.2.4. Let Λ be the set of n equally spaced points from $1/n$ to 1, inclusive. This represents a discretization of the uniform measure $\mu = \mathrm{Unif}([0,1])$. In Section 4.4.1, we will show that for $j \leq \frac{n}{16}$, the set Λ is (δ, ω, j)-equidistributed for $\delta = 1/4$ and $\omega = 4e^{-2}$.

Example 4.2.5. Now consider a set (or multiset) Λ of $n > 0$ points grouped in m equally spaced small clusters. To make this precise, fix two parameters $\varepsilon, g > 0$ and consider $-1 = a_1 \leq b_1 < a_2 \leq b_2 < \cdots < a_m \leq b_m = 1$ such that for every $i = 1, \ldots, m$ we have $b_i - a_i = \varepsilon$ and $a_{i+1} - b_i = g$ (we think of ε as *small* with respect to g and of m as *small* with respect to n). If $\Lambda \subset \bigcup_{i=1}^{m} [a_i, b_i]$ with $|\Lambda \cap [a_i, b_i]| \geq \lfloor \frac{n}{m} \rfloor$ for every $i = 1, \ldots, m$, then Λ is $(\frac{m-j}{m}, g, j)$-equidistributed and $g \approx 2/m$.

Note that in this case we have good equidistribution parameters unless $j \approx m$. In Section 4.4 we give a generalization of this assertion in Observation 4.4.9.

Theorem 4.2.2 yields concentration of the entries of the random matrix $J_k(u)$. In general, controlling the entries of a random matrix does not yield control over its random eigenvalues or eigenvectors. However, since $J_k(u)$ is symmetric we know that its spectrum is stable with respect to small perturbations of the entries, and under some conditions its eigenvectors are stable as well. More precisely, we now invoke two classic results in perturbation theory, Weyl's inequality and the Davis-Kahan theorem [39]. See [128] or [75] for more modern references and (1.65) in [122] for a generalization of this case of Weyl's inequality from ℓ^∞ to ℓ^p that might be of interest in this context.

Lemma 4.2.6 (Weyl)**.** For every matrix X, let $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$ denote the eigenvalues of X. If A and B are $n \times n$ Hermitian matrices, then for all $1 \leq i \leq n$ we have

$$|\lambda_i(A + B) - \lambda_i(A)| \leq \|B\|.$$

Theorem 4.2.7 (Davis-Kahan)**.** Here we use the notation of Lemma 4.2.6. Fix $i \in \{1, \ldots, n\}$ and assume that $\lambda_i(A)$ has multiplicity 1. Define

$$\varepsilon = \min_{j : j \neq i} |\lambda_i(A) - \lambda_j(A)|,$$

and let $\theta \in [0, \pi/2]$ denote the angle between the i-th eigenvectors of A and $A + B$. Then

$$\sin\theta \leq \frac{2\|B\|}{\varepsilon}.$$

Following the notation in Theorem 4.2.2, let $\tilde{J}_k$ be the $k \times k$ Jacobi matrix with entries $\tilde{\alpha}_i$ and $\tilde{\beta}_i$, and denote the eigenvalues of $\tilde{J}_k$ by $\tilde{r}_1 \geq \cdots \geq \tilde{r}_k$. Also, let $\tilde{w}_i$ be the eigenvector of $\tilde{J}_k$ corresponding to $\tilde{r}_i$ and let $w_i(u)$ be the eigenvector of $J_k(u)$ corresponding to $r_i(u)$. Since

$J_k(u)$ concentrates around $\tilde{J}_k$, by the Weyl inequality and the Davis-Kahan theorem, the Ritz values $r_i(u)$ will concentrate around their medians $\tilde{r}_i$ and, provided that $\tilde{r}_i$ is sufficiently separated from the rest of the r_j, $w_i(u)$ will concentrate around $\tilde{w}_i$. Indeed, at the end of Section 4.4 we show the following proposition.

Proposition 4.2.8 (Concentration of the Ritz values)**.** Assume that the spectrum of A is (δ, ω, k)-equidistributed for some $\delta, \omega > 0$ and $k \in \mathbb{N}$. With the notation described above, let $\vec{r} = (\tilde{r}_1, \ldots, \tilde{r}_k)$ and let $\vec{r}(u) = (r_1(u), \ldots, r_k(u))$ be the vector of Ritz values after k iterations. Then

$$\begin{aligned}&\mathbb{P}[\|\vec{r}(u) - \vec{r}\|_\infty \geq t\|A\|] \\ &\leq 4k\left[\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + \exp\left\{-\frac{1}{192}\left(\frac{\omega}{4\|A\|}\right)^{2k}\delta^2 t^2 n\right\}\right].\end{aligned}$$

Proposition 4.2.9 (Concentration of the Ritz vectors)**.** Assume that the spectrum of A is (δ, ω, k)-equidistributed for some $\delta, \omega > 0$ and $k \in \mathbb{N}$ and fix some $i \in \mathbb{N}$ with $1 \leq i \leq k$. With the notation described above, let $\theta \in [0, \pi/2]$ be the angle between $w_i(u)$ and $\tilde{w}_i$ and let $\varepsilon = \min_{j:j\neq i} |\tilde{r}_i - \tilde{r}_j|$. Then for any $0 \leq c < 1/2$ we have

$$\begin{aligned}&\mathbb{P}\left[\sin\theta \geq \frac{2\|A\|}{\varepsilon n^c}\right] \\ &\leq 4k\left[\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + \exp\left\{-\frac{1}{192}\left(\frac{\omega}{4\|A\|}\right)^{2k}\delta^2 n^{1-2c}\right\}\right].\end{aligned}$$

Note: The same result holds for the Ritz vectors, since these are obtained by applying an isometry to the $w_i(u)$.

Theorem 4.2.2, Proposition 4.2.8 and Proposition 4.2.9 above show that the Lanczos algorithm is almost deterministic when the number of iterations is a fraction of the logarithm of the dimension of A.

The results above regard concentration, but do not say anything about the locations of the medians that our Ritz values and Jacobi coefficients concentrate around. The rest of the chapter focuses on studying the locations of these quantities. In section 4.5.1, we show that if k is a certain fraction of $\log n$, the Ritz values obtained after k iterations are contained in a small blow-up of the convex hull of the bulk of the spectrum of A. This complements classical guarantees which show that for some multiple of $\log n$, say K, the Lanczos algorithm approximates with high accuracy the outliers of the spectrum of A when K iterations are performed. Our results are quantitative and use our notion of equidistribution.

Theorem 4.2.10. Suppose the spectrum of A is (δ, ω, j)-equidistributed for some $\delta, \omega > 0$ and $j \in \mathbb{N}$. Let M be the diameter of the spectrum of A. Let R be a real number and let $0 < c < 1/2$, and suppose there are at most $m \leq \min\{0.02n, 2n^\alpha\}$ "outliers", eigenvalues of

A lying above R, for some $\alpha < 1 - c$. Let $g = \max_{1 \le i \le n}\{\lambda_i - R\}$ and let $\kappa > 0$. Then for up to

$$k = \min\left\{j, \frac{1}{2\log\frac{M}{\omega}}\left(c\log n + \log\frac{\kappa\delta}{2mg}\right)\right\}$$

iterations, the probability that the top Ritz value exceeds $R + \kappa$ is at most

$$2\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + 2\exp\left\{-\frac{1}{16}n^{1-2c}\right\}$$

for $n > e^{\frac{1}{1-c-\alpha}}$.

The strength of the above result might be obscured by the appearance of several unintuitive parameters. For the reader's benefit we include an example below, and to provide a slightly different perspective, we include an asymptotic version of the above result, namely Proposition 4.2.12.

Example 4.2.11. Let $n > 0$ and let A be a matrix whose spectrum consists of $n - 1$ equally spaced points from $2/n$ to 1 inclusive, together with an outlier of value 1.1 (compare with Figure 4.1). In Section 4.4.1 we will show that for $j \le n/16$ the spectrum of A is $(1/4, 4e^{-2}, j)$-equidistributed.

In order to apply Theorem 4.2.10, we also note that in this case $M = 1.08$, $m = 1$ and $g = 10^{-1}$. Take $\kappa = 10^{-4}$. Then, for any $0 < c < 1/2$, the Ritz values of the Lanczos algorithm on A after $\lfloor \frac{7c}{10}\log n - 7/2 \rfloor$ iterations will be contained in the interval $[2/n, 1 + 10^{-4}]$ with overwhelming probability.

Proposition 4.2.12. Let $(A_n)_{n=1}^{\infty}$ be a sequence of $n \times n$ Hermitian matrices with uniformly bounded norm. Assume their empirical spectral distributions μ_n converge in distribution to a measure μ with nontrivial absolutely continuous part, and further assume $\mathrm{Kol}(\mu_n, \mu) = O(1/\log n)$. Suppose there exists $m \in \mathbb{N}$ such that each A_n has at most m eigenvalues ("outliers") greater than R, where R denotes the right edge of the support of μ.

Then there exists $c > 0$ such that for every $\kappa > 0$, the Ritz values of Lanczos applied to A_n after $c\log n$ iterations are bounded above by $R + \kappa$ with overwhelming probability for n sufficiently large (depending on how small the gap κ is chosen.)

Finally, we give a result about the locations of the Ritz values and Jacobi coefficients when at most $d\sqrt{\log n}$ iterations are performed, with d depending only on μ and the speed of convergence of the sequence μ_n. Essentially, we show that in this regime the Jacobi matrix after k iterations is sharply concentrated around the kth Jacobi matrix of the measure μ.

Theorem 4.2.13 (Location of Jacobi coefficients)**.** Let $(A_n)_{n=1}^{\infty}$ be a sequence of $n \times n$ Hermitian matrices with uniformly bounded operator norm. Assume their empirical spectral distributions μ_n converge in distribution to a measure μ with nontrivial absolutely continuous part, and further assume $\mathrm{Kol}(\mu_n, \mu) = O(n^{-c})$ for some $c > 0$.

Then there is a constant $d > 0$ dependent on μ and c, such that for any sequence of integers $1 \leq k_n \leq d\sqrt{\log n}$ we have

$$\|J_{k_n}(u) - J_{k_n}(\mu)\| \longrightarrow_P 0,$$

where $J_{k_n}(u)$ denotes the Jacobi matrix output by the Lanczos algorithm applied to A_n under the input $u \sim \text{Unif}(\mathbb{S}^{n-1})$ after k_n iterations, and where $J_{k_n}(\mu)$ is the k_n-th Jacobi matrix of the measure μ.

Note that Theorem 4.2.13 may be of particular relevance in applications where an infinite dimensional operator is discretized with the goal of computing its density. In essence, Theorem 4.2.13 states that, in this situation, the first iterations of the Lanczos algorithm are an accurate approximation of the true Jacobi coefficients of the measure μ, and hence the procedure gives valuable information to recover the limiting measure.

From the above proposition, a standard application of the Weyl eigenvalue perturbation inequality yields the following proposition.

Proposition 4.2.14 (Location of the Ritz values)**.** Using the same notation as in Theorem 4.2.13, let $\vec{r}_{k_n}(u) = (r_1(u), \ldots, r_{k_n}(u))$, where $r_1(u) \geq \cdots \geq r_{k_n}(u)$ are the random Ritz values of the Lanczos algorithm after k_n iterations are performed. Then under the assumptions in Theorem 4.2.13, we have that

$$\|\vec{r}_{k_n}(u) - \vec{r}_{k_n}(\mu)\|_{L^\infty(\mathbb{R}^{k_n})} \longrightarrow_P 0,$$

where $\vec{r}_{k_n}(\mu)$ is the vector whose entries are the roots of the k_n-th orthogonal polynomial with respect to μ in decreasing order.

It remains an open question if similar results can be obtained when $O(\log n)$ iterations are performed. See Section 4.6 for open questions and further research.

4.3 Applying the local Lévy lemma

4.3.1 Strategy

The well known Lévy lemma states, in a quantitative way, that if $f : \mathbb{S}^{n-1} \to \mathbb{R}$ is a Lipschitz function, then $f(u)$ is a random variable concentrated around its median. See Chapter 5.1 in [128] for a detailed discussion. In this direction, the main obstacle for showing concentration of the random variables $\alpha_i(u)$ and $\beta_i(u)$ is that the functions $\alpha_i, \beta_i : \mathbb{S}^{n-1} \to \mathbb{R}$ are not Lipschitz on the entire sphere. However, we will be able to show that these functions are Lipschitz in a large region of the sphere, which is a common idea in geometric functional analysis. We will use a local version of Lévy's lemma, which is recorded as Corollary 5.35 in [6], and which we restate below with explicit universal constants.

Lemma 4.3.1 (Local Lévy lemma)**.** Let $\Omega \subset \mathbb{S}^{n-1}$ be a subset of measure larger than 3/4. Let $f : \mathbb{S}^{n-1} \to \mathbb{R}$ be a function such that the restriction of f to Ω is Lipschitz with constant L. Then, for every $\varepsilon > 0$,

$$\mathbb{P}[|f(u) - \tilde{f}| > \varepsilon] \leq \mathbb{P}[u \in \mathbb{S}^{n-1} \setminus \Omega] + 2\exp\{-4n\varepsilon^2/L^2\},$$

where $\tilde{f}$ is the median of $f(u)$ and where $u \sim \mathbb{S}^{n-1}$.

One may also consider nonuniform random $u \in \mathbb{S}^{n-1}$, provided that there is a Lipschitz map $g : \mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$ such that u is distributed as the pushforward of the uniform measure under g.

In order to identify the correct region of the sphere in which the functions α_i and β_i are Lipschitz, we need a local version of the notion of Lipschitz constant. In what might be a slight departure from standard definitions, we will define *local Lipschitz continuity* as follows.

Definition 4.3.2. Let (X_1, d_1) and (X_2, d_2) be metric spaces. A function $f : X_1 \to X_2$ is said to be locally Lipschitz continuous with constant c at $x_0 \in X_1$, if for every $c' > c$ there is a neighborhood $U \subset X_1$ of x_0 such that

$$d_2(f(x), f(y)) \leq c' d_1(x, y) \quad \forall x, y \in U.$$

It is obvious that if a function is locally Lipschitz with constant c on every point of a convex set, then the function is globally Lipschitz on the set with the same constant c. However, if the convexity assumption is dropped, a similar conclusion is not guaranteed in general and in order to obtain a global Lipschitz constant the geometry of the set should be analyzed.

Definition 4.3.3. Let $K > 0$ and (X, d) be a metric space. We say that $S_1 \subset X$ is K-connected in S_2, with $S_1 \subset S_2 \subset X$, if for every $x, y \in S_1$ there is a rectifiable Jordan arc $\alpha : [0, 1] \to S_2$ with $\alpha(0) = x$ and $\alpha(1) = y$, such that the length of the trace of α is less than or equal to $Kd(x, y)$.

Now that we have introduced the notion of K-connected set we can generalize what we observed for convex sets.

Lemma 4.3.4. Let (X_1, d_1) and (X_2, d_2) be metric spaces. Assume that $S_1 \subset X_1$ is K-connected in $S_2 \subset X_1$ and let $f : X_1 \to X_2$ satisfy that for every $x_0 \in S_2$, f is locally Lipschitz at x_0 with constant c. Then f is globally Lipschitz on S_1 with constant cK.

Proof. Fix $x, y \in S_1$ and $\varepsilon > 0$. We will show that $d_2(f(x), f(y)) \leq (c+\varepsilon)Kd_1(x, y)$. Consider a rectifiable Jordan arc $\alpha : [0, 1] \to X_1$, such that $\alpha(0) = x$, $\alpha(1) = y$, $\alpha([0, 1]) \subset S_2$ and the length of α is at most $Kd_1(x, y)$.

Since the trace of α is contained in S_2, for every $w \in \alpha([0, 1])$ we can take an open ball U_w containing w such that f is $(c + \varepsilon)$-Lipschitz on U_w. Moreover, observe that since α is

continuous and injective, for every $w \in \alpha([0,1])$ we can take U_w small enough such that $\alpha^{-1}(U_w)$ is connected and hence an open interval in $[0,1]$.

By compactness of $\alpha([0,1])$ we may take $w_1, \dots, w_n \in \alpha([0,1])$ such that $\{U_{w_i}\}_{i=1}^n$ is a minimal cover for $\alpha([0,1])$. Now, since each $\alpha^{-1}(U_{w_i})$ is connected, and the cover is minimal, we have that $\alpha^{-1}(U_{w_i}) \cap \alpha^{-1}(U_{w_{i+1}}) \neq \emptyset$ for every $1 \dots, n-1$.

Furthermore, we will now see that we can modify the sequence of w_i such that $w_{i+1} \in U_{w_i}$ for every $i = 1, \dots, n-1$. Assume that this does not hold and let i be the smallest index for which $w_{i+1} \notin U_{w_i}$. Now take some $t \in \alpha^{-1}(U_{w_i}) \cap \alpha^{-1}(U_{w_{i+1}})$ and define $w' = \alpha(t)$. We construct a new sequence $\tilde{w}_1, \dots, \tilde{w}_{n+1} \in \alpha([0,1])$ by taking $\tilde{w}_j = w_j$ for $j < i$, $\tilde{w}_i = w'$, $\tilde{w}_{j+1} = w_j$ for $j \geq i$, and $U_{\tilde{w}_i}$ to be equal to $U_{w_{i+1}}$. Observe that for the new sequence of points $(\tilde{w}_i)_{i=1}^{n+1}$ in $\alpha([0,1])$ and sequence of open balls $U_{\tilde{w}_i}$ it holds that $\tilde{w}_{j+1} \in U_{\tilde{w}_j}$ for all $j \leq i$. By iterating this process we will obtain a finite sequence with the desired property. So, in what follows we can assume without loss of generality that $w_{i+1} \in U_{w_i}$ for every $i = 1, \dots, n-1$. We then will have

$$d_2(f(w_i), f(w_{i+1})) \leq (c + \varepsilon) d_1(w_i, w_i + 1).$$

Using the triangle inequality and the fact that $\sum_i d_1(w_i, w_{i+1})$ is bounded by the length of the trace of α the result follows. □

In the following section the local Lipschitz constants of the functions $\alpha_i(u)$ and $\beta_i(u)$ are shown to be related to the orthogonal polynomials of the measure μ^u.

4.3.2 Local Lipschitz constants for Jacobi coefficients

As it can be seen from Algorithm 1, the dependence of the quantities $\alpha_i(u)$, $\beta_i(u)$ and $v_j(u)$ on u is highly non-linear, which makes it complicated to show that such quantities are stable under perturbations of the input vector u. Here we exploit the fact that during every iteration of the Lanczos algorithm only locally Lipschitz operations are performed. The analysis of the compound effect of iterating the procedure yields a bound on the local Lipschitz constant of the quantities of interests. This bound is exponential in the number of iterations, which is enough to obtain concentration results when $O(\log(n))$ iterations are performed. In what follows, recall that $\gamma_i(u)$ denotes the leading coefficient of the ith orthonormal polynomial with respect to the measure μ^u defined in (1.11).

Proposition 4.3.5. Fix $\tilde{u} \in \mathbb{S}^{n-1}$ and let $v_j(u)$ be as in Algorithm 1. Then, for any $0 \leq j \leq n-1$, the functions $v_j(u)$ are locally Lipschitz at $\tilde{u}$ with constant $(4\|A\|)^j \gamma_j(\tilde{u})$.

Proof. We proceed by induction. For $j = 0$, recall $v_0(u) = u$ and $\gamma_0(\tilde{u}) = 1$; the statement follows. Now assume the proposition is true for some $j \geq 0$. For every $x \in \mathbb{S}^{n-1}$ denote $W_x = \operatorname{span}\{v_0(x) = x, v_1(x), \dots, v_j(x)\}$ and for any subspace $W \leq \mathbb{R}^n$ by Proj_W we mean the orthogonal projection onto W.

Take $x, y \in \mathbb{S}^{n-1}$ in a neighborhood $\mathcal{U}$ of $\tilde{u}$ to be determined and note that

$$\begin{aligned}
&\|\mathrm{Proj}_{W_x^\perp}(Av_j(x)) - \mathrm{Proj}_{W_y^\perp}(Av_j(y))\| \\
&\leq \|\mathrm{Proj}_{W_x^\perp}(A(v_j(x) - v_j(y)))\| + \|(\mathrm{Proj}_{W_x^\perp} - \mathrm{Proj}_{Wy^\perp})(Av_j(y))\| \\
&= \|\mathrm{Proj}_{W_x^\perp}(A(v_j(x) - v_j(y)))\| + \|(\mathrm{Proj}_{W_x} - \mathrm{Proj}_{W_y})(Av_j(y))\|. \qquad (4.2)
\end{aligned}$$

From the induction hypothesis we have that, for any $\varepsilon > 0$, we can choose $\mathcal{U}$ small enough so that

$$\|\mathrm{Proj}_{W_x^\perp}(A(v_j(x) - v_j(y)))\| \leq \|A\|\|v_j(x) - v_j(y)\| \leq \|A\|((4\|A\|)^j\gamma_j(\tilde{u}) + \varepsilon)\|x - y\|. \quad (4.3)$$

On the other hand, from Algorithm 1 it follows that $\beta_i(\tilde{u}) \leq \|A\|$ for every $i = 0, \ldots, n-1$, so in view of (1.7), the $\|A\|^i\gamma_i(\tilde{u})$ form an increasing sequence. It then follows that

$$\sum_{i=0}^{j}(4\|A\|)^i\gamma_i(\tilde{u}) \leq \sum_{i=0}^{j} 4^i\|A\|^j\gamma_j(\tilde{u}) \leq \frac{4^{j+1}\|A\|^j\gamma_j(\tilde{u})}{3}.$$

For any unit vector w, by the triangle inequality, we have that

$$\|\mathrm{Proj}_{W_x}(w) - \mathrm{Proj}_{W_y}(w)\| \leq \sum_{i=0}^{j}\|\langle v_i(x), w\rangle v_i(x) - \langle v_i(y), w\rangle v_i(y)\| \qquad (4.4)$$

and we can bound each term on the right-hand side of (4.4) as follows:

$$\begin{aligned}
\|\langle v_i(x), w\rangle v_i(x) - \langle v_i(y), w\rangle v_i(y)\| &\leq |\langle v_i(x) - v_i(y), w\rangle| + \|v_i(x) - v_i(y)\||\langle v_i(y), w\rangle| \\
&\leq \|v_i(x) - v_i(y)\|\|w\| + \|v_i(x) - v_i(y)\|\|v_i(y)\|\|w\| \\
&\leq 2(4\|A\|)^i\gamma_i(\tilde{u})\|x - y\|.
\end{aligned}$$

Hence, adding over i we obtain

$$\|\mathrm{Proj}_{W_x}(w) - \mathrm{Proj}_{W_y}(w)\| \leq \frac{2}{3} \cdot 4^{j+1}\|A\|^j\gamma_j(\tilde{u})\|x - y\|$$

which implies that $\|\mathrm{Proj}_{W_x} - \mathrm{Proj}_{W_y}\| \leq \frac{2}{3} \cdot 4^{j+1}\|A\|^j\gamma_j(\tilde{u})\|x - y\|$ and hence

$$\|(\mathrm{Proj}_{W_x} - \mathrm{Proj}_{W_y})(Av_j(y))\| \leq \frac{2}{3} \cdot (4\|A\|)^{j+1}\gamma_j(\tilde{u})\|x - y\| \qquad (4.5)$$

Putting together inequalities (4.2), (4.3) and (4.5), we get for any $x, y \in \mathcal{U}$ that

$$\|\mathrm{Proj}_{W_x^\perp}(Av_j(x)) - \mathrm{Proj}_{W_y^\perp}(Av_j(y))\| \leq (4\|A\|)^{j+1}\gamma_j(\tilde{u})\|x - y\|.$$

With this we have established that the function $u \mapsto \mathrm{Proj}_{W_u^\perp}(Av_j(u))$ is locally Lipschitz at $\tilde{u}$ with constant $(4\|A\|)^{j+1}\gamma_j(\tilde{u})$. Now consider the function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$f(x) = x/\|x\|$. It is easy to show that for any $x_0 \neq 0$, f is locally Lipschitz at x_0 with constant $1/\|x_0\|$. Now recall that by definition $\beta_j(\tilde{u}) = \|\mathrm{Proj}_{W_{\tilde{u}}^{\perp}}(Av_j(\tilde{u}))\|$. Since the composition of locally Lipschitz functions is locally Lipschitz with the constant being the product of the constants of each of the functions in the composition, we have that the function

$$u \mapsto v_{j+1}(u) = f(\mathrm{Proj}_{W_u^{\perp}}(Av_j(u)))$$

is locally Lipschitz at $\tilde{u}$ with constant $\frac{(4\|A\|)^{j+1}\gamma_j(\tilde{u})}{\beta_j(\tilde{u})} = (4\|A\|)^{j+1}\gamma_{j+1}(\tilde{u})$, where this equality follows from equation (1.7). □

Proposition 4.3.6. For any $0 \leq j \leq n-1$ and any $\tilde{u} \in \mathbb{S}^{n-1}$, the function $\alpha_j(u)$ is locally Lipschitz at $\tilde{u}$ with constant $\frac{1}{2} \cdot (4\|A\|)^{j+1}\gamma_j(\tilde{u})$, while $\beta_j(u)$ is locally Lipschitz at $\tilde{u}$ with constant $(4\|A\|)^{j+1}\gamma_j(\tilde{u})$.

Proof. We will use the same notation as in Proposition 4.3.5. Recall from Algorithm 1 that $\alpha_j(u) = \langle Av_j(u), v_j(u)\rangle$. Note that the local Lipschitz constant of the function $u \mapsto Av_j(u)$ is obtained by multiplying the local Lipschitz constant of $v_j(u)$ by $\|A\|$. Then, for any ε we can pick $\mathcal{U}$ to be a small enough neighborhood of $\tilde{u}$ such that for any $x, y \in \mathcal{U}$ we have

$$\begin{aligned}
|\alpha_j(x) - \alpha_j(y)| &= |\langle Av_j(x), v_j(x)\rangle - \langle Av_j(y), v_j(y)\rangle| \\
&\leq |\langle A(v_j(x) - v_j(y)), v_j(x)\rangle| + |\langle Av_j(y), v_j(x) - v_j(y)\rangle| \\
&\leq 2 \cdot (4^j\|A\|^{j+1}\gamma_i(\tilde{u}) + \varepsilon)\|x - y\|.
\end{aligned}$$

On the other hand, since $\beta_j(u) = \|\mathrm{Proj}_{W_u^{\perp}}(Av_j(u)))\|$ and we established in the proof of Proposition 4.3.5 that this function is locally Lipschitz with constant $(4\|A\|)^{j+1}\gamma_j(\tilde{u})$, the proof is concluded. □

Remark 4.3.7. The local Lipschitz constants presented in the above statements can be improved; the term 4^j next to $\|A\|^j\gamma_j(\tilde{u})$ was chosen for the sake of exposition. Nevertheless, it seems complicated to show that the quantities $v_j(u)$ are locally Lipschitz at $\tilde{u}$ with a constant of the form $C_j\|A\|^j\gamma_j$ and C_j subexponential. In any case, the term $\|A\|^j\gamma_j$ is typically exponential in j, so an improvement on C_j would not yield an asymptotic improvement to the final result if the same level of generality is considered. However, as we point out in Section 4.6, sharpening our constants is of relevance for applications.

4.3.3 Incompressibility

In Section 4.4, we will see that our upper bounds for the local Lipschitz constants of the Jacobi coefficients go to infinity if u becomes too close to a sparse vector, roughly speaking. So we only have a good local Lipschitz constant in a certain region of the unit sphere that avoids sparse vectors. In order to upgrade our local Lipschitz constant to a global Lipschitz constant, we must prove

1. that this region is large enough to apply the local Lévy lemma (Lemma 4.3.1), and
2. that this region is K-connected for a small enough K.

First we give this region a name. Loosely inspired by the compressed sensing literature (see for example [129]), we say that a vector u in $\mathbb{S}^{n-1}$ is (δ, ε)-*incompressible* if each set of at least δn coordinates carries at least ε of its "L^2 mass." Otherwise, we say that u is (δ, ε)-*compressible*. We denote the set of (δ, ε)-incompressible vectors in $\mathbb{S}^{n-1}$ by $I_n(\delta, \varepsilon)$ and record the formal definition below:

Definition 4.3.8.

$$I_n(\delta, \varepsilon) = \left\{ u \in \mathbb{S}^{n-1} : \sum_{i \in S} u_i^2 > \varepsilon \text{ for all } S \subseteq \{1, 2, \ldots, n\}, |S| \geq \delta n \right\}$$

For incompressible u we prove an adequate bound on the local Lipschitz constant in Proposition 4.4.1. Fortunately, a uniform random unit vector u is incompressible with high probability, as we will now show.

Proposition 4.3.9. Let $u \in \mathbb{S}^{n-1}$ be a uniform random unit vector, and let $0 < \varepsilon < \delta$. Then

$$\mathbb{P}[u \notin I_n(\delta, \varepsilon)] \leq \exp\left\{ 2\delta(1 + \log 1/\delta)n - \left(\frac{\varepsilon}{\delta} - 1\right)^2 n \right\} + \exp\{-\varepsilon^2 n/8\}$$

Corollary 4.3.10. Let $u \in \mathbb{S}^{n-1}$ be a uniform random unit vector, and let $0 < \delta \leq 1/50$. Then

$$\mathbb{P}[u \notin I_n(\delta, \delta/2)] \leq 2\exp\{-\delta^2 n/32\}.$$

Proof. Set $\varepsilon = \delta/2$ in Proposition 4.3.9. Note that $\varepsilon^2/8 = \delta^2/32$ and $2\delta(1+\log 1/\delta) - (1/2)^2 < -1/32$ for $0 < \delta \leq 1/50$. □

The proof of Proposition 4.3.9 consists of two parts. First, we prove a similar proposition where instead of the u_i we have independent Gaussian random variables with the same variance $1/n$. We then use a coupling argument to conclude the desired bound for u drawn uniformly from the unit sphere.

We will need upper and lower tail bounds on the χ^2 distribution. One can get good enough bounds using the Chernoff method, but rather than develop these from scratch we will cite the following corollary of Lemma 1 from Section 4.1 of [84].

Lemma 4.3.11. Let Y be distributed as $\chi^2(k)$ for a positive integer k. Then the following upper and lower tail bounds hold for any $t \geq 0$:

$$\mathbb{P}\left[Y \leq k - 2\sqrt{kt}\right] \leq e^{-t}$$

$$\mathbb{P}\left[Y \geq k + 2\sqrt{kt} + 2t\right] \leq e^{-t}$$

Proof of Proposition 4.3.9. Let $X_1, \dots, X_n$ denote independent Gaussian random variables each with variance $1/n$, and let $X = (X_1, \dots, X_n)$. If we set $u = X/\|X\|$, then u is uniformly distributed on the unit sphere; see e.g. [96].

We seek to upper bound the probability of compressibility $\{u \notin I_n(\delta, \varepsilon)\}$, which is the event that $\sum_{i \in S} u_i^2 < \varepsilon$ for some subset S of coordinates with $|S| \geq \delta n$. This event is contained in the union of the following two events:

1. E, the event that $\sum_{i \in S} X_i^2 \leq 2\varepsilon$ for some $|S| \geq \delta n$, and
2. F, the event that $\sum_{i \in S} X_i^2 \geq \varepsilon + \sum_{i \in S} u_i^2$ for some $|S| \geq \delta n$.

Indeed, if neither of these events hold, then for all $|S| \geq \delta n$ we have

$$2\varepsilon < \sum_{i \in S} X_i^2 < \varepsilon + \sum_{i \in S} u_i^2,$$

so u is incompressible.

To upper bound the probability of E, we use the union bound over all sets of size $k = \lceil n\delta \rceil$:

$$\begin{aligned} \mathbb{P}[E] &\leq \binom{n}{k} \mathbb{P}\left[\sum_{i=1}^{k} X_i^2 \leq 2\varepsilon\right] \\ &\leq (en/k)^k \exp\left\{-\frac{(k - 2n\varepsilon)^2}{4k}\right\} \end{aligned}$$

where in the last step we apply the lower tail bound in Lemma 4.3.11 with t being the solution to $k - 2\sqrt{kt} = 2n\varepsilon$. To avoid bookkeeping of ceiling and floor functions we use the extremely crude inequality $n\delta \leq k \leq 2n\delta$ (valid as long as $\delta n \geq 1$), which will suffice for our purposes:

$$\mathbb{P}[E] \leq \exp\left\{2\delta(1 + \log \delta^{-1})n - \left(\frac{\varepsilon}{\delta} - 1\right)^2 n\right\}.$$

We now upper bound the probability of F:

$$\begin{aligned} \mathbb{P}[F] &= \mathbb{P}\left[\sum_{i \in S} \left(X_i^2 - \frac{X_i^2}{\|X\|^2}\right) \geq \varepsilon \text{ for some } |S| > \delta n\right] \\ &= \mathbb{P}\left[\left(1 - \frac{1}{\|X\|^2}\right) \sum_{i \in S} X_i^2 \geq \varepsilon \text{ for some } |S| > \delta n\right] \\ &\leq \mathbb{P}\left[\left(1 - \frac{1}{\|X\|^2}\right) \|X\|^2 \geq \varepsilon\right] \\ &= \mathbb{P}\left[\|X\|^2 \geq 1 + \varepsilon\right] \end{aligned}$$

Since $Y = n\|X\|^2$ is distributed as $\chi^2(n)$, we may apply the upper tail bound in Lemma 4.3.11 with $t = n\varepsilon^2/8$ to obtain

$$\mathbb{P}[F] \leq \exp\{-n\varepsilon^2/8\}.$$

To conclude, we have $\mathbb{P}[u \notin I_n(\delta, \varepsilon)] \leq \mathbb{P}[E] + \mathbb{P}[F]$, and substituting the bounds we just derived, we obtain the desired inequality. □

4.3.4 K-connectedness of the incompressible region

Having proven that the incompressible region $I_n(\delta, \varepsilon)$ where we have a good local Lipschitz constant is almost the entire sphere, we now turn to proving that the region is K-connected for a small enough K.

One could try to show that any two points in $I_n(\delta, \varepsilon)$ can be connected by a short path contained in $I_n(\delta, \varepsilon)$, but for our purposes it is okay to let the path venture out into the larger region $I_n(4\delta, \varepsilon/\sqrt{2})$. When upgrading to a global Lipschitz constant, we will have to use the slightly worse upper bound for the local Lipschitz constant in this larger region, but this will still be good enough.

Proposition 4.3.12. $I_n(\delta, \varepsilon)$ is $\sqrt{2/\varepsilon}$-connected in $I_n(4\delta, \varepsilon/\sqrt{2})$.

Proof. Let x and y be any two endpoints in $I_n(\delta, \varepsilon)$. The construction will proceed in two steps. First, we will construct a path from x to y in $\mathbb{R}^n$ consisting of $\lceil \delta^{-1} \rceil$ pairwise orthogonal line segments. Then we will project this path radially onto the unit sphere and show that the result indeed lies in $I_n(4\delta, \varepsilon/2)$ and has length at most $(2/\sqrt{\varepsilon})\|x - y\|$.

Roughly speaking, we will partition the coordinates of x into $1/\delta$ blocks of δn coordinates and move the entries of each block linearly from x to y in parallel, one block at a time.

Because basic quantities such as $1/\delta$ and δn may not be integers, we will be content to split up $\mathbb{R}^n$ as the direct sum $\bigoplus_{i=1}^m \mathbb{R}^{n_i}$ where $\delta n \leq n_i \leq 2\delta n$ for all i.[1] Note also that this implies $m \geq \frac{1}{\delta}$. Similarly, for any vector $z \in \mathbb{R}^n$, we will write $z = \bigoplus_{i=1}^m z^{(i)}$, where $z^{(i)} \in \mathbb{R}^{n_i}$.

Now we may formally define the path P_i to be the line segment

$$P_i(t) = x^{(1)} \oplus \cdots \oplus x^{(i-1)} \oplus \left(tx^{(i)} + (1-t)y^{(i)}\right) \oplus y^{(i+1)} \oplus \cdots \oplus y^{(m)},$$

and define P to be the concatenation of the segments $P_1, \ldots, P_m$. The length of P is

$$\sum_{i=1}^m \|x^{(i)} - y^{(i)}\| \leq \sqrt{m}\|x - y\| \leq \sqrt{1/\delta}\|x - y\|,$$

by the Cauchy-Schwarz inequality. Also, $\|P(t)\| \geq \sqrt{\varepsilon/2\delta}$, because

$$\|P_i(t)\|^2 \geq \sum_{j=1}^{i-1} \|x^{(j)}\|^2 + \sum_{j=i+1}^{m} \|y^{(j)}\|^2 \geq (m-1)\varepsilon \geq \frac{\varepsilon}{2\delta}$$

[1] This is possible as long as $n/2 \geq \delta n \geq 1$, which will be true in our regime.

where we use that x and y are (δ, ε)-incompressible.

Furthermore, note that P lies inside the closed ball of radius $\sqrt{2}$, because for any i and t,

$$\|P_i(t)\|^2 \leq \sum_{j=1}^{m} \max\{\|x^{(j)}\|, \|y^{(j)}\|\}^2 \leq \sum_{j=1}^{m} \left(\|x^{(j)}\|^2 + \|y^{(j)}\|^2\right) = 2.$$

The path P currently does not lie in the unit sphere, so we project it onto the unit sphere along radii to get our final path P'. We now show that P' indeed lies in $I_n(4\delta, \varepsilon/\sqrt{2})$.

At this stage, we will dispense with the direct sum decomposition and use ordinary coordinates $z = (z_1, \ldots, z_n)$.

Consider any set S of at least $4\delta n$ coordinates, and consider any point $P_i(t)$ in our path P (before projection). The ith block of coordinates is in motion, and all of the other coordinates are either frozen at their initial value (from x) or their final value (from y).

The ith block consists of at most $2\delta n$ coordinates. Besides these, there are at least $4\delta n - 2\delta n = 2\delta n$ remaining coordinates in our set S. At least δn of them are from x or at least δn of them are from y. By incompressibility of x and y, the sum of the squares of these δn coordinates is at least ε.

After projecting onto the unit sphere, the sum of the same coordinates is still at least $\varepsilon/\sqrt{2}$, because as we saw, the original path had norm at most $\sqrt{2}$ at every point.

Finally, when projecting onto the unit sphere, the length of the path increases by at most a factor of $1/\sqrt{\varepsilon/2\delta}$, because as we saw earlier, originally each segment lay outside the smaller sphere of radius $\sqrt{\varepsilon/2\delta}$. The verification is an exercise in plane geometry (using the fact that $\tan\theta > \theta$ for $0 < \theta < \pi/2$) and also follows from the arc length formula $d\mathbf{s} = \sqrt{r^2 + (dr/d\theta)^2}\, d\theta \geq r\, d\theta$.

Thus, finally, we have shown that the path P' is contained in $I_n(4\delta, \varepsilon/\sqrt{2})$ and has length at most

$$\sqrt{1/\delta}\|x - y\|(1/\sqrt{\varepsilon/2\delta}) = \sqrt{2/\varepsilon}\|x - y\|.$$

□

4.4 Concentration of the Ritz values and Jacobi coefficients

We now analyze the local Lipschitz constant for the entries α_i and β_i of the Jacobi matrix. To simplify notation, in what follows we assume that $\|A\| = 1$ by rescaling A. Recall that this will also rescale the Ritz values and Jacobi coefficients by a factor $1/\|A\|$.

By Corollary 4.3.6, the function $\alpha_i(u)$ has local Lipschitz constant $2 \cdot 4^i \gamma_i(u)$, and $\beta_i(u)$ has local Lipschitz constant $4^{i+1}\gamma_i(u)$. Thus we are naturally led to the question of finding upper bounds for $\gamma_k(u)$. Recall that $\gamma_k(u)$ is defined as the leading coefficient of the kth orthogonal polynomial with respect to the measure $\mu^u = \sum_{i=1}^{n} u_i^2 \delta_{\lambda_i}$, and that π_k^u is the *monic* orthogonal polynomial with respect to the same measure.

The equations (1.6) and (1.11) imply

$$\gamma_k(u) = \left(\sum_{i=1}^{n} u_i^2 \pi_k^u(\lambda_i)^2 \right)^{-\frac{1}{2}}.$$

We seek to upper bound $\gamma_k(u)$ in terms of u, so we need to lower bound the quantity

$$\sum_{i=1}^{n} u_i^2 \pi_k^u(\lambda_i)^2 = \sum_{i=1}^{n} u_i^2 \prod_{j=1}^{k} |\lambda_i - r_j(u)|^2,$$

where $r_1(u), \ldots, r_k(u)$ are the roots of $\pi_k^u(z)$, i.e. the Ritz values.

Now, if it happens to be the case that the n eigenvalues λ_i are all clustered very close to the k Ritz values r_j, then we won't get a good lower bound. However, if $k << n$ and if the λ_i are reasonably spread out, we expect to get a good lower bound for most i. To make this precise, we are led to the notion of equidistribution, which was stated in Section 4.2.3 and which we restate below:

Definition 2.1 (Equidistribution). *Let Λ be any finite set of n real numbers. Let δ and ω be positive real numbers and let j be a natural number. We say that Λ is* (δ, ω, j)-equidistributed *if for any finite set T of at most j real numbers,*

$$\left| \left\{ \lambda \in \Lambda : \frac{1}{|T|} \sum_{t \in T} \log |\lambda - t| \geq \log \omega \right\} \right| \geq \delta n.$$

We will show in Section 4.4.1 that a wide range of spectra are equidistributed.

Now we apply the definition. Returning to our effort to upper bound $\gamma_j(u)$, we see that if we assume the spectrum of A is (δ, ω, k)-equidistributed, then

$$\sum_{i=1}^{n} u_i^2 \prod_{j=1}^{k} |\lambda_i - r_j(u)|^2 \geq \sum_{i \in S} u_i^2 \omega^{2k},$$

where S is some subset of $\{1, \ldots, n\}$ of size at least δn. However, for an arbitrary unit vector u and an arbitrary subset S, we have no lower bound on the sum $\sum_{i \in S} u_i^2$—it could even be zero. This leads to our definition of incompressibility in Section 4.3, which is satisfied by u with high probability.

Indeed, if we assume that the unit vector u is (δ, ε)-incompressible, then the right hand side expression above is greater than $\varepsilon \omega^{2k}$. Putting together the last few equations, we have $\gamma_k(u) \leq (\varepsilon \omega^{2k})^{-1/2}$. We summarize the result in the following proposition:

Proposition 4.4.1. Suppose the spectrum of A is (δ, ω, k)-equidistributed and suppose that u is (δ, ε)-incompressible for some $\delta, \omega, \varepsilon > 0$ and $k \in \mathbb{N}$. Then

$$\gamma_k(u) \leq \frac{1}{\omega^k \sqrt{\varepsilon}}.$$

4.4.1 Equidistribution

In this section we establish sufficient conditions for equidistribution that apply to a wide range of spectra. First, we present an immediate generalization of the notion of equidistribution which applies to measures μ instead of finite sets Λ. The definitions coincide for finite sets if one identifies Λ with the uniform probability distribution on Λ.

Definition 4.4.2 (Equidistribution for measures)**.** Let μ be a probability measure on $\mathbb{R}$. Let $\delta, \omega > 0$ and j be a natural number. We say that μ is (δ, ω, j)-*equidistributed* if for any finite set T of at most j real numbers,

$$\mu\left(\left\{x \in \mathbb{R} : \frac{1}{|T|}\sum_{t\in T} \log|x-t| \geq \log\omega\right\}\right) \geq \delta.$$

If a measure is (δ, ω, j)-equidistributed for every $j \in \mathbb{N}$, we will just say that it is (δ, ω)-equidistributed.

For absolutely continuous measures, we have the following general equidistribution result:

Proposition 4.4.3 (Absolutely continuous measures are equidistributed)**.** Let ν be a compactly supported probability measure on $\mathbb{R}$ with a nontrivial absolutely continuous part. Then there exist constants $\delta, \omega > 0$ such that ν is (δ, ω)-equidistributed.

Proof. By the assumption, we may write $\nu = \nu_1 + \nu_2$ where ν_1 is absolutely continuous with respect to Lebesgue measure. By cutting off the portion where the density of ν_1 is greater than some large $M > 0$ and assigning that mass to ν_2 instead, we may assume without loss of generality that the density function of ν_1 is bounded.

We now utilize a Markov inequality type argument. Let T be any set of j real numbers. Define the logarithmic potential

$$V(x) = -\frac{1}{j}\sum_{t\in T} \log|x-t|.$$

Since ν_1 has a bounded density function, $\log|x-t|$ is integrable against ν_1 for all t, so the integral $\int_{-\infty}^{\infty} V_t(x)\, d\nu_1(x)$ is finite for each $t \in T$. Averaging over all $t \in T$, we find that

$$\frac{1}{\nu_1(\mathbb{R})}\int_{-\infty}^{\infty} V(x) d\nu_1(x) \leq a$$

for some constant $a < \infty$. Then

$$a \geq \frac{1}{\nu_1(\mathbb{R})}\int_{-\infty}^{\infty} V(x) d\nu_1(x) \geq \frac{2a\nu_1(\{x \in \mathbb{R} : V(x) \geq 2a\})}{\nu_1(\mathbb{R})}.$$

Relating this back to the definition of equidistribution, we have

$$\nu_1\left(\left\{x \in \mathbb{R} : \frac{1}{|T|}\sum_{t\in T} \log|x-t| \geq -2a\right\}\right) = \nu_1(\{x \in \mathbb{R} : V(x) \leq 2a\}) \geq \frac{1}{2}\nu_1(\mathbb{R}).$$

Hence we may take $\delta = \frac{1}{2}\nu_1(\mathbb{R})$ and $\omega = e^{-2a}$. □

Given our framework, it will be useful to have a statement relating the equidistribution of an absolutely continuous measure to a discretization of that measure. If the two measures are close in Kolmogorov distance, then we can prove such a statement.

Proposition 4.4.4. Let μ and ν be probability measures. If μ is (δ, ω, j)-equidistributed for some $\delta, \omega > 0$ and $j \in \mathbb{N}$, then ν is $(\delta - \varepsilon, \omega, j)$-equidistributed, where $\varepsilon = 4j\mathrm{Kol}(\mu, \nu)$.

Proof. Let T be any set of at most j real numbers. Since $p(x) = \prod_{t\in T} |x - t|$ is the absolute value of a polynomial of degree j, each of its level sets is a union of at most $2j$ intervals. Hence,

$$|\mu(\{x \in \mathbb{R} : p(x) \geq \omega^{|T|}\}) - \nu(\{x \in \mathbb{R} : p(x) \geq \omega^{|T|}\})| \leq 4j\mathrm{Kol}(\mu, \nu).$$

□

Thus, to prove equidistribution for an atomic measure, it suffices to prove equidistribution for a nearby absolutely continuous measure.

The above propositions immediately yield a useful corollary for analyzing the Lanczos procedure in the regime of $O(\log n)$ iterations:

Corollary 4.4.5. Let μ be a compactly supported probability measure with nontrivial absolutely continuous part. Let $\{\mu_n\}$ be a sequence of probability measures such that $\mathrm{Kol}(\mu_n, \mu) \leq \frac{C}{\log n}$ for some $C > 0$. Then for all n, for all $j \leq \frac{1}{2C} \log n$ we have that μ_n is (δ, ω, j)-equidistributed for some $\delta, \omega > 0$.

Remark 4.4.6. If μ is (δ, ω, j)-equidistributed and ν is the pushforward of μ under the affine map $x \mapsto ax + b$, then ν is $(\delta, a\omega, j)$-equidistributed.

We now compute the equidistribution for a few example measures, following the proof of Proposition 4.4.3.

Example 4.4.7. Let μ denote the uniform measure on $[0, 1]$. Then

$$\int V(x)\, d\mu(x) \leq \int -\log \left| x - \frac{1}{2} \right| \, d\mu(x) = 1 + \log 2.$$

Thus, μ is $(1/2, 4e^{-2})$-equidistributed.

Example 4.4.8. Let ν denote the *semicircle law* $d\nu = \frac{1}{2\pi}\sqrt{(4 - x^2)_+}\, dx$. Then

$$\int V(x)\, d\nu(x) \leq \int -\log |x| \, d\nu(x) = 1/2.$$

Thus, ν is $(1/2, e^{-1})$-equidistributed.

With the above the claims made in the examples of Section 4.2.3 are now trivial.

Proof of Example 4.2.4 and Example 4.2.11. It is enough to put together Proposition 4.4.4 and Example 4.4.7. □

Note that for a given set of points that does not resemble a discretization of an absolutely continuous distribution, it will still be likely that the equidistribution parameters are well behaved (relative to their scale) provided that the points are somewhat spread out. On the other hand, if the points are clustered in a few small clusters the analysis becomes trivial.

Observation 4.4.9. Let Λ be a set (or multiset) of n points. Let $a_1 \leq b_1 < a_2 \leq b_2 < \cdots < a_m \leq b_m$ be such that $\Lambda \subset \bigcup_{i=1}^{m}[a_i, b_i]$. Define $n_i = |\Lambda \cap [a_i, b_i]|$ and let g the minimal gap between clusters, namely $g = \min_{1\leq i\leq m-1} a_{i+1} - b_i$. Then Λ is $(\frac{k_j}{n}, \frac{g}{2}, j)$-distributed, where $k_j = \min_S \sum_{i\in S^c} n_i$ and S runs over all subsets of $\{1, \ldots, m\}$ of size j.

Proof. The proof follows directly from the definition of equidistribution. □

Remark 4.4.10. A particular case of Observation 4.4.9 is when $n_i \geq \lfloor \frac{n}{m} \rfloor$ and $g = a_{i+1} - b_i$ for every $i = 1, \ldots, m$, which yields Example 4.2.5 above. More generally, if each n_i is roughly n/m then k_j will be roughly $m - j$, and hence the δ parameter for the equidistribution of Λ will only degrade when $j \approx m$. In other words, Theorem 4.2.2 is still strong for matrices whose spectrum consists of small clusters if the number of such clusters exceeds the number of iterations of the Lanczos procedure. On the other hand, if the number of iterations exceeds the number of clusters it is not hard to show that the Lanczos procedure will output (with overwhelming probability) at least one Ritz value per cluster.

4.4.2 Proof of Theorem 4.2.2

We now have the necessary tools to prove concentration for the entries of the Jacobi matrix.

Proposition 4.4.11 (Jacobi coefficients are globally Lipschitz)**.** Suppose the spectrum of A_n is $(4\delta, \omega, i)$-equidistributed for some $\delta, \omega > 0$ and $i \in \mathbb{N}$. Then for any $0 < \varepsilon < \delta$, functions $\alpha_i(u)$ and $\beta_i(u)$ are globally Lipschitz on $I_n(\delta, \varepsilon)$ with constant $L_{i,\varepsilon} \leq \frac{4^{i+2}\|A\|^{i+1}}{\omega^i \varepsilon}$.

Proof. Proposition 4.3.6 says that $\alpha_i(u)$ and $\beta_i(u)$ both have local Lipschitz constant at most $4^{i+1}\|A\|^{i+1}\gamma_i(u)$ for all $u \in \mathbb{S}^{n-1}$. Proposition 4.4.1 says that because the spectrum of A_n is $(4\delta, \omega, i)$-equidistributed, $\gamma_i(u) \leq \frac{1}{\omega^i\sqrt{\varepsilon/\sqrt{2}}}$ for all $u \in I_n(4\delta, \varepsilon/\sqrt{2})$. Combining these, we have that $\alpha_i(u)$ and $\beta_i(u)$ are locally Lipschitz with constant

$$\frac{4^{i+1}\|A\|^{i+1}}{\omega^i\sqrt{\varepsilon/\sqrt{2}}}$$

for all $u \in I_n(4\delta, \varepsilon/\sqrt{2})$. Proposition 4.3.12 says that $I_n(\delta,\varepsilon)$ is $\sqrt{2/\varepsilon}$-connected in the larger set $I_n(4\delta, \varepsilon/\sqrt{2})$, so Lemma 4.3.4 implies that $\alpha_i(u)$ and $\beta_i(u)$ are *globally* Lipschitz on $I_n(\delta,\varepsilon)$ with constant

$$L_{i,\varepsilon} = \frac{\sqrt{2}}{\sqrt{\varepsilon}}\left(\frac{4^{i+1}\|A\|^{i+1}}{\omega^i\sqrt{\varepsilon/\sqrt{2}}}\right) \leq \frac{4^{i+2}\|A\|^{i+1}}{\omega^i\varepsilon}.$$

□

We now have the tools to prove our first main theorem, which quantifies the concentration of the Jacobi coefficients around their medians.

Proof of Theorem 4.2.2. The local Lévy lemma (Lemma 4.3.1) yields that $\mathbb{P}[|\alpha_i(u) - \tilde{\alpha}_i| > t\|A\|]$ and $\mathbb{P}[|\beta_i(u) - \tilde{\beta}_i| > t\|A\|]$ are both at most

$$\mathbb{P}[u \notin I_n(\delta,\varepsilon)] + 2\exp\{-4nt^2\|A\|^2/L_{i,\varepsilon}^2\},$$

where $L_{i,\varepsilon}$ is the global Lipschitz constant on $I_n(\delta,\varepsilon)$ obtained in Proposition 4.4.11. Note that if $\delta > 1/50$, then A is still $(1/50,\omega,i)$-equidistributed, so we may set $\varepsilon = \delta/7$ and apply Corollary 4.3.10 to bound $\mathbb{P}[u \notin I_n(\delta,\varepsilon)]$. We obtain the upper bound

$$2\exp\left\{-\frac{\min\{\delta,1/50\}^2}{32}n\right\} + 2\exp\left\{\frac{-4nt^2\|A\|^2\omega^{2i}(\delta/2)^2}{4^{2i+4}\|A\|^{2i+2}}\right\}$$

$$\leq 2\exp\left\{-\frac{\min\{\delta,1/50\}^2}{32}n\right\} + 2\exp\left\{-\frac{1}{64}\left(\frac{\omega}{4\|A\|}\right)^{2i}\delta^2t^2n\right\}$$

as desired. □

Now we show how Theorem 4.2.2 implies Propositions 4.2.8 and 4.2.9.

Proof of Proposition 4.2.8. Throughout this proof we will use the same notation as in the statement of Proposition 4.2.8. Since $\tilde{J}_k$ and $J_k(u)$ are tridiagonal matrices, we may split $J_k - \tilde{J}_k$ into the sum of three matrices consisting of the diagonal, the subdiagonal and the superdiagonal and then use the triangle inequality to obtain

$$\|J_k(u) - \tilde{J}_k\| \leq \max_{0\leq i\leq k-1}\{|\alpha_i(u) - \tilde{\alpha}_i|\} + 2\max_{0\leq i\leq k-2}\{|\beta_i(u) - \tilde{\beta}_i|\}. \tag{4.6}$$

Hence, we deduce that

$$\begin{aligned}\mathbb{P}[\|\vec{r}(u) - \vec{r}\|_\infty \geq t] &\leq \mathbb{P}[\|J_k(u) - \tilde{J}_k\| \geq t] \\ &\leq \mathbb{P}\left[\max_{0\leq i\leq k-1}\{|\alpha_i(u) - \tilde{\alpha}_i|\} + 2\max_{0\leq i\leq k-2}\{|\beta_i(u) - \tilde{\beta}_i|\} \geq t\right]\end{aligned}$$

where the first inequality follows from Lemma 4.2.6 and the second inequality from (4.6). Now observe that the event $\{\max_{0\le i\le k-1}\{|\alpha_i(u)-\tilde{\alpha}_i|\}+2\max_{0\le i\le k-2}\{|\beta_i(u)-\tilde{\beta}_i|\}\ge t\}$ is contained in the event

$$\left\{\max_{0\le i\le k-1}\{|\alpha_i(u)-\tilde{\alpha}_i|\}\ge\frac{t}{3}\right\}\cup\left\{\max_{0\le i\le k-2}\{|\beta_i(u)-\tilde{\beta}_i|\}\ge\frac{t}{3}\right\},$$

which in turn is contained in the event

$$\bigcup_{i=1}^{k}\left\{|\alpha_i(u)-\tilde{\alpha}_i|\ge\frac{t}{3}\right\}\bigcup\left\{|\beta_i(u)-\tilde{\beta}_i|\ge\frac{t}{3}\right\}.$$

Using a union bound and applying Theorem 4.2.2, we obtain the desired result. □

Proof of Proposition 4.2.9. From Theorem 4.2.7 we have that $\sin\theta\le\frac{2\|\tilde{J}_k(u)-\tilde{J}_k(u)\|}{\varepsilon}$ and hence

$$\begin{aligned}\mathbb{P}[\sin\theta\ge t]&\le\mathbb{P}[\|J_k(u)-\tilde{J}_k\|\ge t]\\&\le\mathbb{P}\left[\max_{0\le i\le k-1}\{|\alpha_i(u)-\tilde{\alpha}_i|\}+2\max_{0\le i\le k-2}\{|\beta_i(u)-\tilde{\beta}_i|\}\ge t\right],\end{aligned}$$

where the latter inequality was established in the proof of Proposition 4.2.8. Using the bound obtained in the aforementioned proof and substituting $t=\frac{2}{\varepsilon n^c}$ we obtain the desired result. □

Remark 4.4.12. Using the same techniques one can prove an analogous result to Theorem 4.2.2 in the case where A_n is not Hermitian, and even not normal. In the non-Hermitian case, the Lanczos algorithm is called the *Arnoldi algorithm* and is still used in practice to identify extreme (complex) eigenvalues. If A_n is non-Hermitian, the $k\times k$ matrix output by the Arnoldi algorithm is guaranteed to be *upper Hessenberg*—that is, zero above the superdiagonal—but not necessarily normal. Thus, its eigenvalues may be highly unstable, due to the phenomenon of *pseudospectrum*—see [125] for a discussion of this issue. Thus, even though we have concentration of the entries of the Hessenberg matrix, this does not imply concentration of the Ritz values. Achieving concentration for the Ritz values of a non-Hermitian matrix remains an open question.

Combining the previous theorem with Corollary 4.4.5 we get convergence in probability of the Jacobi matrices in the regime $k=O(\log n)$:

Proposition 4.4.13. Let the spectra μ_n of A_n converge to the spectrum μ of A in Kolmogorov distance with rate $O(1/\log n)$. Suppose μ has a nontrivial absolutely continuous part. Then there exists $c_2>0$ and a sequence $k_n\ge c_2\log n$ such that the Jacobi matrices J_{k_n} output by the Lanczos algorithm after k_n iterations converge to entrywise in probability to deterministic constants.

Proof. By Corollary 4.4.5, we have that μ_n is (δ, ω, k)-equidistributed for all $k \leq c_1 \log n$. Picking $c_2 < c_1$ and applying Theorem 4.2.2, for $i \leq c_2 \log n$ this yields the bound

$$\begin{aligned}\mathbb{P}[|\alpha_i - \tilde{\alpha_i}| > t] &\leq \exp\{-\delta^2 n/32\} + 2\exp\left\{-\frac{4}{4^3}(\omega/4)^{2c_2 \log n} n t^2\right\} \\ &= \exp\{-\delta^2 n/32\} + 2\exp\left\{-\frac{4}{4^3} n^{2c_2 \log(\omega/4)+1} t^2\right\}\end{aligned}$$

so as long as $2c_2 \log(\omega/4) + 1 > 0$, we have convergence in probability of the Jacobi coefficients as $n \to \infty$. But this is certainly true for small enough c_1. The β_i have the same bound as the α_i, so we are done. □

Convergence for *fixed* k to the infinite Jacobi matrix J of μ (with no hypothesis on the rate of convergence of μ_n) is proven in [59], §4. In Proposition 4.4.13 we leave it open to prove that the limit is actually J (see Question 2), but if we reduce the number of iterations from $k = O(\log n)$ to $k = O(\sqrt{\log n})$, we can indeed prove that the limit is J. This is the content of Theorem 4.2.13, proven in Section 4.5.

4.5 Proofs of Proposition 4.2.12 and Theorem 4.2.13

4.5.1 Proof of Proposition 4.2.12

We now prove our theorem about the Lanczos algorithm missing outliers in the spectrum.

Proof of Proposition 4.2.12. By Proposition 4.4.4, we have that μ_n is (δ, ω, j)-equidistributed for some $\delta, \omega > 0$ and all $j < c \log n$. Suppose $u \in I_n(\delta, \varepsilon)$, which happens with overwhelming probability by Proposition 4.3.9. Then by Proposition 4.4.1, we have an upper bound on the leading coefficient of the jth orthogonal polynomial: $\gamma_j(u) \leq \frac{1}{\omega^j \sqrt{\varepsilon}}$. Equivalently, this is a lower bound on the L^2 norm of the jth *monic* orthogonal polynomial: $\|\pi_j^u\|_{L^2(\mu^u)} \geq \omega^j \sqrt{\varepsilon}$. As mentioned in the preliminaries in Section 1.1, it is a classical fact that the monic orthogonal polynomial of any given degree has minimal L^2 norm over all monic polynomials of that degree. Thus, we in fact have

$$\int q(x)^2 \, d\mu^u(x) \geq \varepsilon \omega^{2j} \tag{4.7}$$

for all monic polynomials q of degree j, with equality when $q(x)$ is the kth orthogonal polynomial $p_k^u(x)$.

For all unit vectors u, let $\rho(u)$ denote the top Ritz value, i.e. the maximum root of $p_k^u(x)$. We wish to show that $\rho(u) < R + \kappa$ with high probability.

Take $p_k^u(x)$ and replace its top root by t to form the monic polynomial P_t. By the first-order condition for the variational characterization of p_k^u mentioned above, to show

$\rho(u) \le R + \kappa$ it suffices to show that $\|P_t\|_{L^2(\mu^u)}$ is strictly increasing in t for $t > R + \kappa$. We have

$$\|P_t\|^2_{L^2(\mu^u)} = \int \left(\frac{\pi^u_k(x)}{x - \rho(u)}(x - t) \right)^2 d\mu^u(x) = \sum_{i=1}^{k} u_i^2 (\lambda_i - t)^2 \prod_{j=2}^{k} (\lambda_i - r_j)^2$$

where we let $r_2, \dots, r_k$ denote the roots of $p^u_k(x)$ besides the maximum root $\rho(u)$, and we omit the argument u for brevity. We calculate the derivative

$$\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} = -2\sum_{i=1}^{m} u_i^2(\lambda_i - t)\prod_{j=1}^{k-1}(\lambda_i - r_j)^2 - 2\sum_{i=m+1}^{n} u_i^2(\lambda_i - t)\prod_{j=2}^{k}(\lambda_i - r_j)^2.$$

We wish to show that this quantity is positive whenever $t \ge R + \kappa$. We have assumed that there are only m outliers, so assume $\lambda_i \le R$ for all $i > m$. Then $t - \lambda_i \ge \kappa$ for every $m < i \le n$.

Thus,

$$\begin{aligned}
\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} &\ge -2\sum_{i=1}^{m} u_i^2(\lambda_i - t)\prod_{j=1}^{k-1}(\lambda_i - r_j)^2 + 2\sum_{i=m+1}^{n} u_i^2\kappa\prod_{j=2}^{k}(\lambda_i - r_j)^2 \\
&= -2\sum_{i=1}^{m} u_i^2(\lambda_i - t)\prod_{j=2}^{k}(\lambda_i - r_j)^2 \\
&\quad + \left[2\kappa \int \left(\frac{p^u_k(x)}{x - \rho(u)} \right)^2 d\mu^u(x) - 2\sum_{i=1}^{m} u_i^2\kappa\prod_{j=2}^{k}(\lambda_i - r_j)^2 \right] \\
&\ge -2\sum_{i=1}^{m} u_i^2(\lambda_i - t)\prod_{j=2}^{k}(\lambda_i - r_j)^2 + 2\kappa\varepsilon\omega^{2(k-1)} - 2\sum_{i=1}^{m} u_i^2\kappa\prod_{j=2}^{k}(\lambda_i - r_j)^2
\end{aligned}$$

where in the last step we used the inequality (4.7) on the degree $k-1$ polynomial $p^u_k(x)/(x - \rho(u))$. Simplifying, we have

$$\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} \ge 2\kappa\varepsilon\omega^{2(k-1)} - 2\sum_{i=1}^{m} u_i^2(\lambda_i + \kappa - t)\prod_{j=2}^{k}(\lambda_i - r_j)^2.$$

By uniform boundedness of the spectra, there exists M large such that $\lambda_i - r_j \le M$ for all $1 \le i \le m$. Let g be the maximum of the outlier gaps $\lambda_i - R$ over all $1 \le i \le m$. Recall that $t \ge R + \kappa$, so $\lambda_i + \kappa - t \le \lambda_i - R \le g$ for all $1 \le i \le m$. Finally, we have with overwhelming probability $\sum_{i=1}^{m} u_i^2 < n^{-c}$ for any positive $c < 1/2$; we will defer the proof to Lemma 4.5.3 below. Putting this all together, we have

$$\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} \ge 2\kappa\varepsilon\omega^{2k-2} - 2n^{-c}M^{2k-2}mg.$$

This quantity is strictly positive when

$$\log \kappa\varepsilon + (2k-2)\log\omega > -c\log n + (2k-2)\log M + \log mg$$

Rearranging, we get

$$(2k-2)\log(\omega/M) > -c\log n + \log mg - \log \kappa\varepsilon$$

for n large. Note that $\omega < M$, because ω is a lower bound on geometric means of distances that are all less than M. In conclusion, with high probability, $\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} > 0$ for all $t > R + \kappa$ when

$$2k - 2 < \frac{1}{\log \frac{M}{\omega}} \left(c\log n + \log \frac{\kappa\varepsilon}{mg} \right). \tag{4.8}$$

For n large, we may absorb the constants $m, g, \kappa, \varepsilon, \omega$ (which do not depend on n) into a single constant $c' > 0$, and we get the desired $k \leq c' \log n$.

□

Remark 4.5.1. We have focused on the right hand side of the spectrum for ease of exposition. Similar results hold for outliers on both sides.

Remark 4.5.2. There are several parameters that can be tuned in the above proof. For example, one could envision a situation in which κ converges to zero as $n \to \infty$, at the expense of some other parameter.

Lemma 4.5.3. Let $0 < c < 1/2$ and suppose $m \leq n^\alpha$, where $\alpha < 1 - c$. Then $\sum_{i=1}^m u_i^2 < n^{-c}$ with overwhelming probability. To be precise,

$$\mathbb{P}\left[\sum_{i=1}^m u_i^2 \geq n^{-c} \right] \leq \exp\left\{ -\frac{1}{16}\left(4n^\alpha - 4\sqrt{2}n^{\frac{1}{2}-\frac{c}{2}+\frac{\alpha}{2}} + 2n^{1-c} \right) \right\} + \exp\left\{ -\frac{1}{16} n^{1-2c} \right\}.$$

Proof. We proceed just as in the proof of Proposition 4.3.9. Define X_i as in that proof. Then

$$\mathbb{P}\left[\sum_{i=1}^m u_i^2 > n^{-c} \right] \leq \mathbb{P}\left[\sum_{i=1}^m X_i^2 > \frac{1}{2} n^{-c} \right] + \mathbb{P}\left[\sum_{i=1}^m X_i^2 < -\frac{1}{2} n^{-c} + \sum_{i=1}^m u_i^2 \right].$$

Using Lemma 4.3.11, we solve for the parameter $\sqrt{t} = \frac{-2\sqrt{m} + \sqrt{2} n^{\frac{1}{2}-\frac{c}{2}}}{4}$ (which requires $\alpha < 1 - c$) and then we get

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^m X_i^2 > \frac{1}{2} n^{-c} \right] &\leq \exp\left\{ -\left(\frac{-2\sqrt{m} + \sqrt{2} n^{\frac{1}{2}-\frac{c}{2}}}{4} \right)^2 \right\} \\ &= \exp\left\{ -\frac{1}{16}\left(4n^\alpha - 4\sqrt{2} n^{\frac{1}{2}-\frac{c}{2}+\frac{\alpha}{2}} + 2n^{1-c} \right) \right\}, \end{aligned}$$

which is an overwhelmingly small probability because $\frac{1}{2} - \frac{c}{2} + \frac{\alpha}{2} < 1 - c$ when $\alpha < 1 - c$.

Now following the same coupling argument in the proof of Proposition 4.3.9 and using Lemma 4.3.11 again, we get

$$\mathbb{P}\left[\sum_{i=1}^{m} X_i^2 < -\frac{1}{2}n^{-c} + \sum_{i=1}^{m} u_i^2\right] \leq \exp\left\{-\frac{1}{16}n^{1-2c}\right\}.$$

□

Proof of Theorem 4.2.10. From the proof of Proposition 4.2.12, setting $\varepsilon = \delta/2$ we have that the Ritz values are contained in the desired interval for

$$k \leq \frac{1}{2\log\frac{M}{\omega}}\left(c\log n + \log\frac{\kappa\delta}{2mg}\right)$$

as long as $k \leq j$, $u \in I_n(\delta, \delta/2)$ and $\sum_{i=1}^{m} u_i^2 > n^{-c}$. Applying Corollary 4.3.10, the probability that u violates either condition is at most

$$\begin{aligned}
&\mathbf{P}[u \notin I_n(\delta, \delta/2)] + \mathbb{P}\left[\sum_{i=1}^{m} u_i^2 > n^{-c}\right] \\
&\leq 2\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + \mathbb{P}\left[\sum_{i=1}^{m} u_i^2 > n^{-c}\right] \\
&\leq 2\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + 2\exp\left\{-\frac{1}{16}n^{1-2c}\right\}
\end{aligned}$$

where in the last step, we apply Lemma 4.5.3 and note that for $n \geq e^{\frac{1}{1-c-\alpha}}$ we have $4\sqrt{2}n^{\frac{1-c+\alpha}{2}} \leq n^{1-c}$.

□

4.5.2 Proof of Theorem 4.2.13

For $C > 0$ let $\mathcal{P}_C$ denote the space of Borel probability measures supported on $[-C, C]$. In order to prove Theorem 4.2.13 we will show that the Jacobi coefficients of a measure are locally Lipschitz quantities on the space $\mathcal{P}_C$ equipped with the Kolmogorov metric. Note that in Section 4.3 similar results were obtained in the case in which the space of measures in consideration is restricted to atomic measures supported on n fixed points, namely the eigenvalues of A_n. Since $\mathcal{P}_C$ is a much larger and complicated space we are not able to obtain results as strong as in Proposition 4.3.6. It remains an open question if a better rate can be achieved at this level of generality. Specifically, two natural questions can be asked. Question 1 posed in Section 4.6 may be of independent interest in the area of orthogonal polynomials, while Question 2 is problem-specific.

We will use the following well known result which, for convenience of the reader, we restate as it appears in Lemma 1.1 in [62].

Lemma 4.5.4. Let A and B be two $k \times k$ matrices. Then $\det(A+B)$ is equal to the sum of the determinants of the 2^k matrices obtained by replacing each subset of the columns of A by the corresponding subset of the columns of B.

Proof. The result follows directly from the fact that the determinant is multilinear in the columns of the matrix. □

Lemma 4.5.5. Let A and B be two $k \times k$ matrices. For $1 \leq i \leq k$, let $A^{(i)}$ and $B^{(i)}$ be the ith columns of A and B respectively. Let $C, \varepsilon > 0$ and assume that

$$\|A^{(i)} - B^{(i)}\|_2 \leq \varepsilon \quad \text{and} \quad \max\{\|A^{(i)}\|_2, \|B^{(i)}\|_2\} \leq C. \tag{4.9}$$

Then

$$|\det(A) - \det(B)| \leq \varepsilon k (C+\varepsilon)^{k-1}.$$

Proof. By the assumption in (4.9) we can write $B = A + E$, where E is a matrix with columns of norm less or equal to ε. Then, using Lemma 4.5.4, the inequalities in (4.9) and the fact that the determinant of a matrix is bounded by the product of the Euclidean norms of its columns, we obtain

$$|\det(A+E) - \det(A)| \leq \sum_{k=1}^{n} \binom{n}{k} C^{n-k}\varepsilon^k = (C+\varepsilon)^k - C^k \leq \varepsilon k (C+\varepsilon)^{k-1}$$

where the last inequality follows from the mean value theorem. □

We now argue that the moments of a measure are Lipschitz quantities in $\mathcal{P}_C$, where the constant is exponential in the order of the moment. With this end fix a Borel measure μ on $\mathbb{R}$ and denote

$$m_k(\mu) = \int_{\mathbb{R}} x^k d\mu(x).$$

A standard application of Fubini's theorem yields that if μ is a finite positive Borel measure supported in $[0, \infty)$ then

$$m_k(\mu) = k \int_0^\infty x^{k-1} \mu(x, \infty) dx. \tag{4.10}$$

This identity is enough to obtain the following bound.

Lemma 4.5.6. Let $\mu, \nu \in \mathcal{P}_C$ and $k > 0$, then $|m_k(\mu) - m_k(\nu)| \leq 2C^k \mathrm{Kol}(\mu, \nu)$.

Proof. Start by decomposing μ into μ_+ and μ_- as follows:

$$\mu_+(A) = \mu(A \cap [0, \infty)), \;\; \mu_-(A) = \mu(-A \cap (-\infty, 0)) \;\; \forall A \in \mathcal{B}(\mathbb{R}).$$

Hence $\mu(A) = \mu_+(A) + \mu_-(-A)$. Define ν_+ and ν_- analogously. Note that these new measures are supported on $[0, \infty)$.

Observe that $m_k(\mu) = m_k(\mu_+) + (-1)^k m_k(\mu_-)$ and that the analogous formula holds for $m_k(\nu)$. Hence

$$|m_k(\mu) - m_k(\nu)| \leq |m_k(\mu_+) - m_k(\nu_+)| + |m_k(\mu_-) - m_k(\nu_-)|.$$

Now, for $t \geq 0$ define $F_{\mu_+}(t) = \mu_+(t, \infty)$ and $F_{\nu_+}(t) = \nu_+(t, \infty)$. By definition of Kolmogorov distance we have that

$$|F_{\mu_+}(t) - F_{\nu_+}(t)| \leq \mathrm{Kol}(\mu, \nu).$$

On the other hand, by equation (4.10) we have that

$$\begin{aligned} |m_k(\mu_+) - m_k(\nu_+)| &\leq k \int_0^\infty x^{k-1} |F_{\mu_+}(x) - F_{\nu_+}(x)| dx \\ &\leq k\mathrm{Kol}(\mu, \nu) \int_0^C x^{k-1} dx \\ &= C^k \mathrm{Kol}(\mu, \nu). \end{aligned}$$

In the exact same way we can bound $|m_k(\mu_-) - m_k(\nu_-)|$ to conclude the proof. □

Given $\mu \in \mathcal{P}_C$ we denote the $(k+1) \times (k+1)$ Hankel matrix of μ by $M_k(\mu)$ and define $D_k(\mu) = \det M_k(\mu)$. We will denote the Jacobi coefficients of μ by α_i^μ and β_i^μ. For the proof of the following results, many of the facts stated in Section 1.1 will be used.

Proposition 4.5.7. Let $\mu, \nu \in \mathcal{P}_C$ and let $s_k > 0$ be constants satisfying

$$\min\{D_j(\mu), D_j(\nu)\} \geq s_k$$

for $j = 1, \ldots, k$. Then

$$|\beta_k^\mu - \beta_k^\nu| \leq \frac{\exp\{gk^2\}\mathrm{Kol}(\mu, \nu)}{s_k^2}.$$

for some $g > 0$ dependent of μ and ν but independent of k.

Proof. To shorten notation let $x_j = D_j(\mu)$ and $y_j = D_j(\nu)$. Without loss of generality $C > 1$. A direct application of Lemma 4.5.6 yields a rough bound between the distance in the Euclidean norm of the corresponding columns of the matrices $M_j(\mu)$ and $M_j(\nu)$. Namely, the columns are at distance less than $\sqrt{j+1}C^{2j-1}\mathrm{Kol}(\mu, \nu)$. The same reasoning yields that the norm of any column in $M_j(\mu)$ or $M_j(\nu)$ is bounded by $\sqrt{j+1}C^{2j-1}$. Hence, using Lemma 4.5.5 we get

$$|x_j - y_j| \leq (\sqrt{j+1})^{j+1} j (C^{(2j-1)} + \varepsilon)^{j+1} \mathrm{Kol}(\mu, \nu) \leq \exp\{gj^2\}\mathrm{Kol}(\mu, \nu)$$

for some $g > 0$ independent of k.

In what follows we will bound two other terms whose logarithm is also $O(k^2)$. The implied constants depend only on μ and ν, so we can modify g to be big enough for the following inequalities to hold as well. By the first expression in equation (1.8) we have that

$$\begin{aligned}|\beta_k^\mu - \beta_k^\nu| &= \left| \frac{\sqrt{x_{k-1}x_{k+1}}}{x_k} - \frac{\sqrt{y_{k-1}y_{k+1}}}{y_k} \right| \\ &\leq \frac{1}{x_k} |\sqrt{x_{k-1}x_{k+1}} - \sqrt{y_{k-1}y_{k+1}}| + \sqrt{y_{k-1}y_{k+1}} \left| \frac{1}{x_k} - \frac{1}{y_k} \right|. \end{aligned} \tag{4.11}$$

To bound the first term on the right-hand side of the above inequality we see that

$$\begin{aligned}|\sqrt{x_{k-1}x_{k+1}} - \sqrt{y_{k-1}y_{k+1}}| &= \frac{|x_{k-1}x_{k+1} - y_{k-1}y_{k+1}|}{\sqrt{x_{k-1}x_{k+1}} + \sqrt{y_{k-1}y_{k+1}}} \quad \text{and} \\ |x_{k-1}x_{k+1} - y_{k-1}y_{k+1}| &\leq x_{k-1}|x_{k+1} - y_{k+1}| + y_{k+1}|x_{k-1} - y_{k-1}| \\ &\leq \exp\{ak^2\}\mathrm{Kol}(\mu,\nu)\end{aligned}$$

which yields

$$\frac{1}{x_k} |\sqrt{x_{k-1}x_{k+1}} - \sqrt{y_{k-1}y_{k+1}}| \leq \frac{\exp\{gk^2\}\mathrm{Kol}(\mu,\nu)}{2s_k^2}. \tag{4.12}$$

On the other hand,

$$\sqrt{y_{k-1}y_{k+1}} \left| \frac{1}{x_k} - \frac{1}{y_k} \right| = \sqrt{y_{k-1}y_{k+1}} \frac{|x_k - y_k|}{x_k y_k} \leq \frac{\exp\{gk^2\}\mathrm{Kol}(\mu,\nu)}{2s_k^2}. \tag{4.13}$$

The result then follows from combining the previous inequalities (4.11), (4.12) and (4.13). □

Remark 4.5.8. The constants s_k have already been studied with sophisticated techniques for some families of measures; see [120] for an example. However, using results only from Section 4.4 it will be easy to show that for measures with an absolutely continuous part we have $|\log(s_k)| = O(k^2)$ where the implied constant depends only on μ, which is enough for the proof of Theorem 4.2.13.

In a similar fashion we can show that the coefficients of $p_k^\mu(x)$ are locally Lipschitz:

Proposition 4.5.9. Fix a positive integer k. Let μ, ν and s_k be as in Proposition 4.5.7. Denote the coefficients of x^i in $p_k^\mu(x)$ and $p_k^\nu(x)$ by a_i^μ and a_i^ν respectively. Then

$$|a_i^\mu - a_i^\nu| \leq \left(\frac{2}{s_k} + \frac{1}{s_k^2} \right) \mathrm{Kol}(\mu,\nu) \exp\{gk^2\}$$

for some $g > 0$ dependent on μ and ν but independent of k.

Proof. For $1 \leq i \leq k$ let $M_k^{(i)}(\mu)$ be the matrix obtained by removing the kth row and ith column of $M_k(\mu)$ and let $d_i(\mu) = \det(M_k^{(i)}(\mu))$. From identity (1.10) we have

$$a_i^\mu = \frac{d_i(\mu)}{\sqrt{D_{k-1}(\mu)D_k(\mu)}}.$$

Using the same notation as in the proof of Proposition 4.5.7 we have that

$$\begin{aligned} |a_i(\mu) - a_i(\nu)| &\leq \left| \frac{d_i(\mu)}{\sqrt{x_{k-1}x_k}} - \frac{d_i(\nu)}{\sqrt{y_{k-1}y_k}} \right| \\ &\leq \frac{1}{\sqrt{x_{k-1}x_k}} |d_i(\mu) - d_i(\nu)| + d_i(\nu) \left| \frac{1}{\sqrt{x_{k-1}x_k}} - \frac{1}{\sqrt{y_{k-1}y_k}} \right|. \end{aligned}$$

As before $\frac{1}{\sqrt{x_{k-1}x_k}} \leq \frac{1}{s_k}$, while $|d_i(\mu) - d_i(\nu)| \leq 2\mathrm{Kol}(\mu,\nu)\exp\{gk^2\}$ for some $g > 0$ dependent on μ and ν only. To bound the second term on the right-hand side of the above inequality note that $d_i(\nu) \leq \exp\{gk^2\}$ and that

$$\begin{aligned} \frac{1}{\sqrt{x_{k-1}x_k}} - \frac{1}{\sqrt{y_{k-1}y_k}} &= (x_{k-1}x_k y_{k-1}y_k)^{-\frac{1}{2}} |\sqrt{x_{k-1}x_k} - \sqrt{y_{k-1}y_k}| \\ &\leq \frac{1}{s_k^3} \exp\{gk^2\}\mathrm{Kol}(\mu,\nu). \end{aligned}$$

where the last inequality is a consequence of (4.12). The result follows. □

Corollary 4.5.10. Let μ, ν, s_k be as in Proposition 4.5.7. Then

$$|\alpha_k^\mu - \alpha_k^\nu| \leq \frac{\mathrm{Kol}(\mu,\nu)\exp\{-gk^2\}}{s_k^3}.$$

Proof. Recall that

$$\alpha_k^\mu = \int x p_k^2(x) d\mu(x) = \sum_{i,j=1}^{k} a_i^\mu a_j^\mu m_{i+j+1}(\mu).$$

As mentioned above, the quantities a_i^μ, a_i^ν and $m_i(\mu), n_i(\nu)$ are of size $O(\exp\{gk^2\})$. Putting this together with Proposition 4.5.9 and Lemma 4.5.6 we get that

$$|a_i^\mu a_j^\mu m_{i+j-1}(\mu) - a_i^\nu a_j^\nu m_{i+j-1}(\nu)| \leq \frac{\exp\{gk^2\}}{s_k^3}.$$

By adding over i, j and modifying g the result follows. □

In order to prove Theorem 4.2.13 and Proposition 4.2.14 we need one final lemma, which states that with overwhelming probability, the random measure μ_n^u is close in Kolmogorov distance to μ_n.

Lemma 4.5.11. For n large enough we have that

$$\mathbb{P}[\mathrm{Kol}(\mu_n^u, \mu_n) \geq n^{-\frac{1}{4}}] \leq \exp\{-n^{\frac{1}{4}}/8\}.$$

Proof. We must show that

$$\left| \sum_{i=1}^{k} u_i^2 - \frac{k}{n} \right| \leq n^{-\frac{1}{4}}$$

for all $1 \leq k \leq n$ with probability at least $1 - \exp\{-n^{1/4}/8\}$.

Fix $1 \leq k \leq n$. As in Section 4.3.3 start by considering $X_1, \ldots, X_k$ independent centered Gaussian random variables of variance $\frac{1}{n}$ and let $Z_k = \sum_{i=1}^{k} X_i^2$. Then by Lemma 4.3.11 we have that

$$\mathbb{P}\left[Z_k \geq \frac{k}{n} + n^{-\frac{1}{4}} \right] \leq e^{-t_1} \quad \text{and} \quad \mathbb{P}\left[Z_k \leq \frac{k}{n} - n^{-\frac{1}{4}} \right] \leq e^{-t_2}$$

where t_1 and t_2 the solutions to

$$n^{-\frac{1}{4}} = \frac{2\sqrt{kt_1}}{n} \quad \text{and} \quad n^{-\frac{1}{4}} = \frac{2\sqrt{kt_2} + 2t_2}{n} \tag{4.14}$$

respectively. Since $k \leq n$ it is clear from (4.14) that $\min\{t_1, t_2\} \geq \frac{n^{\frac{1}{4}}}{4}$. This implies that

$$\mathbb{P}\left[\left| Z_k - \frac{k}{n} \right| \geq n^{-\frac{1}{4}} \right] \leq \exp\{-n^{\frac{1}{4}}/4\}.$$

Now, letting k run from 1 to n, a union bound yields that

$$\mathbb{P}\left[\max_{1 \leq k \leq n} \left| Z_k - \frac{k}{n} \right| > n^{-\frac{1}{4}} \right] \leq n \exp\{-n^{\frac{1}{4}}/4\} \leq \frac{1}{2} \exp\{-n^{\frac{1}{4}}/8\},$$

where the last equality holds for n large enough. Now, as in the proof of Proposition (4.3.9) we can show by a standard coupling argument that if we take $u_i = X_i/\sqrt{Z_n}$, we will have that

$$\mathbb{P}\left[\max_{1 \leq k \leq n} \left| Z_k - \sum_{i=1}^{k} u_i^2 \right| \right] \leq \frac{1}{2} \exp\{-n^{\frac{1}{4}}/8\}$$

and the result follows. □

Proof of Theorem 4.2.13. From Lemma 4.5.11, for n large enough, we have that $\mathrm{Kol}(\mu^u, \mu_n) \leq n^{-\frac{1}{4}}$ with overwhelming probability. By the assumption $\mathrm{Kol}(\mu_n, \mu) = n^{-c}$ we then have that $\mathrm{Kol}(\mu^u, \mu) \leq n^{-c'}$ also with overwhelming probability for $c' = \min\{1/4, c\}$. Hence, under the event $\{\mathrm{Kol}(\mu^u, \mu) \leq n^{-c'}\}$ we can apply Proposition 4.5.7 and Corollary 4.5.10 and use the fact that the Jacobi matrices are tridiagonal to obtain that

$$\|J_{k_n}(u) - J_{k_n}(\mu)\| \leq \frac{6C \exp\{d'k^2\}}{n^{c'} \min\{s_k^2, s_k^3\}}.$$

Since μ has an absolutely continuous part we know from Proposition 4.4.3 and Corollary 4.4.5 that $|\log(\gamma_k^\mu)| = O(k)$. Hence, from equation (1.9) we get $|\log s_k| = O(k^2)$, which makes it clear that there exists $d > 0$ and a sequence $k_n \leq d\sqrt{\log n}$ satisfying the theorem statement. □

Proof of Proposition 4.2.14. As mentioned in Section 4.2, this proposition is a direct consequence of Theorem 4.2.13 and Lemma 4.2.6. □

Remark 4.5.12. Observe that the above proofs repeatedly use the fact that moments are Lipschitz quantities on $\mathcal{P}_C$ and that the Jacobi coefficients are an explicit function of the moments. However, going from moments to Jacobi coefficients is an expensive process which we pay for by getting a rate of $O(\sqrt{\log n})$ instead of $\Theta(\log n)$. At a first glance, it may seem that the results in Section 4.3.2 may be used in a similar fashion to obtain a better rate; however, even if we have strong concentration results for the Jacobi coefficients of the random measures μ_n^u, it is a difficult task to control the location of the medians (or means) of $\alpha_j(u)$ and $\beta_j(u)$ and hence it is hard to show that these quantities converge at a good enough rate to the Jacobi coefficients of μ.

4.6 Concluding remarks

Several directions can be pursued to expand the results presented throughout this chapter. Currently, we have only analyzed the Lanczos algorithm in its prototypical form, but have not analyzed the more sophisticated variants that are used in practice. Obtaining similar concentration results and negative results for these modifications, and more generally for Krylov subspace methods, would be of great interest.

The Lanczos algorithm is used in practice for non-Hermitian matrices and even non-normal matrices, despite these cases being far less understood. In this incarnation, the algorithm is referred to as the Arnoldi algorithm. Extending the results of this chapter to the Arnoldi algorithm is a natural direction to pursue. As mentioned in Remark 4.4.12, it is easy to extend Theorem 4.2.2 to the non-Hermitian setting, but no longer so easy to prove concentration of the Ritz values or to say anything about their location.

A less fundamental but still important task is to sharpen the constants in the results in this article. Currently, our concentration inequalities become meaningful when the matrices involved have dimension of the order $n = 10^7$. The main offending term is the coefficient $\left(\frac{\omega}{4\|A\|}\right)^k$ in the exponential, which limits us to k very small. We believe that the constants can be sharpened significantly, which would allow the results to apply to smaller matrices, more iterations, and yield tighter probability bounds.

Finally, Theorem 4.2.13 and Proposition 4.2.14 have natural places for improvement. Below we pose two concrete questions in this setting which we leave open.

Question 1. Is there a natural metric on $\mathcal{P}_C$ inducing a topology for which the set of atomic measures is a dense subset of $\mathcal{P}_C$ and such that the Jacobi coefficients

$$\alpha_j : \mathcal{P}_C \to \mathbb{R} \quad \text{and} \quad \beta_j : \mathcal{P}_C \to \mathbb{R},$$

have a local Lipschitz constant of size at most exponential in j?

Question 2. Do Proposition 4.2.13 and Theorem 4.2.14 still hold if the hypothesis $1 \leq k_n \leq d\sqrt{\log n}$ is replaced by $1 \leq k_n \leq d' \log n$?

Bibliography

[1] Michael Aizenman et al. "Matrix regularizing effects of Gaussian perturbations". In: *Communications in Contemporary Mathematics* 19.03 (2017), p. 1750028.

[2] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Vol. 118. Cambridge university press, 2010.

[3] Diego Armentano and Felipe Cucker. "A randomized homotopy for the Hermitian eigenpair problem". In: *Foundations of Computational Mathematics* 15.1 (2015), pp. 281–312.

[4] Diego Armentano et al. "A stable, polynomial-time algorithm for the eigenpair problem". English. In: *Journal of the European Mathematical Society* 20.6 (2018), pp. 1375–1437. ISSN: 1435-9855. DOI: 10.4171/JEMS/789.

[5] Gérard Ben Arous and Paul Bourgade. "Extreme gaps between eigenvalues of random matrices". In: *The Annals of Probability* 41.4 (2013), pp. 2648–2681.

[6] Guillaume Aubrun and Stanisław J. Szarek. *Alice and Bob Meet Banach: The Interface of Asymptotic Geometric Analysis and Quantum Information Theory*. Vol. 223. American Mathematical Soc., 2017.

[7] Zhaojun Bai and James Demmel. "Using the matrix sign function to compute invariant subspaces". In: *SIAM Journal on Matrix Analysis and Applications* 19.1 (1998), pp. 205–225.

[8] Zhaojun Bai, James Demmel, and Ming Gu. "An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems". In: *Numerische Mathematik* 76.3 (1997), pp. 279–308.

[9] Grey Ballard, James Demmel, and Ioana Dumitriu. "Minimizing communication for eigenproblems and the singular value decomposition". In: *arXiv preprint arXiv:1011.3077* (2010).

[10] Grey Ballard et al. "A Generalized Randomized Rank-Revealing Factorization". In: *arXiv preprint arXiv:1909.06524* (2019).

[11] Jess Banks et al. "Gaussian Regularization of the Pseudospectrum and Davies' Conjecture". In: *arXiv preprint arXiv:1906.11819, to appear in Communications on Pure and Applied Mathematics* (2019).

[12] Jess Banks et al. *Overlaps, Eigenvalue Gaps, and Pseudospectrum under real Ginibre and Absolutely Continuous Perturbations*. 2020. arXiv: 2005.08930 [math.PR].

[13] Jess Banks et al. *Pseudospectral Shattering, the Sign Function, and Diagonalization in Nearly Matrix Multiplication Time*. 2019. arXiv: 1912.08805 [math.NA].

[14] Anirban Basak, Elliot Paquette, and Ofer Zeitouni. "Regularization of non-normal matrices by Gaussian noise - the banded Toeplitz and twisted Toeplitz cases". In: *Forum of Mathematics, Sigma*. Vol. 7. Cambridge University Press. 2019.

[15] Anirban Basak, Elliot Paquette, and Ofer Zeitouni. "Spectrum of random perturbations of Toeplitz matrices with finite symbols". In: *arXiv preprint arXiv:1812.06207* (2018).

[16] Friedrich L. Bauer and Charles T. Fike. "Norms and exclusion theorems". In: *Numerische Mathematik* 2.1 (1960), pp. 137–141.

[17] A. N. Beavers Jr. and E. D. Denman. "A new similarity transformation method for eigenvalues and eigenvectors". In: *Mathematical Biosciences* 21.1-2 (1974), pp. 143–169.

[18] A. N. Beavers and E. D. Denman. "A computational method for eigenvalues and eigenvectors of a matrix with real eigenvalues". In: *Numerische Mathematik* 21.5 (1973), pp. 389–396.

[19] Bernhard Beckermann. "A note on the convergence of Ritz values for sequences of matrices". In: *Publication ANO* 408 (2000).

[20] Mohammed Bellalij, Yousef Saad, and Hassane Sadok. "Further analysis of the Arnoldi process for eigenvalue problems". In: *SIAM Journal on Numerical Analysis* 48.2 (2010), pp. 393–407.

[21] Michael Ben-Or and Lior Eldar. "A Quasi-Random Approach to Matrix Spectral Analysis". In: *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2018.

[22] Florent Benaych-Georges and Ofer Zeitouni. "Eigenvectors of non normal random matrices". In: *Electronic Communications in Probability* 23 (2018).

[23] David Bindel et al. *A fast and stable nonsymmetric eigensolver for certain structured matrices*. Tech. rep. Technical report, University of California, Berkeley, CA, 2005.

[24] Paul Bourgade and Guillaume Dubach. "The distribution of overlaps between eigenvectors of Ginibre matrices". In: *arXiv preprint arXiv:1801.01219* (2018).

[25] Marie-France Bru. "Diffusions of perturbed principal component analysis". In: *Journal of multivariate analysis* 29.1 (1989), pp. 127–136.

[26] Marie-France Bru. "Wishart processes". In: *Journal of Theoretical Probability* 4.4 (1991), pp. 725–751.

[27] Ralph Byers. "Numerical stability and instability in matrix sign function based algorithms". In: *Computational and Combinatorial Methods in Systems Theory*. Citeseer. 1986.

[28] Ralph Byers, Chunyang He, and Volker Mehrmann. "The matrix sign function method and the computation of invariant subspaces". In: *SIAM Journal on Matrix Analysis and Applications* 18.3 (1997), pp. 615–632.

[29] Ralph Byers and Hongguo Xu. "A new scaling for Newton's iteration for the polar decomposition and its backward stability". In: *SIAM Journal on Matrix Analysis and Applications* 30.2 (2008), pp. 822–843.

[30] Jin-yi Cai. "Computing Jordan normal forms exactly for commuting matrices in polynomial time". In: *International Journal of Foundations of Computer Science* 5.03n04 (1994), pp. 293–302.

[31] Daniela Calvetti, Lothar Reichel, and Danny Chris Sorensen. "An implicitly restarted Lanczos method for large symmetric eigenvalue problems". In: *Electronic Transactions on Numerical Analysis* 2.1 (1994), p. 21.

[32] John T. Chalker and Bernhard Mehlig. "Eigenvector statistics in non-Hermitian random matrix ensembles". In: *Physical review letters* 81.16 (1998), p. 3367.

[33] Robert M. Corless et al. "On the Lambert W function". In: *Advances in Computational mathematics* 5.1 (1996), pp. 329–359.

[34] K. R. Davidson, D. A. Herrero, and N Salinas. "Quasidiagonal Operators, Approximation, and C*–Algebras". In: *Indiana University Mathematics Journal* 38.4 (1989), pp. 973–998.

[35] Kenneth R. Davidson and Stanisław J. Szarek. "Local operator theory, random matrices and Banach spaces". In: *Handbook of the geometry of Banach spaces* 1.317-366 (2001), p. 131.

[36] E Brian Davies. "Approximate diagonalization". In: *SIAM Journal on Matrix Analysis and Applications* 29.4 (2007), pp. 1051–1064.

[37] E. Brian Davies. "Approximate diagonalization". In: *SIAM Journal on Matrix Analysis and Applications* 29.4 (2007), pp. 1051–1064.

[38] E. Brian Davies and Mildred Hager. "Perturbations of Jordan matrices". In: *Journal of Approximation Theory* 156.1 (2009), pp. 82–94.

[39] Chandler Davis and William Morton Kahan. "The rotation of eigenvectors by a perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pp. 1–46.

[40] Emilio Defez et al. "An efficient and accurate algorithm for computing the matrix cosine based on new Hermite approximations". In: *Journal of Computational and Applied Mathematics* 348 (2019), pp. 1–13.

[41] Percy Deift. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*. Vol. 3. American Mathematical Soc., 1999.

[42] Percy Deift and Dimitri Gioev. *Random matrix theory: invariant ensembles and universality*. Vol. 18. American Mathematical Soc., 2009.

[43] James W. Demmel. *Applied numerical linear algebra*. Vol. 56. SIAM, 1997.

[44] James W. Demmel. "The probability that a numerical analysis problem is difficult". In: *Mathematics of Computation* 50.182 (1988), pp. 449–480.

[45] James Weldon Demmel. *A numerical analyst's Jordan canonical form*. Tech. rep. UC Berkeley Center for Pure and Applied Mathematics, 1983.

[46] James Weldon Demmel. "On condition numbers and the distance to the nearest ill-posed problem". In: *Numerische Mathematik* 51.3 (1987), pp. 251–289.

[47] James Demmel, Ioana Dumitriu, and Olga Holtz. "Fast linear algebra is stable". In: *Numerische Mathematik* 108.1 (2007), pp. 59–91.

[48] James Demmel et al. "Fast matrix multiplication is stable". In: *Numerische Mathematik* 106.2 (2007), pp. 199–224.

[49] Eugene D. Denman and Alex N. Beavers Jr. "The matrix sign function and computations in systems". In: *Applied mathematics and Computation* 2.1 (1976), pp. 63–94.

[50] Xiaodong Ding and Rangquan Wu. "A new proof for comparison theorems for stochastic differential inequalities with respect to semimartingales". In: *Stochastic Processes and their applications* 78.2 (1998), pp. 155–171.

[51] Ioana Dumitriu. "Smallest eigenvalue distributions for two classes of β-Jacobi ensembles". In: *Journal of Mathematical Physics* 53.10 (2012), p. 103301.

[52] Alan Edelman. "Eigenvalues and condition numbers of random matrices". In: *SIAM Journal on Matrix Analysis and Applications* 9.4 (1988), pp. 543–560.

[53] Alan Edelman and N. Raj Rao. "Random matrix theory". In: *Acta Numerica* 14 (2005), pp. 233–297.

[54] Alan Edelman and Brian D. Sutton. "The beta-Jacobi matrix model, the CS decomposition, and generalized singular value problems". In: *Foundations of Computational Mathematics* 8.2 (2008), pp. 259–285.

[55] László Erdős and Horng-Tzer Yau. "Universality of local spectral statistics of random matrices". In: *Bulletin of the American Mathematical Society* 49.3 (2012), pp. 377–414.

[56] Ohad N. Feldheim, Elliot Paquette, and Ofer Zeitouni. "Regularization of non-normal matrices by Gaussian noise". In: *International Mathematics Research Notices* 2015.18 (2014), pp. 8724–8751.

[57] Peter J. Forrester. *Log-gases and random matrices (LMS-34)*. Princeton University Press, 2010.

[58] Yan V. Fyodorov. "On statistics of bi-orthogonal eigenvectors in real and complex Ginibre ensembles: combining partial Schur decomposition with supersymmetry". In: *Communications in Mathematical Physics* 363.2 (2018), pp. 579–603.

[59] Walter Gautschi. "Construction of Gauss-Christoffel quadrature formulas". In: *Math. Comp* 22.102 (1968), pp. 251–270.

[60] Stephen Ge. "The Eigenvalue Spacing of IID Random Matrices and Related Least Singular Value Results". PhD thesis. UCLA, 2017.

[61] Christel Geiß and Ralf Manthey. "Comparison theorems for stochastic differential equations in finite and infinite dimensions". In: *Stochastic processes and their applications* 53.1 (1994), pp. 23–35.

[62] Chris Godsil. *Algebraic combinatorics*. Routledge, 2017.

[63] Gene Howard Golub and Richard Underwood. "The block Lanczos method for computing eigenvalues". In: *Mathematical software*. Elsevier, 1977, pp. 361–377.

[64] Piotr Graczyk and Jacek Małecki. "Strong solutions of non-colliding particle systems". In: *Electronic Journal of Probability* 19 (2014).

[65] Anne Greenbaum, Ren-cang Li, and Michael L Overton. "First-order perturbation theory for eigenvalues and eigenvectors". In: *SIAM Review* 62.2 (2020), pp. 463–482.

[66] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*. Vol. 2. Springer Science & Business Media, 2012.

[67] Ming Gu and Stanley C. Eisenstat. "Efficient algorithms for computing a strong rank-revealing QR factorization". In: *SIAM Journal on Scientific Computing* 17.4 (1996), pp. 848–869.

[68] Alice Guionnet, Philip Wood, and Ofer Zeitouni. "Convergence of the spectral measure of non-normal matrices". In: *Proceedings of the American Mathematical Society* 142.2 (2014), pp. 667–679.

[69] Uffe Haagerup and Flemming Larsen. "Brown's spectral distribution measure for R-diagonal elements in finite von Neumann algebras". In: *Journal of Functional Analysis* 176.2 (2000), pp. 331–367.

[70] Nicholas J. Higham. *Accuracy and stability of numerical algorithms*. Vol. 80. SIAM, 2002.

[71] Nicholas J. Higham. *Functions of matrices: theory and computation*. Vol. 104. SIAM, 2008.

[72] Nicholas J. Higham. "The matrix sign decomposition and its relation to the polar decomposition". In: *Linear Algebra and its Applications* 212 (1994), pp. 3–20.

[73] Nicholas J. Higham and Lijing Lin. "An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives". In: *SIAM Journal on Matrix Analysis and Applications* 34.3 (2013), pp. 1341–1360.

[74] Walter Hoffmann and Beresford N Parlett. "A new proof of global convergence for the tridiagonal QL algorithm". In: *SIAM Journal on Numerical Analysis* 15.5 (1978), pp. 929–937.

[75] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[76] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. *On the real Davies' conjecture*. 2020. arXiv: 2005.08908 [math.FA].

[77] Shmuel Kaniel. "Estimates for some computational techniques in linear algebra". In: *Mathematics of Computation* 20.95 (1966), pp. 369–378.

[78] Charles S. Kenney and Alan J Laub. "The matrix sign function". In: *IEEE Transactions on Automatic Control* 40.8 (1995), pp. 1330–1348.

[79] Wolfgang König, Neil O'Connell, et al. "Eigenvalues of the Laguerre process as non-colliding squared Bessel processes". In: *Electronic Communications in Probability* 6 (2001), pp. 107–114.

[80] Vladislav Krasin. "Comparison theorem and its applications to finance". PhD thesis. Edmonton, Alberta: University of Alberta, 2010.

[81] Jacek Kuczyński and Henryk Woźniakowski. "Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm". In: *SIAM Journal on Matrix Analysis and Applications* 15.2 (1994), pp. 672–691.

[82] Arno B. J. Kuijlaars. "Convergence analysis of Krylov subspace iterations with methods from potential theory". In: *SIAM review* 48.1 (2006), pp. 3–40.

[83] Arno B. J. Kuijlaars. "Which eigenvalues are found by the Lanczos method?" In: *SIAM Journal on Matrix Analysis and Applications* 22.1 (2000), pp. 306–321.

[84] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics* (2000), pp. 1302–1338.

[85] Huiling Le. "Brownian motions on shape and size-and-shape spaces". In: *Journal of applied probability* 31.1 (1994), pp. 101–113.

[86] Huiling Le. "Singular-values of matrix-valued Ornstein–Uhlenbeck processes". In: *Stochastic processes and their applications* 82.1 (1999), pp. 53–60.

[87] Ruipeng Li et al. "A Thick-Restart Lanczos algorithm with polynomial filtering for Hermitian eigenvalue problems". In: *SIAM Journal on Scientific Computing* 38.4 (2016), A2512–A2534.

[88] Lin Lin, Yousef Saad, and Chao Yang. "Approximating spectral densities of large matrices". In: *SIAM review* 58.1 (2016), pp. 34–65.

[89] Anand Louis and Santosh S Vempala. "Accelerated newton iteration for roots of black box polynomials". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 732–740.

[90] Alexander N. Malyshev. "Parallel algorithm for solving some spectral problems of linear algebra". In: *Linear algebra and its applications* 188 (1993), pp. 489–520.

[91] Per-Gunnar Martinsson and Joel Tropp. "Randomized numerical linear algebra: Foundations & algorithms". In: *arXiv preprint arXiv:2002.01387* (2020).

[92] Francesco Mezzadri. "How to generate random matrices from the classical compact groups". In: *arXiv preprint math-ph/0609050* (2006).

[93] Dragoslav S. Mitrinovic, Josip Pecaric, and Arlington M Fink. *Inequalities involving functions and their integrals and derivatives*. Vol. 53. Springer Science & Business Media, 1991.

[94] Awad H. Al-Mohy, Nicholas J. Higham, and Samuel D. Relton. "Computing the Fréchet derivative of the matrix logarithm and estimating the condition number". In: *SIAM Journal on Scientific Computing* 35.4 (2013), pp. C394–C410.

[95] Awad H. Al-Mohy, Nicholas J. Higham, and Samuel D. Relton. "New algorithms for computing the matrix sine and cosine separately or simultaneously". In: *SIAM Journal on Scientific Computing* 37.1 (2015), A456–A487.

[96] Mervin E Muller. "A note on a method for generating points uniformly on n-dimensional spheres". In: *Communications of the ACM* 2.4 (1959), pp. 19–20.

[97] Prashanth Nadukandi and Nicholas J. Higham. "Computing the wave-kernel matrix functions". In: *SIAM Journal on Scientific Computing* 40.6 (2018), A4060–A4082.

[98] Yuji Nakatsukasa and Roland W. Freund. "Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarev's functions". In: *SIAM Review* 58.3 (2016), pp. 461–493.

[99] Yuji Nakatsukasa and Nicholas J. Higham. "Stable and efficient spectral divide and conquer algorithms for the symmetric eigenvalue decomposition and the SVD". In: *SIAM Journal on Scientific Computing* 35.3 (2013), A1325–A1349.

[100] Hoi Nguyen, Terence Tao, and Van Vu. "Random matrices: tail bounds for gaps between eigenvalues". In: *Probability Theory and Related Fields* 167.3-4 (2017), pp. 777–816.

[101] Christopher Conway Paige. "The computation of eigenvalues and eigenvectors of very large sparse matrices." PhD thesis. University of London, 1971.

[102] Victor Y. Pan and Zhao Q. Chen. "The complexity of the matrix eigenproblem". In: *Proceedings of the thirty-first annual ACM symposium on Theory of computing*. ACM. 1999, pp. 507–516.

[103] Beresford N. Parlett. *The symmetric eigenvalue problem*. Vol. 20. SIAM, 1998.

[104] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Vol. 293. Springer Science & Business Media, 1999.

[105] John Douglas Roberts. "Linear model reduction and solution of the algebraic Riccati equation by use of the sign function". In: *International Journal of Control* 32.4 (1980), pp. 677–687.

[106] Axel Ruhe. "Closest normal matrix finally found!" In: *BIT Numerical Mathematics* 27.4 (1987), pp. 585–598.

[107] Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. Vol. 66. Siam, 2011.

[108] Yousef Saad. "On the rates of convergence of the Lanczos and the block-Lanczos methods". In: *SIAM Journal on Numerical Analysis* 17.5 (1980), pp. 687–706.

[109] Arvind Sankar, Daniel A. Spielman, and Shang-Hua Teng. "Smoothed analysis of the condition numbers and growth factors of matrices". In: *SIAM Journal on Matrix Analysis and Applications* 28.2 (2006), pp. 446–476.

[110] Dai Shi and Yunjiang Jiang. "Smallest Gaps Between Eigenvalues of Random Matrices With Complex Ginibre, Wishart and Universal Unitary Ensembles". In: *arXiv preprint arXiv:1207.4240* (2012).

[111] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. "Tight query complexity lower bounds for PCA via finite sample deformed Wigner law". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2018, pp. 1249–1259.

[112] Johannes Sjöstrand and Martin Vogel. "General Toeplitz matrices subject to Gaussian perturbations". In: *arXiv preprint arXiv:1905.10265* (2019).

[113] Johannes Sjöstrand and Martin Vogel. "Toeplitz band matrices with small random perturbations". In: *arXiv preprint arXiv:1901.08982* (2019).

[114] Steve Smale. "Complexity theory and numerical analysis". In: *Acta Numerica* 6 (1997), pp. 523–551.

[115] Steve Smale. "On the efficiency of algorithms of analysis". In: *Bulletin (New Series) of The American Mathematical Society* 13.2 (1985), pp. 87–121.

[116] Piotr Śniady. "Random regularization of Brown spectral measure". In: *Journal of Functional Analysis* 193.2 (2002), pp. 291–313.

[117] Daniel A. Spielman and Shang-Hua Teng. "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time". In: *Journal of the ACM (JACM)* 51.3 (2004), pp. 385–463.

[118] Ji-Guang Sun. "Perturbation bounds for the Cholesky and QR factorizations". In: *BIT Numerical Mathematics* 31.2 (1991), pp. 341–352.

[119] Stanislaw J. Szarek. "Condition numbers of random matrices". In: *Journal of Complexity* 7.2 (1991), pp. 131–149.

[120] Gabor Szegő. "Hankel forms". In: *American Mathematical Society Translations* 108 (1977).

[121] Gabor Szegő. *Orthogonal polynomials*. Vol. 23. American Mathematical Soc., 1939.

[122] Terence Tao. *Topics in random matrix theory*. Vol. 132. American Mathematical Soc., 2012.

[123] Terence Tao, Van Vu, and Manjunath Krishnapur. "Random matrices: Universality of ESDs and the circular law". In: *The Annals of Probability* 38.5 (2010), pp. 2023–2065.

[124] Lloyd N. Trefethen and David Bau III. *Numerical linear algebra*. Vol. 50. Siam, 1997.

[125] Lloyd N. Trefethen and Mark Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.

[126] Jos L. M. Van Dorsselaer, Michiel E. Hochstenbach, and Henk A Van Der Vorst. "Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method". In: *SIAM Journal on Matrix Analysis and Applications* 22.3 (2001), pp. 837–852.

[127] Jorge Garza Vargas and Archit Kulkarni. "The Lanczos Algorithm Under Few Iterations: Concentration and Location of the Ritz Values". In: *arXiv preprint arXiv:1904.06012* (2019).

[128] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[129] Roman Vershynin. "On the role of sparsity in compressed sensing and random matrix theory". In: *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE. 2009, pp. 189–192.

[130] Jade P. Vinson. "Closest spacing of eigenvalues". In: *arXiv preprint arXiv:1111.2743* (2011).

[131] John Von Neumann and Herman H. Goldstine. "Numerical inverting of matrices of high order". In: *Bulletin of the American Mathematical Society* 53.11 (1947), pp. 1021–1099.

[132] James Hardy Wilkinson. "Global convergence of tridiagonal QR algorithm with origin shifts". In: *Linear Algebra and its Applications* 1.3 (1968), pp. 409–420.

[133] Thomas G. Wright and Lloyd N. Trefethen. "EigTool". In: (2002). Software available at http://www.comlab.ox.ac.uk/pseudospectra/eigtool.

[134] Qiaochu Yuan, Ming Gu, and Bo Li. "Superlinear convergence of randomized block Lanczos algorithm". In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, pp. 1404–1409.

Appendix A

Deferred Proofs

A.1 SDE analysis

The goal of this section is to adapt Śniady's [116] proof of Theorem 3.3.2, as outlined below the statement of Theorem 3.3.2, to the case of real matrices with real Ginibre perturbations.

The stochastic differential equation satisfied by the squared singular values of a real matrix Brownian motion was derived by Bru in her work on Wishart processes [25, 26] and independently by Le in her work on shape theory [85, 86]. The equation reads as follows:

$$d\lambda_i = \frac{2\sqrt{\lambda_i}}{n}\, dB_i + \left(1 + \sum_{j \neq i} \frac{\lambda_i + \lambda_j}{\lambda_i - \lambda_j}\right) dt, \qquad 1 \leq i \leq n. \tag{A.1}$$

The proof strategy of Śniady crucially relies on the existence and uniqueness of strong solutions to the singular value SDE. This is needed in order to obtain two solutions driven by the same Brownian motion, and to assert that the law of each solution indeed matches the law of the singular values of a noncentered Ginibre matrix. See [2] for a definition of strong solution and a rigorous proof of existence and uniqueness of strong solutions for Dyson Brownian motion, the Hermitian analogue of the Ginibre singular values process.

Fortunately, such results are known for the SDE (A.1). Let Λ denote the domain

$$\Lambda \in \mathbb{R}^n := \{\lambda : 0 \leq \lambda_n < \cdots < \lambda_1\}.$$

For any initial data $\lambda(0)$ lying in the closure $\overline{\Lambda}$, it is known that strong solutions to (A.1) exist, are unique, and lie in Λ for all $t > 0$, almost surely [64, Corollary 6.5]. Combining this with [25, Theorem 1], we have that for initial data $\lambda(0)$ lying in Λ, the law of the strong solutions to (A.1) matches the law of the squared singular values process of $A + M/\sqrt{n}$, where M is a matrix of i.i.d. standard real Brownian motions and A has squared singular values $\lambda(0)$. (It should be possible to extend this last statement for initial data in $\overline{\Lambda}$, but the proof may be somewhat involved—cf. [2], which contains a proof of the corresponding extension for Dyson Brownian motion.)

Let $a_i(\lambda) = 1 + \sum_{j \neq i} \frac{\lambda_i + \lambda_j}{\lambda_i - \lambda_j}$ denote the drift coefficient in (A.1). As in Śniady's proof for the complex Ginibre case (Theorem 3.3.2), the key property of a allowing for the comparison theorem is the so-called *quasi-monotonicity* (see [50]) or *Kamke–Ważewski condition* [93, §XI.13] from differential inequalities, which is simply that

$$\text{for all } i,\ a_i(\lambda^{(1)}) \leq a_i(\lambda^{(2)}) \text{ whenever } \lambda_i^{(1)} = \lambda_i^{(2)} \text{ and } \lambda_j^{(1)} \leq \lambda_j^{(2)} \text{ for all } j \neq i. \tag{A.2}$$

One easily checks that a satisfies this condition on the domain Λ.

The nonconstant (indeed, non-Lipschitz) diffusion coefficient $2\sqrt{\lambda_i}/n$ in (A.1) is a technical obstacle which does not appear in the SDE (2.3) for the complex case. Consequently, the final step of Śniady's proof as sketched below Theorem 3.3.2 cannot be repeated naively, because taking the difference of two solutions no longer cancels out the diffusion terms. Fortunately, theory has been developed to handle Hölder-1/2 diffusion coefficients; see [104, §IX.3] for exposition of the one-dimensional case and see [80] for a survey of comparison theorems for SDEs in general.

Quasi-monotonicity and the one-dimensional Hölder-1/2 comparison theory are combined in a rather general multidimensional comparison theorem of Geiß and Manthey [61, Theorem 1.2]. Applied to the SDE (A.1), this theorem provides exactly the right conclusion to replace the final step of Śniady's proof. We state the relevant special case of their theorem below:

Theorem A.1.1 (Geiß-Manthey)**.** Consider the SDE

$$dX_i = \sigma_i(X)\, dB_i + a_i(X)\, dt, \qquad 1 \leq i \leq n,$$

where the B_i are independent standard real Brownian motions, and $\sigma_i, a_i : \mathbb{R}^n \to \mathbb{R}$ are continuous. Suppose the following conditions are satisfied:

1. the drift coefficient a satisfies the quasi-monotonicity condition (A.2)
2. there exists $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ increasing with $\int_0^\varepsilon \rho^{-2}(u)\, du = \infty$ for some $\varepsilon > 0$, such that $|\sigma_i(x) - \sigma_i(y)| \leq \rho(|x_i - y_i|)$ for all i and all $x, y \in \mathbb{R}^n$
3. strong solutions for the SDE exist for all time and are unique.

Suppose initial conditions $X^{(1)}(0)$ and $X^{(2)}(0)$ satisfy the inequality $X_i^{(1)}(0) \leq X_i^{(2)}(0)$ for all i. Then almost surely, $X_i^{(1)}(t) \leq X_i^{(2)}(t)$ for all i and for all $t > 0$.

Setting $\rho(u) := \sqrt{u}$, the SDE (A.1) satisfies the conditions of the Geiß-Manthey theorem, except that our domain for both a_i and σ_i is Λ, not $\mathbb{R}^n$. We address these two coefficients in turn.

First we deal with the drift coefficient a_i, using a standard localization argument already implicit in the proof of Geiß and Manthey. They (implicitly) define the stopping time ϑ_N to be the first time $\|X^{(1)}\| \geq N$ or $\|X^{(2)}\| \geq N$, and use the fact that a is Lipschitz on the restricted domain $\|X\| \leq N$ to show that

$$\mathbb{P}\left[X_i^{(1)}(t) \leq X_i^{(2)}(t) \text{ for all } 0 \leq t \leq \vartheta_N\right] = 1.$$

Since strong solutions exist for all time, we have $\vartheta_N \to \infty$ as $N \to \infty$ almost surely, which proves the theorem. We modify this strategy for our SDE (A.1) in the standard way: Define the stopping time $\tau_{1/m}$ to be the first time either $\lambda^{(1)}$ or $\lambda^{(2)}$ leaves the set

$$\Lambda_{1/m} := \{\lambda \in \Lambda : |\lambda_i - \lambda_{i+1}| > 1/m \text{ for all } 1 \le i \le n-1.\}.$$

Since strong solutions starting in Λ stay in Λ for all $t \ge 0$ and are continuous, we have $\tau_{1/m} \to \infty$ as $m \to \infty$ almost surely. Since our a is Lipschitz on $\Lambda_{1/m}$, the proof of Theorem A.1.1 shows that

$$\mathbb{P}\left[\lambda_i^{(1)}(t) \le \lambda_i^{(2)}(t) \text{ for all } 0 \le t \le \tau_{1/m}\right] = 1$$

for all m. Taking $m \to \infty$, the result follows.

Finally, we address the diffusion coefficient $\sigma_i(\lambda) = 2\sqrt{\lambda_i}/n$. The standard fix is to first modify the SDE to have diffusion coefficients $2\sqrt{|\lambda_i|}/n$ for all i, so that the domain of σ_i is enlarged to $\mathbb{R}^n$ and Theorem A.1.1 may be applied. For this modified SDE, note that the constant zero function $\lambda^{(1)}(t) = 0$ is a strong solution. Now let $\lambda^{(2)}$ be any solution with $\lambda_i^{(2)}(0) \ge 0$ for all i. Applying Theorem A.1.1 to $\lambda^{(1)}$ and $\lambda^{(2)}$, we conclude that in fact, $\lambda^{(2)}(t) \ge 0$ for all $t \ge 0$. Thus, the absolute value bars in the modified SDE can be removed a posteriori. This argument is used, for example, when setting up the SDE for the so-called *Bessel process*, which shares this square-root diffusion coefficient—see [104, §XI.1] for details.

A.2 Deferred Proofs from Section 3.4

Lemma A.2.1 (Restatement of Lemma 3.4.11)**.** Assume the matrix inverse is computed by an algorithm INV satisfying the guarantee in Definition 3.2.3. Then $\mathsf{G}(A) = g(A) + E$ for some error matrix E with norm

$$\|E\| \le \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|\right) 4\sqrt{n}\mathbf{u}. \tag{A.3}$$

Proof. The computation of $\mathsf{G}(A)$ consists of three steps:

1. Form A^{-1} according to Definition 3.2.3. This incurs an additive error of $E_{\mathsf{INV}} = \mu_{\mathsf{INV}}(n) \cdot \mathbf{u} \cdot \kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|$. The result is $\mathsf{INV}(A) = A^{-1} + E_{\mathsf{INV}}$.

2. Add A to $\mathsf{INV}(A)$. This incurs an entry-wise relative error of size $\mathbf{u}$: The result is

   $$(A + A^{-1} + E_{\mathsf{INV}}) \circ (J + E_{add})$$

   where J denotes the all-ones matrix, $\|E_{add}\|_{max} \le \mathbf{u}$, and where $\circ$ denotes the entrywise (Hadamard) product of matrices.

3. Divide the resulting matrix by 2. This incurs an entrywise relative error of size $\mathbf{u}$. The final result is

   $$\mathsf{G}(A) = \frac{1}{2}(A + A^{-1} + E_{\mathsf{INV}}) \circ (J + E_{add}) \circ (J + E_{div})$$

   where $\|E_{div}\|_{\max} \le \mathbf{u}$.

Finally, recall that for any $n \times n$ matrices M and E, we have the relation (3.8)

$$\|M \circ E\| \leq \|M\|\|E\|_{max}\sqrt{n}.$$

Putting it all together, we have

$$\begin{aligned}\|\mathsf{G}(A) - g(A)\| &\leq \frac{1}{2}\left(\|A\| + \|A^{-1}\|\right)(2\mathbf{u} + \mathbf{u}^2)\sqrt{n} + \|E_{\mathsf{INV}}\|(1+\mathbf{u})^2\sqrt{n} \\ &\leq \frac{1}{2}\left(\|A\| + \|A^{-1}\|\right)(2\mathbf{u} + \mathbf{u}^2)\sqrt{n} + \mu_{\mathsf{INV}}(n)\cdot\mathbf{u}\cdot\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|(1+\mathbf{u})^2\sqrt{n} \\ &\leq \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|\right)4\sqrt{n}\mathbf{u}\end{aligned}$$

where we use $\mathbf{u} < 1$ in the last line. □

In what remains of this section we will repeatedly use the following simple calculus fact.

Lemma A.2.2. Let $x, y > 0$, then

$$\log(x+y) \leq \log(x) + \frac{y}{x} \quad \text{and} \quad \lg(x+y) \leq \lg(x) + \frac{1}{\log 2}\frac{y}{x}.$$

Proof. This follows directly from the concavity of the logarithm. □

Lemma A.2.3 (Restatement of Lemma 3.4.15)**.** Let $1/800 > t > 0$ and $1/2 > c > 0$ be given. Then for

$$j \geq \lg(1/t) + 2\lg\lg(1/t) + \lg\lg(1/c) + 1.62,$$

we have

$$\frac{(1-t)^{2^j}}{t^{2j}} < c.$$

Proof of Lemma 3.4.15. An exact solution for j can be written in terms of the *Lambert W-function*; see [33] for further discussion and a useful series expansion. For our purposes, it is simpler to derive the necessary quantitative bound from scratch.

Immediately from the assumption $t < 1/800$, we have $j > \log(1/t) \geq 9$.

First let us solve the case $c = 1/2$. We will prove the contrapositive, so assume

$$\frac{(1-t)^{2^j}}{t^{2j}} \geq 1/2.$$

Then taking log on both sides, we have

$$2j\log(1/t) + 1 \geq -2^j\log(1-t) \geq 2^j t.$$

Taking lg and applying Lemma A.2.2, we obtain

$$1 + \lg j + \lg\log(1/t) + \frac{1}{\log 2}\frac{1}{2j\log(1/t)} \geq j + \lg t.$$

Since $t < 1/800$ we have $\frac{1}{\log 2}\frac{1}{2j\log(1/t)} < 0.01$, so

$$j - \lg j \le \lg(1/t) + \lg\log(1/t) + 1.01 \le \lg(1/t) + \lg\lg(1/t) + 0.49 =: K.$$

But since $j \ge 9$, we have $j - \lg j \ge 0.64j$, so

$$j \le \frac{1}{0.64}(j - \lg j) \le \frac{1}{0.64}K$$

which implies

$$j \le K + \lg j \le K + \lg(1.57K) = K + \lg K + 0.65.$$

Note $K \le 1.39\lg(1/t)$, because $K - \lg(1/t) = \lg\lg(1/t) + 0.49 \le 0.39\lg(1/t)$ for $t \le 1/800$. Thus

$$\lg K \le \lg(1.39\lg(1/t)) \le \lg\lg(1/t) + 0.48,$$

so for the case $c = 1/2$ we conclude the proof of the contrapositive of the lemma:

$$\begin{aligned} j &\le K + \lg K + 0.65 \\ &\le \lg(1/t) + \lg\lg(1/t) + 0.49 + (\lg\lg(1/t) + 0.48) + 0.65 \\ &= \lg(1/t) + 2\lg\lg(1/t) + 1.62. \end{aligned}$$

For the general case, once $(1-t)^{2^j}/t^{2j} \le 1/2$, consider the effect of incrementing j on the left hand side. This has the effect of squaring and then multiplying by t^{2j-2}, which makes it even smaller. At most $\lg\lg(1/c)$ increments are required to bring the left hand side down to c, since $(1/2)^{2^{\lg\lg(1/c)}} = c$. This gives the value of j stated in the lemma, as desired. □

Lemma A.2.4 (Restatement of Lemma 3.4.18)**.** If

$$N = \lceil \lg(1/s) + 3\lg\lg(1/s) + \lg\lg(1/(\beta\varepsilon_0)) + 7.59 \rceil,$$

then

$$N \ge \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\varepsilon_0)) + 1.62.$$

Proof of Lemma 3.4.18. We aim to provide a slightly cleaner sufficient condition on N than the current condition

$$N \ge \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\varepsilon_0)) + 1.62.$$

Repeatedly using Lemma A.2.2, as well as the cruder fact $\lg\lg(ab) \le \lg\lg a + \lg\lg b$ provided $a, b \ge 4$, we have

$$\begin{aligned} \lg\lg(16/(\beta s^2\varepsilon_0)) &\le \lg\lg(16/s^2) + \lg\lg(1/(\beta\varepsilon_0)) \\ &= 1 + \lg(3 + \lg(1/s)) + \lg\lg(1/(\beta\varepsilon_0)) \\ &\le 1 + \lg\lg(1/s) + \frac{3}{\log 2\lg(1/s)} + \lg\lg(1/(\beta\varepsilon_0)) \\ &\le \lg\lg(1/s) + \lg\lg(1/(\beta\varepsilon_0)) + 1.66 \end{aligned}$$

where in the last line we use the assumption $s < 1/100$. Similarly,

$$\begin{aligned}\lg(8/s) + 2\lg\lg(8/s) &\le 3 + \lg(1/s) + 2\lg(3 + \lg(1/s)) \\ &\le 3 + \lg(1/s) + 2\left(\lg\lg(1/s) + \frac{3}{\log 2 \lg(1/s)}\right) \\ &\le \lg(1/s) + 2\lg\lg(2/s) + 4.31\end{aligned}$$

Thus, a sufficient condition is

$$N = \lceil \lg(1/s) + 3\lg\lg(1/s) + \lg\lg(1/(\beta\varepsilon_0)) + 7.59 \rceil.$$

□

A.3 Analysis of SPLIT

Although it has many potential uses in its own right, the purpose of the approximate matrix sign function in our algorithm is to split the spectrum of a matrix into two roughly equal pieces, so that approximately diagonalizing A may be recursively reduced to two sub-problems of smaller size.

First, we need a lemma ensuring that a shattered pseudospectrum can be bisected by a grid line with at least $n/5$ eigenvalues on each side.

Lemma A.3.1. Let A have ε-pseudospectrum shattered with respect to some grid g. Then there exists a horizontal or vertical grid line of g partitioning g into two grids $\mathsf{g}_\pm$, each containing at least $\min\{n/5, 1\}$ eigenvalues.

Proof. We will view g as a $s_1 \times s_2$ array of squares. Write $r_1, r_2, ..., r_{s_1}$ for the number of eigenvalues in each row of the grid. Either there exists $1 \le i < s_2$ such that $r_1 + \cdots + r_i \ge n/5$ and $r_{i+1} + \cdots + r_{s_1} \ge n/5$—in which case we can bisect at the grid line dividing the ith from $(i+1)$st rows—or there exists some i for which $r_i \ge 3/5$. In the latter case, we can always find a vertical grid line so that at least $n/5$ of the eigenvalues in the ith row are on each of the left and right sides. Finally, if $n \le 5$, we may trivially pick a grid line to bisect along so that both sides contain at least one eigenvalue. □

Proof of Theorem 3.5.2. We'll prove first that SPLIT has the advertised guarantees. The main observation is that, given any matrix A, we can determine how many eigenvalues are on either side of any horizontal or vertical line by approximating the matrix sign function. In particular, $\operatorname{Tr}\operatorname{sgn}(A - h) = n_+ - n_-$, where $n_\pm$ are the eigenvalue counts on either side of the line $\Re z = h$.

Running SGN to a final accuracy of β,

$$\begin{aligned}|\operatorname{Tr}, \mathsf{SGN}(A - h) + e_4 - \operatorname{Tr}, \operatorname{sgn}(A - h)| &\le |\operatorname{Tr}, \mathsf{SGN}(A - h) - \operatorname{Tr}, \operatorname{sgn}(A - h)| + |e_4| \\ &\le n\big(\|\mathsf{SGN}(A - h) - \operatorname{sgn}(A - h)\| + \|\mathsf{SGN}(A - h)\|\mathbf{u}\big) \\ &\le n\big(\beta + \beta\mathbf{u} + \|\operatorname{sgn}(A - h)\|\mathbf{u}\big).\end{aligned}$$

SPLIT

Input: Matrix $A \in \mathbb{C}^{n\times n}$, grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$ pseudospectral guarantee ϵ, and a desired accuracy ν.
Requires: $\Lambda_\epsilon(A)$ is shattered with respect to g, and $\beta \leq 0.05/n$.
Algorithm: $(\widetilde{P}_+, \widetilde{P}_-, \mathsf{g}_+, \mathsf{g}_-) = \mathsf{SPLIT}(A, \mathsf{g}, \epsilon, \beta)$

1. $h \leftarrow \Re z_0 + \omega s_1/2$
2. $M \leftarrow A - h + E_2$
3. $\alpha_0 \leftarrow 1 - \frac{\epsilon}{2\,\mathrm{diam}(\mathsf{g})^2}$
4. $\phi \leftarrow \mathrm{round}\,(\mathrm{Tr}\,\mathsf{SGN}(M, \epsilon/4, \alpha_0, \beta) + e_4)$
5. If $|\phi| < \min(3n/5, n-1)$
   a) $\mathsf{g}_- = \mathsf{grid}(z_0, \omega, s_1/2, s_2)$
   b) $z_0 \leftarrow z_0 + h$
   c) $\mathsf{g}_+ = \mathsf{grid}(z_0, \omega, s_1/2, s_2)$
   d) $(\widetilde{P}_+, \widetilde{P}_-) = \frac{1}{2}(1 \pm \mathsf{SGN}(A - h, \beta))$
6. Else, execute a binary search over horizontal grid-line shifts h until $\mathsf{SGN}(A - h, \epsilon/4, \alpha_0, \beta)$, at which point output $\mathsf{g}_\pm$, the subgrids on either side of the shift h, and set $\widetilde{P}_\pm \leftarrow \frac{1}{2}\,(\mathsf{SGN}(h - A, \epsilon/4, \alpha_0, \beta))$.
7. If this fails, set $A \leftarrow iA$, and execute a binary search among vertical shifts from the original grid.

Output: Sub-grids $\mathsf{g}_\pm$, approximate spectral projectors $\widetilde{P}_\pm$, and ranks $n_\pm$.
Ensures: There exist true spectral projectors $P_\pm$ satisfying (i) $P_+ + P_- = 1$, (ii)$\mathrm{rank}(P_\pm) = n_\pm \geq n/5$, (iii) $\|P_\pm - \widetilde{P}_\pm\| \leq \beta$, and (iv) $P_\pm$ are the spectral projectors onto the interiors of $\mathsf{g}_\pm$.

Since we can form $\mathrm{sgn}(A - h)$ by integrating around the boundary of the portions of g on either side of the line $\Re z = h$, the fact that $\Lambda_\epsilon(A)$ is shattered means that

$$\|\mathrm{sgn}(A - h)\| \leq \frac{1}{2\pi}\frac{1}{\epsilon}\omega(2s_1 + 4s_2) \leq 8/\epsilon;$$

in the last inequality we have used that g has side lengths of at most 8. Since we have run SGN to accuracy β, this gives a total additive error of $n(\beta + \beta\mathbf{u} + 8\mathbf{u}/\epsilon)$ in computing the trace. If $\beta \leq 0.1/n$ and $\mathbf{u} \leq \epsilon/100n$, then this error will strictly less than 0.5 and we can

round $\mathrm{Tr}, \mathsf{SGN}(A - h)$ to the nearest real integer. Horizontal bisections work similarly, with $iA - h$ instead.

Since we need only modify the diagonal entries of A when creating M, we incur a *diagonal* error matrix E_3 of norm at most $\mathbf{u}|\max_i A_{i,i} - h|$ when creating M. Using $|A_{i,i}| \leq \|A\| \leq 4$ and $|h| \leq 4$, the fact that $\mathbf{u} \leq \epsilon/100n \leq \epsilon/16$ ensures that the $\epsilon/2$-pseudospectrum of M will still be shattered with respect to a translation of the original grid g that includes a segment of the imaginary axis. Using Lemma 3.4.10 and the fact that $\mathrm{diam}(\mathsf{g})^2 = 128$, we can safely call SGN with parameters $\epsilon_0 = \epsilon/4$ and

$$\alpha_0 = 1 - \frac{\epsilon}{256}.$$

Plugging these in to the Theorem 3.4.9 ($\epsilon < 1/2$ so $1 - \alpha_0 \leq 1/100$, and $\beta \leq 0.05/n \leq 1/12$ so the hypotheses are satisfied) for final accuracy β a sufficient number of iterations is

$$N_{\mathsf{SPLIT}} := \lg \frac{256}{\epsilon} + 3 \lg \lg \frac{256}{\epsilon} + \lg \lg \frac{4}{\beta\epsilon} + 7.59.$$

In the course of these binary searches, we make at most $\lg s_1 s_2$ calls to SGN at accuracy β. These require at most

$$\lg s_1 s_2 \, T_{\mathsf{SGN}} \left(n, \epsilon/2, 1 - \frac{\epsilon}{2\,\mathrm{diam}(\mathsf{g})^2}, \beta \right)$$

arithmetic operations. In addition, creating M and computing the trace of the approximate sign function cost us $O(n \lg s_1 s_2)$ scalar addition operations. We are assuming that g has side lengths at most 8, so $\lg s_1 s_2 \leq 12 \lg 1/\omega(\mathsf{g})$. Combining all of this with the runtime analysis and machine precision of SGN appearing in Theorem 3.4.9, we obtain

$$T_{\mathsf{SPLIT}}(n, \mathsf{g}, \epsilon, \beta) \leq 12 \lg \frac{1}{\omega(\mathsf{g})} \cdot N_{\mathsf{SPLIT}} \cdot \left(T_{\mathsf{INV}}(n, \mathbf{u}) + O(n^2) \right).$$

□

A.4 Analysis of DEFLATE

The algorithm $\mathsf{DEFLATE}$, defined in Section 3.5, can be viewed as a small variation of the randomized rank revealing algorithm introduced in [47] and revisited subsequently in [10]. Following these works, we will call this algorithm RURV.

Roughly speaking, in finite arithmetic, RURV takes a matrix A with $\sigma_r(A)/\sigma_{r+1}(A) \gg 1$, for some $1 \leq r \leq n-1$, and finds nearly unitary matrices U, V and an upper triangular matrix R such that $URV \approx A$. Crucially, R has the block decomposition

$$R = \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix}, \tag{A.4}$$

RURV

Input: Matrix $A \in \mathbb{C}^{n \times n}$
Algorithm: RURV(A)

1. $G \leftarrow n \times n$ complex Ginibre matrix $+E_1$
2. $(V, R) \leftarrow \mathsf{QR}(G)$
3. $B \leftarrow AV^* + E_3$
4. $(U, R) \leftarrow \mathsf{QR}(B)$

Output: A pair of matrices (U, R).
Ensures: $\|R_{22}\| \leq \frac{\sqrt{r(n-r)}}{\theta}\sigma_{r+1}(A)$ with probability $1 - \theta^2$, for every $1 \leq r \leq n-1$ and $\theta > 0$, where R_{22} is the $(n-r) \times (n-r)$ lower-right corner of R.

where $R_{11} \in \mathbb{C}^{r \times r}$ has smallest singular value close to $\sigma_r(A)$, and R_{22} has largest singular value roughly $\sigma_{r+1}(A)$. We will use and analyze the following implementation of RURV.

As discussed in Section 3.5, we hope to use DEFLATE to approximate the range of a projector P with rank $r < n$, given an approximation $\widetilde{P}$ close to P in operator norm. We will show that from the output of RURV$(\widetilde{P})$ we can obtain a good approximation to such a subspace. More specifically, under certain conditions, if $(U, R) = \mathsf{RURV}(\widetilde{P})$, then the first r columns of U carry all the information we need. For a formal statement see Proposition A.4.12 and Proposition A.4.18 below.

Since it may be of broader use, we will work in somewhat greater generality, and define the subroutine DEFLATE which receives a matrix A and an integer r and returns a matrix $S \in \mathbb{C}^{n \times r}$ with nearly orthonormal columns. Intuitively, if A is diagonalizable, then under the guarantee that r is the smallest integer k such that $\sigma_k(A)/\sigma_{k+1}(A) \gg 1$, the columns of the output S span a space close (in some sense) to the span of the top r eigenvectors of A. Our implementation of DEFLATE is as follows.

Throughout this section we use rurv$(\cdot)$ and deflate$(\cdot, \cdot)$ to denote the exact arithmetic versions of RURV and DEFLATE respectively. In Subsection A.4.1 we present a random matrix result that will be needed in the analysis of DEFLATE. In Subsection A.4.3 we state the properties of RURV that will be needed. Finally in Subsections A.4.4 and A.4.5 we prove the main guarantees of deflate and DEFLATE, respectively, that are used throughout this chapter.

A.4.1 Smallest Singular Value of the Corner of a Haar Unitary

We recall the defining property of the Haar measure on the unitary group:

DEFLATE

Input: Matrix $\widetilde{A} \in \mathbb{C}^{n\times n}$ and parameter $r \leq n$
Requires: $1/3 \leq \|A\|$, and $\|\widetilde{A} - A\| \leq \beta$ for some $A \in \mathbb{C}^{n\times n}$ with $\mathrm{rank}(A) = \mathrm{rank}(A^2) = r$, as well as $\beta \leq 1/4 \leq \|\widetilde{A}\|$ and $1 \leq \mu_{\mathsf{MM}}(n), \mu_{\mathsf{QR}}(n), c_{\mathsf{N}}$.

Algorithm: $\widetilde{S} = \mathsf{DEFLATE}(A, r)$.

1. $(U, R) \leftarrow \mathsf{RURV}(A)$
2. $\widetilde{S} \leftarrow$ first r columns of U.
3. Output $\widetilde{S}$

Output: Matrix $S \in \mathbb{C}^{n\times r}$.
Ensures: There exists a matrix $S \in \mathbb{C}^{n\times k}$ whose orthogonal columns span range(A), such that $\|\widetilde{S} - S\| \leq \eta$, with probability at least $1 - \frac{(20n)^3\sqrt{\beta}}{\eta^2\sigma_r(A)}$.

Definition A.4.1. A random $n \times n$ unitary matrix V is *Haar-distributed* if, for any other unitary matrix W, VW and WV are Haar-distributed as well.

For short, we will often refer to such a matrix as a *Haar unitary.*

Let $n > r$ be positive integers. In what follows we will consider an $n \times n$ Haar unitary matrix V and denote by X its upper-left $r \times r$ corner. The purpose of the present subsection is to derive a tail bound for the random variable $\sigma_n(X)$. We begin with the well-known fact that we can always reduce our analysis to the case when $r \leq n/2$.

Observation A.4.2. Let $n > r > 0$ and $V \in \mathbb{C}^{n\times n}$ be a unitary matrix and denote by V_{11} and V_{22} its upper-left $r \times r$ corner and its lower-right $(n-r) \times (n-r)$ corner respectively. If $r \geq n/2$, then $2r - n$ of the singular values of V_{22} are equal to 1, while the remaining $n - r$ are equal to those of V_{11}.

Proposition A.4.3 (σ_n of a submatrix of a Haar unitary)**.** Let $n > r > 0$ and let V be an $n \times n$ Haar unitary. Let X be the upper left $r \times r$ corner of V. Then, for all $\theta > 0$

$$\mathbb{P}\left[\frac{1}{\sigma_n(X)} \leq \frac{1}{\theta}\right] = (1 - \theta^2)^{r(n-r)}. \tag{A.5}$$

In particular, for every $\theta > 0$ we have

$$\mathbb{P}\left[\frac{1}{\sigma_n(X)} \leq \frac{\sqrt{r(n-r)}}{\theta}\right] \geq 1 - \theta^2. \tag{A.6}$$

This exact formula for the CDF of the smallest singular value of X is remarkably simple, and we have not seen it anywhere in the literature. It is an immediate consequence of

substantially more general results of Dumitriu [51], from which one can extract and simplify the density of $\sigma_n(X)$. We will begin by introducing the relevant pieces of [51], deferring the final proof until the end of this subsection.

Some of the formulas presented here are written in terms of the generalized hypergeometric function which we denote by ${}_2F_1^{\beta}(a, b; c; (x_1, \ldots, x_m))$. For our application it is sufficient to know that

$$ {}_2F_1^{\beta}(0, b; c, (x_1, \ldots, x_m)) = 1, \tag{A.7}$$

whenever $c > 0$ and ${}_2F_1$ is well defined. The above equation can be derived directly from the definition of ${}_2F_1^{\beta}$ (see Definition 13.1.1 in [57] or Definition 2.2 in [51]).

The generic results in [51] concern the β-*Jacobi* random matrices, which we have no cause here to define in full. Of particular use to us will be [51, Theorem 3.1], which expresses the density of the smallest singular value of such a matrix in terms of the generalized hypergeometric function:

Theorem A.4.4 ([51])**.** The probability distribution of the smallest eigenvalue λ of the β-Jacobi ensembles of parameters a, b and size m is given by

$$\begin{aligned} f_{\lambda_n}(\lambda) := C_{\beta,a,b,m} \lambda^{\frac{\beta}{2}(a+1)-1} (1-\lambda)^{\frac{\beta}{2}m(b+m)-1} {}_2F_1^{2/\beta} \\ \times \left(1 - \frac{\beta(a+1)}{2}, \frac{\beta(b+m-1)}{2}; \frac{\beta(b+2m-1)}{2} + 1; (1-\lambda)^{m-1}\right), \end{aligned} \tag{A.8}$$

for some normalizing constant $C_{\beta,a,b,m}$.

For a particular choice of parameters, the above theorem can be applied to describe the the distribution of $\sigma_n^2(X)$. The connection between singular values of corners of Haar unitary matrices and β-Jacobi ensembles is the content of [54, Theorem 1.5], which we rephrase below to match our context.

Theorem A.4.5 ([54])**.** Let V be an $n \times n$ Haar unitary matrix and let $r \leq \frac{n}{2}$. Let X be the $r \times r$ upper-left corner of V. Then, the eigenvalues of XX^* distribute as the eigenvalues of a $\beta-$Jacobi matrix of size r with parameters $\beta = 2, a = 0$ and $b = n - 2r$.

In view of the above result, Theorem A.4.4 gives a formula for the density of $\sigma_n^2(X)$.

Corollary A.4.6 (Density of $\sigma_n^2(X)$)**.** Let V be an $n \times n$ Haar unitary and X be its upper-left $r \times r$ corner with $r < n$, then $\sigma_n^2(X)$ has the following density

$$f_{\sigma_n^2}(x) := \begin{cases} r(n-r)\,(1-x)^{r(n-r)-1} & \text{if} \quad 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{A.9}$$

Proof. If $r > n/2$, since we care only about the smallest singular value of X, we can use Observation A.4.2 to analyse the $(n-r) \times (n-r)$ lower right corner of V instead. Hence,

we can assume that $r \leq n/2$. Now, substitute $\beta = 2, a = 0, b = n - 2r, m = r$ in Theorem A.4.4 and observe that in this case

$$f_{\lambda_n}(x) = C(1-x)^{r(n-r)-1}{}_2F_1^1(0, n-r-1; n; (1-x)^{r-1}) = C(1-x)^{r(n-r)-1} \tag{A.10}$$

where the last equality follows from (A.7). Using the relation between the distribution of $\sigma_n^2(X)$ and the distribution of the minimum eigenvalue of the respective β-Jacobi ensemble described in Theorem A.4.5 we have $f_{\sigma_n^2}(x) = f_{\lambda_n}(x)$. By integrating on $[0,1]$ the right side of (A.10) we find $C = r(n-r)$. □

Proof of Proposition A.4.3. From (A.9) we have that

$$\mathbb{P}\left[\sigma_n^2(X) \leq \theta\right] = r(n-r)\int_0^\theta (1-x)^{r(n-r)-1}dx = 1 - (1-\theta)^{r(n-r)},$$

from where (A.5) follows. To prove (A.6) note that $g(t) := (1-t)^{r(n-r)}$ is convex in $[0,1]$, and hence $g(t) \geq g(0) + tg'(0)$ for every $t \in [0,1]$. □

A.4.2 Sampling Haar Unitaries in Finite Precision

It is a well-known fact that Haar unitary matrices can be numerically generated from complex Ginibre matrices. We refer the reader to [53, Section 4.6] and [92] for a detailed discussion. In this subsection we carefully analyze this process in finite arithmetic.

The following fact (see [92, Section 5]) is the starting point of our discussion.

Lemma A.4.7 (Haar from Ginibre)**.** Let G_n be a complex $n \times n$ Ginibre matrix and $U, R \in \mathbb{C}^{n\times n}$ be defined implicitly, as a function of G_n, by the equation $G_n = UR$ and the constraints that U is unitary and R is upper-triangular with nonnegative diagonal entries[1]. Then, U is Haar distributed in the unitary group.

The above lemma suggests that $\mathsf{QR}(\cdot)$ can be used to generate random matrices that are approximately Haar unitaries. While doing this, one should keep in mind that when working with finite arithmetic, the matrix $\widetilde{G_n}$ passed to QR is not exactly Ginibre-distributed, and the algorithm QR itself incurs round-off errors.

Following the discussion in Section 3.2.2 we can assume that we have access to a random matrix $\widetilde{G}_n$, with

$$\widetilde{G}_n = G_n + E,$$

where G_n is a complex $n \times n$ Ginibre matrix and $E \in \mathbb{C}^{n\times n}$ is an adversarial perturbation whose entries are bounded by $\frac{1}{\sqrt{n}}c_{\mathsf{N}}\mathbf{u}$. Hence, we have $\|E\| \leq \|E\|_F \leq \sqrt{n}c_{\mathsf{N}}\mathbf{u}$.

In what follows we use $\mathrm{QR}(\cdot)$ to denote the exact arithmetic version of $\mathsf{QR}(\cdot)$. Furthermore, we assume that for any $A \in \mathbb{C}^{n\times n}$, $\mathrm{QR}(A)$ returns a pair (U, R) with the property that R

[1] G_n is almost surely invertible and under this event U and R are uniquely determined by these conditions.

has nonnegative entries on the diagonal. Since we want to compare $\mathrm{QR}(G_n)$ with $\mathsf{QR}(\widetilde{G}_n)$ it is necessary to have a bound on the condition number of the QR decomposition. For this, we cite the following consequence of a result of Sun [118, Theorem 1.6]:

Lemma A.4.8 (Condition number for the QR decomposition [118])**.** Let $A, E \in \mathbb{C}^{n\times n}$ with A invertible. Furthermore assume that $\|E\|\|A^{-1}\| \leq \frac{1}{2}$. If $(U, R) = \mathrm{QR}(A)$ and $(\widetilde{U}, \widetilde{R}) = \mathrm{QR}(A + E)$, then

$$\|\widetilde{U} - U\|_F \leq 4\|A^{-1}\|\|E\|_F.$$

We are now ready to prove the main result of this subsection. As in the other sections devoted to finite arithmetic analysis, we will assume that $\mathbf{u}$ is small compared to $\mu_{\mathsf{QR}}(n)$; precisely, let us assume that

$$\mathbf{u}\mu_{\mathsf{QR}}(n) \leq 1. \tag{A.11}$$

Proposition A.4.9 (Guarantees for finite-arithmetic Haar unitary matrices)**.** Suppose that QR satisfies the assumptions in Definition 3.2.4 and that it is designed to output upper triangular matrices with nonnegative entries on the diagonal[2]. If $(V, R) = \mathsf{QR}(\widetilde{G}_n)$, then there is a Haar unitary matrix U and a random matrix E such that $\widetilde{V} = U + E$. Moreover, for every $1 > \alpha > 0$ and $t > 2\sqrt{2} + 1$ we have

$$\mathbb{P}\left[\|E\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u}\right] \geq 1 - 2e\alpha^2 - 2e^{-t^2 n}.$$

Proof. From our Gaussian sampling assumption, $\widetilde{G}_n = G_n + E$ where $\|E\| \leq \sqrt{n}c_{\mathsf{N}}\mathbf{u}$. Also, by the assumptions on QR from Definition 3.2.4, there are matrices $\widetilde{\widetilde{G}}_n$ and $\widetilde{V}$ such that $(\widetilde{V}, R) = \mathrm{QR}(\widetilde{\widetilde{G}}_n)$, and

$$\begin{aligned}\|\widetilde{V} - V\| &< \mu_{\mathsf{QR}}(n)\mathbf{u}\\ \|\widetilde{\widetilde{G}}_n - \widetilde{G}_n\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}\|\widetilde{G}_n\| &\leq \mu_{\mathsf{QR}}(n)\mathbf{u}\left(\|G_n\| + \sqrt{n}c_{\mathsf{N}}\mathbf{u}\right).\end{aligned}$$

The latter inequality implies, using (A.11), that

$$\|\widetilde{\widetilde{G}}_n - G_n\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}\left(\|G_n\| + \sqrt{n}c_{\mathsf{N}}\mathbf{u}\right) + \sqrt{n}c_{\mathsf{N}}\mathbf{u} \leq \mu_{\mathsf{QR}}(n)\mathbf{u}\|G_n\| + 2\sqrt{n}c_{\mathsf{N}}\mathbf{u}. \tag{A.12}$$

Let $(U, R') := \mathrm{QR}(G_n)$. From Lemma A.4.7 we know that U is Haar distributed on the unitary group, so using (A.12) and Lemma A.4.8, and the fact that $\|M\| \leq \|M\|_F \leq \sqrt{n}\|M\|$ for any $n \times n$ matrix M, we know that

$$\|U-V\|-\mu_{\mathsf{QR}}(n)\mathbf{u} \leq \|U-V\|-\|\widetilde{V}-V\| \leq \|U-\widetilde{V}\| \leq 4\sqrt{n}c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u}\|G_n\|\|G_n^{-1}\|+10nc_{\mathsf{N}}\mathbf{u}\|G_n^{-1}\|. \tag{A.13}$$

[2]Any algorithm that yields the QR decomposition can be modified in a stable way to satisfy this last condition at the cost of $O^*(n\log(1/\mathbf{u}))$ operations

Now, from $\|G_n^{-1}\| = 1/\sigma_n(G_n)$ and from Theorem 3.3.1 we have that

$$P\left[\|G_n^{-1}\| \geq \frac{n}{\alpha}\right] \leq (\sqrt{2e}\alpha)^2 = 2e\alpha^2.$$

On the other hand, from Lemma 2.2 of [11] we have $P\left[\|G_n\| > 2\sqrt{2} + t\right] \leq e^{-nt^2}$. Hence, under the events $\|G_n^{-1}\| \leq \frac{n}{\alpha}$ and $\|G_n\| \leq 2\sqrt{2} + t$, inequality (A.13) yields

$$\|U - V\| \leq \frac{4n^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}} \mu_{\mathsf{QR}}(n) \mathbf{u} \left(2\sqrt{2} + t + 1\right) + \frac{10n^2}{\alpha} c_{\mathsf{N}} \mathbf{u}.$$

Finally, if $t > 2\sqrt{2} + 1$ we can exchange the term $2\sqrt{2} + t + 1$ for $2t$ in the bound. Then, using a union bound we obtain the advertised guarantee. □

A.4.3 Preliminaries of RURV

Let $A \in \mathbb{C}^{n\times n}$ and $(U, R) = \mathrm{rurv}(A)$. As will become clear later, in order to analyze $\mathsf{DEFLATE}(A, r)$ it is of fundamental importance to bound the quantity $\|R_{22}\|$, where R_{22} is the lower-right $(n-r)\times(n-r)$ block of R. To this end, it will suffice to use Corollary A.4.11 below, which is the complex analog to the upper bound given in equation (4) of [10, Theorem 5.1]. Actually, Corollary A.4.11 is a direct consequence of Lemma 4.1 in the aforementioned paper and Proposition A.4.3 proved above. We elaborate below.

Lemma A.4.10 ([10])**.** Let $n > r > 0$, $A \in \mathbb{C}^{n\times n}$ and $A = P\Sigma Q^*$ be its singular value decomposition. Let $(U, R) = \mathrm{rurv}(A)$, R_{22} be the lower right $(n-r)\times(n-r)$ corner of R, and V be such that $A = URV$. Then, if $X = Q^*V^*$,

$$\|R_{22}\| \leq \frac{\sigma_{r+1}(A)}{\sigma_n(X_{11})},$$

where X_{11} is the upper left $r \times r$ block of X.

This lemma reduces the problem to obtaining a lower bound on $\sigma_n(X_{11})$. But, since V is a Haar unitary matrix by construction and $X = Q^*V$ with Q^* unitary, we have that X is distributed as a Haar unitary. Combining Lemma A.4.10 and Proposition A.4.3 gives the following result.

Corollary A.4.11. Let $n > r > 0$, $A \in \mathbb{C}^{n\times n}$, $(U, R) = \mathrm{rurv}(A)$ and R_{22} be the lower right $(n-r)\times(n-r)$ corner of R. Then for any $\theta > 0$

$$\mathbb{P}\left[\|R_{22}\| \leq \frac{\sqrt{r(n-r)}}{\theta}\sigma_{r+1}(A)\right] \geq 1 - \theta^2.$$

A.4.4 Exact Arithmetic Analysis of DEFLATE

It is a standard consequence of the properties of the QR decomposition that if A is a matrix of rank r, then almost surely deflate(A, r) is a $n \times r$ matrix with orthonormal columns that span the range of A. As a warm-up let's recall this argument.

Let $(U, R) = \text{rurv}(A)$ and V be the unitary matrix used by the algorithm to produce this output. Since we are working in exact arithmetic, V is a Haar unitary matrix, and hence it is almost surely invertible. Therefore, with probability 1 we have $\text{rank}(AV^*) = r$, so if $UR = AV^*$ we will have $R_{22} = 0$ and $R_{11} \in \mathbb{C}^{r\times r}$, where R_{11} and R_{22} are as in (A.4). Writing

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}$$

for the block decomposition of U with $U_{11} \in \mathbb{C}^{r\times r}$, note that

$$AV^* = UR = \begin{pmatrix} U_{11}R_{11} & U_{11}R_{12} + U_{12}R_{22} \\ U_{21}R_{11} & U_{21}R_{12} + U_{22}R_{22} \end{pmatrix}. \tag{A.14}$$

On the other hand, almost surely the first r columns of AV^* span the range of A. Using the right side of equation (A.14) we see that this subspace also coincides with the span of the first r columns of U, since R_{11} is invertible.

We will now prove a robust version of the above observation for a large class of matrices, namely those A for which $\text{rank}(A) = \text{rank}(A^2)$.[3] We make this precise below and defer the proof to the end of the subsection.

Proposition A.4.12 (Main guarantee for deflate)**.** Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that $\|A - \widetilde{A}\| \leq \beta$ and $\text{rank}(A) = \text{rank}(A^2) = r$. Denote $S := \text{deflate}(\widetilde{A}, r)$ and $T := \text{deflate}(A, r)$. Then, for any $\theta \in (0, 1)$, with probability $1 - \theta^2$ there exists a unitary $W \in \mathbb{C}^{r\times r}$ such that

$$\|S - TW^*\| \leq \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\beta}{\theta}}. \tag{A.15}$$

Remark A.4.13 (The projector case)**.** In the case in which the matrix A of Proposition A.4.12 is a (not necessarily orthogonal) projector, $T^*AT = I_r$, and the σ_r term in the denominator of (A.15) becomes a 1.

We begin by recalling a result about the stability of singular values which will be important throughout this section. This fact is a consequence of Weyl's inequalities; see for example [75, Theorem 3.3.16] .

Lemma A.4.14 (Stability of singular values)**.** Let $X, E \in \mathbb{C}^{n\times n}$. Then, for any $k = 1, \ldots, n$ we have

$$|\sigma_k(X + E) - \sigma_k(X)| \leq \|E\|.$$

[3] For example, diagonalizable matrices satisfy this criterion.

We now show that the orthogonal projection $P := \text{deflate}(\widetilde{A}, r)\text{deflate}(\widetilde{A}, r)^*$ is close to a projection onto the range of A, in the sense that $PA \approx A$.

Lemma A.4.15. Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that $\text{rank}(A) = r$ and $\|A - \widetilde{A}\| \leq \beta$. Let $(U, R) := \text{rurv}(\widetilde{A})$ and $S := \text{deflate}(\widetilde{A}, r)$. Then, almost surely

$$\|(SS^* - I_n)A\| \leq \|R_{22}\| + \beta, \tag{A.16}$$

where R_{22} is the lower right $(n-r)\times(n-r)$ block of R.

Proof. We will begin by showing that $\|(SS^* - I_n)\widetilde{A}\|$ is small. Let V be the unitary matrix that was used to generate (U, R). As $\text{deflate}(\cdot,\cdot)$ outputs the first r columns of U, we have the block decomposition $U = \begin{pmatrix} S & U' \end{pmatrix}$, where $S \in \mathbb{C}^{n\times r}$ and $U' \in \mathbb{C}^{n\times(n-r)}$.

On the other hand we have $\widetilde{A} = URV$, so

$$(SS^* - I_n)\widetilde{A} = (SS^* - I)\begin{pmatrix} S & U' \end{pmatrix} RV = \begin{pmatrix} 0 & -U' \end{pmatrix} RV = \begin{pmatrix} 0 & -U'R_{2,2} \end{pmatrix} V.$$

Since $\|U'\| = \|V\| = 1$ from the above equation we get $\|(SS^* - I_n)\widetilde{A}\| \leq \|R_{22}\|$. Now we can conclude that

$$\|(SS^* - I_n)A\| \leq \|(SS^* - I_n)\widetilde{A}\| + \|(SS^* - I_n)(A - \widetilde{A})\| \leq \|R_{22}\| + \beta.$$

□

The inequality (A.16) can be applied to quantify the distance between the ranges of $\text{deflate}(\widetilde{A}, r)$ and $\text{deflate}(A, r)$ in terms of $\|R_{22}\|$, as the following result shows.

Lemma A.4.16 (Bound in terms of $\|R_{22}\|$)**.** Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that $\text{rank}(A) = \text{rank}(A^2) = r$ and $\|A - \widetilde{A}\| \leq \beta$. Denote by $(U, R) := \text{rurv}(\widetilde{A})$, $S := \text{deflate}(\widetilde{A}, r)$ and $T := \text{deflate}(A, r)$. Then, almost surely there exists a unitary $W \in \mathbb{C}^{r\times r}$ such that

$$\|S - TW^*\| \leq 2\sqrt{\frac{\|R_{22}\| + \beta}{\sigma_r(T^*AT)}}, \tag{A.17}$$

where R_{22} is the lower right $(n-r)\times(n-r)$ block of R.

Proof. From Lemma A.4.15 we know that almost surely $\|(SS^* - I_n)A\| \leq \|R_{22}\| + \beta$. We will use this to show that $\|T^*SS^*T - I_r\|$ is small, which can be interpreted as S^*T being close to unitary. First note that

$$\|T^*SS^*T - I_r\| = \sup_{w\in\mathbb{C}^r, \|w\|=1} \|T^*(SS^* - I_r)Tw\| = \sup_{w\in\text{range}(A), \|w\|=1} \|T^*(SS^* - I_r)w\|. \tag{A.18}$$

Now, since $\text{rank}(A) = \text{rank}(A^2)$, if $w \in \text{range}(A)$ then $w = Av$ for some $v \in \text{range}(A)$. So by the Courant-Fischer formula

$$\frac{\|w\|}{\|v\|} = \frac{\|Av\|}{\|v\|} \geq \inf_{u\in\text{range}(A)} \frac{\|Au\|}{\|u\|} = \sigma_r(T^*AT).$$

We can then revisit (A.18) and get

$$\sup_{w\in\text{range}(A),\|w\|=1}\|T^*(SS^*-I_r)w\| = \sup_{v\in\text{range}(A),\|v\|\leq 1}\frac{\|T^*(SS^*-I_r)Av\|}{\sigma_r(T^*AT)} \leq \frac{\|T^*(SS^*-I_r)AT\|}{\sigma_r(T^*AT)}. \tag{A.19}$$

On the other hand $\|T^*(SS^*-I_r)AT\| \leq \|(SS^*-I_r)A\| \leq \|R_{22}\|+\beta$, so combining this fact with (A.18) and (A.19) we obtain

$$\|T^*SS^*T - I_r\| \leq \frac{\|R_{22}\|+\beta}{\sigma_r(T^*AT)}.$$

Now define $X := S^*T$, $\beta' := \frac{\|R_{22}\|+\beta}{\sigma_r(T^*AT)}$ and let $X = W|X|$ be the polar decomposition of X. Observe that

$$\||X| - I_r\| \leq \sigma_1(X) - 1 \leq |\sigma_1(X)^2 - 1| = \|X^*X - I_r\| \leq \beta'.$$

Thus $\|S^*T - W\| = \|X - W\| = \|(|X| - I_n)W\| \leq \beta'$. Finally note that

$$\begin{aligned}\|S - TW^*\|^2 &= \|(S^* - WT^*)(S - TW^*)\| \\ &= \|2I_r - S^*TW^* - WT^*S\| \\ &= \|2I_r - S^*T(T^*S + W^* - T^*S) - (S^*T + W - S^*T)T^*S\| \\ &\leq 2\|I_r - S^*TT^*S\| + \|S^*T(W^* - T^*S)\| + \|(W - S^*T)T^*S\| \leq 4\beta',\end{aligned}$$

which concludes the proof. □

Note that so far our results have been deterministic. The possibility of failure of the guarantee given in Proposition A.4.12 comes from the non-deterministic bound on $\|R_{22}\|$.

Proof of Proposition A.4.12. From Lemma A.4.14 we have $\sigma_{r+1}(\widetilde{A}) \leq \beta$. Now combine Lemma A.4.16 with Corollary A.4.11. □

A.4.5 Finite Arithmetic Analysis of DEFLATE

In what follows we will have an approximation $\widetilde{A}$ of a matrix A of rank r with the guarantee that $\|A - \widetilde{A}\| \leq \beta$.

For the sake of readability we will not present optimal bounds for the error induced by roundoff, and we will assume that

$$4\|A\| \cdot \max\{c_{\mathsf{N}}\mu_{\mathsf{MM}}(n)\mathbf{u}, c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u}\} \leq \beta \leq \frac{1}{4} \leq \|A\| \quad \text{and} \quad 1 \leq \min\{\mu_{\mathsf{MM}}(n), \mu_{\mathsf{QR}}(n), c_{\mathsf{N}}\}. \tag{A.20}$$

We begin by analyzing the subroutine RURV in finite arithmetic. This was done in [47, Lemma 5.4]. Here we make the constants arising from this analysis explicit and take into consideration that Haar unitary matrices cannot be exactly generated in finite arithmetic.

Lemma A.4.17 (RURV analysis)**.** Assume that QR and MM satisfy the guarantees in Definitions 3.2.2 and 3.2.4. Also suppose that the assumptions in (A.20) hold. Then, if $(U, R) := \mathsf{RURV}(A)$ and V is the matrix used to produce such output, there are unitary matrices $\widetilde{U}, \widetilde{V}$ and a matrix $\widetilde{A}$ such that $\widetilde{A} = \widetilde{U}R\widetilde{V}$ and the following guarantees hold:

1. $\|U - \widetilde{U}\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}$.
2. $\widetilde{V}$ is Haar distributed in the unitary group.
3. For every $1 > \alpha > 0$ and $t > 2\sqrt{2} + 1$, the event:

$$\|\widetilde{V} - V\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}} \mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha}\mathbf{u}$$

$$\text{and} \quad \|A - \widetilde{A}\| < \|A\| \left(\frac{9tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}} \mu_{\mathsf{QR}}(n)\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u} \right) \tag{A.21}$$

occurs with probability at least $1 - 2e\alpha^2 - 2e^{-t^2 n}$.

Proof. By definition $V = \mathsf{QR}(\widetilde{G_n})$ with $\widetilde{G}_n = G_n + E$, where G_n is an $n \times n$ Ginibre matrix and $\|E\| \leq \sqrt{n}\mathbf{u}$. A direct application of the guarantees on each step yields the following:

1. From Proposition A.4.9, we know that there is a Haar unitary $\widetilde{V}$ and a random matrix E_0, such that $V = \widetilde{V} + E_0$ and

$$\mathbb{P}\left[\|E_0\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}} \mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u} \right] \geq 1 - 2e\alpha^2 - 2e^{-t^2 n}. \tag{A.22}$$

2. If $B := \mathsf{MM}(A, V^*) = AV^* + E_1$, then from the guarantees for MM we have $\|E_1\| \leq \|A\|\|V\|\mu_{\mathsf{MM}}(n)\mathbf{u}$. Now from the guarantees for QR we know that V is $\mu_{\mathsf{QR}}(n)\mathbf{u}$ away from a unitary, and hence

$$\|V\|\mu_{\mathsf{MM}}(n)\mathbf{u} \leq (1 + \mu_{\mathsf{QR}}(n)\mathbf{u})\mu_{\mathsf{MM}}(n)\mathbf{u} \leq \frac{5}{4}\mu_{\mathsf{MM}}(n)\mathbf{u}$$

where the last inequality follows from the assumptions in (A.20). This translates into

$$\|B\| \leq \|A\|\|V\| + \|E_1\| \leq (1 + \mu_{\mathsf{QR}}(n)\mathbf{u})\|A\| + \|E_1\| \leq \frac{5}{4}\|A\| + \|E_1\|.$$

Putting the above together and using (A.20) again, we get

$$\|E_1\| \leq \frac{5}{4}\|A\|\mu_{\mathsf{MM}}(n)\mathbf{u} \quad \text{and} \quad B \leq \frac{5}{4}\|A\|(1 + \mu_{\mathsf{MM}}(n)\mathbf{u}) < 2\|A\|. \tag{A.23}$$

3. Let $(U, R) = \mathsf{QR}(B)$. Then there is a unitary $\widetilde{U}$ and a matrix $\widetilde{B}$ such that $U = \widetilde{U} + E_2$, $B = \widetilde{B} + E_3$, and $\widetilde{B} = \widetilde{U}R$, with error bounds $\|E_2\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}$ and $\|E_3\| \leq \|B\|\mu_{\mathsf{QR}}(n)\mathbf{u}$. Using (A.23) we obtain

$$\|E_3\| \leq \|B\|\mu_{\mathsf{QR}}(n)\mathbf{u} < 2\|A\|\mu_{\mathsf{QR}}(n)\mathbf{u}. \tag{A.24}$$

4. Finally, define $\widetilde{A} := \widetilde{B}\widetilde{V}$. Note that $\widetilde{A} = \widetilde{U}R\widetilde{V}$ and

$$\widetilde{A} = \widetilde{B}\widetilde{V} = (B - E_3)\widetilde{V} = (AV^* + E_1 - E_3)\widetilde{V} = (A(\widetilde{V} + E_0)^* + E_1 - E_3)\widetilde{V} = A + (AE_0^* + E_1 - E_3)\widetilde{V},$$

which translates into

$$\|A - \widetilde{A}\| \leq \|A\|\|E_0\| + \|E_1\| + \|E_3\|.$$

Hence, on the event described in the left side of (A.22), we have

$$\|A - \widetilde{A}\| \leq \|A\| \left(\frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}} \mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u} + \frac{5}{4}\mu_{\mathsf{MM}}(n)\mathbf{u} + 2\mu_{\mathsf{QR}}(n)\mathbf{u} \right),$$

and using some crude bounds, the above inequality yields the advertised bound.

□

We can now prove a finite arithmetic version of Proposition A.4.12.

Proposition A.4.18 (Main guarantee for DEFLATE)**.** Let $n > r$ be positive integers, and let $\beta, \theta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that $\|A - \widetilde{A}\| \leq \beta$ and $\operatorname{rank}(A) = \operatorname{rank}(A^2) = r$. Let $S := \mathsf{DEFLATE}(\widetilde{A}, r)$ and $T := \operatorname{deflate}(A, r)$. If QR and MM satisfy the guarantees in Definitions 3.2.2 and 3.2.4, and (A.20) holds, then, for every $t > 2\sqrt{2}+1$ there exist a unitary $W \in \mathbb{C}^{r\times r}$ such that

$$\|S - TW^*\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u} + 12\sqrt{\frac{tn^2\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\beta}{\theta^2}}, \tag{A.25}$$

with probability at least $1 - 7\theta^2 - 2e^{-t^2 n}$.

Proof. Let $(U, R) = \mathsf{RURV}(\widetilde{A})$. From Lemma A.4.17 we know that there exist $\widetilde{U}, \widetilde{\widetilde{A}} \in \mathbb{C}^{n\times n}$, such that $\|U - \widetilde{U}\|$ and $\|\widetilde{A} - \widetilde{\widetilde{A}}\|$ are small, and $(\widetilde{U}, R) = \operatorname{rurv}(\widetilde{\widetilde{A}})$ for the respective realization of an exact Haar unitary matrix. Then, from $\|\widetilde{A}\| \leq \|A\| + \beta$ and (A.21), for every $1 > \alpha > 0$ and $t > 2\sqrt{2} + 1$ we have

$$\left\|A - \widetilde{\widetilde{A}}\right\| \leq \left\|\widetilde{\widetilde{A}} - \widetilde{A}\right\| + \|\widetilde{A} - A\| \leq (\|A\| + \beta) \left(\frac{9tn^{\frac{3}{2}}}{\alpha} \mu_{\mathsf{QR}}(n) c_{\mathsf{N}}\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u} \right) + \beta, \tag{A.26}$$

with probability $1 - 2e\alpha^2 - 2e^{-t^2 n}$.

Now, from (A.20) we have $\mathbf{u} \leq \beta \leq \frac{1}{4}$ and $c_{\mathsf{N}}\|A\|\mu\mathbf{u} \leq \beta$ for $\mu = \mu_{\mathsf{QR}}(n), \mu_{\mathsf{MM}}(n)$, so we can bound the respective terms in (A.26) by β:

$$(\|A\| + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u}\right) + \beta$$

$$\leq (1+\beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha}\beta + 2\beta + \frac{10n^2}{\alpha}\beta\right) + \beta \leq \frac{(12t+16)}{\alpha} n^2\beta, \tag{A.27}$$

where the last crude bound uses $1 \leq n^{\frac{3}{2}} \leq n^2, 1+\beta \leq \frac{5}{4}$ and $t > 2$.

Observe that $\widetilde{S} = \mathrm{deflate}(\widetilde{\widetilde{A}}, r)$ is the matrix formed by the first r columns of $\widetilde{U}$, and that by Proposition A.4.12 we know that for every $\theta > 0$, with probability $1 - \theta^2$ there exists a unitary W such that

$$\|\widetilde{S} - TW^*\| \leq \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\left\|A - \widetilde{\widetilde{A}}\right\|}{\theta}}. \tag{A.28}$$

On the other hand, S is the matrix formed by the first r columns of U. Hence

$$\|S - \widetilde{S}\| \leq \|U - \widetilde{U}\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}.$$

Putting the above together we get that under this event

$$\|S - TW^*\| \leq \|S - \widetilde{S}\| + \|\widetilde{S} - TW^*\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u} + \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\left\|A - \widetilde{\widetilde{A}}\right\|}{\theta}}. \tag{A.29}$$

Now, taking $\alpha = \theta$, we note that both events in (A.26) and (A.28) happen with probability at least $1 - (2e+1)\theta^2 - 2e^{-t^2 n}$. The result follows from replacing the constant $2e+1$ with 7, using $t > 2\sqrt{2} + 1$ and replacing $8(12t+16)$ with $144t$, and combining the inequalities (A.26), (A.27) and (A.29). □

We end by proving Theorem 3.5.3, the guarantees on DEFLATE that we will use when analyzing the main algorithm.

Proof of Theorem 3.5.3. As Remark A.4.13 points out, in the context of this theorem we are passing to DEFLATE an approximate projector $\widetilde{P}$, and the above result simplifies. Using this fact, as well as the upper bound $r(n-r) \leq n^2/4$, we get that

$$\|S - TW^*\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{12\sqrt{tn^3\beta}}{\theta}.$$

with probability at least $1 - 7\theta^2 - 2e^{-t^2 n}$ for every $t > 2\sqrt{2}$. If our desired quality of approximation is $\|S - TW^*\| = \eta$, then some basic algebra gives the success probability as at least

$$1 - 1008\frac{n^3 t\beta}{(\eta - \mu_{\mathsf{QR}}(n)\mathbf{u})^2} - 2e^{-t^2 n}.$$

Since $\beta \leq 1/4$, we can safely set $t = \sqrt{2/\beta}$, giving

$$1 - 1426\frac{n^3\sqrt{\beta}}{(\eta - \mu_{\mathsf{QR}}(n)\mathbf{u})^2} - 2e^{-2n/\beta}.$$

To simplify even further, we'd like to use the upper bound $2e^{-2n/\beta} \leq \frac{n^3\sqrt{\beta}}{(\eta - \mu_{\mathsf{QR}}(n)\mathbf{u})^2}$. These two terms have opposite curvature in β on the interval $(0, 1)$, and are equal at zero, so it suffices to check that the inequality holds when $\beta = 1$. The terms only become closer by setting $n = 1$ everywhere except in the argument of $\mu_{\mathsf{QR}}(\cdot)$, so we need only check that

$$\frac{2}{e^2} \leq \frac{1}{(\eta - \mu_{\mathsf{QR}}(n)\mathbf{u})^2}.$$

Under our assumptions $\eta, \mu_{\mathsf{QR}}(n)\mathbf{u} \leq 1$, the right hand side is greater than one, and the left hand less. Thus we can make the replacement, use $\mathbf{u} \leq \frac{\eta}{2\mu_{QR}(n)}$, and round for readability to a success probability of no worse than

$$1 - 6000\frac{n^3\sqrt{\beta}}{\eta^2};$$

the constant here is certainly not optimal.

Finally, for the running time, we need to sample n^2 complex Gaussians, perform two QR decompositions, and one matrix multiplication; this gives the total bit operations as

$$T_{\mathsf{DEFLATE}}(n) = n^2 T_{\mathsf{N}} + 2T_{\mathsf{QR}}(n) + T_{\mathsf{MM}}(n).$$

□

Remark A.4.19. Note that the exact same proof of Theorem 3.5.3 goes through in the more general case where the matrix in question is not necessarily a projection, but any matrix close to a rank-deficient matrix A. In this case an extra $\sigma_r(T^*AT)$ term appears in the probability of success (see the guarantee given in the box for the Algorithm DEFLATE that appears in this appendix).