Title
Rigorous Guarantees for Randomized Diagonalization Algorithms

Author
Garza Vargas, Jorge

Publication Date
2022

Peer reviewed|Thesis/dissertation

Rigorous Guarantees for Randomized Diagonalization Algorithms

by

Jorge Garza-Vargas

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:

Associate Professor Nikhil Srivastava, Co-chair
Professor Dan-Virgil Voiculescu, Co-chair
Professor James Demmel
Associate Professor Shirshendu Ganguly

Fall 2022

Rigorous Guarantees for Randomized Diagonalization Algorithms

Abstract

Rigorous Guarantees for Randomized Diagonalization Algorithms

by

Jorge Garza-Vargas

Doctor of Philosophy in Mathematics

University of California, Berkeley

Associate Professor Nikhil Srivastava, Co-chair

Professor Dan-Virgil Voiculescu, Co-chair

We rigorously analyze numerical methods for (approximately) computing the eigenvalues and eigenvectors of a matrix. Our main results are the following:

- *Spectral bisection.* We show that a backward-stable solution to the eigenvalue problem can be obtained by a randomized algorithm in nearly matrix-multiplication time on any input. More specifically, we show that for every $\delta > 0$, there is a randomized version of the spectral bisection algorithm that, on any input $A \in \mathbb{C}^{n\times n}$, with probability $1 - O(1/n)$, finds matrices $V, D \in \mathbb{C}^{n\times n}$ with V invertible and D diagonal such that $\|A - VDV^{-1}\| \leq \delta\|A\|$, using $O(T_{\mathsf{MM}}(n)\log^2(n/\delta))$ arithmetic operations, in finite arithmetic with $O(\log^4(n/\delta)\log n)$ bits of precision. 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.

- *The Hessenberg shifted QR algorithm.* We introduce a randomized shifting strategy for the Hessenberg QR algorithm, for which we prove (in the finite precision arithmetic model) rapid global convergence. It follows from our results, that for any $\delta, \phi > 0$, a randomized version of the shifted Hessenberg QR algorithm can be used to compute, on any input $A \in \mathbb{C}^{n\times n}$, with probability at least $1 - \phi$, matrices $U, T \in \mathbb{C}^{n\times n}$ with U unitary and T upper triangular, such that $\|A - UTU^*\| \leq \delta\|A\|$, using $O(n^3 \log(n/(\delta\phi))^2 \log\log(n/(\delta\phi)))$ arithmetic operations performed using $O(\log^2(n/(\delta\phi))\log\log(n/(\delta\phi)))$ bits of precision. This provides the first rapidly globally convergent shifting strategy in the nonsymmetric case, which was an open problem even under the assumption of exact arithmetic.

- *The Lanczos algorithm.* We analyze the Lanczos algorithm where the initial vector u is sampled uniformly at random from the unit sphere $\mathbb{S}^{n-1}$, and show that when run on a Hermitian input $A \in \mathbb{C}^{n\times n}$ for $c_1 \log n$ iterations (where c_1 depends on some coarse global property of the spectrum of A but not on n) its outputs (both the Jacobi coefficients and the Ritz values) are exponentially concentrated around their medians. Our techniques also allow us to understand the location of the output in the regime of $k = O(\log n)$ iterations, and in particular we use them to show that the Lanczos algorithm can fail to identify outlying eigenvalues when run for less than $c_2 \log n$, where c_2 depends again on the same global property of the spectrum of A used to determine c_1.

A key ingredient in showing the first two results is a smoothed analysis of the eigenvector condition number and minimum eigenvalue gap of a matrix. In this direction we show a general result that might be of independent interest in random matrix theory: if one adds independent (absolutely continuous with respect to the Lebesgue measure) random variables of small variance to the entries of an arbitrary matrix $A \in \mathbb{C}^{n\times n}$, with high probability, the resulting matrix A' will have (relatively) stable eigenvenvalues and eigenvectors.

To my parents, Rubi and Jorge, for their unconditional support.

Contents

Acknowledgments

First of all I'd like to thank my advisors, Dan and Nikhil, who have been essential to my development during the last five years. I thank Dan for sharing his wisdom and for his continuous encouragement and endorsement from the very beginning. I thank Nikhil for his guidance and his infinite generosity, both in terms of time and ideas, which greatly shaped the way I think about research and the way I appreciate and approach problems.

I also want to thank Archit, Jess, and Nikhil for the wonderful collaboration that we established, without which the results in this thesis would be non-existent.

Thanks also to Benson, Daniel, Octavio, and Satyaki, who were also my collaborators during my PhD, and from whom I learned a great deal while having a great time. More in general, I'd like to thank my friends and family, for all of their support, encouragement, and fun times that made this process more enjoyable.

Finally, I'd like to thank the members of my thesis committee for their careful reading and useful suggestions which greatly improved this dissertation. Thanks also to Steve for agreeing to be in my qual committee, and to Shirshendu and Marc for their attention and guidance earlier in my PhD.

Chapter 1

Introduction

Given its unquestionable relevance, for the last many decades, several mathematicians, scientists and engineers have undertaken the challenging task of designing efficient and accurate algorithms for numerically diagonalizing a matrix (computing its eigenvalues and eigenvectors), which is known as *solving the eigenvalue problem*.

On the practical side, nowadays most medium-scale eigenvalue problems can routinely be solved by personal computers to a satisfactory level of accuracy. However, industrial and scientific applications oftentimes require dealing with large-scale problems that are beyond the reach of classical numerical methods.

On the other hand, despite Wilkinson's monumental book [175], other excellent more modern references [55, 125, 157], and important research developments throughout the years [68, 69, 99, 174, 52, 23, 37, 167, 53, 3], the theoretical aspects of the eigenvalue problem are still underdeveloped and lag behind their practical counterparts. In fact, to this day, many foundational questions in the area remain open: to give an example, no formal guarantees of success (on arbitrary inputs) have been yet given for some of the most widely used algorithms (which occasionally happen to fail even in small and medium-sized problems).

In this thesis we develop mathematical tools that allow for the rigorous study of some of the most popular diagonalization algorithms, with the goal of bridging theory and practice. The hope is that, on the one hand, furthering theory can eventually lead to informing practice in problems that are currently out of its reach, and on the other hand, a deepened understanding of theory can simplify and improve (both conceptually and in terms of performance and implementation) already existing algorithms.

The following algorithms will be discussed here: spectral bisection, shifted QR iteration, and Lanczos iteration. For each of these algorithms our analysis will have to address the relevance and effect of randomness in the performance of the algorithm, the consequences of round-off error on the accuracy of the algorithm (we will not discuss this for the Lanczos algorithm, where inputs are Hermitian so their eigenvalues are robust to perturbations), and how the aforementioned phenomena feed into the dynamical system defined by each of these procedures. We refer the reader to Sections 1.2-1.5 for a detailed overview of the main results in this dissertation.

From the theoretical perspective, the scope of the work presented here is ambitious in that it seeks to provide (for the first time in some instances) close to optimal rigorous algorithmic guarantees. On the other hand, when it comes to practical matters, it remains humble and does not seek to be a prescription for practitioners. More experimentation and theory development would be required to hope to improve the existing (extremely sophisticated) libraries for diagonalization, which throughout the years have been fine-tuned to enhance their performance and efficiency.

Bibliographical note. The content presented in Chapter 2 (*Spectral Stability Under Random Perturbations*) is joint work with Jess Banks, Archit Kulkarni, and Nikhil Srivastava [10]. The content in Chapter 3 (*Spectral Bisection*) is joint work with the same set of authors in [11]. The content in Chapter 4 (*Hessenberg QR Algorithm*) is joint work with Jess Banks and Nikhil Srivastava, and appears in the sequence of papers [12, 13, 14]. The content in Chapter 5 (*The Lanczos Algorithm Under Few Iterations*) is joint work with Archit Kulkarni [71].

Reader's guide for Chapter 1. The purpose of this chapter is to give a condensed overview of the main results, motivation, proof techniques, and mathematical tools appearing in this dissertation. The main theorems from each of the subsequent chapters will be discussed here, in the corresponding section, where the main ingredients of the corresponding proofs will be outlined. For example, Chapter 2 will be surveyed in Section 1.2, which starts by presenting the general random matrix phenomena in question, followed by the main formal statements that will be proven in Chapter 2, and a rough sketch of the ideas used in the proofs of these theorems; the standard arguments will be laid out for the convenience of the reader and the main technical complications that will be addressed in Chapter 2 will be highlighted.

Sections 1.2 through Section 1.5 have been written in a mostly independent way with the purpose of allowing the reader to skip those that are not of their interest.

Section 1.1 contains all of the technical preliminaries that are needed for the remainder of Chapter 1, which I believe are also indispensable tools for any theoretician that seeks to work in this general research direction.

Finally, in Section 1.6 I have tried to encapsulate the common themes appearing in this work and to compare the different main mechanisms that are used throughout the dissertation.

In Section 1.7 a short list of open problems is given for the interested reader.

1.1 Preliminaries

1.1.1 Notation, Definitions, and Conventions

Models of Computation

The three models of computation that are most commonly used when analyzing numerical algorithms are: 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) also known as *floating point*. The first one is used when the quantities handled by the algorithm in question are numerically stable and it is believed that errors coming from machine computations do not have a significant impact on the output; in particular, this is the model that will be used in this dissertation when discussing the Lanczos algorithm. The last two models are more realistic, and do take into account the unavoidable limitations of computers when handling real numbers. Therefore, these two models can be used to obtain actual Boolean complexity bounds [1], and yield concrete accuracy bounds for the output. Among these two, the finite precision arithmetic model is the one that is used more often, since it is in line with most implementations, and this will be our model of choice when analyzing the shifted QR and spectral bisection algorithms[2].

Finite precision arithmetic model. We will assume that 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 $n \in \mathbb{N}$ and desired accuracy $\delta > 0$. 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

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

Thus, the outcomes of all operations are adversarially noisy due to roundoff. More generally, throughout this dissertation, 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 so that the errors are small enough to still guarantee success.

We will also assume that the bit lengths of numbers stored 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 $O^*(\lg(1/\mathbf{u}))$, the running time of standard floating point arithmetic, where the $*$ suppresses $\lg\lg(1/\mathbf{u})$ factors. We will state all running times

[1] That is, describing the number of Boolean operations that are required to execute a particular algorithm.

[2] More specifically, we will use the standard floating point axioms from [85].

in terms of arithmetic operations accompanied by the required number of bits of precision, which thereby immediately imply bit complexity bounds. To simplify exposition, as it is customary, we will ignore overflow and underflow effects.

Linear Algebra

Throughout this dissertation we will use $\mathbb{C}^{n\times n}$ to refer to the space of $n \times n$ matrix with complex entries. All vector norms will be ℓ^2-norms, and given $A \in \mathbb{C}^{n\times n}$ we will use $\|A\|$ to denote the ℓ^2-operator norm of A, and $\|A\|_F$ to denote the Frobenius norm of A.

Oftentimes, if $z \in \mathbb{C}, A \in \mathbb{C}^{n\times n}$, and I_n is the identity matrix, to lighten notation, we will write $z - A$ instead of $zI_n - A$.

Eigenpairs. Given a matrix $A \in \mathbb{C}^{n\times n}$, a vector $v \in \mathbb{C}^n$, and a complex number $\lambda \in \mathbb{C}$ we will say that (v, λ) is a *right (resp. left) eigenpair* of A if $Av = \lambda v$ (resp. $v^*A = \lambda v^*$). When $\|v\| = 1$ we will say that v is normalized, or that (v, λ) is a normalized eigenpair. The collection of eigenvalues of A will be referred to as the spectrum of A and will be denoted by $\mathrm{Spec}(A)$.

Diagonalization and spectral decomposition. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be diagonalizable if one can find a collection of (say) right eigenpairs $(\lambda_1, v_1), \ldots, (\lambda_n, v_n)$ for which $\{v_1, \ldots, v_n\}$ forms a basis for $\mathbb{C}^n$. This is equivalent to the existence of an invertible matrix $V \in \mathbb{C}^{n\times n}$ and a diagonal matrix $D \in \mathbb{C}^{n\times n}$ such that

$$A = VDV^{-1}.$$

Such representation is referred to as a *diagonalization of* A. Observe that diagonalizations are unique up to multiplying by a scalars the columns of V, and that given a diagonalization, the columns of V, say $\{v_1, \ldots, v_n\}$ and the rows of V^{-1}, say $\{w_1, \ldots, w_n\}$, are respectively left and right eigenvectors for A, from which one can form the *spectral decomposition*

$$A = \sum_{i=1}^{n} \lambda_i v_i w_i^*, \tag{1.1}$$

where the λ_i are the eigenvalues of A (which also happen to be the diagonal entries of D). Note that with the above setup the v_i and w_i form a dual basis, that is

$$w_i^* v_i = 1 \quad \text{and} \quad w_i^* v_j = 0,$$

for $i \neq j$. In the particular case in which A is normal, if one normalizes the v_i one gets that $v_i = w_i$ and that V is a unitary matrix.

Spectral Projectors. For an arbitrary $A \in \mathbb{C}^{n\times n}$ we say that a matrix P is a *spectral projector* for A if $AP = PA$ and $P^2 = P$. For instance, when A is diagonalizable, each

of the terms $v_i w_i^*$ appearing in the spectral decomposition (1.1) is a spectral projector, as $Av_i w_i^* = \lambda_i v_i w_i^* = v_i w_i^* A$ and $w_i^* v_i = 1$.

More in general, if P is a spectral projector of A, observe that since it commutes with A, its range agrees exactly with an invariant subspace of A, which will be generated by the generalized eigenvectors associated to a subset of $\mathrm{Spec}(A)$. Conversely, every subset of $\mathrm{Spec}(A)$ has an associated spectral projector of A.

Schur Form. Any matrix $A \in \mathbb{C}^{n\times n}$ admits a *Schur decomposition*, that is, it can be written in the form

$$A = UTU^*,$$

where $U \in \mathbb{C}^{n\times n}$ is unitary and $T \in \mathbb{C}^{n\times n}$ is an upper triangular matrix. In this case, the eigenvalues of A will be the diagonal elements of T, and when A is diagonalizable, the eigenvector matrix V can be recovered from U and T.

Labeling of Eigenvalues and Singular Values. Given a matrix $A \in \mathbb{C}^{n\times n}$ we will denote (and order) its singular values by

$$\sigma_1(A) \geq \cdots \geq \sigma_n(A).$$

Sometimes we will use the notation $\sigma_{\min}(A) := \sigma_n(A)$ to emphasize that we are referring to the smallest singular value of A. Similarly, when A is Hermitian, its eigenvalues will be real and will be denoted by

$$\lambda_1(A) \geq \cdots \geq \lambda_n(A),$$

and we will occasionally use $\lambda_{\min}(A) := \lambda_n(A)$ and $\lambda_{\max}(A) = \lambda_1(A)$. In this case, $v_i(A)$ will denote the normalized eigenvector of A corresponding to $\lambda_i(A)$.

Ramdom matrix theory. Throughout this dissertation we will assume that the reader is familiar with basic concepts and arguments in probability and random matrix theory. For the most part, our probabilistic arguments will not utilize anything beyond the notions of independence, conditioning, concentration and anti-concentration. Similarly, the arguments specific to random matrix theory appearing here are not technical and are common in the non-asymptotic literature.

Finally, we introduce some important notation. If n is a positive integer, we will use $[n]$ to denote the set $\{1, \ldots, n\}$. For $z \in \mathbb{C}$ and $r > 0$, we will use $D(z, r)$ to denote the open disk in $\mathbb{C}$ with center at z and radius r.

A *normalized complex (resp. real) Ginibre matrix* G_n is an $n \times n$ random matrix with independent complex (resp. real) Gaussian entries of variance $\frac{1}{n}$. This normalization is chosen so that $\mathbb{E}\|G_n\| = O(1)$ as n goes to infinity.

1.1.2 The Eigenvalue Problem

When working within the finite precision arithmetic model one can only hope to obtain approximations of the eigenvalues and eigenvectors of a matrix, up to a desired accuracy

$\delta > 0$. In the numerical analysis literature there are two standard notions of approximation, which we define below.

Forward approximation. Let $A \in \mathbb{C}^{n\times n}$, $\delta > 0$ and (λ, v) be an eigenpair (left or right) of A. We say that (λ', v') is δ-forward approximation of (λ, v) if

$$|\lambda - \lambda'| \leq \delta \quad \text{and} \quad \|v - v'\| \leq \delta. \tag{1.2}$$

It makes sense to use the above definition in contexts where the exact solution is meaningful; e.g. the matrix is of theoretical/mathematical origin (as opposed to being constructed e.g. from approximate data), and unstable (in the entries) quantities such as eigenvalue multiplicity, or the eigenvalues themselves, can have a significant meaning.

The δ in (1.2) quantifies in absolute terms the accuracy of the approximations λ' and v'. However, when talking about eigenvalues sometimes we will be interested in notions of relative accuracy. In these cases we will compare the size of the error $|\lambda - \lambda'|$ to $\|A\|$,[3] or alternatively use the notion of forward approximation defined above and assume (without loss of generality) that $\|A\| \leq 1$.

Backward Approximation. Using the same notation as above, we say that (λ', v') is a δ-backward eigenpair of A if it is the true eigenpair of some matrix A' with

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

Again, sometimes we will be interested in relative (as opposed to absolute) accuracy, in which case we will either assume that $\|A\| \leq 1$ and pursue finding backward approximations in the above sense, or not assume anything about the norm of A and look for an exact eigenpair of a matrix A' satisfying $\|A - A'\| \leq \delta\|A\|$.

The notion of backward approximation 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 when storing the matrix would destroy this exactness).

Approximate Diagonalization. Given an input matrix $A \in \mathbb{C}^{n\times n}$, 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. To be precise, we will be aiming to solve the following (backward) version of the eigenvalue problem.

Definition 1.1.1 (The backward eigenvalue problem)**.** Given an input matrix $A \in \mathbb{C}^{n\times n}$ and an accuracy parameter $\delta > 0$, find an invertible matrix $V \in \mathbb{C}^{n\times n}$ and a diagonal matrix $D \in \mathbb{C}^{n\times n}$ such that

$$\|A - VDV^{-1}\| \leq \delta\|A\|. \tag{1.3}$$

[3] Oftentimes, when discussing relative accuracy the size of the error $|\lambda - \lambda'|$ is compared to $|\lambda|$. We will not be using this notion here, since we will only discuss accuracy in the context of algorithms that yield full eigendecompositions, and their performance will depend on global properties of the spectrum (and not on the value of individual eigenvalues).

Remark 1.1.2 (Finding the approximate Schur form)**.** When the eigenvectors of A are known to be highly unstable (in a sense that will be made precise below), instead of solving the above formulation of the eigenvalue problem, it makes more sense to compute the *approximate Schur form* of the matrix A. That is, finding a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a triangular matrix $T \in \mathbb{C}^{n\times n}$ such that

$$\|A - UTU^*\| \leq \delta\|A\|.$$

Note that a backward solution to the eigenvalue problem in the above sense is also a forward solution but for a worse (sometimes meaningless) accuracy parameter. The two notions of approximation can be related precisely if one has quantitative knowledge of the level of sensitivity of the eigenvalues and eigenvectors of the matrix. To this end it is useful to define the following condition numbers for the eigenvalue problem.

Condition numbers. For diagonalizable A, the *eigenvector condition number* of A, denoted $\kappa_V(A)$, is defined as:

$$\kappa_V(A) := \inf_{V:A=VDV^{-1}} \|V\|\|V^{-1}\|.$$

The *eigenvalue gaps* and *minimum eigenvalue gap* of A are defined as

$$\mathrm{gap}_i(A) := \min_{j:j\neq i} |\lambda_i - \lambda_j| \quad \text{and} \quad \mathrm{gap}(A) := \min_{i\in[n]} \mathrm{gap}_i(A),$$

where the λ_i are the eigenvalues of A (with multiplicity). And we define the *condition number of the eigenproblem* to be[4]:

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

It then follows from the proposition below (whose proof we defer to Section 1.1.3) that a δ-backward approximate solution to the eigenproblem is a $6n\kappa_{\mathrm{eig}}(A)\delta$-forward approximate solution.[5]

Proposition 1.1.3. *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 normalized eigenpairs of A, A' with distinct eigenvalues, and $\delta < \frac{\mathrm{gap}(A)}{8\kappa_V(A)}$, then*

$$\|v_i' - v_i\| \leq 2n\kappa_{\mathrm{eig}}(A)\delta \quad \text{and} \quad |\lambda_i' - \lambda_i| \leq \kappa_V(A)\delta \leq 2\kappa_{\mathrm{eig}}(A)\delta \quad \forall i = 1, \ldots, n, \tag{1.4}$$

after possibly multiplying the v_i by phases.

Note that $\kappa_{\mathrm{eig}} = \infty$ if and only if A has a multiple eigenvalue; in this case, a relation like (1.4) is not possible since different infinitesimal changes to A can produce macroscopically different eigenpairs.

[4]This quantity is inspired by but not identical to the "reciprocal of the distance to ill-posedness" for the eigenproblem considered by Demmel [57], to which it is polynomially related. See also [162] for another natural definition of eigenvector condition number similar in spirit to that of Demmel.

[5]In fact, it can be shown that $\kappa_{\mathrm{eig}}(A)$ is related by a $\mathrm{poly}(n)$ factor to the smallest constant for which (1.4) holds for all sufficiently small $\delta > 0$.

Accuracy vs. Precision and the Meaning of Stability. The gold standard of *backward stability* in numerical analysis defines algorithms as backward stable, if when ran in a computer with machine precision $\mathbf{u}$, on an input of size n, they output a δ-backward approximation to the correct answer where

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

for some moderate constant c. That is, according to the gold standard, for an algorithm to be considered backward stable, its accuracy should have only a few (depending on the size of the problem) bits less than the precision used. The relaxed notion of "logarithmic stability" introduced in [54] expands the notion of stable algorithms, to those satisfying

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

for some constant c, where κ is an appropriate condition number for the problem in question. In this work we relax this notion even further, and consider an algorithm to be backward stable if $\mathbf{u}$ is polylogarithmic in $1/\delta$ and n (for some moderate polynomial, without any reference to a condition number). The reason for this, is that if a fast algorithm (i.e. one that executes few arithmetic operations) is backward stable in the latter sense, then it can be implemented in a computer with high enough precision to guarantee its accuracy, and theoretically (if hardware considerations are ignored) the precision required will still be low enough so that this implementation remains fast (since the number of bit operations that will be performed will be nearly as many, i.e. a polylogarithmic multiple, as the arithmetic operations used by the algorithm).

1.1.3 Tools for Analyzing Nonnormal Matrices

The Holomorphic Functional Calculus

When a matrix $A \in \mathbb{C}^{n\times n}$ is nonnormal the notion of spectral measure is no longer available. However, one can still make sense of functions applied to A. When $A = VDV^{-1}$ is diagonalizable, for any function $f : \mathrm{Spec}(A) \to \mathbb{C}$ we can simply define $f(A) := Vf(D)V^{-1}$, where $f(D)$ means applying f to the diagonal entries of D (note that $f(A)$ is independent of the choice of V). And when A is non-diagonalizable we can recur to the well-known holomorphic functional calculus.

Proposition 1.1.4 (Holomorphic Functional Calculus)**.** *Let $A \in \mathbb{C}^{n\times n}$, $\mathcal{B} \supset \mathrm{Spec}(A)$ be an open neighborhood of its spectrum (not necessarily connected), and $\Gamma_1, ..., \Gamma_k$ be simple closed positively oriented rectifiable curves in $\mathcal{B}$ whose interiors together contain all of* $\mathrm{Spec}(A)$*. Then if f is holomorphic on $\mathcal{B}$, the definition*

$$f(A) := \frac{1}{2\pi i}\sum_{j=1}^{k}\oint_{\Gamma_j} f(z)(z-A)^{-1}\mathrm{d}z \tag{1.5}$$

determines a unital algebra homomorphism *from the space of holomorphic functions on $\mathcal{B}$ to the algebra of $n \times n$ matrices. Moreover, when $A = VDV^{-1}$ is diagonalizable, this definition of $f(A)$ coincides with the definition $f(A) = Vf(D)V^{-1}$.*

One can use the holomorphic functional calculus to extract spectral projectors of A by taking a function f whose restriction to $\mathrm{Spec}(A)$ is an indicator function (equivalently, by integrating $f(z)(z-A)^{-1}$ for the constant function $f \equiv 1$ over a single contour that encloses a subset of the relevant eigenvalues of $\mathrm{Spec}(A)$). In particular, if λ_i is a simple eigenvalue of A with associated spectral projector P_i, and 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 not hard to show that

$$P_i = \frac{1}{2\pi i} \oint_{\Gamma_i} (z-A)^{-1} \mathrm{d}z, \tag{1.6}$$

by taking the Jordan normal form of the the *resolvent* $(z-A)^{-1}$ and applying Cauchy's integral formula. More in general, if we integrate the resolvent over some region of the complex plane bounded by a simple closed positively oriented rectifiable curve Γ, we will be obtaining the spectral projector onto the invariant subspace spanned by those eigenvectors whose eigenvalues lie inside Γ.

Finally, we emphasize that the holomorphic functional calculus is instrumental even when dealing with diagonalizable matrices, since it allows to turn bounds on the operator norm of the resolvent of a matrix into bounds on the norm of functions applied to the matrix, which is a crucial technique in perturbation theory. In this context, although obvious, the *resolvent identity*

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

proves to be extremely useful.

Perturbation Theory

The Hermitian Case. The perturbation theory for spectral quantities of a Hermitian matrix is quite simple. The eigenvalue displacement under a Hermitian perturbation of a Hermitian matrix can be controlled via Weyl's inequality [89, Theorem 4.3.1], while the eigenvector displacements can be controlled using the Davis-Kahan theorem [46].

Lemma 1.1.5 (Weyl's inequality)**.** *If $A, E \in \mathbb{C}^{n\times n}$ are Hermitian, then for all $1 \le i \le n$*

$$|\lambda_i(A+E) - \lambda_i(A)| \le \|E\|.$$

Lemma 1.1.6 (Davis-Kahan)**.** *If $A, E \in \mathbb{C}^{n\times n}$ are Hermitian and $\lambda_i(A)$ has multiplicity one, then*

$$\sin \angle(v_i(A), v_i(A+E)) \le \frac{2\|E\|}{\mathrm{gap}_i(A)}$$

where $\angle(v_i(A), v_i(A+E)) \in [0, \pi/2]$ denotes the angle between $v_i(A)$ and $v_i(A+E)$.

Note that the above two lemmas are the Hermitian version of Proposition 1.1.3, where no constraints on the size of the perturbation E are needed and for the eigenvector displacement bound the dimension dependence can be removed. For nonnormal matrices the perturbation theory for spectral quantities is much more delicate, and other concepts and definitions are required for a detailed understanding.

Eigenvalue Condition Numbers. If $A \in \mathbb{C}^{n\times n}$ has distinct eigenvalues $\lambda_1, \dots, \lambda_n$ and spectral decomposition as in (1.1), we define the *eigenvalue condition number of* λ_i to be

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

These quantities can be used to provide local (nonnormal) versions of Lemmas 1.1.5 and 1.1.6. To be precise, consider a smooth trajectory $A(t)$ in $\mathbb{C}^{n\times n}$ for which $A(t)$ has distinct eigenvalues at all times t. One can show (see [80] for details) that there exist smooth trajectories $w_i(t), v_i(t) \in \mathbb{C}^n$ and $\lambda_i(t) \in \mathbb{C}$ with $w_i(t)^* v_i(t) = 1$ such that the spectral decomposition of $A(t)$ is given by $\sum_{i=1}^n \lambda_i(t) v_i(t) w_i(t)^*$. Then, one can control the rate of change of the eigenvalues and eigenvectors along the trajectory using the following inequalities

$$|\dot{\lambda}_i(t)| \le \kappa(\lambda_i(t))\|\dot{A}(t)\| \quad \text{and} \quad \frac{\|\dot{v}_i(t)\|}{\|v_i(t)\|} \le \frac{\kappa_V(A(t))\|\dot{A}(t)\|}{\mathrm{gap}_i(A(t))}, \tag{1.7}$$

which can be easily proven from explicit formulas for $\dot{\lambda}_i(t)$ and $\dot{v}_i(t)$ derived in [80, Theorems 1 and 2].

Moreover, for an arbitrary diagonalizable matrix $A = VDV^{-1}$ with eigenvalues $\lambda_1, \dots, \lambda_n$, if one takes V to have columns of norm one, the following useful relationship can be derived

$$\max_{i\in[n]} \kappa(\lambda_i) \le \kappa_V(A) \le \|V\|\|V^{-1}\| \le \|V\|_F\|V^{-1}\|_F \le \sqrt{n \cdot \sum_{i=1}^n \kappa(\lambda_i)^2}. \tag{1.8}$$

Therefore, to a first order, the rates of change of the eigenvalues and eigenvectors under a perturbation can be controlled purely in terms of the eigenvalue condition numbers and the eigenvalue gaps. However, oftentimes, when one seeks to provide rigorous guarantees for an algorithm in the finite precision arithmetic model, it is often necessary to deal with macroscopic perturbations and provide an explicit ball around a matrix where a given inequality will hold (just like in Proposition 1.1.3). For this endeavour the notion of pseudospectrum becomes crucial.

The ϵ-pseudospectrum. Given $A \in \mathbb{C}^{n\times n}$ and $\epsilon \ge 0$ we define the ϵ-pseudospectrum of A as the set

$$\Lambda_\epsilon(A) := \{\lambda \in \mathbb{C} : \|(\lambda - A)^{-1}\| \ge 1/\epsilon\}. \tag{1.9}$$

In particular $\mathrm{Spec}(A) \subset \Lambda_\epsilon(A)$ for every $\epsilon > 0$ and $\mathrm{Spec}(A) = \Lambda_0(A)$. Moreover, one can show (see [158, Theorem 2.1]) that

$$\Lambda_\epsilon(A) = \{\lambda \in \mathbb{C} : \lambda \in \mathrm{Spec}(A + E) \text{ for some } \|E\| \le \epsilon\},$$

and as direct consequence derive the following two standard properties.

Lemma 1.1.7. *For any $A, E, U \in \mathbb{C}^{n\times n}$ with $\|E\| \leq \epsilon$ and U unitary, the following hold*

i) $\Lambda_\epsilon(UAU^*) = \Lambda_\epsilon(A)$.

ii) $\Lambda_{\epsilon-\|E\|}(A+E) \subset \Lambda_\epsilon(A)$.

We refer the reader to the book [158] for an elegant and comprehensive treatment on the notion of pseudospectrum. Here, we will only need the following basic lemmas.

Lemma 1.1.8 ([158], Theorem 4.3)**.** *For any $A \in \mathbb{C}^{n\times n}$ and any $\epsilon > 0$, every connected component of $\Lambda_\epsilon(A)$ has a non-empty intersection with* Spec(A).

Lemma 1.1.9 ([158], Theorems 2.2 and 2.3)**.** *For every $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$,*

$$\bigcup_{i=1}^{n} D(\lambda_i, \epsilon) \subset \Lambda_\epsilon(A) \subset \bigcup_{i=1}^{n} D(\lambda_i, \epsilon\kappa_V(A)). \tag{1.10}$$

Where for $z \in \mathbb{C}$ and $r > 0$, $D(z, r)$ is used to denote the circle of radius r centered at z.

Note that as $\epsilon \to 0$, (1.10) can be refined by invoking (1.7) so that the radius of the i-th disk solely depends on the stability of λ_i, that is

$$\Lambda_\epsilon(A) \subset \bigcup_{i=1}^{n} D(\lambda_i, \epsilon\kappa(\lambda_i) + O(\epsilon^2)).$$

In fact, when the eigenvalues of A are distinct, it turns out that as $\epsilon \to 0$ the set $\Lambda_\epsilon(A)$ asymptotically looks like a disjoint union of disks of radius $\epsilon\kappa(\lambda_i)$, in the following sense.

Lemma 1.1.10 ([15], Lemma 3.2)**.** *Let $A \in \mathbb{C}^{n\times n}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\mathcal{B} \subset \mathbb{C}$ be an open set. Then*

$$\lim_{\epsilon\to 0^+} \frac{\text{Area}(\Lambda_\epsilon(A) \cap \mathcal{B})}{\pi\epsilon^2} = \sum_{i:\lambda_i\in\mathcal{B}} \kappa(\lambda_i)^2.$$

Perturbation of Spectral Projectors. Importantly, the results discussed above can be used to obtain bounds on the displacement of spectral projections under perturbations of their matrix. Although the proof is trivial, given its importance, we record the bound below as a lemma.

Lemma 1.1.11 (Stability of spectral projectors)**.** *Let $A \in \mathbb{C}^{n\times n}$ be arbitrary, λ_i be one of its simple eigenvalues and $\epsilon > 0$. Let Γ_i be a contour in the complex plane separating λ_i from the rest of the spectrum of A, and assume $\Lambda_\epsilon(A) \cap \Gamma_i = \emptyset$. Then, for any $A' \in \mathbb{C}^{n\times n}$ with $\|A - A'\| < \eta < \epsilon$ we have that A' has a unique eigenvalue, say λ'_i, in the region enclosed by Γ_i, and that*

$$\|P_i - P'_i\| \leq \frac{\ell(\Gamma_i)\eta}{2\pi\epsilon(\epsilon-\eta)}.$$

where P_i and $P_i'^*$ are the spectral projectors of A and A' corresponding to λ_i and λ'_i respectively.

Proof. Combining Lemmas 1.1.7 ii) and 1.1.8 we get that $\Lambda_{\epsilon-\eta}(A') \cap \Gamma_i = \emptyset$ and that A' has a unique eigenvalue λ_i' in the region enclosed by Γ_i. Therefore

$$\begin{aligned}
\|P_i - P_i'\| &= \frac{1}{2\pi}\left\| \oint_{\Gamma_i} (z-A)^{-1} - (z-A')^{-1} \mathrm{d}z \right\| \\
&= \frac{1}{2\pi}\left\| \oint_{\Gamma_i} (z-A')^{-1}(A'-A)(z-A)^{-1} \mathrm{d}z \right\| \\
&\leq \frac{1}{2\pi} \oint_{\Gamma_i} \|(z-A')^{-1}\| \|A'-A\| \|(z-A)^{-1}\| dz \\
&\leq \frac{\ell(\Gamma_i)}{2\pi} \frac{1}{\epsilon - \eta} \cdot \eta \cdot \frac{1}{\epsilon}
\end{aligned}$$

Where in the last inequality we used the assumption $\Lambda(A) \cap \Gamma_i = \emptyset$ and $\Lambda_{\epsilon-\eta}(A') \cap \Gamma_i = \emptyset$ to upper bound the norms of the respective resolvents. □

We are now ready to prove Proposition 1.1.3.

Proof of Proposition 1.1.3. For $t \in [0,1]$ define $A(t) = (1-t)A + tA'$. Since $\delta < \frac{\mathrm{gap}(A)}{8\kappa_V(A)}$ the Bauer-Fike theorem (which follows from (1.10)) implies that $A(t)$ has distinct eigenvalues for all t, and in fact $\mathrm{gap}(A(t)) \geq \frac{3\mathrm{gap}(A)}{4}$. Standard results in perturbation theory (for instance [80, Theorem 1] or any of the references therein) imply that for every $i = 1, \ldots, n$, $A(t)$ has a unique eigenvalue $\lambda_i(t)$ such that $\lambda_i(t)$ is a differentiable trajectory, $\lambda_i(0) = \lambda_i$ and $\lambda_i(1) = \lambda_i'$. Let $P_i(t)$ be the associated spectral projector of $\lambda_i(t)$ and write $P_i = P_i(0)$.

Let Γ_i be the positively oriented contour forming the boundary of the closed disk centered at λ_i with radius $\mathrm{gap}(A)/2$, and define $\epsilon = \frac{\mathrm{gap}(A)}{2\kappa_V(A)}$. Lemma 1.1.9 implies $\Lambda_\epsilon(A)$ is contained in the union of these disks over all $i \in [n]$, and for fixed $t \in [0,1]$, since $\|A - A(t)\| < t\delta \leq \epsilon/4$, Lemma 1.1.6 ii) gives the same containment for $\Lambda_{3\epsilon/4}(A(t))$. Since these disks intersect only in their boundaries (if they do at all), $\|(z-A)^{-1}\| \leq 1/\epsilon$ and $\|(z-A(t))^{-1}\| \leq 4/3\epsilon$ for $z \in \Gamma_i$. So, from the definition of eigenvector condition number and Lemma 1.1.11 we have

$$|\kappa(\lambda_i) - \kappa(\lambda_i(t))| \leq \|P_i(t) - P_i\| \leq \frac{\ell(\Gamma_i)}{2\pi} \cdot \frac{1}{\epsilon} \cdot \frac{4}{3\epsilon} \cdot \frac{\epsilon}{4} = \frac{\mathrm{gap}(A)}{2} \frac{2\kappa_V(A)}{3\mathrm{gap}(A)} = \frac{\kappa_V(A)}{3}$$

and hence $\kappa(\lambda_i(t)) \leq \kappa(\lambda_i) + \kappa_V(A)/3 \leq 4\kappa_V(A)/3$. Combining this with (1.8) we obtain

$$\kappa_V(A(t)) \leq 2\sqrt{n \cdot \sum_i \kappa(\lambda_i)^2} < 4n\kappa_V(A)/3.$$

There exist smooth functions $v_i(t)$ satisfying $v_i(0) = v_i$ and $A(t)v_i(t) = \lambda_i(t)v_i(t)$ for all $i \in [n]$ and $t \in [0,1]$ (see [80]), which furthermore admit the bound given in (1.7). However, these $v_i(t)$ need not in general be unit vectors (see [80, Section 3.4] and references for discussion of

various normalizations). Therefore set $\hat{v}_i(t) = \|v_i(t)\|^{-1} v_i(t)$, and note that by an application of the chain rule,

$$\|\dot{\hat{v}}_i(t)\| \leq \frac{\delta \kappa_V(A(t))}{\mathrm{gap}(A(t))}.$$

It then follows that the vectors $v_i' = \hat{v}_i(1)$ for $i \in [n]$ satisfy the conclusion of the theorem, by bounding $\kappa_V(A(t)) \leq 4n\kappa_V(A)/3$ and $\mathrm{gap}(A(t)) \geq \frac{3\mathrm{gap}(A)}{4}$, and integrating the resulting upper bound $\|\dot{\hat{v}}_i(t)\| \leq \frac{16n\delta\kappa_V(A)}{9\mathrm{gap}(A)}$ from $t = 0$ to $t = 1$.

□

Eigenvalues vs Singular Values

When working with nonnormal matrices, specially random nonnormal matrices, it will be useful to understand the relation between the singular values and eigenvalues of the matrix in question. For this, the log-majorization property (see [90, Theorem 3.3.2]) is often useful.

Lemma 1.1.12 (Log-majorization)**.** *For any $A \in \mathbb{C}^{n\times n}$ and any $1 \leq k \leq n$,*

$$\prod_{i=1}^{k} |\lambda_i(A)| \leq \prod_{i=1}^{k} \sigma_i(A).$$

where $|\lambda_n(A)| \leq \cdots \leq |\lambda_1(A)|$ denote the eigenvalues of A ordered by their size. Moreover, because $\prod_{i=1}^{n} \sigma_i(A) = |\det(A)| = \prod_{i=1}^{n} |\lambda_i(A)|$ the above inequality can be rewritten as

$$\prod_{i=1}^{k} \sigma_{n-i+1}(A) \leq \prod_{i=1}^{k} |\lambda_{n-i+1}(A)|.$$

The above shows how to compare the modulus of an aggregate of eigenvalues with their corresponding singular values. If one is willing to lose a factor of $\kappa_V(A)$ it is possible to do a one by one comparison (with upper and lower bounds) as the following lemma shows.

Lemma 1.1.13. *Let $A \in \mathbb{C}^{n\times n}$ and $|\lambda_n(A)| \leq \cdots \leq |\lambda_1(A)|$ be the eigenvalues of A. Then, for any $k \in [n]$*

$$\kappa_V(A)^{-1} |\lambda_k(A)| \leq \sigma_k(A) \leq \kappa_V(A) |\lambda_k(A)|.$$

Proof. Write $A = VDV^{-1}$ for V attaining $\|V\|\|V^{-1}\| = \kappa_V(A)$. Using the Courant-Fischer min-max formulas we get:

$$\begin{aligned}
\sigma_k(A) &= \min_{S:\dim(S)=n-k+1} \max_{x\in S\setminus\{0\}} \frac{\|VDV^{-1}x\|}{\|x\|} \\
&= \min_{S:\dim(S)=n-k+1} \max_{y\in V(S)\setminus\{0\}} \frac{\|VDy\|}{\|Vy\|} && \text{setting } y = Vx \\
&= \min_{S:\dim(S)=n-k+1} \max_{y\in S\setminus\{0\}} \frac{\|VDy\|}{\|Vy\|} && \text{since } V \text{ is invertible}
\end{aligned}$$

$$\leq \min_{S:\dim(S)=n-k+1} \max_{y\in S\setminus\{0\}} \frac{\|V\|\|Dy\|}{\sigma_n(V)\|y\|}$$
$$= \kappa_V(M)\sigma_k(D).$$

Since $D = \operatorname{diag}(\lambda_1(A), \ldots, \lambda_n(A))$ we have $\sigma_k(D) = |\lambda_k(A)|$. Similarly, reusing the above equations we have that

$$\sigma_k(A) = \min_{S:\dim(S)=n-k+1} \max_{y\in S\setminus\{0\}} \frac{\|VDy\|}{\|Vy\|}$$
$$\geq \min_{S:\dim(S)=n-k+1} \max_{y\in S\setminus\{0\}} \frac{\sigma_n(V)\|Dy\|}{\|V\|\|y\|}$$
$$= \kappa_V(A)^{-1}\sigma_k(D).$$

□

1.2 Chapter 2: Spectral Stability Under Random Perturbations

In this dissertation randomness plays an instrumental role when establishing general algorithm performance guarantees that hold on arbitrary inputs. In regards to the eigenvalue problem (in the spirit of smoothed analysis [144] and as suggested by Davies in [45]) we study and exploit a random matrix phenomenon: if one adds independent random variables of small variance to the entries of an arbitrary matrix $A \in \mathbb{C}^{n\times n}$, with high probability, the resulting matrix A' will have (relatively) stable eigenvenvalues and eigenvectors. Moreover, by taking the distribution of the entries of the random perturbation to have fast decaying tails (e.g. taking Gaussian or bounded entries), with high probability, A' will be close to A, and therefore it will still be possible to find a valid solution to the backward eigenvalue problem (see Definition 1.1.1) by running the algorithm of our choice on A'. Concretely, here we will discuss statements of the following form.

Proto-Statement 1.2.1. Let M_n be a random matrix with independent entries of variance one. Then, for any $A \in \mathbb{C}^{n\times n}$ and $\gamma > 0$, under certain assumptions about the distributions of the entries of M_n, for any $t \in (0, 1)$ it holds that

$$\mathbb{P}\left[\kappa_V(A + \gamma M_n) \geq \frac{1}{t}\right] \leq \operatorname{poly}_1(\|A\|, \gamma^{-1}, t, n) \tag{1.11}$$

and

$$\mathbb{P}\left[\operatorname{gap}(A + \gamma M_n) \leq t\right] \leq \operatorname{poly}_2(\gamma^{-1}, t, n) \tag{1.12}$$

where poly_1 and poly_2 are certain explicit multivariate polynomials that converge to 0 as $t \to 0$ if the other parameters remain fixed.

Remark 1.2.2. Note that by putting (1.11) and (1.12) together one gets that κ_{eig} is controlled with high probability, which by Proposition 1.1.3 ensures a certain degree of stability for the eigenvalues and eigenvectors of the resulting matrix.

Formulas for the expectation of certain condition numbers of $A+\gamma M_n$ in the case when M_n is a complex Ginibre were derived in [3], and from those formulas one can obtain inequalities of the form (1.11) and (1.12) when M_n is a complex Ginibre (see Appendix A for a detailed discussion). However, this approach breaks down when the entries of M_n are not complex Gaussians, posing the question of if this phenomenon is still present when the entries of M_n are distributed differently.

In [15] a different approach for controlling $\kappa_V(A+\gamma M_n)$ was introduced (see below for details), and it is not hard to see that the results from this paper can be applied easily to derive an inequality of the form (1.11) in the particular case when M_n is a complex Ginibre matrix (see Appendix A). Moreover, by complementing the arguments from [15] with standard random matrix techniques one can obtain an inequality of the form (1.11) in the more general case when M_n has entries whose distributions are absolutely continuous with respect to the Lebesgue measure on $\mathbb{C}$. However, this approach breaks down once absolutely continuity over $\mathbb{C}$ is removed (e.g. it is not possible to derive from those ideas a tail bound when M_n is a *real* Ginibre matirx), posing the question of independent interest in random matrix theory of if stability of eigenvectors and eigenvalues can also be ensured when using real perturbations. Note that this question also has certain relevance in the context of numerical analysis, since often times it is convenient to maintain all computations within the real numbers if the input is real. So, in [10] we studied the following random matrix model.

Assumption 1.2.3. Let M_n be an $n \times n$ random matrix. We will assume that M_n has independent real entries, each with density on $\mathbb{R}$ upper bounded almost everywhere by $\sqrt{n}K > 0$, for some constant K.

Remark 1.2.4. The $\sqrt{n}$ term in the upper bound on the density of the entries of M_n is accounting for the fact that M_n may be a normalized random matrix (i.e. $\mathbb{E}\|M_n\|$ is upper bounded by a universal constant, independent of n). For example, if M_n is a normalized real Ginibre matrix, then the density of its entries will be upper bounded by $\sqrt{n/2\pi}$, that is, in this case $K = 1/\sqrt{2\pi}$. Our aim will be to provide inequalities of the form (1.11) and (1.12) where the coefficients of poly_1 and poly_2 will a function of K.

We warn the reader that depending on the entry distributions of M_n, the $\sqrt{n}$ in Assumption 1.2.3 need not be the appropriate normalization so that $\mathbb{E}\|M_n\| = O(1)$. However, this holds in the case when the entries of M_n have bounded fourth moment, and we include this explicit scaling for easier comparison to the Gaussian case.

We will find it useful to state our results in terms of the L^p-norms

$$B_{M_n,p} := \mathbb{E}\left[\|M_n\|^p\right]^{1/p}, \tag{1.13}$$

which are uniformly bounded (over n for fixed p) when the entries of M_n have rapid decay. For example, $B_{M_n,p} \leq 9$ for all $p \leq 2n$ when M_n is a normalized real Ginibre matrix.

Theorem 1.2.5 (Minimum Eigenvalue Gap)**.** *Let $n \geq 16$, $A \in \mathbb{R}^{n\times n}$ be deterministic, and M_n be a random matrix satisfying Assumption 1.2.3 with parameter $K > 0$. For any $0 < \gamma < K$ and $r < 1 < R$:*

$$\begin{aligned}&\mathbb{P}\left[\operatorname{gap}(A+\gamma M_n) \leq r\right]\\ \leq& C_{1.2.5} R^2 \left(\gamma B_{M_n,8} + \|A\| + R\right)(K/\gamma)^{5/2} n^4 r^{2/7} + \mathbb{P}\left[\|A+\gamma M_n\| \geq R\right],\end{aligned} \tag{1.14}$$

where $C_{1.2.5}$ is an explicitly computable (moderate) universal constant.

Theorem 1.2.6 (Eigenvector condition number)**.** *Let $n \geq 9$. Let $A \in \mathbb{R}^{n\times n}$ be deterministic, and let M_n satisfy Assumption 1.2.3 with parameter $K > 0$. Let $0 < \gamma < K \min\{1, \|A\| + R\}$ and $R > \mathbb{E}\|\gamma M_n\|$. Then for any $\epsilon_1, \epsilon_2 > 0$, with probability at most*

$$2\epsilon_1 + O\left(\frac{R(R+\|A\|)^{3/5} K^{8/5} n^{14/5} \epsilon_2^{3/5}}{\gamma^{8/5}}\right) + 2\mathbb{P}[\gamma\|M_n\| > R],$$

we have

$$\kappa_V(A+\gamma M_n) \leq \epsilon_1^{-1}\sqrt{\log(1/\epsilon_2)} C_{1.2.6} K^{3/2} n^3 \cdot \frac{(\|A\|+R)^{3/2}}{\gamma^{3/2}}.$$

where $C_{1.2.6}$ is a universal constant.

In Chapter 2 we will also present upper bounds for the tails of the (random) eigenvalue condition numbers of $A + \gamma M_n$, where it will be crucial to distinguishing between the real and truly complex eigenvalues. Very similar results, using essentially different methods, were obtained by other authors in the independent and concurrent work [93]. We refer the reader to Chapter 2 for a comparison between the two works and more in general for a detailed discussion of the relevant literature. For now, we give a brief overview of the proof techniques used to prove Theorems 1.2.5 and Theorem 1.2.6.

Proof techniques: Hermitization

First note that Assumption 1.2.3 is not imposing any constraints on the mean of the entries of M_n or on the size of K, so, to simplify notation, when working with $A + \gamma M_n$ we can "absorb" A and γ into M_n, and just focus on proving results about M_n for general K.

Now, when dealing with the spectrum of nonnormal random matrices, such as M_n, it is a common trick (that goes back at least to Girko [74]) to reduce the problem to understanding the singular values of $z - M_n$ for all $z \in \mathbb{C}$. Since with this, one turns the problem of studying the eigenvalues of a nonnormal matrix into analyzing the eigenvalues of a family of Hermitian random matrices, this technique is often referred to as *Hermitization*.

Hermitization for minimum eigenvalue gaps. Here we will summarize our approach from Chapter 2, which was inspired by the work of Ge [73], although it is significantly simpler. The main observation is the following.

Lemma 1.2.7. *Let $A \in \mathbb{C}^{n\times n}$, $z \in \mathbb{C}$ and $r > 0$. If A has two eigenvalues in $D(z, r)$, then*

$$\sigma_n(z - A)\sigma_{n-1}(z - A) \leq r^2.$$

Proof. Applying Lemma 1.1.12 to $z - A$, we get

$$\sigma_n(z - A)\sigma_{n-1}(z - A) \leq |\lambda_n(z - A)||\lambda_{n-1}(z - A)| \leq r^2.$$

□

The above result tells us that if M_n has two eigenvalues λ_i and λ_j that are close to each other, this will manifest in the product $\sigma_n(z-M_n)\sigma_{n-1}(z-M_n)$ being small for any z close to λ_i and λ_j. So, the idea is that given $r > 0$ one can judiciously choose $\epsilon > 0$ and an ϵ-covering[6] $\mathcal{N}$ on the relevant region of $\mathbb{C}$, so that the event $\{\text{gap}(M_n) \leq r\}$ is contained in the event $\{\sigma_n(z - M_n)\sigma_{n-1}(z - M_n) \leq r^2 : \text{ for some } z \in \mathcal{N}\}$, and using a union bound one can obtain

$$\mathbb{P}[\text{gap}(M_n) \leq r] \leq \sum_{z\in\mathcal{N}} \mathbb{P}[\sigma_n(z - M_n)\sigma_{n-1}(z - M_n) \leq r^2]. \tag{1.15}$$

On the other hand, by separately controlling the left tails

$$\mathbb{P}[\sigma_n(z - M_n) \leq r_1] \quad \text{and} \quad \mathbb{P}[\sigma_{n-1}(z - M_n) \leq r_2] \tag{1.16}$$

for any z, one can obtain easily obtain a bound on each of the terms appearing on the right-hand side of (1.15), ultimately yielding an upper bound on $\mathbb{P}[\text{gap}(M_n) \leq r]$. Certainly, for this bound to be meaningful, the control on the tails in (1.16) should be strong enough to subdue the size of the net $\mathcal{N}$, which will be $O(1/\epsilon^2)$, which in turn is a function of r (since ϵ itself is a function of r). In Chapter 2 we will show that balancing these parameters to obtain a meaningful bound is possible if M_n satisfies Assumption 1.2.3. For now, to give the reader some idea on what bounds on tails of singular values will look like, we recall the case when $M_n = A + \gamma G_n$ for $A \in \mathbb{C}^{n\times n}$ deterministic, $\gamma > 0$, and G_n a normalized complex Ginibre matrix. In this particular case, a combination of the results in [147] and [143] (see [15, Section 2] for details) yields

$$\mathbb{P}[\sigma_{n-k+1}(z - A - \gamma G_n) \leq r] \leq \left(\frac{Cnr}{\gamma k}\right)^{2k^2} \tag{1.17}$$

for some universal constant C and for all $k \leq n$ (for this discussion only $k = 1, 2$ are relevant).

Remark 1.2.8 (Overcrowding estimates)**.** Note that the same argument can be repeated looking at the k smallest singular values (instead of just the smallest two) to prove results about clusters of k eigenvalues in the spectrum of M_n, such as the ones appearing in [115].

[6] Given a region $\Omega \subset \mathbb{C}$ and $\epsilon > 0$, an ϵ-covering of Ω is a (usually finite) set of points $\mathcal{N}$ in Ω for which $\Omega \subset \bigcup_{z\in\mathcal{N}} D(z, \epsilon)$.

Hermitization for eigenvalue condition numbers. In [15] Banks, Kulkarni, Mukherjee and Srivastava used (1.17) (in the particular case of $k = 1$) to show the following.

Theorem 1.2.9 (Theorem 1.5 in [15]). *Let $A \in \mathbb{C}^{n\times n}$, $\gamma \in (0, \|A\|)$, and G_n be a complex normalized Ginibre matrix. If $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ are the random eigenvalues of $A + \gamma G_n$, for every measurable open set $\mathcal{B} \subset \mathbb{C}$*

$$\mathbb{E}\left[\sum_{\lambda_i \in \mathcal{B}} \kappa(\lambda_i)^2\right] \leq \frac{n^2}{\pi\gamma^2}\text{Area}(\mathcal{B}).$$

From a statement like the one above one can easily derive a bound of the form (1.11). The idea is that if one takes $\mathcal{B}$ to be, say, $D(0, \|A\| + 4\gamma)$, then with exponentially high probability all of the eigenvalues of $A + \gamma G_n$ will be contained in $\mathcal{B}$, so one can use Theorem 1.2.9 and Markov's inequality to obtain a tail bound on $\sum_{i=1}^n \kappa(\lambda_i)^2$ which in turn, by (1.8), implies a tail bound for $\kappa_V(A + \gamma G_n)$.

We now present the arguments in [15] for the proof of Theorem 1.2.9, after which we discuss where exactly the assumption of complex Gaussian entries is used and what would one need to prove a more general result.

Proof of Theorem 1.2.9. Start by controlling the expected area of the ϵ-pseudospectrum of $A + \gamma G_n$. To do this observe that

$$\Lambda_\epsilon(A + \gamma G_n) = \{z \in \mathbb{C} : \|(z - A - \gamma G_n)^{-1}\| \geq 1/\epsilon\} = \{z \in \mathbb{C} : \sigma_{\min}(z - A - \gamma G_n) \leq \epsilon\}.$$

Hence

$$\begin{aligned}
\mathbb{E}\left[\text{Area}(\Lambda_\epsilon(A + \gamma G_n) \cap \mathcal{B})\right] &= \mathbb{E}\left[\int_{\mathcal{B}} \mathbb{1}\left\{\sigma_n(z - A - \gamma G_n) \leq \epsilon\right\} dz\right] \\
&= \int_{\mathcal{B}} \mathbb{P}[\sigma_n(z - A - \gamma G_n) \leq \epsilon] dz && \text{by Fubini} \\
&\leq \int_{\mathcal{B}} \frac{Cn^2\epsilon^2}{\gamma^2} dz && \text{by (1.17)} \\
&= \frac{Cn^2\epsilon^2}{\gamma^2}\text{Area}(\mathcal{B}).
\end{aligned}$$

Moreover, in [62, Section 5] it was shown that (1.17) holds for $C = 1$ when $k = 1, A = 0$ and $\gamma = 0$, and combining this with the results from [143], one gets that (1.17) holds for $C = 1$ when $k = 1$ and for arbitrary A and γ. Therefore, if we divide by $\pi\epsilon^2$ the first and last expression in the above chain of inequalities we get

$$\mathbb{E}\left[\frac{\text{Area}(\Lambda_\epsilon(A + \gamma G_n) \cap \mathcal{B})}{\pi\epsilon^2}\right] \leq \frac{n^2\text{Area}(\mathcal{B})}{\pi\gamma^2}$$

The proof is then concluded by taking $\epsilon \to 0$ and applying Lemma 1.1.10. □

Observe that in the above proof it was crucial that the ϵ-dependence in the upper bound on $\mathbb{P}[\sigma_n(z-A-\gamma G_n) \leq \epsilon]$ was of the form $O(\epsilon^2)$. As it will be discussed below, this is typical of random matrix models M_n whose entries are absolutely continuous with respect to the Lebesgue measure on $\mathbb{C}$, however when the entries of M_n are real one should expect that the left tail of $\sigma_n(M_n)$ decays as $O(\epsilon)$, rendering the above approach incomplete.

Proof techniques: Invertibility via Distance

The literature on tails of least singular values of random matrices is immense (see [154] for a recent survey). However, the vast majority of the papers in the subject have as a common starting point (some form of) the by now well known *invertibility via distance argument*: let M_n be a random matrix with independent entries and $z \in \mathbb{C}$. Let $R_1, \ldots, R_n$ be the rows of M_n, let $C_1, \ldots, C_n$ be the columns of $(z - M_n)^{-1}$, and let $V_i := \text{Span}\{R_j : j \in [n] \setminus i\}$. Then basic linear algebra yields that $\|C_i\| = \mathsf{dist}(R_i, V_i)^{-1}$ and therefore

$$\|(z - M_n)^{-1}\|^2 \leq \sum_{i=1}^{n} \sigma_i(z - M_n)^{-2} = \|(z - M_n)^{-1}\|_F^2 = \sum_{i=1}^{n} \|C_i\|^2 = \sum_{i=1}^{n} \mathsf{dist}(R_i, V_i)^{-2}.$$

Using the above inequality and a union bound we obtain

$$\mathbb{P}[\sigma_n(z - M_n) \leq r] = \mathbb{P}[\|(z - M_n)^{-1}\|^2 \geq 1/r^2] \leq \sum_{i=1}^{n} \mathbb{P}[\mathsf{dist}(R_i, V_i) \leq r]. \tag{1.18}$$

To conclude the argument, note that if N_i is a unit vector orthogonal to V_i, then

$$\mathsf{dist}(R_i, V_i) = |R_i^* N_i|. \tag{1.19}$$

Moreover, because the entries of M_n are independent, N_i will be independent of R_i, so to upper bound $\mathbb{P}[\mathsf{dist}(R_i, V_i) \leq r]$ it is sufficient to upper bound $\mathbb{P}[|R_i^* X| \leq r]$ for every deterministic unit vector X. Now note that, again because the entries of M_n are independent, $R_i^* X$ is a sum of n independent random variables, so if the distribution of each of them has a density on $\mathbb{C}$ upper bounded say by $\sqrt{n}K$, $R_i^* X$ will have a density on $\mathbb{C}$, and it is not hard to show (e.g. see the proof of Lemma 4.12 in [29]) that in fact this density will be upper bounded by nK (regardless of the choice of X, as long as $\|X\| = 1$). Therefore we obtain the *anticoncentration* bound

$$\mathbb{P}[|R_i^* X| \leq r] \leq n\pi K r^2.$$

Combining this with (1.18) one gets

$$\mathbb{P}[\sigma_n(z - M_n) \leq r] \leq n^2 \pi K r^2.$$

And since this bound has r-dependence $O(r^2)$ one can easily adapt the proof of Theorem 1.2.9 to show an inequality of the form (1.11) in the case where the entries of M_n have a bounded density on $\mathbb{C}$.

However, in the case when the entries of M_n are real, the situation is not straight forward and anti-concentration bounds for $|R_i^* X|$ heavily depend on the imaginary part of the vector X. As a consequence, $\mathrm{Im}(z)$ plays an important role in determining the kind of distribution that (1.19) has on $\mathbb{C}$ and hence on the kind of bound one can obtain on $\sigma_n(z - M_n)$. In this regard, from the work of Ge [73] it follows that if M_n satisfies Assumption 1.2.3 one has

$$\mathbb{P}[\sigma_n(z - M_n) \leq r \quad \text{and} \quad \|M_n\| \leq R] \leq \frac{Cn^2r^2}{\mathrm{Im}(z)} + e^{-cn}$$

for some universal constants c, C and R. However, since the right-hand side of the above upper bounds does not go to zero as $r \to 0$, it is still not clear how to complete the Hermitization argument for the eigenvalue condition numbers of M_n with a bound of this sort. Moreover, the Hermitization argument for the eigenvalue gaps of M_n also requires control of the left tails of $\sigma_{n-1}(z - M_n)$. With this in mind, in [10] we showed the following results (which were instrumental in proving Theorems 1.2.5 and 1.2.6 discussed above).

Theorem 1.2.10 (Singular Values of Complex Shifts)**.** *Let $z \in \mathbb{C} \setminus \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and let M_n satisfy Assumption 1.2.3 with parameter $K > 0$. Then, for every $k \leq \sqrt{n} - 2$,*

$$\begin{aligned}&\mathbb{P}\left[\sigma_{n-k+1}\left(z - (A + M_n)\right) \leq \epsilon\right] \\ \leq &(1 + k^2)\binom{n}{k}^2 \left(C_{1.2.10} k^2 (nK)^3 \left((B_{M_n,2k^2} + \|A\| + |\operatorname{Re} z|)^2 + |\operatorname{Im} z|^2\right) \frac{\epsilon^2}{|\operatorname{Im} z|}\right)^{k^2},\end{aligned}$$

where $C_{1.2.10} = 8\sqrt{3}(e\pi)^{3/2}$.

The main technical tools for the proof of the above theorem will be developed in Sections 2.3 and 2.4. One of these ingredients will be a restricted invertibility lemma that will play the role of the inequality $\|(z - M_n)^{-1}\| \leq \sum_{n=1}^{n} \mathsf{dist}(R_i, V_i)^{-2}$ in the more general context of controlling the tail of the k-th smallest singular value. The other ingredient will be an anticoncentration result for random quadratic forms which will provide tail bounds for the quantities that will replace the expressions (1.19) appearing above. These two ingredients will be combined via a delicate analysis of the interaction between the real and imaginary parts of the resolvent $(z - M_n)^{-1}$.

1.3 Chapter 3: Spectral Bisection

A myriad of papers have been written on different aspects of the eigenvalue problem, both in the computer science and numerical analysis communities. While there are provably fast and accurate algorithms in the Hermitian case and a large body of work for various structured matrices (see, e.g., [27]), the general case is not nearly as well-understood, and in 1997 Demmel remarked in his well-known textbook [55]: "...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 [3]. They exhibited an algorithm (see their Theorem 2.28) which given any $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$ and $\sigma > 0$ produces in $O(n^9/\sigma^2)$ expected arithmetic operations the diagonalization of the nearby random perturbation $A + \sigma G_n$ where G_n is an unnormalized complex Ginibre matrix. By setting σ sufficiently small, this may be viewed as a solution to the *backward eigenvalue problem* – in particular, by setting $\sigma = \delta/\sqrt{n}$ and noting that $\|G_n\| = O(\sqrt{n})$ with very high probability, their result implies an expected running time of $O(n^{10}/\delta^2)$ for an output of backward accuracy δ in our setting.

The main goal of Chapter 3 will be to show that a variant of the well-known spectral bisection algorithm of Beavers and Denman [23], combined with a pre-prosessing step that consists of adding a random perturbation to the input matrix (as discussed in Section 1.2), can be implemented in a backward stable way (in the relaxed sense of stability defined in Section 1.1.2) using only a polylogarithmic number of calls to a given matrix-multiplication oracle. From the complexity theory perspective, this is a significant improvement to the result of Armentano et al. and shows that the complexity of the (randomized) eigenvalue problem (defined as in Definition 1.1.1) is nearly that of matrix-multiplication. To be precise, we will show the following.

Theorem 1.3.1 (Backward Approximation Algorithm)**.** *There is a randomized algorithm* EIG *which on any input matrix $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$ and a desired accuracy parameter $\delta > 0$ outputs a diagonal D and invertible V such that*

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

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 $1 - 14/n$. Here $T_{\mathsf{MM}}(n)$ refers to the running time of a numerically stable matrix multiplication algorithm (detailed in Chapter 3).

Of course, in view of Proposition 1.1.3 the above result also guarantees the existence of an algorithm that can give forward approximations, that is, by invoking EIG with specified accuracy $\frac{\delta}{6n\kappa_{\text{eig}}}$ one can obtain the following guarantee.

Corollary 1.3.2 (Forward Approximation Algorithm)**.** *There is a randomized algorithm which on any input matrix $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$, 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 $1-1/n-12/n^2$.

Below we outline the main ingredients that will be used to show Theorem 1.3.1.

Computing the Matrix Sign Function

As we will see later, computing the sign function of a matrix is the key subroutine of the spectral bisection algorithm.

The matrix sign function. The sign function of a number $z \in \mathbb{C}$ with $\text{Re}(z) \neq 0$ is defined as $+1$ if $\text{Re}(z) > 0$ and -1 if $\text{Re}(z) < 0$. 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

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

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 [6] defined the following condition number for the sign function:

$$\kappa_{\text{sign}}(M) := \inf\left\{1/\epsilon^2 : \Lambda_\epsilon(M) \text{ does not intersect the imaginary axis}\right\}, \tag{1.20}$$

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

Roberts' Newton Iteration. Roberts [131] showed that the simple iteration defined by

$$A_0 = A \quad \text{and} \quad A_{k+1} = \frac{A_k + A_k^{-1}}{2} \tag{1.21}$$

converges globally and quadratically to $\text{sgn}(A)$ in exact arithmetic, but his proof relied on the fact that all iterates of the algorithm are simultaneously diagonalizable (see below for more details), a property which is destroyed in finite arithmetic since inversions can only be done approximately.[7] 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. Our main result in this direction will be the following.

[7] 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.

Theorem 1.3.3 (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* $K \geq \kappa_{\mathrm{sign}}(A)$*, and a desired accuracy* $\beta \in (0, 1/12)$*, outputs an approximation* $\mathsf{SGN}(A)$ *with*

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

in

$$O((\log K + \log\log(1/\beta))T_{\mathsf{INV}}(n)) \tag{1.22}$$

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).*

Proof Technique and Complications. The idea of Roberts' proof in exact arithmetic is that when $A = VDV^{-1}$ is diagonalizable one can recursively show that each A_k can be written as VD_kV^{-1} for some diagonal matrix D_k, by recursively writing

$$A_{k+1} = \frac{A_k + A_k^{-1}}{2} = \frac{VD_kV^{-1} + VD_kV^{-1}}{2} = V\left(\frac{D_k + D_k^{-1}}{2}\right)V^{-1},$$

which moreover provides the recursion $D_{k+1} = \frac{1}{2}(D_k + D_k^{-1})$. Once this has been observed it is clear that the role of V is immaterial and that one can view Roberts' iteration as a dynamic over diagonal matrices. Moreover, because there is no interaction between the distinct diagonal entries of these matrices, one can further reduce the problem to studying the dynamics on $\mathbb{C}$ defined by the recursion $z_{k+1} = \frac{z_k + z_k^{-1}}{2}$. It is then not hard to show that for any $z \in \mathbb{C}$ with $\mathrm{Re}(z) \neq 0$, if $z_0 = z$ then z_k converges quadratically to $\mathrm{sgn}(z)$ as $k \to \infty$.

Now, as mentioned above, when working in finite arithmetic, because inversion is not exact, the computed $\widetilde{A}_{k+1}$ will be a perturbation of what A_{k+1} would be if all computations were executed in exact arithmetic. Therefore, the $\widetilde{A}_k$ will not be simultaneously diagonalizable anymore, and moreover, if such matrices have unstable eigenvalues the spectrum of $\widetilde{A}_k$ can be quite far from the spectrum of A_k, and therefore the eigenvector matrices now play an important role in the dynamics. For these reasons, analyzing Roberts iteration in finite arithmetic is significantly harder than the exact arithmetic case.

The main new idea in the proof of Theorem 1.3.3 is to view the iteration as a dynamic on the pseudospectrum of the matrices in sequence (rather than a dynamic on their spectrum). This allows for more robust arguments that are resilient to inexact computations. To be more precise we control the evolution of the pseudospectra $\Lambda_{\epsilon_k}(\widetilde{A}_k)$ 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 will be introduced in Sections 3.4.1 and 3.4.2, and carried out in finite arithmetic in Section 3.4.3, yielding Theorem 1.3.3.

Spectral Bisection and Pseudospectral Shattering

Diagonalization by Spectral Bisection. Given an algorithm for computing the sign function, there is a natural and well-known approach to the eigenvalue problem pioneered in [23]. The main observation is that the matrices $P_{\pm} := \frac{1}{2}(I \pm \mathrm{sgn}(A))$ are the spectral projectors onto the invariant subspaces corresponding to the eigenvalues of A in the left and right open half planes of $\mathbb{C}$. To exploit this, note that this implies that for $r \in \mathbb{R}$, $P_{\pm}(r) := \frac{1}{2}(I \pm \mathrm{sgn}(A - r))$ are the spectral projectors corresponding to the half planes $\{z \in \mathbb{C} : \mathrm{Re}(z) > r\}$ and $\{z \in \mathbb{C} : \mathrm{Re}(z) < r\}$. So if the shifted matrix $A - r$ has roughly the same number of eigenvalues on the left and right sides of the complex plane then $P_+(r)$ and $P_-(r)$ can be used to compress A into two smaller subproblems (each of which will be roughly of the same size) appropriate for recursion. Note that when no such r exists one can compute the sign function on matrices of the form $iA - r$, which corresponds to trying to find a horizontal line in $\mathbb{C}$ that roughly evenly splits $\mathrm{Spec}(A)$, and that for any configuration of $\mathrm{Spec}(A)$ there is either a horizontal or vertical line that will work well for this purpose.

Grid Implementation. The way we will implement this algorithm is by initially choosing a grid $\mathtt{g}$ that divides $\mathbb{C}$ into small squares. Then, we will choose (in a way that will be made precise in Chapter 3) a line in $\mathtt{g}$ that splits the spectrum of the matrix in a somewhat balanced way, and as explained above this line will be used to reduce the problem to two subproblems, each with their spectra contained in one of the two sides of $\mathtt{g}$. By recursively continuing this procedure, in each step only retaining the part of $\mathtt{g}$ that is relevant to each subproblem, one will end up computing the spectral projectors for invariant subspaces whose corresponding eigenvalues are all contained in a common square of $\mathtt{g}$. Moreover, if $\mathtt{g}$ is taken so that there is at most one eigenvalue of the input matrix A per square, then the computed spectral projectors will be rank one and can be used to compute the eigenvectors of A; moreover, any point in the corresponding square can be used as a forward approximation for the eigenvalue corresponding to the computed eigenvector (and hence the resolution of the grid will determine the accuracy of the final output).

Difficulties. The difficulties in carrying out the above approach are: (a) finding a balanced splitting along an axis that is well-separated from the spectrum (this requires taking the initial grid with lines well-separated from the spectrum) (b) efficiently and accurately computing the sign function (c) ensuring that solving the subproblems obtained after compression (which have been corrupted by machine errors) would yield an accurate solution for the original problem. Note that (a) and (b) are nontrivial even in exact arithmetic, since the iteration (1.21) converges slowly if (a) is not satisfied, even without roundoff error. Moreover, for (b) and (c) to be carried out in finite arithmetic one has to make sure that throughout the algorithm the matrices that one works with have reasonable spectral stability properties. To this end, the initial prepossessing step with a random perturbation will be crucial.

Pseudospectral Shattering for a Grid. The results discussed in Section 1.2 imply that random matrices of the form $A + \gamma M_n$, with A deterministic and M_n random satisfying

Assumption 1.2.3, have with high probability relatively small eigenvalue condition numbers and relatively large eigenvalue gaps.

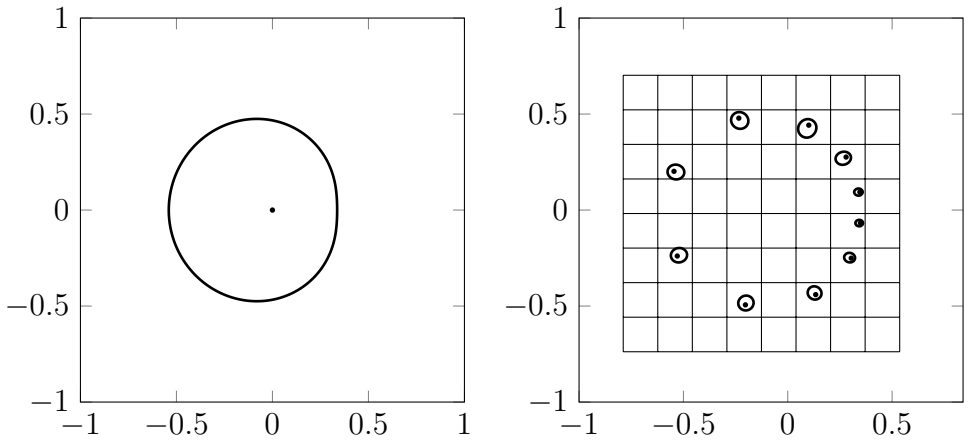

Figure 1.1: T is a sample of an upper triangular 10×10 Toeplitz matrix with zeros on the diagonal and an independent standard real Gaussian repeated along each diagonal above the main diagonal. G is a sample of a 10×10 complex Ginibre matrix with unit variance entries. Using the MATLAB package EigTool [176], the boundaries of the ϵ-pseudospectrum of T (left) and $T + 10^{-6}G$ (right) for $\epsilon = 10^{-6}$ are plotted along with the spectra. The latter pseudospectrum is shattered with respect to the pictured grid.

So, given the relation between the pseudospectrum and the eigenvalue condition numbers (e.g. see Lemma 1.1.9 and the discussion around it), this means that the pseudospectrum of $A + \gamma M_n$ is *shattered* into small pieces with high probability. In particular, in the context of the spectral bisection algorithm, to address the difficulties mentioned above we will want this small pseudospectral pieces to lie inside the chosen grid g (see Figure 1.1), and show that this property is maintained throughout the algorithm.

Definition 1.3.4 (Grid Pseudospectral Shattering)**.** Let $A \in \mathbb{C}^{n \times n}$ and $\epsilon > 0$. The 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. $\Lambda_\epsilon(A) \cap \mathsf{g} = \emptyset$.

The above property will not only be used to ensure that the eigenvalues of the matrix are far from the bisecting line whenever the subroutine for computing the sign function is called, but it will also be used to deduce spectral stability of the matrix via Lemma 1.1.11.

In Section 3.3 we will translate tail bounds for the minimum eigenvalue gaps and eigenvalue condition numbers of $A + \gamma M_n$ into high probability pseudospectral shattering guarantees with respect to a randomized grid g. To simplify our analyze we will limit the discussion to

the case when M_n is a complex Ginibre matrix, but as discussed in Section 1.2 this could be done more generally for a larger class of random perturbations.

1.4 Chapter 4: Hessenberg QR Algorithm

The Hessenberg Shifted QR Algorithm, discovered in the late 1950's independently by Francis [68, 69] and Kublanovskaya [99], has been for several decades the most widely used method for approximately computing all of the eigenvalues of a dense matrix, and moreover finding its approximate Schur form[8]. It is implemented in all of the major software packages for numerical linear algebra and was listed as one of the "Top 10 algorithms of the twentieth century," along with the Metropolis algorithm and the Simplex algorithm [59, 123]. The algorithm is specified by a *shifting strategy*, which is an efficiently computable function

$$\mathrm{Sh} : \mathbb{H}^{n\times n} \to \mathcal{P}_k,$$

where $\mathbb{H}^{n\times n}$ is the set of $n \times n$ complex Hessenberg[9] matrices and $\mathcal{P}_k$ is the set of monic complex univariate polynomials of degree k, for some $k = k(n)$ typically much smaller than n. The word "shift" comes from the fact that when $k = 1$ we have $p_t(H_t) = H_t - s_t I$ for some $s_t \in \mathbb{C}$. The algorithm then consists of the following discrete-time isospectral nonlinear dynamical system on $\mathbb{H}^{n\times n}$, given an initial condition H_0:

$$\begin{aligned} Q_t R_t &= p_t(H_t) \qquad \text{where } p_t = \mathrm{Sh}(H_t) \quad \text{whenever } p_t(H_t) \text{ is invertible,} \\ H_{t+1} &= Q_t^* H_t Q_t, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = 0, 1, 2, \ldots. \end{aligned} \tag{1.23}$$

The first step in (1.23) is a QR decomposition so that Q_t is unitary. It is not hard to see that each iteration preserves the Hessenberg structure; we ignore the case when $p_t(H_t)$ is singular in this overview (see Chapter 4 for a discussion of the singular case).

The relevance of this iteration to the eigenvalue problem stems from two facts. First, every matrix $A \in \mathbb{C}^{n\times n}$ is unitarily similar to a Hessenberg matrix H_0, and in exact arithmetic such a similarity can be computed exactly in $O(n^3)$ operations. Second, it was shown in [68, 99] that for the trivial "unshifted" strategy $p(z) = z$, the iterates H_t under some mild genericity conditions always converge to an upper triangular matrix H_∞; this is because the unshifted QR iteration can be precisely related to the (inverse) power iteration (see e.g. [157]). Combining the unitary similarities accumulated during the iteration, these two facts yield a Schur factorization $A = Q^* H_\infty Q$ of the original matrix, from which the eigenvalues of A can be read off. The unshifted QR iteration does *not* give an efficient algorithm, however, as it is easy to see that convergence can be arbitrarily slow if the ratios of the magnitudes of the eigenvalues of H_0 are close to 1. The role of the shifting strategy is to adaptively improve

[8] In the sense of backward error, see Definition 1.1.1 and Remark 1.1.2.

[9] A matrix H is (upper) Hessenberg if $H(i, j) = 0$ whenever $i > j + 1$. Such matrices are "almost" upper triangular.

these ratios and thereby accelerate convergence. The challenge is that this must be done efficiently without prior knowledge of the eigenvalues.

We quantify the rate of convergence of a sequence of iterates of (1.23) in terms of its *δ-decoupling time* $\mathsf{dec}_\delta(H_0)$, which is defined as the smallest t at which some subdiagonal entry of H_t satisfies

$$|H_t(i+1,i)| \leq \delta \|H_t\|.$$

In this context, "rapid" convergence means that $\mathsf{dec}_\delta(H_0)$ is a very slowly growing function of n and $1/\delta$, ideally logarithmic or polylogarithmic.

Remark 1.4.1 (Arithmetic Complexity from Decoupling Time)**.** The motivation for the particular measure of convergence above is that there is a procedure called *deflation* which zeroes out the smallest subdiagonal entry of a δ-decoupled Hessenberg matrix and obtains a nearby block upper triangular matrix, which allows one to pass to subproblems of smaller size incurring a backward error of $\delta\|H_0\|$. Repeating this procedure n times (and exploiting the special structure of Hessenberg matrices to compute the Q_t efficiently) yields an algorithm for computing a triangular T and unitary Q such that $\|H_0 - Q^*TQ\| \leq n\delta\|H_0\|$ in a total of $O(n^3\mathsf{dec}_\delta(H_0))$ arithmetic operations [170]. Thus, the interesting regime is to take $\delta \ll 1/n$.

In a celebrated work, Wilkinson [174] proved global convergence[10] of shifted QR on all *real symmetric* tridiagonal[11] matrices using the shifting strategy that now carries his name. The linear convergence bound $\mathsf{dec}_\delta(H_0) \leq O(\log(1/\delta))$ for this shifting strategy was then obtained by Dekker and Traub [52] (in the more general setting of Hermitian matrices), and reproven by Hoffman and Parlett [88] using different arguments. Other than these results for Hermitian matrices, there is no known bound on the worst-case decoupling time of shifted QR for any large class of matrices or any other shifting strategy.[12] Shifted QR is nonetheless the most commonly used algorithm in practice for the nonsymmetric eigenproblem on dense matrices. The strategies implemented in standard software libraries heuristically converge very rapidly on "typical" inputs, but occasionally examples of nonconvergence are found [47, 109] and dealt with in ad hoc ways.

Accordingly, the main theoretical question concerning shifted QR, which has remained open since the 1960s, is:

Question D. *Is there a shifting strategy for which the Hessenberg shifted QR iteration provably and rapidly decouples on nonsymmetric matrices?*

Question D was asked in various forms e.g. by Parlett [121, 124], Moler [110, 109], Demmel [55, Ch. 4], Higham [84, p. IV.10], and Smale [141] (who referred to it as a "great challenge").

[10] i.e., from any initial condition H_0.

[11] i.e., arising as the Hessenberg form of symmetric matrices.

[12] For nonnormal matrices, it is not even known if there is a shifting strategy which yields global convergence regardless of an effective bound on the decoupling time. A thorough discussion of related work will be given in Chapter 4.

Chapter 4 will be devoted a positive answer to Question D which is quantified in terms of the degree of nonnormality of the input matrix H_0. Moreover, we show, within the finite precision arithmetic model, that the shifted QR algorithm can be used to efficiently solve the backward eigenvalue problem. We do this in three steps:

1. *Decoupling in Exact Arithmetic.* Let $\mathbb{H}_B^{n\times n}$ be the set of diagonalizable complex Hessenberg matrices H_0 with eigenvector condition number $\kappa_V(H_0) \leq B$. In Section 4.1 we will exhibit a two parameter family of deterministic shifting strategies $\mathrm{Sh}_{k,B}$ indexed by a degree parameter $k = 2, 4, 8 \ldots$ and a nonnormality parameter $B \geq 1$ such that:

   (i) The strategy $\mathrm{Sh}_{k,B}$ satisfies $\mathsf{dec}_\delta(H_0) \leq O(\log(1/\delta))$ for every $H_0 \in \mathbb{H}_B^{n\times n}$ and $\delta > 0$.

   (ii) $\mathrm{Sh}_{k,B}$ has degree k and can be computed in roughly $O((\log k + B^{\frac{\log k}{k}})kn^2)$ arithmetic operations, which is simply $O(n^2 k \log k)$ for the judicious setting $k = \Omega(\log B \log\log B)$.

   Thus, the computational cost of the shifting strategy required for convergence blows up as the input matrix becomes more and more nonnormal, but the dependence on nonnormality is very mild.

   We remark that such a result was not previously known even in the case $B = 1$ of normal matrices. Further, in the spirit of smoothed analysis and the discussion from Section 1.2, a tiny random perturbation of any $H_0 \in \mathbb{H}^{n\times n}$ is likely to be an element of $\mathbb{H}_B^{n\times n}$ for small B (not depending on H_0). Thus, while our theorem does not give a single shifting strategy which works for all matrices, it does give a strategy which works for a tiny random perturbation of every matrix (with high probability, where "tiny" and "small" must be quantified appropriately).

2. *Numerical Stability.* With the analysis of the dynamics in exact arithmetic in hand, in Section 4.2 we will show that both the correctness and rapid convergence of these strategies continue to hold in finite arithmetic with an appropriate implementation, and prove a bound on the number of bits of precision needed, for matrices with controlled condition number κ_{eig}. To do so, we develop some general tools enabling rigorous finite arithmetic analysis of the shifted QR iteration with any shifting strategy which uses Ritz values as shifts, of which $\mathsf{Sh}_{k,B}$ (which will be used to denote the finite arithmetic implementation of $\mathrm{Sh}_{k,\mathcal{B}}$) is a special case.

   The main challenge here is the forward instability of QR steps as the Ritz values start to converge towards the eigenvalues of the input matrix. This phenomenon heavily complicates the dynamics analysis in the finite arithmetic model, since it becomes hard to reason about the computed iterates when decoupling is close to taking place. To solve this, inspired by the work of Parlett and Le [126] and in the spirit of aggressive early deflation [31, 32], we show a dichotomy: either each QR step of our shifting strategy is (forward) stable enough for our analysis to be valid or the information provided by Ritz values in such step can be used to decouple the problem immediately.

3. *Ritz Value Finder with Provable Guarantees.* The *Ritz values of order k* of an upper Hessenberg matrix H are equal to the eigenvalues of its bottom right $k \times k$ corner $H_{(k)}$; they are also defined variationally as the zeros of the monic degree k polynomial p_k minimizing $\|e_n^* p_k(H)\|$, where e_n is an elementary basis vector. All of the higher order shifting strategies we are aware of are defined in terms of these Ritz values. However, we are not aware of any theoretical analysis of how to compute the Ritz values (approximately) in the case of nonsymmetric $H_{(k)}$, nor a theoretical treatment of which notion of approximation is appropriate for their use in the shifted QR iteration.[13] In Section 4.3 we will give a method, based on inverse iteration, for provably computing the eigenvalues of any small matrix.

Below we detail our main technical results regarding each of the steps discussed above. Comprehensive discussions about proof techniques, difficulties, and related literature will be deferred to Chapter 4.

Decoupling in Exact Arithmetic

Since computing eigenvalues exactly is impossible when $k \geq 5$, even when working in the framework of exact arithmetic, we only assume access to a method for computing approximate Ritz values, in the sense encapsulated in the following definition.

Definition 1.4.2 (θ-Optimal Ritz values and Ritz value finders)**.** Let $\theta \geq 1$. We call $\mathcal{R} = \{r_1, \ldots, r_k\} \subset \mathbb{C}$ a set of *θ-optimal Ritz values* of a Hessenberg matrix H if

$$\left\| e_n^* \prod_{i \leq k} (H - r_i) \right\|^{1/k} \leq \theta \min_{p \in \mathcal{P}_k} \|e_n^* p(H)\|^{1/k}. \tag{1.24}$$

A *Ritz value finder* is an algorithm $\mathrm{OptRitz}(H, k, \theta)$ that takes as inputs a Hessenberg matrix $H \in \mathbb{C}^{n \times n}$, a positive integer k and an accuracy parameter $\theta > 1$, and outputs a set $\mathcal{R} = \{r_1, \ldots, r_k\}$ of θ-optimal Ritz values of H whenever the right hand side of (1.24) is nonzero. Let $T_{\mathrm{OptRitz}}(k, \theta, \delta)$ be the maximum number of arithmetic operations used by $\mathrm{OptRitz}(H, k, \theta)$ over all inputs H such that the right hand side of (1.24) satisfies[14]

$$\min_{p \in \mathcal{P}_k} \|e_n^* p(H)\|^{1/k} \geq \delta \|H\|.$$

A Ritz value finder satisfying Definition 1.4.2 can be efficiently instantiated using polynomial root finders (e.g. [119]) or other provable eigenvalue computation algorithms (e.g. [11, 14]) with guarantees of type $T_{\mathrm{OptRitz}}(\theta, k, \delta) = O(k^c \log(\frac{1}{\delta(\theta-1)}))$. We defer a detailed

[13] In practice, and in the current version of LAPACK, the prescription is to run the shifted QR algorithm itself on $H_{(k)}$, but there are no proven guarantees for this approach.

[14] Such a lower bound is needed, since otherwise we could use OptRitz to compute the eigenvalues of $H_{(k)}$ to arbitrary accuracy in finite time.

discussion of numerical issues surrounding this implementation to Sections 4.2 and 4.3. The subtlety of not being able to compute Ritz values exactly is secondary to the dynamical phenomena which are the focus of our first step of the analysis of the QR algorithm. So on first reading of Section 4.1 it is recommended to assume $\theta = 1$ (i.e., Ritz values are computed exactly), even though this is unrealistic when $k > 4$. The theorem below is stated with $\theta = 2$, which is also the parameter setting that will be used in Section 4.2.

Our main theorem regarding the dynamics of the algorithm is the following.[15]

Theorem 1.4.3. *There is a family of deterministic shifting strategies* $\mathrm{Sh}_{k,B}$ *(described in Section 4.1.6) parameterized by degree* $k = 2, 4, 8, \ldots$ *and nonnormality bound* $B \geq 1$ *with the following properties.*

1. *(Rapid Decoupling) If* $H_0 \in \mathbb{H}_B^{n\times n}$ *, then, assuming exact arithmetic, for every* $\delta > 0$*, the QR iteration with strategy* $\mathrm{Sh}_{k,B}$ *satisfies*

$$\mathsf{dec}_\delta(H_0) \leq 4 \lg(1/\delta). \tag{1.25}$$

2. *(Cost Per Iteration Before Decoupling) Given a Ritz value finder* $\mathrm{OptRitz}(H, k, \theta)$ *with complexity* $T_{\mathrm{OptRitz}}(k, \theta, \delta)$*, an accuracy parameter* $\delta > 0$*, and a Hessenberg matrix* $H_t \in \mathbb{H}_B^{n\times n}$*, computing* H_{t+1} *given* H_t *has a cost per iteration of at most*

$$\left(\lg k + N_{\mathsf{net}}\left(0.002\, B^{-\frac{8\lg k+4}{k-1}}\right)\right) \cdot T_{\mathsf{lQR}}(k, n) + T_{\mathrm{OptRitz}}(k, 2, \delta) + \lg k \tag{1.26}$$

arithmetic operations for all iterations before (1.25) is satisfied, where $N_{\mathsf{net}}(\epsilon) = O(\epsilon^{-2})$ *is number of points in an efficiently computable* ϵ*-net of the unit disk and* $T_{\mathsf{lQR}}(k, n) \leq 7kn^2$ *is an upper bound on the arithmetic cost of a degree* k *implicit QR step (see Section 4.1.3).*

The term involving N_{net} captures the the cost of performing certain "exceptional shifts" (see Section 4.1.8) used in the strategy. The tradeoff between the nonnormality of the input matrix and the efficiency of the shifting strategy appears in the cost of the exceptional shift, where it is seen that setting

$$k = \Omega(\log B \log\log B) \tag{1.27}$$

yields a total running time of $O(n^2 k \log k)$ operations per iteration. Note that the bound $B \geq \kappa_V(H_0)$ must be known in advance in order to determine how large a k is needed to make the cost of the exceptional shift small. One may also take k to be a constant independent of B, but this causes the arithmetic complexity of each iteration to depend polynomially on B rather than logarithmically. Note that for normal matrices one may take $k = 2$ and $B = 1$.

Remark 1.4.4 (Higher Degree Shifts)**.** A QR step with a degree k shift

$$p(z) = (z - r_1) \ldots (z - r_k)$$

[15] All logarithms are base 2.

is identical to a sequence of k steps with degree 1 shifts $(z-r_1),(z-r_2),\ldots,(z-r_k)$ (see e.g. [170] for a proof), so any degree k strategy can be simulated by a degree 1 strategy while increasing the iteration count by a factor of k.[16] We choose to present our strategy as higher degree for conceptual clarity. The efficiency of using degrees as high as $k = 180$ has been tested in the past [31, Section 3] and $k = 50$ is often used in practice [96].

Numerical Stability

Theorem 1.4.3 above (which assumes exact arithmetic) only gives a description of the algorithm up to when decoupling takes place, but this is not sufficient (even in exact arithmetic) to obtain running times and accuracy guarantees when the full eigendecomposition of the input matrix is desired. This due to our assumption of an upper bound B on the eigenvector condition number of the input matrix, which does not necessarily bound the eigenvector condition number of the submatrices that appear in the recursion after deflation.

In Section 4.2 we will take into consideration numerical errors, and factor into our analysis the deflation step. To do this, we will assume access to a black box algorithm SmallEig, which will be called upon when approximate Ritz values are needed, with the following guarantee on a matrix A of dimension k or smaller. (The notion of forward error here is absolute, instead of relative — this will simplify some of the analysis later on).

Definition 1.4.5. A *small eigenvalue solver* $\mathsf{SmallEig}(A,\beta,\phi)$ takes as input a matrix A of size at most $k\times k$, and with probability at least $1-\phi$, outputs $\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_k\in\mathbb{C}$ such that $|\widetilde{\lambda}_i-\lambda_i|\leq\beta$ for each of $\lambda_1,\ldots,\lambda_k\in\mathrm{Spec}(A)$.

Our main result in this direction will be the following.

Theorem 1.4.6. *Let H be an $n\times n$ upper Hessenberg matrix and $B\geq 2\kappa_V(H)$ and $\Gamma\leq \mathrm{gap}(H)/2$ upper and lowerbounds on its eigenvector condition number and minimum eigenvalue gap. For a certain $k=O(\log B\log\log B)$ — which will be specified in Section 4.2 — the shifting strategy $\mathrm{Sh}_{k,B}$ can be implemented in finite arithmetic to give a randomized shifted QR algorithm, ShiftedQR, with the following guarantee: for any $\delta>0$ $\mathsf{ShiftedQR}(H,\delta,\phi)$ produces the eigenvalues of a matrix H' with $\|H-H'\|\leq\delta\|H\|$, with probability at least $1-\phi$, using*

- $O\left(n^3\left(\log\frac{nB}{\delta\Gamma}\cdot k\log k+k^2\right)\right)$ *arithmetic operations on a floating point machine with* $O\left(k\log\frac{nB}{\delta\Gamma\phi}\right)$ *bits of precision; and*
- $O(n\log\frac{nB}{\delta\Gamma})$ *calls to* SmallEig *with accuracy* $\Omega(\frac{\delta^2\Gamma^2}{n^4B^4\Sigma})$ *and failure probability tolerance* $\Omega\left(\frac{\phi}{n^2\log\frac{nB}{\delta\Gamma}}\right)$

[16] This also has some important advantages with regards to numerical stability, which are discussed in [13].

Remark 1.4.7 (Constants)**.** The constants on arithmetic operations and precision hidden in the asymptotic notation above are modest and can be read off by unpacking the expressions for $T_{\mathsf{ShiftedQR}}$ in equation (4.55) and $\mathbf{u}_{\mathsf{ShiftedQR}}$ in equation (4.54), respectively.

Remark 1.4.8 (Computing Eigenvalues of an Arbitrary Matrix)**.** The algorithm ShiftedQR can be used to compute backward approximations of the eigenvalues of an arbitrary matrix $A \in \mathbb{C}^{n\times n}$ with a backward error of $\delta\|A\|$ as follows:

1. Add a random complex Gaussian perturbation of norm $\delta\|A\|/2$ to the input matrix, which yields $\log(B/\Gamma) = O(\log(n/\delta))$ with high probability (this follows from the discussion in Section 1.2).

2. Put the resulting matrix in Hessenberg form using Householder reflectors. This step is backward stable when performed in finite arithmetic [155], and thus approximately preserves the bounds on B, Γ by the results that will be proven in Section 4.2.

3. Apply Theorem 1.4.6 with accuracy $\delta/2$, noting that the bound on $\log(B/\Gamma)$ from step 1 implies that $k = O(\log(n/\delta)\log\log(n/\delta))$ is sufficient.

Remark 1.4.9 (Hermitian Matrices)**.** For the important case of Hermitian tridiagonal matrices there is no difficulty in maintaining $\kappa_V(H) = 1$, so we may take $k = 2$ and $B = 1$. A minimum eigenvalue gap of $\Gamma \geq (\delta/n)^c$ may be guaranteed by adding a diagonal Gaussian perturbation of size $\delta/2$ [2] to the matrix (or by adding a GUE perturbation and then tridiagonalizing the matrix). The Ritz values in this case can be computed to sufficient accuracy using the quadratic formula. The amount of precision required by Theorem 1.4.6 is consequently simply $O(\log(n/\delta))$ and the number of arithmetic operations used is $O(n^3 + n^2\log(n/\delta))$, which is asymptotically the same as in the exact arithmetic analysis of tridiagonal QR with Wilkinson shift.

Ritz Values via Inverse Shifted Iteration

In Section 4.3 we will provide an algorithm SmallEig based on shifted inverse iteration. For this algorithm we will show the following.

Theorem 1.4.10. *Given $A \in \mathbb{C}^{k\times k}$ with $\|A\| \leq \Sigma$, there is an algorithm,* SmallEig*, which solves the forward eigenvalue problem in the sense of Definition 1.4.5, using at most*

$$O\big(k^5\log(k\Sigma/\beta\phi)^2 + k^2\log(k\Sigma/\beta\phi)^2\log(k\log(k\Sigma/\beta\phi))\big)$$

arithmetic operations on a floating point machine with $O(k^2\log(k\Sigma/\beta\phi)^2)$ bits of precision.

Note that the algorithm SmallEig uses higher precision than we require anywhere else in this analysis, but because it is called infrequently and on $k \times k$ matrices only, the total Boolean operations are still subdominant.

Remark 1.4.11 (Final running times)**.** This yields a total worst-case complexity bound of $O(n^3 \log^2(n/\delta)(\log\log(n/\delta))^2)$ arithmetic operations with $O(\log^2(n/\delta)\log\log(n/\delta))$ bits of precision *plus* $O(n\log(n/\delta)\cdot\log^7(n/\delta)\log\log(n/\delta)^5)$ operations with $O(\log^4(n/\delta)(\log\log(n/\delta)^2))$ bits of precision for the calls to SmallEig. The Boolean cost of calls to SmallEig is subdominant whenever $n \geq \log^{7/2}(n/\delta)(\log\log(n/\delta))^2$.

While this asymptotic complexity guaranteed by Remark 1.4.11 is significantly higher than the nearly matrix multiplication time spectral bisection algorithm discussed in Chapter 3, that algorithm uses $O(\log^4(n/\delta)\log(n))$ bits of precision, moreover with a larger hidden constant. On the other hand, the algorithm of [3] uses $O(n^{10}/\delta^2)$ arithmetic operations but with only $O(\log(n/\delta))$ bits of precision (as is stated but not formally proven in [3]).

1.5 Chapter 5: The Lanczos Algorithm Under Few Iterations

The Lancozs algorithm is one of the most widely used numerical methods for solving problems pertaining to large Hermitian matrices. In particular, it is invoked in applications that only require knowing specific features of the spectrum of a very large Hermitian matrix $A \in \mathbb{C}^{n\times n}$, and where computing the full eigendecomposition of A would be wasteful and in occasions prohibitively expensive (both in terms of storage and running time). Although it can be viewed as an iterative method for approximating the eigenvalues of a matrix, the Lanczos algorithm is fundamentally different than the two algorithms discussed above (spectral bisection and QR iteration), and in the context of this dissertation it may be better thought of as a dimensional reduction technique.

The Lanczos Algorithm. We will think of the Lanczos algorithm as an algorithm that receives three inputs: a matrix $A \in \mathbb{C}^{n\times n}$, a vector $u \in \mathbb{C}^n$, and an integer $k \in [n]$. Given these inputs, the procedure runs for k iterations, each of which utilizes only matrix-vector and vector-vector multiplications, and when terminated outputs a $k \times k$ tridiagonal matrix J called the *Jacobi matrix*[17], we refer the reader to Section 5.1 for details.

The nontrivial entries of the the Jacobi matrix are called the *Jacobi coefficients* of the matrix; the diagonal ones will be denoted by α_i and the off-diagonal ones by β_i (sometimes $\alpha_i(u)$ and $\beta_i(u)$ when it is important to emphasize the u-dependence), its eigenvalues are known as the k *Ritz values* of A, which we denote by r_i (similarly sometimes we will use $r_i(u)$), and its eigenvectors can be used to compute the so called *Ritz vectors*, but we will only treat the latter in passing.

If k is set to $k = n$ the Jacobi matrix will be unitarily similar to the input A and therefore the Ritz values will be precisely the eigenvalues of A. However, the use case for this algorithm

[17]Technically speaking this $k \times k$ matrix is in fact the $k \times k$ corner of the actual Jacobi matrix of the spectral measure of A associated to u, but for simplicity, when k is fixed we will refer to J as the Jacobi matrix.

is when n is large and $k \ll n$, and therefore computing the full eigendecomposition of the Jacobi matrix[18] (i.e. computing the Ritz values and Ritz vectors) can be done quickly and accurately.

Approximating outlying eigenvalues using Ritz values. The success of the Lanczos algorithm resides to some extent in its ability to find (by looking at the Ritz values) the *outliers* of the spectrum of the matrix A using very few iterations. By outliers, we mean the eigenvalues distant from the region in which the majority of the spectrum accumulates (the *bulk*). As the reader may be well aware of, there are several applications, both in science and engineering, where the value of these outlying eigenvalues is extremely informative.

Understanding the bulk of the spectrum using Jacobi coefficients. Lanczos-type methods can also be used to approximate the global spectral density of large matrices, also known as density of states; for a survey of techniques see [103]. In applied mathematics, large matrices can arise as discretizations of infinite-dimensional operators such as the Laplacian or as finite-dimensional representations of an infinite-dimensional Hamiltonian. Computing the eigenvalues and Jacobi coefficients of the finite-dimensional operator then yields information about the infinite-dimensional operator and the underlying continuous system. For an example, see [138], or Section 7 of [161] for numerical experiments and bounds for the Lanczos algorithm applied to an explicit discretized Laplace operator.

In the setting described above, the Jacobi coefficients contain all the information of the spectral density of the infinite-dimensional operator in question and even the first few coefficients are of use. To give an example, in [83] the Haydock method (as it is termed today) was introduced. This method exploits the fact the resolvent of an operator admits a continued fraction expansion where the coefficients are precisely the Jacobi coefficients, and hence knowing these quantities is fundamental to understanding the spectral density of the operator—see [103, Section 3.2.2] for a summary of the Haydock method.

Using a slightly different perspective, note that from the $k \times k$ Jacobi matrix of an operator one can obtain the $[k-1, k]$ Padé approximation of its resolvent [160]. In particular, knowing the $k \times k$ Jacobi matrix is enough to compute the first $2k-1$ moments of the spectral density of the infinite-dimensional operator.

Main results

Often times, the initial vector $u \in \mathbb{C}^n$ that is fed into the Lanczos algorithm as an input is taken uniformly at random from the sphere $\mathbb{S}^{n-1}$. In this case, when the algorithm is run for $k \ll n$ iterations, there are two non-trivial fundamental questions that arise:

1. How much does the (random) output vary?

2. How many iterations are necessary and sufficient to obtain a satisfactory approximation for the problem in question?

[18] This is usually done by running the QR iteration with Wilkinson's shift on the Jacobi matrix.

Regarding the first question, one should note that two independent vectors drawn from $\mathbb{S}^{n-1}$ might be quite different (and in fact will be close to orthogonal if n is large), and therefore when run more than once on the same matrix, the Lanczos algorithm could (a priori) output two very distinct outputs. This inherent feature of randomness hurts reproducibility and poses a concern when safety relies on the accuracy and consistency of the algorithm. To mitigate this potential issue, sometimes practitioners run the Lanczos algorithm several times (with independent initial vectors) on the same matrix, and report as final output the average of these preliminary outputs. In this regard, the first question is also relevant to efficiency, since it can inform practice when deciding how many samples are needed to ensure reproducibility.

The second question is relevant to large scale problems, where each extra iteration is significantly costly and one wants to find the minimum k for which k iterations will suffice to solve the problem in hand.

Assumptions and complications. One of the main complications that arises when addressing the questions posed above, is that the behavior of the Jacobi coefficients and Ritz values varies substantially according to the properties of the spectrum of the input matrix A. It is therefore challenging to show general results about arbitrary inputs A, and such results most necessarily be quantified, to some extent, in terms of certain features of the spectrum of the input matrix. For example, in the influential paper of Saad [135], many of the results about rates of convergence for the Ritz values were stated in terms of the (not very explicit) quantities

$$t_i^{(k)} = \min_{p \in \mathcal{P}_{k-1}^{(i)}} \max_{j:j\neq i} |p(\lambda_j)|,$$

where $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues of A and $\mathcal{P}_{k-1}^{(i)}$ denotes the set of all polynomials of degree not exceeding $k-1$ and satisfying $p(\lambda_i) = 1$. In the present work we use a more geometric notion about the spectrum to state our results.

Definition 1.5.1 (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

$$\left| \left\{ \lambda \in \Lambda : \frac{1}{|T|} \sum_{t \in T} \log |\lambda - t| \geq \log \omega \right\} \right| \geq \delta n.$$

Intuitively, the spectrum of the input matrix A is equidistributed if it is not grouped in a small number of tight clusters (see Examples 1.5.2 and 1.5.3 below). As we will show in Section 5.3.1, the family of well equidistributed point sets includes, but is not limited to, those sets obtained by discretizing an absolutely continuous distribution.

Example 1.5.2. 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 5.3.1 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 1.5.3. 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 we give a generalization of this assertion in Observation 5.3.9.

Finally, since there are no numerical stability concerns for the questions we will be addressing, throughout the discussion of this work we will assume exact arithmetic.

Concentration of the output. In Chapter 5 we will show that when the spectrum of the input matrix $A \in \mathbb{C}^{n \times n}$ is reasonably equidistributed, and the Lanczos algorithm is run for a few iterations, the output is exponentially concentrated (see Figure 1.2). To be more precise, in this direction our main result will be the following.

Theorem 1.5.4 (Concentration of Jacobi coefficients after i iterations)**.** *Assume the initial vector u is sampled uniformly at random from $\mathbb{S}^{n-1}$, and assume 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 probabilities $\mathbb{P}[|\alpha_i(u) - \tilde{\alpha}_i| > t\|A\|]$ and $\mathbb{P}[|\beta_i(u) - \tilde{\beta}_i| > t\|A\|]$ are both bounded above by*

$$2 \exp\left\{ -\frac{\min\{\delta, 1/50\}^2}{32} n \right\} + 2 \exp\left\{ -\frac{1}{64} \left(\frac{\omega}{4\|A\|} \right)^{2i} \delta^2 t^2 n \right\}. \tag{1.28}$$

Remark 1.5.5. The equidistribution parameters 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. Note that $\omega \leq \|A\|$ (by taking $T = \{0\}$ in Definition 1.5.1) and that ω scales linearly with A. As a result, $\omega/\|A\|$ is typically of constant size independent of n in applications. Since $\omega/\|A\| < 1$, Theorem 1.5.4 yields concentration for i at most logarithmic in n.

Remark 1.5.6 (Ritz values)**.** Using Weyl's inequality and the Davis-Kahan theorem (see Lemmas 1.1.5 and 1.1.6 above) Theorem 1.5.4 can be used to easily obtain concentration results for the Ritz values and Ritz eigenvectors. We defer this discussion to Sections 5.3.3 and 5.3.4.

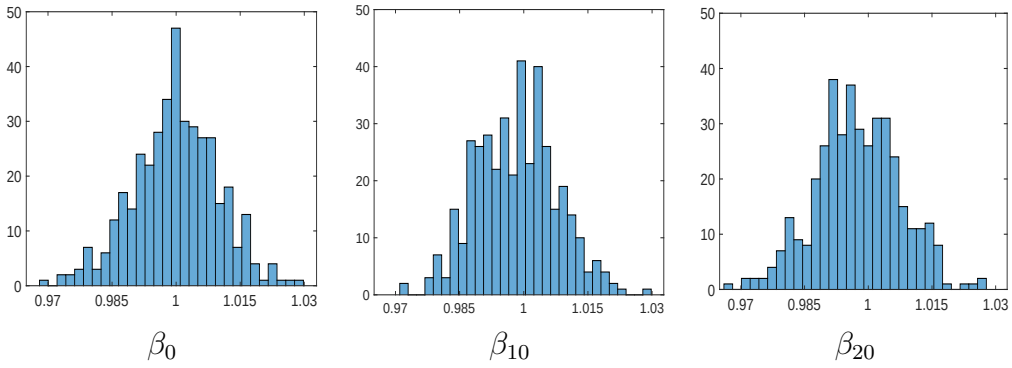

Figure 1.2: Here A is a fixed $n \times n$ matrix drawn from the Gaussian orthogonal ensemble (GOE) with $n = 5000$. Since the empirical spectral distribution of A will be close to the semicircular law it is expected that the Jacobi coefficients β_i of A will be approximately 1. The above histograms show the values β_0, β_{10}, and β_{20} obtained by running the Lanczos algorithm 400 times on the input A. Note that in each of these cases, β_i appears to be concentrated.

To the best of our knowledge, there is no previous work studying the concentration properties of the output of the Lanczos algorithm when the initial vector is taken at random.

Undetected outliers. In Section 5.4 we will 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, and hence Lanczos fails to detect outlying eigenvalues if only k iterations are performed. This complements classical guarantees which show that for some other 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.

Theorem 1.5.7. *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 $g = \max_{1 \leq i \leq n}\{\lambda_i - R\}$ 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 convenience we include Example 1.5.8 below (see also an asymptotic version of the above result, proven in Section 5.4).

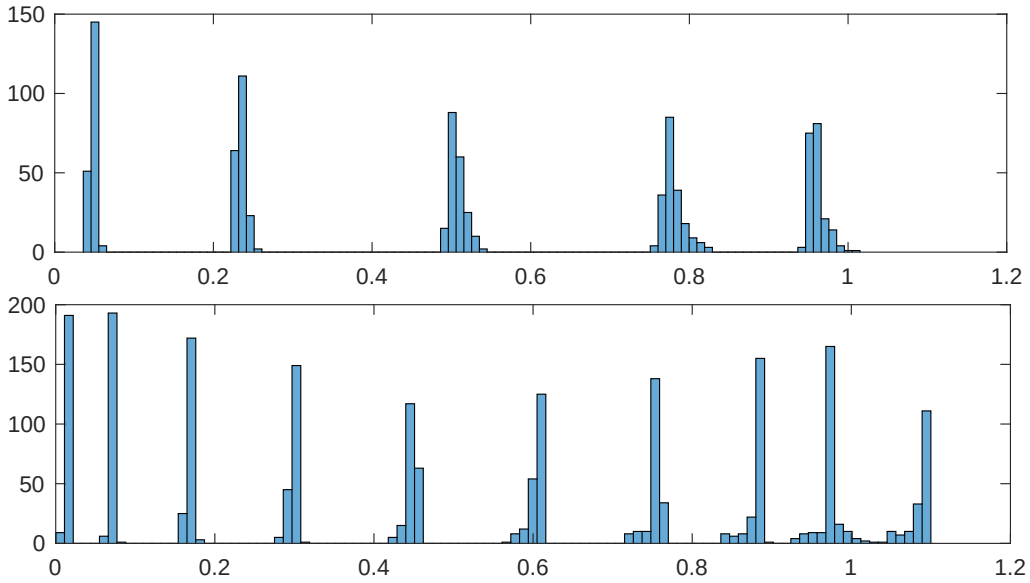

Figure 1.3: A is a 2000×2000 diagonal matrix with entries $\{0, 1/2000, 2/2000, \ldots, 1999/2000, 1.1\}$. This represents a discretization of Unif($[0, 1]$) plus an outlier at 1.1. Plotted is a histogram of the Ritz values output by Lanczos after $k = 5$ iterations (above) and after $k = 10$ iterations (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). However, even in the regime of $k = 5$ iterations the output appears to be concentrated.

Example 1.5.8. 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 1.3). In Section 5.3.1 we will show that for $j \leq n/16$ the spectrum of A is $(1/4, 4e^{-2}, j)$-equidistributed. In order to apply Theorem 1.5.7, 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.

To put Theorem 1.5.7 into context, we note that a lot has been written on the location of Ritz values (as a function of the input vector and input matrix). However, most of this literature is devoted to providing an *upper bound* on the number of iterations required to obtain an accurate approximation of outlying eigenvalues (see [94, 118, 135]). Roughly speaking, previous literature provides inequalities that state that $k \geq C \log n$ iterations suffice for the Lanczos algorithm to approximate the true extreme eigenvalues of the input matrix $A \in \mathbb{C}^{n \times n}$ very well, making the use of $O(\log n)$ iterations common in practice (see [100] or [161] 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 [177] for an example or [24] for a survey.

Regarding negative results, i.e. *lower bounds* on the number of iterations needed to detect outliers, the only work we are aware of is [140], where 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.

Asymptotic location of the output. For this discussion the notion of empirical spectral distribution will be relevant. Given $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, the empirical spectral distribution of A is defined to be the probability measure

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

We will work in the following setting: we consider a probability measure μ and a sequence of matrices A_n whose empirical spectral distributions μ_n converge to μ, and 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 1.5.9 (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 $c > 0$.*[19]

Then there is a constant $d > 0$ dependent on μ and c, such that for any sequence of integers $1 \le k_n \le d\sqrt{\log n}$ 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 $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$[20] *after k_n iterations, $J_{k_n}(\mu)$ is the k_n-th Jacobi matrix of the measure μ, and $\longrightarrow_P$ denotes convergence in probability.*

Note that Theorem 1.5.9 may be of particular relevance in applications where an infinite-dimensional operator is discretized with the goal of computing its density. In essence, Theorem 1.5.9 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 (see Figure 1.4).

[19] Hereafter, for two probability measures ν_1 and ν_2, we will use $\mathrm{Kol}(\nu_1, \nu_2)$ to denote the Kolmogorov-Smirnov distance between the two measures.

[20] From now on we will use $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ to denote that u is chosen uniformly at random from $\mathbb{S}^{n-1}$.

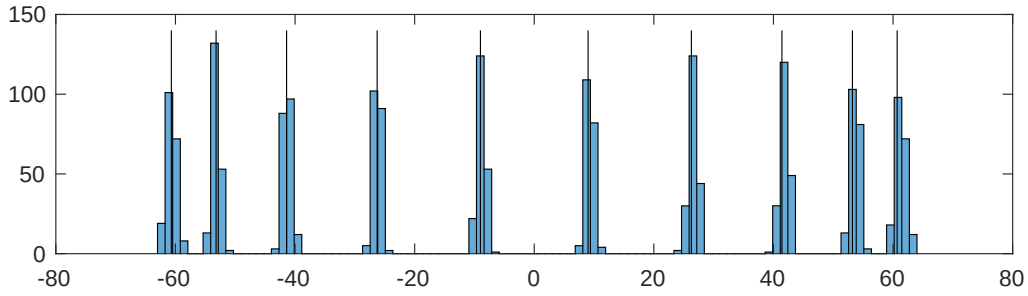

Figure 1.4: A is a fixed $n \times n$ matrix drawn from the GOE with $n = 2000$. Plotted is the histogram of the Ritz values after 200 repetitions of the Lanczos algorithm with $k = 10$ iterations. Also plotted are the roots of the 10th orthogonal polynomial with respect to the (suitably rescaled) semicircle law, which is the limit of the distribution of eigenvalues for GOE as $n \to \infty$.

Proposition 1.5.10 (Location of the Ritz values)**.** *Using the same notation as in Theorem 1.5.9, let $\vec{r}_{k_n}(u) = (r_1(u), \dots, r_{k_n}(u))$, 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 1.5.9, 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.

Proof techniques

The main tool for showing Theorem 1.5.4 is Levy's concentration lemma, which roughly speaking states that if $f : \mathbb{S}^{n-1} \to \mathbb{R}$ is a Lipschitz function and $u \sim \text{Unif}(\mathbb{S}^{n-1})$, then $f(u)$ is exponentially (in n) concentrated around its median. In view of this, to prove Theorem 1.5.4 it would be enough to show that, on a fixed input matrix A, the Jacobi coefficients are Lipschitz when viewed as the functions $u \mapsto \alpha_i(u)$ and $u \mapsto \beta_i(u)$ on $\mathbb{S}^{n-1}$. However, the functions $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ can have singularity points close to which their modulus of continuity blows up, and therefore this naive approach has no chance of succeeding. Instead, we will have to recur to the following refinement of Levy's lemma, which allows one to obtain concentration results by arguing that the function in question has a controlled Lipschitz constant on "most" of the sphere.

Lemma 1.5.11 (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 (with respect to the geodesic metric on the sphere). 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 $u \sim \text{Unif}(\mathbb{S}^{n-1})$.

With this lemma in hand, our proof uses elements from the theory of orthogonal polynomials to control the local Lipschitz constant of the functions $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ at any given point in $\mathbb{S}^{n-1}$. Then, by appealing to basic probability theory arguments, we argue that the set of "bad points" in $\mathbb{S}^{n-1}$ (i.e. the points for which the Jacobi coefficients have a large local Lipschitz constant) have small measure, and by studying the geometry of this bad region we then argue that locally controlling the Lipschitz constant is enough to ensure a global control of the Lipschitz constant on the full "good region".

The techniques from the theory of orthogonal polynomials mentioned above turn out to be powerful enough that they also allow to understand the locations of the (medians) of the Jacobi coefficients and Ritz values, and therefore allow us to also prove Theorems 1.5.7 and 1.5.9.

1.6 Overview: Mechanisms and Key Phenomena

From a technical perspective, the proofs of the main results in this dissertation are relatively sophisticated, and in some cases (specially when dealing with round off errors) they are also intricate. However, the underlying mechanisms and techniques are conceptually simple and insightful, so we thought worth it to highlight them in an abstracted explicit manner.

1.6.1 Spectral Measures and the Functional Calculus

In noncommutative probability (which was inspired by quantum mechanics) it is common to regard linear operators T on a Hilbert space $\mathcal{H}$ as (noncommutative) random variables. In the particular case when T is bounded and normal, and $u \in \mathcal{H}$ with $\|u\| = 1$, classical machinery from functional analysis allows one to define a probability distribution supported on $\mathrm{Spec}(T)$, which we will denote by $\mu_{T,u}$, and refer to as the spectral measure of T associated to u. Moreover, one can define, via the functional calculus, a bounded normal operator $f(T)$ for any continuous function $f : \mathrm{Spec}(T) \to \mathbb{C}$, which satisfies (among other things)

$$\langle u, f(T)u \rangle = \int_{\mathbb{C}} f(x) d\mu_{T,u}(x),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathcal{H}$. Therefore, once u is chosen, one can identify T with a (classical) random variable X_T distributed as $\mu_{T,u}$, and rewrite the above as

$$\mathbb{E}[f(X_T)] = \langle u, f(T)u \rangle.$$

When $\mathcal{H}$ is finite dimensional these notions are very concrete: if $A \in \mathbb{C}^{n \times n}$ is normal with spectral decomposition $A = \sum_{i=1}^{n} \lambda_i v_i v_i^*$ and $u \in \mathbb{C}$ with $\|u\| = 1$ then

$$\mu_{A,u}(\lambda_i) = |u^* v_i|^2$$

is clearly a probability measure (since u has unit norm and the v_i are orthogonal), and one can easily see that

$$u^* f(A) u = \sum_{i=1}^{n} f(\lambda_i)|u^* v_i|^2 = \int_{\mathbb{C}} f(x) d\mu_{A,u}(x),$$

for any $f : \mathrm{Spec}(A) \to \mathbb{C}$.

Now, for bounded nonnormal operators T one no longer has the notion of spectral measure, but as explained in Section 1.1 there is still the notion of holomorphic functional calculus. In some instances this will suffice for our algorithm analysis, and in some others we will need to find a replacement for the notion of spectral measure.

The Lanczos algorithm and orthogonal polynomials. In Chapter 5, to analyze the Lanczos algorithm on the input matrix $A \in \mathbb{C}^{n\times n}$ and starting vector $u \in \mathbb{C}^n$, we adopt the above perspective and consider the spectral measure $\mu_{A,u}$. It is a well-known fact that the Jacobi coefficients $\alpha_i(u)$ and $\beta_i(u)$ generated by the Lanczos algorithm are the terms appearing in the three-term recursion (cf. Favard's theorem) that defines the sequence of orthogonal polynomials $\pi_0^u(x), \pi_1^u(x), \ldots, \pi_n^u(x)$ with respect to $\mu_{A,u}$. Moreover, the roots of $\pi_k^u(x)$ are precisely the Ritz values obtained after k iterations of the Lanczos algorithm. In Chapter 5 we exploit these connections in the case when $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ is random, to analyze the (random) output of the Lanczos algorithm in terms of features of the (random) measure $\mu_{A,u}$.

Spectral measures of normal Hessenberg matrices. Given that we have just discussed the Lanczos algorithm (which assumes the input to be Hermitian) in the context of orthogonal polynomials, let us extend this discussion to normal matrices, since this is relevant to our analysis of the QR algorithm. As mentioned in Section 1.5, when the Lanczos algorithm is run on a self-adjoint matrix $A \in \mathbb{C}^{n\times n}$ and starting vector $u \in \mathbb{C}^n$ for n iterations, one obtains the $n \times n$ Jacobi matrix J, which is a symmetric tridiagonal matrix that is unitarily equivalent to A. Furthermore, it is not hard to see that $\mu_{J,e_n} = \mu_{A,u}$, where $e_n = (0, \ldots, 0, 1)$[21]. In the case when A is normal (but not self-adjoint), and one uses the Arnoldi algorithm[22] instead, the output after n iterations will now be an upper Hessenberg matrix H that is unitarily equivalent to A. Once again, the entries of H will correspond to the recursion coefficients for the orthogonal polynomials $\pi_k(z)$ with respect to $\mu_{A,u}$ (the recurrence is no longer a three-term recurrence), and $\mu_{A,u} = \mu_{H,e_n}$. Moreover, the quantity

$$\psi_k(H) := |h_{n-k,n-k-1} \cdots h_{n,n-1}|^{1/k}$$

which we will use to measure progress in the shifted Hessenberg QR algorithm, can be expressed as $\psi_k(H) = \|\pi_k(z)\|^{1/k}_{L^2(\mu_{H,e_n})}$. Note that from this perspective, it is natural to turn

[21] Depending on the implementation one either has that $\mu_{A,u} = \mu_{J,e_1}$ or $\mu_{A,u} = \mu_{J,e_n}$. Here we have in mind the implementation that produces the matrix J from bottom to top (i.e. $J(n,n) = \alpha_0$ and $\beta_1 = J(n, n-1) = J(n-1, n)$) and this corresponds to $\mu_{A,u} = \mu_{J,e_n}$.

[22] The non-Hermitian version of the Lanczos algorithm.

to $\psi_k(H)$ as a measure of progress whenever the shifting strategy uses Ritz values, since when $\psi_k(H) = 0$ we get that $d\mu_{H,e_n}$ is supported on k eigenvalues of H, which can be seen (from $\|\pi_k(z)\|_{L^2(\mu_{H,e_n})} = 0$) to also be the roots of $\pi_k(z)$, i.e. the Ritz values of H obtained after k iterations of the Arnoldi procedure.

The nonnormal Hessemberg QR algorithm and the approximate functional calculus. Even if the notion of spectral measure is not defined for nonnormal matrices, one can use the following substitute motivated by the above discussion. For $H \in \mathbb{C}^{n\times n}$ a diagonalizable upper Hessenberg matrix, write $H = VDV^{-1}$ with V chosen[23] so that $\|V\| = \|V^{-1}\| = \sqrt{\kappa_V(H)}$ and D to be a diagonal matrix with $D_{i,i} = \lambda_i$ being the i-th eigenvalue of H. Then let Z_H be the random variable supported on $\mathrm{Spec}(H)$, with distribution

$$\mathbb{P}[Z_H = \lambda_i] = \frac{|e_n^* V e_i|^2}{\|e_n^* V\|^2}$$

so that $\mathbb{P}[Z_H = \lambda_i] = 1$ exactly when e_n^* is a left eigenvector with eigenvalue λ_i. From the above, note that when H is normal, the distribution of Z_H is the spectral measure of H associated to e_n, and by the functional calculus we have $\|e_n^* f(H)\| = \mathbb{E}[|f(Z_H)|^2]^{\frac{1}{2}}$. Moreover, when H is nonnormal, we will still have an approximate version of this identity (see Lemma 4.1.5 and its proof), namely:

$$\frac{\|e_n^* f(H)\|}{\kappa_V(H)} \leq \mathbb{E}\left[|f(Z_H)|^2\right]^{\frac{1}{2}} \leq \kappa_V(H)\|e_n^* f(H)\|.$$

We will refer to this statement as the *approximate functional calculus*. The key observation here, is that when $k \gg \log \kappa_V(H)$, then $\kappa_V(H)^{1/k} \approx 1$, and hence quantities such as the $\psi_k(H)$ defined above, which can be represented in the form $\|e_n^* p(H)\|^{1/k}$ (for some polynomial of degree k), admit an approximate analytic representation, namely $\mathbb{E}[|p(Z_H)|^2]^{1/2k}$.

Spectral bisection and the holomorphic functional calculus. When working with the spectral bisection algorithm we were able to avoid the use of any measure on the spectrum of the (nonnormal) matrix in question, but we still heavily relied on the holomorphic functional calculus as it has been explained above.

1.6.2 The Effect of κ_V and gap on Convergence and Deflation

As explained in Section 1.1, for $A \in \mathbb{C}^{n\times n}$, $\kappa_V(A)$ can be used to quantify "how nonnormal A is" and as an upper bound for the sensitivity of the eigenvalues of A. On the other hand, to control the sensitivity of the eigenvectors of A, a lower bound on $\mathrm{gap}(A)$ is also required. So it is to be expected that these quantities will appear in some form in the finite arithmetic analysis of any diagonalization algorithm.

[23] In the event that there are multiple such choices of V it does not matter which we choose, only that it remains fixed throughout the analysis.

Interestingly, when it comes to our analysis of the spectral bisection and the Hessenberg QR algorithm, κ_V and gap play an important role even if exact arithmetic is assumed. In short, κ_V affects the number of iterations needed for convergence to take place (it follows that to guarantee quick convergence an upper bound on κ_V will be required). On the other hand gap becomes relevant in the deflation step: eigenvalues need to be well separated for a successful reduction of the problem to take place.

The effect of κ_V on convergence. Let us start by discussing the spectral bisection algorithm. As explained in Section 1.3, a key subroutine of this algorithm is Robert's Newton iteration, which on an input matrix A produces a sequence of matrices $A_0 = A, A_1, A_2, \dots$ that converge (in exact arithmetic) to $\operatorname{sgn}(A)$. In Chapter 3, to prove convergence in finite arithmetic we will show that for some suitably chosen decreasing sequence of positive numbers ϵ_N, the set $\Lambda_{\epsilon_N}(A_N)$ converges (for any reasonable definition of convergence for sets) to the two-point set $\{-1, 1\}$. Importantly, the speed of this convergence is not affected by the size of $\kappa_V(A)$, however, the size of the ϵ_N is determined by $\kappa_V(A)$ and in turn this dependence makes an appearance on the upper bound on $\|A_N - \operatorname{sgn}(A)\|$ that one can infer as a result of $\Lambda_{\epsilon_N}(A)$ clustering around $\{-1, 1\}$ (see Proposition 3.4.8).

In the case of the shifted Hessenberg QR algorithm, the dependence on $\kappa_V(H)$ in our analysis comes from the use of the approximate functional calculus, which as explained above becomes effective only when applied to polynomials of degree $k \gg \kappa_V(H)$. This forces the use of high order k polynomial shifts, each of which can be viewed as k steps of degree 1 shifts. Therefore, if speed of convergence is measured in terms of the number of degree 1 QR steps that it takes for decoupling to occur, in our analysis the speed of convergence will depend (at least logarithmically) on κ_V.

The effect of gap *on deflation.* In the case of the spectral bisection algorithm it is apparent why a lower bound on $\operatorname{gap}(A)$ is needed to ensure deflation: if two or more eigenvalues of the input matrix A lie in the same square grid when it terminates, the spectral bisection algorithm will only be able to compute the spectral projector corresponding to the span of the eigenvectors associated to such eigenvalues, and will fail to compute each of the individual eigenvectors. On the other hand, when the position of the grid is randomized, in order to have that with high probability no two eigenvalues lie in the same square grid, it is necessary for $\operatorname{gap}(A)$ to be bigger than the resolution of the grid (which ultimately determines the accuracy of the solution).

When it comes to the QR algorithm, the relevance of $\operatorname{gap}(H)$ is more subtle. To understand this recall what occurs during deflation. If H is a δ-decoupled Hessenberg matrix, i.e. $|h_{j,j-1}| < \delta$ for some $j = 2, \dots, n$, then H is deflated by setting $h_{j,j-1}$ to be 0, turning the matrix into an block upper triangular matrix H' from where we can extract the smaller subproblems $H^{(1)}$ (a $j \times j$ Hessenberg matrix) and $H^{(2)}$ (an $(n-j) \times (n-j)$ Hessenberg matrix). Now, as explained above, in order to ensure fast convergence (i.e. rapid decoupling) for each of the subproblems, we will need an upper bound on $\kappa_V(H^{(1)})$ and $\kappa_V(H^{(2)})$, which will be inherited from an upper bound on $\kappa_V(H')$. However, the initial assumption is only

a bound on $\kappa_V(H)$, which we would like to translate into a bound on $\kappa_V(H')$ by exploiting that $\|H - H'\| < \delta$, and, in view of Proposition 1.1.3 this is possible provided that δ is small compared to $\mathrm{gap}(H)$, which enforces a lower bound on $\mathrm{gap}(H)$. Alternatively one can think of $\mathrm{gap}(H)$ as a given quantity, and δ a parameter to be chosen, in which case a smaller gap eigenvalue gap results in a smaller setting of δ, which in turn forces the algorithm to run for longer (since the decoupling requirement is stronger) and the required computational precision to be higher to still obtain meaningful answers at such small scales[24].

1.6.3 The Role of Randomness

All of the algorithms analyzed in this dissertation heavily rely on randomness. However, the goals of randomness and the mechanism behind achieving these goals vary from case to case.

Psedospectral shattering. As it has been discussed already, in the case of the spectral bisection and the QR algorithms, the eigenvalue condition number and the minimum eigenvalue gap of the input matrix determine (1) the number of bits of precision required to guarantee an accurate solution and (2) the speed of convergence of the algorithm and the scale at which deflation should be performed (which in turn also affects the overall running time of the algorithm). In order to obtain uniform running time bounds that are satisfied under general precision requirements independent of the input matrix $A \in \mathbb{C}^{n\times n}$, as explained in Section 1.2, a small random perturbation γM_n (of scale γ) is added to A. This guarantees, with high probability, a polynomial in n and γ^{-1} upper bound on κ_V (henceforth denoted as $\mathrm{poly}(n, \gamma^{-1})$) and a $\mathrm{poly}(1/n, \gamma)$ lower bound on $\mathrm{gap}(A)$; translating into a $\mathrm{polylog}(n, \gamma^{-1})$ upper bound for the number of bits of precision that are required to obtain an accurate output (both in the spectral bisection and in the QR case), and a $\mathrm{polylog}(n, \gamma^{-1})$ number of calls to a matrix-multiplication algorithm (in the case of spectral bisection) or a $n \cdot \mathrm{polylog}(n, \gamma^{-1})$ number of calls to an implicit QR algorithm (in the case of the Hessenberg QR algorithm) for the execution of these algorithms.

Randomized grid and Ritz value perturbation. The number of iterations needed for Robert's Newton iteration to converge on the input matrix A depends logarithmically on α^{-1} for $\alpha(A) := \min_{\lambda\in\mathrm{Spec}(A)} |\mathrm{Re}(\lambda)|$. As mentioned above, in the implementation of the spectral bisection algorithm suggested in this dissertation, the position of the grid used to perform the bisection is randomized in an absolutely continuous way (say by shifting it by a random complex number taken uniformly in a small disk). The purpose of this is to ensure (by anti-concentration) that the distance of the grid to the eigenvalues of A is at least $\eta > 0$, with probability $1 - O(\eta/\omega)$ where ω is the length of the side of any of the squares in the grid. In this way, when the sign function is called as a subroutine of the spectral bisection algorithm on some matrix A', we are guaranteed a high probability quantitative lower bound

[24]Note that the analog of this phenomenon in the case of the spectral bisection algorithm would be to make the squares in the initial grid smaller, which again would result in slower convergence of the sign function and a higher required precision.

on $\alpha(A')$. This is not only relevant in terms of convergence, but also in terms of numerical stability (cf. the condition number of the sign function defined in Section 1.3).

Randomness is used in an analogous way in our suggested implementation of the shifted QR algorithm, where approximate Ritz values (in the sense of Definition 1.4.2) $r_1, \ldots, r_k$ are used to define the polynomial shifts in the iteration. Here, instead of working directly with the r_i, we will work with $r_1 + w_1, \ldots, r_k + w_k$ where the w_i are random complex numbers drawn uniformly from a small disk around zero (where the scale of the disk is suitably chosen). With this, one can guarantee (again by anti-concentration) that with high probability, the $r_i + w_i$ will be η away (for some suitable parameter $\eta > 0$) from any eigenvalue of the matrix in question H. The point being, that the condition number of the QR decomposition of a matrix A is proportional to $\|A^{-1}\|$ (see Lemma B.2.8), so to guarantee that the implicit QR algorithm run on $H - s$, for a given $s \in \mathbb{C}$, will be forward stable one needs some control on $\|(H - s)^{-1}\|$; for our main shifting strategy s will be of the form $s = r_i + w_i$.

Concentration. In contrast with the use of randomness discussed above, which relies on anti-concentration, the randomness present in the Lanczos algorithm solely relies on concentration. Conceptually, the phenomenon that we are exploiting in this case, is that if $A \in \mathbb{C}^{n\times n}$ is the input matrix and $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, then $\mu_{A,u}$ will concentrate around the empirical spectral distribution of A, so it is to be expected that its Jacobi concentrate as well.

1.7 Future Directions

Regarding the use of a random perturbation of the input matrix to ensure spectral stability, which was discussed in Section 1.2, there are many natural follow-up (open) random matrix questions (see [10, Section 7], [40, Conjecture 2.4], [11, Conjecture D.6]). However, perhaps the one that is most relevant to numerical linear algebra, is if the same pseudospectral shattering phenomenon is achieved when the random perturbation preserves the structure of the input matrix. Here we state some concrete questions in this direction.

Problem 1 (Structured random perturbations)**.** Let $A \in \mathbb{C}^{n\times n}$ be deterministic, $\gamma > 0$ and M_n random. Can one obtain tail bounds on $\kappa_V(A + \gamma M_n)$ and $\mathrm{gap}(A + \gamma M_n)$ like the ones given in Proto-Statement 1.2.1 when:

- A is sparse and $M_n = G_n \circ B_n$,[25] where G_n is a normalized complex Ginibre matrix and B_n is an independent matrix of i.i.d. Bernoulli random variables of parameter p_n (determined by the level of sparsity of A)?
- A is an upper Hessenberg matrix and $M_n = G_n \circ H$, where G_n is a normalized complex Ginibre matrix and H is a deterministic all-ones upper Hessenberg matrix (i.e. $h_{ij} = 1$ whenever $i \leq j + 1$ and $h_{ij} = 0$ otherwise)?

[25] Where $\circ$ denotes the Hadamard product.

- When A is a Toeplitz matrix and M_n is a random Toeplitz matrix with independent (modulo the Toeplitz structure) complex Gaussian entries of variance $1/n$?

Provided that one can find a solution to any of the above questions, do the proof techniques extend to distributions that are not complex Gaussian?

Recall, from Section 1.3, that our algorithm analysis of the spectral bisection algorithm shows that the eigenvalue problem can be solved in nearly matrix-multiplication time. This raises the natural (to the best of our knowledge open) question of if this is a nearly optimal upper bound on the running time.

Problem 2 (Complexity of diagonalization)**.** Is the bit complexity of the eigenvalue problem (in the sense of Definition 1.1.1) lower bounded by that of stable matrix-multiplication?

As discussed in Section 1.4, our analysis of the shifted QR algorithm, which produces a sequence of Hessenberg matrices $H_0, H_1, H_2, \ldots$, requires reasoning about the subdiagonal entries of H_{n+1} in terms of H_n, which in turn requires having forward stability guarantees for each (degree k) implicit QR step. The latter forces one to use higher precision computations, resulting in the $\Omega(\log^2(n/(\delta\phi)) \log\log(n/(\delta\phi)))$ bits of precision required to obtain a solution to the eigevalue problem with accuracy δ and probability of success $1-\phi$. This motivates the following question, whose solution would presumably require an essentially different approach.

Problem 3 (Optimal precision for the QR algorithm)**.** Is there a shifting strategy for the QR algorithm, for which one can prove the same guarantees as above assuming only $O(\log(n/(\delta\phi)))$ bits of precision?

Finally, we bring to the attention of the reader the block Lanczos algorithm [159, 78, 135], which is a commonly used extension of the Lanczos algorithm. Interestingly, for this block version there is a connection with matrix-valued orthogonal (with respect to a matrix-valued measure) polynomials on the real line. Given this connection, the techniques presented in Chapter 5 suggest an approach for analyzing the block Lanczos algorithm, although new ideas and additional substantial work would be needed to extend those techniques to the matrix-valued case.

Problem 4 (Block Lanczos)**.** Can analogous results to those discussed in Section 1.5 be obtained in the analysis of the block Lanczos algorithm?

Chapter 2

Spectral Stability Under Random Perturbations

In this chapter we will prove the results discussed in Section 1.2. Throughout the chapter $\boldsymbol{M}_n$ will be an $n \times n$ real random matrix satisfying Assumption 1.2.3. And, for the sake of clarity, we will use boldface to denote random quantities and distinguish them from deterministic ones.

Before delving into details we remind the reader that, although $\mathbb{R}$ has Lebesgue measure zero inside $\mathbb{C}$, non-normal *real* random matrices can have real eigenvalues (in fact many) with high probability (see [63]). Moreover, we will see that the behavior of the real eigenvalues will differ from that of the truly complex ones, and we will have to handle them separately. In particular, to prove Theorem 1.2.6 mentioned above, will the following separate tail bounds for the eigenvalue condition numbers of real and complex eigenvalues.

Theorem 2.0.1 (Eigenvalue and Eigenvector Condition Numbers)**.** *Let $n \geq 9$. Let $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{M}_n$ satisfy Assumption 1.2.3 with parameter $K > 0$. Let $0 < \gamma < K \min\{1, \|A\| + R\}$, and write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \gamma\boldsymbol{M}_n$. Let $R > \mathbb{E}\|\gamma\boldsymbol{M}_n\|$. Then for any $\epsilon_1, \epsilon_2 > 0$, with probability at least*

$$1 - 2\epsilon_1 - O\left(\frac{R(R + \|A\|)^{3/5}K^{8/5}n^{14/5}\epsilon_2^{3/5}}{\gamma^{8/5}}\right) - 2\mathbb{P}[\gamma\|\boldsymbol{M}_n\| > R],$$

we have

$$\sum_{\boldsymbol{\lambda}_i\in\mathbb{R}} \kappa(\boldsymbol{\lambda}_i) \leq \epsilon_1^{-1}C_{2.0.1}Kn^2\frac{\|A\| + R}{\gamma},$$

$$\sum_{\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}} \kappa(\boldsymbol{\lambda}_i)^2 \leq \epsilon_1^{-1}\log(1/\epsilon_2)C_{2.0.1}K^3n^5 \cdot \frac{(\|A\| + R)^3}{\gamma^3}, \qquad \textit{and}$$

$$\kappa_V(A + \gamma\boldsymbol{M}_n) \leq \epsilon_1^{-1}\sqrt{\log(1/\epsilon_2)}C_{2.0.1}K^{3/2}n^3 \cdot \frac{(\|A\| + R)^{3/2}}{\gamma^{3/2}},$$

Similarly, it will not be enough to have the upper bound given in Theorem 1.2.10 for the tails of the smallest singular values of complex shifts $z - \boldsymbol{M}_n$, but we will also have to handle real shifts separately. To this end we will also show the following.

Theorem 2.0.2 (Singular Values of $\boldsymbol{M}_n$)**.** *Let $\boldsymbol{M}_n \in \mathbb{R}^{n\times n}$ be a random matrix satisfying Assumption 1.2.3 with parameter $K > 0$. Then*

$$\mathbb{P}\left[\sigma_{n-k+1}(\boldsymbol{M}_n) \leq \epsilon\right] \leq \binom{n}{k}\left(\sqrt{2}K\epsilon\sqrt{kn(n-k+1)}\right)^{k^2} \leq n^{k^2+k}k^{\frac{1}{2}k^2}(\sqrt{2}K)^{k^2}\epsilon^{k^2}.$$

Note that Theorem 2.0.2 includes as a special case matrices of type $z - (A + \gamma\boldsymbol{M}_n)$ for real z and A, as such matrices themselves satisfy Assumption 1.

2.1 Related Work

2.1.1 Eigenvalue condition numbers

In the physics literature, the eigenvalue condition numbers of some diagonalizable $A \in \mathbb{C}^{n\times n}$ are referred to as the (diagonal) overlaps of A. More in general, the $n \times n$ *overlap matrix* of

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

is defined as $\mathscr{O}(A)_{i,j} = v_j^* v_i \overline{w_j^* w_i}$, so that $\mathscr{O}(A)_{i,i} = \kappa(\lambda_i)^2$.

For complex Ginibre matrices, much is known about diagonal overlaps and off-diagonal overlaps. In the seminal work of Chalker and Mehlig [38] explicit formulas were given for the limiting expected overlaps as $n \to \infty$, conditioned on the locations of the participating eigenvalues. Since then there has been significant progress; here we mention a few recent milestones. In [30], a formula for the limiting distribution of the diagonal overlaps was proved, as well as asymptotic formulas for the expected value of all overlaps, and for correlations between overlaps. Using a different approach, in [70], an explicit nonasymptotic formula for the joint density of an eigenvalue and its diagonal overlap was proved.

For the real Ginibre ensemble, results are more limited. The same paper [70] gives an analogous joint density formula for real Ginibre matrices, but only for real eigenvalues.[1] Compared to a joint density formula, our Theorem 2.0.1 (a polynomial upper bound with high probability) is rather coarse, but our theorem holds for general random matrices with absolutely continuous entries. More recent work [40] gave an optimal bound for the tails of the diagonal overlaps of the *complex* eigenvalues of a real Ginibre matrix.

As mentioned in Section 1.2, in [15] tail bounds for the diagonal overlaps of $\boldsymbol{M}_n = A+\gamma\boldsymbol{G}_n$ where obtained in the case where $\boldsymbol{G}_n$ is a complex Ginibre matrix. And similar results can

[1] Fyodorov [70] writes: "The approach suggested in the present paper can be certainly adjusted for addressing overlaps of left/right eigenvectors corresponding to complex eigenvalues of the real Ginibre ensemble, although in this way one encounters a few challenging technical problems not yet fully resolved."

be obtained from the integration formulas provided in [4]. Finally, we discuss the concurrent work of Jain, Sah and Sawhney [93]:

Concurrent and Independent Work. After completing [10], we learned of the independent work [92] which obtains results similar to ours regarding the eigenvector condition number and minimum eigenvalue gap. Their bound on κ_V improves Theorem 1.2.6 by a factor of $O(n/(\sqrt{\gamma}\log(n/\gamma)))$, thus almost matching the dependence on γ in Davies' conjecture [45]; their bound on the minimum eigenvalue gap is also better than that supplied by Theorem 1.2.5 by a poly(n/γ) factor. They do not obtain specific control on the $\kappa(\lambda_i)$ for real and complex λ_i separately, and our bound for the sum of the real $\kappa(\lambda_i)$ in Theorem 2.0.1 implies a bound for the maximum which is slightly better than their κ_V bound alone.

The techniques used by both papers focus on deriving tail bounds for the least singular value with the correct scaling in ϵ, but the proofs are essentially different. In particular, our proof relies on studying the entries of the resolvent, whereas theirs is more geometric. We obtain bounds on the kth smallest singular values of real and complex shifts (Theorems 2.0.2 and 1.2.10) with the correct ϵ^{k^2} and ϵ^{2k^2} scaling, whereas they derive bounds for $k = 1, 2$, but with better dependence on n.

They do not take the limit as $\epsilon \to 0$ to derive bounds on $\kappa_V(\boldsymbol{M}_n)$, relying instead on a bootstrapping scheme, while we do.

2.1.2 Singular Values of Real Matrices with Complex Shifts.

As already discussed in Section 1.2, in the Ph.D. thesis of Ge [73] it was shown that when $\boldsymbol{M}_n$ is a real matrix with i.i.d. entries of mean zero and variance $1/n$ satisfying a standard anticoncentration condition, one has

$$\mathbb{P}\left[\sigma_n(\boldsymbol{M}_n - z) \leq \epsilon \text{ and } \|\boldsymbol{M}_n\| \leq R\right] \leq \frac{Cn^2\epsilon^2}{|\operatorname{Im}(z)|} + e^{-cn} \tag{2.1}$$

for all z, where R, C and c are universal constants, independent of n. It is important to remark here, that the additional exponential term is an essential feature of the proof technique of considering "compressible" and "incompressible" vectors in a net argument, and does not go away if one additionally assumes that the entries are absolutely continuous.

In the case of real Ginibre matrices, the following finer result was obtained by Cipolloni, Erdös and Schröder in [42]:

$$\mathbb{P}\left[\sigma_n(\boldsymbol{G}_n - z) \leq \epsilon\right] \leq C(n^2(1 + |\log\epsilon|)\epsilon^2 + n\epsilon e^{-\frac{1}{2}n(\operatorname{Im} z)^2}) \tag{2.2}$$

for $|z| \leq 1 + O(1/\sqrt{n})$, with an improved n-dependence at the edge $|z - 1| = O(1/\sqrt{n})$. In later work [41], the same authors showed that when $\boldsymbol{M}_n$ has real i.i.d. entries with unit variance and $|\operatorname{Im} z| \sim 1$, the statistics of the small singular values $z - \boldsymbol{M}_n$ agree with those

Result	Bound	Setting
[62]	$\mathbb{P}[\sigma_n(\boldsymbol{M}_n) < \epsilon] \le n\epsilon$	real Ginibre
[133]	$\mathbb{P}[\sigma_n(\boldsymbol{M}_n) < \epsilon] \le Cn\epsilon + e^{-cn}$	real i.i.d. subgaussian
[150]	$\mathbb{P}[\sigma_n(\boldsymbol{M}_n) < \epsilon] \le n\epsilon + O(n^{-c})$	real i.i.d., finite moment assumption
[137]	$\mathbb{P}[\sigma_n(A + \boldsymbol{M}_n) < \epsilon] \le Cn\epsilon$	real Ginibre, A real
[153]	$\mathbb{P}[\sigma_n(A + \boldsymbol{M}_n) < \epsilon] \le Cn\epsilon$	real ind. rows with log-concave law, A real
[15]	$\mathbb{P}[\sigma_n(A + \boldsymbol{M}_n) < \epsilon] \le n\epsilon$	real Ginibre, A real

Table 2.1: Some bounds on σ_n for real $\boldsymbol{M}_n$ and A. Entries of $\boldsymbol{M}_n$ have variance $1/n$.

of the complex Ginibre ensemble.[2] And in more recent work [40] they improved the bound given in (2.2).

As remarked in Section 1.2, the key feature of our bounds is that we obtain a strict ϵ^2 dependence for nonreal z, without any additive terms. Our approach is essentially different from the above two approaches, and relies on exploiting a certain conditional independence (Observation 2.5.2) between submatrices of the real and imaginary parts of the resolvent.

2.1.3 Singular Values of Real Matrices with Real Shifts.

In the more general non-Gaussian case, there are a number of recent results in the literature. The most relevant recent result is that of Nguyen [116], who proves a tail bound for all singular values for non-centered ensembles with potentially discrete entries. In the particular case of continuous entries, Nguyen shows that if $\boldsymbol{M}_n$ satisfies Assumption 1.2.3 with parameter $K > 0$,

$$\mathbf{P}\left[\sigma_{n-k+1}(\boldsymbol{M}_n) \le \epsilon\right] \le n^{k(k-1)}(CkK\epsilon)^{(k-1)^2}, \tag{2.3}$$

in addition to a bound greatly improving the dependence in k at the expense of the dependence on ϵ and n, as well as results for symmetric Wigner matrices and perturbations thereof.

The exponent of ϵ in (2.3) is suboptimal, which renders (2.3) incompatible with the approach outlined in Section 1.2. Below, in Theorem 2.0.2 we will obtain the optimal exponent of ϵ, namely k^2, in exchange for a worse exponent of n. The key ingredient in doing this is a simple "restricted invertibility" type estimate (Lemma 2.4.1) tailored to our setting.

For bounds on the least singular value alone, there is a substantial literature; see Table 2.1 for a non-exhaustive summary.

[2] They further write, "It is expected that the same result holds for all (possibly n-dependent) z as long as $|\operatorname{Im}(z)| \gg n^{-1/2}$, while in the opposite regime $|\operatorname{Im}(z)| \ll n^{-1/2}$ the local statistics of the real Ginibre prevails with an interpolating family of new statistics which emerges for $|\operatorname{Im}(z)| \sim n^{-1/2}$."

2.1.4 Minimum Eigenvalue Gap

Bounds on the minimum eigenvalue gap of random non-Hermitian matrices have seen rapid progress in the last few years. Ge shows in the thesis [73] that when $\boldsymbol{M}_n$ has i.i.d. entries with zero mean and variance $1/n$, satisfying a standard anticoncentration condition,

$$\mathbb{P}[\mathrm{gap}(\boldsymbol{M}_n) < s] = O\left(\delta n^{2+o(1)} + \frac{s^2 n^{4+o(1)}}{\delta^2}\right) + e^{-cn} + \mathbb{P}[\|\boldsymbol{M}_n\| \geq R]$$

for every $C > 0$ and every $\delta > s > n^{-C}$. In more recent work, Luh and O'Rourke [106] build on Ge's result, dropping the mean zero assumption and extending the range of s all the way down to 0:

$$\mathbb{P}[\mathrm{gap}(\boldsymbol{M}_n) \leq s \text{ and } \|\boldsymbol{M}_n\| \leq M] \leq Cs^{2/3}n^{16/15} + Ce^{-cn} + \mathbb{P}[\|\boldsymbol{M}_n\| \geq M]. \tag{2.4}$$

However, (2.4) still requires the entries of $\boldsymbol{M}_n$ to be identically distributed, so, for example, it does not imply a gap bound for the noncentered Ginibre ensemble $A + \boldsymbol{G}_n$ unless A is a scalar multiple of the all-ones matrix.

In our work [11] (which will be discussed in Chapter 3), a complex Gaussian perturbation was crucially used in a preprocessing step in a numerically stable diagonalization algorithm for non-Hermitian matrices. This paper identified the minimum eigenvalue gap as a key feature controlling the stability of the algorithm, and proved:

Theorem 2.1.1 ([11, Corollary 3.7])**.** *Suppose $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$, and $\boldsymbol{G}_n$ is a normalized* complex *Ginibre matrix. For every $\delta \in (0, 1/2)$,*

$$\mathbb{P}[\mathrm{gap}(A + \delta\boldsymbol{G}_n) < s] \leq 42(n/\gamma)^{16/5}s^{6/5} + 2e^{-2n}.$$

Each of the gap results above are proved by way of tail bounds on the smallest two singular values of $z - \boldsymbol{G}_n$. The only other work we are aware of proving gap bounds for the case of matrices with i.i.d. entries is [139], which proves an inverse polynomial lower bound for the complex Ginibre ensemble.

2.2 Probabilistic Preliminaries

Many of our probabilistic arguments hinge on the phenomenon of *anticoncentration*, whereby a random vector is unlikely to lie in a small region. An elementary way to extract quantitative information about such behavior is by controlling the density function of the random vector. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random vector with absolutely continuous distribution with respect to the Lebesgue on $\mathbb{R}^d$, and let $f_{\boldsymbol{x}}$ be its density. We will denote

$$\delta_\infty(\boldsymbol{x}) := \|f_{\boldsymbol{x}}\|_\infty \tag{2.5}$$

We will repeatedly use two basic observations about the quantity δ_∞. First, for any $v \in \mathbb{R}^d$,

$$\mathbb{P}\left[\|\boldsymbol{x} - v\| \leq \epsilon\right] \leq \frac{(\pi\epsilon^2)^{d/2}}{\Gamma(d/2+1)}\delta_\infty(\boldsymbol{x}) \leq \frac{1}{\sqrt{\pi d}}\left(\frac{2e\pi\epsilon^2}{d}\right)^{d/2}\delta_\infty(\boldsymbol{x}), \tag{2.6}$$

where in the first inequality we use the formula for the volume of a ball in $\mathbb{R}^d$, and in the second inequality we use Stirling's approximation for the gamma function. Second, δ_∞ is preserved under convolution:

Observation 2.2.1 (Convolution Bound)**.** Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ be independent random vectors with absolutely continuous distributions. Then

$$\delta_\infty(\boldsymbol{x} + \boldsymbol{y}) \leq \min\{\delta_\infty(\boldsymbol{x}), \delta_\infty(\boldsymbol{y})\}.$$

We will require as well a much more general result of Rudelson and Vershynin quantifying the deterioration of δ_∞ after orthogonal projection.[3]

Theorem 2.2.2 ([132])**.** *Let $\boldsymbol{x} \in \mathbb{R}^d$ have independent entries, each with density pointwise almost surely bounded by K. Let $P \in \mathbb{R}^{k\times d}$ denote a deterministic orthogonal projection onto a subspace of dimension $k \leq d$. Then there exists a universal constant $C_{\mathrm{RV}} > 0$ such that*

$$\delta_\infty(P\boldsymbol{x}) \leq (C_{\mathrm{RV}}K)^k.$$

In [104], it is shown that Theorem 2.2.2 holds with $C_{\mathrm{RV}} = \sqrt{2}$, and that this is sharp. Moreover, if $\boldsymbol{x}$ has independent standard real Gaussian entries, one may take $C_{\mathrm{RV}} = 1$ and $K = (2\pi)^{-1/2}$.

Many of our results on real random matrices whose independent entries have bounded density—in other words, matrices satisfying Assumption 1.2.3—can be strengthened for real Ginibre matrices, mainly via the use of results in [143] and [147]. In this dissertation we will focus on the general case and refer the reader to [10] for improved bounds of the results presented here in the particular case when the entries are Gaussian.

2.3 Anticoncentration for Quadratic Forms

In this section we study the anticoncentration properties of certain quadratic functions of rectangular matrices with independent entries. These will be necessary in Section 3 to extract singular value tail bounds.

Theorem 2.3.1 (Density of Quadratic Forms)**.** *Assume that $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times k}$ are random matrices with independent entries, each with density on $\mathbb{R}$ bounded a.e. by $K > 0$. Let*

[3] Throughout the chapter, we will refer to a rectangular matrix with orthonormal columns as an "orthogonal projection" although this is not standard.

$Z \in \mathbb{R}^{n\times n}$, $U, V \in \mathbb{R}^{n\times k}$, and $W \in \mathbb{R}^{k\times k}$ be deterministic, and write $q(\boldsymbol{X}, \boldsymbol{Y}) := \boldsymbol{X}^\intercal Z \boldsymbol{Y} + \boldsymbol{X}^\intercal U + V^\intercal \boldsymbol{Y} + W$. Then

$$\delta_\infty \left(q(\boldsymbol{X}, \boldsymbol{Y})\right) \leq (1+k^2)\left(2K^2\sqrt{2e\pi k} \min_{j>k^2+k+1} \frac{1}{\sqrt{j-k+1}\sigma_j(Z)}\right)^{k^2}.$$

Whenever $\sigma_j(Z)$ is zero, we interpret $1/\sigma_j(Z) = \infty$; thus the above theorem has content only when $\operatorname{rank}(Z) > k^2 + k + 1$. Before proving the above theorem let us begin with some observations that we will use in the proof to come.

Lemma 2.3.2. *Consider measurable functions $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^r$ and $c : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$. Let $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^q$ be independent random vectors with densities bounded almost everywhere. Assume that for almost all $y \in \mathbb{R}^r$ it holds that $\delta_\infty \left(f(\boldsymbol{x}, y)\right) \leq c(y)$. Then*

$$\delta_\infty \left(f(\boldsymbol{x}, \boldsymbol{y})\right) \leq \mathbb{E}[c(\boldsymbol{y})].$$

Proof. Let $\mathrm{Leb}_{\mathbb{R}^r}$ denote the Lebesgue measure on $\mathbb{R}^r$. Note that it is enough to show that for every measurable set $E \subset \mathbb{R}^r$ one has

$$\mathbb{P}[f(\boldsymbol{x}, \boldsymbol{y}) \in E] \leq \mathrm{Leb}_{\mathbb{R}^r}(E)\mathbb{E}[c(\boldsymbol{y})].$$

On the other hand, by assumption, we have $\mathbb{P}[f(\boldsymbol{x}, y) \in E] \leq \mathrm{Leb}_{\mathbb{R}^r}(E)c(y)$ for all y. From the fact that $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent and have a density it follows that

$$\mathbb{P}[f(\boldsymbol{x}, \boldsymbol{y}) \in E] = \mathbb{E}[\mathbb{1}\left\{f(\boldsymbol{x}, \boldsymbol{y}) \in E\right\}] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{f(\boldsymbol{x}, \boldsymbol{y}) \in E\right\} | \boldsymbol{y}\right]\right] \leq \mathbb{E}\left[\mathrm{Leb}_{\mathbb{R}^r}(E)c(\boldsymbol{y})\right],$$

as we wanted to show. □

We will also require the following left tail bound on the smallest singular value of certain rectangular random matrices, which is a consequence of Theorem 2.2.2.

Lemma 2.3.3. *Let $\boldsymbol{Y}$ be a $n \times k$ random matrix whose entries are independent and have density on $\mathbb{R}$ bounded a.e. by $K > 0$. Furthermore, for some $k \leq j \leq n$ let V be a $j \times n$ projector. Then*

$$\mathbb{P}[\sigma_k(V\boldsymbol{Y}) \leq s] \leq k\frac{(\sqrt{2}K\sqrt{\pi k}s)^{j-k+1}}{\Gamma((j-k+3)/2)} := C_{j,k}s^{j-k+1} \tag{2.7}$$

Proof. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ be the columns of $\boldsymbol{Y}$ and for every $i = 1, \ldots, k$ let $\boldsymbol{W}_i$ be the $(j-k+1)\times j$ orthogonal projector onto the subspace orthogonal to the span of $\{V\boldsymbol{y}_l\}_{l\neq i}$. Applying the "negative second moment identity" [151], we have

$$k\left(\min_{i\in[k]} \|\boldsymbol{W}_i V \boldsymbol{y}_i\|\right)^{-2} \geq \sum_{i=1}^k \|\boldsymbol{W}_i V \boldsymbol{y}_i\|^{-2} \geq \sum_{i=1}^k \sigma_i(V\boldsymbol{Y})^{-2} \geq k\sigma_k(V\boldsymbol{Y})^{-2},$$

which implies

$$\sigma_k(\boldsymbol{Y}) \geq \frac{\min_i \|\boldsymbol{W}_i V \boldsymbol{y}_i\|}{\sqrt{k}}.$$

Since $\boldsymbol{W}_i V$ is itself an orthogonal projector, and is independent of $\boldsymbol{y}_i$, Theorem 2.2.2 and Observation 2.3.2 ensure that the density of $\|\boldsymbol{W}_i V \boldsymbol{y}_i\|$ is bounded by $(\sqrt{2}K)^{j-k+1}$. Applying a union bound and recalling again the formula for a ball,

$$\mathbb{P}[\sigma_k(\boldsymbol{Y}) \leq s] \leq \mathbb{P}[\min_i \|\boldsymbol{W}_i V \boldsymbol{y}_i\| \leq \sqrt{k}s] \leq \sum_{i=1}^{k} \mathbb{P}[\|\boldsymbol{W}_i V \boldsymbol{y}_i\| \leq \sqrt{k}s] \leq k \frac{(\sqrt{2}K\sqrt{\pi k}s)^{j-k+1}}{\Gamma((j-k+3)/2)}.$$

□

With these two tools in hand, we can proceed with the proof of the main result of this section.

Proof of Theorem 2.3.1. For any deterministic $Y \in \mathbb{R}^{n\times k}$ one has

$$\delta_\infty(q(\boldsymbol{X}, Y)) = \delta_\infty(\boldsymbol{X}^T(ZY + U)),$$

since δ_∞ is agnostic to deterministic translations. By the polar decomposition we can write $ZY + U = VS$, where $V \in \mathbb{R}^{n\times k}$ is a partial isometry and $S \succeq 0$. By Theorem 2.2.2, the density of the random matrix $\boldsymbol{X}^\intercal V$ in $\mathbb{R}^{k\times k}$ is at most $(\sqrt{2}K)^{k^2}$, and thus the density of $\boldsymbol{X}^\intercal VS$ is at most $(\sqrt{2}K)^{k^2}(\det S)^{-k}$; moreover

$$\det S = \prod_{i=1}^{k} \sigma_i(S) = \prod_{i=1}^{k} \sigma_i(ZY + U).$$

Therefore by Lemma 2.3.2,

$$\delta_\infty(q(\boldsymbol{X}, \boldsymbol{Y})) \leq (\sqrt{2}K)^{k^2} \mathbb{E}\left[\prod_{i\in k} \sigma_i(Z\boldsymbol{Y} + U)^{-k}\right]. \tag{2.8}$$

We now compute this expectation.

Choose $j \geq k$ so that $\sigma_j(Z) > 0$, and write the SVD of Z in the following block form,

$$Z = P^T \Sigma Q = \begin{pmatrix} P_1^\intercal & P_2^\intercal \end{pmatrix} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}, \tag{2.9}$$

where Σ_1 is a diagonal matrix containing the largest j singular values, and P, Q are orthogonal matrices. This gives

$$Z\boldsymbol{Y} + U = \begin{pmatrix} P_1^\intercal & P_2^\intercal \end{pmatrix} \begin{pmatrix} \Sigma_1 Q_1 \boldsymbol{Y} + P_1 U \\ \Sigma_2 Q_2 \boldsymbol{Y} + P_2 U \end{pmatrix}.$$

By interlacing of singular values, $\sigma_i(Z\boldsymbol{Y}+U) \geq \sigma_i(\Sigma_1 Q_1 \boldsymbol{Y} + P_1 U)$ for each $i = 1, ..., k$, so we are free to study

$$\mathbb{E}\left[\prod_{i\in[k]} \sigma_i(\Sigma_1 Q_1 \boldsymbol{Y} + P_1 U)^{-k}\right] \leq \sigma_j(\Sigma_1)^{-k^2} \mathbb{E}\left[\prod_{i\in[k]} \sigma_i(Q_1 \boldsymbol{Y} + \Sigma_1^{-1} P_1 U)^{-k}\right]. \tag{2.10}$$

Now, since Q_1 is a partial isometry, we can select a matrix $\tilde{U}$ so that $Q_1\tilde{U} = \Sigma_1^{-1}P_1U$, and observe that

$$\mathbb{E}\prod_{i\in[k]} \sigma_i(\Sigma_1 Q_1 \boldsymbol{Y} + P_1 U)^{-k} \leq \sigma_j(Z)^{-k^2} \sigma_k(Q_1(\boldsymbol{Y}+\tilde{U}))^{-k^2}.$$

The random matrix $\boldsymbol{Y}+\tilde{U}$ satisfies the conditions of Lemma 2.3.3, so we can apply the tail formula for expectation to obtain

$$\begin{aligned}
\mathbb{E}\left[\sigma_k(Q_1(\boldsymbol{Y}+\tilde{U}))^{-k^2}\right] &= \int_0^\infty \mathbb{P}\left[\sigma_k(Q_1(\boldsymbol{Y}+\tilde{U}))^{-k^2} \geq t\right] dt \\
&\leq \lambda + C_{j,k}\int_\lambda^\infty t^{-\frac{j-k+1}{k^2}} dt && C_{j,k} \text{ from (2.7)} \\
&= \lambda + C_{j,k}\frac{k^2}{j-k^2-k+1}\lambda^{\frac{k^2+k-j-1}{k^2}} && \text{if } j-k+1 > k^2.
\end{aligned}$$

Optimizing the above bound in λ, we set $\lambda = C_{j,k}^{\frac{k^2}{j-k+1}}$ and evaluate $C_{j,k}$ to find

$$\begin{aligned}
&\mathbb{E}\left[\sigma_k(Q_1(\boldsymbol{Y}+\tilde{U}))^{-k^2}\right] \\
\leq &\left(\frac{k(\sqrt{2}K\sqrt{\pi k})^{j-k+1}}{\Gamma((j-k+3)/2)}\right)^{\frac{k^2}{j-k+1}} \left(1+\frac{k^2}{j-k^2-k+1}\right) \\
\leq &(\sqrt{2}K\sqrt{\pi k})^{k^2}\left(\frac{k}{\Gamma((j-k+3)/2)}\right)^{\frac{k^2}{j-k+1}}(1+k^2) && j-k+1 > k^2 \\
\leq &(\sqrt{2}K\sqrt{\pi k})^{k^2}\left(\frac{k}{\sqrt{\pi(j-k+1)}}\right)^{\frac{k^2}{j-k+1}}\left(\frac{\sqrt{2e}}{\sqrt{j-k+1}}\right)^{k^2}(1+k^2) && \text{Stirling} \\
\leq &\left(\frac{\sqrt{2}K\sqrt{2e\pi k}}{\sqrt{j-k+1}}\right)^{k^2}(1+k^2) && j-k+1 > k^2
\end{aligned}$$

where we have repeatedly used that $j-k+1 > k^2$, as well as Stirling's approximation, $\Gamma(z+1) \geq \sqrt{2\pi z}(z/e)^z$, valid for real $z \geq 2$. To complete the proof, we combine the above with equation (2.8). □

To end this section, we offer an improvement of the above result for the case $k = 1$.

Corollary 2.3.4. *In the case $k = 1$, the conclusion of Theorem 2.3.1 may be improved to*

$$\delta_\infty\left(q(\boldsymbol{X}, \boldsymbol{Y})\right) \leq 2(C_{\mathrm{RV}}K)^2\sqrt{2e\pi}\min_{j\geq 2}\frac{1}{\sqrt{j}\prod_{i\in[j]}\sigma_i(Z)^{1/j}}.$$

Where $C_{\mathrm{RV}} = \sqrt{2}$.

Proof. The discussion between equations (2.8) and (2.10) in this case tells us

$$\delta_\infty((q(\boldsymbol{X}, \boldsymbol{Y})) \leq \sqrt{2}K\mathbb{E}\left[\|\Sigma_1 Q_1 \boldsymbol{Y} + P_1 U\|^{-1}\right].$$

The random vector $\Sigma_1 Q_1 \boldsymbol{Y} + P_1 U$ has density on $\mathbb{R}^j$ bounded by $(\sqrt{2}K)^j \det \Sigma_1^{-1}$, so we have the tail bound

$$\mathbb{P}\left[\|\Sigma_1 Q_1 \boldsymbol{Y} + P_1 U\| \leq s\right] \leq \det \Sigma_1^{-1}\frac{(\sqrt{2}K\sqrt{\pi}s)^j}{\Gamma(j/2+1)} = \det \Sigma_1^{-1}\cdot C_{j,1}s^j.$$

Replacing in the remainder of the proof $C_{j,k}$ with $\det \Sigma_1^{-1}C_{j,1}$, and recalling $\det \Sigma_1 = \sigma_1(Z)\cdots\sigma_j(Z)$, will give

$$\delta_\infty\left(q(\boldsymbol{X}, \boldsymbol{Y})\right) \leq \sqrt{2}K\mathbb{E}\left[\|\Sigma_1 Q_1 Y + P_1 U\|^{-1}\right] \leq 2\frac{(\sqrt{2}K)^2\sqrt{2e\pi}}{\sqrt{j}\prod_{i\in[j]}\sigma_i(Z)^{1/j}}$$

whenever $j \geq 2$. □

We believe that Theorem 2.3.1 should hold, for every k, with the jth singular value of Z exchanged for the geometric mean of the top j. The main obstacle in showing this seems to be that Theorem 2.2.2 cannot tightly bound the density of $A\boldsymbol{y}$, where $\boldsymbol{y} \in \mathbb{R}^n$ is a random vector with independent entries and bounded density, and $A \in \mathbb{R}^{n\times k}$ is an arbitrary matrix.

2.4 Singular Value Bounds for Non-Centered Real Matrices

In this section, we discuss singular value tail bounds for real matrices with independent absolutely continuous entries. In particular, our study of minimum eigenvalue gap and eigenvalue condition numbers will require tail bounds on the least two singular values for shifted random matrices of the form $z - (A + \boldsymbol{M}_n)$, where $z \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ are deterministic, and $\boldsymbol{M}_n$ satisfies Assumption 1.2.3.

For matrices with i.i.d. subgaussian entries, results similar to the tail bounds in (1.17) are known, but they are accompanied by additive error terms of the form e^{-cn} and therefore do not yield useful results in the limit as the tail parameter ϵ goes to zero. The closest result to ours appears in a paper of Nguyen [116]; it excises the additive error terms, but contains a sub-optimal exponent on ϵ. We will add one key insight to Nguyen's proof that allows one to obtain the correct ϵ-dependence.

2.4.1 A Restricted Invertibility Lemma

The device we add to Nguyen's argument, and which we will return to at several points throughout the chapter, is the following lemma, which shows that the kth largest eigenvalue of a PSD matrix is approximately witnessed by the smallest eigenvalue of some principal $k \times k$ submatrix. Given a matrix $A \in \mathbb{C}^{n\times n}$ and $S, T \subset [n]$ we use $A_{S,T}$ to denote the $|S| \times |T|$ matrix determined by looking at the intersection of the rows of A with indices in S with the columns with indices in T.

Lemma 2.4.1 (Principal Submatrix with Large σ_k)**.** *Let $X \in \mathbb{C}^{n\times n} \setminus \{0\}$ be positive semidefinite. Then for every $1 \le k \le n$, there exists an $k \times k$ principal submatrix $X_{S,S}$ such that*

$$\lambda_k(X_{S,S}) \ge \frac{\mathrm{Tr}(X)}{\sum_{i=1}^k \lambda_i(X)} \cdot \frac{\lambda_k(X)}{k(n-k+1)}. \tag{2.11}$$

Proof. Examining the coefficient of λ^k in the characteristic polynomial $\det(\lambda - X)$, we have

$$\sum_{|S|=k} \det X_{S,S} = e_k(\lambda_1(X), \lambda_2(X), \ldots, \lambda_n(X)),$$

where e_k denotes the k-th elementary symmetric function, and the sum runs over subsets of $[n]$. We may now have the upper bound:

$$\begin{aligned}
e_k(X) &= \sum_{|S|=k} \det(X_{S,S}) \\
&= \sum_{|S|=k} \lambda_k(X_{S,S})\lambda_{k-1}(X_{S,S}) \ldots \lambda_1(X_{S,S}) \\
&\le \sum_{|S|=k} \lambda_k(X_{S,S}) e_{k-1}(X_{S,S}) && \text{since } \lambda_i(X_{S,S}) \ge 0 \text{ by interlacing} \\
&\le \max_S \lambda_k(X_{S,S}) \cdot \sum_{|S|=k} \sum_{T\subset S, |T|=k-1} \det(X_{S',S'}) \\
&= \max_S \lambda_k(X_{S,S}) \cdot (n-k+1) e_{k-1}(X).
\end{aligned}$$

It now remains to furnish a complementary lower bound on $e_k(X)$ in terms of $e_{k-1}(X)$. Recall the routine fact that

$$k e_k(X) = k \sum_{|S|=k} \prod_{i\in S} \lambda_i(X) = \sum_{|T|=k-1} \sum_{j\notin T} \lambda_j(X) \prod_{i\in T} \lambda_i(X).$$

Now, for each $|T| = k-1$,

$$\sum_{j\in[k]} \lambda_j(X) \sum_{\ell\notin T} \lambda_\ell(X) = \sum_{j\in[k]} \lambda_j(X) \left(e_1(X) - \sum_{j\in T} \lambda_j(X) \right)$$

$$= \lambda_k(X)e_1(X) + \left(\sum_{j\in[k-1]} \lambda_j(X)\right) e_1(X) - \left(\sum_{j\in T} \lambda_j(X)\right)\left(\sum_{j\in[k]} \lambda_j(X)\right)$$
$$\geq \lambda_k(X)e_1(X),$$

since $\sum_{j\in[k-1]} \lambda_j(X) \geq \sum_{j\in T} \lambda_j(X)$, and $e_1(X) \geq \sum_{j\in[k]} \lambda_j(X)$. Thus

$$k\sum_{j\in[k]} \lambda_j(X)\cdot e_k(X) \geq \sum_{|T|=k-1} \lambda_k(X)e_1(X)\prod_{i\in T}\lambda_i(X) = \lambda_k(X)e_1(X)e_{k-1}(X).$$

Putting everything together, and recalling $e_1(X) = \mathrm{Tr}X$,

$$\max_S \lambda_k(X_{S,S}) \geq \frac{e_k(X)}{(n-k+1)e_{k-1}(X)} \geq \frac{\mathrm{Tr}(X)}{\sum_{i\in[k]}\lambda_i(X)}\frac{\lambda_k(X)}{k(n-k+1)}$$

as desired. □

We will employ Lemma 2.4.1 in the form of the corollary below.

Corollary 2.4.2. *Let $1 \leq k \leq n$. For every matrix $R \in \mathbb{C}^{n\times k}$, there exists a $k\times k$ submatrix Q of R such that*

$$\sigma_k(Q) \geq \frac{\sigma_k(R)}{\sqrt{k(n-k+1)}}. \tag{2.12}$$

Similarly, for every matrix $A \in \mathbb{C}^{n\times n}$, there are subsets $S, T \subset [n]$ of size k such that

$$\sigma_k(A_{S,T}) \geq \frac{\|A\|_F}{\sqrt{\sum_{i\in[k]}\sigma_i(A)^2}}\frac{\sigma_k(A)}{k(n-k+1)} \geq \frac{\sigma_k(A)}{k(n-k+1)} \tag{2.13}$$

This generalizes the elementary fact that the operator norm of an $n\times n$ matrix is bounded above by n times the maximal entry. Corollary 2.4.2 additionally sits within a much larger literature on *restricted invertibility*; see [114] for a comprehensive introduction. Most notably, the main result in [75] states that for any $R \in \mathbb{C}^{n\times k}$ of rank k, there exist a $k\times k$ submatrix Q of R, such that

$$\frac{1}{\sum_{i=1}^k \sigma_i(Q)^{-2}} \geq \frac{1}{(n-k+1)\sum_{i=1}^k \sigma_i(R)^{-2}}. \tag{2.14}$$

Note that neither (2.12) implies (2.14) nor (2.14) implies (2.12). However, from (2.12) one can derive an inequality very similar to (2.14) that has a slightly weaker dependence on k, and vice versa. The proof in [75] shares some features with our proof of Lemma 2.3.3, but differs in that it does not exploit the fact that coefficients of the characteristic polynomial can be written both in terms of the eigenvalues and in terms of the entries of the matrix. This allows us to obtain a result for general $n\times n$ matrices, namely (2.13), which is not clear how to obtain from (2.14).

2.4.2 Proof of Theorem 2.0.2

Finally, we may prove the desired tail bound:

Proof of Theorem 2.0.2. We give a similar argument to that of Nguyen [116], but using Corollary 2.4.2 where Nguyen uses the restricted invertibility theorem of [114].

Suppose $\sigma_{n-k+1}(\boldsymbol{M}_n) \leq \epsilon$. By the minimax formula for singular values, there exist orthogonal unit vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k \in \mathbb{R}^n$ such that $\|\boldsymbol{M}_n \boldsymbol{z}_i\| \leq \epsilon$. Letting $\boldsymbol{Z} \in \mathbb{R}^{n \times k}$ be the matrix whose columns are $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k$, we can bound $\|\boldsymbol{M}_n \boldsymbol{Z}\|_F \leq \epsilon\sqrt{k}$. Since $\sigma_k(\boldsymbol{Z}) = 1$, by Corollary 2.4.2, there is a $k \times k$ submatrix $\boldsymbol{Z}_1$ of $\boldsymbol{Z}$ for which

$$\|\boldsymbol{Z}_1^{-1}\| \leq \sqrt{k(n-k+1)}.$$

Denote by $\boldsymbol{Z}$ the subset of rows of $\boldsymbol{Z}$ participating in $\boldsymbol{Z}_1$; by permuting if necessary we can write

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{M}_n = \begin{pmatrix} \boldsymbol{M}_1 & \boldsymbol{M}_2 \end{pmatrix},$$

observing that

$$\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}_1^{-1} = \begin{pmatrix} \boldsymbol{M}_1 & \boldsymbol{M}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{pmatrix} \boldsymbol{Z}_1^{-1} = \boldsymbol{M}_1 + \boldsymbol{M}_2 \boldsymbol{Z}_2 \boldsymbol{Z}_1^{-1}. \tag{2.15}$$

Denote the columns of $\boldsymbol{M}_n$ by $\boldsymbol{m}_1, \dots, \boldsymbol{m}_n$ and let $\boldsymbol{H}$ denote the orthogonal projector onto the k-dimensional subspace orthogonal to the span of $\{\boldsymbol{m}_i\}_{i \notin \boldsymbol{S}}$, so that $\boldsymbol{H}\boldsymbol{M}_2 = 0$. Thus we have

$$\sum_{i \in \boldsymbol{S}} \|\boldsymbol{H}\boldsymbol{m}_i\|^2 = \|\boldsymbol{H}\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}_1^{-1}\|_F^2 \leq \|\boldsymbol{M}_n \boldsymbol{Z}\boldsymbol{Z}_1^{-1}\|_F^2 \leq \|\boldsymbol{M}_n \boldsymbol{Z}\|_F^2 \|\boldsymbol{Z}_1^{-1}\|^2 \leq \epsilon^2 k^2 (n-k+1).$$

Since the entries of $\boldsymbol{M}_n$ are independent, with densities on $\mathbb{R}$ bounded by $\sqrt{n}K$, by Theorem 2.2.2 the above event occurs with probability at most

$$\prod_{i=1}^{k} \mathbb{P}\left[\|\boldsymbol{H}\boldsymbol{m}_i\| \leq \epsilon k \sqrt{n-k+1}\right] < \left(\sqrt{2}K\sqrt{n} \cdot \epsilon\sqrt{k(n-k+1)}\right)^{k^2}.$$

Performing a union bound over all possibilities for the subset $\boldsymbol{S}$ of rows of $\boldsymbol{Z}$, we finally obtain

$$\mathbb{P}\left[\sigma_{n-k+1}(\boldsymbol{M}_n) \leq \epsilon\right] \leq \binom{n}{k} \left(\sqrt{2}K\epsilon\sqrt{kn(n-k+1)}\right)^{k^2} \leq n^{k^2+k} k^{\frac{1}{2}k^2} (\sqrt{2}K)^{k^2} \epsilon^{k^2}.$$

□

Comparing with the tail bounds in (1.17), we conclude that the exponent of ϵ in Theorem 2.0.2 is optimal, and if not for the factor of $\binom{n}{k}$ arising from the union bound, the exponent of n would be optimal as well. Since we made no requirement that $\boldsymbol{M}_n$ is centered, the following corollary is immediate:

Corollary 2.4.3. *Let $z \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and $\boldsymbol{M}_n$ satisfy Assumption 1.2.3 with parameter $K > 0$. Then*

$$\mathbb{P}[\sigma_{n-k+1}(z - (A + \boldsymbol{M}_n)) \le \epsilon] \le n^{\frac{1}{2}k^2+k} k^{\frac{1}{2}k^2} (\sqrt{2}K)^{k^2} \epsilon^{k^2}.$$

2.5 Singular Value Bounds for Real Matrices with Complex Shifts

As mentioned above, in order to control the eigenvalue gaps and pseudospectrum of random real perturbations, we need to understand the smallest singular values of real random matrices with complex scalar shifts, so our main goal of this section will be to prove Theorem 1.2.10.

As discussed in the introduction, our results will be stated in terms of the quantities

$$B_{\boldsymbol{M}_n,p} := [\mathbb{E}\|\boldsymbol{M}_n\|^p]^{1/p},$$

and important features of the bounds in our context are (1) the optimal dependence on ϵ as $\epsilon \to 0$, and (2) the factor $\frac{1}{|\operatorname{Im} z|}$ controlling the necessary deterioration of the bound as z approaches the real line.

2.5.1 Proof of Theorem 1.2.10

In view of Corollary 2.4.2, we can study the kth smallest singular value of $z - (A + \boldsymbol{M}_n)$ by examining the smallest singular value of every $k \times k$ submatrix of its inverse. In particular, we will show momentarily that Theorem 1.2.10 may be reduced to the following lemma, which we will prove in Section 2.5.2.

Lemma 2.5.1 (Tail bound for corner of the resolvent)**.** *Let $\delta \in \mathbb{R}$, let U be a permutation matrix, and let $\boldsymbol{M}_n$ satisfy Assumption 1.2.3 with parameter $K > 0$. Denote the upper-left $k \times k$ corner of $(\delta i U - \boldsymbol{M}_n)^{-1}$ by $\boldsymbol{N}_k$. If $n \ge (k+2)^2$,*

$$\mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \ge 1/\epsilon\right] \le (1+k^2)\left(8\sqrt{3}(e\pi)^{3/2}K^3 n \frac{\epsilon^2}{|\delta|}\right)^{k^2} \mathbb{E}\left[\left(\|\boldsymbol{M}_n\|^2 + \delta^2\right)^{k^2}\right]. \tag{2.16}$$

We now show that Lemma 2.5.1 implies Theorem 1.2.10. The proof of Lemma 2.5.1 is deferred to Section 2.5.2 and is the main technical work of the proof.

Proof of Theorem 1.2.10 assuming Lemma 2.5.1. Applying Corollary 2.4.2 and a union bound,

$$\begin{aligned}
&\mathbb{P}\left[\sigma_{n-k+1}(z - (A + \boldsymbol{M}_n)) \le \epsilon\right] \\
=&\mathbb{P}\left[\sigma_k\left((z - (A + \boldsymbol{M}_n))^{-1}\right) \ge 1/\epsilon\right] \\
\le&\mathbb{P}\left[\max_{S,T\subset[n],|S|=|T|=k} \sigma_k\left((z - (A + \boldsymbol{M}_n))^{-1}_{S,T}\right) \ge \frac{1}{k(n-k+1)\epsilon}\right]
\end{aligned}$$

$$\leq \sum_{S,T\subset[n],|S|=|T|=k} \mathbb{P}\left[\sigma_k\left((z-(A+\boldsymbol{M}_n))^{-1}_{S,T}\right) \geq \frac{1}{k(n-k+1)\epsilon}\right]. \tag{2.17}$$

Fixing $S, T \subset [n]$ of size k, there are permutation matrices P and Q such that

$$\begin{aligned}(z-(A+\boldsymbol{M}_n))^{-1}_{S,T} &= \left(Q^\intercal(z-(A+\boldsymbol{M}_n))^{-1}P\right)_{[k],[k]} \\ &= (PQ^\intercal i\operatorname{Im} z + P(\operatorname{Re} z-(A+\boldsymbol{M}_n))Q^\intercal)^{-1}_{[k],[k]}.\end{aligned}$$

As $PQ^\intercal$ is a permutation matrix and $P(\operatorname{Re} z-(A+\boldsymbol{M}_n))Q^\intercal$ satisfies Assumption 1.2.3 with parameter $K>0$, we can apply Lemma 2.5.1. Defining

$$C_{1.2.10} := 8\sqrt{3}(e\pi)^{3/2}, \tag{2.18}$$

this gives

$$\begin{aligned}&\mathbb{P}\left[\sigma_k\left((z-(A+\boldsymbol{M}_n))^{-1}_{S,T}\right) \geq \frac{1}{k(n-k+1)\epsilon}\right] \\ =&\mathbb{P}\left[\sigma_k\left(i\operatorname{Im} zPQ^\intercal - P(\operatorname{Re} z-(A+\boldsymbol{M}_n))Q^\intercal\right)^{-1}_{[k],[k]} \geq \frac{1}{k(n-k+1)\epsilon}\right] \\ \leq&(1+k^2)\left(C_{1.2.10}K^3n\frac{k^2(n-k+1)^2\epsilon^2}{|\operatorname{Im} z|}\right)^{k^2}\mathbb{E}\left[\left(\|P(\operatorname{Re} z-A+M_n)Q^\intercal\|^2+|\operatorname{Im} z|^2\right)^{k^2}\right] \\ \leq&(1+k^2)\left(C_{1.2.10}k^2n^3K^3\frac{\epsilon^2}{|\operatorname{Im} z|}\right)^{k^2}\mathbb{E}\left[\left(\|P(\operatorname{Re} z-(A+\boldsymbol{M}_n))Q^\intercal\|^2+|\operatorname{Im} z|^2\right)^{k^2}\right],\end{aligned}$$

where we have bounded $n-k+1\leq n$. By Jensen, $B_{\boldsymbol{M},s}\leq B_{\boldsymbol{M},t}$ for any random matrix $\boldsymbol{M}$ and $s\leq t$, and thus expanding out with the binomial theorem gives $B_{A+\boldsymbol{M},s}\leq B_{\boldsymbol{M},s}+\|A\|$ for every deterministic A. Finally,

$$\begin{aligned}\mathbb{E}\left[\left(\|P(\operatorname{Re} z-(A+\boldsymbol{M}_n))Q^\intercal\|^2+|\operatorname{Im} z|^2\right)^{k^2}\right] &= \mathbb{E}\left[\left(\|\operatorname{Re} z-(A+\boldsymbol{M}_n)\|^2+|\operatorname{Im} z|^2\right)^{k^2}\right] \\ &= \sum_{r=0}^{k^2}\binom{k^2}{r}B^{2r}_{\operatorname{Re} z-(A+\boldsymbol{M}_n),2r}|\operatorname{Im} z|^{2k^2-2r} \\ &\leq (B^2_{\operatorname{Re} z-(A+\boldsymbol{M}_n),2k^2}+|\operatorname{Im} z|^2)^{k^2} \\ &\leq \left((B_{\boldsymbol{M}_n,2k^2}+\|A\|+|\operatorname{Re} z|)^2+|\operatorname{Im} z|^2\right)^{k^2}.\end{aligned}$$

We finish by combining this with the previous equation, and multiplying by $\binom{n}{k}^2$ for the union bound over pairs of size-k subsets S and T. □

2.5.2 Proof of Lemma 2.5.1

In what follows we use the notation and assumptions of Lemma 2.5.1. In particular, $\boldsymbol{M}_n$ satisfies Assumption 1.2.3 with parameter $K > 0$, U is a permutation matrix, and $\delta \in \mathbb{R}$. Once again writing $\boldsymbol{N}_k$ for the upper left $k \times k$ block of $(\delta i U + \boldsymbol{M}_n)^{-1}$, we need to show that $\mathbb{P}[\|\boldsymbol{N}_k^{-1}\| \le \epsilon] = O(\epsilon^{2k^2})$. One would expect this behavior if the real and imaginary parts of $\boldsymbol{N}_k^{-1}$ were independent, and each had a density on $\mathbb{R}^{k\times k}$. We will not be quite so lucky, but we *will* be able to separate the randomness in its real and imaginary parts, obtaining the $O(\epsilon^{2k^2})$ behavior by conditioning on some well-chosen entries of $\boldsymbol{M}_n$. To make this precise, we will need some notation.

Let us write $\boldsymbol{M}_n$ and δU in the following block form:

$$\boldsymbol{M}_n = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{pmatrix} \quad \text{and} \quad \delta U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \tag{2.19}$$

where $\boldsymbol{M}_{11}$ and U_{11} are $k \times k$ matrices. Define as well the $(n-k) \times (n-k)$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$\boldsymbol{X} := \operatorname{Re}(\boldsymbol{M}_{22} + iU_{22})^{-1} \quad \text{and} \quad \boldsymbol{Y} := \operatorname{Im}(\boldsymbol{M}_{22} + iU_{22})^{-1}. \tag{2.20}$$

Applying the Schur complement formula to the block decomposition in (2.19), we get

$$\begin{aligned} \boldsymbol{N}_k^{-1} &= \boldsymbol{M}_{11} + iU_{11} - (\boldsymbol{M}_{12} + iU_{12})(\boldsymbol{M}_{22} + iU_{22})^{-1}(\boldsymbol{M}_{21} + iU_{21}) \\ &= \boldsymbol{M}_{11} + iU_{11} - (\boldsymbol{M}_{12} + iU_{12})(\boldsymbol{X} + i\boldsymbol{Y})(\boldsymbol{M}_{21} + iU_{21}), \end{aligned}$$

meaning that

$$\operatorname{Re} \boldsymbol{N}_k^{-1} = \boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{X}\boldsymbol{M}_{21} + U_{12}\boldsymbol{Y}\boldsymbol{M}_{21} - \boldsymbol{M}_{12}\boldsymbol{Y}U_{21} + U_{12}\boldsymbol{X}U_{21} \tag{2.21}$$

$$\operatorname{Im} \boldsymbol{N}_k^{-1} = U_{11} - \boldsymbol{M}_{12}\boldsymbol{Y}\boldsymbol{M}_{21} - \boldsymbol{M}_{12}\boldsymbol{X}U_{21} - U_{12}\boldsymbol{X}\boldsymbol{M}_{21} + U_{12}\boldsymbol{Y}U_{21}. \tag{2.22}$$

Examining these two formulae, and recalling that the entries of $\boldsymbol{M}_n$ are independent and have a joint density on $\mathbb{R}^{n\times n}$, we arrive at the key observation of this section:

Observation 2.5.2. The imaginary part $\operatorname{Im} \boldsymbol{N}_k^{-1}$ is independent of $\boldsymbol{M}_{11}$. Moreover, conditional on $\boldsymbol{M}_{12}, \boldsymbol{M}_{21}$ and $\boldsymbol{M}_{22}$, the real part $\operatorname{Re} \boldsymbol{N}_k^{-1}$ has independent entries, each with density on $\mathbb{R}$ bounded by $K\sqrt{n}$.

Writing this conditioning explicitly,

$$\begin{aligned} &\mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \ge 1/\epsilon\right] \\ =&\mathbb{P}\left[\|\boldsymbol{N}_k^{-1}\| \le \epsilon\right] \\ \le&\mathbb{P}\left[\|\operatorname{Re} \boldsymbol{N}_k^{-1} + i \operatorname{Im} \boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right] \\ \le&\mathbb{P}\left[\|\operatorname{Re} \boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}, \|\operatorname{Im} \boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right] \\ =&\mathbb{E}\,\mathbb{E}\left[\mathbb{1}\left\{\|\operatorname{Re} \boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right\} \mathbb{1}\left\{\|\operatorname{Im} \boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right\} \,\middle|\, \boldsymbol{M}_{12}, \boldsymbol{M}_{21}, \boldsymbol{M}_{22}\right] \end{aligned}$$

$$=\mathbb{E}\left[\mathbb{1}\left\{\|\operatorname{Im}\boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right\}\mathbb{E}\left[\mathbb{1}\left\{\|\operatorname{Re}\boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right\}\,\middle|\,\boldsymbol{M}_{12},\boldsymbol{M}_{21},\boldsymbol{M}_{22}\right]\right]. \tag{2.23}$$

We can bound the inner conditional expectation using Observation 2.5.2:

$$\mathbb{E}\left[\mathbb{1}\left\{\|\operatorname{Re}\boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right\}\,\middle|\,\boldsymbol{M}_{12},\boldsymbol{M}_{21},\boldsymbol{M}_{22}\right] \le \frac{(\sqrt{\pi k n}K\epsilon)^{k^2}}{\Gamma(k^2/2+1)} \le \left(\frac{\sqrt{2e\pi n}K\epsilon}{\sqrt{k}}\right)^{k^2} \tag{2.24}$$

In the final two steps we have used the volume of a Frobenius norm ball in $\mathbb{R}^{k\times k}$, and Stirling's approximation. Plugging into (2.23) gives

$$\mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \ge 1/\epsilon\right] \le \mathbb{P}\left[\|\operatorname{Im}\boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\right]\left(\frac{\sqrt{2e\pi n}K\epsilon}{\sqrt{k}}\right)^{k^2},$$

and we now turn to the more serious task of the requisite small-ball probability estimate for $\operatorname{Im}\boldsymbol{N}_k^{-1}$. This calculation is facilitated by a second key observation, which is an immediate consequence of the full expression (2.22) for $\operatorname{Im}\boldsymbol{N}_k^{-1}$.

Observation 2.5.3. Conditional on $\boldsymbol{M}_{22}$, the imaginary part $\operatorname{Im}\boldsymbol{N}_k^{-1}$ is a quadratic function in $\boldsymbol{M}_{12}$ and $\boldsymbol{M}_{21}$, of the type studied in Section 2.3.

In particular, for any deterministic $(n-k)\times(n-k)$ matrices Y and X, and j satisfying $n-k \ge j > k^2+k+1$, Theorem 2.3.1 implies

$$\begin{aligned}\mathbb{P}\left[\|U_{12}-\boldsymbol{M}_{12}Y\boldsymbol{M}_{21}-\boldsymbol{M}_{12}XU_{21}-U_{12}X\boldsymbol{M}_{21}+U_{12}YU_{21}\|_F \le \epsilon\sqrt{k}\right]&\\ \le (1+k^2)\left(\frac{2K^2n\sqrt{2e\pi k}}{\sqrt{j-k+1}\sigma_j(\boldsymbol{Y})}\right)^{k^2}\left(\frac{\sqrt{2e\pi}\epsilon}{\sqrt{k}}\right)^{k^2}&\\ = (1+k^2)\left(\frac{4K^2n\cdot e\pi\cdot\varepsilon}{\sqrt{j-k+1}\sigma_j(\boldsymbol{Y})}\right)^{k^2},&\end{aligned} \tag{2.25}$$

(again using the volume of a Frobenius norm ball). Since $\boldsymbol{Y}$ depends only on the randomness in $\boldsymbol{M}_{22}$, and is thus independent of $\boldsymbol{M}_{12}$ and $\boldsymbol{M}_{21}$, conditioning and integrating over $\boldsymbol{M}_{22}$ gives us

$$\mathbb{P}\left[\|\operatorname{Im}\boldsymbol{N}_k^{-1}\| \le \epsilon\right] \le (1+k^2)\left(\frac{4K^2n\cdot e\pi\cdot\varepsilon}{\sqrt{j-k+1}}\right)^{k^2}\mathbb{E}\left[\sigma_j(\boldsymbol{Y})^{-k^2}\right]. \tag{2.26}$$

To finish the proof, we now need to bound this remaining expectation for a suitable choice of j, satisfying $n-k \ge j > k^2+k+1$. In (2.20), we defined $\boldsymbol{Y} = \operatorname{Im}(\boldsymbol{M}_{22}+iU_{22})^{-1}$, and we now require a more explicit formula. Using the representation of $\mathbb{C}^{(n-1)\times(n-1)}$ as a set of block matrices in $\mathbb{R}^{2(n-1)\times 2(n-1)}$, and again applying the Schur complement formula,

$$\begin{pmatrix}\boldsymbol{X} & -\boldsymbol{Y}\\ \boldsymbol{Y} & \boldsymbol{X}\end{pmatrix} = \begin{pmatrix}\boldsymbol{M}_{22} & -U_{22}\\ U_{22} & \boldsymbol{M}_{22}\end{pmatrix}^{-1}$$

$$= \begin{pmatrix} (\boldsymbol{M}_{22} + U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1} & (\boldsymbol{M}_{22} + U_{2,2}\boldsymbol{M}_{22}^{-1}U_{22})^{-1}U_{22}\boldsymbol{M}_{22}^{-1} \\ -(\boldsymbol{M}_{22} + U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1}U_{22}\boldsymbol{M}_{22}^{-1} & (\boldsymbol{M}_{2,2} + U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1} \end{pmatrix}$$

and hence

$$\boldsymbol{Y} = -(\boldsymbol{M}_{22} + U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1}U_{22}\boldsymbol{M}_{22}^{-1}. \tag{2.27}$$

If we could invert U_{22}, we could rewrite this as $-(\boldsymbol{M}_{22}U_{22}^{-1}\boldsymbol{M}_{22} + U_{22})^{-1}$ and set $j = n - k$, giving

$$\sigma_{n-k}\left(\boldsymbol{Y}\right)^{-k^2} = \|\boldsymbol{M}_{22}U_{22}\boldsymbol{M}_{22} + U_{22}\|^{k^2} \leq \left(|\delta|^{-1}\|\boldsymbol{M}_{22}\|^2 + |\delta|\right)^{k^2} \leq \left(|\delta|^{-1}\|\boldsymbol{M}_n\|^2 + |\delta|\right)^{k^2}.$$

However, not every principal block of a permutation matrix is invertible, so we will need to work a bit harder.

Since U is a permutation matrix, and U_{22} is an $(n-k)\times(n-k)$ block of δU, by the usual interlacing of singular values for submatrices [89, Corollary 7.3.6], we can be sure that $\sigma_1(U_{22}) = \cdots = \sigma_{n-2k}(U_{22}) = |\delta|$. Hence, there exists a matrix E of rank at most $2k$ such that $\hat{U}_{22} := U_{22} + E$ is invertible, with all singular values equal to $|\delta|$. We can therefore write

$$\boldsymbol{Y} = -(\boldsymbol{M}_{22} + U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1}U_{22}\boldsymbol{M}_{22}^{-1} = -(\boldsymbol{M}_{22} + \hat{U}_{22}\boldsymbol{M}_{22}^{-1}U_{22} + \boldsymbol{E}_1)^{-1}\hat{U}_{22}\boldsymbol{M}_{22}^{-1} + \boldsymbol{E}_2$$

where $\boldsymbol{E}_1 = -E\boldsymbol{M}_{22}^{-1}U_{22}$ and $\boldsymbol{E}_2 = -(\boldsymbol{M}_{22} - U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1}E\boldsymbol{M}_{22}^{-1}$. Since $\operatorname{rank}(\boldsymbol{E}_2) \leq \operatorname{rank}(E) \leq 2k$, interlacing of singular values upon low-rank updates [152, Theorem 1] ensures

$$\sigma_j(\boldsymbol{Y}) \geq \sigma_{j+2k}\left((\boldsymbol{M}_{22} + \hat{U}_{22}\boldsymbol{M}_{22}^{-1}U_{22} + \boldsymbol{E}_1)^{-1}\hat{U}_{22}\boldsymbol{M}_{22}^{-1}\right). \tag{2.28}$$

On the other hand

$$(\boldsymbol{M}_{22} + \hat{U}_{22}\boldsymbol{M}_{22}^{-1}U_{22} + \boldsymbol{E}_1)^{-1}\hat{U}_{22}\boldsymbol{M}_{22}^{-1} = (\boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22} + M_{22}\hat{U}_{22}^{-1}\boldsymbol{E}_1)^{-1}, \tag{2.29}$$

and since $\operatorname{rank}(\boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{E}_1) \leq \operatorname{rank}(\boldsymbol{E}_1) \leq \operatorname{rank}(E) \leq 2k$, a further application of the low-rank update bound tells us

$$\sigma_{j+2k}\left((\boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22} + \boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{E}_1)^{-1}\right) \geq \sigma_{j+4k}\left((\boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22})^{-1}\right). \tag{2.30}$$

Putting together (2.28), (2.29), and (2.30), we get

$$\sigma_j(\boldsymbol{Y}) \geq \sigma_{j+4k}\left((\boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22})^{-1}\right),$$

and finally, setting $j = n - 5k$, and recalling $\|U_{2,2}\| = |\delta|$, $\|\hat{U}_{2,2}^{-1}\| = |\delta|^{-1}$, and $\|\boldsymbol{M}_{22}\| \leq \|\boldsymbol{M}_n\|$, we have

$$\sigma_{n-5k}(\boldsymbol{Y})^{-k^2} \leq \left\|\boldsymbol{M}_{22}\hat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22}\right\|^{k^2} \leq \left(|\delta|^{-1}\|\boldsymbol{M}_n\|^2 + |\delta|\right)^{k^2} \tag{2.31}$$

We now assemble our work so far:

$$\begin{aligned}
&\mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \geq 1/\epsilon\right] \\
\leq &\mathbb{P}\left[\|\operatorname{Im} \boldsymbol{N}_k^{-1}\| \leq \epsilon\right] \left(\frac{\sqrt{2e\pi n}K\epsilon}{\sqrt{k}}\right)^{k^2} \\
\leq &(1+k^2)\left(\frac{4K^2 n \cdot e\pi \cdot \epsilon}{\sqrt{j-k+1}}\right)^{k^2} \left(\frac{\sqrt{2e\pi n}K\epsilon}{\sqrt{k}}\right)^{k^2} \mathbb{E}\left[\sigma_j(\boldsymbol{Y})^{-k^2}\right] \quad \forall n-k \geq j \geq k^2+k+1 \\
\leq &(1+k^2)\left(\frac{4\sqrt{2}K^3(e\pi n)^{3/2}}{\sqrt{k(n-6k+1)}}\right)^{k^2} \left(\frac{\epsilon^2}{|\delta|}\right)^{k^2} \mathbb{E}\left(\|\boldsymbol{M}_n\| + \delta^2\right)^{k^2} \quad \text{setting } j = n-5k.
\end{aligned}$$

For this to go through, we need $n \geq \max\{6k, (k+2)^2\} = (k+2)^2$. Finally, we can use $1/(n-6k+1) \leq 6k/n$ to obtain the final result.

2.6 Lower Bounds on the Minimum Eigenvalue Gap

This section is devoted to several results regarding eigenvalue gaps of real random matrices with independent entries, and its main goal is to prove Theorem 1.2.5.

As discussed in Section 1.2 the proof strategy for obtaining tail bounds on $\operatorname{gap}(A + \boldsymbol{M}_n)$ is to use an ϵ-net argument to show that in the event that $\operatorname{gap}(A + M_n)$ is small, there will be some z in the net for which $\sigma_n(z - \boldsymbol{M}_n)\sigma_{n-1}(z - \boldsymbol{M}_n)$ is small, after which one uses the tail bounds for $\sigma_n(z - A - \boldsymbol{M}_n)$ and $\sigma_{n-1}(z - A - \boldsymbol{M}_n)$ obtained above to argue that this is an unlikely event. The main complication here is that our tail bounds on the singular values of $z - A - \boldsymbol{M}_n$ depend on the shift z: on the real line they are governed by Theorem 2.0.2, and away from it by Theorem 1.2.10. To handle this, we will use a combination of nets, exploiting the fact that real matrices have conjugate-symmetric spectra. Specifically, this symmetry means that we can think of small gaps as arising in one of three different ways: gaps in which at least one eigenvalue is real, gaps between a conjugate pair of eigenvalues with small imaginary part, and gaps between complex eigenvalues away from the real line. Thus motivated, let us define, for any matrix $M \in \mathbb{R}^{n\times n}$ and $\delta > 0$,

$$\begin{aligned}
\operatorname{gap}_{\mathbb{R}}(M) &:= \min\left\{|\lambda_i(M) - \lambda_j(M)| : i \neq j \text{ and } \lambda_i(M) \in \mathbb{R}\right\}, \\
\operatorname{Im}_{\min}(M) &:= \min\left\{|\operatorname{Im}\lambda_i(M)| : \lambda_i(M) \notin \mathbb{R}\right\}, \\
\operatorname{gap}_{\operatorname{Im}\geq\delta}(M) &:= \min\left\{|\lambda_i(M) - \lambda_j(M)| : i \neq j \text{ and } |\operatorname{Im}\lambda_i(M)|, |\operatorname{Im}\lambda_j(M)| \geq \delta\right\}.
\end{aligned}$$

Proof of Theorem 1.2.5 . For most of the proof, let us absorb γ into the constant K—the condition $\gamma < 1/K$ will not be relevant until the end.

First observe that if $\delta > 0$,

$$\{\operatorname{gap}(A + \boldsymbol{M}_n) \leq s\} \tag{2.32}$$

$$=\{\mathrm{gap}_{\mathbb{R}}(A+\boldsymbol{M}_n)\le s\}\cup\{\mathrm{Im}_{\min}(A+\boldsymbol{M}_n)\le\delta\}\cup\{\mathrm{gap}_{\mathrm{Im}\ge\delta}(A+\boldsymbol{M}_n)\le s\}.$$

Now choose a covering of the region $D(0,R)\subset\mathbb{C}$ with disks, whose centers will form the net, with the property that any pair of eigenvalues at distance less than s must both lie in at least one of them. In view of (2.32), we will set up a separate net to union bound each of the events appearing on the right-hand side: let

$$\begin{aligned}\mathcal{N}_\eta^{\mathbb{R}}&:=\{j\eta : j\in\mathbb{Z}\}\cap[-R,R]\\ \mathcal{N}_{\delta,\eta}^{\mathbb{C}}&:=\{\eta j+i(\delta+\eta k) : j,k\in\mathbb{Z}\}\cap B(0,R).\end{aligned}$$

Then, judiciously choosing the spacing and radii of disks, for any $\delta>0$ we have:

$$\begin{aligned}\mathbb{P}\left[\mathrm{gap}(A+\boldsymbol{M}_n)\le s\right]\le&\sum_{z\in\mathcal{N}_{2s}^{\mathbb{R}}}\mathbb{P}\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,3s/2)|\ge 2\right]\\&+\sum_{z\in\mathcal{N}_{\delta}^{\mathbb{R}}}\mathbb{P}\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,\sqrt{2}\delta)|\ge 2\right]\\&+\sum_{z\in\mathcal{N}_{\delta,s}^{\mathbb{C}}}\mathbb{P}\left[\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,\sqrt{5/4}s)|\ge 2\right]\\&+\mathbb{P}\left[\|A+\boldsymbol{M}_n\|\ge R\right].\end{aligned}\tag{2.33}$$

The first line controls $\mathrm{gap}_{\mathbb{R}}$, the second one $\mathrm{Im}_{\min}$, the third one $\mathrm{gap}_{\mathrm{Im}\ge\delta}$, and the final one the event that some eigenvalue lies outside the region covered by our net. One could further optimize the above in the pursuit of tighter constants, but we optimize for simplicity. The remainder of the proof consists of bounding these events with Theorems 2.0.2 and 1.2.10—the constants and exponents become somewhat unwieldy, and on a first reading we recommend following the argument at a high level to avoid being bogged down in technicalities.

Step 1: Gaps on the Real Line. We first must bound the probability

$$\mathbb{P}\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,3s/2)|\ge 2\right]$$

for $z\in\mathbb{R}$. To use Lemma 1.2.7, we need tail bounds for the *product* of the two smallest singular values of $z-(A+\boldsymbol{M}_n)$, whereas Theorem 2.0.2 concerns individual singular values. To get around this, note that for every $z\in\mathbb{R}$ and $x>0$,

$$\begin{aligned}&\mathbb{P}\left[\sigma_n(z-(A+\boldsymbol{M}_n))\sigma_{n-1}(z-(A+\boldsymbol{M}_n))\le r^2\right]\\\le&\mathbb{P}\left[\sigma_n(A+\boldsymbol{M}_n)\le rx\right]+\mathbb{P}\left[\sigma_{n-1}(A+\boldsymbol{M}_n)\le r/x\right]\\\le&2Kn^2rx+16K^4n^6r^4/x^4.\end{aligned}$$

Optimizing in x, we have

$$\mathbb{P}\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,r)|\ge 2\right]$$

$$\leq \left(4^{1/5} + 4^{-4/5}\right)(2Kr)^{8/5} n^{14/5} \leq 3n^{14/5}(\sqrt{2}Kr)^{8/5}. \tag{2.34}$$

The rough bound $\left|\mathcal{N}_{2s}^{\mathbb{R}}\right| \leq (R/s+1) \leq 3R/2s$ now gives

$$\begin{aligned}\sum_{z\in\mathcal{N}_s^{\mathbb{R}}} \mathbb{P}\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,3s/2)|\geq 2\right] &\leq \left|\mathcal{N}_s^{\mathbb{R}}\right|\cdot 3n^{14/5}(3\sqrt{2}Ks/2)^{8/5} \\ &\leq 9R(\sqrt{2}K)^{8/5}n^{14/5}s^{3/5}. \end{aligned}\tag{2.35}$$

Step 2: Eigenvalues Near the Real Line. Using (2.34) and imitating the remainder of Step 1,

$$\sum_{z\in\mathcal{N}_\delta^{\mathbb{R}}} P\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,\sqrt{2}\delta)|\geq 2\right] \leq 8R(\sqrt{2}K)^{8/5}n^{14/5}\delta^{3/5} \tag{2.36}$$

This directly implies a stand-alone tail bound on $\mathrm{Im}_{\min}$, which we record for use in Section 4.2.4,:

$$\mathbb{P}\left[\mathrm{Im}_{\min}(A+\boldsymbol{M}_n)\leq\delta\right] \leq 8R(\sqrt{2}K)^{8/5}n^{14/5}\delta^{3/5} + \mathbb{P}[\|\boldsymbol{M}_n\|\geq R]. \tag{2.37}$$

Step 3: Eigenvalues Away from the Real Line. We finally turn to non-real z. As in Step 1, observe that for any $z\in\mathbb{C}\setminus\mathbb{R}$, $r>0$, and $n\geq 16$, Theorem 1.2.10 implies

$$\begin{aligned}&\mathbb{P}\left[|\|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,r)|\geq 2\right] \qquad (2.38)\\ \leq &\min_{x>0}\left\{\mathbb{P}\left[\sigma_n(A+\boldsymbol{M}_n)\leq rx\right]+\mathbb{P}\left[\sigma_{n-1}(A+\boldsymbol{M}_n)\leq r/x\right]\right\}\\ \leq &\min_{x>0}\left\{2C_{1.2.10}K^3n^5\left(\left(B_{\boldsymbol{M}_n,2}+\|A\|+|\operatorname{Re}z|\right)^2+|\operatorname{Im}z|^2\right)\frac{(rx)^2}{|\operatorname{Im}z|}\right.\\ &\qquad\left.+640C_{1.2.10}^4K^{12}n^{14}\left(\left(B_{\boldsymbol{M}_n,8}+\|A\|+|\operatorname{Re}z|\right)^2+|\operatorname{Im}z|^2\right)^4\frac{r^8}{x^8|\operatorname{Im}z|^4}\right\}\\ \leq &C_{(2.39)}\left(\frac{\left(B_{\boldsymbol{M}_n,8}+\|A\|+|\operatorname{Re}z|\right)^2+|\operatorname{Im}z|^2}{|\operatorname{Im}z|}\right)^{8/5}K^{24/5}r^{16/5}n^{34/5} \qquad (2.39)\end{aligned}$$

where we have used $B_{\boldsymbol{M}_n,1}\leq B_{\boldsymbol{M}_n,8}$ and defined $C_{(2.39)} = 11C_{1.2.10} = 88\sqrt{3}(e\pi)^{3/2}$.

Finally, observing that every $z\in\mathcal{N}_{\delta,s}^{\mathbb{C}}$ has $|\operatorname{Im}z|>\delta$ and $|z|\leq R$, we have

$$\begin{aligned}&\sum_{z\in\mathcal{N}_{\delta,s}^{\mathbb{C}}}\mathbb{P}\left[|\mathrm{Spec}(A+\boldsymbol{M}_n)\cap D(z,\sqrt{5/4}s)|\geq 2\right]\\ &\qquad\leq 6(R/s)^2C_{(2.39)}\left(\frac{(B_{\boldsymbol{M}_n,8}+\|A\|+R)^2}{\delta}\right)^{8/5}K^{24/5}(\sqrt{5}s/2)^{16/5}n^{34/5}\\ &\qquad\leq C_{(2.40)}R^2(B_{\boldsymbol{M}_n,8}+\|A\|+R)^{16/5}\frac{K^{24/5}s^{6/5}n^{34/5}}{\delta^{8/5}}\end{aligned}\tag{2.40}$$

where $C_{(2.40)} := 6(5/4)^{8/5}C_{(2.39)} = 528(5/4)^{8/5}\sqrt{3}(e\pi)^{3/2}$.

Step 4: Conclusion. We now put together the three steps above, substituting (2.35), (2.36), and (2.40) into (2.33), and adding back in the γ scaling. Using the fact that $\psi\delta^{3/5} + \phi s^{6/5}\delta^{-8/5} \leq 2\psi^{8/11}\phi^{3/11}s^{18/55}$, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\operatorname{gap}(A+\gamma \boldsymbol{M}_n) \leq s\right] \qquad (2.41)\\
\leq& 9R(\sqrt{2}K/\gamma)^{8/5}n^{14/5}s^{3/5}\\
&+ 2\left(C_{(2.40)}R^2(\gamma B_{\boldsymbol{M}_n,8} + \|A\| + R)^2(K/\gamma)^{24/5}n^{34/5}\right)^{3/11}\left(8(\sqrt{2}K/\gamma)^{8/5}n^{14/5}\right)^{8/11}s^{18/55}\\
&+ \mathbb{P}\left[\|A+\gamma \boldsymbol{M}_n\| \geq R\right]\\
\leq& C_{1.2.5}R^{14/11}\left(\gamma B_{\boldsymbol{M}_n,8} + \|A\| + R\right)^{6/11}(K/\gamma)^{136/55}n^{214/55}s^{18/55} + \mathbb{P}\left[\|A+\boldsymbol{M}_n\| \geq R\right]\\
\leq& C_{1.2.5}R^2\left(\gamma B_{\boldsymbol{M}_n,8} + \|A\| + R\right)(K/\gamma)^{5/2}n^4 s^{2/7} + \mathbb{P}\left[\|A+\boldsymbol{M}_n\| \geq R\right], \qquad (2.42)
\end{aligned}$$

where

$$C_{1.2.5} := 2C_{(2.40)}^{3/11}\cdot 8^{8/11}\sqrt{2}^{64/55} + 9\sqrt{2}^{8/5} < 250. \qquad (2.43)$$

□

2.7 Upper Bounds on the Eigenvalue Condition Numbers

In this section, we follow the Hermitization strategy presented in Section 1.2 to convert our probabilistic lower bounds on the least singular value into upper bounds on the mean eigenvalue condition numbers. As mentioned above, this makes use of the relationship between the eigenvalue condition numbers and the area of the ϵ-pseudospectrum given in Lemma 1.1.10. However, since we will treat separately real and complex eigenvalues, in addition to Lemma 1.1.10, we will need an easy variant relating pseudospectrum on the real line to the condition numbers of *real* eigenvalues.

Lemma 2.7.1 (Limiting Length of Pseudospectrum on Real Line)**.** *Let $M \in \mathbb{R}^{n\times n}$ have n distinct eigenvalues $\lambda_1, ..., \lambda_n$. Let $\ell_{\mathbb{R}}$ denote the Lebesgue measure on $\mathbb{R}$, and let $\Omega \subset \mathbb{R}$ be an open set. Then*

$$2\sum_{\lambda_i\in\Omega}\kappa(\lambda_i) \leq \liminf_{\epsilon\to 0}\frac{\ell_{\mathbb{R}}\left(\Lambda_\epsilon(M)\cap\Omega\right)}{\epsilon}$$

Proof. For each $z \in \mathbb{C}$ and $r \geq 0$, let $D(z, r)$ denote the closed disk centered at z of radius r. In the proof of [15, Lemma 3.2] it is shown that if M has n distinct eigenvalues,

$$\bigcup_{i=1}^{n} D(\lambda_i, \kappa(\lambda_i)\epsilon - O(\epsilon^2)) \subseteq \Lambda_\epsilon(M) \subseteq \bigcup_{i=1}^{n} D(\lambda_i, \kappa(\lambda_i)\epsilon + O(\epsilon^2)).$$

In particular, each $\lambda_i \in \Omega$ contributes at least $2\kappa(\lambda)\epsilon - O(\epsilon^2)$ to the measure of $\Lambda_\epsilon \cap \Omega$. Taking $\epsilon \to 0$ yields the conclusion. □

In both Lemma 1.1.10 and Lemma 2.7.1, if the boundary of Ω contains none of the eigenvalues, one actually has equality, the limit inferior can be replaced by the limit, and Ω need only be measurable, but we will not need this fact.

2.7.1 Bounds in Expectation

We now come to the first main proposition of this section.

Proposition 2.7.2 ($\kappa(\lambda_i)$ on the real line)**.** *Let $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{M}_n$ satisfy Assumption 1.2.3 with parameter $K > 0$. Write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \gamma\boldsymbol{M}_n$. Then for every open set $\Omega \subset \mathbb{R}$,*

$$\mathbb{E}\sum_{\boldsymbol{\lambda}_i\in\Omega} \kappa(\boldsymbol{\lambda}_i) \le \frac{Kn^2}{\sqrt{2}\gamma} \cdot \ell_{\mathbb{R}}(\Omega).$$

Proof. When z is real, $z - A$ is also real, so we may apply the tail bound in Corollary 2.4.3. In particular, setting $k = 1$, we obtain the following tail bound for real z:

$$\mathbb{P}[\sigma_n((z - A) + \gamma(-\boldsymbol{M}_n)) \le \epsilon] < \frac{\sqrt{2}Kn^2\epsilon}{\gamma}.$$

Since the eigenvalues of $z - (A + \gamma\boldsymbol{M}_n)$ are distinct with probability 1, we have

$$\begin{aligned}
2\mathbb{E}\sum_{\lambda_i\in\Omega} \kappa(\lambda_i) &\le \mathbb{E}\liminf_{\epsilon\to 0} \epsilon^{-1}\ell_{\mathbb{R}}\left(\Lambda_\epsilon(A + \gamma\boldsymbol{M}_n) \cap \Omega\right) && \text{Lemma 2.7.1}\\
&\le \liminf_{\epsilon\to 0} \epsilon^{-1}\mathbb{E}\int_\Omega \mathbb{1}\left\{z \in \Lambda_\epsilon(A + \gamma\boldsymbol{M}_n)\right\} dz && \text{Fatou's lemma}\\
&= \liminf_{\epsilon\to 0} \epsilon^{-1}\int_\Omega \mathbb{P}[z \in \Lambda_\epsilon(A + \gamma\boldsymbol{M}_n)]\, dz && \text{Fubini's theorem}\\
&= \liminf_{\epsilon\to 0} \epsilon^{-1}\int_\Omega \mathbb{P}[\sigma_n(z - (A + \gamma\boldsymbol{M}_n)) < \epsilon]\, dz\\
&\le \frac{\sqrt{2}Kn^2}{\gamma}\ell_{\mathbb{R}}(\Omega). && \text{Corollary 2.4.3}
\end{aligned}$$

□

We now give the analogous proposition for the nonreal eigenvalues.

Proposition 2.7.3 ($\kappa(\lambda_i)$ away from real line)**.** *Let $n \geq 9$. Let $A \in \mathbb{R}^{n\times n}$ be deterministic. Let $\boldsymbol{M}_n$ satisfy Assumption 1.2.3 with parameter $K > 0$. Let $\gamma > 0$, and write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \gamma \boldsymbol{M}_n$. Then for every open set $\Omega \subseteq \mathbb{C} \setminus \mathbb{R}$,*

$$\mathbb{E} \sum_{\boldsymbol{\lambda}_i \in \Omega} \kappa(\boldsymbol{\lambda}_i)^2 \leq \frac{C_{1.2.10} K^3 n^5}{\gamma^3} \int_\Omega \frac{(\gamma \mathbb{E}\|\boldsymbol{M}_n\| + \|A\| + |\operatorname{Re} z|)^2 + |\operatorname{Im} z|^2}{|\operatorname{Im} z|}\, dz.$$

In the special case where $\boldsymbol{M}_n$ is real Ginibre, one may take $n \geq 7$ and replace the term $C_{1.2.10}K^3$ with $\frac{\sqrt{7e}}{4\pi}$.

Proof. In the proof of Proposition 2.7.2, since $\Omega \subseteq \mathbb{C} \setminus \mathbb{R}$ we replace Lemma 2.7.1 with Lemma 1.1.10. Since z is no longer real we must also replace the singular value tail bound in Corollary 2.4.3 with the one in Theorem 1.2.10. □

2.7.2 Bounds with high probability: Proof of Theorem 2.0.1

In the notation of Theorem 2.0.1, $R, \|A\|, K$, and γ will be $\Theta(1)$ in most applications, so ϵ_1 and ϵ_2 may be set to $1/n^D$ for sufficiently high D.

Proof of Theorem 2.0.1. From here on out, assume that each of

$$\sum_{\boldsymbol{\lambda}_i \in \mathbb{R}} \kappa(\boldsymbol{\lambda}_i) \quad \text{and} \quad \sum_{\boldsymbol{\lambda}_i \in \mathbb{C}\setminus\mathbb{R}} \kappa(\boldsymbol{\lambda}_i)^2$$

is at most ϵ_1^{-1} times its expectation; by Markov's inequality and a union bound this happens with probability at least $1 - 2\epsilon_1$.

Let $\delta \in (0, R)$ be a small parameter to be optimized later. Let $L := \|A\| + R$, and define the regions $\Omega_\mathbb{R}$ and $\Omega_\mathbb{C}$ as follows:

$$\Omega_\mathbb{R} := \{x \in \mathbb{R} : |x| < L\}$$

$$\Omega_\mathbb{C} := \{x + yi : x \in \mathbb{R} \text{ and } \delta < |y| < L.\}$$

Write E_{bound} for the event that $\gamma\|\boldsymbol{M}_n\| < R$ and let E_{strip} denote the event that $\operatorname{Im}_{\min}(A + \gamma \boldsymbol{M}_n) > \delta$. Then with probability at least $1 - 2\epsilon_1 - \mathbb{P}[E_{\text{bound}}] - \mathbb{P}[E_{\text{strip}}]$, all eigenvalues of $A + \gamma M_n$ are contained in $\Omega_\mathbb{R} \cup \Omega_C$, so

$$\sum_{\boldsymbol{\lambda}_i \in \mathbb{R}} \kappa(\boldsymbol{\lambda}_i) = \sum_{\boldsymbol{\lambda}_i \in \Omega_\mathbb{R}} \kappa(\boldsymbol{\lambda}_i) \leq \frac{Kn^2}{\sqrt{2}\gamma} \ell_\mathbb{R}(\Omega_\mathbb{R}) \leq \frac{\sqrt{2}Kn^2 L}{\gamma}$$

and

$$\sum_{\boldsymbol{\lambda}_i \in \mathbb{C}\setminus\mathbb{R}} \kappa(\boldsymbol{\lambda}_i)^2 = \sum_{\boldsymbol{\lambda}_i \in \Omega_\mathbb{C}} \kappa(\boldsymbol{\lambda}_i)^2$$

$$
\begin{aligned}
&\leq \frac{CK^3n^5}{\gamma^3}\int_{\Omega_{\mathbb{C}}} \frac{(\gamma\mathbb{E}\|M_n\| + \|A\| + |\operatorname{Re} z|)^2 + |\operatorname{Im} z|^2}{|\operatorname{Im} z|}\,dz \\
&\leq 2\frac{CK^3n^5}{\gamma^3}\int_{\delta}^{L}\int_{-L}^{L} \frac{(\gamma\mathbb{E}\|M_n\| + \|A\| + |x|)^2 + y^2}{y}\,dx\,dy \\
&\leq 2\frac{CK^3n^5}{\gamma^3}\int_{\delta}^{L} 2L\frac{(2L)^2 + L^2}{y}\,dy \\
&= 20\frac{CK^3n^5}{\gamma^3}L^3(\ln L + \ln(1/\delta)).
\end{aligned}
$$

Recall from (2.37) that

$$\mathbb{P}[E_{\text{strip}}] = O(RK^{8/5}n^{14/5}\delta^{3/5}/\gamma^{8/5}) + \mathbb{P}[\gamma\|\boldsymbol{M}_n\| \geq R],$$

so setting $\delta = L\epsilon_2$ yields the result. □

The proof of Theorem 1.2.6 advertised in the introduction is now trivial.

Proof of Theorem 1.2.6. To get a bound on κ_V invoke inequality (1.8) to obtain

$$\kappa_V(A + \gamma\boldsymbol{M}_n) \leq \sqrt{n\sum_{i=1}^{n}\kappa(\boldsymbol{\lambda}_i)^2} \leq \sqrt{n}\sqrt{\left(\sum_{\boldsymbol{\lambda}_i\in\mathbb{R}}\kappa_i(\boldsymbol{\lambda}_i)\right)^2 + \sum_{\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}}\kappa(\boldsymbol{\lambda}_i)^2}.$$

Then apply Theorem 2.0.1. □

Chapter 3

Spectral Bisection

Here we will provide the proofs for Theorems 1.3.1 and 1.3.3 discussed in Section 1.3.

After discussing the related work in Section 3.1, in Section 3.2 we will discuss some finite arithmetic considerations and define the subroutines that will be used as a black-box when analyzing the main algorithm. In Section 3.3 we will tailor the smoothed analysis results on gap and κ_V to this context. The main technical part of this chapter will be presented in Section 3.4, where we give rigorous guarantees for computing the sign function via Roberts' iteration. Finally, in Section 3.5 we will put everything together to provide general guarantees for the spectral bisection algorithm. The analysis of some of the subroutines will be deferred to Appendix B.

3.1 Related Work

Smoothed Analysis. The study of numerical algorithms on Gaussian random matrices (i.e., the case $A = 0$ of smoothed analysis) dates back to [165, 142, 56, 62]. 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 [144], in the context of the simplex algorithm. Sankar, Spielman, and Teng [137] showed that adding real Gaussian noise to any matrix yields a matrix with polynomially-bounded condition number; [15] 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 [3]) 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 [143] and Haagerup and Larsen [82] 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 [22, 23, 58] and Roberts [131], who used it first to solve the algebraic Ricatti and Lyapunov equations and then as an approach to the eigenproblem; see [95] for a broad survey of its early history. The numerical stability of Roberts' Newton iteration was investigated by Byers [34], who identified some cases where it is and isn't stable. Malyshev [107], Byers, He, and Mehrmann [35], Bai, Demmel, and Gu [7], and Bai and Demmel [6] 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[1], but did not prove convergence in finite arithmetic and left this as an open question.[2] 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 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. –[86, Section 5.7]

This is precisely the problem solved by Theorem 1.3.3, 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 [87] established efficient reductions between the sign function and the polar decomposition. Byers and Xu [36] 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 [113] (see also the improvement [112]) 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). As mentioned in Section 1.3, the work of Armentano, Beltrán, Bürgisser, Cucker, and Shub [3] was the first to provide a way of solving the eigenvalue problem with provable guarantees, which ensure a running time of $O(n^{10}/\delta^2)$, where δ is the final accuracy. 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.

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

[2] [35] states: "A priori backward and forward error bounds for evaluation of the matrix sign function

Result	Error	Arithmetic Ops	Boolean Ops
[125]	B	$n^3 + n^2 \log(1/\delta)$	$n^3 \log(n/\delta) + n^2 \log(1/\delta) \log(n/\delta)$
[3]	B	n^{10}/δ^2	$n^{10}/\delta^2 \cdot \text{polylog}(n/\delta)$[a]
[25]	B	$n^{\omega+1}\text{polylog}(n)\log(1/\delta)$	$n^{\omega+1}\text{polylog}(n)\log(1/\delta)$
Thm. 1.3.1 [b]	B	$T_{\mathsf{MM}}(n)\log^2(n/\delta)$	$T_{\mathsf{MM}}(n)\log^6(n/\delta)\log(n)$
Cor. 1.3.2	F	$T_{\mathsf{MM}}(n)\log^2(n\kappa_{\text{eig}}/\delta)$	$T_{\mathsf{MM}}(n)\log^6(n\kappa_{\text{eig}}/\delta)\log(n)$

B=Backward, F=Forward.
[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. The works appearing in the first and third item only apply to Hermitian matrices.

Result	Model	Error	Arithmetic Ops	Boolean Ops
[37]	Rat.	F[a]	$\text{poly}(a, n, \log(1/\delta))$[b]	$\text{poly}(a, n, \log(1/\delta))$
[120]	Rat.	F	$n^\omega + n \log\log(1/\delta)$	$n^{\omega+1}a + n^2 \log(1/\delta)\log\log(1/\delta)$
[105]	Fin.[c]	F	$n^\omega \log(n)\log(1/\delta)$	$n^\omega \log^4(n)\log^2(n/\delta)$

Rat.=Rational, Fin.=Finite, F=Forward.
[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. The work in the second item only addresses how to compute eigenvalues (not eigenvectors), and uses a custom bit representation for intermediate quantities. The work in the last item applies to Hermitian matirces only, and it is only about computing λ_1.

Non-Hermitian Eigenproblem (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 [37] that given a rational $n \times n$ matrix A with integer entries of bit length a, one can find an δ-forward approximation to its Jordan Normal Form $A = VJV^{-1}$ in time $\text{poly}(n, a, \log(1/\delta))$, 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.

remain elusive."

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 [120] 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.[3]

Finally, the important work of Demmel, Dumitriu, and Holtz [53] (see also the followup [8]), which we rely on heavily, does not claim to provably solve the eigenproblem either—it 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 [174], who exhibited a globally convergent diagonalization algorithm, and the work of Dekker and Traub [52] who quantified the rate of convergence of Wilkinson's algorithm and from which it follows that the Hermitian eigenproblem can be solved with backward error δ in $O(n^3 + n^2 \log(1/\delta))$ arithmetic operations in exact arithmetic.[4]. We refer the reader to [125, §8.10] for the simplest and most insightful proof of this result, due to Hoffman and Parlett [88]

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 [105] 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 [25] 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 3.1.1 (Davies' Conjecture)**.** The beautiful paper [44] 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.

[3] "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." – [120, Section 12].

[4] We are not aware a published analysis of this algorithm in finite arithmetic, but believe that it can be carried out with $O(\log(n/\delta))$ bits of precision. The only issue that needs to be handled is forward instability of the QR step when the Wilkinson shift is very close to an eigenvalue of the matrix, which can be resolved e.g. by a small random perturbation of the Wilkinson shift.

The existence of an approximate diagonalization (in the sense of Definition 1.3) 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 [44], which was recently solved by some of the authors and Mukherjee in [15]. The existence of such a V is a pre-requisite 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 1.3.1 can be viewed as an efficient algorithmic answer to Davies' question.

Remark 3.1.2 (Subsequent work in Random Matrix Theory)**.** Since the first version of [11] was made public there have been some advances in random matrix theory [10, 93] that prove analogues of Corollary 3.3.2 in the case where G_n is replaced by a perturbation with random real independent entries (which have been discussed in detail in Chapter 2).

3.2 Finite Arithmetic Considerations

We start by elaborating on the axioms for floating-point arithmetic given in Section 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,

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

where $\circ$ denotes the Hadamard product. 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, operator norm (i.e. 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

$$\|A \circ \Delta\| \leq \sqrt{n} \max_i \|(A \circ \Delta)e_i\| \leq \sqrt{n} \max_{i,j} |\Delta_{i,j}| \max_i \|Ae_i\| \leq \mathbf{u}\sqrt{n}\|A\|. \tag{3.1}$$

For more complicated operations such as matrix-matrix multiplication and matrix inversion, we use existing error guarantees from the literature.

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 [85, Chapters 3-4]) and the results in this chapter are highly insensitive to the details. For simplicity, calling $\hat{x} := x/\|x\|$, we will assume that

$$|\mathsf{fl}\,(\mathrm{Tr}A) - \mathrm{Tr}A| \leq n\|A\|\mathbf{u} \tag{3.2}$$

$$\|\mathsf{fl}(\hat{x}) - \hat{x}\| \leq n\mathbf{u}. \tag{3.3}$$

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.

3.2.1 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 $T_{\mathsf{N}} > 0$.

Note that, since the Gaussian distribution has unbounded support, one should only expect the sampler $\mathsf{N}(\sigma)$ to have a relative error guarantee of the sort $|\widetilde{G} - G| \leq c_{\mathsf{N}} \sigma |G| \cdot \mathbf{u}$. However, as it will become clear below, we only care about realizations of Gaussians satisfying $|G| < R$, for a certain prespecified $R > 0$, and the rare event $|G| > R$ will be accounted for in the failure probability of the algorithm. So, for the sake of exposition we decided to omit the $|G|$ in the bound on $|\widetilde{G} - G|$.

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 paper, we will omit issues of underflow or overflow. Throughout this chapter, to simplify some of our bounds, we will also assume that $c_{\mathsf{N}} \geq 1$.

3.2.2 Black-box Error Assumptions for Multiplication, Inversion, and QR Decomposition

Our algorithm will use 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 [53]. 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* $\mathsf{MM}(\cdot,\cdot)$ 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 $A \in \mathbb{C}^{n\times n}$ 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

$$Q'A' = R, \tag{3.4}$$

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 $c_{\mathsf{INV}} = 1$ [85]. 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 [53]. 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 [54]. (2) is Theorem 3.3 (see also equation (9) above its statement) of [53]. 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 [53]. □

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 [54] and its references for details), so these may also be used as a black box, though we will not do this here.

3.3 Pseudospectral Shattering

This section will be devoted to obtaining quantitative pseudospectral shattering results that are tailored for the analysis of the spectral bisection algorithms.

In Chapter 2 we have obtained separate tail bounds for $\mathrm{gap}(A+\gamma M_n)$ and $\kappa_V(A+\gamma M_n)$, for M_n a random matrix satisfying Assumption 1.2.3. In order for this results to be used as smoothed analysis input guarantees for an algorithm, one needs to combine them to obtain a probability bound for an event of the form

$$\{\mathrm{gap}(A+\gamma M_n) \geq r, \quad \kappa_V(A+\gamma M_n) \leq t, \quad \|M_n\| \leq C\}.$$

This is straight forward, but because the bounds are stronger and simpler in the case when M_n is a complex Ginibre, we will limit our attention to this case, where we can prove the following.

Theorem 3.3.1 (Multiparameter Tail Bound)**.** *Let* $A \in \mathbb{C}^{n\times n}$*. Assume* $\|A\| \leq 1$ *and* $\gamma < 1/2$*, and let* $M_n := A + \gamma G_n$ *where* G_n *is a complex Ginibre matrix. For every* $t, r > 0$*:*

$$\mathbb{P}\big[\kappa_V(M_n) < t,\ \mathrm{gap}(M_n) > r,\ \|G_n\| < 4\big] \geq 1 - \left(\frac{144}{r^2}\cdot 4(trn/\gamma)^8 + (9n^3/\gamma^2t^2) + 2e^{-2n}\right). \quad (3.5)$$

Proof. Write $\mathrm{Spec}(M_n) := \{\lambda_1, \ldots, \lambda_n\}$ for the (random) eigenvalues of $M_n := A + \gamma G_n$, 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 $(3 \cdot 4/r)^2 = 144/r^2$. 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 $\mathrm{gap}(M_n) < r$, then there are two eigenvalues in the disk of radius r centered at some point $y \in \mathcal{N}$.

Therefore, consider the events

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

Lemma 1.1.13 applied to each $y \in \mathcal{N}$ with $k = n-1$ reveals that

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

whence

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

By a union bound, we have

$$\mathbb{P}[E_{\mathrm{gap}} \cup E_\kappa] \leq \mathbb{P}[E_D \cup E_\kappa] + |\mathcal{N}| \max_{y \in \mathcal{N}} \mathbb{P}[E_y]. \tag{3.6}$$

From the tail bound on the operator norm of a Ginibre matrix in [15, Lemma 2.2],

$$\mathbb{P}[E_D] \leq \mathbb{P}[E_G] \leq 2e^{-(4-2\sqrt{2})^2 n} \leq 2e^{-2n}. \tag{3.7}$$

Observe that by (1.8),

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

since the inequality in the left hand event must reverse when we sum over all $\lambda_i \in \mathrm{Spec}(M_n)$; thus

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

Theorem 1.2.9 and Markov's inequality yields

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

Thus, we have

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

The singular values in (1.17) applied to $M = -y + A$ give 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 (3.6) we have

$$\mathbb{P}[E_{\text{gap}} \cup E_\kappa \cup E_D] \leq \mathbb{P}[E_{\text{gap}} \cup E_\kappa \cup E_G] \leq \frac{144}{r^2} \cdot 4\left(\frac{trn}{\gamma}\right)^8 + \frac{9n^3}{\gamma^2 t^2} + 2e^{-2n},$$

as desired. □

As a corollary we obtain the following.

Corollary 3.3.2 (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 $\mathcal{N}(0, 1_{\mathbb{C}}/n)$ entries, and let $M_n := A + \gamma G_n$. Then*

$$\kappa_V(M_n) \leq \frac{n^2}{\gamma}, \quad \text{gap}(M_n) \geq \frac{\gamma^4}{n^5}, \quad \text{and} \quad \|G_n\| \leq 4, \tag{3.8}$$

with probability at least $1 - 12/n$.

Now we use the above result to show that by adding a random perturbation, with high probability, we can achieve pseudospectral shattering (in the sense of Definition 1.3.4) with respect to a random grid. First we introduce some notation.

Definition 3.3.3 (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 $\text{diam}(\mathsf{g}) := \omega\sqrt{s_1^2 + s_2^2}$ for the diameter of the grid.

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

Lemma 3.3.4. *If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, and $\Lambda_\epsilon(A)$ is shattered with respect to a grid g with side length ω, then every eigenvalue condition number satisfies $\kappa(\lambda_i) \leq \frac{2\omega}{\pi\epsilon}$.*

Proof. Let v, w^* be a right/left eigenvector pair for some eigenvalue λ_i of A, 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

$$\kappa(\lambda_i) = \|vw^*\| = \left\| \frac{1}{2\pi i} \oint_\Gamma (z - A)^{-1} \mathrm{d}z \right\| \leq \frac{1}{2\pi} \oint_\Gamma \left\| (z - A)^{-1} \right\| \mathrm{d}z \leq \frac{2\omega}{\pi\epsilon}. \tag{3.9}$$

In the final step we have used the fact that $\Lambda_\epsilon(A) \cap \mathsf{g} = \emptyset$ means $\|(z - A)^{-1}\| \leq 1/\epsilon$ on g. □

We can now show the following.

Theorem 3.3.5 (Exact Arithmetic Shattering)**.** *Let $A \in \mathbb{C}^{n\times n}$ and $M_n := A + \gamma G_n$ for G_n a complex Ginibre matrix. Assume $\|A\| \leq 1$ 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, $\kappa_V(M_n) \leq n^2/\gamma$, $\|A - M_n\| \leq 4\gamma$, and $\Lambda_\epsilon(M_n)$ is shattered with respect to g for*

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

with probability at least $1 - 13/n$.

Proof. Condition on the event in Corollary 3.3.2, so that

$$\kappa_V(M_n) \leq \frac{n^2}{\gamma}, \quad \|M_n - A\| \leq 4\gamma, \quad \text{and } \mathrm{gap}(M_n) \geq \frac{\gamma^4}{n^5} = 4\omega.$$

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 M_n is contained in one square of g with probability 1. Moreover, since $\mathrm{gap}(M_n) > 4\omega$, no square can contain two eigenvalues. Let

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

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

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

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

$$\mathbb{P}[\min_{i\leq n} \mathsf{dist}_\mathsf{g}(\lambda_i) > \omega/4n^2] > 1 - 1/n, \tag{3.10}$$

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

We now recall from (1.10) that

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

Thus, on the events (3.8) and (3.10), we see that $\Lambda_\epsilon(M_n)$ is shattered with respect to g as long as

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

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

$$1 - \frac{1}{n} - \frac{13}{n} = 1 - \frac{13}{n},$$

as desired. □

Since, adding a random perturbation to the input matrix will be a subroutine of our algorithm, we record this as such below, and prove a result similar to the one above but taking into account numerical errors.

SHATTER

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

1. $G_{ij} \leftarrow \mathsf{N}(1/n)$ for $i, j = 1, \ldots, n$.
2. $M \leftarrow A + \gamma G + E$.
3. Let g be a random grid with $\omega = \frac{\gamma^4}{4n^5}$ and bottom left corner z chosen as in Theorem 3.3.5.
4. $\epsilon \leftarrow \frac{1}{2} \cdot \frac{\gamma^5}{16n^9}$

Output: Matrix $M \in \mathbb{C}^{n \times n}$, grid g, shattering parameter $\epsilon > 0$.
Ensures: $\|M - A\| \le 4\gamma$, $\kappa_V(M) \le n^2/\gamma$, and $\Lambda_\epsilon(M)$ is shattered with respect to g, with probability at least $1 - 13/n$.

Theorem 3.3.6 (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*

$$\mathbf{u} \le \frac{1}{2} \frac{\gamma^5}{16n^9} \cdot \frac{1}{(3 + c_{\mathsf{N}})\sqrt{n}}, \tag{3.11}$$

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

Thus, as long as the precision satisfies (3.11), we have

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

where shatter$(\cdot)$ refers to the (exact arithmetic) outcome of Theorem 3.3.5. The correctness of SHATTER now follows from Lemma 1.1.7. 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 chapter is the analysis, in finite arithmetic, of a well-known iterative method for approximating to the matrix sign function. Recall from Section 1.3 that if A is a matrix whose spectrum avoids the imaginary axis, then

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

where the P_+ and P_- 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

$$g(z) := \frac{1}{2}(z + z^{-1}). \tag{3.12}$$

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

We denote the specific finite-arithmetic implementation used in our algorithm by SGN; the pseudocode is provided below.

In Section 3.4.1 we briefly discuss the specific preliminaries that will be used throughout this section. In Section 3.4.2 we give a *pseudospectral* proof of the rapid global convergence of this iteration when implemented in exact arithmetic. In Section 3.4.2 we show that the proof provided in Section 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.

SGN

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

1. $N \leftarrow \lceil \lg(1/(1-\alpha)) + 3\lg\lg(1/(1-\alpha)) + \lg\lg(1/(\beta\epsilon)) + 7.59 \rceil$
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: $\|S - \mathrm{sgn}(A)\| \leq \delta$

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 with a fixed ratio of distances to two focal points. By fixing the focal 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.

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 set

$$m(z) := \frac{1-z}{1+z}$$

to be 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 $\partial\mathsf{C}_\alpha^+$ and $\partial\mathsf{C}_\alpha^-$ for their boundaries, and $\mathsf{C}_\alpha = \mathsf{C}_\alpha^+ \cup \mathsf{C}_\alpha^-$ for their union. See Figure 3.1 for an illustration.

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 the circles are not concentric so this is not strictly speaking an annulus, and note also that in our notation

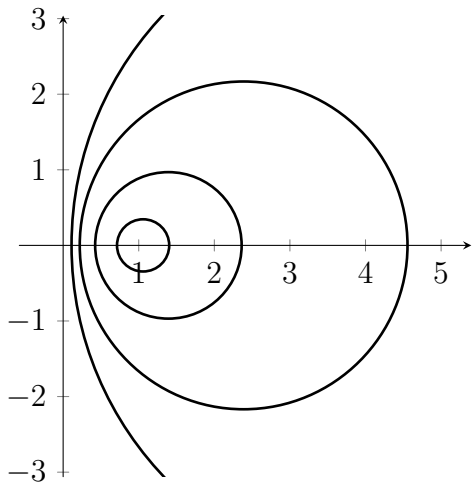

Figure 3.1: Apollonian circles appearing in the analysis of the Newton iteration. Depicted are $\partial \mathsf{C}^+_{\alpha^{2^k}}$ for $\alpha = 0.8$ and $k = 0, 1, 2, 3$, with smaller circles corresponding to larger k.

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 ([131])**.** 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,

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

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

In this section, we set $A_0 := A \in \mathbb{C}^{n\times n}$ and $A_{k+1} := g(A_k)$ for all $k \geq 0$. As explained in Section 1.3, 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 detail this argument (due to [131]) below.

Proposition 3.4.2. *Assume that A is a diagonalizable matrix and that* $\mathrm{Spec}(A) \subset \mathsf{C}_\alpha$ *for some $\alpha \in (0, 1)$. Then for every $N \in \mathbb{N}$ we have the guarantee*

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

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

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

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 $\mathrm{Spec}(A_N) \subset \mathsf{C}_{\alpha^{2^N}}$ and $\operatorname{sgn}(\lambda_i) = \operatorname{sgn}(\lambda_i^{(N)})$. Moreover, A_N and $\operatorname{sgn}(A)$ have the same eigenvectors. Hence

$$\|A_N - \operatorname{sgn}(A)\| \leq \left\| \sum_{\operatorname{Re}(\lambda_i)>0} (\lambda_i^{(N)} - 1) v_i w_i^* \right\| + \left\| \sum_{\operatorname{Re}(\lambda_i)<0} (\lambda_i^{(N)} + 1) v_i w_i^* \right\|. \tag{3.13}$$

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 (3.13) 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, also due to [131], yields a proof of convergence in exact arithmetic even when A is not diagonalizable. However, that 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 for certain $\epsilon_k > 0$, 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. For example, the following lemma shows how to obtain a bound on $\|A_N - \operatorname{sgn}(A)\|$ solely using information from the pseudospectrum of A_N.

Lemma 3.4.3 (Pseudospectral Error Bound)**.** *Let $A \in \mathbb{C}^{n\times n}$ be arbitrary and A_N be the Nth iterate of the Newton iteration under exact arithmetic. Assume that $\epsilon_N > 0$ and $\alpha_N \in (0,1)$ satisfy $\Lambda_{\epsilon_N}(A_N) \subset \mathsf{C}_{\alpha_N}$. Then we have the guarantee*

$$\|A_N - \operatorname{sgn}(A)\| \leq \frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\epsilon_N}. \tag{3.14}$$

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

$$\begin{aligned}
&\|A_N - \operatorname{sgn}(A_N)\| \\
=&\left\| \frac{1}{2\pi i} \oint_{\partial \mathsf{C}_{\alpha_N}} z(z-A_N)^{-1}\,dz - \frac{1}{2\pi i}\left(\oint_{\partial \mathsf{C}^+_{\alpha_N}} (z-A_N)^{-1}\,dz - \oint_{\partial \mathsf{C}^-_{\alpha_N}} (z-A_N)^{-1}\,dz \right) \right\| \\
=&\left\| \frac{1}{2\pi i} \oint_{\partial \mathsf{C}^+_{\alpha_N}} z(z-A_N)^{-1} - (z-A_N)^{-1}\,dz + \frac{1}{2\pi i} \oint_{\partial \mathsf{C}^-_{\alpha_N}} z(z-A_N)^{-1} + (z-A_N)^{-1}\,dz \right\| \\
\leq&\frac{1}{2\pi} \left\| \oint_{\partial \mathsf{C}^+_{\alpha_N}} (z-1)(z-A_N)^{-1}\,dz \right\| + \frac{1}{2\pi} \left\| \oint_{\partial \mathsf{C}^-_{\alpha_N}} (z+1)(z-A_N)^{-1}\,dz \right\| \\
\leq&2 \cdot \frac{1}{2\pi} \ell(\partial \mathsf{C}^+_{\alpha_N}) \sup\{|z-1| : z \in \mathsf{C}^+_{\alpha_N}\} \frac{1}{\epsilon_N} \\
=&\frac{4\alpha_N}{1-\alpha_N^2}\left(\frac{1+\alpha_N}{1-\alpha_N} - 1 \right) \frac{1}{\epsilon_N} \\
=&\frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\epsilon_N}.
\end{aligned}$$

□

In view of Lemma 3.4.3, we would now like 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 rapidly to zero. The dependence of this quantity on the *square* of α_k turns out to be crucial. As we will see below, we can find such a sequence with ϵ_k shrinking roughly at the same rate as α_k. This yields quadratic convergence, which will be necessary for our bound on the required machine precision in the finite arithmetic analysis of Section 3.4.3.

The lemma below is instrumental in determining the sequences α_k, ϵ_k.

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 C_α. 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 annulus $\mathsf{A}_{\alpha,\alpha'}$.

Let $w \in \mathsf{A}_{\alpha,\alpha'}$; see Figure 3.2 for an illustration. We must show that $w \notin \Lambda_{\epsilon'}(g(A))$. 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 $\mathrm{Spec}(A) \subset \Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, Observation 3.4.1 also guarantees that $\mathrm{Spec}(g(A)) \subset \mathsf{C}_{\alpha^2}$. 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:

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

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:

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

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 $\mathrm{Spec}(g(A)) \subset \mathsf{C}_{\alpha^2}$, Lemma 1.1.8 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 $g(A)$. □

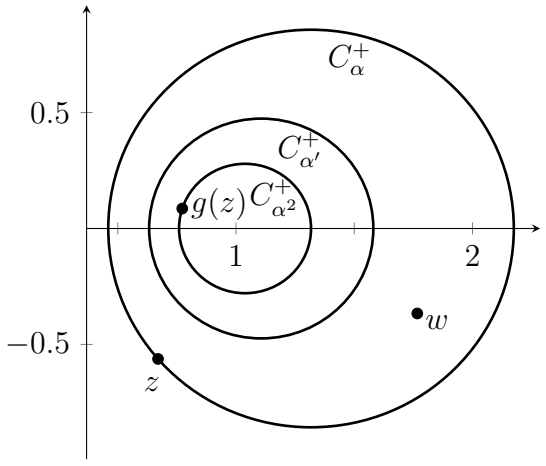

Figure 3.2: Illustration of the proof of Lemma 3.4.4

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

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

□

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*

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

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

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

□

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 statement. □

We are now ready to prove the main result of this section, 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*

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

Proof. Using the choice of α_k and ϵ_k given in the proof of Lemma 3.4.7 and the bound (3.14), we get that

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

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 become prohibitively small, so we will only need $\operatorname{polylog}(1-\alpha_0, \epsilon_0, \beta)$ 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 $\alpha_0, \beta, \mu_{\mathsf{INV}}(n)$ and c_{INV} are in place only to simplify the analysis of constants; 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* $\epsilon_0 \in (0, 1), \beta \in (0, 1/12)$*, assume* $\mu_{\mathsf{INV}}(n) \geq 1$ *and* $c_{\mathsf{INV}} \log n \geq 1$*, and assume* $A = \widetilde{A}_0$ *is a floating-point matrix with* ϵ_0*-pseudospectrum contained in* C_{α_0} *where* $0 < 1 - \alpha_0 < 1/100$*. Run* SGN *with*

$$N = \lceil \lg(1/(1-\alpha_0)) + 3 \lg \lg(1/(1-\alpha_0)) + \lg \lg(1/(\beta\epsilon_0)) + 7.59 \rceil$$

iterations (as specified 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

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

corresponding to at most

$$\lg(1/\mathbf{u}_{\mathsf{SGN}})) = O(\log n \log^3(1/(1-\alpha_0))(\log(1/\beta) + \log(1/\epsilon_0)))$$

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* $\Lambda_{\epsilon_0}(A) \subseteq \mathsf{C}_{\alpha_0}$ *where* $\epsilon_0 = \epsilon/2$ *and*

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

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

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

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

From what we just observed about the pseudospectrum, we can take $|\operatorname{Re} z| \geq \epsilon/2$. To bound the denominator, we can use the crude bound that 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*

$$\|E\| \leq \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}} \log n}\|A^{-1}\|\right) 2\sqrt{n}\mathbf{u}. \tag{3.15}$$

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, $\|E_{add}\|_{max} \leq \mathbf{u}$, and where $\circ$ denotes the entrywise (Hadamard) product of matrices.

3. Divide the resulting matrix by 2, which is an exact operation in our floating-point model as we can simply decrement the exponent. The final result is

   $$\mathsf{G}(A) = \frac{1}{2}(A + A^{-1} + E_{\mathsf{INV}}) \circ (J + E_{add}).$$

Finally, recall that for any $n \times n$ matrices M and E, we have the relation (3.1)

$$\|M \circ E\| \leq \|M\|\|E\|_{max}\sqrt{n}.$$

Putting it all together, we have

$$\|\mathsf{G}(A) - g(A)\| \leq \frac{1}{2}\left(\|A\| + \|A^{-1}\|\right)\mathbf{u}\sqrt{n} + \|E_{\mathsf{INV}}\|(1+\mathbf{u})\sqrt{n}$$

$$\leq \frac{1}{2}\left(\|A\| + \|A^{-1}\|\right)\mathbf{u}\sqrt{n} + \mu_{\mathsf{INV}}(n)\cdot\mathbf{u}\cdot\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|(1+\mathbf{u})\sqrt{n}$$
$$\leq \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|\right)2\sqrt{n}\mathbf{u}$$

where we use $\mathbf{u} < 1$ in the last line. □

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 $s := 1 - \alpha_0$, which should be thought of as a small parameter. As in the exact arithmetic case, for $k \geq 1$, we will recursively define decreasing sequences α_k and ϵ_k maintaining the property

$$\Lambda_{\epsilon_k}(\widetilde{A}_k) \subset \mathsf{C}_{\alpha_k} \qquad \text{for all } k \geq 0 \tag{3.16}$$

by induction as follows:

1. The base case $k = 0$ holds because by assumption, $\Lambda_{\epsilon_0} \subset \mathsf{C}_{\alpha_0}$.

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

$$\alpha_k \leq (1 - s/2)^{2^k}. \tag{3.17}$$

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 $\Lambda_{\epsilon'}\left(g(\widetilde{A}_k)\right) \subset \mathsf{C}_{\alpha_{k+1}}$, where

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

Thus in particular

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

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, Lemma 1.1.7 ii) ensures that if we set

$$\epsilon_{k+1} := \epsilon_k \frac{s^2\alpha_k}{32} - \|E_k\| \tag{3.18}$$

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

We now need to show that the ϵ_k do not decrease too fast as k increases. In view of (3.18), it will be helpful to set the machine precision small enough to guarantee that $\|E_k\|$ is a small fraction of $\epsilon_k \frac{\alpha_k s^2}{32}$.

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 (3.15) on $\|E_k\|$ from Lemma 3.4.11, as functions of ϵ_k. By Corollary 3.4.6, we have

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

Thus, we may write the coefficient of $\mathbf{u}$ in the bound (3.15) as

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

so that Lemma 3.4.11 reads

$$\|E_k\| \leq K_{\epsilon_k}\mathbf{u}. \tag{3.19}$$

Plugging this into the definition (3.18) of ϵ_{k+1}, we have

$$\epsilon_{k+1} \geq \epsilon_k\frac{s^2\alpha_k}{32} - K_{\epsilon_k}\mathbf{u}. \tag{3.20}$$

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

$$K_{\epsilon_k}\mathbf{u} \leq \frac{1}{3}\epsilon_k\frac{s^2\alpha_k}{32}. \tag{3.21}$$

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

$$\epsilon_{k+1} \geq \frac{2}{3}\epsilon_k\frac{s^2\alpha_k}{32} = \frac{1}{48}\epsilon_k s^2\alpha_k, \tag{3.22}$$

which implies

$$\|E_k\| \leq \frac{1}{2}\epsilon_{k+1}; \tag{3.23}$$

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 (Preliminary lower bound on ϵ_k)**.** *Let $k \geq 0$, and for all $0 \leq i \leq k-1$, assume $\mathbf{u}$ satisfies the requirement (3.21):*

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

Then we have

$$\epsilon_k \geq e_k := \epsilon_0\left(\frac{s^2}{50}\right)^k\alpha_k.$$

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

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

Since (3.21) implies (3.22), we may apply (3.22) repeatedly to obtain

$$\begin{aligned}
\epsilon_k &\geq \epsilon_0(s^2/48)^k \prod_{i=0}^{k-1} \alpha_i \\
&= \epsilon_0(s^2/48)^k(1+s/4)^{2^k-1-k}\alpha_0^{2^k-1} && \text{by the definition of } \alpha_i \\
&= \epsilon_0 \left(\frac{s^2}{48(1+s/4)}\right)^k \frac{\alpha_k}{\alpha_0} \\
&\geq \epsilon_0 \left(\frac{s^2}{50}\right)^k \alpha_k. && \alpha_0 \leq 1, s < 1/8
\end{aligned}$$

□

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 $\epsilon_i \geq e_i$ a priori, so to avoid circularity we must use a short inductive argument.

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

$$K_{e_i}\mathbf{u} \leq \frac{1}{3} e_i \frac{s^2\alpha_i}{32} \tag{3.24}$$

where e_i is defined in Lemma 3.4.12. Then we have

$$\epsilon_k \geq e_k.$$

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 and K_{e_i} is increasing in i. Assuming the full hypothesis of this lemma, we prove $\epsilon_i \geq e_i$ for $0 \leq i \leq k$ by induction on i. For the base case, we have $\epsilon_0 \geq e_0 = \epsilon_0\alpha_0$.

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

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

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

Lemma 3.4.14 (Main accuracy bound)**.** *Suppose* $\mathbf{u}$ *satisfies the requirement (3.21) for all $0 \leq k \leq N$. Then*

$$\|\widetilde{A}_N - \operatorname{sgn}(A)\| \leq \frac{8}{s}\sum_{k=0}^{N-1} \frac{\|E_k\|}{\epsilon_{k+1}^2} + \frac{8\cdot 50^N}{s^{2N+2}\epsilon_0}(1-s/2)^{2^N}. \tag{3.25}$$

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:

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

where we use the definition (1.9) of pseudospectrum and Lemma 1.1.7, together with the property (3.16). Ultimately, this chain of inequalities implies

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

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

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

where in the last step we use $\alpha_k \leq 1$ and $1 - \alpha_{k+1}^2 \geq s$, as well as (3.23).

By Lemma 3.4.3 (to be precise, by repeating the proof of that lemma with $\widetilde{A_N}$ substituted for A_N), we have

$$\|\widetilde{A_N} - \operatorname{sgn}(\widetilde{A_N})\| \leq \frac{8\alpha_N^2}{(1 - \alpha_N)^2(1 + \alpha_N)\epsilon_N}$$

$$\begin{aligned}
&\leq \frac{8}{s^2}\alpha_N \frac{\alpha_N}{\epsilon_N} \\
&\leq \frac{8}{s^2}\alpha_N \frac{1}{\epsilon_0}\left(\frac{50}{s^2}\right)^N \\
&\leq \frac{8}{s^2\epsilon_0}(1-s/2)^{2^N}\left(\frac{50}{s^2}\right)^N \\
&\leq \frac{8\cdot 50^N}{s^{2N+2}\epsilon_0}(1-s/2)^{2^N}.
\end{aligned}$$

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 $\|\widetilde{A_N} - \operatorname{sgn}(A)\|$ is at most β, the desired accuracy parameter. The upper bound (3.25) 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 then give a sufficient condition on the machine precision $\mathbf{u}$. This will be the content of Lemmas 3.4.17 and 3.4.18.

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

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

Before proving the above lemma, and other results of the sort, we state a trivial but powerful observation.

Lemma 3.4.16. *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. □

We can now show Lemma 3.4.15.

Proof of Lemma 3.4.15. An exact solution for j can be written in terms of the *Lambert W-function*; see [43] 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 of both sides and applying the second inequality in Lemma 3.4.16 with $x = 2j\log(1/t)$ and $y = 1$, using $\lg x = 1 + \lg j + \lg\log(1/t)$, 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

$$j - \lg j \leq \lg(1/t) + \lg\log(1/t) + 1.01 \leq \lg(1/t) + \lg\lg(1/t) + 0.49 =: K.$$

But since $j \geq 9$, we have $j - \lg j \geq 0.64j$, so

$$j \leq \frac{1}{0.64}(j - \lg j) \leq \frac{1}{0.64}K$$

which implies

$$j \leq K + \lg j \leq K + \lg(1.57K) = K + \lg K + 0.65.$$

Note $K \leq 1.39\lg(1/t)$, because $K - \lg(1/t) = \lg\lg(1/t) + 0.49 \leq 0.39\lg(1/t)$ for $t \leq 1/800$. Thus

$$\lg K \leq \lg(1.39\lg(1/t)) \leq \lg\lg(1/t) + 0.48,$$

so for the case $c = 1/2$ we conclude the proof of the contrapositive of the lemma:

$$\begin{aligned} j &\leq K + \lg K + 0.65 \\ &\leq \lg(1/t) + \lg\lg(1/t) + 0.49 + (\lg\lg(1/t) + 0.48) + 0.65 \\ &= \lg(1/t) + 2\lg\lg(1/t) + 1.62. \end{aligned}$$

For the general case, once $(1-t)^{2^j}/t^{2j} \leq 1/2$, 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. □

We use this to now show.

Lemma 3.4.17 (Bound on second term of (3.25))**.** *Suppose we have*

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\epsilon_0)) + 1.62.$$

Then

$$\frac{8\cdot 50^N}{s^{2N+2}\epsilon_0}(1 - s/2)^{2^N} \leq \beta/2.$$

Proof. It is sufficient that

$$\frac{8 \cdot 64^N}{s^{2N+2}\epsilon_0}(1 - s/8)^{2^N} \leq \beta/2.$$

The result now follows from applying Lemma 3.4.15 with $c = \beta s^2 \epsilon_0/16$ and $t = s/8$. □

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

Lemma 3.4.18 (Bound on first term of (3.25))**.** *Suppose*

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\epsilon_0)) + 1.62,$$

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

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

Then we have

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

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

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

In view of (3.19), which says $\|E_k\| \leq K_{\epsilon_k}\mathbf{u}$, it is sufficient to have for all $0 \leq k \leq N-1$

$$\mathbf{u} \leq \frac{1}{K_{\epsilon_k}}\frac{\beta\epsilon_{k+1}^2 s}{16N}. \tag{3.26}$$

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

$$\mathbf{u} \leq \frac{1}{K_{e_k}}\frac{\beta e_{k+1}^2 s}{16N}. \tag{3.27}$$

Indeed, on the one hand, since $\beta < 1/6$ and by the loose bound $e_{k+1} < s\alpha_{k+1} < s\alpha_k$ we have that (3.27) implies $\mathbf{u} \leq \frac{1}{3K_{e_k}}\frac{s^2 e_k}{32}$, which means that the assumption in Corollary 3.4.13 is satisfied. On the other hand Corollary 3.4.13 yields $e_k \leq \epsilon_k$ for all $0 \leq k \leq N$, which in turn, combined with (3.27) would give (3.26) and conclude the proof.

We now show that (3.27) holds for all $0 \leq k \leq N-1$. 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:

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

where we use the bound $\frac{1}{e_N} \leq \frac{4}{s^2 e_N^2}$ without much loss, and we also use our assumption $\mu_{\mathsf{INV}}(n) \geq 1$ 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

$$\mathbf{u} \leq \frac{\beta}{192 \mu_{\mathsf{INV}}(n) \sqrt{n} N} \left(\frac{\epsilon_0 (s^2/50)^N \alpha_N s^2}{4}\right)^{c_{\mathsf{INV}} \log n + 3}.$$

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

$$\mathbf{u} \leq \frac{\beta}{192 \mu_{\mathsf{INV}}(n) \sqrt{n} N} \left(\frac{\epsilon_0 (s^2/50)^N (1-s)^{2^N} s^2}{4}\right)^{c_{\mathsf{INV}} \log n + 3}.$$

Now we finally will use our hypothesis on the size of N to simplify this expression. Applying Lemma 3.4.17, we have

$$\epsilon_0 (s^2/50)^N / 4 \geq \frac{4(1-s)^{2^N}}{s^2 \beta}.$$

Thus, our sufficient condition becomes

$$\mathbf{u} \leq \frac{\beta}{192 \mu_{\mathsf{INV}}(n) \sqrt{n} N} \left(\frac{4(1-s)^{2^{N+1}}}{\beta}\right)^{c_{\mathsf{INV}} \log n + 3}.$$

To make the expression simpler, since $c_{\mathsf{INV}} \log n + 3 \geq 4$ we may pull out a factor of $4^4 > 192$ and remove the occurrences of β to yield the sufficient condition

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

□

Matching the statement of Theorem 3.4.9, we give a slightly cleaner sufficient condition on N that implies the hypothesis on N appearing in the above lemmas.

Lemma 3.4.19 (Final sufficient condition on N)**.** *If*

$$N = \lceil \lg(1/s) + 3\lg\lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) + 7.59 \rceil,$$

then

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\epsilon_0)) + 1.62.$$

Proof. We aim to provide a slightly cleaner sufficient condition on N than the current condition

$$N \geq \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\epsilon_0)) + 1.62.$$

Repeatedly using Lemma 3.4.16, as well as the cruder fact $\lg\lg(ab) \leq \lg\lg a + \lg\lg b$ provided $a, b \geq 4$, we have

$$\begin{aligned}
\lg\lg(16/(\beta s^2\epsilon_0)) &\leq \lg\lg(16/s^2) + \lg\lg(1/(\beta\epsilon_0)) \\
&= 1 + \lg(3 + \lg(1/s)) + \lg\lg(1/(\beta\epsilon_0)) \\
&\leq 1 + \lg\lg(1/s) + \frac{3}{\log 2 \lg(1/s)} + \lg\lg(1/(\beta\epsilon_0)) \\
&\leq \lg\lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) + 1.66
\end{aligned}$$

where in the last line we use the assumption $s < 1/100$. Similarly,

$$\begin{aligned}
\lg(8/s) + 2\lg\lg(8/s) &\leq 3 + \lg(1/s) + 2\lg(3 + \lg(1/s)) \\
&\leq 3 + \lg(1/s) + 2\left(\lg\lg(1/s) + \frac{3}{\log 2 \lg(1/s)}\right) \\
&\leq \lg(1/s) + 2\lg\lg(2/s) + 4.31
\end{aligned}$$

Thus, a sufficient condition is

$$N = \lceil \lg(1/s) + 3\lg\lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) + 7.59 \rceil.$$

□

Taking the logarithm of the machine precision yields the number of bits required:

Lemma 3.4.20 (Bit length computation)**.** *Suppose*

$$N = \lceil \lg(1/s) + 3\lg\lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) + 7.59 \rceil$$

and

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

Then

$$\lg(1/\mathbf{u}_{\mathsf{SGN}}) = O\big(\log n \log(1/s)^3(\log(1/\beta) + \log(1/\epsilon_0))\big).$$

Proof. In the course of the proof, for convenience we also record a nonasymptotic bound (for $s < 1/100$, $\beta < 1/12$, $\epsilon_0 < 1$ and $c_{\mathsf{INV}} \log n > 1$ as in the hypothesis of Theorem 3.4.9), at the cost of making the computation somewhat messier.

Immediately we have

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

Note that $\log(1/(1-s)) < s$ for $s < 1/2$. Also, $2^{N+1} \leq (1/s)\lg(1/s)^3(\lg(1/\beta) + \lg(1/\epsilon_0))2^{9.59}$. Putting this together, we have

$$\lg(1/\mathbf{u}_{\mathsf{SGN}}) \leq \lg \mu_{\mathsf{INV}}(n) + \frac{1}{2}\lg n + \lg N + 1000(c_{\mathsf{INV}} \log n + 3)\lg(1/s)^3(\lg(1/\beta) + \lg(1/\epsilon_0)).$$

We now crudely bound $\lg N$. Note that for $s < 1/100$ we have $\lg(1/s) + 3\lg\lg(1/s) + 7.59 \leq 1/s$. Thus,

$$\begin{aligned}
\lg N &\leq \lg(1/s + \lg\lg(1/(\beta\epsilon_0))) \\
&\leq \lg(1/s + \lg(1/(\beta\epsilon_0))) \\
&\leq \lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) && \lg(a+b) \leq \lg a + \lg b \text{ for } a, b > 2 \\
&\leq \lg(1/s)^3 \lg(1/(\beta\epsilon_0)).
\end{aligned}$$

Combining the above, we may fold the $\lg N$ and $\lg n$ terms into the final term to obtain

$$\lg(1/\mathbf{u}_{\mathsf{SGN}}) \leq \lg \mu_{\mathsf{INV}}(n) + 5000 c_{\mathsf{INV}} \log n \lg(1/s)^3(\lg(1/\beta) + \lg(1/\epsilon_0)) \tag{3.28}$$

where we use that $c_{\mathsf{INV}} \log n > 1$ and therefore $c_{\mathsf{INV}} \log n + 3 < 4 c_{\mathsf{INV}} \log n$.

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

This completes the proof of Theorem 3.4.9. Finally, we may prove the theorem advertised in Section 1.3.

Proof of Theorem 1.3.3. Set $\epsilon := \min\{\frac{1}{K}, 1\}$. Then $\Lambda_\epsilon(A)$ does not intersect the imaginary axis, and furthermore $\Lambda_\epsilon(A) \subseteq D(0,2)$ because $\|A\| \leq 1$. Thus, we may apply Lemma 3.4.10 with $\mathrm{diam}(\mathsf{g}) = 4\sqrt{2}$ to obtain parameters α_0, ϵ_0 with the property that $\log(1/(1-\alpha_0))$ and $\log(1/\epsilon_0)$ are both $O(\log K)$. Theorem 3.4.9 now yields the desired conclusion. □

3.5 Analysis of the Spectral Bisection Algorithm

In this section we will prove Theorem 1.3.1. As discussed in Section 1.3, 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 $\mathrm{Spec}(A_+) \sqcup \mathrm{Spec}(A_-)$, 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 B.1 and B.2. 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

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

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 B.1 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 $\max\{n/5, 1\}$. 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* $\epsilon \leq 0.5$*,* $\beta \leq 0.05/n$*, and* $\|A\| \leq 4$ *and* g *have side lengths of at most* 8*, and define*

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

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: $(\widetilde{P_\pm}, \mathsf{g}_\pm, n_\pm) = \mathsf{SPLIT}(A, \epsilon, \mathsf{g}, \beta)$

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

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

$$\mathbf{u} \le \mathbf{u}_{\mathsf{SPLIT}} := \min\left\{\frac{\left(1 - \frac{\epsilon}{256}\right)^{2^{N_{\mathsf{SPLIT}}+1}(c_{\mathsf{INV}}\log n + 3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{SPLIT}}}, \frac{\epsilon}{100n}, \frac{\epsilon^2}{512}\right\},$$

Using at most

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

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 [81], or probabilistically in matrix-multiplication time with a variant of the method of [53]; we will use the latter.

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 $1 - \frac{(20n)^3\sqrt{\beta}}{\eta^2}$.

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:*

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

The number of arithmetic operations is at most:

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

Remark 3.5.4. The proof of the above theorem, which is deferred to Appendix B.2, closely follows and builds on the analysis of the randomized rank revealing factorization algorithm (RURV) introduced in [53] and further studied in [9]. 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 [53] and [9] 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 B.2.6) 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 [9]. 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 and 3—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.

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 $m \leq n$.
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. $\widetilde{V} \leftarrow \text{normalize}(\widetilde{V}) + E_9$
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 $\lambda_i \in \text{Spec}(A)$, and each column $\widetilde{v}_i$ of $\widetilde{V}$ has norm $1 \pm n\mathbf{u}$, and satisfies $\|\widetilde{v}_i - v_i\| \leq \delta$ for some exact unit right eigenvector $Av_i = \lambda_i v_i$.

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 $\epsilon < 1/2$, 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*

$$N_{\mathsf{EIG}} := \lg \frac{256n}{\epsilon} + 3 \lg \lg \frac{256n}{\epsilon} + \lg \lg \frac{(5n)^{26}}{\theta^2 \delta^4 \epsilon^9} + 7.59.$$

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

$$\begin{aligned}&\lg 1/\mathbf{u}\\ \geq &\max\left\{\lg^3\frac{n}{\epsilon}\lg\left(\frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8}\right)2^{9.6}(c_{\mathsf{INV}}\log n+3)+\lg N_{\mathsf{EIG}},\lg\frac{(5n)^{30}}{\theta^2\delta^4\epsilon^8}+\lg(\{\mu_{\mathsf{MM}}(n)\vee\mu_{\mathsf{QR}}(n)\right\}\\ =&O\left(\log^3\frac{n}{\epsilon}\log\frac{n}{\theta\delta\epsilon}\log n\right).\end{aligned}$$

The number of arithmetic operations is at most

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

Remark 3.5.6. We have not fully optimized the large constant $2^{9.59}$ appearing in the bit length above.

Theorem 3.5.5 easily implies Theorem 1.3.1 when combined with SHATTER.

Proof of 1.3.1. Given A and δ, consider the following two step algorithm:

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

$$\delta':=\frac{\delta^3}{n^{4.5}\cdot 6\cdot 128\cdot 2}. \tag{3.29}$$

With probability at least $1-13/n$, $\mathsf{SHATTER}(A,\delta/8)$ succeeds, in which case the output $(X,\mathsf{grid},\epsilon)$ output easily satisfy the assumptions in Theorem 3.5.5: $\delta'\leq\delta<1$, $\epsilon=\frac{(\delta/8)^5}{32n^9}\leq 1/2$, g is defined by SHATTER to have side length 8, $\|M\|\leq\|A\|+\|M-A\|\leq 1+4(\delta/8)\leq 3.5$, and M has ϵ-pseudospectrum shattered with respect to g. On this event, $M=WCW^{-1}$, and (using the proof of Theorem 3.3.1) if we normalize W to have unit length columns, then $\kappa(W)=\|W\|\|W^{-1}\|\leq 8n^2/\delta$.

We will show that the choice of δ' in (3.29) guarantees

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

Since $\|M\|\leq\|A\|+\|A-M\|\leq 1+4\gamma\leq 3$ from Theorem 3.3.6, the hypotheses of Theorem 3.5.5 are satisfied. Thus EIG succeeds with probability at least $1-1/n$, and by a union bound, both EIG and SHATTER succeed with probabiility at least $1-14/n$. On this event, 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 the much cruder bound of

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

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:

$$\begin{aligned}
\|V^{-1} - W^{-1}\| &= \|W^{-1}(W - V)V^{-1}\| \\
&\leq \kappa(W)\|W - V\|\|V^{-1}\| \\
&\leq \frac{8n^2}{\delta} \cdot \delta'\sqrt{n} \cdot \frac{16n^2}{\delta} \\
&\leq \frac{128n^{4.5}\delta'}{\delta^2}.
\end{aligned}$$

The grid output by $\mathsf{SHATTER}(A, \delta/8)$ has $\omega = \frac{\delta^4}{4*8^4*n^5} \leq \frac{\delta}{\sqrt{2}}$ provided $\delta < 1$. Thus the guarantees on EIG in Theorem 3.5.5 tell us each eigenvalue of $M = WCW^{-1}$ shares a grid square with exactly one diagonal entry of D, which means that $\|C - D\| \leq \sqrt{2}\omega \leq \delta$. So, we have:

$$\begin{aligned}
\|VDV^{-1} - WCW^{-1}\| &\leq \|(V - W)DV^{-1}\| + \|W(D - C)V^{-1}\| + \|WC(V^{-1} - W^{-1})\| \\
&\leq \delta'\sqrt{n} \cdot 5 \cdot \frac{16n^2}{\delta} + \sqrt{n}\delta'\frac{16n^2}{\delta} + \sqrt{n} \cdot 5 \cdot \frac{128n^{4.5}\delta'}{\delta^2} \\
&= \frac{\delta' n^{4.5}}{\delta}\left(5 \cdot 16 + 16 + \frac{5 \cdot 128}{\delta}\right) \\
&\leq \frac{\delta' n^{4.5}}{\delta^2} \cdot 6 \cdot 128
\end{aligned}$$

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

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

with probability $1 - 14/n$ as desired.

To compute the running time and precision, we observe that $\mathsf{SHATTER}$ 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. The main ingredient in the proof will be the spectral projector stability result proven in Lemma 1.1.11.

Lemma 3.5.7 (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 $\|\widetilde{A}-A\|\leq\eta<\epsilon$. 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*

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

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 $w^*v=\widetilde{w}^*\widetilde{v}=1$. Let Γ be the boundary of the grid square containing the eigenvalues associated to v and $\widetilde{v}$ respectively. Then, using a contour integral along Γ as in Lemma 1.1.11, one gets

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

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

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

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

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

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 3.5.8 (Shattering is preserved after compression)**.** *Suppose P is a spectral projector of $A \in \mathbb{C}^{n\times n}$ of rank k. Let $Q \in \mathbb{C}^{n\times k}$ be such that $Q^*Q = I_k$ and that its columns span the same space as the columns of P. Then for every $\epsilon > 0$,*

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

Alternatively, the same pseudospectral inclusion holds if again $Q^*Q = I_k$ and, instead, the columns of Q span the same space as the rows of P.

Proof. We will first analyze the case when the columns of Q span the same space as the columns of P. To begin, note that if $z \in \Lambda_\epsilon(Q^*AQ)$ then there exists $v \in \mathbb{C}^k$ satisfying $\|(z-Q^*AQ)v\| \leq \epsilon\|v\|$. Since $I_k = Q^*I_nQ$ we have

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

And, because Q^* acts as an isometry on range(Q) (the span of the columns of Q) and by assumption this space is invariant under P (and hence under $(z-A)$), we have that $(z-A)Qv \in$ range(Q), and therefore $\|Q^*(z-A)Qv\| = \|(z-A)Qv\|$. From where we obtain

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

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

For the case in which the columns of Q span the rows of P, the above proof can be easily modified by now taking v with the property that $\|v^*Q^*(z-A)Q\| \leq \epsilon\|v\|$. □

Observation 3.5.9. 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. 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 3.5.10 (Accuracy of $\widetilde{\lambda}_i$)**.** *When* DEFLATE *succeeds, each eigenvalue of A shares a square of* g *with a unique eigenvalue of either $\widetilde{A_+}$ or $\widetilde{A_-}$, and furthermore $\Lambda_{4\epsilon/5}(\widetilde{A_\pm}) \subset \Lambda_\epsilon(A)$.*

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, on the event that DEFLATE succeeds, the approximation $\widetilde{Q_\pm}$ that it outputs satisfies $\|\widetilde{Q_\pm} - Q_\pm\| \leq \eta$, so in particular $\|\widetilde{Q_\pm}\| \leq 2$ as $\eta \leq 1$. The error $E_{6,\pm}$ from performing the matrix multiplications necessary to compute $\widetilde{A_\pm}$ admits the bound

$$\begin{aligned}\|E_{6,\pm}\| &\leq \mu_{\mathsf{MM}}(n)\|\widetilde{Q_\pm}\|\|A\widetilde{Q_\pm}\|\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\|\widetilde{Q_\pm}A\|\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\|\widetilde{Q_\pm}\|^2\|A\|\mathbf{u} \\ &\leq 16\left(\mu_{\mathsf{MM}}(n)\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\mathbf{u}^2\right) \\ &\leq 3\eta\end{aligned}$$

Where the second inequality follows from $\|A\| \leq 4$ and $\|\widetilde{Q_\pm}\| \leq 1 + \eta \leq 1.02 \leq \sqrt{2}$, while the last one follows from the assumption $\mathbf{u} \leq \frac{\eta}{10\mu_{\mathsf{MM}}(n)^2}$. Iterating the triangle inequality, we obtain

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

We can now apply Lemma 3.5.7. □

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 3.5.7, 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

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

and e', a column E_9 incurred by performing the normalization in floating point; in our initial discussion of floating point arithmetic we assumed in (3.3) that $\|e'\| \leq n\mathbf{u}$.

First, since $\widetilde{v} - e'$ and $\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e$ are parallel, the distance between them is just the difference in their norms:

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

Inductively $\|\widetilde{\widetilde{v}_\pm} - \widetilde{\widetilde{v}_\pm}\| \leq 4\delta/5$, and since $\|A_\pm - \widetilde{A_\pm}\| \leq \epsilon/5$ and $A_\pm$ has shattered ϵ-pseudospectrum from Lemma 3.5.8, Lemma 3.5.7 ensures

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

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

$$\begin{aligned}
&\|\widetilde{v} - v\| \\
&= \left\| \frac{\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e}{\|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\|} + e' - Q_\pm v_\pm \right\| \\
&\leq \left\| \frac{\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e}{\|\widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e\|} - \widetilde{Q_\pm}\widetilde{\widetilde{v}_\pm} + e \right\| + \|e'\| + \|e\| + \|(\widetilde{Q_\pm} - Q_\pm)\widetilde{\widetilde{v}_\pm}\| \\
&\quad + \|Q_\pm(\widetilde{\widetilde{v}_\pm} - \widetilde{v}_\pm)\| + \|Q_\pm(\widetilde{v}_\pm - v_\pm)\|
\end{aligned}$$

$$\begin{aligned}
&\leq 4\eta + n\mathbf{u} + \mu_{\mathsf{MM}}(n)\mathbf{u} + \eta(1 + n\mathbf{u}) + 4\delta/5 + \delta/10 \\
&\leq 8\eta + 4\delta/5 + \delta/10 && n\mathbf{u}, \mu_{\mathsf{MM}}(n)\mathbf{u} \leq \eta \\
&\leq \delta && \eta \leq \delta/200.
\end{aligned}$$

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 3.5.11 (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

$$d(n) = \begin{cases} n & n \leq 5 \\ 1 + \max_{\theta \in [1/5, 4/5]} d(\theta n) & n > 5 \end{cases} \tag{3.30}$$

As $n \leq \log_{5/4} n$ for $n \leq 5$, the result is immediate from induction. □

We pause briefly to verify that the assumptions $\delta < 1$, $\epsilon < 1/2$, grid has side lengths at most 9, and $\|A\| \leq 3.5$ in Theorem 3.5.5 ensure that every call to SPLIT throughout the algorithm satisfies the hypotheses of Theorem 3.5.2, namely that $\epsilon \leq 0.5, \beta \leq 0.05/n, \|A\| \leq 4$, and grid has side lengths of at most 8. Since δ, ϵ, and β are non-increasing as we travel down the recursion tree of EIG — with β monotonically decreasing in δ and ϵ — we need only verify that the hypotheses of Theorem 3.5.2 hold on the initial call to EIG. The condition on ϵ is immediately satisfied; for the one on β, we have

$$\beta = \frac{\eta^4\theta^2}{(20n)^6 \cdot 4n^8} = \frac{\theta^2\delta^4\epsilon^8}{200^4(20n)^6 \cdot 4n^8},$$

which is clearly at most $0.05/n$.

On each new call to EIG the grid only decreases in size, so the initial assumption is sufficient. Finally, we need that every matrix passed to SPLIT throughout the course of the algorithm has norm at most 4. Lemma 3.5.10 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 an additive $\epsilon/5$, where ϵ is the input to the outermost call to EIG. Since ϵ decreases by a factor of $4/5$ on each recursion step, this means that by the end of the algorithm the norm of the matrix passed to EIG will increase by at most an additive $(\epsilon + (4/5)\epsilon + (4/5)^2\epsilon + \cdots)/5 = \epsilon \leq 1/2$. Thus we will be safe if our initial matrix has norm at most 3.5, as assumed.

Lemma 3.5.12 (Lower Bounds on the Parameters)**.** *Assume* EIG *is run on an $n \times n$ matrix, with some parameters δ and ϵ. Throughout the algorithm, on every recursive call to* EIG*, the corresponding parameters δ' and ϵ' satisfy*

$$\delta' \geq \delta/n \qquad \epsilon' \geq \epsilon/n.$$

On each such call to EIG*, the parameters η' and β' passed to* SPLIT *and* DEFLATE *satisfy*

$$\eta' \geq \frac{\delta\epsilon^2}{200n^3} \qquad \beta' \geq \frac{\theta^2\delta^4\epsilon^8}{(5n)^{26}}.$$

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. The parameters η' and β' are computed directly from ϵ' and δ'. □

Lemma 3.5.13 (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

$$\begin{aligned} T_{\mathsf{EIG}}(n, \delta, \mathsf{g}, \epsilon, \theta, n) \leq\ & T_{\mathsf{SPLIT}}(n, \epsilon, \beta) + T_{\mathsf{DEFLATE}}(n, \beta, \eta) + 2T_{\mathsf{MM}}(n) \\ & + T_{\mathsf{EIG}}(n_+, 4\delta/5, \mathsf{g}_+, 4\epsilon/5, \theta, n) + T_{\mathsf{EIG}}(n_-, 4\delta/5, \mathsf{g}_-, 4\epsilon/5, \theta, n) \\ & + 2T_{\mathsf{MM}}(n) + O(n^2). \end{aligned}$$

All of T_{SPLIT}, T_{DEFLATE}, and T_{MM} are of the form $\mathrm{polylog}(n)\mathrm{poly}(n)$, with all coefficients nonnegative and exponents in the $\mathrm{poly}(n)$ no smaller than 2. So, for any $n_+ + n_- = n$ and $n_\pm \geq 4n/5$, holding all other parameters fixed, $T_{\mathsf{SPLIT}}(n_+, ...) + T_{\mathsf{SPLIT}}(n_-, ...) \leq ((4/5)^2 + (1/5)^2)\, T_{\mathsf{SPLIT}}(n, ...) = (17/25) T_{\mathsf{SPLIT}}(n, ...)$ and the same holds for T_{DEFLATE} and T_{MM}. Applying this recursively, with all parameters other than n set to their lower bounds from Lemma 3.5.12, we then have

$$\begin{aligned} T_{\mathsf{EIG}}(n, \delta, \mathsf{g}, \epsilon, \theta, n) &\leq \frac{1}{1 - 17/25} \left(T_{\mathsf{SPLIT}} \left(n, \epsilon/n, \mathsf{g}, \frac{\delta^4\epsilon^8\theta^2}{(5n)^{26}} \right) \right. \\ &\qquad \left. + T_{\mathsf{DEFLATE}} \left(n, \beta/n, \epsilon/n, \frac{\delta^4\epsilon^8\theta^2}{(5n)^{26}} \right) + 4T_{\mathsf{MM}}(n) + O(n^2) \right) \\ &= \frac{25}{8} \left(12 N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})} \left(T_{\mathsf{INV}}(n) + O(n^2) \right) + 2T_{\mathsf{QR}}(n) \right. \end{aligned}$$

$$+ 5T_{\mathsf{MM}}(n) + n^2 T_{\mathsf{N}} + O(n^2)\Big)$$
$$\leq 60 N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})} \left(T_{\mathsf{INV}}(n) + O(n^2)\right) + 10T_{\mathsf{QR}}(n) + 25T_{\mathsf{MM}}(n),$$

where

$$N_{\mathsf{EIG}} := \lg \frac{256n}{\epsilon} + 3 \lg \lg \frac{256n}{\epsilon} + \lg \lg \frac{(5n)^{26}}{\theta^2 \delta^4 \epsilon^9} + 7.59.$$

In the above inequalities, we've substituted in the expressions for T_{SPLIT} and T_{DEFLATE} from Theorems 3.5.2 and 3.5.3, respectively; N_{EIG} is defined by recomputing N_{SPLIT} with the parameter lower bounds, and the ϵ^9 is not an error. The final inequality uses our assumption $T_{\mathsf{N}} = O(1)$. Thus using the fast and stable instantiations of MM, INV, and QR from Theorem 3.2.6, we have

$$T_{\mathsf{EIG}}(n, \delta, \mathsf{g}, \epsilon, \theta, n) = O\left(\log \frac{1}{\omega(\mathsf{g})} \left(\log \frac{n}{\epsilon} + \log \log \frac{1}{\theta\delta}\right) T_{\mathsf{MM}}(n, \mathbf{u})\right); \tag{3.31}$$

exact constants can be extracted by analyzing N_{EIG} and opening Theorem 3.2.6.

Required Bits of Precision. We will need the following bound on the norms of all spectral projectors.

Lemma 3.5.14 (Sizes of Spectral Projectors)**.** *Throughout the algorithm, every approximate spectral projector $\widetilde{P}$ given to* 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 3.5.12, we need $\mathbf{u}$ to satisfy

$$\mathbf{u} \leq \min \left\{ \frac{\left(1 - \frac{\epsilon}{256n}\right)^{2^{N_{\mathsf{EIG}}+1}(c_{\mathsf{INV}} \log n + 3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n} N_{\mathsf{EIG}}}, \right.$$
$$\frac{\epsilon}{100n^2}, \frac{\theta^2\delta^4\epsilon^8}{(5n)^{26}}, \frac{1}{4\|\widetilde{P}\| \max\{\mu_{\mathsf{QR}}(n), \mu_{\mathsf{MM}}(n)\}},$$
$$\left. \frac{\delta\epsilon^2}{100n^3 \cdot 2\mu_{\mathsf{QR}}(n)}, \frac{\delta\epsilon^2}{100n^3 \max\{4\mu_{\mathsf{MM}}(n), n, 2\mu_{\mathsf{QR}}(n)\}} \right\}$$

From Lemma 3.5.14, $\|\widetilde{P}\| \leq 10n/\epsilon$, so the conditions in the second two lines are all satisfied if we make the crass upper bound

$$\mathbf{u} \leq \frac{\theta^2\delta^4\epsilon^8}{(5n)^{30}} \frac{1}{\max\{\mu_{\mathsf{QR}}(n), \mu_{MM}(n), n\}}, \tag{3.32}$$

i.e. if $\lg 1/\mathbf{u} \geq O\left(\lg \frac{n}{\theta\delta\epsilon}\right)$. Unpacking the first requirement, using the definition $N_{\mathsf{EIG}} := \lg \frac{256n}{\epsilon} + 3\lg\lg \frac{256n}{\epsilon} + \lg\lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^9} + 7.59$ from Theorem 3.5.5, and recalling that $\epsilon \leq 1/2$, $n \geq 1$, and $(1-x)^{1/x} \geq 1/4$ for $x \in (0, 1/512)$, we have

$$\begin{aligned}
\frac{\left(1 - \frac{\epsilon}{256n}\right)^{2^{N_{\mathsf{EIG}}+1}(c_{\mathsf{INV}} \log n + 3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}} &= \frac{\left(\left(1 - \frac{\epsilon}{256n}\right)^{\frac{256n}{\epsilon}}\right)^{\lg^3 \frac{256n}{\epsilon} \lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8} 2^{8.59}(c_{\mathsf{INV}} \log n + 3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}} \\
&\geq \frac{4^{-\lg^3 \frac{256n}{\epsilon} \lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8} 2^{8.59}(c_{\mathsf{INV}} \log n + 3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}},
\end{aligned}$$

so setting $\mathbf{u}$ smaller than the final expression is sufficient to guarantee EIG and all subroutines can execute as advertised. This gives

$$\begin{aligned}
\lg 1/\mathbf{u} &\geq \lg^3 \frac{n}{\epsilon} \lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8} 2^{9.59}(c_{\mathsf{INV}} \log n + 3) + \lg N_{\mathsf{EIG}} \\
&= O\left(\log^3 \frac{n}{\epsilon} \log \frac{n}{\theta\delta\epsilon} \log n\right).
\end{aligned}$$

This dominates the precision requirement from (3.32), and completes the proof of Theorem 3.5.5.

Remark 3.5.15. A constant may be extracted directly from the expression above — leaving ϵ, δ, θ fixed, a crude bound on it is $2^{9.59} \cdot 26 \cdot 8 \cdot c_{\mathsf{INV}} \approx 160303 c_{\mathsf{INV}}$. This can certainly be optimized, the improvement with the highest impact would be tighter analysis of SPLIT, with the aim of eliminating the additive 7.59 term in N_{SPLIT}.

Chapter 4

Hessenberg QR Algorithm

This chapter is divided in three sections. The first two are devoted, respectively, to the understanding of the dynamics and numerical stability of the shifted QR algorithm. The last one presents a self-contained diagonalization algorithm, perhaps of independent interest, that will be used to compute the Ritz values that will be used in our shifting strategy.

Throughout the chapter, in most occasions, we will use the letter H when referring to upper Hessenberg matrices, and for $k \leq \dim(H)$ we will use $H_{(k)}$ to denote the lower-right $k \times k$ corner of H. We denote the distance between two sets $\mathcal{R}, \mathcal{S} \subset \mathbb{C}$ as

$$\mathsf{dist}(\mathcal{R}, \mathcal{S}) := \inf_{r \in \mathcal{R},\, s \in \mathcal{S}} |r - s|.$$

4.1 Dynamics

4.1.1 Proof Techniques and Complications

There are two distinct phenomena which make analyzing the dynamics of shifted QR challenging.

1. *Transient behavior due to nonnormality.* In the nonnormal case, the iterates H_t can behave chaotically on short time scales,[1] lacking any kind of obvious algebraic or geometric monotonicity properties (which are present in the symmetric case). This lack of monotonicity makes it hard to reason about convergence.

2. *Fixed points and periodic orbits due to symmetry.* The most natural shifting strategies define $p_t(z)$ as a simple function of the entries of H_t, typically a function of the characteristic polynomial of the bottom right $k \times k$ corner $(H_t)_{(k)}$ of H_t (see Section 4.1.2 for more details). These strategies typically have attractive fixed points and cycles which are not upper triangular, leading to slow convergence or nonconvergence (e.g. see

[1] We measure time not as the number of QR steps, but as the number of QR steps of degree 1, so for example a QR step with a degree k shift corresponds to k time steps.

[122, 17, 47]). The conceptual cause of these fixed points is symmetry — at a very high level, the dynamical system "cannot decide which invariant subspace to converge to." This feature is seen even in normal matrices, and in fact its most severe manifestation occurs in the case of unitary matrices.

Example 4.1.1. Both pathologies are seen in the instructive family of $n \times n$ examples

$$M = \begin{pmatrix} & & & & \beta_n \\ \beta_1 & & & & \\ & \beta_2 & & & \\ & & \ddots & & \\ & & & \beta_{n-1} & \end{pmatrix}$$

where $\beta_1, \ldots, \beta_n \in (0, 1)$. Observe that, for $k \leq n-1$, the characteristic polynomial of $M_{(k)}$ is just z^k, so any naïve shifting strategy based on it will yield the trivial shift. One can verify that a QR step with the trivial strategy applied to M cyclically permutes the β_i, while leaving the zero pattern of M intact. This means that for adversarially chosen $\beta_1, ..., \beta_n$, the bottom few subdiagonal entries of M — the traditional place to look for monotonicity in order to prove convergence — exhibit arbitrary behavior over a small number of QR steps. At very long time scales of n steps, the behavior becomes periodic and predictable, but there is still no convergence.

Previous approaches to showing rapid decoupling have been essentially algebraic (relying on examining entries of the iterates, their resolvents, or characteristic polynomials of their submatrices) or geometric (viewing the iteration as a flow on a manifold), and have been unable to surmount these difficulties in the nonsymmetric case.

In contrast, we take an essentially analytic approach. The key idea is to associate a measure μ_t, similar in spirit to the notion of spectral measure of a normal matrix, with the not necessarily normal iterates H_t. When the eigenvector condition number $\kappa_V(H_0)$ is bounded, the dynamics of shifted QR can be understood in terms these measures: it turns out that while the μ_t may evolve erratically on short time scales, they must behave in a predictable way over time scales of k degree 1 QR steps when $k \gg \log \kappa_V(H_0)$ as in (1.27) — essentially, this is enough to "damp" the transient behavior due to nonnormality (this is articulated precisely in Section 4.1.3). Specifically, the behavior of the μ_t can be related to the *geometric mean* of the bottom k subdiagonal entries of H_t, which we show satisfies an *approximate monotonicity* property and use as a *potential function* to track convergence.

To see this phenomenon in action, if we impose a bound on $\kappa_V(M)$ in Example 4.1.1, it can be seen that the ratios of the β_i cannot be arbitrary and the geometric mean of the bottom $\log \kappa_V(M)$ subdiagonal entries of M must remain almost-constant on intervals of k unshifted QR steps.

The above insights are sufficient to handle the transience issue but not the symmetry issue. For the latter, we carefully design a shifting strategy which satisfies the following dichotomy: either (i) a certain QR step of degree k significantly decreases the potential

function defined above, or, (ii) the measure μ_t associated to the current iterate H_t must have a special structure (essentially, being well-supported on an annulus of a particular radius). In the second case (which corresponds to the symmetry case discussed above) we exploit the structure to design a simple exceptional shift which is guaranteed to significantly reduce the potential, yielding linear convergence in either case. Thus, our proof articulates that transients and symmetry are the only obstacles to rapid convergence of the shifted QR iteration on nonsymmetric matrices.

4.1.2 History and Related Work

The literature on shifted QR is vast, so we mention only the most relevant works — in particular, we omit the large body of experimental work and do not discuss the many works on local convergence of shifted QR (i.e., starting from an H_0 which is already very close to decoupling). The reader is directed to the excellent surveys [18, 141, 39] or [123, 171, 77] for a dynamical or numerical viewpoint, respectively, or to the books [79, 157, 55, 170] for a comprehensive treatment.

Most of the shifting strategies studied in the literature are a combination of the following three types. The motivation for considering shifts depending on $H_{(k)}$ is closely related to Krylov subspace methods, see e.g. [170]. Below H denotes the current Hessenberg iterate.

1. *k-Francis Shift.* Take $p(z) = \det(z - H_{(k)})$ for some k. The case $k = 1$ is called Rayleigh shift.

2. *Wilkinson Shift.* Take $p(z) = (z - a)$ where a is the root of $\det(z - H_{(2)})$ closer to $H_{(1)}$.

3. *Exceptional Shift.* Let $p(z) = (z - x)$ for some x chosen randomly or arbitrarily, perhaps with a specified magnitude (e.g. $|x| = 1$ for unitary matrices in [61, 166, 167, 168]).

Shifting strategies which combine more than one of these through some kind of case analysis are called "mixed" strategies.

Symmetric Matrices. Jiang [66] showed that the geometric mean of the bottom k subdiagonal entries is monotone for the k-Francis strategy in the case of symmetric tridiagonal matrices. Aishima et al. [1] showed that this monotonicity continues to hold for a "Wilkinson-like" shift which chooses $k - 1$ out of k Ritz values. Both of these results yield global convergence on symmetric tridiagonal matrices (without a rate).

Rayleigh Quotient Iteration and Normal Matrices. The behavior of shifted QR is well known to be related to shifted inverse iteration (see e.g. [157]). In particular, the Rayleigh shifting strategy corresponds to a vector iteration process known as Rayleigh Quotient Iteration (RQI). Parlett [124] (building on [117, 33, 127]) showed that RQI converges globally (but without giving a rate) on almost every normal matrix and investigated how to generalize this to the nonnormal case.

Batterson [17] studied the convergence of 2-Francis shifted QR on 3×3 normal matrices with a certain exceptional shift and showed that it always converges. The subsequent work [16] showed that 2-Francis shifted QR converges globally on almost every real $n \times n$ normal matrix (without a rate). In Theorem 6 of that paper, it was shown that the same potential that we consider is monotone-decreasing when the k-Francis shift is run on normal matrices, which was an inspiration for our proof of almost-monotonicty for nonnormal matrices.

Nonnormal Matrices. Parlett [122] showed that an unshifted QR step applied to a singular matrix leads to immediate 0-decoupling, taking care of the singularity issue that was glossed over in the introduction, and further proved that all of the fixed points of an extension of the 2-Francis shifted QR step (for general matrices) are multiples of unitary matrices.

In a sequence of works, Batterson and coauthors investigated the behavior of RQI and 2-Francis on nonnormal matrices from a dynamical systems perspective. Batterson and Smillie [21, 20] showed that there are real matrices such that RQI fails to converge for an open set of real starting vectors. The latter paper also established that RQI exhibits chaotic behavior on some instances, in the sense of having periodic points of infinitely many periods. Batterson and Day [19] showed that 2-Francis shifted QR converges globally and linearly on a certain conjugacy class of 4×4 Hessenberg matrices.

In the realm of periodicity and symmetry breaking, Day [47], building on an example of Demmel, showed that there is an open set of 4×4 matrices on which certain mixed shifting strategies used in the EISPACK library fail to converge rapidly in exact arithmetic; such an example was independently discovered by Moler [109] who described its behavior in finite arithmetic. These examples are almost normal in the sense that they satisfy $\kappa_V \leq 2$, so the reason for nonconvergence is symmetry, and our strategy $\mathrm{Sh}_{k,B}$ with modest parameters $k = B = 2$ is guaranteed to converge rapidly on them (in exact arithmetic).

Using topological considerations, Leite et al. [102] proved that no continuous shifting strategy can decouple on every symmetric matrix. Accordingly (in retrospect), the most successful shifting strategy for symmetric matrices, the Wilkinson Shift, is discontinuous in the entries of the matrix and explicitly breaks symmetry when it occurs. Our strategy $\mathrm{Sh}_{k,B}$ is also discontinuous in the entries of the matrix.

Mixed and Exceptional Shifts. Eberlein and Huang [61] showed global convergence (without a bound on the rate) of a certain mixed strategy for unitary Hessenberg matrices; more recently, the works [166, 167, 168] exhibited mixed strategies which converge globally for unitary Hessenberg matrices with a bound on the rate, but this bound depends on the matrix in a complicated way and is not clearly bounded away from 1. Our strategy $\mathrm{Sh}_{k,B}$ is also a mixed strategy which in a sense combines all three types above. Our choice of exceptional shift was in particular inspired by the work of [61, 167] — the difference is that the size of the exceptional shift is naturally of order 1 in the unitary case, but in the general case it must be chosen carefully at the correct spectral scale.

Higher Degree Shifts. The idea of using higher degree shifts was already present in [68, 52], but was popularized in by Bai and Demmel in [5], who observed that higher order shifts can sometimes be implemented more efficiently than a sequence of lower order ones; see [5, Section 3] for a discussion of various higher order shifting strategies which were considered in the 1980s.

Integrable Systems. The unshifted QR algorithm on Hermitian matrices is known to correspond to evaluations of an integrable dynamical system called the Toda flow at integer times [49]; such a correspondence is not known for any nontrivial shifting scheme or for nonnormal matrices. See [39] for a detailed survey of this connection. More recently, the line of work [130, 51, 50] studied the universality properties of the decoupling time of unshifted QR on random matrices, and used the connection to Toda flow to prove universality in the symmetric case; it was experimentally observed that such universality continues to hold for shifted QR.

We defer a detailed discussion of the extensive related work on numerical issues related to shifted QR as well as a comparison to other algorithms for computing eigenvalues (in particular, [3] and [11]) to Section 4.2 below.

4.1.3 Notation and Basic Lemmas

Throughout the remainder of Section 4.1, $H = (h_{i,j})_{i,j\in[n]}$ will denote an $n \times n$ upper Hessenberg matrix, $B \geq \kappa_V(H)$ an upper bound on its eigenvector condition number and and $k \geq 2$ a power of two, which the reader may consider for concreteness to be on the order of $\log B \log\log B$; all logarithms will be taken base two for simplicity. As above, we use $H_{(k)}$ and $\chi_k(z)$ to denote the lower-right $k \times k$ corner of H and its characteristic polynomial respectively. All matrix norms are operator norms, denoted by $\|\cdot\|$.

We will use the geometric mean of the last k subdiagonal entries of the H to track convergence of the Shifted QR iteration, since we are guaranteed δ-decoupling once this quantity is smaller than $\delta\|H\|$. More explicitly, define the *potential* $\psi_k(H)$ of H to be

$$\psi_k(H) := |h_{n-k,n-k-1} \cdots h_{n,n-1}|^{\frac{1}{k}}.$$

Fixing some $\gamma \in (0,1)$, we will show that our shifting strategy guarantees *potential reduction*: the efficient computation of a Hessenberg matrix $\widehat{H}$, unitarily equivalent to H, with the property that

$$\psi_k(\widehat{H}) \leq \gamma\psi_k(H). \tag{4.1}$$

Since $\psi_k(H) \leq \|H\|$, it follows immediately that we can achieve δ-decoupling in $\frac{\log\delta}{\log\gamma}$ iterations. Note that the relationship (1.27) between k and B is *not* required for the proof of potential reduction, but impacts the cost of performing each iteration. The table below collates several constants which will appear throughout Section 4.1.

We assume black box access to a routine for efficiently performing a QR step in $O(kn^2)$ arithmetic operations rather than $O(kn^3)$.

Symbol	Meaning	Typical Scale
H	Upper Hessenberg matrix	
B	Eigenvector condition bound	$B \geq \kappa_V(H)$
k	Shift degree	$O(\log B \log\log B)$
δ	Decoupling parameter	
γ	Decoupling rate	0.8
θ	Approximation parameter for Ritz values	2
α	Promising Ritz value parameter	$B^{4k^{-1}\log k} = 1 + o(1)$

Definition 4.1.2 (Implicit QR Algorithm)**.** For $k \leq n$, an exact implicit QR algorithm $\mathsf{iqr}(H, p(z))$ takes as inputs a Hessenberg matrix $H \in \mathbb{C}^{n\times n}$ and a polynomial $p(z) = (z - s_1)\cdots(z - s_k)$ and outputs a Hessenberg matrix $\widehat{H}$ satisfying

$$\widehat{H} = Q^* H Q,$$

where Q is a unitary matrix such that $p(H) = QR$ for some upper triangular matrix R, as well as the number $\|e_n^* p^{-1}(H)\|$ whenever $p(H)$ is invertible. It runs in at most

$$T_{\mathsf{IQR}}(k, n) \leq 7kn^2 \tag{4.2}$$

operations.

See e..g [171, Section 3] for a proof in exact arithmetic of the existence of an efficient implicit QR algorithm.

Before introducing and analyzing our shifting strategy, we pause to prove three simple and essential lemmas relating the potential $\psi_k(H)$, the Hessenberg structure of H, its eigenvector condition number $\kappa_V(H)$, and certain measures associated with H. The first is well known and gives a variational characterization of the potential (see [157, Theorem 34.1]).

Lemma 4.1.3 (Variational Formula for ψ_k)**.** *Let $H \in \mathbb{C}^{n\times n}$ be any Hessenberg matrix. Then, for any k*

$$\psi_k(H) = \min_{p\in\mathcal{P}_k} \|e_n^* p(H)\|^{\frac{1}{k}},$$

with the minimum attained for $p = \chi_k$.

Proof. Since H is upper Hessenberg, for any polynomial $p \in \mathcal{P}_k$ we have

$$p(H)_{n,n-j} = \begin{cases} p(H_{(k)})_{k,k-j+1} & j = 0, \ldots, k-1, \\ h_{n-k,n-k-1}\cdots h_{n,n-1} & j = k, \\ 0 & j \geq k+1. \end{cases}$$

Thus for every such p,

$$\min_{p\in\mathcal{P}_k} \|e_n^* p(H)\| \geq |h_{n-k,n-k-1}\cdots h_{n,n-1}| = \psi_k(H)^k,$$

and the bound will be tight for any polynomial whose application to $H_{(k)}$ zeroes out the last row; by Cayley-Hamilton, the matrix $\chi_k(H_{(k)})$ is identically zero. □

It will be useful to have a mechanism for proving upper bounds on the potential of $\widehat{H}$ produced from H by an implicit QR step. To this end, let $p \in \mathcal{P}_k$ and define

$$\tau_p(H) := \|e_n^* p(H)^{-1}\|^{-\frac{1}{k}}, \tag{4.3}$$

when $p(H)$ is invertible, and $\tau_p(H) = 0$ otherwise. The special case $k = 1$ of this quantity has been used to great effect in previous work studying linear shifts (e.g. [88]), and our next lemma shows that it bounds the potential of $\widehat{H} = \mathsf{iqr}(H, p(z))$ for shift polynomials p of arbitrary degree.

Lemma 4.1.4 (Upper Bounds on $\psi_k(\widehat{H})$)**.** *Let $H \in \mathbb{C}^{n\times n}$ be a Hessenberg matrix, $p(z)$ a monic polynomial of degree k and $\widehat{H} = \mathsf{iqr}(H, p(z))$. Then*

$$\psi_k(\widehat{H}) \leq \tau_p(H).$$

Proof. Assume first that $p(H)$ is singular. In this case for any QR decomposition $p(H) = QR$, the entry $R_{n,n} = 0$, and because $p(\widehat{H}) = Q^* p(H) Q = RQ$, the last row of $p(\widehat{H})$ is zero as well. In particular $\psi_k(\widehat{H}) = |p(\widehat{H})_{1,k+1}|^{\frac{1}{k}} = 0 = \tau_p(H)$. When $p(H)$ is invertible, applying Lemma 4.1.3 and using repeatedly that Q is unitary, R is triangular, and $p(H) = QR$,

$$\psi_k(\widehat{H})^k \leq \|e_n^* p(\widehat{H})\| = \|e_n^* Q^* p(H)\| = \|e_n^* R\| = \|e_n^* R^{-1} Q^*\|^{-1} = \|e_n^* p(H)^{-1}\|^{-1} = \tau_p(H)^k.$$

□

Lemma 4.1.4 ensures that given H, we can reduce the potential with an implicit QR step by producing a polynomial p with $\|e_n^* p(H)^{-1}\|^{\frac{1}{k}} \leq \gamma \psi_k(H)$. To do so, we will require a final lemma relating quantities of this form to the moments of a certain measure associated to H which quantifies the overlap of the vector e_n^* with the left eigenvectors of H.

4.1.4 Approximate Functional Calculus

For $H \in \mathbb{C}^{n\times n}$ upper Hessenberg and diagonalizable, recall the definition of the associated random variable Z_H defined in Section 1.6. We will now prove the claim made about the approximate functional calculus.

Lemma 4.1.5 (Approximate Functional Calculus)**.** *For any upper Hessenberg H and complex function f whose domain includes the eigenvalues of H,*

$$\frac{\|e_n^* f(H)\|}{\kappa_V(H)} \leq \mathbb{E}\left[|f(Z_H)|^2\right]^{\frac{1}{2}} \leq \kappa_V(H) \|e_n^* f(H)\|.$$

Proof. By the definition of Z_H above,

$$\mathbb{E}\left[|f(Z_H)|^2\right]^{\frac{1}{2}} = \frac{\|e_n^* f(H) V\|}{\|e_n^* V\|} \leq \|e_n^* f(H)\| \|V\| \|V^{-1}\| = \|e_n^* f(H)\| \kappa_V(H),$$

and the left hand inequality is analogous. □

Using this lemma with some carefully chosen rational functions f of degree k, we will be able to probe the distribution of Z_H for each iterate H of the algorithm by examining the observable quantities $\|e_n^* f(H)\|^{\frac{1}{k}}$ — for appropriately large k, these are related to $(\mathbb{E}|f(Z_H)|^2)^{\frac{1}{k}}$ by a multiplicative factor of $\kappa_V(H)^{\frac{1}{k}} \approx 1$, so we obtain accurate information about Z_H, which enables a precise understanding of convergence. Since the iterates are all unitarily similar, κ_V is preserved with each iteration, so the k required is an invariant of the algorithm. Thus the use of a sufficiently high-degree shifting strategy is both an essential feature and unavoidable cost of our approach.

4.1.5 Promising Ritz Values and Almost Monotonicity of the Potential

In the same spirit as Wilkinson's shift, which chooses a particular Ritz value (out of two), but using a different criterion, our shifting strategy will begin by choosing a Ritz value (out of k) that has the following property for some $\alpha \geq 1$.

Definition 4.1.6 (α-promising Ritz value)**.** Let $\alpha \geq 1$, $\mathcal{R} = \{r_1, ..., r_k\}$ be a set of θ-approximate Ritz values for H, and $p(z) = \prod_{i=1}^{k}(z - r_i)$. We say that $r \in \mathcal{R}$ is *α-promising* if

$$\mathbb{E}\frac{1}{|Z_H - r|^k} \geq \frac{1}{\alpha^k}\mathbb{E}\frac{1}{|p(Z_H)|}. \tag{4.4}$$

Note that there is at least one 1-promising Ritz value in every set of approximate Ritz values, since

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\frac{1}{|Z_H - r_i|^k} = \mathbb{E}\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|Z_H - r_i|^k} \geq \mathbb{E}\frac{1}{|p(Z_H)|} \tag{4.5}$$

by linearity of expectation and AM/GM. The notion of α-promising Ritz value is a relaxation which can be computed efficiently from the entries of H (in fact, as we will explain in Section 4.1.7, using a small number of implicit QR steps with Francis-like shifts of degree $k/2$).

As a warm-up for the analysis of the shifting strategy, we will first show that if $k \gg \log \kappa_V(H)$ and r is a promising Ritz value, the potential is *almost monotone* under the shift $(z - r)^k$. This justifies the intuition from Section 4.1.1 and suggests that promising Ritz values should give rise to good polynomial shifts, but is not actually used in the proof of our main theorem. Subsequent proofs will instead use the stronger property (4.6) established below.

Lemma 4.1.7 (Almost-monotonicity and Moment Comparison)**.** *Let $\mathcal{R} = \{r_1, \dots, r_k\}$ be a set of θ-optimal Ritz values and assume that $r \in \mathcal{R}$ is α-promising. If $\widehat{H} = \mathsf{iqr}(H, (z-r)^k)$ then*

$$\psi_k(\widehat{H}) \leq \kappa_V(H)^{\frac{2}{k}} \alpha\theta\psi_k(H),$$

and moreover

$$\mathbb{E}\left[|Z_H - r|^{-2k}\right] \geq \mathbb{E}\left[|Z_H - r|^{-k}\right]^2 \geq \frac{1}{\kappa_V(H)^2(\alpha\theta\psi_k(H))^{2k}}. \tag{4.6}$$

Proof. Let $p(z) = \prod_{i=1}^{k}(z - r_i)$. The claim follows from the following chain of inequalities:

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|Z_H - r|^{-2k}\right]} &\geq \mathbb{E}\left[|Z_H - r|^{-k}\right] && \text{Jensen, } x \mapsto x^2 \\
&\geq \frac{1}{\alpha^k}\mathbb{E}[|p(Z_H)|^{-1}] && r \text{ is } \alpha\text{-promising} \\
&\geq \frac{1}{\alpha^k}\frac{1}{\sqrt{\mathbb{E}[|p(Z_H)|^2]}} && \text{Jensen, } x \mapsto x^2 \\
&\geq \frac{1}{\alpha^k}\frac{1}{\|e_n^* p(H)\|\kappa_V(H)} && \text{Lemma 4.1.5} \\
&\geq \frac{1}{\alpha^k}\frac{1}{\theta^k\|e_n^* \chi_k(H)\|\kappa_V(H)} && \text{Definition 1.4.2 of } \theta\text{-optimal} \\
&= \frac{1}{\alpha^k}\frac{1}{\theta^k\psi_k(H)^k\kappa_V(H)} && \text{Lemma 4.1.3.}
\end{aligned}$$

This already shows (4.6). For the other claim, rearrange both extremes of the above inequality to get

$$\begin{aligned}
\alpha\theta\kappa_V(H)^{\frac{1}{k}}\psi_k(H) &\geq \mathbb{E}\left[|Z_H - r|^{-2k}\right]^{-\frac{1}{2k}} \\
&\geq \frac{\tau_{(z-r)^k}(H)}{\kappa_V(H)^{\frac{1}{k}}} && \text{Lemma 4.1.5} \\
&\geq \frac{\psi_k(\widehat{H})}{\kappa_V(H)^{\frac{1}{k}}} && \text{Lemma 4.1.4}
\end{aligned}$$

which concludes the proof. □

In 4.1.6, we will see that when the shift associated with a promising Ritz value does not reduce the potential, Lemma 4.1.7 can be used to provide a two-sided bound on the quantities $\mathbb{E}[|Z_H - r|^{-2k}]$ and $\mathbb{E}[|Z_H - r|^{-k}]^2$. This is the main ingredient needed to obtain information about the distribution of Z_H when potential reduction is not achieved.

4.1.6 Shifting Strategy

An important component of our shifting scheme, discussed in detail in Section 4.1.7, is a simple subroutine, "Find," guaranteed to produce an α-promising Ritz value with $\alpha = \kappa_V(H)^{4k^{-1}\log k}$. Guarantees for this subroutine are stated in the lemma below and proved in Section 4.1.7.

Lemma 4.1.8 (Guarantees for Find)**.** *The subroutine* Find *specified in Section 4.1.7 produces a $\kappa_V(H)^{4k^{-1}\log k}$-promising Ritz value, using at most $12k\log kn^2 + \log k$ arithmetic operations.*

Our strategy is then built around the following dichotomy, which crucially uses the α-promising property: in the event that a degree k implicit QR step with the α-promising Ritz value output by Find does *not* achieve potential reduction, we show that there is a modestly sized set of exceptional shifts, one of which is *guaranteed* to achieve potential reduction. These exceptional shifts are constructed by the procedure "Exc" described in Section 4.1.8. The overall strategy is specified below.

$\mathrm{Sh}_{k,B}$

Input: Hessenberg H and a set $\mathcal{R}$ of θ-approximate Ritz values of H
Output: Hessenberg $\widehat{H}$.
Requires: $0 < \psi_k(H)$ and $\kappa_V(H) \le B$
Ensures: $\psi_k(\widehat{H}) \le \gamma\psi_k(H)$ and $\kappa_V(\widehat{H}) \le B$

1. $r \leftarrow \mathrm{Find}(H, \mathcal{R})$
2. If $\psi_k(\mathsf{iqr}(H, (z-r)^k)) \le \gamma\psi_k(H)$, output $\widehat{H} = \mathsf{iqr}(H, (z-r)^k)$
3. Else, $\mathcal{S} \leftarrow \mathrm{Exc}(H, r, B)$
4. For each $s \in \mathcal{S}$, if $\psi_k(\mathsf{iqr}(H, (z-s)^k)) \le \gamma\psi_k(H)$, output $\widehat{H} = \mathsf{iqr}(H, (z-s)^k)$

The failure of line (2) of $\mathrm{Sh}_{k,B}$ to reduce the potential gives useful quantitative information about the distribution of Z_H, articulated in the following lemma. This will then be used to design the set $\mathcal{S}$ of exceptional shifts produced by Exc in line (3) and prove that at least one of them makes progress in line (4).

Lemma 4.1.9 (Stagnation Implies Support)**.** *Let $\gamma \in (0,1)$ and $\theta \ge 1$, and let $\mathcal{R} = \{r_1, \ldots, r_k\}$ be a set of θ-approximate Ritz values of H. Suppose $r \in \mathcal{R}$ is α-promising and assume*

$$\psi_k\left(\mathsf{iqr}(H, (z-r)^k)\right) \ge \gamma\psi_k(H) > 0. \tag{4.7}$$

Then Z_H is well-supported on an disk of radius approximately $\alpha\psi_k(H)$ centered at r in the

following sense: for every $t \in (0,1)$:

$$\mathbb{P}\left[|Z_H - r| \le \theta\alpha \left(\frac{\kappa_V(H)}{t}\right)^{\frac{1}{k}} \psi_k(H)\right] \ge (1-t)^2 \frac{\gamma^{2k}}{\alpha^{2k}\theta^{2k}\kappa_V(H)^4}. \tag{4.8}$$

Proof. Observe that $H - r$ is invertible since otherwise, for $\widehat{H} = \mathsf{iqr}(H, (z-r)^k)$, we would have $\psi_k(\widehat{H}) = 0$ by Lemma 4.1.4. Our assumption implies that that:

$$\begin{aligned}
\gamma\psi_k(H) &\le \psi_k(\widehat{H}) && \text{hypothesis}\\
&\le \tau_{(z-r)^k}(H) && \text{Lemma 4.1.4}\\
&= \|e_n^*(H-r)^{-k}\|^{-\frac{1}{k}} && \text{definition}\\
&\le \left(\frac{\kappa_V(H)}{\mathbb{E}\left[|Z_H - r|^{-2k}\right]^{\frac{1}{2}}}\right)^{1/k} && \text{Lemma 4.1.5.}
\end{aligned}$$

Rearranging and using (4.6) from Lemma 4.1.7 we get

$$\frac{\kappa_V(H)^2}{(1-\gamma)^{2k}\psi_k(H)^{2k}} \ge \mathbb{E}\left[|Z_H - r|^{-2k}\right] \ge \mathbb{E}\left[|Z_H - r|^{-k}\right]^2 \ge \frac{1}{\alpha^{2k}\theta^{2k}\psi_k(H)^{2k}\kappa_V(H)^2}, \tag{4.9}$$

which upon further rearrangement yields the "reverse Jensen" type bound:

$$\frac{\mathbb{E}[|Z_H - r|^{-2k}]}{\mathbb{E}[|Z_H - r|^{-k}]^2} \le \left(\frac{\alpha\theta}{\gamma}\right)^{2k} \kappa_V(H)^4. \tag{4.10}$$

We now have

$$\begin{aligned}
\mathbb{P}\left[|Z_H - r| \le \frac{\alpha}{t^{1/k}}\theta\psi_k(H)\kappa_V^{1/k}\right] &= \mathbb{P}\left[|Z_H - r|^{-k} \ge t\frac{1}{\alpha^k\theta^k\psi_k(H)^k\kappa_V}\right]\\
&\ge \mathbb{P}\left[|Z_H - r|^{-k} \ge t\mathbb{E}[|Z_H - r|^{-k}]\right] && \text{by (4.9)}\\
&\ge (1-t)^2\frac{\mathbb{E}[|Z_H - r|^{-k}]^2}{\mathbb{E}[|Z_H - r|^{-2k}]} && \text{Paley-Zygmund}\\
&\ge (1-t)^2\frac{\gamma^{2k}}{\alpha^{2k}\theta^{2k}\kappa_V(H)^4} && \text{by (4.10),}
\end{aligned}$$

establishing (4.8), as desired. □

In Section 4.1.8, we will use Lemma 4.1.9 to prove the following guarantee on Exc.

Lemma 4.1.10 (Guarantees for Exc)**.** *The subroutine* Exc *specified in Section 4.1.8 produces a set $\mathcal{S}$ of exceptional shifts, one of which achieves potential reduction. If $\theta \le 2$, $\gamma = 0.8$, and $\alpha = B^{4\log k/k}$, then both the arithmetic operations required for* Exc*, and the size of $\mathcal{S}$, are at most*

$$N_{\mathsf{net}}\left(0.002B^{-\frac{8\log k+4}{k}}\right),$$

where $N_{\mathsf{net}}(\epsilon) = O(\epsilon^{-2})$ denotes number of points in an efficiently computable ϵ-net of the unit disk. In the normal case, taking $B = \alpha = \theta = 1$, $k = 4$, $\gamma = 0.8$, the arithmetic operations required and the size of $|\mathcal{S}|$ are both bounded by 50.

Remark 4.1.11 (Improving the Disk to an Annulus)**.** Control on the other tail of $|Z_H - r|$ can be achieved by using Markov's inequality and the upper bound (4.10) on the inverse moment $\mathbb{E}[|Z_H - r|^{-2k}]$. Then, for $k \gg \log \kappa_V(H)$, the control on both tails yields that the distribution of Z_H has significant mass on a thin annulus (the inner and outer radii are almost the same).[2] In this scenario one can take a net $\mathcal{S}$ with fewer elements when calling the exceptional shift, which would reduce the running time of $T_{\mathsf{exc}}(k, B)$. However, following this path would complicate the analysis and for the sake of exposition we do not pursue it any further in dissertation.

We are now ready to prove Theorem 1.4.3.

Proof of Theorem 1.4.3. Rapid convergence. In the event that we choose a α-promising Ritz value in step (1) that does not achieve potential reduction in step (2), Lemma 4.1.10 then guarantees we achieve potential reduction in (3). Thus each iteration decreases the potential by a factor of at least γ, and since $\psi_k(H_0) \le \|H\|$ we need at most

$$\frac{\log(1/\delta)}{\log(1/\gamma)} \le 4 \log(1/\delta)$$

iterations before $\psi_k(H_t) \le \delta \|H_0\|$, which in particular implies δ-decoupling.

Arithmetic Complexity. Computing a full set $\mathcal{R}$ of θ-approximate Ritz values of H has a cost $T_{\mathrm{OptRitz}}(k, \theta, \delta)$. Then, using an efficient implicit QR algorithm (cf. Definition 4.1.2) each computation of $\mathsf{iqr}(H, (z - r_i)^k)$ has a cost of $7kn^2$. By Lemma 4.1.8, we can produce a promising Ritz value in at most $12k \log kn^2 + \log k$ arithmetic operations. Then, in the event that the promising shift fails to reduce the potential the algorithm calls Exc, which takes $N_{\mathsf{net}}(0.002B^{-\frac{8 \log k + 4}{k-1}})$ arithmetic operations to specify the set $\mathcal{S}$ of exceptional shifts. Some exceptional shift achieves potential reduction, and we pay $7kn^2$ operations for each one that we check. □

4.1.7 Efficiently Finding a Promising Ritz Value

In this section we show how to efficiently find a promising Ritz value, in $O(n^2 k \log k)$ arithmetic operations. Note that it is trivial to find a $\kappa_V(H)^{2/k}$-promising Ritz value in $O(n^2 k^2)$ arithmetic operations simply by computing $\|e_n^*(H - r_i)^{-k/2}\|$ for $i = 1, \ldots, k$ with k calls to $\mathsf{iqr}(H, (z - r_i)^{k/2})$, choosing the maximizing index i, and appealing to Lemma 4.1.5. The

[2] We note in passing (cf. [122]) that when H is normal, $\alpha = 1$, $\theta = \gamma = 0$, and $k = 1$, the above arguments can be modified to show that, under the assumption of Lemma 4.1.9, Z_H is fully supported on a circle with center r and radius $\psi_k(H)$, and hence $\frac{1}{\psi_k(H)}(H - r)$ is a unitary matrix.

content of Lemma 4.1.8 below that this can be done considerably more efficiently if we use a binary search type procedure. This improvement has nothing to do with the dynamical properties of our shifting strategy so readers uninterested in computational efficiency may skip this section.

Find

Input: Hessenberg H, a set $\mathcal{R} = \{r_1, \ldots, r_k\}$ of θ-optimal Ritz values of H.
Output: A complex number $r \in \mathcal{R}$
Requires: $\psi_k(H) > 0$
Ensures: r is α-promising for $\alpha = \kappa_V(H)^{\frac{4 \log k}{k}}$.

1. For $j = 1, ..., \log k$
   a) Evenly partition $\mathcal{R} = \mathcal{R}_0 \sqcup \mathcal{R}_1$, and for $b = 0, 1$ set $p_{j,b} = \prod_{r \in \mathcal{R}_b}(z - r)$
   b) $\mathcal{R} \leftarrow \mathcal{R}_b$, where b maximizes $\|e_n^* p_{j,b}(H)^{-2^{j-1}}\|$
2. Output $\mathcal{R} = \{r\}$

Proof of Lemma 4.1.8 (Guarantees for Find*).* First, observe that $\|e_n^* q(H)\| \neq 0$ for every polynomial appearing in the definition of Find, since otherwise we would have $\psi_k(H) = 0$.

On the first step of the subroutine $p_{1,0}p_{1,1} = p$, the polynomial whose roots are the full set of approximate Ritz values, so

$$\begin{aligned}
\max_b \|e_n^* p_{1,b}(H)^{-1}\| &\geq \frac{1}{\kappa_V(H)^2} \mathbb{E}\left[\frac{1}{2}\left(|p_{1,0}(Z_H)|^{-2} + |p_{1,1}(Z_H)|^{-2}\right)\right] && \text{Lemma 4.1.5} \\
&\geq \frac{1}{\kappa_V(H)^2} \mathbb{E}[|p(Z_H)|^{-1}] && \text{AM/GM.}
\end{aligned}$$

On each subsequent step, we've arranged things so that $p_{j+1,0}p_{j+1,1} = p_{j,b}$, where b maximizes $\|e_n^* p_{j,b}(H)^{-2^{j-1}}\|$, and so by the same argument

$$\begin{aligned}
&\max_b \|e_n^* p_{j+1,b}(H)^{-2^j}\|^2 \\
\geq &\frac{1}{\kappa_V(H)^2} \mathbb{E}\left[\frac{1}{2}\left(|p_{j+1,0}(Z_H)|^{-2^{j+1}} + |p_{j+1,1}(Z_H)|^{-2^{j+1}}\right)\right] && \text{Lemma 4.1.5} \\
\geq &\frac{1}{\kappa_V(H)^2} \mathbb{E}\left[|p_{j+1,0}(Z_H)p_{j+1,1}(Z_H)|^{-2^j}\right] && \text{AM/GM} \\
\geq &\frac{1}{\kappa_V(H)^4} \|e_n^*(p_{j+1,0}(H)p_{j+1,1}(H))^{-2^{j-1}}\| && \text{Lemma 4.1.5} \\
= &\frac{1}{\kappa_V(H)^4} \max_b \|e_n^* p_{j,b}(H)^{-2^{j-1}}\|.
\end{aligned}$$

Paying a further $\kappa_V(H)^2$ on the final step to convert the norm into an expectation, we get

$$\mathbb{E}\left[|Z_H - r|^{-k}\right] \geq \frac{1}{\kappa_V(H)^{4\log k}} \mathbb{E}\left[|p(Z_H)|^{-1}\right]$$

as promised.

For the runtime, we can compute every $\|e_n^* p_{j,b}(H)^{-2^{j-1}}\|$ by running an implicit QR step with the polynomials $p_{j,b}^{2^{j-1}}$, all of which have degree $k/2$. There are $2\log k$ such computations throughout the subroutine, and each one requires $6kn^2$ arithmetic operations. Beyond that we need only compare the two norms on each of the $\log k$ steps. □

Remark 4.1.12 (Opportunism and Judicious Partitioning)**.** In practice, it may be beneficial to implement Find *opportunistically*, meaning that in each iteration one should check if the new set of Ritz values gives potential reduction (this can be combined with the computation of $\|e_n^* p_{j,b}(H)^{-2^{j-1}}\|$ and implemented with no extra cost). Moreover, note that Find does not specify a way to partition the set of Ritz values obtained after each iteration, and as can be seen from the above proof, the algorithm works regardless of the partitioning choices. It is conceivable that a judicious choice of the partitioning could be used to obtain further improvements.

4.1.8 Analysis of the Exceptional Shift

To conclude our analysis, it remains only to define the subroutine "Exc," which produces a set $\mathcal{S}$ of possible exceptional shifts in the event that an α-promising Ritz value does not achieve potential reduction. The main geometric intuition is captured in the case when H is normal and $\kappa_V(H) = 1$. Here, Find gives us a 1-promising Ritz value r and Lemma 4.1.9 with $t = 1/2$ tells us that if r does not achieve potential reduction, than Z_H has measure at least $\frac{1}{4}(\gamma/\theta)^{2k}$ on a disk of radius $R := 2^{1/k}\theta\psi_k(H)$.

For any $\epsilon > 0$, we can easily construct an *$R\epsilon$-net* $\mathcal{S}$ contained in this disk — i.e., a set with the property that every point in the disk is at least $R\epsilon$-close to a point in $\mathcal{S}$ — with $O(1/\epsilon)^2$ points. One can then find a point $s \in \mathcal{S}$ satisfying

$$\begin{aligned}
\tau_{(z-s)^k}(H)^{-2k} &= \|e_n^*(H-s)^{-k}\|^2 \\
&= \mathbb{E}[|Z_H - s|^{-2k}] \\
&\geq \frac{\mathbb{P}[|Z_H - s| \leq \psi_k(H)]}{|\mathcal{S}|(R\epsilon)^{2k}} \\
&\approx \frac{1}{4}\left(\frac{\gamma}{\theta}\right)^{2k} \frac{1}{R^{2k}\epsilon^{2k-2}},
\end{aligned}$$

where the first equality is by normality of H, and second inequality comes from choosing $s \in \mathcal{S}$ to maximize $|Z_H - s|^{-2k}$. Since $\psi_k(\mathsf{iqr}(H, (z-s)^k)) \leq \tau_{(z-s)^k}(H)$, we can ensure potential reduction by setting $\epsilon \approx \frac{\gamma^2 R}{\theta\psi_k(H)} \approx (\gamma/\theta)^2$.

When H is nonnormal, the chain of inequalities above hold only up to factors of $\kappa_V(H)$, and Find is only guaranteed to produce a $\kappa_V(H)^{4\log k/k}$-promising Ritz value. The necessary adjustments are addressed below in the implementation of Exc and the subsequent proof of its guarantees.

Exc

Input: Hessenberg H, a θ-approximate Ritz value r, a condition number bound B, promising parameter α
Output: A set $\mathcal{S} \subset \mathbb{C}$
Requires: $\kappa_V(H) \le B$, r is α-promising, and $\psi_k(\mathsf{iqr}(H, (z-r)^k) \ge \gamma\psi_k(H)$
Ensures: For some $s \in \mathcal{S}$, $\psi_k(\mathsf{iqr}(H, (z-s)^k) \le \gamma\psi_k(H)$

1. $R \leftarrow 2^{1/k}\theta\alpha B^{1/k}\psi_k(H)$
2. $\epsilon \leftarrow \left(\frac{\gamma^2}{(12B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}$
3. $\mathcal{S} \leftarrow \epsilon R$-net of $R\psi_k(H)$.

Proof of Lemma 4.1.10: Guarantees for Exc. Instantiating $t = 1/2$ in equation (4.8), we find that for the setting of R in line (1) of Exc,

$$\mathbb{P}\left[|Z_H - r| \le D(r, R)\right] \ge \frac{1}{4B^4}\left(\frac{\gamma}{\alpha\theta}\right)^{2k}.$$

Let $\mathcal{S}$ be an ϵR-net of $D(r, R)$; it is routine that such a net has at most $(1 + 2/\epsilon)^2 \le 9/\epsilon^2$ points. By Lemma 4.1.4, to show that some $s \in \mathcal{S}$ achieves potential reduction, it suffices to find one for which

$$\|e_n^*(H - s)^{-k}\|^2 \ge \frac{1}{\gamma^{2k}\psi_k(H)^{2k}}.$$

We thus compute

$$\begin{aligned}
\max_{s\in\mathcal{S}} \|e_n^*(H-s)^{-k}\|^2 &\ge \frac{1}{\kappa_V(H)^2|\mathcal{S}|}\sum_{s\in\mathcal{S}} \mathbb{E}\left[|Z_H - s|^{-2k}\right] \\
&\ge \frac{\epsilon^2}{9B^2}\mathbb{E}\left[\sum_{s\in\mathcal{S}} |Z_H - s|^{-2k} \cdot \mathbb{1}\left\{Z_H \in D(r,R)\right\}\right] && \text{Fubini, } \kappa_V(H) \le B \\
&\ge \frac{\epsilon^2}{9B^2}\mathbb{E}\left[\max_{s\in\mathcal{S}} |Z_H - s|^{-2k} \cdot \mathbb{1}\left\{Z_H \in D(r,R)\right\}\right] \\
&\ge \frac{\epsilon^2}{9B^2}\mathbb{E}\left[\frac{\mathbb{1}\left\{Z_H \in D(r,R)\right\}}{(\epsilon R)^{2k}}\right] && \mathcal{S} \text{ is an } \epsilon R\text{-net} \\
&\ge \frac{\mathbb{P}[Z_H \in D(r,R)]}{9B^2R^{2k}\epsilon^{2k-2}}
\end{aligned}$$

$$\geq \frac{1}{\gamma^{2k}\psi(H)^{2k}}$$

with the second to last line following from the fact that some $s \in \mathcal{S}$ is at least ϵR-close to Z_H whenever the latter is in $D(r, R)$, and the final inequality holding provided that

$$\epsilon \leq \left(\frac{\mathbb{P}[|Z_H - r| \leq R\psi_k(H)]\gamma^{2k}\psi_k(H)^{2k}}{9B^2R^{2k}}\right)^{\frac{1}{2k-2}}.$$

Expanding the probability and using the definition of R in line 1, it suffices to set ϵ smaller than

$$\left(\frac{\gamma^{2k}}{4B^4\alpha^{2k}\theta^{2k}} \cdot \frac{\gamma^{2k}\psi_k(H)^{2k}}{9B^2} \cdot \frac{1}{4B^2\alpha^{2k}\theta^{2k}\psi_k(H)^{2k}}\right)^{\frac{1}{2k-2}} = \left(\frac{\gamma^2}{(12B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}},$$

which is the quantity appearing in line 2. Setting $\theta = 2$, $\gamma = 0.8$, and $\alpha = B^{4\log k/k}$, and using $k \geq 2$, we obtain the expression appearing in $N_{\mathsf{net}}(\cdot)$ in the statement of Lemma 4.1.10.

However, a more practical choice (and the one that we will use in Section 4.2) is an equilateral triangular lattice with spacing $\sqrt{3}\epsilon$, intersected with the $D(r, (1+\epsilon)R)$. Such a construction is optimal as $\epsilon \to 0$, and can be used to give a better bound on $N_{\mathsf{net}}(\epsilon)$ when ϵ is small. For instance, by adapting an argument of [3, Lemma 2.6] one can show that this choice of $\mathcal{S}$ satisfies

$$N_{\mathsf{net}}(\epsilon) \leq \frac{2\pi}{3\sqrt{3}}(1 + 1/\epsilon)^2 + \frac{4\sqrt{2}}{\sqrt{3}}(1 + 1/\epsilon) + 1.$$

In the normal case, when $B = \alpha = \theta = 1$, $k = 4$, and $\gamma = 0.8$, the above bound gives

$$|\mathcal{S}| \leq N_{\mathsf{net}}\left(\left(\frac{0.8^2}{12^{1/4}}\right)^{4/3}\right) \leq 49.9.$$

□

4.2 Numerical Stability

4.2.1 Results and Organization

When trying to obtain guarantees for the shifted QR algorithm in finite arithmetic, the following two interrelated issues arise:

1. *Forward Stability of QR Steps.* Consider a degree k shifted QR step:

$$p(H) = QR \qquad H = Q^*HQ,$$

where $p(z) = (z - r_1) \dots (z - r_k)$ is a monic polynomial of degree k and H is an upper Hessenberg matrix. It is well-known that such a step can be implemented in a way which is backward stable, in the sense that the finite arithmetic computation produces a matrix H which is the unitary conjugation of a matrix near H [155]. Backward stability is sufficient to prove correctness of the shifted QR algorithm in finite arithmetic, i.e., whenever it converges in a small number of iterations, the backward error is controlled. However, it is insufficient for proving an upper bound on the number of iterations before decoupling, which requires showing that certain subdiagonal entries of the Hessenberg iterates decay rapidly — to reason about these entries, some form of forward stability is required. The issue is that a shifted QR step is *not* forward stable when $p(H)$ is nearly singular (which can occur before decoupling). Thus, the existing convergence proofs break down in finite arithmetic whenever this situation occurs. As far as we know, there is no complete and published proof of rapid convergence of the implicitly shifted QR algorithm with any shifting strategy in finite arithmetic, even on symmetric matrices (see Section 4.2.2 for a detailed discussion).

2. *Computation of (Good Quality) Approximate Ritz Values.* The Ritz values of order k of an upper Hessenberg matrix H are equal to the eigenvalues of its bottom right $k \times k$ corner $H_{(k)}$; they are also defined variationally as the zeros of the monic degree k polynomial p_k minimizing $\|e_n^* p_k(H)\|$, where e_n is an elementary basis vector. All of the higher order shifting strategies we are aware of are defined in terms of these Ritz values. However, we are not aware of any theoretical analysis of how to compute the Ritz values (approximately) in the case of nonsymmetric $H_{(k)}$, nor a theoretical treatment of which notion of approximation is appropriate for their use in the shifted QR iteration.[3]

Section 4.2 contains the following contributions, which address the above complications and in conjunction provide a proof of Theorem 1.4.6 advertised in Section 1.4.

(i) Forward Stability by Regularization. We handle the first issue above simply by replacing any given shifts $r_1, \dots, r_k$ in a QR step by random perturbations $r_1 + w_1, \dots, r_k + w_k$ where the w_i are independent random numbers of an appropriate size (which depends on $\kappa_V(H)$ and $\text{gap}(H)$). We refer to this technique as *shift regularization* and show in Section 4.2.4 (Lemma 4.2.13) that it yields forward stability of an implicit QR step with high probability, for any Hessenberg matrix H with an upperbound on $\kappa_V(H)$ and a lowerbound on $\text{gap}(H)$, and any shifts $r_1, \dots, r_k$. Note that here the w_i must be large enough to ensure forward stability, but small enough to preserve the convergence properties of the QR iteration, which are presumably tied to the $r_1, \dots, r_k$ being approximate Ritz values. The precise notion of "approximate" used thus determine how constrained we are in choosing our shifts while maintaining good convergence properties.

[3] In practice, and in the current version of LAPACK, the prescription is to run the shifted QR algorithm itself on $H_{(k)}$, but there are no proven guarantees for this approach.

The proof of forward stability requires us to establish stronger backward stability of implicit QR steps than was previously recorded in the literature; this appears in Section 4.2.4 and may be of independent interest.

(ii) Optimal Ritz Values/Early Decoupling Dichotomy. The second issue is more subtle. The notion of approximate Ritz values relevant for analyzing $\mathsf{Sh}_{k,B}$ is the following variational one. Recall from Section 1.4 that $\{r_1, \ldots, r_k\} \subset \mathbb{C}$ is called a set of *θ-optimal Ritz values* of a Hessenberg matrix H if:

$$\|e_n^*(H - r_1)\ldots(H - r_k)\|^{1/k} \leq \theta \min_p \|e_n^* p(H)\|^{1/k}, \tag{4.11}$$

where the minimization is over monic polynomials of degree k. Thus, the true Ritz values are 1-optimal.

It is not immediately clear how to efficiently compute a set of θ-optimal Ritz values, so we reduce this task to the more standard one of computing forward-approximate Ritz values, which are just forward-approximations of the eigenvalues of $H_{(k)}$ with an appropriately chosen accuracy parameter β roughly proportional to the right hand side of (4.11). Our key result (Theorem 4.2.14) is the following *dichotomy*: if a set of β-forward approximate Ritz values $r_1, \ldots, r_k$ of H is *not* θ-optimal, then one of the Ritz values r_j must be close to an eigenvalue of H and the corresponding right eigenvector of H must have a large inner product with e_n. In the latter scenario we show that a single degree k implicit QR step using the culprit Ritz value r_j as a shift must lead to immediate decoupling, which we refer to as *early decoupling*.

Importantly, this dichotomy is compatible with the random regularizing perturbation used in (i), since the property of being a β-forward approximate Ritz value is preserved (with a slight increase in β) under small perturbations $r_i \to r_i + w_i$ when $|w_i| \ll \beta$. Thus, as long as we can compute β-forward approximations $r_1, \ldots, r_k$ of the eigenvalues of $H_{(k)}$, the combination of (i) and the dichotomy guarantees that with high probability, $r_1 + w_1, \ldots, r_k + w_k$ are θ-optimal Ritz values *and* the corresponding QR step is forward stable (which is exactly what is needed in order to analyze convergence of the iteration) — *or* we achieve early decoupling.

Example 4.2.1 (Necessity of Forward Error for Ritz Value Optimality)**.** It is natural to ask whether the weaker property of being a β-backward approximation of the eigenvalues of $H_{(k)}$ is sufficient for producing $O(1)$-optimal Ritz values when the right hand side of (4.11) is of scale β. The following example shows that this is not in general the case: let T be an $n \times n$ Hessenberg Toeplitz matrix with 1s on the superdiagonal, δs on the subdiagonal, and $T(1, n) = 1$. Let the bottom right $k \times k$ corner of T be $T_{(k)}$ and let $T'_{(k)} = T_{(k)} + \beta e_k e_1^*$. An explicit computation of characteristic polynomials shows that if $r_1, \ldots, r_k$ are the eigenvalues of $T_{(k)}$ and $r'_1 \ldots, r'_k$ are the eigenvalues of $T'_{(k)}$ (which are β-backward approximations of the r_i) then

$$\delta = \|e_n^*(T - r_1)\ldots(T - r_k)\|^{1/k} \ll \|e_n^*(T - r'_1)\ldots(T - r'_k)\|^{1/k} \approx \beta^{1/k},$$

unless $\beta = O(\delta^k)$. But this latter condition is enough to guarantee that the $r'_1 \ldots, r'_k$ are δ-forward approximations of the Ritz values of T, which is what we require. Since T is close

to normal when $\delta \ll 1/n$, this example also highlights that while we may have control of the nonnormality of H, this does not imply any control on the nonnormality of $H_{(k)}$ in general.

To produce a complete eigenvalue algorithm, we also need the following auxiliary ingredients.

(iii) Analysis of Deflation. Once the shifting strategy $\mathsf{Sh}_{k,B}$ has been used to achieve decoupling, it is typical to *deflate* the resulting matrix by zeroing out small subdiagonal elements. The outcome of this procedure is a block upper triangular matrix whose diagonal blocks are themselves upper Hessenberg, allowing one to recursively apply $\mathsf{Sh}_{k,B}$. Because our analysis of $\mathsf{Sh}_{k,B}$ relies on $\kappa_V(H)$ and $\mathrm{gap}(H)$ being controlled, it is critical that we can preserve these quantities when deflating and passing to a submatrix. This will be handled in Section 4.2.7.

4.2.2 Related Work

The need for a finite arithmetic convergence analysis of shifted QR in the case of symmetric tridiagonal matrices was noted in the remarkable thesis of Sanderson [136], who observed that it does not follow from the exact arithmetic analysis of Wilkinson [174]. Sanderson formally proved the convergence of the tridiagonal QR algorithm with explicit (as opposed to implicit) QR steps using Wilkinson shift under certain additional assumptions, one of which [136, Section 4] is that the "computation of the [Wilkinson shift] be done more accurately [i.e., in exact arithmetic]". Sanderson left open the question of analyzing implicit shifted QR and gave an example for which its convergence breaks down unless the machine precision is sufficiently small in relation to the subdiagonal entries of the matrix. These insightful observations of Sanderson are consistent with the approach taken in this section, and Sanderson's question is resolved by Remark 1.4.9, albeit with a different shifting strategy.

Forward Stability of Shifted QR. An important step towards understanding and addressing the two issues mentioned at the beginning of the introduction was taken by Parlett and Le [126], who showed that for symmetric tridiagonal matrices, high sensitivity of the next QR iterate to the shift parameter (a form of forward instability) is always accompanied by "premature deflation", which is a phenomenon specific to "bulge-chasing" implementations of the implicit QR algorithm on tridiagonal matrices. Our dichotomy is distinct from but was inspired by their paper, and carries the same conceptual message: if the behavior of the algorithm is highly sensitive to the choice of shifts, then one must already be close to convergence in some sense.

Watkins [169] argued informally (but did not prove) that the implicit QR iteration should in many cases converge rapidly even in the presence of forward instability. This is an intriguing direction for further theoretical investigation, and could potentially lead to provable guarantees for the shifted QR algorithm with lower precision than required here.

Aggressive Early Deflation. The classical criterion for decoupling/deflation in shifted QR algorithms is the existence of small subdiagonal entries of H. The celebrated papers [31, 32]

introduced an additional criterion called aggressive early deflation which yields significant improvements in practice. Kressner [98] showed that this criterion is equivalent to checking for converged Ritz values (i.e., Ritz pairs which are approximate eigenpairs of H), and "locking and deflating them" (i.e., deflating while preserving the Hessenberg structure of H) using Stewart's Krylov-Schur algorithm [145].

The early decoupling procedure introduced in this section is similar in spirit to aggressive early deflation — in that it detects Ritz values which are close to eigenvalues of H and enables decoupling even when the subdiagonal entries of H are large — but different in that it does not require the corresponding Ritz vector to have a small residual, and it ultimately produces classical decoupling in the sense of a small subdiagonal entry.

Shift Blurring. The shifting strategies considered in Section 4.1 use shift polynomials $p(z) = (z - r_1) \dots (z - r_k)$ of degree k where k is roughly proportional to $\log \kappa_V(H)$. It was initially proposed [5] that such higher degree shifts should be implemented via "large bulge chasing", a procedure which computes the QR decomposition of $p(H)$ in a single implicit QR step. This procedure was found to have poor numerical stability properties, which was referred to as "shift blurring" and explained by Watkins [172] and further by Kressner [97] by relating it to some ill-conditioned eigenvalue and pole placement problems.

To avoid these issues, we implement all degree k QR steps in this section as a sequence of k degree-1 "small bulge" QR steps. However, since our analysis requires establishing forward stability of each degree k step, the amount of numerical precision required for provable $\delta-$decoupling increases as a function of k, roughly as $O(k \log(n/\delta))$ bits. This increase in precision is sufficient to avoid shift blurring. We suspect that forward stability of large bulge chasing can be established given a similar increase in precision, and leave this as a direction for further work.

4.2.3 Preliminaries

Finite Precision Arithmetic.

We will use the floating point axioms discussed in Section 1.1 (ignoring overflow and underflow as is customary), and use $\mathbf{u}$ to denote the unit roundoff.

Our implementation of implicit QR steps is based on *Givens rotations*. If $x \in \mathbb{R}^2$, write $\mathsf{giv}(x)$ for the 2×2 Givens rotation mapping $\mathsf{giv}(x) : x \mapsto \|x\| e_1$. It is routine [85, Lemmas 19.7-19.8, e.g.] that, assuming $\mathbf{u} \leq 1/24$, one can compute the norm of x with relative error $2\mathbf{u}$ and apply $\mathsf{giv}(x)$ to a vector $y \in \mathbb{R}^2$ in floating point so that

$$\left| \widetilde{(\mathsf{giv}(x)y)}_i - (\mathsf{giv}(x)y)_i \right| \leq \|y\| \frac{6\mathbf{u}}{1 - 6\mathbf{u}} \leq \|y\| \cdot 8\mathbf{u} \qquad i = 1, 2.$$

For some tasks, our algorithm and many of its subroutines need to set certain scalar parameters in order to know when to halt, at what scale to perform certain operations and how many iterations to perform. In this context, sometimes the algorithm will have to

compute k-th roots for moderate values of k — even though these operations are not directly used on the matrices in question. We will assume that the following elementary functions can be computed accurately and relatively quickly.

Lemma 4.2.2 (kth Roots)**.** *There exist small universal constants $C_{\mathsf{root}}, c_{\mathsf{root}} \geq 1$, such that whenever $kc_{\mathsf{root}}\mathbf{u} \leq \epsilon \leq 1/2$ and for any $a \in \mathbb{R}^+$, there exists an algorithm that computes $a^{1/k}$ with relative error ϵ in at most*

$$T_{\mathsf{root}}(k, \epsilon) := C_{\mathsf{root}} k \log(k \log(1/\epsilon))$$

arithmetic operations.

Proof sketch. Use Newton's method, with starting point found via bisection. □

Random Sampling Assumptions.

As discussed above, we will repeatedly regularize our shifts by replacing each with uniformly random point on a small surrounding disk of radius $O(\delta^2)$, where δ is the accuracy. To simplify the presentation, we will assume that these perturbations can be executed in *exact* arithmetic. Importantly, this assumption's only impact is on the failure probability of the algorithm, and its effect is quite mild. We will see below that the algorithm fails when one of our randomly perturbed shifts happens to land too close to an eigenvalue, and we bound the failure probability by computing the area of the 'bad' subset of the disk where this occurs. If the random perturbation was instead executed in finite arithmetic, the probability of landing in the bad set differs from this estimate by $O(\mathbf{u}/\delta^2)$. Since we will set $\mathbf{u} = o(\delta^2)$, this discrepancy can reasonably be neglected.

Definition 4.2.3 (Efficient Perturbation Algorithm)**.** An efficient random perturbation algorithm takes as input $s \in \mathbb{C}$ and $R > 0$, and generates a random $w \in \mathbb{C}$ distributed uniformly in the disk $D(s, R)$ using C_{D} arithmetic operations.

Reader Guide and Parameter Settings

There are many algorithm inputs, constants, and parameters that the reader will encounter in Section 4.2; we will collect them here, along with some typical settings. We will regard our main algorithm ShiftedQR in fact as a family of algorithms, indexed by several defining parameters; these in turn used to set a number of global constants used by the algorithm and its subroutines. The most important of the former is the "nonnormality" or condition number bound B, from which we define the shift degree k to be the smallest power of 2 for which

$$B^{\frac{8\log k+3}{k-1}} \cdot (2B^4)^{\frac{2}{k-1}} \leq 3, \tag{4.12}$$

which makes $k = O(\log B \log\log B)$. We further define the auxiliary constants

$$\alpha := (1.01B)^{4\log k/k} \in [1,2], \qquad \theta := \tfrac{1.01}{0.998^{1/k}}(2B^4)^{1/2k} \in [1,2] \qquad \gamma := 0.2, \tag{4.13}$$

Defining Parameter	Meaning	Typical Setting
B	Eigenvector Condition Number Bound	$B \geq 2\kappa_V(H)$
Γ	Minimum Gap Bound	$\Gamma \leq \text{gap}(H)/2$
Σ	Operator Norm Bound	$\Sigma \geq 2\|H\|$
k	Shift Degree	$O(\log B \log\log B)$
Global Constant		
α	Ritz Value Promising-ness	$\alpha \in [1, 2]$
θ	Ritz Value Optimality	$\theta \in [1, 2]$
γ	Decoupling Rate	0.2

Table 4.1: Global Data for ShiftedQR

Input Parameter	Meaning	Typical Setting
H	Upper Hessenberg Matrix	
δ	Accuracy	
ϕ	Failure Probability Tolerance	
Internal Parameter		
ϵ	Working Accuracy	$\Omega\left(\min\{\delta n^{-2}, \Gamma n^{-3/2} B^{-2}\}\right)$
φ	Working Failure Probability Tolerance	$\Omega\left(\frac{\phi}{\log(\epsilon/\Sigma)}\right)$
η_1, η_2	Regularization Parameters	$\Omega(\epsilon^2), \Omega(\epsilon^2 \phi^{-1/2} \Sigma^{-1})$
β	Forward Accuracy for Ritz Values	$\Omega(\epsilon^2 \Sigma^{-1})$
$\mathcal{R}$	Approximate Ritz Values	
$\mathcal{S}$	Exceptional Shifts	

Table 4.2: Input and Internal Parameters for ShiftedQR

which depend only on B.

Table 4.2 contains the input parameters for ShiftedQR, as well as internal parameters used by its subroutines. The setting of the working accuracy below is to ensure that the norm, eigenvector condition number, and minimum eigenvalue gap are controlled for every matrix H' encountered in the course of the algorithm, in the sense that

$$\kappa_V(H') \leq 2\kappa_V(H) \leq B \qquad \|H'\| \leq 2\|H\| \leq \Sigma \qquad \text{gap}(H') \geq \text{gap}(H)/2 \geq \Gamma.$$

We will not include the defining parameters or global constants as input to ShiftedQR or its subroutines, and instead assume that all subroutines have access to them; however, we will for clarity keep track of which of this *global data* each subroutine uses, and any constraints

that it places on their inputs. Table 4.3 lists the main subroutines (note that we will write $\mathsf{Sh}_{k,B}$ for the finite arithmetic implementation of $\mathrm{Sh}_{k,B}$).

Subroutine	Action	Output	Input	Global Data
IQR	Implicit QR Step	$\widetilde{H}, \widetilde{R}$	$H, p(z)$	
Tau^m	Approx. $\tau^m_{p(z)}(H) = \|e_n^* p(H)^{-1}\|$	$\widetilde{\tau^m}$	$H, p(z)$	
Optimal	Check Ritz Value Optimality	opt	$H, \mathcal{R}$	θ
RitzOrDecouple	Compute θ-Optimal Ritz Values	$H, \mathcal{R}$, dec	H, ϵ, ϕ	Σ, Γ, θ
Find	Find a α-Promising Ritz Value	r	$H, \mathcal{R}$	α
Exc	Compute Exceptional Shifts	$\mathcal{S}$	H, r, ϵ, ϕ	$B, \Sigma, \gamma, \theta, \alpha$
$\mathsf{Sh}_{k,B}$	Shifting Strategy to Reduce ψ_k	H	$H, \mathcal{R}, \epsilon, \phi$	$B, \Sigma, \gamma, \theta, \alpha$
deflate	Deflate a Decoupled Matrix	$H_1, H_2, ...$	H, ϵ	

Table 4.3: Subroutines of ShiftedQR

Absolute vs. Relative Decoupling. Because ShiftedQR and its subroutines do not have direct access to the norms of matrices, we will find it useful for the remainder of the section to work with an *absolute* notion of decoupling, instead of the relative one used in Section 4.1. In particular, we will say that a matrix H is ϵ-*decoupled* if one of its k bottom subdiagonal entries is smaller than ϵ (as opposed to $\epsilon\|H\|$), and ϵ-*unreduced* if every one of its k bottom subdiagonal entries is larger than ϵ.

4.2.4 Implicit QR: Implementation, Forward Stability, and Regularization

Here we present a standard implementation (called "IQR") of a degree 1 (i.e., single shift) implicit QR step using Givens rotations (see [55, Section 4.4.8]) and provide an analysis of its backward stability which is slightly stronger than the guarantees of [155][4]. We then use this to give a corresponding backward error bound for a degree k IQR step. We suspect much of this material is already known to experts, but we could not find it in the literature so we record it here.

We will prove bounds on the forward error of a degree k IQR step in terms of the distance of the shifts to the spectrum; we will accordingly refer to shifts which are appropriately far away from the spectrum as *forward stable*. We also record a forward error bound on the bottom right entry R_{nn} of the QR factorization, which is used in analyzing many shifting strategies.

[4] [155] uses Householder reflectors instead of Givens rotations. We have chosen the latter for simplicity of exposition, but the stronger backward stability analysis obtained in Lemma 4.2.8 can also be shown for Householder reflectors.

We show in Section 4.2.4 that a sufficiently large random perturbation of any choice of shifts is commensurately forward stable, with high probability.

Description and Backward Stability of IQR

We begin with some preliminaries on implicit QR steps in exact arithmetic.

Definition 4.2.4. The *QR decomposition* of an invertible matrix A is the unique factorization $A = QR$ where Q is unitary and R is upper triangular with positive diagonal entries. We will use

$$[Q, R] = \mathrm{qr}(A)$$

to signal that Q and R are the matrices coming from the QR decomposition of A.

Given a polynomial $p(z)$ and a Hessenberg matrix H, $\mathsf{iqr}(H, p(z))$ will denote the matrix $H = Q^*HQ$ where $[Q, R] = p(H)$. When $p(z) = z - s$ we will use $\mathsf{iqr}(H, s)$ as a shorthand notation for $\mathsf{iqr}(H, z - s)$. We will also denote by $\varkappa(A) := \|A\|\|A^{-1}\|$ the condition number of a matrix A. We pause to verify a fundamental composition property of iqr; the proof is standard (e.g. see [155, Section 2.3]), but we will need to adapt it in the sequel so we include it for the reader's convenience.

Lemma 4.2.5. *For any invertible H and polynomial $p(z) = (z - r_1) \cdots (z - r_k)$,*

$$\mathsf{iqr}(H, p(z)) = \mathsf{iqr}(\cdots \mathsf{iqr}(\mathsf{iqr}(H, r_1), r_2), ..., r_k). \tag{4.14}$$

Moreover, if $[Q, R] = \mathrm{qr}(p(H))$, $H_1 = H$, and for each $\ell \in [k]$ we set $[Q_\ell, R_\ell] := \mathrm{qr}(H_\ell - r_\ell)$ and $H_{\ell+1} := Q_\ell^* H_\ell Q_\ell$, then

$$Q = Q_1 \cdots Q_k \quad \textit{and} \quad R = R_k R_{k-1} \cdots R_1. \tag{4.15}$$

Proof. Repeatedly using definition of Q_ℓ, R_ℓ, and H_ℓ for each $\ell \in [k]$, we can compute

$$\begin{aligned}
p(H) = p(H_1) &= (H_1 - r_k) \cdots (H_1 - r_1) \\
&= (H_1 - r_k) \cdots (H_1 - r_2) Q_1 R_1 && H_1 - r_1 = Q_1 R_1 \\
&= (H_1 - r_k) \cdots Q_1 (H_2 - r_2) R_1 && H_2 = Q_1^* H_1 Q_1 \\
&= (H_1 - r_k) \cdots (H_1 - r_3) Q_1 Q_2 R_2 R_1 && H_2 - r_2 = Q_2 R_2 \\
&= Q_1 Q_2 \cdots Q_k R_k R_{k-1} \cdots R_1, && \text{etc.}
\end{aligned}$$

where in the final equality we continue passing $Q_1 \cdots Q_\ell$ across the term $H_1 - r_\ell$ and then replace the resulting $H_\ell - r_\ell = Q_\ell R_\ell$. Since each Q_ℓ is unitary and R_ℓ has positive diagonal entries, uniqueness of the QR decomposition gives $Q = Q_1 \cdots Q_k$ and $R = R_k \cdots R_1$ as desired. The composition property (4.14) is then immediate. □

The following corollary will be repeatedly useful.

Lemma 4.2.6. *Under the hypotheses of Lemma 4.2.5,*

$$\|e_n^* p(H)^{-1}\|^{-1} = R_{nn} = (R_1)_{nn} \cdots (R_k)_{nn} \tag{4.16}$$

Proof. Maintaining the notation of Lemma 4.2.5, we have

$$\|e_n^* p(H)^{-1}\| = \|e_n^* R^{-1} Q^*\| = \|e_n^* R^{-1}\| = \frac{1}{R_{n,n}},$$

and the proof is concluded by observing that (4.15) implies $R_{n,n} = (R_1)_{n,n} \cdots (R_k)_{n,n}$. □

We will require the following definition of backward stability for a degree 1 implicit QR step. The difference between this and the backward stability condition considered in [155] is the additional second equation below.

Definition 4.2.7 (Backward-Stable Degree 1 Implicit QR Algorithm)**.** A $\nu_{\mathsf{IQR}}(n)$-stable single-shift implicit QR algorithm takes as inputs a Hessenberg matrix $H \in \mathbb{C}^{n\times n}$ and a shift $s \in \mathbb{C}$ and outputs a Hessenberg matrix $\widetilde{H}$ and an exactly triangular matrix $\widetilde{R}$, for which there exists a unitary $\widetilde{Q}$ satisfying

$$\left\| \widetilde{H} - \widetilde{Q}^* H \widetilde{Q} \right\| \leq \|H - s\| \nu_{\mathsf{IQR}}(n) \mathbf{u} \tag{4.17}$$

$$\left\| H - s - \widetilde{Q}\widetilde{R} \right\| \leq \|H - s\| \nu_{\mathsf{IQR}}(n) \mathbf{u} \tag{4.18}$$

We now verify that there is a suitable backward-stable implicit QR algorithm. The pseucodode of IQR given below is a standard implementation based on Givens rotations. We use sans serif fonts to indicate subroutines implemented in finite arithmetic.

Lemma 4.2.8 (Backward Stability of Degree 1 IQR)**.** *Assuming*

$$\mathbf{u} \leq \min\left\{ \frac{1}{24}, \frac{\log 2}{8n^{5/2}} \right\} = 2^{-O(\log n)}, \tag{4.19}$$

IQR *satisfies its guarantees and uses at most* $7n^2$ *arithmetic operations. In particular, it is a* $\nu_{\mathsf{IQR}}(n)$*-stable implicit QR algorithm for* $\nu_{\mathsf{IQR}}(n) = 32n^{3/2}$.

The straightforward proof is deferred to Appendix C.1].

We now extend the definition of IQR to shifts of higher degree. We take the straightforward approach of composing many degree 1 QR steps to obtain a higher degree one. Given a Hessenberg matrix H, an implicit QR algorithm IQR satisfying Definition 4.2.7, and shifts $s_1, \ldots, s_k$, we will define

$$\mathsf{IQR}(H, \{s_1, \ldots, s_k\}) := \mathsf{IQR}(\mathsf{IQR}(\cdots \mathsf{IQR}(\mathsf{IQR}(H, s_1), s_2), \cdots), s_k), \tag{4.20}$$

which can be executed in $T_{\mathsf{IQR}}(n, k) = 7kn^2$ arithmetic operations. We will sometimes use the notation

$$\mathsf{IQR}(H, p(z)) = \mathsf{IQR}(H, \{s_1, \ldots, s_k\})$$

IQR

Input: Hessenberg H, shift $s \in \mathbb{C}$
Output: Hessenberg $\widetilde{H}$ and triangular $\widetilde{R}$
Ensures: $\|\widetilde{H}\| \leq \|H\| + 32n^{3/2}\mathbf{u} \cdot \|H - s\|$, and there exists unitary $\widetilde{Q}$ for which $\|\widetilde{H} - \widetilde{Q}^* H \widetilde{Q}\| \leq 32n^{3/2}\mathbf{u} \cdot \|H - s\|$ and $\|H - s - \widetilde{Q}\widetilde{R}\| \leq 16n^{3/2}\mathbf{u} \cdot \|H - s\|$

1. $\widetilde{R} \leftarrow H - s$
2. **For** $i = 1, 2, ..., n-1$
   a) $X_{1:2,i} \leftarrow \widetilde{R}_{i:i+1,i}$
   b) $\widetilde{R}_{i:i+1,i+1,n} \leftarrow \mathsf{giv}(X_{1:2,i})^* \widetilde{R}_{i:i+1,i+1:n} + E_{2,i,b}$
   c) $\widetilde{R}_{i:i+1,i} \leftarrow \begin{pmatrix} \|X_{1:2,i}\| + E_{2,i,c} \\ 0 \end{pmatrix}$
3. $\widetilde{H} \leftarrow \widetilde{R}$
4. **For** $i = 1, 2, ...n-1$
   a) $\widetilde{H}_{1:n,i:i+1} \leftarrow \widetilde{H}_{1:n,i:i+1}\mathsf{giv}(X_{1:2,i}) + E_{4,i}$
5. $\widetilde{H} \leftarrow \widetilde{H} + s$

where $p(z) = (z - s_1)\dots(z - s_k)$, though it is understood that IQR takes the roots of p and not its coefficients as input. Lemma 4.2.8 is readily adapted to give backward stability guarantees for $\mathsf{IQR}(H, p(z))$.

Lemma 4.2.9 (Backward Error Guarantees for Higher Degree IQR)**.** *Fix $C > 0$ and let $p(z) = \prod_{\ell \in [k]}(z - s_\ell)$, where $\mathcal{S} = \{s_1, ..., s_k\} \subset D(0, C\|H\|)$. Write $\left[\widetilde{H}, \widetilde{R}_1, ..., \widetilde{R}_k\right] = \mathsf{IQR}(H, p(z))$, and let $\widetilde{Q}_\ell$ be the unitary guaranteed by Definition 4.2.7 to the ℓth internal call to IQR. Assuming*

$$\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq 1/4,$$

the outputs $\widetilde{R} = \widetilde{R}_k \cdots \widetilde{R}_1$ and $\widetilde{Q} = \widetilde{Q}_1 \cdots \widetilde{Q}_k$ satisfy

$$\left\|\widetilde{H} - \widetilde{Q}^* H \widetilde{Q}\right\| \leq 1.4k(1 + C)\|H\|\nu_{\mathsf{IQR}}(n)\mathbf{u} \tag{4.21}$$

$$\left\|p(H) - \widetilde{Q}\widetilde{R}\right\| \leq 4\Big(2(1 + C)\|H\|\Big)^k \nu_{\mathsf{IQR}}(n)\mathbf{u}. \tag{4.22}$$

The straightforward proof is deferred to Appendix C.1. □

Forward Stability of Higher Degree IQR

In this subsection we prove forward error guarantees for $\mathsf{IQR}(H, p(z))$ using the backward error guarantees of the previous section. Let us first recall the following bound on the condition number of the QR decomposition [146, Theorem 1.6].

Lemma 4.2.10 (Condition Number of the QR Decomposition)**.** *Let $A, E \in \mathbb{C}^{n\times n}$ with A invertible. Furthermore assume that $\|E\|\|A^{-1}\| \le \frac{1}{2}$. If $[Q, R] = \mathrm{qr}(A)$ and $[\widetilde{Q}, \widetilde{R}] = \mathrm{qr}(A+E)$, then*

$$\|\widetilde{Q} - Q\|_F \le 4\|A^{-1}\|\|E\|_F \quad \text{and} \quad \|\widetilde{R} - R\| \le 3\|A^{-1}\|\|R\|\|E\|.$$

The main result of this subsection, which will be used throughout, is the following.

Lemma 4.2.11 (Forward Error Guarantees for IQR)**.** *Under the hypotheses of Lemma 4.2.9, and assuming further that $[Q, R] = \mathrm{qr}(p(H))$, $H = Q^*HQ$, and*

$$\mathbf{u} \le \mathbf{u}_{\mathsf{IQR}}(n, k, \|H\|, \kappa_V(H), \mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))) := \frac{1}{8\kappa_V(H)\nu_{\mathsf{IQR}}(n)} \left(\frac{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}{\|H\|}\right)^k = 2^{-O\left(\log n\kappa_V(H) + k\log\frac{\|H\|}{\mathsf{dist}(\mathcal{S},\mathrm{Spec}(H))}\right)}, \tag{4.23}$$

we have the forward error guarantees:

$$\|\widetilde{Q} - Q\|_F \le 16\kappa_V(H)\left(\frac{(2+2C)\|H\|}{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}\right)^k n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u} \tag{4.24}$$

$$\|\widetilde{R} - R\| \le 12\kappa_V(H)\left(\frac{(2+2C)^2\|H\|^2}{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}\right)^k \nu_{\mathsf{IQR}}(n)\mathbf{u} \tag{4.25}$$

$$\left\|\widetilde{H} - H\right\|_F \le 32\kappa_V(H)\|H\|\left(\frac{(2+2C)\|H\|}{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}\right)^k n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u}. \tag{4.26}$$

Proof. The first two assertions are immediate from applying Lemma B.2.8 to $M = p(H)$, computing

$$\|M^{-1}\| = \|p(H)^{-1}\| \le \frac{\kappa_V(H)}{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))^k},$$

bounding $\|p(H)\| \le (2+2C)^k\|H\|^k$, and finally using Lemma 4.2.9 to control $\|E\| \le 2(2+2C)^k\|H\|^k\nu_{\mathsf{IQR}}(n)\mathbf{u}$. For the third, observe that

$$\|\widetilde{Q}^*H\widetilde{Q} - Q^*HQ\|_F \le \|\widetilde{Q}^*H(\widetilde{Q} - Q)\|_F + \|(\widetilde{Q}^* - Q^*)HQ\|_F \le 2\|H\|\|\widetilde{Q} - Q\|_F,$$

and use the first assertion again. □

We close the subsection by giving forward error bounds for computing

$$\tau_p(H)^k = \|e_n^* p(H)^{-1}\|^{-1}$$

indirectly, from the R's output by $\mathsf{IQR}(H, p(z))$, for p a polynomial of degree k.

Tau^k

Input: Hessenberg $H \in \mathbb{C}^{n\times n}$, polynomial $p(z) = (z - s_1)\cdots(z - s_k)$
Output: $\widetilde{\tau^k} \geq 0$
Ensures: $|\widetilde{\tau^k} - \tau_p(H)^k| \leq 0.001\tau_p(H)^k$

1. $[\widetilde{\hat{H}}, \widetilde{R}_1, \ldots, \widetilde{R}_k] \leftarrow \mathsf{IQR}(H, p(z))$
2. $\widetilde{\tau^k} \leftarrow \mathsf{fl}\left((\widetilde{R}_1)_{nn}\cdots(\widetilde{R}_k)_{nn}\right)$

Lemma 4.2.12 (Guarantees for Tau^k). *If $\mathcal{S} = \{s_1, ..., s_k\} \subset D(0, C\|H\|)$ and*

$$\begin{aligned}
\mathbf{u} &\leq \mathbf{u}_{\mathsf{Tau}}(n, k, C, \|H\|, \kappa_V(H), \mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))) && (4.27)\\
&:= \frac{1}{6\cdot 10^3 \kappa_V(H)\nu_{\mathsf{IQR}}(n)}\left(\frac{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}{(2+2C)\|H\|}\right)^{2k} && (4.28)\\
&= 2^{-O\left(\log n\kappa_V(H) + k\log\frac{\|H\|}{\mathsf{dist}(\mathcal{S},\mathrm{Spec}(H))}\right)},
\end{aligned}$$

then Tau^k satisfies its guarantees, and runs in

$$T_{\mathsf{Tau}}(n, k) := T_{\mathsf{IQR}}(n, k) + k = O(kn^2)$$

arithmetic operations.

Proof. Let $[Q, R] = \mathrm{qr}(p(H))$ and recall that (4.16) shows that $\tau_p(H)^k = R_{nn}$. As (4.27) implies $\mathbf{u}_{\mathsf{IQR}}(n, k, \|H\|, \kappa_V(H), \mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H)))$, we can apply Lemma 4.2.11: the matrix $\widetilde{R} = \widetilde{R}_k \cdots \widetilde{R}_1$ satisfies

$$\begin{aligned}
&|\widetilde{R}_{n,n} - R_{n,n}|\\
\leq&\|\widetilde{R} - R\|\\
\leq&12\kappa_V(H)\left(\frac{(2+2C)^2\|H\|^2}{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}\right)^k \nu_{\mathsf{IQR}}(n)\mathbf{u} && \text{Lemma 4.2.11}\\
\leq&\frac{0.0005}{\|p(H)^{-1}\|} && (4.27),\ \|p(H)^{-1}\| \leq \kappa_V(H)\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))^{-k}\\
\leq&0.0005\,\sigma_{\min}(R) && p(H) = QR
\end{aligned}$$

$\leq 0.0005\, R_{n,n}.$ $\qquad\qquad$ $\sigma_{\min}(R) \leq \|e_n^* R\| = R_{n,n}$

Now, because $\widetilde{\tau^k}$ is the result of computing the product of the $(\widetilde{R}_i)_{n,n}$ in floating point arithmetic, we have $\left|\widetilde{\tau^k} - \widetilde{R}_{n,n}\right| \leq k\mathbf{u}\widetilde{R}_{n,n}$, whence

$$\begin{aligned}\left|\widetilde{\tau^k} - R_{n,n}\right| &\leq \left|\widetilde{\tau^k} - \widetilde{R}_{n,n}\right| + \left|\widetilde{R}_{n,n} - R_{n,n}\right| \\ &\leq k\mathbf{u}\widetilde{R}_{n,n} + 0.0005\, R_{n,n} \\ &\leq (1.0005k\mathbf{u} + 0.0005)R_{n,n} \\ &\leq 0.001 R_{n,n}.\end{aligned}$$

It will also be useful to observe that

$$\left|\frac{1}{\widetilde{\tau^k}} - \frac{1}{R_{n,n}}\right| \leq \frac{0.001}{|\widetilde{\tau^k}|} \leq \frac{0.001}{|\widetilde{\tau^k}|} \leq \frac{0.001}{\left|R_{n,n} - |\widetilde{\tau^k} - R_{n,n}|\right|} \leq \frac{0.001}{0.99 R_{n,n}} \leq \frac{0.0011}{R_{n,n}}.$$

□

Shift Regularization

The forward error bounds on our shifts are controlled by the inverse of the distance to $\mathrm{Spec}(H)$; to ensure that this is not too large, we *regularize* the shifts $r_1, \ldots, r_k$ by randomly perturbing them.

Lemma 4.2.13 (Regularization of shifts)**.** *Let $\mathcal{R} = \{r_1, ..., r_k\} \subset \mathbb{C}$ and $\eta_2 \geq \eta_1 > 0$. Assume*

$$\eta_1 + \eta_2 \leq \frac{\mathrm{gap}(H)}{2}.$$

Let $w_1, ..., w_k \sim$ Unif$(D(0, \eta_2))$ be i.i.d. and $\check{\mathcal{R}} = \{\check{r}_1, ..., \check{r}_k\} = \{r_1 + w_1, ..., r_k + w_k\}$. Then with probability at least $1 - k\,(\eta_1/\eta_2)^2$, we have $\mathsf{dist}(\check{\mathcal{R}}, \mathrm{Spec}(H)) \geq \eta_1$.

Proof. Define the bad region $\mathcal{B} \subset \mathbb{C}$ as the union of disks $\mathcal{B} := \bigcup_{\lambda \in \mathrm{Spec}(H)} D(\lambda, \eta_1)$. The assumption $\eta_1 + \eta_2 \leq \mathrm{gap}(H)/2$ implies that for each r_i, the disk $D(r_i, \eta_2)$ intersects at most one disk in $\mathcal{B}$; since $\check{r}_i$ is distributed uniformly in $D(r_i, \eta_2)$ we have

$$\mathbb{P}[\check{r}_i \in \mathcal{B}] \leq \left(\frac{\eta_1}{\eta_2}\right)^2,$$

and the total probability that at least one $\check{r}_i$ lies in the bad region is at most k times this by a union bound.

□

4.2.5 Finding Forward Stable Optimal Ritz Values (or Decoupling Early)

The shifting strategy $\mathsf{Sh}_{k,B}$ in Section 4.1 uses a specific notion of approximation for Ritz values, namely θ-*optimality* as defined in (4.11). In Section 4.1 we assumed the existence of a black box algorithm for computing such optimal values. In this section we will show how to compute θ-optimal Ritz values which are forward stable in the sense of Section 4.2.4 (or guarantee immediate decoupling).

The procedure consists of two steps, and relies on the black box algorithm SmallEig for computing forward approximations of the eigenvalues of a $k \times k$ or smaller matrix, in the sense of Definition 1.4.5. The first step of our approximation procedure is simply to compute forward approximations to the Ritz values using SmallEig. Second, we show the following dichotomy: for appropriately set parameters, any forward-approximate set of Ritz values $\mathcal{R}$ of a Hessenberg matrix H is either (i) θ-optimal *or* (ii) contains a Ritz value which can be used to decouple the matrix in a single degree k implicit QR step (in fact, the proof shows that this Ritz value must be close to an eigenvalue of H, see Remark 4.2.16). This is the content of Theorem 4.2.14, which is established in Section 4.2.5. We give a finite arithmetic implementation of this dichotomy in Section 4.2.5.

The Dichotomy in Exact Arithmetic

In this subsection we show that for β small enough and θ large enough, any set $\mathcal{R} = \{r_1, \dots, r_k\}$ of β-forward approximate Ritz values of H either yields a θ-optimal set of Ritz values, or one of the $r_i \in \mathcal{R}$ has a small value of $\tau_{(z-r_i)^k}(H)$.

Theorem 4.2.14 (Dichotomy)**.** *Let* $\mathrm{P} = \{\rho_1, \dots, \rho_k\}$ *be the Ritz values of* H *and assume that* $\mathcal{R} = \{r_1, \dots, r_k\}$ *satisfies* $|\rho_i - r_i| \leq \beta$ *for all* $i \in [k]$*. If*

$$\theta \geq (2\kappa_V^4(H))^{1/2k} \quad \text{and} \quad \frac{\beta}{\operatorname{gap}(H)} \leq \frac{1}{2}\left(\frac{\theta}{(2\kappa_V^4(H))^{1/2k}} - 1\right) =: c \tag{4.29}$$

then at least one of the following is true:

i) $\mathcal{R}$ is a set of θ-optimal Ritz values of H.

ii) There is an $r_i \in \mathcal{R}$ for which

$$\|e_n^*(H - r_i)^{-k}\|^{1/k} \geq \frac{1}{2\kappa_V(H)^{2/k}} \cdot \left(\frac{\psi_k(H)}{\|H\| + \beta}\right) \cdot \left(\frac{1 - \frac{(2\kappa_V^4)^{1/2k}}{\theta}}{\beta}\right). \tag{4.30}$$

The remainder of this subsection is dedicated to the proof of Theorem 4.2.14. Let $\mathrm{P} = \{\rho_1, \dots, \rho_k\}$ and $\mathcal{R} = \{r_1, \dots, r_k\}$ be as in Lemma 4.2.14, and set $\chi(z) = (z - \rho_1) \cdots (z - \rho_k)$ and $p(z) = (z - r_1) \cdots (z - r_k)$. Of course, by construction $\chi(z)$ is the characteristic polynomial

of $H_{(k)}$. Our strategy in proving Theorem 4.2.14 will be to show that if i) does not hold, then ii) does; assuming the former, we can get that

$$\begin{aligned}
\mathbb{E}[|p(Z_H)|^2] &\geq \frac{\|e_n^* p(H)\|^2}{\kappa_V(H)^2} && \text{Lemma 4.1.5} \\
&\geq \frac{\theta^{2k}\|e_n^* \chi(H)\|^2}{\kappa_V(H)^2} && \text{Negation of i)} \\
&\geq \frac{\theta^{2k}\mathbb{E}[|\chi(Z_H)|^2]}{\kappa_V(H)^4} && \text{Lemma 4.1.5} \qquad (4.31) \\
&= 2(1+2c)^{2k}\mathbb{E}[|\chi(Z_H)|^2] && (4.29) \qquad (4.32)
\end{aligned}$$

In other words, $\mathbb{E}[|p(Z_H)|^2]$ is much larger than $\mathbb{E}[|\chi(Z_H)|^2]$. On the other hand, by the assumptions in Theorem 4.2.14, the roots of $p(z)$ and $\chi(z)$ are quite close. Intuitively, because Z_H is supported on the eigenvalues of H, these two phenomena can only occur simultaneously if some root of $p(z)$ is close to an eigenvalue of of H with significant mass under the distribution of Z_H. The following lemma, whose proof we will briefly defer, articulates this precisely. The lemma does not require any particular properties of p and χ other than that their roots are close, so we will phrase it in terms of two generic polynomials q and $\check{q}$; when we apply the lemma, we will set $q = \chi$ and $\check{q} = p$.

Lemma 4.2.15. *Assume that* $\frac{\beta}{c} \leq \mathrm{gap}(H)$ *with* c *defined as in* (4.29), $q(z) := (z - s_1)\cdots(z - s_k)$ *for some* $\mathcal{S} = \{s_1, \ldots, s_k\} \subset D(0, \|H\|)$, *and let* $\check{q}(z) := (z - \check{s}_1)\cdots(z - \check{s}_k)$ *with* $\check{s}_1, ..., \check{s}_k \in \mathbb{C}$ *satisfying*

$$\max_{i \in [k]} |s_i - \check{s}_i| \leq \beta.$$

Then

$$\mathbb{P}\left[\mathrm{dist}(Z_H, \{s_1, ..., s_k\}) \leq \frac{\beta}{2c}\right] \geq \frac{\mathbb{E}[|\check{q}(Z_H)|^2] - (1+2c)^{2k}\mathbb{E}[|q(Z_H)|^2]}{(2(\|H\| + \beta)(1+2c))^{2k}}.$$

Lemma in hand, we can now complete the proof.

Proof of Theorem 4.2.14. Using Lemma 4.2.15 with $q(z) = \chi(z) = (z - \rho_1)\ldots(z - \rho_k)$ and $\check{q}(z) = p(z) = (z - r_1)\ldots(z - r_k)$, we find that

$$\begin{aligned}
\mathbb{P}\left[\mathrm{dist}(Z_H, \mathrm{P}) \leq \frac{\beta}{2c}\right] &\geq \frac{\mathbb{E}[|p(Z_H)|^2] - (1+2c)^{2k}\mathbb{E}[|\chi(Z_H)|^2]}{(2(\|H\| + \beta)(1+2c))^{2k}} \\
&\geq \frac{\mathbb{E}[|\chi(Z_H)|^2]}{2^{2k}(\|H\| + \beta)^{2k}} && (4.32) \\
&\geq \frac{\|e_n^* \chi(H)\|^2}{2^{2k}\kappa_V(H)^2(\|H\| + \beta)^{2k}} && \text{Lemma 4.1.5} \\
&= \frac{\psi_k^{2k}(H)}{2^{2k}\kappa_V(H)^2(\|H\| + \beta)^{2k}} && \text{Lemma 4.1.3.}
\end{aligned}$$

Since the right hand side is nonzero and Z_H is supported on the spectrum of H (and since $c \leq 1/2$ by assumption) this implies that for some $i \in [k]$ and $\lambda \in \mathrm{Spec}(H)$

$$|\rho_i - \lambda| \leq \frac{\beta}{2c}.$$

On the other hand, as we are assuming $\beta/c \leq \mathrm{gap}(H)$, there can be at most one eigenvalue within $\beta/2c$ of each ρ_i — otherwise by the triangle inequality two such eigenvalues would be at distance less that $\beta/c \leq \mathrm{gap}(H)$ from one another. Since there are only k of the ρ_i's, at least one of the eigenvalues, say λ, that is at least $\beta/2c$-close to of one of them must satisfy

$$\mathbb{P}\left[Z_H = \lambda\right] \geq \frac{1}{k}\left(\frac{\psi_k(H)}{2\kappa_V(H)^{1/k}(\|H\| + \beta)}\right)^{2k}. \tag{4.33}$$

By the triangle inequality, we then have

$$|r_i - \lambda| \leq |r_i - \rho_i| + |\rho_i - \lambda| \leq \beta\left(1 + \frac{1}{2c}\right). \tag{4.34}$$

Finally,

$$\begin{aligned}
&\|e_n^*(H - r_i)^{-k}\|^{1/k} \\
\geq &\frac{\mathbb{E}\left[|Z_H - r|^{-2k}\right]^{1/2k}}{\kappa_V(H)^{1/k}} && \text{Lemma 4.1.5} \\
\geq &\frac{1}{\kappa_V(H)^{1/k}} \cdot \frac{1}{(2k)^{1/2k}}\left(\frac{\psi_k(H)}{\kappa_V(H)^{1/k}(\|H\| + \beta)}\right) \cdot \left(\frac{2c}{(2c+1)\beta}\right),
\end{aligned}$$

where the second inequality uses $\mathbb{E}\left[|Z_H - r_i|^{-2k}\right] \geq \frac{\mathbb{P}[Z_H = \lambda]}{|\lambda - r|^{2k}}$ and (4.33), (4.34). This yields the conclusion by substituting c and noting that $(2k)^{1/2k} \leq 2$. □

Remark 4.2.16. By (4.33) and (4.34), the above proof shows that the culprit Ritz value r_i is close to an eigenvalue of H and the corresponding right eigenvector has a large inner product with e_n. This could alternatively be used to decouple the matrix using other techniques such as inverse iteration.

Proof of Lemma 4.2.15. We begin by partitioning set $\mathcal{S} = \{s_1, ..., s_k\}$ according to which eigenvalue of H is the closest: relabelling $\mathrm{Spec}(H) = \{\lambda_1, ..., \lambda_n\}$ as necessary, write $\mathcal{S} = S_1 \sqcup \cdots \sqcup S_\ell$, where S_j consists of those s_i whose closest eigenvalue is λ_j (breaking ties arbitrarily).

Now, recursively define a sequence of polynomials $q_0, \ldots, q_l$ with $l \leq k$ given by $q_0(z) = q(z)$ and

$$q_{j+1}(z) := \frac{\prod_{i \in S_{j+1}}(z - \check{s}_i)}{\prod_{i \in S_{j+1}}(z - s_i)} q_j(z);$$

in other words, the q_j interpolate between q and $\check{q}$ by exchanging the original roots $s_1, ..., s_k$ for the perturbed ones $\check{s}_1, ..., \check{s}_k$, doing so in batches according to the partition $\mathcal{S} = S_1 \sqcup \cdots \sqcup S_\ell$. The proof reduces to the following bound on $\mathbb{E}[|q_j(Z_H)|^2]$ in terms of $\mathbb{E}[|q_{j-1}(Z_H)|^2]$, which we will prove shortly.

Claim 4.2.17. For each $j = 1, ..., \ell$, we have

$$\mathbb{E}[|q_j(Z_H)|^2] \leq (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] + (2(\|H\|+\beta))^{2k}\mathbb{P}[Z_H = \lambda_j]\mathbf{1}\left[\operatorname{dist}(\lambda_j, \mathcal{S}) \leq \tfrac{\beta}{2c}\right].$$

In view of the claim, we can inductively assemble these bounds to compare $\mathbb{E}[|q(Z_H)|^2]$ and $\mathbb{E}[|\check{q}(Z_H)^2]$:

$$\begin{aligned}
\mathbb{E}[|\check{q}(Z_H)|^2] &= \mathbb{E}[|q_\ell(Z_H)|^2] \\
&\leq (1+2c)^{2|S_\ell|}\mathbb{E}[|q_{\ell-1}(Z_H)|^2] + (2(\|H\|+\beta))^{2k}\mathbb{P}[Z_H = \lambda_\ell]\mathbf{1}\left[\operatorname{dist}(\lambda_\ell, \mathcal{S}) \leq \tfrac{\beta}{2c}\right] \\
&\leq (1+2c)^{2k}\mathbb{E}[|q_0(Z_H)|^2] \\
&\quad + \sum_{i\in[\ell]}(2(\|H\|+\beta))^{2k}(1+2c)^{2\sum_{j=1}^{i}|S_i|}\mathbb{P}[Z_H = \lambda_i]\mathbf{1}\left[\operatorname{dist}(\lambda_i, \mathcal{S}) \leq \tfrac{\beta}{2c}\right] \\
&\leq (1+2c)^{2k}\left(\mathbb{E}[|q(Z_H)|^2] + (2(\|H\|+\beta))^{2k}\sum_{i\in[\ell]}\mathbb{P}[Z_H = \lambda_i]\mathbf{1}\left[\operatorname{dist}(\lambda_i, \mathcal{S}) \leq \tfrac{\beta}{2c}\right]\right) \\
&\leq (1+2c)^{2k}\left(\mathbb{E}[|q(Z_H)|^2] + (2(\|H\|+\beta))^{2k}\mathbb{P}\left[\operatorname{dist}(Z_H, \mathcal{S}) \leq \tfrac{\beta}{2c}\right]\right).
\end{aligned}$$

Rearranging gives the bound advertised in the lemma.

It remains to prove Claim 4.2.17. To lighten notation, we'll write s and $\check{s}$ for an arbitrary element in $S_j \subset \mathcal{S}$, and its perturbation, respectively. For any $m \in [n] \setminus j$ and $s \in S_j$, we have $|\lambda_m - s| \geq \frac{\operatorname{gap}(H)}{2}$, so

$$\left|\frac{\lambda_m - \check{s}}{\lambda_m - s}\right| \leq 1 + \left|\frac{s - \check{s}}{\lambda_m - s}\right| \leq 1 + \frac{2|s-\check{s}|}{\operatorname{gap}(H)} \leq 1 + 2c,$$

and hence

$$\prod_{s\in S_j}\left|\frac{\lambda_m - \check{s}}{\lambda_m - s}\right| \leq (1+2c)^{|S_j|}.$$

Using the above, the definition of q_j in terms of q_{j-1}, and expanding the expectation as a sum, we find

$$\begin{aligned}
\mathbb{E}[|q_j(Z_H)|^2] &= \mathbb{P}[Z_H = \lambda_j]|q_j(\lambda_j)|^2 + \sum_{m\in[n]\setminus j}\mathbb{P}[Z_H = \lambda_m]|q_{j-1}(\lambda_m)|^2\prod_{s\in S_{j+1}}\left|\frac{\lambda_m - \check{s}}{\lambda_m - s}\right|^2 \\
&\leq \mathbb{P}[Z_H = \lambda_j]|q_j(\lambda_j)|^2 + (1+2c)^{2|S_j|}\sum_{m\in[n]\setminus j}\mathbb{P}[Z_H = \lambda_m]|q_{j-1}(\lambda_m)|^2
\end{aligned}$$

$$\leq \mathbb{P}[Z_H = \lambda_j]\left(|q_j(\lambda_j)|^2 - (1+2c)^{2|S_j|}|q_j(\lambda_{j-1})|^2\right) + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] \tag{4.35}$$

$$\leq \mathbb{P}[Z_H = \lambda_j]|q_{j-1}(\lambda_j)|^2\left(\prod_{s\in S_j}\left(1+\left|\frac{s-\check{s}}{\lambda_j - s}\right|\right)^2 - (1+2c)^{2|S_j|}\right) + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] \tag{4.36}$$

We have defined S_j so that λ_j is the closest eigenvalue to every $s \in S_j$, so $\mathsf{dist}(\lambda_j, \mathcal{S}) = \mathsf{dist}(\lambda_j, S_j)$. Thus when $\mathsf{dist}(\lambda_j, \mathcal{S}) > \frac{\beta}{2c}$, we can rearrange to see that

$$\begin{aligned} 0 &\geq \left(1 + \frac{\beta}{\mathsf{dist}(\lambda_j, S_j)}\right)^{2|S_j|} - (1+2c)^{2|S_j|} \\ &\geq \prod_{s\in S_j}\left(1+\frac{|s-\check{s}|}{|\lambda_j - s|}\right)^2 - (1+2c)^{2|S_j|}; \end{aligned}$$

the latter is a factor of the first term on the right hand side of (4.36), so in the event $\mathsf{dist}(\lambda_j, \mathcal{S}) > \frac{\beta}{2c}$ we have

$$\mathbb{E}[|q_j(Z_H)|^2] \leq (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2].$$

On the other hand, independent of $\mathsf{dist}(\lambda_j, \mathcal{S})$ (and thus in particular when $\mathsf{dist}(\lambda_j, \mathcal{S}) \leq \frac{\beta}{2c}$) from (4.35) we know that

$$\begin{aligned} \mathbb{E}[|q_j(Z_H)|^2] &\leq \mathbb{P}[Z_H = \lambda_j]|q_j(\lambda_j)|^2 + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] \\ &\leq \mathbb{P}[Z_H = \lambda_j](2(\|H\| + \beta))^{2k} + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2. \end{aligned}$$

For the final inequality, note that $\lambda_j \in D(0, \|H\|)$ and, because $\mathcal{S} \subset D(0, \|H\|)$, and $|\check{s} - s| \leq \beta$ for every $s \in \mathcal{S}$, the roots of each q_j are contained in $D(0, \|H\| + \beta)$. Combining the bounds on $\mathbb{E}[|q_j(Z_H)|^2]$ in the cases $\mathsf{dist}(\lambda_j, \mathcal{S}) > \frac{\beta}{2c}$ and $\mathsf{dist}(\lambda_j, \mathcal{S}) \leq \frac{\beta}{2c}$, we find that

$$\mathbb{E}[|q_j(Z_H)|^2] \leq (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] + (2(\|H\| + \beta))^{2k}\mathbb{P}[Z_H = \lambda_j]\mathbf{1}\left[\mathrm{dist}(\lambda_j, \mathcal{S}) \leq \tfrac{\beta}{2c}\right],$$

establishing the claim. □

Finite Arithmetic Implementation of RitzOrDecouple

In this subsection we combine Theorem 4.2.14 and the regularization procedure of Lemma 4.2.13 to obtain a finite arithmetic algorithm RitzOrDecouple for finding θ-optimal Ritz values in the sense of Definition 1.4.2, for θ set as in (4.13), and with the additional property of being forward stable. The first step is testing whether a set of putative approximate Ritz values are θ-optimal.

Optimal

Input: Hessenberg $H \in \mathbb{C}^{n\times n}$, $\{s_1, \ldots, s_k\} = \mathcal{S} \subset \mathbb{C}$
Global Data: Optimality parameter θ
Output: Optimality flag `opt`
Ensures: If `opt` = `true`, then $\mathcal{S}$ are θ-optimal; if `opt` = `false`, then they are not $(.998^{1/k}\theta)$-optimal

1. $\widetilde{v_0} \leftarrow e_n$
2. **For** $j = 0, \ldots, k-1$
   a) $\widetilde{v_{j+1}} \leftarrow \mathsf{fl}\left((H - s_{j+1})^* \widetilde{v_j}\right)$
3. **If** $\mathsf{fl}(\|\widetilde{v_k}\|) \geq .999\theta^k \psi_k^k(H)$, `opt` $\leftarrow$ `false`, **else** `opt` $\leftarrow$ `true`

Lemma 4.2.18 (Guarantees for Optimal)**.** *Assume that* $s_1, \ldots, s_k \in D(0, C\|H\|)$ *and*

$$\mathbf{u} \leq \mathbf{u}_{\mathsf{Optimal}}(n, k, C, \|H\|, \theta) := \frac{1}{2 \cdot 10^3 n^2} \left(\frac{\psi_k(H)}{\theta(2+2C)\|H\|}\right)^k = 2^{-O\left(\log n + k \log \frac{\theta\|H\|}{\psi_k(H)}\right)}; \quad (4.37)$$

then Optimal *satisfies its guarantees and runs in at most* $T_{\mathsf{Optimal}}(k) := 4k^2 = O(k^2)$ *arithmetic operations.*

Proof of Lemma 4.2.18. From our initial floating point assumptions, we have $\widetilde{v_i} = (H - s_i)\widetilde{v_{i-1}} + \Delta_i$, where Δ is supported only on its $i+1$ final coordinates, each of which has magnitude at most $(1+C)\|H\|\|\widetilde{v_{i-1}}\| \cdot n\mathbf{u}$, giving the crude bound $\|\Delta_i\| \leq (1+C)\|H\|\|\widetilde{v_{i-1}}\| \cdot n^{3/2}\mathbf{u}$. Thus inductively

$$\|\widetilde{v_i}\| \leq \left((1+C)\|H\|(1+n^{3/2}\mathbf{u})\right)^i$$

and given $\mathbf{u} \leq n^{-3/2}$,

$$\begin{aligned}
|\mathsf{fl}\left(\|\widetilde{v_k}\|\right) - \|e_n^* p(H)\|| &\leq n\mathbf{u}\|\widetilde{v_k}\| + |\|\widetilde{v_k}\| - \|e_n^* p(H)\|| \\
&\leq n\mathbf{u}\left((1+C)\|H\|(1+n^{3/2}\mathbf{u})\right)^k + kn^{3/2}\mathbf{u} \cdot \left((1+C)\|H\|(1+n^{3/2}\mathbf{u})\right)^k \\
&\leq 2n^2(2+2C)^k\|H\|^k\mathbf{u}.
\end{aligned}$$

Thus if $\mathsf{fl}(\|\widetilde{v_k}\|) \geq .999\theta^k\psi_k^k(H)$, our assumption on $\mathbf{u}$ ensures

$$\|e_n^* p(H)\| \geq .999\theta^k\psi_k^k(H) - 2(1+C)^k\|H\|^k k^2 n^{3/2}\mathbf{u} \geq .998\theta^k\psi_k^k(H).$$

On the other hand, if $\mathsf{fl}(\|\widetilde{v_k}\|) \leq .999\theta^k\psi_k^k(H)$, then analogously we have

$$\|e_n^* p(H)\| \leq \theta^k\psi_k^k(H).$$

For the running time, each $\widetilde{v_i}$ is supported only on $i+2$ coordinates, so each multiplication $(H - s_i)\widetilde{v_{i-1}}$ requires $3i+3$ arithmetic operations, for a total of $3k(k+1)/2$; we then require a further $2k$ to compute $\|\widetilde{v_k}\|$, giving $3k(k+1)/2 + 2k \leq 4k^2$ arithmetic operations overall. □

We now specify RitzOrDecouple in full.

RitzOrDecouple

Input: Hessenberg H, working accuracy ϵ, failure probability ϕ
Global Data: Norm bound Σ, optimality parameter θ as in (4.13)
Requires: H is ϵ-unreduced, $\|H\| \le \Sigma$, $\mathrm{gap}(H) \ge \frac{2\epsilon^2}{\Sigma}$, $k/\phi \ge 2$
Output: Hessenberg H, θ-approximate Ritz values $\check{\mathcal{R}}$, decoupling flag `dec`
Ensures: With probability at least $1-\phi$, $\mathsf{dist}(\check{\mathcal{R}}, \mathrm{Spec}(H)) \ge \eta_1$ (as defined in line 1) *and* one of the following holds:

- `dec` $=$ `false`, $H = H$, and $\check{\mathcal{R}}$ is an exact set of θ-optimal Ritz values of H, satisfying $\check{\mathcal{R}} \subset D(0, 1.1\|H\|)$.
- `dec` $=$ `true` and for some $\check{r} \in \check{\mathcal{R}}$, $H = \mathsf{IQR}(H, (z-\check{r})^k)$ is ϵ-decoupled.

1. $\beta \leftarrow \frac{\epsilon^2}{16\cdot 101\cdot \Sigma}, \eta_2 \leftarrow \frac{\beta}{2}, \eta_1 \leftarrow \frac{\eta_2}{\sqrt{2k/\phi}} = \frac{\epsilon^2\sqrt{\phi}}{32\cdot 101\cdot \Sigma\sqrt{2k}}$
2. $\mathcal{R} \leftarrow \mathsf{SmallEig}\left(H_{(k)}, \beta/2, \phi/2\right)$
3. $\{\check{r}_1, \ldots, \check{r}_k\} = \check{\mathcal{R}} \leftarrow \{r_1 + w_1, \ldots, r_k + w_k\}$, where the w_i are i.i.d samples from $\mathrm{Unif}\big(D(0, \eta_2)\big)$
4. **If** $\mathsf{Optimal}(\check{\mathcal{R}}, H, \theta) =$ `true`, set $H \leftarrow H$ and `dec` $\leftarrow$ `false`
5. **Else if** $\mathsf{Optimal}(\check{\mathcal{R}}, H, \theta) =$ `false`, **for** $i = 1, ..., k$
   a) $H \leftarrow \mathsf{IQR}(H, (z-\check{r}_i)^k)$
   b) **If** $H_{j+1,j} \le \epsilon$ for any $j \in \{n-k, n-k+1, \ldots, n-1\}$, set `dec` $\leftarrow$ `true` and halt

Lemma 4.2.19 (Guarantees for RitzOrDecouple)**.** *Assuming that*

$$\begin{aligned}\mathbf{u} &\le \mathbf{u}_{\mathsf{RitzOrDecouple}}(n,k,\Sigma,B,\theta,\epsilon,\phi)\\ &:= \min\left\{\mathbf{u}_{\mathsf{Optimal}}(n,k,1.1,\Sigma,\theta), \frac{\epsilon}{8n^{1/2}\Sigma}\mathbf{u}_{\mathsf{IQR}}\left(n,k,1.1,\Sigma,B,\frac{\epsilon^2\sqrt{\phi}}{32\cdot 101\cdot\Sigma\sqrt{2k}}\right)\right\} \quad (4.38)\\ &= 2^{-O\left(\log nB + k\log\frac{\theta\|H\|\cdot k\Sigma}{\epsilon\phi}\right)} \quad (4.39)\end{aligned}$$

then RitzOrDecouple *satisfies its guarantees and its running time depends on the value of the decoupling flag. In either case it makes one call to* SmallEig*, in addition to that call*

1. *if* `dec` $=$ `false`*,* RitzOrDecouple *uses at most*

$$T_{\mathsf{RitzOrDecouple}}(n,k,\texttt{false}) := kC_{\mathsf{D}} + k + T_{\mathsf{Optimal}}(k) = O(k^2)$$

arithmetic operations.

2. otherwise, RitzOrDecouple *uses at most*

$$T_{\mathsf{RitzOrDecouple}}(n, k, \texttt{true}) := T_{\mathsf{Optimal}}(k) + k(T_{\mathsf{IQR}}(n, k) + k + C_{\mathsf{D}} + 1) = O(k^2 n^2)$$

arithmetic operations.

Proof. First, the assumptions of RitzOrDecouple on its input parameters imply that

$$\eta_1 + \eta_2 \leq \beta \leq \frac{\epsilon^2}{\Sigma} \leq \operatorname{gap}(H)/2$$

so we can apply Lemma 4.2.13 to find that $\mathsf{dist}(\check{\mathcal{R}}, \operatorname{Spec}(H)) \geq \eta_1$ with probability at least

$$1 - k\left(\frac{\eta_1}{\eta_2}\right)^2 = 1 - k\left(\sqrt{\frac{\phi}{2k}}\right)^2 \geq 1 - \phi/2.$$

By the black box assumptions on SmallEig, $\mathcal{R}$ is a set of $\beta/2$-forward approximate Ritz values with probability at least $1 - \phi/2$. The perturbed set $\check{\mathcal{R}}$ are in this case β-forward approximate Ritz values, and we further have

$$\beta \leq 0.1\epsilon \leq 0.1\|H\|$$

so the set $\check{\mathcal{R}}$ is contained in a disk of radius $1.1\|H\|$.

The assumption $\mathbf{u} \leq \mathbf{u}_{\mathsf{Optimal}}(1.1, k, n, H)$ means that if $\mathsf{Optimal}(\check{\mathcal{R}}, H, \theta) = \texttt{true}$ we are guaranteed that $\check{\mathcal{R}}$ is indeed a set of θ-optimal Ritz values for H. On the other hand if $\mathsf{Optimal}(\check{\mathcal{R}}, H, \theta) = \texttt{false}$, then by Lemma 4.2.18 the $\check{\mathcal{R}}$ fail to be $0.998^{1/k}\theta$-optimal. Examining the definitions of θ and β, we verify the hypotheses of Theorem 4.2.14:

$$c = \frac{1}{2}\left(\frac{0.998^{1/k}\theta}{(2\kappa_V(H)^4)^{1/2k}} - 1\right) \geq \frac{1}{2}\left(\frac{\frac{101}{100}(2B^4)^{1/2k}}{(2B^4)^{1/2k}} - 1\right) = \frac{1}{200} \geq \frac{\beta}{\operatorname{gap}(H)},$$

and conclude that there is some $\check{r} \in \check{\mathcal{R}}$ for which

$$\begin{aligned}
&\|e_n^*(H - \check{r})^{-k}\|^{1/k} \\
\geq &\frac{1}{2\kappa_V(H)^{2/k}} \cdot \left(\frac{\psi_k(H)}{\|H\| + \beta}\right) \cdot \left(\frac{1 - \frac{(2\kappa_V^4)^{1/2k}}{0.998^{1/k}\theta}}{\beta}\right) \\
\geq &\frac{1}{4} \cdot \left(\frac{\epsilon}{2\Sigma}\right) \cdot \left(\frac{1 - \frac{100}{101}}{\beta}\right) && B^{2/k} \leq 2,\ \psi_k(H) \leq \epsilon,\ \beta \leq \|H\| \\
\geq &\frac{2}{\epsilon}
\end{aligned}$$

by the definition of β in line 1. In the event that $\mathsf{dist}(\check{\mathcal{R}}, \mathrm{Spec}(H)) \geq \eta_1$, our choice of $\mathbf{u}$ in (4.38) means that we can apply Lemma 4.2.11 to $H = \mathsf{IQR}(H, (z-\check{r})^k)$ with $C = 1.1$, giving

$$\|H - \mathsf{iqr}(H, (z-\check{r})^k)\|_F \leq 32\kappa_V(H)\|H\| \left(\frac{4.2\|H\|}{\mathsf{dist}(\check{r}, \mathrm{Spec}(H))}\right)^k n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq \epsilon/2.$$

Using $\psi_k(\mathsf{iqr}(H, (z-\check{r})^k) \leq \tau_{(z-\check{r})^k}(H) \leq \epsilon/2$ (which was verified in Lemma 4.1.4) we find that $\mathsf{iqr}(H, (z-\check{r})^k)$ has a subdiagonal entry smaller than $\epsilon/2$, so H must have a subdiagonal entry smaller than ϵ, completing the proof of correctness.

To analyze the running time, note that when $\mathtt{dec} = \mathtt{false}$ other than the call to $\mathsf{SmallEig}$, in line 3 k samples are taken from $\mathrm{Unif}(D(0, \eta_2))$ and k additions are made which amounts to $C_{\mathsf{D}}k + k$ operations, and in line 4 $\mathsf{Optimal}$ is called once, adding $T_{\mathsf{Optimal}}(k)$ to the running time. In addition to that, when $\mathtt{dec} = \mathtt{true}$, at most k calls to IQR with degree k are made and each time k subdiagonals of H are checked, adding $kT_{\mathsf{IQR}}(n, k) + k^2$ operations. □

4.2.6 Finite Arithmetic Analysis of One Iteration of $\mathsf{Sh}_{k,B}$

In this section we provide the finite arithmetic analysis of a single iteration of the shifting strategy $\mathrm{Sh}_{k,B}$ introduced in Section 4.1; we assume familiarity with the context and notions introduced there. In exact arithmetic, $\mathrm{Sh}_{k,B}$ takes as input a Hessenberg matrix H with $\kappa_V(H) \leq B$, and a set $\mathcal{R}$ of θ-optimal Ritz values for H, and ouputs a new Hessenberg matrix $\hat{H}$ unitarily equivalent to H, with $\psi_k(\hat{H}) \leq (1-\gamma)\psi_k(H)$. Along the way, it first uses a subroutine Find to generate a promising Ritz value $r \in \mathcal{R}$ and then — in the event that the shift $(z-r)^k$ does not reduce the potential — uses a subroutine Exc to produce a set of exceptional shifts $\mathcal{S}$, one of which is guaranteed to achieve potential reduction. Let us now specify these subroutines in finite arithmetic and state their guarantees.

Computation of τ and ψ. The shifting strategy $\mathrm{Sh}_{k,B}$ needs access to both $\tau_p(H)$ and $\psi_k(H)$. The former can be computed using Lemma 4.2.12. For the latter, we will assume for simplicity that $\psi_k^k(H)$ (which is simply a product of k entries of H) can be computed *exactly* (this could for instance be achieved by temporary use of moderately increased precision). On the other hand, in some places it will be important to account for the error in computing the k-th root of $\psi_k^k(H)$, so we will denote

$$\widetilde{\psi_k}(H) := \mathsf{fl}\left(\left(\psi_k^k(H)\right)^{1/k}\right),$$

and assume

$$|\widetilde{\psi_k}(H) - \psi_k(H)| \leq (1 - 0.999^{1/k})\psi_k(H) \leq 0.001\psi_k(H), \tag{4.40}$$

which as per Lemma 4.2.2 can be computed in $T_\psi(k) := k + T_{\mathsf{root}}(k, 1-0.999^{1/k})$ arithmetic operations provided that

$$\mathbf{u} \leq \mathbf{u}_\psi(k) := \frac{1 - 0.999^{1/k}}{k(c_{\mathsf{root}} + 1 - 0.999^{1/k})} = 2^{-O(\log k)}. \tag{4.41}$$

This setting of the accuracy of $\widetilde{\psi_k}$ will be convenient for the analysis of Exc below.

Analysis of Find. To produce a promising Ritz value with Find, we will proceed as in the exact arithmetic case, using Tau^k to guide our binary search procedure. The guarantees on Tau^k are only strong enough to ensure that we discover a $(1.01\kappa_V(H))^{\frac{4\log k}{k}}$-promising Ritz value — as opposed the $\kappa_V(H)^{\frac{4\log k}{k}}$-optimality we are guaranteed in the exact case.

Find

Input: Hessenberg H, a set $\mathcal{R} = \{r_1, \ldots, r_k\} \subset \mathbb{C}$
Global Data: Promising parameter $\alpha = (1.01B)^{\frac{4\log k}{k}}$ as in (4.13)
Output: A complex number $r \in \mathcal{R}$
Requires: $\psi_k(H) > 0$
Ensures: r is α-promising

1. **For** $j = 1, ..., \log k$
   a) Evenly partition $\mathcal{R} = \mathcal{R}_0 \sqcup \mathcal{R}_1$, and **for** $b = 0, 1$ set $p_{j,b} = \prod_{r \in \mathcal{R}_b}(z - r)$
   b) $\mathcal{R} \leftarrow \mathcal{R}_{\widetilde{b_j}}$, where $\widetilde{b_j}$ is the b that minimizes $\mathsf{Tau}^{k/2}(H, p_{j,b}^{2^{j-1}})$
2. Output $\mathcal{R} = \{r\}$

Lemma 4.2.20 (Guarantees for Find)**.** *Assume that* $\mathcal{R} \subset D(0, C\|H\|)$ *and*

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{\mathsf{Find}}(n, k, C, \|H\|, \kappa_V(H), \mathsf{dist}(\mathcal{R}, \mathrm{Spec}(H))) \\ &:= \mathbf{u}_{\mathsf{Tau}}(n, k/2, C, \|H\|, \kappa_V(H), \mathsf{dist}(\mathcal{R}, \mathrm{Spec}(H))) \\ &= 2^{-O\left(\log n\kappa_V(H) + k\log \frac{\|H\|}{\mathsf{dist}(\mathcal{R}, \mathrm{Spec}(H))}\right)}. \end{aligned} \tag{4.42}$$

Then Find *satisfies its guarantees, and runs in*

$$T_{\mathsf{Find}}(n, k) := 2\log k T_{\mathsf{Tau}}(n, k/2) + \log k = O(k \log k \cdot n^2)$$

arithmetic operations.

Proof. The definition of $\mathbf{u}_{\mathsf{Find}}$ is sufficient to let us invoke Lemma 4.2.12 and conclude that it satisfies its guarantees throughout Find. On each step of the iteration, write b_j for the $b \in \{0, 1\}$ maximizing $\|e_n^* p_{j,b}(H)^{-1}\|$. Applying Lemma 4.2.12, for each $b \in \{0, 1\}$ we have

$$\left|\mathsf{Tau}^{k/2}(H, p_{j,b}) - \|e_n^* p_{j,b}(H)^{-1}\|^{-1}\right| \leq 0.0011\|e_n^* p_{j,b}(H)^{-1}\|^{-1},$$

and thus it always holds that

$$\|e_n^* p_{j,\widetilde{b_j}}(H)^{-1}\|^2 \geq (1 - 0.0022)^2\|p_{j,b_j}(H)^{-1}\|^2 \geq \frac{1}{2.02}\left(\|p_{j,0}(H)^{-1}\|^2 + \|p_{j,1}(H)^{-1}\|^2\right).$$

We now mirror the proof of the analogous Lemma 2.7 in Part 1 of this work, which analyzes Find in exact arithmetic. On each step of the iteration, we have defined thing so that

$$p_{j,\widetilde{b_j}}(z) = p_{j+1,0}(z)p_{j+1,1}(z). \tag{4.43}$$

On the first step of the subroutine, this identity becomes $p(z) = p_{1,0}(z)p_{1,1}(z)$, where $p(z)$ is the polynomial whose roots are the full set $\mathcal{R}$ of approximate Ritz values, so

$$\begin{aligned}
\|e_n^* p_{1,\widetilde{b_1}}(H)^{-1}\|^2 &\geq \frac{1}{2.02}\left(\|e_n^* p_{1,0}(H)^{-1}\|^2 + \|e_n^* p_{1,1}(H)^{-1}\|^2\right) \\
&\geq \frac{1}{1.01\kappa_V(H)^2}\mathbb{E}\left[\frac{1}{2}\left(|p_{1,0}(Z_H)|^{-2} + |p_{1,1}(Z_H)|^{-2}\right)\right] && \text{Lemma 4.1.5} \\
&\geq \frac{1}{1.01\kappa_V(H)^2}\mathbb{E}[|p(Z_H)|^{-1}] && \text{AM/GM and (4.43)}
\end{aligned}$$

Applying the same argument to each subsequent step,

$$\begin{aligned}
&\|e_n^* p_{j+1,\widetilde{b_{j+1}}}(H)^{-2^j}\|^2 \\
\geq &\frac{1}{1.01\kappa_V(H)^2}\mathbb{E}\left[\frac{1}{2}\left(|p_{j+1,0}(Z_H)|^{-2^{j+1}} + |p_{j+1,1}(Z_H)|^{-2^{j+1}}\right)\right] && \text{Lemma 4.1.5} \\
\geq &\frac{1}{1.01\kappa_V(H)^2}\mathbb{E}\left[|p_{j+1,0}(Z_H)p_{j+1,1}(Z_H)|^{-2^j}\right] && \text{AM/GM} \\
\geq &\frac{1}{1.01\kappa_V(H)^4}\|e_n^*(p_{j+1,0}(H)p_{j+1,1}(H))^{-2^{j-1}}\| && \text{Lemma 4.1.5} \\
= &\frac{1}{1.01\kappa_V(H)^4}\|e_n^* p_{j,\widetilde{b_j}}(H)^{-2^{j-1}}\|. && (4.43)
\end{aligned}$$

Paying a further $\kappa_V(H)^2$ on the final step to convert the norm into an expectation, we get

$$\mathbb{E}\left[|Z_H - r|^{-k}\right] \geq \left(\frac{1}{1.01\kappa_V(H)}\right)^{4\log k}\mathbb{E}\left[|p(Z_H)|^{-1}\right]$$

as promised.

For the runtime, we make $2\log k$ calls to $\mathsf{Tau}^{k/2}$ and $\log k$ comparisons of two floating point numbers. □

Analysis of Exc. We now come to the exceptional shift, effectuated by the subroutine Exc in the event that a promising Ritz value fails to achieve potential reduction. In finite arithmetic, we will again proceed similarly to the exact arithmetic setting — however, we will additionally need to ensure that all of our exceptional shifts are forward stable in the sense of Section 4.2.4, and to achieve this we will apply a random perturbation in the same spirit as Section 4.2.4.

Let us first pause to prove a key lemma ensuring potential reduction in finite arithmetic for sufficiently forward stable shifts. In particular, we will use the forward error guarantee of Lemma 4.2.11 to analyze the potential of $\mathsf{IQR}(H, p(z))$, by directly comparing it to that of $\mathsf{iqr}(H, p(z))$.

Lemma 4.2.21. *Let $p(z) = (z - s_1)...(z - s_m)$ for some floating point complex numbers $\mathcal{S} = \{s_1, ..., s_m\} \subset D(0, C\|H\|)$, and assume that for some $\epsilon > 0$,*

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{4.2.21}(n, k, C, \|H\|, \kappa_V(H), \mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H)), \epsilon) \\ &:= 0.001\epsilon \cdot \frac{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))^k}{32\kappa_V(H)\|H\|^{k+1}(2+2C)^k n^{1/2}\nu_{\mathsf{IQR}}(n)} \\ &= 2^{-O\left(\log \frac{n\kappa_V(H)}{\epsilon} + k \log \frac{\|H\|}{\mathsf{dist}(\mathcal{S}, \mathrm{Spec}(H))}\right)}. \end{aligned} \tag{4.44}$$

Then at least one of the following holds:

1. *(ϵ-Decoupling) Some subdiagonal of $\mathsf{IQR}(H, p(z))$ is smaller than ϵ.*
2. *(Potential Approximation) $\psi_k(\mathsf{IQR}(H, p(z))) \leq 1.0011\psi_k(\mathsf{iqr}(H, p(z)))$.*

Proof. Calling $\widetilde{H} = \mathsf{IQR}(H, p(z))$ and $H = \mathsf{iqr}(H, p(z))$, one of two cases are possible. If $H_{i+1,i} < 0.999\epsilon$ for some $i \in [n-1]$, then applying Lemma 4.2.11 and our assumption on $\mathbf{u}$,

$$\widetilde{H}_{i+1,i} < H_{i+1,i} + 0.001\epsilon < \epsilon.$$

On the other hand, if for every $i \in [n-1]$ we have $H_{i+1,i} \geq 0.999\epsilon$, then

$$\psi_k\left(\widetilde{H}\right) \leq \psi_k(H) \left(\prod_{i \in [n-1]} \left(1 + \frac{0.001\epsilon}{H_{i+1,i}}\right)\right)^{1/k} \leq 1.0011\psi_k(H).$$

□

Lemma 4.2.22 (Guarantees for Exc)**.** *Assume that $|r| + 1.001\theta\alpha B^{1/k}\psi_k(H) \leq C\|H\|$ and*

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{\mathsf{Exc}}(n, k, C, \Sigma, B, \theta, \epsilon, \phi, \gamma, \xi, \alpha) \\ &:= \min\left\{\mathbf{u}_\psi(k), \frac{0.1\varepsilon \cdot 1.998\theta\alpha B\epsilon}{4(\varepsilon + 2(1+\varepsilon)C\Sigma)}, \right. \\ &\qquad \left. \mathbf{u}_{4.2.21}\left(n, k, C, \Sigma, B, \left(\frac{\xi(1-\gamma)}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}} \cdot \frac{1.998\,\theta\alpha B^{1/k}\epsilon\sqrt{\phi}}{\sqrt{3n}}, \epsilon\right)\right\} \qquad (4.45) \\ &= 2^{-O\left(k \log \frac{n\Sigma B\alpha\theta}{\xi(1-\gamma)\epsilon\phi}\right)}. \qquad (4.46) \end{aligned}$$

Then Exc (defined below) satisfies its guarantees and runs in at most

$$T_{\mathsf{Exc}}(n, k, \xi, \gamma, B, \alpha, \theta)$$

$$:= T_\psi(k) + 2S\left(\left(\frac{\xi(1-\gamma)}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\right) + C_{\mathsf{D}} + O(1) = O\left(B^{\frac{8}{k-1}}\left(\frac{\alpha^2\theta^2}{\xi(1-\gamma)}\right)^{\frac{2k}{k-1}}\right)$$

arithmetic operations and

$$|\mathcal{S}| \le S\left(\left(\frac{\xi(1-\gamma)}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\right) = O\left(B^{\frac{8}{k-1}}\left(\frac{\alpha^2\theta^2}{\xi(1-\gamma)}\right)^{\frac{2k}{k-1}}\right)$$

where the function $S(\varepsilon) = O(\varepsilon^{-2})$ is defined in (4.48).

Exc

Input: Hessenberg H, initial shift r, working accuracy ϵ, stagnation ratio ξ, failure probability tolerance ϕ
Global Data: Condition number bound B, decoupling rate γ, norm bound Σ, optimality parameter θ, promising parameter α
Output: Finite subset $\mathcal{S} \subset \mathbb{C}$.
Requires: H is ϵ-unreduced, $\kappa_V(H) \le B$, $\|H\| \le \Sigma$, r is a θ-approximate, α-promising Ritz value, and $\tau_{(z-r)^k}(H) \ge \xi\psi_k(H)$
Ensures: With probability at least $1-\phi$, some $s \in \mathcal{S}$ satisfies at least one of

- (ϵ-Decoupling) A subdiagonal of $\mathsf{IQR}(H, (z-s)^k)$ is smaller than ϵ
- (Potential Reduction) $\psi_k(\mathsf{IQR}(H, (z-s)^k)) \le 1.0011(1-\gamma)\psi_k(H)$

1. $\widetilde{R} \leftarrow 2^{1/k}\alpha B^{1/k}\theta\widetilde{\psi_k}(H)$
2. $\varepsilon \leftarrow \left(\frac{\xi(1-\gamma)}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}$
3. $\mathcal{S}_0 \leftarrow$ maximal 0.99ε-net of $D\big(0, 1+\varepsilon\big)$
4. $w \sim \mathrm{Unif}\left(D\big(0, \varepsilon\widetilde{R}\big)\right)$
5. $\mathcal{S} \leftarrow \mathsf{fl}\left((r + w + \widetilde{R}\mathcal{S}_0)\right) \cap D\big(r, \widetilde{R}\big)$

Proof of Lemma 4.2.22. From (4.40), the fact that $\mathbf{u} \le \mathbf{u}_\psi(k)$ we can bound

$$1.998\,\theta\alpha B\psi_k(H) \le (2\cdot.999)^{1/k}\theta\alpha B\psi_k(H) \le \widetilde{R} \le 1.001\cdot\theta\alpha B^{1/k}\psi_k(H), \tag{4.47}$$

meaning that (as $\psi_k(H) \le \|H\|$) the set $\mathcal{S}$ is contained in a disk of radius $|r| + 1.001\theta\alpha B^{1/k}\|H\| = C\|H\|$. We can then obtain that

$$\mathbb{P}\left[Z_H \in D(r, \widetilde{R})\right] \ge \mathbb{P}\left[|Z_H - r| \le 1.998\,\theta\alpha\kappa_V^{1/k}(H)\psi_k(H)\right] \quad \text{by (4.47)}$$

$$\begin{aligned}
&\geq \left(1 - \frac{1}{1.998}\right)^2 \frac{\xi^{2k}}{\kappa_V(H)^4 \alpha^{2k} \theta^{2k}} && \text{Lemma 4.1.9 with } t = \frac{1}{1.998} \\
&\geq \frac{0.24\,\xi^{2k}}{B^4 \alpha^{2k} \theta^{2k}} \\
&:= P.
\end{aligned}$$

When we shift and scale each point $s_0 \in \mathcal{S}_0$ in finite arithmetic,

$$\begin{aligned}
|\mathsf{fl}(r + w + \widetilde{R}s_0) - r + w + \widetilde{R}s_0| &\leq \frac{3\mathbf{u}}{1 - 3\mathbf{u}} |r + w + \widetilde{R}s_0| \\
&\leq 4\mathbf{u} \left(|r| + \varepsilon + (1+\varepsilon) 1.001 \theta \alpha B^{1/k} \psi_k(H)\right) \\
&\leq 4\mathbf{u} \left(\varepsilon + 2(1+\varepsilon) C\Sigma\right) \\
&\leq 0.1\varepsilon \cdot 1.998\, \theta \alpha B \epsilon \\
&\leq 0.1\varepsilon \widetilde{R}
\end{aligned}$$

from our assumption on $\mathbf{u}$, which means that the computed $\mathcal{S}$ still contains a $\varepsilon\widetilde{R}$-net of $D(r, \widetilde{R})$. We will assume for simplicity that one can perform the intersection in the final line of Exc while preserving the property that $\mathcal{S}$ is a maximal ε-net of $D(r, \widetilde{R})$) —this can be achieved, e.g., by intersecting with a slightly larger set and projecting all points outside $D(r, \widetilde{R})$) to this latter set. Since $\mathcal{S}$ is a maximal ε-net of $D(r, \widetilde{R})$), it has size at most $9/\varepsilon^2$, and we may recycle a calculation from Section 4.1,

$$\max_{s \in \mathcal{S}} \tau_{(z-s)^k}^{-2k}(H) \geq \frac{P}{9B^2 \varepsilon^{2k-2} \widetilde{R}^{2k}} \geq \frac{1}{(1-\gamma)^{2k} \psi_k^{2k}(H)}$$

provided that ε is no larger than

$$\begin{aligned}
\left(\frac{P(1-\gamma)^{2k}\psi_k^{2k}(H)}{9B^2\widetilde{R}^{2k}}\right)^{\frac{1}{2k-2}} &\geq \left(\frac{0.24\xi^{2k}(1-\gamma)^{2k}}{B^6\alpha^{2k}\theta^{2k} \cdot 9 \cdot 2.001^2 \theta^{2k} \alpha^{2k} B^2}\right)^{\frac{1}{2k-2}} \\
&\geq \left(\frac{\xi(1-\gamma)}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}},
\end{aligned}$$

which is the expression appearing in line 2 of Exc.

On the other hand, after the random translation, one can quickly show that every $s \in \mathcal{S}$ is forward stable with high probability. Because the net is maximal (meaning that no two of the points in it are within $\varepsilon\widetilde{R}$ of one another) each eigenvalue $\lambda \in \mathrm{Spec}(H)$ lies within distance $\varepsilon\widetilde{R}$ of at most three points in the net, so the probability that $\mathsf{dist}(\lambda, \mathcal{S}) < \eta$ after the random translation is at most $3\eta^2/\varepsilon^2\widetilde{R}^2$. Thus the probability that $\mathsf{dist}(\mathrm{Spec}(H), \mathcal{S}) < \eta$ after the random translation is at most $3n\eta^2/\varepsilon^2\widetilde{R}^2$. To ensure that this is smaller than the failure probability ϕ, we can safely set

$$\eta = \frac{\varepsilon\widetilde{R}\sqrt{\phi}}{\sqrt{3n}} \geq \left(\frac{\xi(1-\gamma)}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}} \cdot \frac{1.998\theta\alpha B^{1/k}\epsilon\sqrt{\phi}}{\sqrt{3n}}.$$

In the event that the shifts are all forward stable, the definition of $\mathbf{u}_{\mathsf{Exc}}$ means that we can invoke Lemma 4.2.21: either some subdiagonal of $\mathsf{IQR}(H,(z-s)^k)$ is smaller than ϵ, or $\mathsf{IQR}(H,(z-s)^k)$ satisfies

$$\psi_k(\mathsf{IQR}(H,(z-s)^k)) < 1.0011\psi_k(\mathsf{iqr}(H,(z-s)^k)) \leq 1.0011\tau_{(z-s)^k}(H) \leq 1.0011(1-\gamma)\psi_k(H).$$

One practical choice of the of the initial $.99\varepsilon$-net of $D(0,(1+\varepsilon))$ is to take an equilateral triangular lattice with spacing $\sqrt{3}\varepsilon$ and intersect it with $D(0,(1+1.99\varepsilon))$; since this lattice gives an optimal planar sphere packing, it is the optimal choice of net as $\varepsilon \to 0$. Other choices may be more desirable when ε is large. Adapting an argument of [4, Lemma 2.6] (which in turn uses [28, Theorem 3, p327]) one can show that with this choice of $\mathcal{S}_0$,

$$\begin{aligned} |\mathcal{S}| \leq |\mathcal{S}_0| &\leq \frac{2\pi}{3\sqrt{3}}\left(1.99+\frac{1}{0.99\varepsilon}\right)^2 + \frac{4\sqrt{2}}{\sqrt{3}}\left(1.99+\frac{1}{0.99\varepsilon}\right)+1 \\ &:= S(\varepsilon) \end{aligned} \tag{4.48}$$

We will see that every time Exc is called in the course of the full algorithm $\mathsf{ShiftedQR}$, the same ε is used, depending only on the global data. Thus the original net of $D(0,1+\varepsilon)$ need only be computed once, and can be regarded a fixed overhead cost of the algorithm. Given the original net, computing $\mathcal{S}$ costs one arithmetic operation to add $r+w$, followed by $|\mathcal{S}_0|$ each to scale and shift by $r+w$. Add to this the operations to compute $\widetilde{\psi_k}(H)$ and $\widetilde{R}$, and the cost of obtaining the single random sample, and we get a total of

$$2|\mathcal{S}_0| + C_{\mathsf{root}}k\log(k\log\tfrac{1}{1-0.999^{1/k}}) + O(1)$$

arithmetic operations. Bounding $|\mathcal{S}_0| \leq S(\varepsilon)$ yields the assertion of the lemma. □

Analysis of $\mathsf{Sh}_{k,B}$. We now specify and analyze the complete shifting strategy $\mathsf{Sh}_{k,B}$.

Lemma 4.2.23 (Guarantees for $\mathsf{Sh}_{k,B}$)**.** *Assume that* $|r| + 1.001\theta\alpha B^{1/k}\psi_k(H) \leq C\|H\|$ *and*

$$\mathbf{u} \leq \mathbf{u}_{\mathrm{Sh}}(n,k,C,\Sigma,B,\mathsf{dist}(\mathcal{R},\mathrm{Spec}(H)),\theta,\epsilon,\phi,\gamma,\alpha) \tag{4.49}$$

$$\begin{aligned} &:= \min\Big\{\mathbf{u}_{\mathsf{Find}}(n,k,C,\Sigma,B,\mathsf{dist}(\mathcal{R},\mathrm{Spec}(H))), \\ &\qquad \mathbf{u}_{\mathsf{Exc}}\left(n,k,C,\Sigma,B,\theta,\epsilon,\phi,\gamma,0.999(1-\gamma),\alpha\right), \\ &\qquad \mathbf{u}_{4.2.21}(n,k,C,\Sigma,B,\mathsf{dist}(\mathcal{R},\mathrm{Spec}(H)),\epsilon)\Big\} \end{aligned} \tag{4.50}$$

$$= 2^{-O\left(k\log\frac{n\Sigma B\theta\alpha}{(1-\gamma)\epsilon\phi\mathsf{dist}(\mathcal{R},\mathrm{Spec}(H))}\right)}$$

Then, $\mathsf{Sh}_{k,B}$ *(defined below) satisfies its guarantees, and runs in at most*

$$T_{\mathrm{Sh}}(n,k,\gamma,B,\alpha,\theta) := T_{\mathsf{Find}}(n,k) + T_{\mathsf{Tau}}(n,k) + T_{\mathsf{Exc}}(n,k,0.999(1-\gamma),(1-\gamma),B,\alpha,\theta)$$

$$+ S\left(\left(\frac{0.999(1-\gamma)^2}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\right)(T_{\mathsf{IQR}}(n,k)+T_\psi(n,k))$$
$$= O\left(kn^2B^{\frac{8}{k-1}}\left(\frac{\alpha\theta}{(1-\gamma)}\right)^{\frac{4k}{k-1}}\right)$$

arithmetic operations.

$\mathsf{Sh}_{k,B}$

Input: Hessenberg H, θ-optimal Ritz values $\mathcal{R}$ of H, working accuracy ϵ, failure probability tolerance ϕ.
Global Data: Condition number bound B, decoupling rate γ, norm bound Σ, optimality parameter θ, promising parameter α
Output: Hessenberg H.
Requires: H is ϵ-unreduced and $\kappa_V(H) \le B$
Ensures: With probability at least $1-\phi$, either H is ϵ-decoupled or $\psi_k(H) \le 1.002(1-\gamma)\psi_k(H)$

1. $r \leftarrow \mathsf{Find}(H, \mathcal{R})$
2. **If** $\mathsf{Tau}^k(H,(z-r)^k) < (1-\gamma)^k\psi_k^k(H)$, output $H = \mathsf{IQR}(H,(z-r)^k)$.
3. **Else**, $\mathcal{S} \leftarrow \mathsf{Exc}(H, r, \epsilon, 0.999(1-\gamma), \phi)$.
4. **For** each $s \in \mathcal{S}$, **if** $\psi_k(\mathsf{IQR}(H,(z-s)^k)) < 1.002(1-\gamma)\psi_k(H)$ or some subdiagonal of $\mathsf{IQR}(H,(z-s)^k)$ is smaller than ϵ, output $H = \mathsf{iqr}(H,(z-s)^k)$

Proof of Lemma 4.2.23. The definition of $\mathbf{u}_{\mathrm{Sh}}$ ensures that Exc and Find (and therefore Tau) satisfy their guarantees when called in the course of Sh; the analysis of Sh is accordingly straightforward. In line 1, Find produces an α-promising, θ-approximate Ritz value r for $\alpha = (1.01B)^{\frac{4\log k}{k}}$ as in Table 4.1; in line 2 — because every subdiagonal of H is assumed larger than ϵ — we know from definition of $\mathbf{u}_{\mathrm{Sh}}$ and Lemma 4.2.21 that if $\mathsf{Tau}^k(H,(z-r)^k) \le (1-\gamma)^k\psi_k^k(H)$, then

$$\begin{aligned}\psi_k(\mathsf{IQR}(H,(z-r)^k)) &\le 1.0011\psi_k(\mathsf{iqr}(H,(z-r)^k))\\ &\le 1.0011\tau_{(z-r)^k}(H)\\ &\le 1.0011\cdot\left(1.001\mathsf{Tau}^k(H,(z-r)^k)\right)^{1/k}\\ &\le 1.002(1-\gamma)\psi_k(H).\end{aligned}$$

On the other hand, if $\mathsf{Tau}^k(H,(z-r)^k) > (1-\gamma)^k\psi_k^k(H)$ in line 2, then using the guarantees for Tau^k,

$$\tau^k_{(z-r)^k}(H) > 0.999\mathsf{Tau}^k(H,(z-r)^k) \ge 0.999(1-\gamma)^k\psi_k(H).$$

Finally, Exc satisfies its guarantees from Lemma 4.2.22 when called with $\alpha = (1.01B)^{\frac{4\log k}{k}}$ and $\xi = 0.999^{1/k}(1-\gamma)$. Thus with probability at least $1-\phi$ at least one exceptional shift $s \in \mathcal{S}$ satisfies either decoupling (some subdiagonal smaller than ϵ) or potential reduction $(\psi_k(\mathsf{IQR}(H,(z-s)^k)) \leq 1.0011(1-\gamma)\psi_k(H) \leq 1.002(1-\gamma)\psi_k(H))$.

For the arithmetic operations, $\mathsf{Sh}_{k,B}$ requires one call to Find, one to Tau^k, one to Exc with stagnation ratio $\xi = 0.999(1-\gamma)$, and finally $|\mathcal{S}|$ calls to degree-k IQR. We can bound $|\mathcal{S}| \leq S(\varepsilon)$, where ε is defined in the course of Exc with stagnation ratio parameter $\xi = 0.999(1-\gamma)$, and $S(\cdot)$ is defined in (4.48). Since and checking every shift in $\mathcal{S}$ for potential reduction dominates the arithmetic operations, we get that

$$T_{\mathrm{Sh}}(n,k,B,\gamma,\alpha,\theta) = O\left(kn^2 \cdot B^{\frac{8}{k-1}}\left(\frac{\alpha\theta}{(1-\gamma)}\right)^{\frac{4k}{k-1}}\right).$$

□

4.2.7 Finite Arithmetic Analysis of ShiftedQR

Preservation of gap **and** κ_V

Lemma 4.2.24. *Suppose A has distinct eigenvalues. Then for any E satisfying*

$$\|E\| \leq \frac{\mathrm{gap}(A)}{8n^2 \cdot \kappa_V^3(A)} \tag{4.51}$$

we have

$$\mathrm{gap}(A+E) \geq \mathrm{gap}(A) - 2\kappa_V(A)\|E\| \tag{4.52}$$

and

$$\kappa_V(A+E) \leq \kappa_V(A) + 6n^2\frac{\kappa_V^3(A)}{\mathrm{gap}(A)}\|E\|. \tag{4.53}$$

Proof. The assertion in (4.52) is an immediate consequence of the Bauer-Fike theorem. For (4.53), let V be scaled so that $\|V\| = \|V^{-1}\| = \kappa_V(A)$, with (not necessarily unit) columns $v_1, ..., v_n$ satisfying $Av_i = \lambda_i v_i$ for each $i \in [n]$. It follows from [11, Proposition 1.1] that whenever $\|E\| \leq \frac{\mathrm{gap}(A)}{8\kappa_V(A)}$, there exists a matrix V' with columns $v_1', ..., v_n'$ diagonalizing $A' := A + E$, such that

$$\|v_i - v_i'\| \leq 2n\frac{\kappa_V(A)}{\mathrm{gap}(A)}\|E\|\|v_i\|,$$

which implies

$$\|V - V'\| \leq \|V - V'\|_F \leq 2n^{3/2}\frac{\kappa_V(A)}{\mathrm{gap}(A)}\|E\|\|V\|_F \leq 2n^2\frac{\kappa_V(A)}{\mathrm{gap}(A)}\|V\|.$$

It is standard that each singular value of V' satisfies $|\sigma_i(V') - \sigma_i(V)| \leq \|V - V'\|$, so using $\|V\| = \|V^{-1}\| = \sqrt{\kappa_V(A)}$, we have

$$\begin{aligned}
\kappa_V(A') &\leq \|V'\|\|(V')^{-1}\| \\
&\leq \frac{\|V\| + \|V - V'\|}{\|V^{-1}\|^{-1} - \|V - V'\|} \\
&\leq \kappa_V(A) \frac{1 + 2n^2 \frac{\kappa_V(A)}{\mathrm{gap}(A)}\|E\|}{1 - 2n^2 \frac{\kappa_V^2(A)}{\mathrm{gap}(A)}\|E\|} \\
&\leq \kappa_V(A) + \frac{8}{3} n^2 (1 + \kappa_V(A)) \frac{\kappa_V^2(A)}{\mathrm{gap}(A)} \|E\|,
\end{aligned}$$

where in the final line we have used (4.51) to argue that $2n^2 \frac{\kappa_V^2(A)}{\mathrm{gap}(A)}\|E\| \leq 1/4$, and convexity of the function $f(x) = \frac{1 + x/\kappa_V(A)}{1-x}$ to bound by the linear interpolation between $x = 0$ and $x = 1/4$. The advertised bound then follows from applying $\kappa_V(A) \geq 1$ and bounding $16/3 \leq 6$. □

Lemma 4.2.25. *If A is block upper triangular and A' is a diagonal block, then $\kappa_V(A') \leq \kappa_V(A)$ and $\mathrm{gap}(A') \geq \mathrm{gap}(A)$.*

Proof. The gap assertion is immediate since $\mathrm{Spec}A' \subset \mathrm{Spec}A$. For κ_V, assume without loss of generality that A is diagonalizable (otherwise the inequality is trivial) and

$$A = \begin{pmatrix} A' & * \\ 0 & * \end{pmatrix}.$$

We claim that every V diagonalizing A is of the form

$$V = \begin{pmatrix} V' & * \\ 0 & * \end{pmatrix},$$

where V' diagonalizes A'. To see this, if $AV = VD$, then block upper triangularity gives $A'V' = V'D'$ for D' the upper left block of D. Moreover, V invertible implies V' is as well, and quantitatively $\|V'\|\|(V')^{-1}\| \leq \|V\|\|V^{-1}\|$. Choosing V so that $\kappa_V(A) = \|V\|\|V^{-1}\|$, we have

$$\kappa_V(A') \leq \|(V')\|\|(V')^{-1}\| \leq \|V\|\|V^{-1}\| = \kappa_V(A).$$

□

The Full Algorithm

We are now ready to analyze, in finite arithmetic, how the shifting strategy $\mathsf{Sh}_{k,B}$ introduced in Section 4.1 can be used to approximately find all eigenvalues of a Hessenberg matrix H. One simple subroutine is required in addition to the ones described in the preceding sections: `deflate`(H, ϵ, k) takes as input a Hessenberg matrix H, deletes any of the bottom $k - 1$

subdiagonal entries smaller than ϵ, and outputs the resulting diagonal blocks $H_1, H_2, \ldots$. It runs in $T_{\texttt{deflate}}(H, \epsilon, k) = k$ arithmetic operations.

ShiftedQR

Input: Hessenberg matrix H, accuracy δ, failure probability tolerance ϕ
Global Data: Eigenvector condition number bound B, eigenvalue gap bound Γ, matrix norm bound Σ, original matrix dimension n
Requires: $\Sigma \geq 2\|H\|$, $B \geq 2\kappa_V(H)$, $\Gamma \leq \text{gap}(H)/2$, $\delta \leq \Sigma$
Output: A multiset $\Lambda \subset \mathbb{C}$
Ensures: With probability at least $1-\phi$, Λ are the eigenvalues of some $\widetilde{H}$ with $\|\widetilde{H} - H\| \leq \delta$

1. $\epsilon \leftarrow \frac{1}{4n} \min\left\{\delta, \frac{\Gamma}{8n^2B^2}\right\}$, $\varphi \leftarrow \frac{\phi}{3n^2} \frac{\log \frac{1}{1.002(1-\gamma)}}{\log \frac{\Sigma}{\epsilon}}$
2. **If** $\dim(H) \leq k$, $\Lambda \leftarrow \mathsf{SmallEig}(H, \delta, \phi)$, output Λ and halt
3. **Else** $\Lambda \leftarrow \emptyset$ and
   a) **While** $\min_{n-k+1 \leq i \leq n} H_{i,i-1} > \epsilon$
      i. $[\mathcal{R}, H, \texttt{dec}] = \mathsf{RitzOrDecouple}(H, \epsilon, \varphi)$
      ii. **If** $\texttt{dec} = \texttt{true}$, $H \leftarrow H$ and end while
      iii. **Else if** $\texttt{dec} = \texttt{false}$, $H \leftarrow \mathsf{Sh}_{k,B}(H, \mathcal{R}, \epsilon, \varphi)$
   b) $[H_1, H_2, \ldots H_\ell] = \texttt{deflate}(H, \epsilon)$
   c) **For** each $j \in [\ell]$
      i. **If** $\dim(H_j) \leq k$, $\Lambda \leftarrow \Lambda \sqcup \mathsf{SmallEig}(H_j, \delta/n, \phi/3n)$
      ii. **Else** repeat lines 3a-3c on H_j

Theorem 4.2.26 (Guarantees for ShiftedQR)**.** *Let k, θ, α, and γ be set in terms of B as in (4.12), N_{dec} be defined in (4.57), and ϵ and φ be defined in line 1 of* ShiftedQR*. Assuming*

$$\begin{aligned}\mathbf{u} &\leq \mathbf{u}_{\mathsf{ShiftedQR}}(n, k, \Sigma, B, \delta) \\ &:= \min\left\{ \frac{\epsilon}{4.5kN_{\mathsf{dec}} \cdot n\nu_{\mathsf{IQR}}(n)\Sigma}, \mathbf{u}_{\mathsf{RitzOrDecouple}}(n, k, \Sigma, B, \theta, \epsilon, \varphi), \right. \\ &\qquad \left. \mathbf{u}_{\mathrm{Sh}}\left(n, k, 3, \Sigma, B, \frac{\epsilon^2\sqrt{\varphi}}{32 \cdot 101 \cdot \Sigma\sqrt{2k}}, \theta, \epsilon, \varphi, \gamma, \alpha\right)\right\} \qquad (4.54) \\ &= 2^{-O\left(k \log \frac{n\Sigma B}{\delta\Gamma\phi}\right)},\end{aligned}$$

ShiftedQR *satisfies its guarantees and runs in at most*

$$T_{\mathsf{ShiftedQR}}(n, k, \delta, B, \Sigma, \gamma) \leq n\Big(T_{\mathsf{RitzOrDecouple}}(n, k, \texttt{true})$$

$$+ N_{\mathsf{dec}}\Big(T_{\mathsf{RitzOrDecouple}}(n,k,\texttt{false}) + T_{\mathrm{Sh}}(n,k,\gamma,B,\alpha,\theta)\Big) \tag{4.55}$$
$$+ T_{\mathit{deflate}}(k)\Big)$$
$$= O\left(\left(\log\frac{nB\Sigma}{\delta\Gamma}k\log k + k^2\right)n^3\right)$$

arithmetic operations, plus $O(n\log\frac{nB\Sigma}{\delta\Gamma})$ *calls to* SmallEig *with accuracy* $\Omega(\frac{\Gamma^2}{n^4B^4\Sigma})$ *and failure probability tolerance* $\Omega(\frac{\phi}{n^2\log\frac{nB\Sigma}{\delta\Gamma}})$.

In the above result, we assume access to an upper bound $\Sigma \geq 2\|H\|$ and show that ShiftedQR can approximate the eigenvalues of H with (absolute) backward error δ, whereas in our main Theorem 1.4.6, we ask for (relative) backward error δ. To prove Theorem 1.4.6 from Theorem 4.2.26, we need only compute an upper bound $\Sigma \geq 2\|H\|$ and call ShiftedQR with accuracy δ/Σ. Such a bound can be computed (for instance) using random vectors or, at the cost of a factor of $\sqrt{n}$, by taking the Frobenius norm of H. In either case, the arithmetic cost and precision are dominated by the requirements for ShiftedQR itself.

Proof of Theorem 4.2.26. At a high level, ShiftedQR is given an input matrix H, ϵ-decouples H to a unitarily similar matrix H via a sequence of applications of $\mathsf{RitzOrDecouple} + \mathsf{Sh}_{k,B}$, deflates H to a block upper triangular matrix with diagonal blocks $H_1, ..., H_\ell$, then repeats this process on each block H_j with dimension larger than $k \times k$. Since the effect of RitzOrDecouple and $\mathsf{Sh}_{k,B}$ on any input matrix H' is approximately a unitary conjugation, it will be fruitful for the analysis to regard each of the blocks $H_1, ..., H_\ell$ as embedded in the original matrix, and promote the approximate unitary conjugation actions of the subroutines on each block to unitary conjugations of the full matrix. The same goes once each of $H_1, ..., H_\ell$ is decoupled and deflated and we pass to further submatrices of each one. Importantly, this viewpoint is necessary *only* for the analysis: the algorithm need not actually manipulate the entries outside the blocks $H_1, ..., H_\ell$. In this picture, the end point of the algorithm is a matrix of the form

$$\begin{pmatrix} L_1 & * & * \\ & L_2 & * \\ & & \ddots \end{pmatrix}, \tag{4.56}$$

where $L_1, L_2, ...$ are all $k \times k$ or smaller matrices on which SmallEig can be called directly, and the $*$ entries are unknown and irrelevant to the algorithm. By the guarantees on SmallEig (and the fact that β-forward approximation of eigenvalues implies β-backward approximation), the output of the algorithm is thus

$$\bigsqcup_j \mathsf{SmallEig}(L_j,\epsilon,\varphi) = \bigsqcup_j \mathrm{Spec}(\widetilde{L}_j) = \mathrm{Spec}\begin{pmatrix} \widetilde{L}_1 & * & * \\ & \widetilde{L}_2 & * \\ & & \ddots \end{pmatrix}$$

where $\widetilde{L}_1, \widetilde{L}_2, ...$ are some matrices satisfying $\|L_j - \widetilde{L}_j\| \leq \delta/n$, and the remaining entries are identical to those in (4.56). Our goal in the proof will thus be to show that for some unitary $\widetilde{Q}$,

$$\left\| \begin{pmatrix} L_1 & * & * \\ & L_2 & * \\ & & \ddots \end{pmatrix} - \widetilde{Q}^* H \widetilde{Q} \right\| \leq \delta - \delta/n,$$

where the left hand matrix is a block upper triangular matrix with the blocks $L_1, L_2, ...$ on the diagonal. This will in turn imply that

$$\begin{aligned} \left\| \begin{pmatrix} \widetilde{L}_1 & * & * \\ & \widetilde{L}_2 & * \\ & & \ddots \end{pmatrix} - \widetilde{Q}^* H \widetilde{Q} \right\| &\leq \left\| \begin{pmatrix} L_1 - \widetilde{L}_1 & * & * \\ & L_2 - \widetilde{L}_2 & * \\ & & \ddots \end{pmatrix} \right\| + \delta - \delta/n \\ &\leq \max_i \|L_i - \widetilde{L}_i\| + \delta - \delta/n \leq \delta, \end{aligned}$$

as desired.

We begin by analyzing the while loop in line 3a.

Lemma 4.2.27. *Assume that at during the execution of* ShiftedQR*, the while loop in line 3a is initialized with a matrix H' satisfying $\|H'\| \leq (1 - 1/2n)\Sigma$, $\kappa_V(H') \leq (1 - 1/2n)B$, and $\mathrm{gap}(H') \geq (1 + 1/2n)\Gamma$. Let*

$$N_{\mathsf{dec}} := \frac{\log \frac{\Sigma}{\epsilon}}{\log \frac{1}{1.002(1-\gamma)}}. \tag{4.57}$$

If

$$\mathbf{u} \leq \mathbf{u}_{\mathsf{ShiftedQR}}(n, k, \Sigma, B, \delta),$$

then the loop terminates in at most N_{dec} iterations, having produced a ϵ-decoupled matrix H' at most ϵ-far from a unitary conjugate of H'.

Proof of Lemma. Let us write H'' for the matrix produced by several runs through lines 3(a)i-3(a)iii, after the while loop has been initialized with H', and assume that all prior calls to RitzOrDecouple or $\mathsf{Sh}_{k,B}$ during the loop have satisfied their guarantees, and moreover that all prior shifts have had modulus at most $4.5\|H'\|$ in the complex plane. We will show inductively that this last condition holds throught the while loop.

Because the prior calls to RitzOrDecouple and $\mathsf{Sh}_{k,B}$ satisfy their guarantees, each previous run through lines 3(a)i-3(a)iii has either effected immediate decoupling or potential reduction by a multiplicative $1.002(1 - \gamma)$. Since $\epsilon \leq \psi_k(H') \leq \|H'\| \leq \Sigma$, there can have been at most N_{dec} runs through lines 3(a)i-3(a)iii so far, each of which we can regard as an IQR step of degree k, meaning that we can think of H'' as being produced from H' by a *single* IQR step of degree kN_{dec}.[5] Thus by Lemma 4.2.9, our inductive assumption on the prior

[5]This is because we have simply defined a higher degree IQR step as a composition of many degree 1 IQR steps.

shifts, and the hypothesis on $\mathbf{u}$, the distance from H'' to a unitary conjugate of H' is at most $4.5\|H\|kN_{\mathsf{dec}}\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq \epsilon$. If H'' is ϵ-decoupled, then the while loop terminates, and the proof is complete.

Otherwise H'' is not ϵ-decoupled. By the definition of ϵ and the fact that $\epsilon \leq \delta/2n \leq \Sigma/2n$, we can apply Lemma 4.2.24 to find

$$\begin{aligned}
\|H''\| &\leq \|H'\| + \epsilon \leq (1 - 1/2n)\Sigma + \Sigma/2n \leq \Sigma \\
\kappa_V(H'') &\leq \kappa_V(H') + 6n^2 \frac{\kappa_V^3(H')}{\mathrm{gap}(H')}\epsilon \leq (1 - 1/2n)B + B/2n \leq B \\
\mathrm{gap}(H'') &\geq \mathrm{gap}(H') - 2\kappa_V(H')\epsilon \geq (1 + 1/2n)\Gamma - \Gamma/2n \geq \Gamma,
\end{aligned}$$

and we furthermore have $2\epsilon^2/\Sigma \leq 2\epsilon \leq \Gamma \leq \mathrm{gap}(H'')$ by the above and the definition of ϵ. This means $\mathsf{RitzOrDecouple}(H'', \epsilon, \varphi)$ meets its requirements, and from our assumption on $\mathbf{u}$ we can apply Lemma 4.2.19 to conclude that it satisfies its guarantees. If this call to $\mathsf{RitzOrDecouple}$ outputs $\mathtt{dec} = \mathtt{true}$, then the matrix it outputs is indeed decoupled and the while loop terminates.

If on the other hand $\mathtt{dec} = \mathtt{false}$, then $\mathsf{RitzOrDecouple}$ outputs H'' and θ-approximate Ritz values $\mathcal{R}$ contained in in a disk of radius $1.1\|H''\|$, and $\mathsf{RitzOrDecouple}$ guarantees

$$\mathsf{dist}(\mathcal{R}, H'') \geq \frac{\epsilon^2\sqrt{\varphi}}{32 \cdot 101 \cdot \Sigma\sqrt{2k}}.$$

The bound on $\kappa_V(H'')$ in the previous paragraph ensures that the requirements of the algorithm $\mathsf{Sh}_{k,B}(H, \mathcal{R}, \epsilon, \varphi)$ have been met, and the parameter settings in (4.12)-(4.13) give us

$$\begin{aligned}
1.001\theta\alpha B^{1/k}\psi_k(H'') &= 1.001\frac{1.01}{0.998^{1/k}}(2B^4)^{1/2k}(1.01B)^{\frac{4\log k}{k}}B^{1/k}\psi_k(H'') \\
&= 1.04 \cdot 2^{1/2k}B^{\frac{4\log k + 3}{k}}\psi_k(H'') \\
&\leq 1.04 \cdot \sqrt{2^{\frac{2}{k-1}}B^{\frac{8\log k + 11}{k-1}}}\|H''\| \\
&\leq 1.04\sqrt{3}\|H''\| \\
&\leq 1.9\|H''\|,
\end{aligned}$$

so every exceptional shift has modulus at most $3\|H''\|$ in the complex plane. Our assumption on $\mathbf{u}$ lets us invoke Lemma 4.2.23 to conclude that $\mathsf{Sh}_{k,B}$ achieves potential reduction by a multiplicative factor of $1.002(1 - \gamma)$. Moreover, the shifts executed by $\mathsf{RitzOrDecouple}$ and Sh in the above run through the while loop had modulus at most

$$3\|H''\| \leq 3\|H'\|(1 + 4.5kN_{\mathsf{dec}}\nu_{\mathsf{IQR}}(n)\mathbf{u}) \leq 3\|H'\| \cdot (1 + \epsilon/\Sigma) \leq 4.5\|H'\|,$$

again since $\epsilon \leq \delta/2n \leq \Sigma$.

The proof above ensures that for each of its first N_{dec} iterations, the while loop either produces decoupling or potential reduction by a multiplicative $1.002(1-\gamma)$, and our earlier discussion implies that it therefore terminates after after at most N_{dec} iterations. When it does, the proof above additionally tells us that the final matrix H' is at most ϵ-far from a unitary conjugate of H', as desired. □

We next check that each time the while loop begins in the course of ShiftedQR, the hypotheses of Lemma 4.2.27 are satisfied. This is immediate the first time the loop begins, where the requirements of ShiftedQR give $\|H\| \leq \Sigma/2$, $\kappa_V(H) \leq B/2$, and $\mathrm{gap}(H) \geq 2\Gamma$. If H' is a matrix passed to the while loop, and each of the while loops in its production has satisfied the conclusion of Lemma 4.2.27, then H' is the result of at most $n-1$ of decouplings-and-deflations, each of which caused the norm, eigenvector condition number, and gap to deteriorate by at worst an additive 2ϵ. Thus, finally using the full force of the $1/4n$ factor in the definition of ϵ,

$$\begin{aligned}
\|H'\| &\leq \|H\| + 2(n-1)\epsilon \leq (1-1/2n)\Sigma \\
\kappa_V(H') &\leq \kappa_V(H) + 6n^2 \frac{\kappa_V^3(H)}{\mathrm{gap}(H)} \cdot 2(n-1)\epsilon \leq (1-1/2n)B \\
\mathrm{gap}(H') &\geq \mathrm{gap}(H) - 2\kappa_V(H) \cdot 2(n-1)\epsilon \geq (1+1/2n)\Gamma
\end{aligned}$$

by the definition of ϵ.

This ensures that *every* execution of the while loop throughout ShiftedQR satisfies the conclusion of Lemma 4.2.27, which means that the set of 'base case' matrices $L_1, L_2, \ldots$ are produced by a tree of alternating decouplings and deflations with depth at most $n-1$, and moreover that

$$\left\| \begin{pmatrix} L_1 & * & * \\ & L_2 & * \\ & & \ddots \end{pmatrix} - \widetilde{Q}^* H \widetilde{Q} \right\| \leq 2(n-1)\epsilon \leq \delta - \delta/n,$$

for some unitary $\widetilde{Q}$, as we had set out to show.

Failure Probability. We have already shown that RitzOrDecouple and $\mathsf{Sh}_{k,B}$ satisfy their guarantees (including their failure probability) throughout ShiftedQR whenever the hypotheses of Theorem 4.2.26; these, plus the base calls to SmallEig, are the only sources of randomness in the algorithm. There are at most $n^2 \cdot N_{\mathsf{dec}}$ calls each to RitzOrDecouple and $\mathsf{Sh}_{k,B}$ over the course of the algorithm, each failing with probability φ, and at most n calls to SmallEig, each failing with probability at most $\phi/3n$. By a union bound and the definition of φ, the total failure probability is at most ϕ.

Arithmetic Operations and Calls to SmallEig. ShiftedQR recursively runs through line 3 many times in the course of the algorithm; write $T_3(m, k, \delta, B, \Sigma, \Gamma)$ for the arithmetic

operations required to execute this line on some matrix of size $m \times m$ during the algorithm, with the convention that this quantity is zero when $m \leq k$. Then we have

$$\begin{aligned} T_{\mathsf{ShiftedQR}}(n,k,\delta,B,\Sigma,\Gamma) &= T_3(n,k,\delta,B,\Sigma,\Gamma) \\ &\leq T_{\mathsf{RitzOrDecouple}}(n,k,\texttt{true}) \\ &\quad + N_{\mathsf{dec}}\Big(T_{\mathsf{RitzOrDecouple}}(n,k,\texttt{false}) + T_{\mathrm{Sh}}(n,k,\delta,B,\Sigma,\Gamma)\Big) \\ &\quad + T_{\texttt{deflate}}(k) + \max_{\sum_i n_i = n} \sum_i T_3(n_i,k,\delta,B,\Sigma,\Gamma). \end{aligned}$$

Since each of the expressions $T_\square(\cdot)$ is a polynomial of degree at most two in n, the maximum in the third line can be bounded by $T_3(n-1,k,\delta,B,\Sigma,\Gamma)$. Losing only a bit in the constant, we can bound as

$$\begin{aligned} T_{\mathsf{ShiftedQR}}(n,k,\delta,B,\Sigma,\gamma) &\leq n\Big(T_{\mathsf{RitzOrDecouple}}(n,k,\texttt{true}) \\ &\quad + N_{\mathsf{dec}}\Big(T_{\mathsf{RitzOrDecouple}}(n,k,\texttt{false}) + T_{\mathrm{Sh}}(n,k,\delta,B,\Sigma,\Gamma)\Big) \\ &\quad + T_{\texttt{deflate}}(k)\Big) \\ &= O\left(\left(\log\frac{nB\Sigma}{\delta\Gamma}k\log k + k^2\right)n^3\right). \end{aligned}$$

In addition, ShiftedQR requires at most $O(n\log\frac{nB\Sigma}{\delta\Gamma})$ calls to SmallEig with accuracy $\Omega(\epsilon^2/\Sigma)$ and failure probability tolerance φ in the course of the calls to RitzOrDecouple, plus $O(n)$ 'base case' calls with accuracy δ/n and failure probability tolerance $\phi/3n$; the latter calls to SmallEig are asymptotically dominated by the former. The estimates in the theorem statement come from bounding ϵ and φ. □

4.3 Finding Ritz Values

Section 4.3 will be devoted to showing the following theorem.

Theorem 4.3.1. *On any input matrix $A \in \mathbb{C}^{n\times n}$, accuracy parameter $\delta > 0$, and failure probability tolerance $\phi > 0$, the algorithm* SmallEig(A,δ,ϕ) *produces, with probability $1-\phi$, the eigenvalues of a matrix $\widetilde{A} \in \mathbb{C}^{n\times n}$ with $\|A-\widetilde{A}\| \leq \delta\|A\|$, using at most*

$$O\big(n^4 + n^3\log(n/\delta\phi)^2 + \log(n/\delta\phi)^2\log\log(n/\delta\phi)\big)$$

arithmetic operations on a floating point machine with $O(\log(n/\delta\phi)^2)$ bits of precision.

The above theorem shows that, when implemented on a floating point machine with $O(\log(n/\delta\phi)^2)$ bits of precision, the algorithm SmallEig is δ-backward stable. However, as mentioned above, in the context of this analysis the above algorithm will be used to find

forward approximations of the eigenvalues of a (small) matrix, namely the $k \times k$ corner of a Hessenberg matrix. The following result [26, Theorem 39.1] turns any backward error algorithm for the eigenproblem into a forward error algorithm, at the cost of multiplying the number of bits of precision by a roughly n (which might be tolerable for small n, but prohibitively expensive otherwise).

Lemma 4.3.2. *Let $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be any two matrices. Then there are labellings $\lambda_1, ..., \lambda_n$ and $\widetilde{\lambda}_1, ..., \widetilde{\lambda}_n$ of the eigenvalues of A and $\widetilde{A}$, respectively, so that*

$$\max_i |\lambda_i - \widetilde{\lambda}_i| \leq 4(\|A\| + \|\widetilde{A}\|)^{1-1/n}\|A - \widetilde{A}\|^{1/n}.$$

In particular, for every $\beta \leq 1$ one can produce β-forward approximate eigenvalues by calling SmallEig with accuracy

$$\delta = \left(\frac{\beta}{12}\right)^n,$$

as $\|\widetilde{A}\| \leq \|A\| + \delta \leq 2\|A\|$. This yields the following corollary.

Corollary 4.3.3. *On any input matrix $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, any accuracy parameter $\beta > 0$, and failure probability tolerance $\phi > 0$, one can use* SmallEig *to find, with probability $1 - \phi$, approximate eigenvalues $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n \in \mathbb{C}^{n\times n}$ such that*

$$\max_i |\lambda_i - \widetilde{\lambda}_i| \leq \beta\|A\|,$$

using at most

$$O\big(n^5 \log(n/\beta\phi)^2 + n^2 \log(n/\beta\phi)^2 \log(n \log(n/\beta\phi))\big)$$

arithmetic operations on a floating point machine with $O(n^2 \log(n/\beta\phi)^2)$ bits of precision.

And a direct translation of the above corollary gives Theorem 4.3.1 advertised in Section 1.4.

4.3.1 Overview of the Algorithm and Intermediate Results

The main subroutine of SmallEig, which we call findOne, is a form of *shifted* inverse iteration that on a diagonalizable input $A \in \mathbb{C}^{n\times n}$ and an input accuracy parameter $\beta \geq 0$, produces a β-forward approximation $\widetilde{\lambda} \in \mathbb{C}$ of an eigenvalue of A. The precision required to ensure stability of this subroutine and its running time are a function of n and the eigenvector condition number of A, i.e. of

$$\kappa_V(A) := \inf_{V:A=VDV^{-1}} \|V\|\|V^{-1}\|.$$

The shifting strategy in findOne crucially relies on a subroutine distSpec, which allows us to estimate the distance of any given point $s \in \mathbb{C}$ to the spectrum of A (henceforth denoted by

Spec A) up to relative distance 0.1. The subroutine distSpec is in itself a form of *unshifted* inverse iteration on $A - s$ and its required precision and running time are also a function of n and $\kappa_V(A)$.

Once a β-forward approximation $\widetilde{\lambda} \in \mathbb{C}$ of A is obtained, the algorithm calls a subroutine `dec`, which essentially uses inverse iteration on $A - \widetilde{\lambda}$ to find a vector $v \in \mathbb{C}^n$ which is close to the right eigenvector of A associated to the eigenvalue which is closest to $\widetilde{\lambda}$. Then, the subroutine `deflate` is called to reduce the problem A to a smaller instance.

All of the subroutines used in the algorithm require some control on $\kappa_V(A)$, and some additionally require a lower bound on $\mathrm{gap}(A)$. As usual, in order for SmallEig to work on any matrix, we pre-process the input matrix by adding a small random perturbation. We refer the reader to Section 4.3.7 for a detailed discussion on how this works in this context.

Below we elaborate on the main subroutines of SmallEig and discuss the technical results proven in this section.

*Computing the Distance to the Spectrum (*distSpec*).* Let $A \in \mathbb{C}^{n\times n}$ be a diagonalizable matrix with spectral decomposition

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

and fix $s \in \mathbb{C} \setminus \mathrm{Spec}\, A$. The main idea behind distSpec is simple: if $u \in \mathbb{C}^n$ is a vector sampled uniformly at random from the complex unit sphere $\mathbb{S}^{n-1}$ then $\|u^*(s-A)^{-m}\|^{-\frac{1}{m}}$ converges (with probability one) as m goes to infinity, to the distance from s to the spectrum of A, which we will denote by $\mathrm{dist}(s, \mathrm{Spec}\, A)$. Indeed:

$$\begin{aligned}
\lim_{m\to\infty} \|u^*(s-A)^{-m}\|^{-\frac{1}{m}} &= \lim_{m\to\infty} \Big\| \sum_{i=1}^{n} (s-\lambda_i)^{-m} u^* v_i w_i^* \Big\|^{-\frac{1}{m}} \\
&= \lim_{m\to\infty} \mathrm{dist}(s, \mathrm{Spec}\, A) \Big\| \sum_{i=1}^{n} \left(\frac{\mathrm{dist}(s, \mathrm{Spec}\, A)}{s-\lambda_i} \right)^m u^* v_i w_i^* \Big\|^{-\frac{1}{m}} \\
&= \mathrm{dist}(s, \mathrm{Spec}\, A) \qquad (4.58)
\end{aligned}$$

where the last equality holds almost surely. In Section 4.3.4 we will prove a quantitative version of this fact, and show that when $m = \Omega\big(\log(n\kappa_V(A))\big)$ one obtains an approximation of $\mathrm{dist}(s, \mathrm{Spec}\, A)$ up to a relative error of 0.1. We will then conclude that distSpec can be implemented with a running time of at most

$$O(\log(n\kappa_V(A))n^2 + \log(n\kappa_V(A)) \log\log(n\kappa_V(A)))$$

arithmetic operations and prove its backward error guarantees, which depend on $\mathrm{dist}(s, \mathrm{Spec}\, A)$ and $\kappa_V(A)$.

*Finding One Eigenvalue (*findOne*).* With distSpec in hand, findOne generates a sequence of complex numbers $s_0, s_1, \ldots$ that converges linearly to an eigenvalue of A. This sequence

is recursively generated as follows: at time t, the algorithm uses distSpec to compute an estimate $\tau_t \approx \mathsf{dist}(s_t, \operatorname{Spec} A)$ with relative error of at most 0.1. This guarantees that there is at least one eigenvalue of A inside the annulus

$$\mathcal{A}_{s_t,\tau_t} := \{z \in \mathbb{C} : 0.9\tau_t \leq |z - s_t| \leq 1.12\tau_t\}, \tag{4.59}$$

and hence if $\mathcal{N}_{s_t,\tau_t}$ is a fine enough net of $\mathcal{A}_{s_t,\tau_t}$ (we will show that nets of six points suffice), we will be able to guarantee that

$$\min_{s \in \mathcal{N}_{s_t,\tau_t}} \mathsf{dist}(s, \operatorname{Spec} H) \leq 0.6\, \mathsf{dist}(s_t, \operatorname{Spec} H).$$

Given the above guarantee, findOne then uses distSpec again, now to estimate the distances of the points $s \in \mathcal{N}_{s_t,\tau_t}$ to the spectrum of A, and chooses a point $s \in \mathcal{N}$ for which

$$\mathsf{distSpec}(s, \operatorname{Spec} A) \leq \gamma\tau_t$$

for some suitably chosen parameter $\gamma \in (0, 1)$ (we will show that when $\gamma = 0.66$ the above inequality is guaranteed for some point in the net). For such an s, findOne sets $s_{t+1} := s$ and $\tau_{t+1} := \mathsf{distSpec}(s, \operatorname{Spec} A)$, after which the iteration is repeated (see Figure 4.1 for an example).

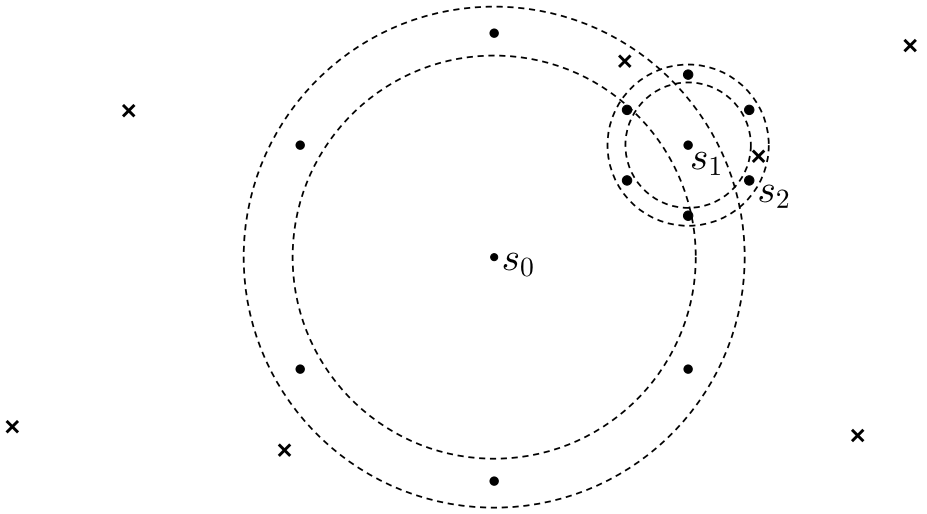

Figure 4.1: The locations of the eigenvalues of A are represented by an $\times$. The figure illustrates the first steps of the iteration which produce s_0, s_1 and s_2. The annuli $\mathcal{A}_{s_0,\tau_0}$ and $\mathcal{A}_{s_1,\tau_1}$ are signaled with dotted lines, and the corresponding nets of six points on them are marked with solid dots.

Clearly, the s_t will converge linearly to an eigenvalue of A and hence finding a point that is at distance at most β from the spectrum of A will take $O(\log(1/\beta))$ calls to distSpec. This will be discussed in detail in Section 4.3.4.

Remark 4.3.4. Intuitively, findOne is a shifting strategy for inverse iteration where each shift is an *exceptional shift* (cf. [61, 167, 12]) chosen from a net of six points.

Remark 4.3.5. Note that even if the subroutine findOne provides a β-*forward* approximation of an eigenvalue of the matrix, the ultimate algorithm SmallEig will only be able to provide an $O(\beta)$-*backward* set of approximate eigenvalues. This is because in order to obtain the full eigendecomposition one needs to *deflate* the problem once a converged eigenvalue is obtained (see the next paragraph for more details on this process), and after deflation we are only able to control the backward error of the eigenvalues that are subsequently obtained.

Implementation of the Subroutines (Tau^m, `dec`, `deflate`*).* There are many ways to implement the subroutines distSpec and findOne described above. In this section, for several reasons, we have decided to operate with matrices in their Hessenberg form (similar to what the shifted QR algorithm does). One of the advantages of doing this is that, in the Hessenberg setting, instead of computing the quantity $\|u^*(s-A)^{-m}\|^{-\frac{1}{m}}$ mentioned in the analysis of distSpec one need to compute

$$\tau_{(z-s)^m}(H) := \|e_n^*(s-H)^{-m}\|^{-\frac{1}{m}}$$

where H is a Hessenberg matrix that is unitarily equivalent to A (or *almost* unitarily equivalent when finite arithmetic is taken into account). Computing the above quantity, as shown in Section 4.2, can be done directly from running the implicit QR algorithm IQR on H (see Section 4.2.4 for a definition of IQR and the subroutine Tau^m defined by it). So in essence, when working with Hessenberg matrices the subroutine distSpec can be easily implemented by calling IQR with a suitable degree.

The second advantage of working with a Hessenberg matrix H is that once a forward approximate eigenvalue $\widetilde{\lambda}$ of H is found (which is the purpose of findOne), reducing the problem H to a smaller instance becomes easier. Indeed, in Section 4.3.5 we will show that if $H_\ell := \mathsf{IQR}(H, (z-\widetilde{\lambda})^\ell)$, then one is guaranteed to have $|(H_\ell)_{n,n-1}| = O(\beta)$ for some $\ell = O(n \log \kappa_V(A))$. This will allow us to decouple and then deflate the problem.

Remark 4.3.6 (Comparison to Shifted QR)**.** One reason why our algorithm is not an actual shifted QR algorithm is that we have chosen to *maintain* the same Hessenberg matrix H throughout the computation of the shifts $s_1, s_2, \ldots$ done by findOne, as opposed to *updating* the Hessenberg matrix in each iteration to produce a sequence of Hessenberg matrices $H_0 = H, H_1, \ldots$ hand in hand with the computation of each s_t (as a standard shifted QR algorithm would do). During this process we are using the Hessenberg structure merely as a device for a fast implementation of inverse iteration, and not any of its more subtle properties as in Sections 4.1 and 4.2. Between calls to findOne the Hessenberg structure is further used to deflate the matrix in a convenient manner.

More substantially, findOne requires as input a Hessenberg matrix whose right eigenvectors all have *reasonably large* (say $1/\mathrm{poly}(n)$) inner products with the vector e_n; this is roughly

because our analysis is based on the power method and not the more sophisticated potential-based arguments of Section 4.1 which require no assumptions whatsoever. We guarantee the inner product condition by computing a Hessenberg form with respect to a *random* vector. Unfortunately this must be redone after each deflation, which inflicts a cost in the running time of $O(n^4)$, as opposed to the $O(n^3)$ achieved by algorithms that do not need to repeatedly recompute the Hessenberg form.

*Randomness in the Algorithm (*RHess, $\mathrm{Unif}(D(0,\eta_2)), G_n$*).* Our algorithm uses randomness in three different ways. The first one is related to the inverse iteration described above when discussing distSpec. In the Hessenberg setting, the equivalent of running inverse iteration on a randomly chosen vector is to compute a *random* Hessenberg matrix H that is unitarily equivalent to the initial matrix A, where the randomness is uniform (in some suitable sense) among the set of Hessenberg matrices that are uniformly equivalent to A. The source of randomness in this case is also a unit vector distributed uniformly on the complex unit sphere $\mathbb{S}_{\mathbb{C}}^{n-1}$. We refer the reader to Section 4.3.3 for the details on the sampling assumptions made here, and to Section 4.3.6 for an analysis of the subroutine RHess which on an input matrix A returns a Hessenberg matrix H chosen at random from the unitary equivalence class (up to machine error) of A.

The second use of randomness is related to the forward stability of $\mathsf{IQR}(H,(z-s)^m)$, which as discussed in Section 4.2, is a function of $\mathrm{dist}(s,\mathrm{Spec}\,H)$. As in Section 4.2, before every call to IQR we will add a small random perturbation to the desired shift s, i.e. we define $\check{s} := s+w$ with w chosen uniformly at random from the disk centered at zero of radius η_2 — henceforth denoted by $w \sim \mathrm{Unif}(D(0,\eta_2))$ — and run $\mathsf{IQR}(H,(z-\check{s})^m)$ instead of $\mathsf{IQR}(H,(z-s)^m)$. The point of doing this is to ensure that with high probability $\mathrm{dist}(\check{s},\mathrm{Spec}\,H) \geq \eta_1$, for some appropriately chosen (as a function of the desired probability) tolerance parameter η_1 that will ultimately determine the precision required for IQR to be numerically forward stable, a necessary condition for our running time guarantees on SmallEig to hold.

Finally, the third way in which we use randomness is to randomly perturb the matrix that is given as input to SmallEig, with the purpose of having high probability upper and lower bounds on κ_V and gap (as explained in Section 1.2) when running the subroutines of SmallEig. For this we assume access to a Gaussian sampler that allows us to generate (once) an $n \times n$ complex Ginibre matrix G_n.

To conclude this section we make some comments about our analysis and presentation.

Pseudospectrum vs gap *and* κ_V*.* Although all of the requirements, actions, and guarantees of the subroutines used by the main algorithm can be phrased in terms of the minimum eigenvalue gap and eigenvector condition number of the matrices in question, in some cases we have decided to instead work with the notion of pseudospectrum. This treatment simplifies the analysis of the effects of roundoff error, since the perturbation theory for the pseudospectrum of a matrix is significantly simpler than that for the eigenvalue gap and eigenvector condition number.

Use of Global Data. As in Section 4.2 we will use the notion of *global data* when presenting the pseudocode of the algorithms. Here, the global data will be composed of four quantities that all of the subroutines can access if needed. More specifically, the global data will be given by n the dimension of the original input matrix, Σ an approximation of the norm of the matrix, and two parameters ϵ and ζ which will be used to control the pseudospectrum.

4.3.2 Related Work and Discussion

Inverse iteration has been used since the 1940's [173] as a method for computing an eigenvector when an approximation of the corresponding eigenvalue is known; a detailed survey of its history and properties may be found in [163, 129, 128, 91]. In contrast, this section uses inverse iteration along with a simple shifting strategy to find the eigenvalues from scratch.

As discussed in the references above, two situations in which the behavior of inverse iteration in finite arithmetic is known to be tricky to analyze are: (1) matrices with tiny eigenvalue gaps (2) nonnormal matrices which exhibit transient behavior. We deal with these issues by assuming *a priori* bounds on the eigenvalue gaps and nonnormality of our input matrix (see Definition 4.3.10) and always dealing with high enough powers of the inverse to dampen transient effects. Assuming such bounds is not restrictive because they may be guaranteed with high probability by adding a small random perturbation, as discussed above.

The algorithm SmallEig presented here is, at the time of writing, one of four known provable algorithms for computing backward approximations of the eigenvalues of an arbitrary complex matrix in floating point arithmetic, along with [3, 11, 13]. The strengths of the algorithm are its simplicity and use of $O(\log^2(n/\delta))$ bits of precision, which is better than [11] but worse than [3] (however [3] has the drawback of running in $O(n^{10}/\delta)$ arithmetic operations). The main weakness of this algorithm compared to [11, 13] is its use of $O(n^4)$ arithmetic operations for repeatedly computing the Hessenberg form. We do not know any example where this recomputation after deflation is actually needed, but are not able to prove that it is not (with high probability). Doing so would entirely remove the $O(n^4)$ factor from the running time in Theorem 4.3.1 and is worthy of further investigation.

4.3.3 Preliminaries

In Section 4.3 many of ingredients from Section 4.2 will be used, and in particular we will maintain the finite precision arithmetic assumptions from Section 4.2.3, and repeatedly use the results and implementations from Section 4.2.4.

In the remaining of Section 4.3 we will use $\mathsf{fl}(*)$ to denote that the expression $*$ is computed in finite arithmetic.

Random Sampling Assumptions

In Section 4.3.1 we enlisted the three different ways in which randomness is used in SmallEig. Here we specify the assumptions we make about the algorithms used to generate the desired

random objects.

Definition 4.3.7 (Efficient Unif($\mathbb{S}_{\mathbb{C}}^{n-1}$) Sampler)**.** An efficient random vector algorithm takes as input a positive integer n and generates a random unit vector $u \in \mathbb{C}^n$ distributed uniformly in the complex unit n-sphere $\mathbb{S}^{n-1}$ and runs in $C_{\mathsf{U}} n$ arithmetic operations, for some universal constant C_{U}.

Definition 4.3.8 (Efficient Unif($D(0,R)$) Sampler)**.** An efficient random perturbation algorithm takes as input an $R > 0$, and generates a random $w \in \mathbb{C}$ distributed uniformly in the disk $D(0,R)$, and runs in C_{D} arithmetic operations, for some universal constant C_{D}.

Definition 4.3.9 (Efficient Ginibre Sampler)**.** An efficient Ginibre sampler takes as input a positive integer n and generates a random matrix $G_n \in \mathbb{C}^{n\times n}$, where the entries of G_n independent centered complex Gaussians of variance $1/n$, and runs in $C_{\mathsf{G}} n^2$ arithmetic operations.

Note that the roundoff error in the algorithm coming from using finite precision when sampling any of these random objects only affects (in a negligible way) the failure probabilities reported in the analysis of the algorithm, and not the quantities handled by the algorithm itself. So, for simplicity we will assume that the samples can be drawn from their exact distribution.

ζ-Shattered Pseudospectra

When analyzing the algorithm in finite arithmetic it will be necessary to have some control on the eigenvector condition number and minimum eigenvalue gap of the matrices produced by the algorithm. For this, we will use the notion of ζ-shattered pseudospectrum, which is very similar to the notion of shattered pseudospectra that we discussed in Section 1.3, but without referencing a grid.

Definition 4.3.10 (ζ-shattered pseudospectrum)**.** Let $\epsilon, \zeta > 0$ and $A \in \mathbb{C}^{n\times n}$. We say that $\Lambda_\epsilon(A)$ is ζ-shattered if there exist n disjoint disks $D_1, \ldots, D_n$ of radius ζ such that

i) (Containment) $\Lambda_\epsilon(A) \subset \bigcup_{i=1}^n D_i$.

ii) (Separation) Any two disks are at distance at least ζ, that is, $\mathsf{dist}(D_i, D_j) \geq \zeta$ for all $i \neq j$.

In what can be thought as a converse of Lemma 1.1.9, the shattering parameter can be used to control the eigenvector condition number of a matrix and its minimum eigenvalue gap.

Lemma 4.3.11 (κ_V and gap from ζ and ϵ)**.** *Let $\epsilon, \zeta > 0$ and $A \in \mathbb{C}^{n\times n}$. If $\Lambda_\epsilon(A)$ is ζ-shattered, then*

i) $\kappa_V(A) \leq \frac{n\zeta}{\epsilon}$.

ii) $\operatorname{gap}(A) \geq \zeta$.

Proof. First note that ii) follows from the fact that $\operatorname{Spec}(A) \subset \Lambda_\epsilon(A)$ and the definition of ζ-shattering. To show i) let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. A trivial modification of the proof of Lemma 3.3.4 yields that $\kappa(\lambda_i) \leq \frac{\zeta}{\epsilon}$. Then, by (1.8) we have

$$\kappa_V(A) \leq \sqrt{n \sum_{i=1}^{n} \kappa(\lambda_i)^2} \leq \frac{n\zeta}{\epsilon}.$$

□

4.3.4 The Shifting Strategy

Analysis of distSpec

We define the subroutine $\mathsf{distSpec}(H, s, m)$ as follows and prove its guarantees below.

distSpec

Input: Hessenberg $H \in \mathbb{C}^{n\times n}$, $s \in \mathbb{C}$, $m \in \mathbb{N}$
Output: $\tau \geq 0$
Ensures: $\frac{0.998}{\kappa_V(H)^{\frac{1}{m}}}\operatorname{dist}(s, \operatorname{Spec} H) \leq \tau \leq \frac{1.003\kappa_V(H)^{\frac{1}{m}}}{\mathbb{P}\left[|Z_H - s| = \operatorname{dist}(s, \operatorname{Spec} H)\right]^{\frac{1}{2m}}}\operatorname{dist}(s, \operatorname{Spec} H)$

1. $\widetilde{\tau^k} \leftarrow \mathsf{Tau}^m(H, (z-s)^m)$
2. $\tau \leftarrow \mathsf{fl}\left((\widetilde{\tau^k})^{\frac{1}{m}}\right)$

Proposition 4.3.12 (Guarantees for distSpec)**.** *Let $C > 0$ and assume that $s \in D(0, C\|H\|)$. Then, the algorithm* distSpec *runs in*

$$T_{\mathsf{distSpec}}(n, m) := T_{\mathsf{Tau}}(n, m) + T_{\mathsf{root}}(m, 10^{-3}) = O(mn^2 + m \log m)$$

arithmetic operations and satisfies its guarantees provided that

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{\mathsf{distSpec}}(n, m, C, \|H\|, \kappa_V(H), \mathsf{dist}(s, \operatorname{Spec} H)) \\ &:= \frac{1}{c_{\mathsf{root}}}\mathbf{u}_{\mathsf{Tau}}(n, m, C, \|H\|, \kappa_V(H), \mathsf{dist}(s, \operatorname{Spec} H)). \end{aligned} \tag{4.60}$$

Proof. First note that

$$\tau_{(z-s)^m}(H) = \|e_n^*(H-s)^{-m}\|^{-\frac{1}{m}}$$

$$\leq \frac{\kappa_V(H)^{\frac{1}{m}}}{\mathbb{E}\left[|Z_H - r|^{-2m}\right]^{\frac{1}{2m}}} \qquad \text{Lemma 4.1.5}$$

$$\leq \frac{\kappa_V(H)^{\frac{1}{m}} \operatorname{dist}(r, \operatorname{Spec} H)}{\mathbb{P}\Big[|Z_H - r| = \operatorname{dist}(r, \operatorname{Spec} H)\Big]^{\frac{1}{2m}}}. \tag{4.61}$$

Similarly, to lower bound $\tau_{(z-s)^m}(H)$ use Lemma 4.1.5 again to obtain

$$\tau_{(z-s)^m}(H) = \|e_n^*(H-s)^{-m}\|^{-1/m} \geq \frac{1}{\kappa_V(H)^{\frac{1}{m}}\mathbb{E}[|Z_H - s|^{-2m}]^{\frac{1}{2m}}} \geq \frac{\operatorname{dist}(s, \operatorname{Spec} H)}{\kappa_V(H)^{\frac{1}{m}}}.$$

So, it only remains to control $|\tau - \tau_{(z-s)^m}(H)|$, where τ is the output of $\mathsf{distSpec}$. Since by assumption (4.60) holds, we can apply Lemma 4.2.12 to get

$$0.999\tau_{(z-s)^m}(H)^m \leq \widetilde{\tau^k} \leq 1.001\tau_{(z-s)^m}(H)^m.$$

Similarly, we can apply Lemma 4.2.2 to get that $\mathsf{fl}((\widetilde{\tau^k})^{\frac{1}{m}})$ can be computed to relative accuracy $\epsilon = 10^{-3}$, using at most $T_{\mathsf{root}}(m, 10^{-3})$ arithmetic operations. Hence

$$0.999(\widetilde{\tau^k})^{\frac{1}{m}} \leq \mathsf{fl}((\widetilde{\tau^k})^{\frac{1}{m}}) \leq 1.001(\widetilde{\tau^k})^{\frac{1}{m}},$$

which combined with all of the above yields the advertised guarantees. To compute the final running time, add to $T_{\mathsf{root}}(m, 10^{-3})$ the $T_{\mathsf{Tau}}(n, m)$ arithmetic operations needed to compute Tau^m. □

Analysis of findOne

For every $s \in \mathbb{C}$ and $\tau > 0$, on the annulus $\mathcal{A}_{s,\tau} = \{z \in \mathbb{C} : 0.9\tau \leq |z - s| \leq 1.12\tau\}$ we will define the set $\mathcal{N}_{s,\tau}$ of six points given by

$$\mathcal{N}_{s,\tau} := \left\{s + \tau e^{i\pi\ell/3} : \ell = 1, \ldots, 6\right\}.$$

As explained in Section 4.3.1, at time t, $\mathsf{findOne}$ will call $\mathsf{distSpec}$ on the the locations given by the points in a net on $\mathcal{A}_{s_t,\tau_t}$ for some s_t and τ_t. So, to give accuracy guarantees on the output provided by $\mathsf{distSpec}$, we will choose the net to be the randomly perturbed set

$$\check{\mathcal{N}}_{s_t,\tau_t} := \{s_t + w : s_t \in \mathcal{N}_{s_t,\tau_t}\}, \quad \text{where} \quad w \sim \operatorname{Unif}(D(0, \eta_2)),$$

(cf. the discussion on shift regularization in Section 4.2.4).

We begin by noting that for any $s \in \mathbb{C}$ and $\tau > 0$, $\check{\mathcal{N}}_{s,\tau}$ is a net on $\mathcal{A}_{s,\tau}$ in the following sense.

Observation 4.3.13. Using the above notation, if $\eta_2 \leq .03\tau$ then for any realization of $\check{\mathcal{N}}_{s,\tau}$ we have

$$\sup_{z \in \mathcal{A}_{s,\tau}} \operatorname{dist}\big(z, \check{\mathcal{N}}_{s,\tau}\big) \leq 0.6\tau.$$

Proof. Basic trigonometry shows that because $z \in \mathcal{A}_{s,\tau}$ we can guarantee $\operatorname{dist}(z, \mathcal{N}_{s,\tau}) \leq .57\tau$. Then, because any realization of $w \sim D(0, \eta_2)$ (which yields a realization of $\check{\mathcal{N}}_{s,\tau}$) satisfies $|w| \leq \eta_2 \leq .03\tau$, the result follows from the triangle inequality. □

We can now define the algorithm.

findOne

Input: $H \in \mathbb{C}^{n\times n}$ Hessenberg, accuracy $\beta > 0$, failure probability tolerance φ, eigenvalue mass lower bound p
Global Data: Norm bound Σ, pseudospectral parameter ϵ, shattering parameter ζ
Output: $[\widetilde{\lambda}, \mathsf{correctness}]$ with $\widetilde{\lambda} \in \mathbb{C}$ and $\mathsf{correctness} \in \{\mathtt{true}, \mathtt{false}\}$
Requires: $\beta \leq 1/2$, $\Lambda_\epsilon(H)$ is ζ-shattered, $\mathbb{P}[Z_H = \lambda] \geq p$ for all $\lambda \in \operatorname{Spec} H$, $10\beta \leq \|H\| \leq 2\Sigma$
Ensures: With probability at least $1 - \varphi$, findOne terminates successfully, that is $\mathsf{correctness} = \mathtt{true}$ and $\widetilde{\lambda}$ satisfies $\eta_1 \leq \operatorname{dist}(\widetilde{\lambda}, \operatorname{Spec} H) \leq \beta$, where η_1 is defined in line 1

1. $m \leftarrow \left\lceil 12\left(\log\left(\frac{n\zeta}{\epsilon}\right) + \frac{1}{2}\log\left(\frac{1}{p}\right)\right)\right\rceil$, $\eta_2 \leftarrow \frac{\beta}{5} \wedge \frac{\zeta}{3}$, $\eta_1 \leftarrow \eta_2\left(\frac{\varphi}{12\log(3\Sigma/10\beta)}\right)^{1/2}$
2. $w \sim \operatorname{Unif}(D(0,\eta_2))$, $\check{s} \leftarrow H_{nn} + w$, $\tau \leftarrow \mathsf{distSpec}(\check{s}, H, m)$
3. **While** $\tau > 0.9\beta$
   a) $w \sim \operatorname{Unif}(D(0,\eta_2))$, $\check{\mathcal{N}} \leftarrow \{\check{s}^{(1)}, \ldots, \check{s}^{(6)}\} = \mathcal{N}_{\check{s},\tau} + w$
   b) $\tau' \leftarrow \min_{j\in[6]} \mathsf{distSpec}(\check{s}^{(j)}, H, m)$
   c) **If** $\tau' \leq 0.66\tau$
   $\check{s} \leftarrow \check{s}^{(j)}$, $\tau \leftarrow \tau'$, $\mathsf{correctness} \leftarrow \mathtt{true}$
   d) **Else** $\mathsf{correctness} \leftarrow \mathtt{false}$, terminate findOne and output $[\check{s}, \mathtt{false}]$.
4. $\widetilde{\lambda} \leftarrow \check{s}$, output $[\widetilde{\lambda}, \mathtt{true}]$

Remark 4.3.14 (About the correctness Flag)**.** Although small, there is a positive probability that while running findOne the subroutine distSpec is called on a complex number $s \in \mathbb{C}$ for which $\operatorname{dist}(s, \operatorname{Spec} H) < \eta_1$. When this happens there will be no guarantee that the output of distSpec is relatively accurate, and the information provided by it might be misleading, giving rise to an update of $\check{s}$ for which the distance to $\operatorname{Spec} H$ might be even larger than what it was for its previous value. In view of this, the purpose of the flag correctness is to identify when as a consequence of an inaccurate output of distSpec it is no longer possible to decrease the variable τ at a geometric rate, in which case the algorithm halts and outputs error[6].

[6]Of course, one could try to formulate a dichotomy as in Section 4.2 in which one leverages that errors can only be made once the shifts that are being used are very close to $\operatorname{Spec} H$, and have a mechanism that

Before proving the main result about findOne, we observe that in line 1 of this algorithm, m is set so that $\mathsf{distSpec}(s, H, m)$ will yield an accurate approximation of $\operatorname{dist}(s, \operatorname{Spec} H)$ all throughout the iteration (provided that s is not too close to $\operatorname{Spec} H$).

Observation 4.3.15 (m is large enough)**.** Let $C > 0$, $s \in D(0, C\|H\|)$ and m be as in line 1 of findOne. Assume that the requirements of findOne are satisfied and that

$$\mathbf{u} \leq \mathbf{u}_{\mathsf{distSpec}}(n, m, C, \|H\|, \kappa_V(H), \mathsf{dist}(s, \operatorname{Spec} H)). \tag{4.62}$$

Then

$$0.9\operatorname{dist}(s, \operatorname{Spec} H) \leq \mathsf{distSpec}(H, s, m) \leq 1.1\operatorname{dist}(s, \operatorname{Spec} H).$$

Proof. Let $\tau = \mathsf{distSpec}(H, s, m)$. Since $\mathbf{u} \leq \mathbf{u}_{\mathsf{distSpec}}$ we can apply Proposition 4.3.12 to get

$$\frac{0.998}{\kappa_V(H)^{\frac{1}{m}}}\operatorname{dist}(s, \operatorname{Spec} H) \leq \tau \leq \frac{1.003\kappa_V(H)^{\frac{1}{m}}\operatorname{dist}(s, \operatorname{Spec} H)}{\mathbb{P}\big[|Z_H - s| = \operatorname{dist}(s, \operatorname{Spec} H)\big]^{\frac{1}{2m}}}.$$

Then, it suffices to show that

$$0.9 \leq \frac{0.998}{\kappa_V(H)^{\frac{1}{m}}} \quad \text{and} \quad \frac{1.003\kappa_V(H)^{\frac{1}{m}}}{\mathbb{P}\big[|Z_H - s| = \operatorname{dist}(s, \operatorname{Spec} H)\big]^{\frac{1}{2m}}} \leq 1.1,$$

or equivalently

$$m \geq \frac{\log(\kappa_V(H))}{\log\left(0.998/0.9\right)} \quad \text{and} \quad m \geq \frac{\log(\kappa_V(H)) + \frac{1}{2}\log(1/\mathbb{P}\big[|Z_H - s| = \operatorname{dist}(s, \operatorname{Spec} H)\big])}{\log(1.1/1.003)}.$$

Finally, using that

$$\mathbb{P}\big[|Z_H - s| = \operatorname{dist}(s, \operatorname{Spec} H)\big] \geq \min_{\lambda \in \operatorname{Spec} H} \mathbb{P}[Z_H = \lambda] \geq p$$

and $\kappa_V(H) \leq \frac{n\zeta}{\epsilon}$ (which follows from Lemma 4.3.11), it is clear that this m satisfies the above inequalities. □

Now we observe that in line 1 of findOne, the parameters η_1 and η_2 are set to be small enough that we can apply Lemma 4.2.13.

Observation 4.3.16. Let η_1, η_2 be as in line 1 and assume that the requirements of findOne are satisfied. Then

$$\eta_1 + \eta_2 \leq \frac{\operatorname{gap}(H)}{2} \quad \text{and} \quad \eta_2 \leq 0.02\|H\|.$$

outputs a forward approximate eigenvalue even when distSpec provides inaccurate answers. Since this proved to be intricate, for the sake of clarity we have decided to settle for this simpler, but efficient enough, version of the algorithm.

Proof. Since $\Lambda_\epsilon(H)$ is ζ-shattered we have $\zeta \leq \operatorname{gap}(H)$, and by definition of the parameters we have $2\eta_1 \leq \eta_2 \leq \zeta/3$, from where $\eta_1 + \eta_2 \leq \operatorname{gap}(H)/2$. To prove the other assertion, note that the requirements of findOne imply that $\beta \leq 0.1\|H\|$, on the other hand by definition $\eta_1 \leq \beta/5$, so the proof is concluded by combining both bounds. □

We now state the main result of this section.

Proposition 4.3.17 (Guarantees for findOne)**.** *Assume that the requirements of* findOne *are satisfied, let m and η_1 be as defined in line 1 of* findOne *and assume that*

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{\mathsf{findOne}}(n, \Sigma, \epsilon, \zeta, p, \beta, \varphi) \\ &:= \mathbf{u}_{\mathsf{distSpec}}\big(n, m, 10, 2\Sigma, n\zeta/\epsilon, \eta_1\big). \end{aligned} \tag{4.63}$$

Then, with probability at least $1 - \varphi$, findOne *outputs a $\widetilde{\lambda} \in \mathbb{C}$ satisfying*

$$\eta_1 \leq \operatorname{dist}(\widetilde{\lambda}, \operatorname{Spec} H) \leq \beta, \tag{4.64}$$

using at most

$$\begin{aligned} &T_{\mathsf{findOne}}(n, \Sigma, \epsilon, \zeta, p, \beta) \\ :=&(6\lceil 2\log(\Sigma/5\beta)\rceil + 1)T_{\mathsf{distSpec}}(n, m) + \lceil 2\log(\Sigma/5\beta)\rceil(C_{\mathsf{D}} + 16) + O(1) \\ =&O\big(\log(\Sigma/\beta)\log(n\zeta/\epsilon p)(n^2 + \log\log(n\zeta/\epsilon p))\big) \end{aligned}$$

arithmetic operations.

Proving Proposition 4.3.17. It is clear that the exact arithmetic version of findOne would satisfy the advertised guarantees. The challenge is in arguing that in finite arithmetic, with high probability, each call to distSpec yields an accurate enough answer, and that the aggregate roundoff errors and failure probabilities is not too large. Since distSpec is based on the subroutine IQR, inaccuracies can only arise when the input $s \in \mathbb{C}$ is either too close to $\operatorname{Spec} H$ or $|s|$ is too large. This is quantified in the following observation, which we will use repeatedly throughout the proof.

Observation 4.3.18 (Conditions for accuracy)**.** For any $s \in D(0, 10\|H\|)$ with

$$\operatorname{dist}(s, \operatorname{Spec} H) \geq \eta_1$$

the following guarantee holds

$$0.9\operatorname{dist}(s, \operatorname{Spec} H) \leq \mathsf{distSpec}(H, s, m) \leq 1.1\operatorname{dist}(s, \operatorname{Spec} H).$$

Proof. Since $\Lambda_\epsilon(H)$ is ζ-shattered by assumption, Lemma 4.3.11 shows that $\kappa_V(H) \leq \frac{n\zeta}{\epsilon}$, and using the assumption $\|H\| \leq 2\Sigma$, we get that (4.63) implies

$$\mathbf{u} \leq \mathbf{u}_{\mathsf{distSpec}}\Big(n, m, 10, \|H\|, \kappa_V(H), \eta_1\Big).$$

So, for any $s \in D(0, 10\|H\|)$ with $\operatorname{dist}(s, \operatorname{Spec} H) \geq \eta_1$, $\mathbf{u}$ will satisfy inequality (4.62), which by Observation 4.3.15 yields the desired inequalities. □

Let $s_0, s_1, \dots$ be the values acquired by the variable $\check{s}$ throughout the algorithm, $\tau_0, \tau_1, \dots$ be the values acquired by τ, and $w_0, w_1, \dots$ be the values acquired by w. We will now show that, by the structure of the algorithm, the only real obstruction to obtaining accuracy is the possibility of the s_i being to close to $\operatorname{Spec} H$.

Lemma 4.3.19 (Accuracy of the τ_i)**.** *Let $t \geq 0$ and assume that* findOne *does not terminate in the first t while loops*[7]*, and that $\operatorname{dist}(s_i, \operatorname{Spec} H) \geq \eta_1$ for all $i = 0, \dots, t$. Then, for all $i = 0, \dots, t$ we have that*

$$0.9\operatorname{dist}(s_i, \operatorname{Spec} H) \leq \tau_i \leq 1.1\operatorname{dist}(s_i, \operatorname{Spec} H), \tag{4.65}$$

$s_i \in D(0, 10\|H\|)$, and moreover $\check{\mathcal{N}}_{s_i,\tau_i} \subset D(0, 10\|H\|)$.

Proof. We proceed by induction. First we will prove the statement for $t = 0$. In this case, because of the way $\check{s}$ is initialized (see line 2 of findOne), $s_0 = H_{nn} + w_0$ for $w_0 \sim D(0, \eta_2)$. So, by definition, $|s_0| \leq \|H\| + \eta_2$, and by Observation 4.3.16 we have $s_0 \in D(0, C\|H\|)$ for $C = 1.02$. It follows, by Observation 4.3.18, that τ_0 satisfies the inequalities in (4.65). Therefore

$$\tau_0 \leq 1.1\operatorname{dist}(s_0, \operatorname{Spec} H) \leq 1.1 \cdot 2.02\|H\| \leq 2.3\|H\|$$

which we record for later use.

Now take $k \leq t$ and assume that (4.65) holds for $i = 0, \dots, k$, we will then show that it also holds for $k + 1$. First note that by the assumption that findOne does not terminate in the first t while loops, we have that $\tau_{i+1} \leq .66\tau_i$ and $.9\beta \leq \tau_i$ for all $i = 0, \dots, k$. Hence, by construction of the sequence $s_0, s_1, \dots,$ for any $s \in \check{\mathcal{N}}_{s_k,\tau_k}$ we can obtain

$$\begin{aligned}
|s| &\leq |s_0| + |s_1 - s_0| + \cdots + |s_k - s_{k-1}| + |s - s_k| \\
&\leq |s_0| + \tau_0 + |w_1| + \cdots + \tau_k + |w_{k+1}| && \text{since } s_{i+1} \in \check{\mathcal{N}}_{s_i,\tau_i},\ s \in \check{\mathcal{N}}_{s_k,\tau_k} \\
&\leq |s_0| + 1.3(\tau_0 + \cdots + \tau_k) && \tau_i \geq 0.9\beta \text{ and } \eta_2 \leq \frac{\beta}{5} \\
&\leq |s_0| + 1.3 \cdot 2.3\|H\|(1 + 0.66 + 0.66^2 + \cdots) && \tau_{i+1} \leq 0.66^i\tau_0 \leq 0.66^i 2.3\|H\| \\
&\leq |s_0| + 8.8\|H\| \\
&\leq 10\|H\| && |s_0| \leq 1.02\|H\|.
\end{aligned}$$

This proves that $\check{\mathcal{N}}_{s_k,\tau_k} \subset D(0, 10\|H\|)$. So, when $k \leq t-1$ we get get that $s_{k+1} \in D(0, 10\|H\|)$, and because we also know that $\operatorname{dist}(s_{k+1}, \operatorname{Spec} H) \geq \eta_1$, we can apply Observation 4.3.18 to show that (4.65) holds for $i = k + 1$. □

In the above lemma we assumed that findOne did not terminate in the first t calls to the while loop, which tacitly assumes that the the flag correctness was set back to `true` in each of those loops. We now show that if τ_t is sufficiently accurate and the elements in $\check{\mathcal{N}}_{s_t,\tau_t}$ are far enough from $\operatorname{Spec} H$, then there is a guarantee that in the while loop $t + 1$ the flag correctness will be set back to `true`.

[7]Here, terminating in the while loop $t = 0$ means that that the first while loop was never started.

Lemma 4.3.20 (Guaranteeing correctness = `true`)**.** *Assume that* $\operatorname{dist}(s_i, \operatorname{Spec} H)$ *for* $i = 1, \dots, t$ *and moreover that each* $s \in \check{\mathcal{N}}_{s_t,\tau_t}$ *satisfies that* $\operatorname{dist}(s, \operatorname{Spec} H) \geq \eta_1$*. Then*

$$\min_{s \in \check{\mathcal{N}}_{s_t,\tau_t}} \mathsf{distSpec}(s, H, m) \leq .66\tau_t,$$

where m *is defined as in line 1 of* findOne*.*

Proof. Because τ_t satisfies (4.65) we know that there is at least one eigenvalue of H in $\mathcal{A}_{s_t,\tau_t}$. By Observation 4.3.13 there is at least one $s \in \check{\mathcal{N}}_{s_{t+1},\tau_{t+1}}$ for which $\operatorname{dist}(s, \operatorname{Spec} H) \leq 0.6\tau_t$. Moreover, by assumption, for such s we know that $\operatorname{dist}(s, \operatorname{Spec} H) \geq \eta_1$, and by Lemma 4.3.19 we also know that $s \in D(0, 10\|H\|)$. Hence Observation 4.3.18 implies that

$$\mathsf{distSpec}(H, s, m) \leq 1.1\operatorname{dist}(s, \operatorname{Spec} H) \leq 0.66\tau_t,$$

as we wanted to show. □

Lemmas 4.3.19 and 4.3.20 imply that as long as all of the values of $\check{s}$ and $\check{s}^{(j)}$ for $j = 1, \dots, 6$ satisfy that $\operatorname{dist}(\check{s}, \operatorname{Spec} H) \geq \eta_1$ and $\operatorname{dist}(\check{s}^{(j)}, \operatorname{Spec} H) \geq \eta_1$, we will have accurate τ_i and the flag correctness will always be set back to `true`. We can now conclude the proof.

Probability of success. Take $t = \lceil 2\log(\Sigma/5\beta) \rceil$, which is set so that $4.6 \cdot 0.66^t/0.9 \leq \beta/\Sigma$.

For $i = 1, \dots, t$ and $j = 1, \dots, 6$ let $s_i^{(j)}$ be the value acquired by the variable $\check{s}^{(j)}$ during the while loop i. Using Lemma 4.2.13 and taking a union bound we have that the probability that

$$\operatorname{dist}(s_0, \operatorname{Spec} H) \geq \eta_1 \quad \text{and} \quad \operatorname{dist}(s_i^{(j)}, \operatorname{Spec} H) \geq \eta_1, \quad \forall i \in [t]\, \forall j \in [6]$$

is at least $1 - (6t+1)(\eta_1/\eta_2)^2$. And from the above discussion we know that under this event findOne will not terminate in the first t while loops with correctness = `false`, and moreover $\tau_0 \leq 2.3\|H\|$ and $\tau_{i+1} \leq .66\tau_i$. Therefore, because $\|H\| \leq 2\Sigma$ and the way we have chosen t,

$$\tau_t \leq 0.66^t\tau_0 \leq 0.66^t 2.3\|H\| \leq 0.66^t \cdot 4.6\Sigma \leq 0.9\beta.$$

This ensures that the algorithm terminates with correctness = `true` sometime in the first t while loops with probability at least $1 - (6t+1)(\eta_1/\eta_2)^2$. Moreover, when it terminates, say at time t_0, we are guaranteed that $\operatorname{dist}(s_{t_0}, \operatorname{Spec} H) \geq \eta_1$, and because τ_{t_0} is accurate we have that

$$.9\operatorname{dist}(s_{t_0}, \operatorname{Spec} H) \leq \tau_{t_0} \leq .9\beta,$$

which implies that $\operatorname{dist}(s_{t_0}, \operatorname{Spec} H) \leq \beta$.

On the other hand

$$(6t+1)(\eta_1/\eta_2)^2 = (6\lceil 2\log(\Sigma/5\beta)\rceil + 1)(\eta_1/\eta_2)^2 \leq 12\log(3\Sigma/10\beta)(\eta_1/\eta_2)^2 = \varphi,$$

that is, the failure probability is upper bounded by φ.

Running time. Finally, we give an upper bound on the running time. First note that each iteration of the while loop calls distSpec six times, draws one sample from $\mathrm{Unif}(D(0,\eta_2))$, and at most other 16 arithmetic operations are done. Since, in the successful event, there are at most $\lceil 2\log(\Sigma/5\beta)\rceil$ while loops, this gives us the count of

$$\lceil 2\log(\Sigma/5\beta)\rceil(T_{\mathsf{distSpec}}(n,m)+C_{\mathsf{D}}+16).$$

Before the while loops distSpec is called once, and other than that at most $O(1)$ operations are done. This yields the advertised result.

4.3.5 Decoupling via Inverse Iteration

The following results are the basis of the subroutine we use to decouple a Hessenberg matrix once a forward approximate eigenvalue of H is obtained.

Lemma 4.3.21 (Decoupling in Exact Arithmetic)**.** *Let $s\in\mathbb{C}$ and $H\in\mathbb{C}^{n\times n}$ be a Hessenberg matrix. Consider the sequence given by $H_0 := H$ and $H_{\ell+1} := R_\ell Q_\ell + s$ for $[Q_\ell, R_\ell] := \mathrm{QR}(H_\ell - s)$. Then, for any $m\geq 1$ there is some $1\leq \ell\leq m$ for which*

$$|(H_\ell)_{n,n-1}| \leq \frac{\kappa_V(H)^{\frac{1}{m}}\operatorname{dist}(s,\operatorname{Spec}H)}{\mathbb{P}\Big[|Z_H - s| = \operatorname{dist}(s,\operatorname{Spec}H)\Big]^{\frac{1}{2m}}}. \tag{4.66}$$

Proof. Because by definition: R_ℓ is upper triangular, all the entries of Q_ℓ are bounded by 1, and $H_{\ell+1} = R_\ell Q_\ell + s$, we know that

$$|(H_{\ell+1})_{n,n-1}| \leq |(R_\ell)_{n,n}|. \tag{4.67}$$

On the other hand

$$\begin{aligned}
|(R_0)_{n,n}\cdots(R_{m-1})_{n,n}|^{\frac{1}{m}} &= \|e_n^*(H-s)^{-m}\|^{-\frac{1}{m}} && \text{by (4.16)}\\
&\leq \frac{\kappa_V(H)^{\frac{1}{m}}}{\mathbb{E}\left[|Z_H-s|^{-2m}\right]^{\frac{1}{2m}}} && \text{Lemma 4.1.5}\\
&\leq \frac{\kappa_V(H)^{\frac{1}{m}}\operatorname{dist}(s,\operatorname{Spec}H)}{\mathbb{P}\Big[|Z_H - s| = \operatorname{dist}(s,\operatorname{Spec}H)\Big]^{\frac{1}{2m}}}. && (4.68)
\end{aligned}$$

So, combining (4.67) and (4.68) we get that (4.66) holds for some $1\leq\ell\leq m$. □

Using the forward error guarantees for IQR given in Lemma 4.2.11 we can easily get a finite arithmetic version of the above result.

Lemma 4.3.22 (Decoupling in Finite Arithmetic)**.** *Let $H \in \mathbb{C}^{n\times n}$ be a Hessenberg matrix and $s \in D(0, C\|H\|)$. For every ℓ define $\widetilde{H_\ell} = \mathsf{IQR}(H, (z-s)^\ell)$. Then, for each $m \geq 1$, if*

$$\mathbf{u} \leq \min_{\ell\in[m]} \mathbf{u}_{\mathsf{IQR}}\big(n, \ell, \|H\|, \kappa_V(H), \mathsf{dist}(s, \mathrm{Spec}H)\big) \tag{4.69}$$

there is some $\ell \in [m]$ for which

$$\begin{aligned}&|(\widetilde{H_\ell})_{n,n-1}|\\ \leq&\frac{\kappa_V(H)^{\frac{1}{m}}\mathrm{dist}(s,\mathrm{Spec}\,H)}{\mathbb{P}\Big[|Z_H - s| = \mathrm{dist}(s,\mathrm{Spec}\,H)\Big]^{\frac{1}{2m}}} + 32\kappa_V(H)\|H\|\left(\frac{(2+2C)\|H\|}{\mathsf{dist}(s,\mathrm{Spec}H)}\right)^\ell n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u}.\end{aligned}$$

Proof. Let $H_0, \ldots, H_m$ be as in the statement of Lemma 4.3.21, and let $\ell \in [m]$ be such that (4.66) holds. Now, (4.69) ensures that we can apply Lemma 4.2.11 for the ℓ we have specified, yielding

$$\Big|(H_\ell)_{n,n-1} - (\widetilde{H_\ell})_{n,n-1}\Big| \leq \Big\|H_\ell - \widetilde{H_\ell}\Big\|_F \leq 32\kappa_V(H)\|H\|\left(\frac{(2+2C)\|H\|}{\mathsf{dist}(s,\mathrm{Spec}H)}\right)^\ell n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u}.$$

Combining this with (4.66) the advertised bound follows. □

Analysis of `dec`

In view of the above results we define the subroutine `dec` as follows.

`dec`

Input: Hessenberg $H \in \mathbb{C}^{n\times n}$, $\widetilde{\lambda} \in \mathbb{C}$, and decoupling parameter $\epsilon > 0$
Output: $H \in \mathbb{C}^{n\times n}$ Hessenberg matrix
Requires: $0 < \mathrm{dist}(\widetilde{\lambda}, \mathrm{Spec}\,H) \leq \omega/2$
Ensures: $|H_{n,n-1}| \leq \epsilon$ and there exists a unitary Q with $\|\hat{H} - Q^*HQ\| \leq 3.5m\|H\|\nu_{\mathsf{IQR}}(n)\mathbf{u}$, for m defined as in the statement of Proposition 4.3.23

1. $\hat{H} \leftarrow H$
2. **While** $|H_{n,n-1}| > \omega$
   (i) $H \leftarrow \mathsf{IQR}(H, z - \widetilde{\lambda})$
3. Output $\hat{H}$

Proposition 4.3.23 (Guarantees for `dec`)**.** *Assume that the requirements of `dec` are satisfied, that H is diagonalizable, and that $d := \mathrm{dist}(\widetilde{\lambda}, \mathrm{Spec}\,H)$ and $p := \mathbb{P}\big[|Z_H - \widetilde{\lambda}| = d\big]$ are positive. If*

$$\mathbf{u} \leq \mathbf{u}_{dec}\big(n, \|H\|, \kappa_V(H), p, d\big) \tag{4.70}$$

$$:= \frac{\mathbf{u}_{\mathsf{IQR}}(n, m, \|H\|, \kappa_V(H), d)\omega}{16 \cdot 5^m \cdot n^{1/2}\|H\|},$$

for $m = \left\lceil \frac{\log(\kappa_V(H)^2/p)}{2\log(3\omega/4d)} \right\rceil$, *then* `dec` *satisfies its guarantees and halts after at most* m *calls to* IQR. *Hence, it runs in at most*

$$T_{\texttt{dec}}(n, \kappa_V(H), p, d) := mT_{\mathsf{IQR}}(n, m) = O\left(\log(\kappa_V(H)/p)^2 n^2\right)$$

arithmetic operations.

Proof. First, if $\omega \geq \|H\|$ the while loop in line 2 terminates immediately and `dec` satisfies its guarantees after one arithmetic operation. Hence, we can assume $\omega \leq \|H\|$, which combined with the assumption $d \leq \omega/2$ gives $d \leq \|H\|/2$ and $\widetilde{\lambda} \in D(0, 1.5\|H\|)$.

Now, for every ℓ define $\widetilde{H_\ell} := \mathsf{IQR}\big(H, (z - \widetilde{\lambda})^\ell\big)$, and note that (4.70) implies that

$$\mathbf{u} \leq \mathbf{u}_{\mathsf{IQR}}(n, m, \|H\|, \kappa_V(H), d) = \min_{\ell \in [m]} \mathbf{u}_{\mathsf{IQR}}\big(n, \ell, \|H\|, \kappa_V(H), d\big),$$

where the last equality follows from $d \leq \|H\|/2$. Therefore, we can apply Lemma 4.3.22 to get that there is some $\ell \in [m]$ for which

$$|(\widetilde{H_\ell})_{n,n-1}| \leq \left(\frac{\kappa_V(H)^2}{p}\right)^{\frac{1}{2m}} d + 32\kappa_V(H)\|H\| \left(\frac{5\|H\|}{d}\right)^\ell n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u}.$$

Now, by our choice of m we have that

$$\left(\frac{\kappa_V(H)^2}{p}\right)^{\frac{1}{2m}} d \leq \frac{3\omega}{4},$$

and by (4.70), because $\ell \leq m$ and $\omega \leq \|H\|$, we have that

$$32\kappa_V(H)\|H\| \left(\frac{5\|H\|}{d}\right)^\ell n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq \frac{\omega}{4}.$$

Combining the above inequalities we get that $|(\widetilde{H_\ell})_{n,n-1}| \leq \omega$ as we wanted to show. To prove the remaining claim use again that $\widetilde{\lambda} \in D(0, C\|H\|)$ for $C = 1.5$, and apply Lemma 4.2.9 to get that there is a unitary Q for which

$$\|\widetilde{H_\ell} - Q^* H Q\| \leq 1.4\ell(1 + C)\|H\|\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq 3.5m\|H\|\nu_{\mathsf{IQR}}(n)\mathbf{u},$$

as we wanted to show. □

4.3.6 Randomized Hessenberg Form

Some of the most common and well understood subroutines in numerical linear algebra are those used to put an arbitrary matrix $A \in \mathbb{C}^{n\times n}$ into a Hessenberg form H (e.g. see [55, 85]). The only reason why we have decided to include this section is that we were not able to find in the literature a rigorous result about the effect of randomizing the Hessenberg form H that could allow us to conclude an explicit probabilistic lower bound on $\min_{\lambda\in \mathrm{Spec}(H)} \mathbb{P}[Z_H = \lambda]$. Here, in our analysis we assume access to a deterministic algorithm that uses Householder reflectors to obtain the Hessenberg form (see Definition 4.3.24 below for details), and to a random unit vector generator satisfying the assumptions from Definition 4.3.7 above.

Householder Reflectors

Computing Householder reflectors is essential to many numerical linear algebra algorithms and a thorough analysis of the numerical errors involved can be found in [85, Section 19.3]. In short, Householder reflectors are matrices $P \in \mathbb{C}^{n\times n}$ of the form $P = I - \beta vv^*$ with $v \in \mathbb{C}^n \setminus \{0\}$ and $\beta := \frac{2}{v^*v}$.[8] In practice, given v, instead of computing P it is more convenient to simply store v, which for any vector x allows to compute Px by just computing $x - \beta(v^*x)v$ and this takes

$$T_{\mathsf{hous}}(n) = O(n)$$

arithmetic operations.

With this in mind, given $x, v \in \mathbb{C}^n$ we will use $\mathsf{hous}(v, x)$ to denote the finite arithmetic computation of Px following the procedure outlined above. Similarly, given $A \in \mathbb{C}^{n\times n}$ we will use $\mathsf{hous}(v, A)$ to denote the finite arithmetic computation of PA, where the i-th column of PA is computed as $\mathsf{hous}(v, A^{(i)})$ where $A^{(i)}$ denotes the i-th column of A.

In [85, Lemma 19.2] it was shown that there exists a small universal constant c_{h} for which, provided that $c_{\mathsf{h}} n\mathbf{u} < 1/2$, one has

$$\mathsf{hous}(v, x) = (P + E)x \quad \text{for} \quad \|E\|_F \leq 2c_{\mathsf{h}} n\mathbf{u}, \tag{4.71}$$

for any $x \in \mathbb{C}^n$. This will be used later in the analysis of RHess.

Hessenberg Form

The standard way in which a matrix $A \in \mathbb{C}^{n\times n}$ is put into Hessenberg form using Householder reflectors is by using a *left-to-right* approach, where one generates a sequence of Householder reflectors $P_1, \ldots, P_{n-2}$, that ensure that $H := P_{n-2} \cdots P_1 A P_1 \cdots P_{n-2}$ is Hessenberg, and where each P_i is used to set to zero the entries in *column* i of the working matrix that are *below* the subdiagonal.

However, since we will be interested in randomizing the relative position of e_n with respect to the eigenbasis of H, it will be convenient to instead use a *bottom-up* approach, and choose

[8] It is easy to see that P is a reflection over the hyperplane $\{v\}^\perp$.

each P_i to set to zero the entries in *row* i that are to the *left* of the corresponding subdiagonal. In this way, when acting on the left of the matrix, the P_i leave the n-th row of the working matrix invariant and, in particular, we will have $e_n^* P_i = e_n^*$. Since the left-to-right and bottom-up approaches are essentially equivalent, the results from [156, Theorem 2] and [55, Section 4.4.6] apply in both situations, and in particular imply the existence of an efficient and backward stable algorithm in the following sense.

Definition 4.3.24 (Bottom-up Hessenberg Form Algorithm)**.** A c_{H}-stable bottom-up Hessenberg form algorithm HessBU, is an algorithm that takes as input a matrix $A \in \mathbb{C}^{n\times n}$ and outputs a Hessenberg matrix $H \in \mathbb{C}^{n\times n}$ satisfying that there exists a unitary Q with

$$\|H - Q^* A Q\| \leq c_{\mathsf{H}} \|A\| n^{5/2} \mathbf{u}$$

and such that $Qe_n = e_n$. We say that HessBU is efficient if it runs in at most

$$T_{\mathsf{HessBU}}(n) := \frac{10}{3} n^3 + O(n^2)$$

arithmetic operations.

Analysis of RHess

As mentioned above, the only source of randomness for HessBU is a random vector uniformly sampled from the complex unit sphere. Our main technical tool for the analysis will be the following standard anti-concentration result.

Lemma 4.3.25 (Anti-Concentration for Random Vectors)**.** *Let* $u \sim \mathrm{Unif}(\mathbb{S}_{\mathbb{C}}^{n-1})$ *and* $v \in \mathbb{C}$ *with* $\|v\| = 1$*. Then for all* $t \in [0, 1]$

$$\mathbb{P}\left[|u^* v| \leq \frac{t}{\sqrt{n-1}}\right] \leq t^2.$$

Proof. Because the distribution of u is unitarily invariant and $\|v\| = 1$, we have $u^* v =_d u^* e_i = u(i)$[9] for every $i \in [n]$. So, for concreteness we will take $i = 1$ and bound $\mathbb{P}[|u(1)| \leq t]$ for any $t \geq 0$.

Now recall that if $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are independent *real* standard Gaussians, then

$$u =_d \frac{(X_1 + iY_1, \ldots, X_k + iY_n)}{\sqrt{X_1^2 + Y_1^2 + \cdots + X_k^2 + Y_n^2}}$$

and in particular $|u(1)|^2 = \frac{Z_1}{Z_1+Z_2}$ where $Z_1 \sim \chi^2(2)$ and $Z_2 \sim \chi^2(2n-2)$ are independent. Then, we use the well known fact that $\frac{Z_1}{Z_1+Z_2}$ has a $\mathrm{Beta}(1, n-1)$ distribution, and hence its

[9] Given two random variables X and Y, we use $X =_d Y$ to denote that they have the same distribution.

probability density function is given $f_{\mathrm{Beta}(1,n-1)}(s) = (n-1)(1-s)^{n-2} \cdot 1_{\{0\le s\le 1\}}$. It follows that, for $t \in [0,1]$

$$\mathbb{P}[|u(1)| \le t] = \mathbb{P}[|u(1)|^2 \le t^2] = (n-1)\int_0^{t^2} (1-s)^{n-2} ds = 1 - (1-t^2)^{n-1} \le (n-1)t^2,$$

where the last inequality follows from Bernoulli's inequality. □

We can now define the algorithm and proof its guarantees.

RHess

Input: $A \in \mathbb{C}^{n\times n}$
Output: $H \in \mathbb{C}^{n\times n}$
Requires: $\Lambda_\epsilon(A)$ is ζ-shattered
Ensures: H is Hessenberg, $\|H - Q^*AQ\| \le c_{\mathsf{RH}}\|A\|n^{5/2}\mathbf{u}$ for some unitary Q, $\Lambda_{\epsilon'}(H)$ is ζ-shattered for $\epsilon' = \epsilon - c_{\mathsf{RH}}\|A\|n^{5/2}\mathbf{u}$. Moreover, for any t, with probability at least $1 - nt^2$ it holds that $\mathbb{P}[Z_H = \lambda] \ge \left(\frac{\epsilon' t}{n^{3/2}\zeta}\right)^2$ for all $\lambda \in \operatorname{Spec} H$

1. $u \sim \mathrm{Unif}(\mathbb{S}^{n-1}_{\mathbb{C}})$
2. $H \leftarrow \mathsf{hous}(u - e_n, A)$
3. $H \leftarrow \mathsf{hous}(u - e_n, H^*)^*$
4. $H \leftarrow \mathsf{HessBU}(H)$

Proposition 4.3.26 (Guarantees for randomized Hessenberg form)**.** *Assume that*

$$\mathbf{u} \le \mathbf{u}_{\mathsf{RHess}}(n) := \frac{1}{20 c_{\mathsf{h}} n^{3/2}}. \tag{4.72}$$

Then, RHess *satisfies its guarantees for* $c_{\mathsf{RH}} = 3(c_{\mathsf{H}} + c_{\mathsf{h}})$ *and can be instantiated using at most*

$$T_{\mathsf{RHess}}(n) := T_{\mathsf{HessBU}}(n) + 2nT_{\mathsf{hous}}(n) + C_{\mathsf{U}}n = O(n^3).$$

arithmetic operations.

Proof. The case $n = 1$ is trivial so we assume $n \ge 2$. Let H be the output of $\mathsf{RHess}(A)$, A_1 and A_2 be the matrices computed in lines 2 and 3 of RHess, $P = I - \beta vv^*$ for $v = u - e_n$ (and $\beta = \frac{2}{v^*v}$), and define $E_1 := A_1 - PA$ and $E_2 := A_2 - A_1P$. From (4.71) it is easy to see that

$$\|E_1\| \le 2c_{\mathsf{h}}\|A\|n^{3/2}\mathbf{u} \quad \text{and} \quad \|E_2\| \le 2c_{\mathsf{h}}\|A_1\|n^{3/2}\mathbf{u}.$$

Using the first inequality and (4.72) we get that $\|A_1\| \leq \|E_1\| + \|A\| \leq 1.1\|A\|$. Then, combining this with the second inequality we get $\|E_2\| \leq 2.2c_{\mathsf{h}}\|A\|n^{3/2}\mathbf{u}$. Hence

$$\|A_2 - PAP\| \leq \|A_2 - A_1P\| + \|A_1P - PAP\| = \|E_1\| + \|E_2\| \leq 4.2c_{\mathsf{h}}\|A\|n^{3/2}\mathbf{u}. \quad (4.73)$$

Again because of (4.72) the above inequality implies that $\|A_2\| \leq 1.3\|A\|$. So, by Definition 4.3.24 we get that $\|H - Q^*A_2Q\| \leq 1.3c_{\mathsf{H}}\|A\|n^{5/2}\mathbf{u}$, for some unitary Q satisfying $Qe_n = e_n$, which combined with (4.73) yields

$$\|H - Q^*PAPQ\| \leq (1.3c_{\mathsf{H}}n^{5/2} + 4.2c_{\mathsf{h}}n^{3/2})\|A\|\mathbf{u} \leq c_{\mathsf{RH}}\|A\|n^{5/2}\mathbf{u},$$

proving the first claim. Now, because $\Lambda_\epsilon(A)$ is ζ-shattered, the above inequality and Lemma 1.1.7 imply that $\Lambda_{\epsilon'}(H)$ is ζ-shattered for $\epsilon' = \epsilon - c_{\mathsf{RH}}\|A\|n^{5/2}\mathbf{u}$.

It remains to prove the anti-concentration statement for Z_H. To do this let $E \in \mathbb{C}^{n\times n}$ be such that $H = Q^*P(A+E)PQ$, and let $A + E = VDV^{-1}$ with $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and V chosen so that $\|V\| = \|V^{-1}\| = \sqrt{\kappa_V(A+E)}$. Now note that Q^*PV is an eigenvector matrix for H, and because P and Q are unitary $\|Q^*PV\| = \|V\| = \sqrt{\kappa_V(A+E)} = \sqrt{\kappa_V(H)}$. So

$$\begin{aligned}
\mathbb{P}[Z_H = \lambda_i] &= \frac{|e_n^*Q^*PVe_i|^2}{\|e_n^*Q^*PV\|^2} && \text{definition of } \mathbb{P}[Z_H = \lambda_i] \\
&= \frac{|e_n^*PVe_i|^2}{\|e_n^*PV\|^2} && e_n^*Q^* = e_n^* \\
&= \frac{|u^*Ve_i|^2}{\|u^*V\|^2} && u = Pe_n \text{ by definition of } P.
\end{aligned}$$

To simplify notation define $v_i := \frac{Ve_i}{\|Ve_i\|}$. We then have

$$\begin{aligned}
\frac{|u^*Ve_i|^2}{\|u^*V\|^2} &= \frac{|u^*v_i|^2\|Ve_i\|^2}{\|u^*V\|^2} \\
&\geq \frac{|u^*v_i|^2}{\|V\|^2\|V^{-1}\|^2} && \|Ve_i\| \geq \frac{1}{\|V^{-1}\|} \text{ and } \|u^*V\| \leq \|V\| \\
&= \frac{|u^*v_i|^2}{\kappa_V(H)^2} && \kappa_V(A+E) = \kappa_V(H) \\
&\geq \left(\frac{\epsilon'|u^*v_i|}{n\zeta}\right)^2 && \Lambda_{\epsilon'}(H) \text{ is } \zeta\text{-shattered and Lemma 4.3.11.}
\end{aligned}$$

Now, because $\|v_i\| = 1$, we can apply Lemma 4.3.25 to get that for any $t \geq 0$

$$\mathbb{P}\left[|u^*v_i| \geq \frac{t}{\sqrt{n-1}}\right] \geq 1 - t^2.$$

Which, in conjunction with the above gives that

$$\mathbb{P}[Z_H = \lambda_i] \geq \left(\frac{\epsilon' t}{n^{3/2}\zeta}\right)^2,$$

with probability at least $1 - t^2$. The advertised claim then follows from taking a union bound over all v_i. The claim about the running tie follows trivially.

□

4.3.7 The Main Algorithm

So far, all but one of the subroutines required to define SmallEig have been discussed. The remaining subroutine that will be needed is the one used for deflation, denoted here by $\texttt{deflate}(H, \omega)$, which on a Hessenberg input $H \in \mathbb{C}^{n \times n}$ sets to zero any of the $n-1$ subdiagonals of H that are less or equal (in absolute value) to ω, and returns the diagonal blocks $H_1, H_2, \ldots$ of the resulting matrix.

We are now ready to define the main algorithm and prove its guarantees. Note that n refers to the dimension of the original input matrix, which is used to set parameters throughout the recursive calls to findRitz.

findRitz

Input: Complex matrix A, accuracy δ, failure probability tolerance ϕ
Global Data: Dimension n, norm estimate Σ, pseudospectral parameter ϵ, shattering parameter ζ
Output: A multiset $\Lambda \subset \mathbb{C}$
Ensures: Λ is the spectrum of a matrix $\widetilde{A}$ with $\|A - \widetilde{A}\| \leq \delta$.

1. $\Delta \leftarrow \frac{\delta\Sigma}{2}$, $\omega \leftarrow \frac{\epsilon \wedge \Delta}{3n}$, $\beta \leftarrow \frac{\omega}{20}$, $p \leftarrow \frac{\phi\epsilon^2}{2n^5\zeta^2}$, $\varphi \leftarrow \frac{\phi}{2n}$, $\mathsf{correctness} \leftarrow \texttt{false}$
2. $H \leftarrow \mathsf{RHess}(A)$
3. **While** $\mathsf{correctness} = \texttt{false}$
   $[\widetilde{\lambda}, \mathsf{correctness}] \leftarrow \mathsf{findOne}(H, \beta, \varphi, p)$
4. $H \leftarrow \texttt{dec}(H, \widetilde{\lambda}, \omega)$
5. $[A_1, A_2, \ldots] \leftarrow \texttt{deflate}(H, \omega)$
6. $\Lambda \leftarrow \bigsqcup_i \mathsf{findRitz}\big(A_i, \delta, \phi\big)$

Theorem 4.3.27. *Let A be the input matrix and $\delta \in (0,1)$. Let β, Δ and p be as in line 1 of* SmallEig*. Assume that the global data satisfies $n = \dim(A)$, $\Sigma/2 \leq \|A\| \leq \Sigma$, and that $\Lambda_{2\epsilon}(A)$ is ζ-shattered. If*

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{\mathsf{SmallEig}}(n, \Sigma, \epsilon, \zeta, \delta, \phi) \\ &:= \frac{\epsilon}{6 \cdot 10^3 (c_{\mathsf{h}} \vee c_{\mathsf{H}} \vee c_{\mathsf{root}}) \nu_{\mathsf{IQR}}(n) n \zeta} \left(\frac{\eta_1}{44\Sigma}\right)^{2m_1} \end{aligned} \tag{4.74}$$

where

$$m_1 = \lceil 12\log(n\zeta/\epsilon) + 6\log(1/p)\rceil = O\big(\log(n\zeta/\epsilon\phi)\big)$$
$$\text{and} \quad \eta_1 = \left(\frac{\epsilon\wedge\Delta}{300n}\right)\left(\frac{\phi}{24n\log(18\Sigma n/(\epsilon\wedge\Delta))}\right)^{1/2} = O\left(\frac{(\epsilon\wedge\Delta)\phi^{1/2}}{n\log(\Sigma n/(\epsilon\wedge\Delta))^{1/2}}\right) \tag{4.75}$$

Then, with probability at least $1-\phi$ SmallEig *satisfies its guarantees, and in this event runs in at most*

$$\begin{aligned} T_{\mathsf{SmallEig}}(n,\epsilon,\zeta,\delta) &= (n-1)(T_{\mathsf{RHess}}(n) + T_{\mathsf{findOne}}(n,\Sigma,\epsilon,\zeta,p,\beta) + T_{dec}(n,\zeta n/\epsilon,p,\beta)) \\ &= O\Big(n^4 + n^3\log(\zeta n/\epsilon\phi)\big(\log(\Sigma n/(\epsilon\wedge\Delta)) + \log(\zeta n/\epsilon\phi)\big) \\ &\qquad + \log(\Sigma n/(\epsilon\wedge\Delta))\log(\zeta n/\epsilon\phi)\log\log(\zeta n/\epsilon\phi)\Big) \end{aligned}$$

arithmetic operations.

Preservation of the Norm and Pseudospectral Parameters

Before delving into the analysis of SmallEig we will show that the global data provides valuable information throughout the execution of the algorithm. The first observation here is that the only subroutine of SmallEig that accesses the global data is findOne, so, to ensure correctness, the only requirements regarding the global data that need to be fulfilled are the ones ensured by the following lemma.

Lemma 4.3.28. *Suppose that the assumptions of Theorem 4.3.27 are satisfied and let* H' *and* A' *be any values acquired by the variables* H *and* A*. If every while loop (line 3) involved in the production of* H' *and* A' *ended with* findOne *being terminated successfully, then:*

i) $\frac{\omega}{2} \le \min\{\|H'\|, \|A'\|\} \le \max\{\|H'\|, \|A'\|\} \le 2\Sigma$.

ii) $\Lambda_\epsilon(H')$ *and* $\Lambda_\epsilon(A')$ *are* ζ*-shattered.*

To prove the above lemma we will need the following result, whose proof consists of combining Lemmas 1.1.7 and 3.5.8, to control the pseudospectral parameters after each deflation step.

Lemma 4.3.29 (Pseudospectrum After Deflation)**.** *Let* $H \in \mathbb{C}^{n\times n}$ *be a Hessenberg matrix and and* $1 \le r \le n-1$ *. Let* H_- *and* H_+ *be its upper-left and lower-right* $r\times r$ *and* $(n-r)\times(n-r)$ *corners respectively. If* $|H_{r+1,r}| \le \epsilon'$ *then*

$$\Lambda_{\epsilon-\epsilon'}(H_-) \cup \Lambda_{\epsilon-\epsilon'}(H_+) \subset \Lambda_\epsilon(H).$$

Proof. Let H_0 be the matrix obtained by zeroing out the $(r+1,r)$ entry of H. By Lemma 1.1.7 and the assumption $|H_{r+1,r}| \le \epsilon'$ we get $\Lambda_{\epsilon-\epsilon'}(H_0) \subset \Lambda_\epsilon(H)$.

We will begin by showing that $\Lambda_{\epsilon-\epsilon'}(H_+) \subset \Lambda(H_0)$. Let $w \in \mathbb{C}^r$ be any left eigenvector of H_+ and note that, since H_0 is block upper triangular, $0_{n-r} \oplus w \in \mathbb{C}^{n\times n}$ is a left eigenvector of H_0. Hence, there is a spectral projector P of H_0 for which its left eigenvectors (equivalently its rows) span the space $\operatorname{span}\{e_{n-r+1}, \dots, e_n\}$. Hence the span of the columns of the $n \times (n-r)$ matrix

$$S = \begin{pmatrix} 0 \\ I_{n-r} \end{pmatrix}$$

coincides the span of the rows of P. So, by Lemma 3.5.8, $\Lambda_{\epsilon-\epsilon'}(H_+) = \Lambda_{\epsilon-\epsilon'}(SHS^*) \subset \Lambda_{\epsilon-\epsilon'}(H_0)$.

The proof that $\Lambda_{\epsilon-\epsilon'}(H_-) \subset \Lambda_{\epsilon-\epsilon'}(H_0)$ is very similar, with the sole difference that this time one should look at the right eigenvectors of H_-, and work with columns (rather than rows) of the spectral projector. □

We can now proceed to the proof of the lemma.

Proof of Lemma 4.3.28. First note that in each call to SmallEig the working matrix gets modified exactly once by each of the subroutines RHess, `dec` and `deflate`. So, there is a sequence of the form

$$A = A_1, F_1, F_1', A_2, F_2, F_2' \dots$$

that ends in H' (respectively A'), and such that $F_i = \mathsf{RHess}(A_i), F_i' = \mathtt{dec}(F_i, , \widetilde{\lambda}_i, \omega)$ and A_{i+1} is one of the matrices in the output of $\mathtt{deflate}(F_i')$. Moreover, by the assumption that findOne terminated successfully at the end of each while loop, we have that

$$\eta_1 \le \operatorname{dist}(\widetilde{\lambda}_i, \operatorname{Spec} F_i) \le \beta. \tag{4.76}$$

We will show by induction that for every $i \le n$ the pseudospectra $\Lambda_{2\epsilon-\epsilon_{i,0}}(A_i), \Lambda_{2\epsilon-\epsilon_{i,1}}(F_i)$ and $\Lambda_{2\epsilon-\epsilon_{i,2}}(F_i')$ are ζ-shattered, where

$$\epsilon_{i,j} := (3(i-1)+j)\omega = \frac{3(i-1)+j}{3n}(\Delta \wedge \epsilon),$$

and that $\|A_i\| \le \Sigma + \epsilon_{i,0}, \|F_i\| \le \Sigma + \epsilon_{i,1}$ and $\|F_i'\| \le \Sigma + \epsilon_{i,2}$. Note that in particular this will imply that ϵ-pseudospectra of the A_i, F_i and F_i' are ζ-shattered, and their norms are bounded by 2Σ (since $\epsilon \le \Sigma$).

That $A_1 = A$ has the advertised pseudospectral and norm properties follows from the assumption about the global data. We can then induct:

- *Effect of* RHess. Assume that $\Lambda_{2\epsilon-\epsilon_{i,0}}(A_i)$ is ζ-shattered and $\|A_i\| \le \Sigma + \epsilon_{i,0}$. Because $F_i = \mathsf{RHess}(A_i)$, and since (4.63) implies that

  $$\mathbf{u} \le \mathbf{u}_{\mathsf{RHess}}(n) \le \mathbf{u}_{\mathsf{RHess}}(\dim(A_i)),$$

  we can apply Proposition 4.3.26 to get that $\Lambda_{2\epsilon-\epsilon_{i,0}-\epsilon'}(F_i)$ is ζ-shattered for

  $$\epsilon' = c_{\mathsf{RH}}\|A_i\| \dim(A_i)^{5/2}\mathbf{u}$$

$$\begin{aligned} &\leq 2c_{\mathsf{RH}}\Sigma n^{5/2}\mathbf{u} && \dim(A_i) \leq n,\ \|A_i\| \leq 2\Sigma \\ &\leq \omega && \text{by (4.74)}. \end{aligned}$$

So, it follows that $\Lambda_{2\epsilon-\epsilon_{i,1}}(F_1)$ is ζ-shattered. And in the same way we can get $\|F_i\| \leq \Sigma + \epsilon_{i,1}$.

- *Effect of* `dec`. Now assume that $\Lambda_{2\epsilon-\epsilon_{i,1}}(F_i)$ is ζ-shattered and $\|F_i\| \leq \Sigma + \epsilon_{i,1}$. Let p and β be as in line 1 of SmallEig and define

$$m_2 := \left\lceil \frac{\log(\zeta^2 n^2/p\epsilon^2)}{2\log(3\omega/4\beta)} \right\rceil = \left\lceil \frac{\log(\zeta^2 n^2/p\epsilon^2)}{2\log(15)} \right\rceil. \tag{4.77}$$

Now, because $m_2 \leq \lceil .3\log(\zeta n/\epsilon) + .15\log(1/p)\rceil$, it is clear that $m_2 \leq m_1$, and then it is easy to see that (4.74) implies

$$\mathbf{u} \leq \mathbf{u}_{\mathtt{dec}}(2\Sigma, \zeta n/\epsilon, p, \eta_1).$$

So, because $\|F_i\| \leq 2\Sigma$ (by assumption), $\kappa_V(F_i) \leq \zeta\epsilon/n$ (since $\Lambda_\epsilon(F_i)$ is ζ-shattered and by Lemma 4.3.11), and $\operatorname{dist}(\widetilde{\lambda}, \operatorname{Spec} F_i) \geq \eta_1$ (by the assumption in (4.76)), we can apply Proposition 4.3.23 to get that there exists a unitary matrix Q for which

$$\begin{aligned} \|F_i' - Q^* F_i Q\| &\leq 3.5 m_1 \|F_i\| \nu_{\mathsf{IQR}}(n)\mathbf{u} \\ &\leq 7 m_1 \Sigma \nu_{\mathsf{IQR}}(n)\mathbf{u} && \|F_i\| \leq 2\Sigma \\ &\leq \omega && \text{by (4.74)}. \end{aligned}$$

Then, by Lemma 1.1.7 and the assumption that $\Lambda_{2\epsilon-\epsilon_{i,1}}(F_i)$ is ζ-shattered, it follows that $\Lambda_{2\epsilon-\epsilon_{i,2}}(F_i')$ is ζ-shattered. And because the norm is preserved under unitary conjugation we also get that $\|F_i'\| \leq \Sigma + \epsilon_{i,2}$.

- *Effect of* `deflate`. Assume that $\Lambda_{2\epsilon-\epsilon_{i,2}}(F_i')$ is ζ-shattered, and recall that A_{i+1} is an output of $\mathtt{deflate}(F_i', \omega)$. Then, by Lemma 4.3.29 we have that

$$\Lambda_{2\epsilon-\epsilon_{i+1,0}}(A_{i+1}) = \Lambda_{2\epsilon-\epsilon_{i,2}-\omega}(A_{i+1}) \subset \Lambda_{2\epsilon-\epsilon_{i,2}}(F_i')$$

and hence $\Lambda_{2\epsilon-\epsilon_{i+1,0}}(A_{i+1})$ is ζ-shattered. Similarly, we can note that $\|A_{i+1}\| \leq \|F_i''\| + \omega \leq \Sigma + \epsilon_{i+1,0}$, which concludes the induction.

Now, since the depth of the recursion tree of SmallEig is at most n, and we have proven the above claim for any A_i, F_i, F_i'' with $i \leq n$, we can conclude that $\Lambda_\epsilon(H')$ is ζ-shattered (resp. $\Lambda_\epsilon(A')$) and $\|H'\| \leq 2\Sigma$ (resp. $\|A'\| \leq 2\Sigma$), as we wanted to show.

Finally, to show that $\omega/2 \leq \|H'\|$ (resp. $\omega/2 \leq \|A'\|$), first note that $\omega \leq \|A_i\|$ for every i. Indeed, when $i = 1$ we can use the assumption $\delta \leq 1$, which yields $\Delta \leq \|A\|$, and combine this with $\omega \leq \Delta$. For $i > 1$ note that A_i is an output of $\mathtt{deflate}(F_{i-1}'', \omega)$, and hence its subdiagonals are guaranteed to have absolute value at least ω, which implies that $\omega \leq \|A_i\|$. We can then proceed as above (using slightly stronger bounds) to show that $\|A_i - F_i\| \leq \omega/2$ and $\|A_i - F_i''\| \leq \omega/2$. So the proof is concluded. □

Analysis of SmallEig

We are now ready to prove Theorem 4.3.27. For clarity, let us divide the proof in several parts.

Backward stability. Assume that SmallEig terminates and outputs Λ. Moreover, assume that when running SmallEig, at the end of all the while loops from line 3, the subroutine findOne terminated successfully (later we will prove that this occurs with probability at least $1-\phi$).

We will show that Λ is the spectrum of a matrix $\widetilde{A}$ with $\|\widetilde{A}-A\| \leq \Delta$ (which combined with the assumption about the global data gives $\|\widetilde{A}-A\| \leq \delta\|A\|$). To be precise we will show an equivalent statement, namely that Λ is the spectrum of a matrix that is at distance at most Δ from the class of matrices that are unitarily equivalent to A. To do this, for the purpose of the analysis, it will be convenient to imagine that during the deflation process (after setting to zero the small subdiagonals) instead of cutting out the blocks on the diagonal and considering them as separate subproblems, one keeps the full $n \times n$ matrix and continues to operate on the full matrix in the obvious way. With this view point the algorithm terminates when the working matrix becomes an upper triangular matrix, and its diagonal elements are precisely the elements of Λ.

In the proof of Lemma 4.3.28 it was shown that the only subroutines that deviate the working matrix from the unitary orbit of the original matrix are RHess, `dec` and `deflate`. Moreover, it was shown that when each of these subroutines is applied, the corresponding backward error incurred is at most of size ω. So we need only to give an upper bound for the number of times these subroutines are called. To do this consider $\mathcal{T}_n(A)$, the recursion tree of SmallEig, where the input matrix A is placed at the root, and then the children of any vertex v are in one-to-one correspondence with the matrices outputted after running `deflate` on the matrix associated to v. It is clear from the construction that leaves correspond to matrices of dimension 1, and internal vertices (vertices that are not leaves) correspond to higher dimensional matrices. Now note that for any internal vertex v it holds that the sum of the dimensions of the matrices associated to the children of v equals the dimension of the matrix associated to v. Then, by induction on n it follows that $\mathcal{T}_n(A)$ has at most $n-1$ internal vertices. And, since the relevant subroutines are only called once at times corresponding to internal leaves, we conclude that each of these subroutines was called at most $n-1$ times. Hence, the ultimate deviation from the original unitary equivalence class is at most

$$3(n-1)\omega = \frac{3(n-1)}{3n}(\epsilon \wedge \Delta) \leq \Delta,$$

as we wanted to show.

Precision requirements. To ensure that the precision has been set to be small enough, so that the precision requirements of each subroutine are satisfied throughout the iteration, we will show that

$$\mathbf{u}_{\mathsf{SmallEig}}(n,\Sigma,\epsilon,\zeta,\delta,\phi,n) \leq \min\{\mathbf{u}_{\mathsf{RHess}}(n), \mathbf{u}_{\mathsf{findOne}}(n,2\Sigma,\epsilon,\zeta,p,\beta,\varphi), \mathbf{u}_{\mathsf{dec}}(n,2\Sigma,\zeta n/\epsilon,p,\eta_1)\}.$$

First, that $\mathbf{u}_{\mathsf{SmallEig}}(n,\Sigma,\epsilon,\zeta,\delta,\phi,n) \leq \mathbf{u}_{\mathsf{RHess}}(n)$ is trivial. On the other hand, by definition we have

$$\begin{aligned}\mathbf{u}_{\mathsf{findOne}}(n,\Sigma,\epsilon,\zeta,p,\beta,\phi) &= \mathbf{u}_{\mathsf{distSpec}}\big(n,m_1,10,2\Sigma,n\zeta/\epsilon,\eta_1\big) && \text{for } m_1,\eta_1 \text{ as in (4.75)}\\ &\geq \frac{\epsilon}{6\cdot 10^3 c_{\mathsf{root}}\cdot \nu_{\mathsf{IQR}}(n)n\zeta}\left(\frac{\eta_1}{44\Sigma}\right)^{2m_1} && \text{(4.63) and (4.27)}\end{aligned}$$

So from the (4.74) it is clear that $\mathbf{u}_{\mathsf{SmallEig}}(n,\Sigma,\epsilon,\zeta,\delta,\phi,n) \leq \mathbf{u}_{\mathsf{findOne}}(n,2\Sigma,\epsilon,\zeta,p,\beta,\phi)$. Finally

$$\begin{aligned}\mathbf{u}_{\mathsf{dec}}\big(n,2\Sigma,\zeta n/\epsilon,p,d\big) &= \frac{\mathbf{u}_{\mathsf{IQR}}(n,m_2,2\Sigma,\zeta n/\epsilon,\eta_1)\omega}{16\cdot 5^{m_2}\cdot n^{1/2}2\Sigma} && \text{for } m_2 \text{ as in (4.77)}\\ &= \frac{\omega\epsilon}{16\cdot 8\nu_{\mathsf{IQR}}(n)n^{3/2}\zeta\cdot 2\Sigma}\left(\frac{\eta_1}{5\cdot 2\Sigma}\right)^{m_2} && \text{from (4.23)}\end{aligned}$$

And because $m_2 \leq m_2$, from (4.74) it is clear that

$$\mathbf{u}_{\mathsf{SmallEig}}(n,\Sigma,\epsilon,\zeta,\delta,\phi,n) \leq \mathbf{u}_{\mathsf{dec}}\big(n,2\Sigma,\zeta n/\epsilon,p,d\big).$$

Probability of success. Observe that the only randomized subroutines of SmallEig are RHess and findOne. First we will provide a lower bound for the probability that the guarantees of RHess and findOne are satisfied every time these subroutines are called.

Combining Lemma 4.3.28 and Proposition 4.3.26 we get that, if findOne has succeeded every time it has been called, then for any value H' acquired by the variable H in line 2 of SmallEig we have for any $t>0$, with probability $1-nt^2$, that

$$\min_{\lambda\in \mathrm{Spec}H'} \mathbb{P}[Z_{H'}=\lambda] \geq \left(\frac{\epsilon t}{n^{3/2}\zeta}\right)^2.$$

In particular (for $t^2=\phi/2n^2$) we get that with probability at least $1-\phi/2n$ it holds that

$$\min_{\lambda\in \mathrm{Spec}H'} \mathbb{P}[Z_{H'}=\lambda] \geq p$$

for p defined as in line 1. Under this event, and because of Lemma 4.3.28 and because the precision is high enough, the requirements of findOne will be met in line 3, and hence (for this call) findOne will succeed with probability at least $1-\varphi = 1-\phi/2n$.

Therefore, every time SmallEig is called, both RHess and findOne will satisfy their guarantees with probability at least $1-\phi/n$. Moreover, from the backward stability proof we know that the recursion tree for SmallEig has at most $n-1$ internal vertices. Therefore, we can conclude that all the calls to RHess and findOne will succeed with probability at least $1-\phi$, as we wanted to show.

Now, under the assumption that RHess and findOne succeed every time, we have that the values of the variables H and $\widetilde{\lambda}$ that are passed every time to `dec` satisfy the requirements of

this subroutine, and by our previous discussion we know that the precision requirements for dec are also met. Therefore, we can apply Proposition 4.3.23 to argue that the matrix H will be decouple in a finite amount of time, and by Lemma 4.3.28 we know that the pseudospectral parameters and norm guarantees will also be maintained.

Running time. From the above discussion we know that with probability at least $1-\phi$, SmallEig terminates successfully and moreover, throughout the algorithm, every call to RHess, findOne and dec will be successful, and the requirements of these subroutines will always be met. Under this event (recalling that each subroutine is called at most $n-1$ times) by Propositions 4.3.26, 4.3.17 and 4.3.23 and using the the running times of the subroutine are monotone in the dimension of the input, we get that the running time of SmallEig is at most

$$(n-1)(T_{\mathsf{RHess}}(n) + T_{\mathsf{findOne}}(n,\Sigma,\epsilon,\zeta,p,\beta) + T_{\mathsf{dec}}(n,\zeta n/\epsilon,p,\beta)).$$

The proof is concluded by writing p,β and η_1 as a function of ϵ,ζ, Σ and δ, and using the big-O bounds provided in Propositions 4.3.26, 4.3.17 and 4.3.23.

Pseudospectral Shattering and Proof of the Main Result

Note that Theorem 4.3.27 assumes that SmallEig has access to the parameters ϵ and ζ in the global data, which control both the minimum eigenvalue gap and the eigenvector condition number of the input matrix $A \in \mathbb{C}^{n\times n}$. In order to ensure that SmallEig works on every input (without having access to ϵ and ζ), instead of running the algorithm on A we will run it on $A+\gamma G_n$ (for $\gamma = \Theta(\delta)$ and G_n a normalized complex Ginibre matrix[10]), and exploit the following result, whose proof we defer to Section C.2 in Appendix C.[11]

Lemma 4.3.30 (Shattering)**.** *For any $A \in \mathbb{C}^{n\times n}$ and $\varphi \in (0,1/2), \gamma \in (0,\|A\|/2)$, we have that, with probability at least $1-\varphi$, $\Lambda_\epsilon(A+\gamma G_n)$ is ζ-shattered for*

$$\zeta := \frac{\varphi^{1/2}\gamma}{2\sqrt{3}n^{3/2}} \quad \textit{and} \quad \epsilon := \frac{\gamma^2\varphi}{180\sqrt{2}\|A\|\log(1/\varphi)n^3}$$

The main result of this section then follows from combining Lemma 4.3.30 with Theorem 4.3.27.

Proof of Theorem 4.3.1. Start by recalling the following the well-known tail bound for the norm of a Ginibre matrix (e.g. see [15, Lemma 2.2])

$$\mathbb{P}[\|G_n\| \geq t] \leq 2\exp\big(-n(t-2\sqrt{2})^2\big), \quad \forall t \geq 2\sqrt{2}. \tag{4.78}$$

[10] That is, an $n\times n$ random matrix with independent centered complex Gaussian entries of variance $1/n$.

[11] Note that a version of this result has already been proven and used in a similar context in Chapter 3. However, since the notion of *shattered pseudospectrum* from that chapter differs from the one used here, we cannot directly apply it.

Then, for $W := 2\sqrt{2} + \frac{1}{n^{1/2}} \log(6/\phi)^{1/2}$ we have that

$$\mathbb{P}[\|G_n\| \leq W] \geq 1 - \phi/3.$$

Then, given a norm estimate Σ satisfying $\Sigma/2 \leq \|A\|(1 \pm \delta/2) \leq \Sigma$, we will choose $\gamma := \frac{\delta\Sigma}{4W}$, so that

$$\begin{aligned}
\mathbb{P}\left[\gamma\|G_n\| \leq \frac{\delta\|A\|}{2}\right] &= \mathbb{P}\left[\|G_n\| \leq \frac{2\|A\|W}{\Sigma}\right] \\
&\geq \mathbb{P}\left[\|G_n\| \leq W\right] && \Sigma/2 \leq \|A\| \\
&\geq 1 - \phi/3.
\end{aligned}$$

Moreover, for this choice of γ, by Lemma 4.3.30 we have that, with probability at least $1 - \phi/3$, $\Lambda_\epsilon(A + \gamma G_n)$ is ζ-shattered for

$$\zeta := \frac{\phi^{1/2}\gamma}{2\sqrt{6}n^{3/2}} \quad \text{and} \quad \epsilon := \frac{\gamma^2\phi}{540\sqrt{2}\|A\| \log(1/\phi)n^3}.$$

On the other hand, conditioning on $\|G_n\| \leq W$ and $\Lambda_\epsilon(A + \gamma G_n)$ being ζ-shattered, we have that $\mathsf{SmallEig}(A + \gamma G_n, \delta/2, \phi/3)$ succeeds with probability at least $1 - \phi/3$ when using n, ϵ, ζ and Σ as global data and provided that $\mathbf{u}$ satisfies (4.74)), in which case the output Λ will be a δ-backward approximation of the spectrum of A.

Hence, using a union bound we get that with probability $1-\phi$, $\mathsf{SmallEig}(A+\gamma G_n, \delta/2, \phi/3)$ provides a δ-accurate answer, and from Theorem 4.3.27 we have that the running time and required bits of precision are as in the statement of of Theorem 4.3.1. □

Chapter 5

The Lanczos Algorithm Under Few Iterations

5.1 Preliminaries

Throughout this chapter only elementary facts about orthogonal polynomials are used. For the convenience of the reader below we include a succinct summary of the results that will be used throughout the chapter. We refer the reader to Chapter 2 in [149] and Chapters 2 and 3 in [48] for a more detailed discussion of these tools.

In order to establish context and notation, we will also explicitly define the Lanczos algorithm and its interpretation in terms of orthogonal polynomials. Some standard references for this matter are Chapter 6 in [157] and Chapter 6 in [134].

5.1.1 Orthogonal Polynomials

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

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

The hypothesis that $|\text{supp}(\mu)| = \infty$ 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

$$\gamma_k = \left(\int_{\mathbb{R}} \pi_k^2(x) d\mu(x) \right)^{-\frac{1}{2}}. \tag{5.1}$$

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,

$$\int_{\mathbb{R}} \pi_k^2(x) d\mu(x) = \min_{q \in \mathcal{P}_k} \int_{\mathbb{R}} q^2(x) d\mu(x),$$

where $\mathcal{P}_k$ 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:

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

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

$$\gamma_k = \left(\prod_{i=0}^{k-1} \beta_i \right)^{-1}. \tag{5.2}$$

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 [48] 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 [48], Section 3.1) that if we define $D_k = \det M_k$, then

$$\beta_k = \frac{\sqrt{D_{k-1} D_{k+1}}}{D_k} \quad \text{and} \quad \gamma_k = \sqrt{\frac{D_{k-1}}{D_k}}, \quad k \geq 0, \tag{5.3}$$

where we define $D_{-1} = 1$. Note that the second identity in (5.3) implies

$$D_k = \prod_{i=0}^{k} \gamma_i^{-2}. \tag{5.4}$$

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:

$$p_k(x) = \frac{\det \tilde{M}_k(x)}{\sqrt{D_{k-1}D_k}}. \tag{5.5}$$

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 $k \leq n$.

5.1.2 The Lanczos Algorithm

As discussed in Section 1.5 we understand the Lanczos algorithm as a randomized procedure 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 $1 \leq k \leq n$. Then, the procedure outputs 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 $r_1 \geq \cdots \geq r_k$. On the other hand, the eigenvectors of J_k give rise (after an orthonormal change of basis determined by the v_j below) to the Ritz vectors, that is, the approximations for the eigenvectors of A. We now describe how the procedure generates the Jacobi coefficients α_i and β_i.

Lanczos

Input: Hermitian $A \in \mathbb{C}^{n\times n}$, integer $k \in [n]$, $u \in \mathbb{S}^{n-1}$.
Output: $J_k \in \mathbb{C}^{k\times k}$

1. Initialize: $v_0 = u$.
2. **For** $j = 0, \dots, k-1$
   a) $W_j = \mathrm{span}\{v_0, \dots, v_j\}$.
   b) $\alpha_j = \langle Av_j, v_j\rangle$.
   c) $\beta_j = \|\mathrm{Proj}_{W_j^\perp}(Av_j)\|_2$.
   d) **If** $\beta_j = 0$ **Halt**.
   e) **Else** $v_{j+1} = \frac{\mathrm{Proj}_{W_j^\perp}(Av_j)}{\|\mathrm{Proj}_{W_j^\perp}(Av_j)\|_2}$.
3. Store the α_i and β_i in the tridiagonal matrix J_k in the standard way.

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 $u_1, \ldots, u_n$ be the coordinates of u when written in the eigenbasis of A. We define the probability measure

$$\mu^u = \sum_{i=1}^{n} u_i^2 \delta_{\lambda_i}. \tag{5.6}$$

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) continuous functions 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 outputed by the Lanczos algorithm are the Jacobi coefficients of the measure μ^u, and the Ritz values after k iterations are the roots of $p_k(x)$.

Note that this algorithm scales linearly in the input A, so to simplify notation, in some of the proofs below we will assume that $\|A\| = 1$.

5.2 Applying the Local Lévy Lemma

5.2.1 Strategy

As discussed in Section 1.5 we will identify a region of $\mathbb{S}^{n-1}$ on which the functions α_i and β_i have a controlled Lipschitz constant. For this, we first introduce 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 5.2.1. 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

$$d_2(f(x), f(y)) \leq c' d_1(x, y) \quad \forall x, y \in U.$$

Remark 5.2.2. For f defined on an open subset of $\mathbb{S}^{n-1}$, we have that f is locally Lipschitz continuous with constant c with respect to the geodesic metric if and only if it is locally Lipschitz continuous with the same constant with respect to the Euclidean ("chordal") metric.

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 5.2.3. 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 5.2.4. *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 $w_1, \ldots, w_n \in \alpha([0, 1])$ 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 $1 \ldots, n - 1$.

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 $i = 1, \ldots, n - 1$. 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 $\tilde{w}_1, \ldots, \tilde{w}_{n+1} \in \alpha([0, 1])$ 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 $i = 1, \ldots, n - 1$. 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.

5.2.2 Local Lipschitz Constants for Jacobi Coefficients

As can be seen from the definition of the Lanczos algorithm, the dependence of the quantities $\alpha_i(u)$, $\beta_i(u)$, and $v_j(u)$ on u is highly nonlinear, 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 (5.6).

Proposition 5.2.5. *Fix $\tilde{u} \in \mathbb{S}^{n-1}$ and let $v_j(u)$ be as above. 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 $W_x = \operatorname{span}\{v_0(x) = x, v_1(x), \ldots, v_j(x)\}$ 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

$$\begin{aligned}
&\|\operatorname{Proj}_{W_x^\perp}(Av_j(x)) - \operatorname{Proj}_{W_y^\perp}(Av_j(y))\| \\
&\leq \|\operatorname{Proj}_{W_x^\perp}(A(v_j(x) - v_j(y)))\| + \|(\operatorname{Proj}_{W_x^\perp} - \operatorname{Proj}_{Wy^\perp})(Av_j(y))\| \\
&= \|\operatorname{Proj}_{W_x^\perp}(A(v_j(x) - v_j(y)))\| + \|(\operatorname{Proj}_{W_x} - \operatorname{Proj}_{W_y})(Av_j(y))\|. \qquad (5.7)
\end{aligned}$$

From the induction hypothesis we have that, for any $\varepsilon > 0$, we can choose $\mathcal{U}$ small enough so that

$$\|\operatorname{Proj}_{W_x^\perp}(A(v_j(x) - v_j(y)))\| \leq \|A\|\|v_j(x) - v_j(y)\| \leq \|A\|((4\|A\|)^j \gamma_j(\tilde{u}) + \varepsilon)\|x - y\|. \qquad (5.8)$$

On the other hand, from the definition of the Lanczos algorithm it follows that $\beta_i(\tilde{u}) \leq \|A\|$ for every $i = 0, \ldots, n-1$, so in view of (5.2), 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

$$\|\operatorname{Proj}_{W_x}(w) - \operatorname{Proj}_{W_y}(w)\| \leq \sum_{i=0}^{j} \|\langle v_i(x), w\rangle v_i(x) - \langle v_i(y), w\rangle v_i(y)\| \qquad (5.9)$$

and we can bound each term on the right-hand side of (5.9) as follows:

$$\|\langle v_i(x), w\rangle v_i(x) - \langle v_i(y), w\rangle v_i(y)\| \leq |\langle v_i(x) - v_i(y), w\rangle| + \|v_i(x) - v_i(y)\||\langle v_i(y), w\rangle|$$

$$\begin{aligned}&\leq \|v_i(x) - v_i(y)\|\|w\| + \|v_i(x) - v_i(y)\|\|v_i(y)\|\|w\| \\ &\leq 2(4\|A\|)^i\gamma_i(\tilde{u})\|x - y\|.\end{aligned}$$

Hence, adding over i we obtain

$$\|\text{Proj}_{W_x}(w) - \text{Proj}_{W_y}(w)\| \leq \frac{2}{3} \cdot 4^{j+1}\|A\|^j\gamma_j(\tilde{u})\|x - y\|,$$

which implies that $\|\text{Proj}_{W_x} - \text{Proj}_{W_y}\| \leq \frac{2}{3} \cdot 4^{j+1}\|A\|^j\gamma_j(\tilde{u})\|x - y\|$ and hence

$$\|(\text{Proj}_{W_x} - \text{Proj}_{W_y})(Av_j(y))\| \leq \frac{2}{3} \cdot (4\|A\|)^{j+1}\gamma_j(\tilde{u})\|x - y\|. \tag{5.10}$$

Putting together inequalities (5.7), (5.8), and (5.10), we get for any $x, y \in \mathcal{U}$ that

$$\|\text{Proj}_{W_x^\perp}(Av_j(x)) - \text{Proj}_{W_y^\perp}(Av_j(y))\| \leq (4\|A\|)^{j+1}\gamma_j(\tilde{u})\|x - y\|.$$

With this we have established that the function $u \mapsto \text{Proj}_{W_u^\perp}(Av_j(u))$ 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 $\beta_j(\tilde{u}) = \|\text{Proj}_{W_{\tilde{u}}^\perp}(Av_j(\tilde{u}))\|$. 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

$$u \mapsto v_{j+1}(u) = f(\text{Proj}_{W_u^\perp}(Av_j(u)))$$

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 (5.2). □

Proposition 5.2.6. *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. Recall 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 $\beta_j(u) = \|\text{Proj}_{W_u^\perp}(Av_j(u)))\|$ and we established in the proof of Proposition 5.2.5 that this function is locally Lipschitz with constant $(4\|A\|)^{j+1}\gamma_j(\tilde{u})$, the proof is concluded. □

Remark 5.2.7. 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 6, sharpening our constants is of relevance for applications.

5.2.3 Incompressibility

In Section 5.3, 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 the measure of this region is large enough to apply the local Lévy lemma (Lemma 1.5.11).
2. That this region is K-connected for a small enough K.

First we give this region a name. Loosely inspired by the random matrix literature (see, for example, [164]), 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 5.2.8.

$$I_n(\delta,\varepsilon) = \left\{ u \in \mathbb{S}^{n-1} : \sum_{i\in S} u_i^2 > \varepsilon \text{ for all } S \subseteq \{1,2,\dots,n\}, |S| \geq \delta n \right\}.$$

For incompressible u we will prove an adequate bound on the local Lipschitz constant in Proposition 5.3.1. Fortunately, a uniform random unit vector u is incompressible with high probability, as we will now show.

Proposition 5.2.9. *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 5.2.10. *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 5.2.9. 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 the Proposition 5.2.9 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 [101].

Lemma 5.2.11. *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 5.2.9. Let $X_1, \ldots, X_n$ denote independent Gaussian random variables each with variance $1/n$, and let $X = (X_1, \ldots, X_n)$. If we set $u = X/\|X\|$, then u is uniformly distributed on the unit sphere; see e.g. [111].

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 5.2.11 with t being the solution to $k - 2\sqrt{kt} = 2n\varepsilon$. To avoid the bookkeeping of ceiling and floor functions we use

the extremely crude inequality $n\delta \le k \le 2n\delta$ (valid as long as $\delta n \ge 1$), which will suffice for our purposes:

$$\mathbb{P}[E] \le \exp\left\{2\delta(1+\log\delta^{-1})n - \left(\frac{\varepsilon}{\delta}-1\right)^2 n\right\}.$$

We now upper bound the probability of F:

$$\begin{aligned}
\mathbb{P}[F] &= \mathbb{P}\left[\sum_{i\in S}\left(X_i^2 - \frac{X_i^2}{\|X\|^2}\right) \ge \varepsilon \text{ for some } |S| > \delta n\right] \\
&= \mathbb{P}\left[\left(1-\frac{1}{\|X\|^2}\right)\sum_{i\in S} X_i^2 \ge \varepsilon \text{ for some } |S| > \delta n\right] \\
&\le \mathbb{P}\left[\left(1-\frac{1}{\|X\|^2}\right)\|X\|^2 \ge \varepsilon\right] \\
&= \mathbb{P}\left[\|X\|^2 \ge 1+\varepsilon\right].
\end{aligned}$$

Since $Y = n\|X\|^2$ is distributed as $\chi^2(n)$, we may apply the upper tail bound in Lemma 5.2.11 with $t = n\varepsilon^2/8$ to obtain

$$\mathbb{P}[F] \le \exp\{-n\varepsilon^2/8\}.$$

To conclude, we have $\mathbb{P}[u \notin I_n(\delta,\varepsilon)] \le \mathbb{P}[E] + \mathbb{P}[F]$, and substituting the bounds we just derived, we obtain the desired inequality. □

5.2.4 *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 5.2.12. *$I_n(\delta,\varepsilon)$ 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 $(2/\sqrt{\varepsilon})\|x-y\|$, which is at most $(2/\sqrt{\varepsilon})d(x,y)$, where d denotes the geodesic distance on $\mathbb{S}^{n-1}$.

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 $\delta n \le n_i \le 2\delta n$ for all i.[1] Note also that this implies $m \ge \frac{1}{\delta}$. 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

$$\sum_{i=1}^m \|x^{(i)} - y^{(i)}\| \le \sqrt{m}\|x - y\| \le \sqrt{1/\delta}\|x - y\|,$$

by the Cauchy-Schwarz inequality. Also, $\|P(t)\| \ge \sqrt{\varepsilon/2\delta}$, because

$$\|P_i(t)\|^2 \ge \sum_{j=1}^{i-1} \|x^{(j)}\|^2 + \sum_{j=i+1}^{m} \|y^{(j)}\|^2 \ge (m-1)\varepsilon \ge \frac{\varepsilon}{2\delta},$$

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,

$$\|P_i(t)\|^2 \le \sum_{j=1}^{m} \max\{\|x^{(j)}\|, \|y^{(j)}\|\}^2 \le \sum_{j=1}^{m} \left(\|x^{(j)}\|^2 + \|y^{(j)}\|^2\right) = 2.$$

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 $d\mathbf{s} = \sqrt{r^2 + (dr/d\theta)^2}\, d\theta \ge r\, d\theta$.

[1] This is possible as long as $n/2 \ge \delta n \ge 1$, which will be true in our regime.

Thus, finally, we have shown that the path P' is contained in $I_n(4\delta, \varepsilon/\sqrt{2})$ and has length at most

$$\sqrt{1/\delta}\|x-y\|(1/\sqrt{\varepsilon/2\delta}) = \sqrt{2/\varepsilon}\|x-y\|.$$

□

5.3 Concentration of the Output

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 5.2.6, 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 (5.1) and (5.6) 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 $k \ll n$ 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 recur to the definition of equidistribution (Definition 1.5.1) given in Section 1.5. Moreover, we will show in Section 5.3.1 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 from Section 5.2, 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 5.3.1. *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*

$$\gamma_k(u) \leq \frac{1}{\omega^k \sqrt{\varepsilon}}.$$

5.3.1 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 5.3.2 (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 5.3.3 (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

$$\frac{1}{\nu_1(\mathbb{R})}\int_{-\infty}^{\infty} V(x)d\nu_1(x) \le a$$

for some constant $a < \infty$. Then

$$a \ge \frac{1}{\nu_1(\mathbb{R})}\int_{-\infty}^{\infty} V(x)d\nu_1(x) \ge \frac{2a\nu_1(\{x \in \mathbb{R} : V(x) \ge 2a\})}{\nu_1(\mathbb{R})}.$$

Relating this back to the definition of equidistribution, we have

$$\nu_1\left(\left\{x \in \mathbb{R} : \frac{1}{|T|}\sum_{t\in T}\log|x-t| \ge -2a\right\}\right) = \nu_1(\{x \in \mathbb{R} : V(x) \le 2a\}) \ge \frac{1}{2}\nu_1(\mathbb{R}).$$

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 5.3.4. *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,

$$|\mu(\{x \in \mathbb{R} : p(x) \ge \omega^{|T|}\}) - \nu(\{x \in \mathbb{R} : p(x) \ge \omega^{|T|}\})| \le 4j\mathrm{Kol}(\mu, \nu).$$

□

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 5.3.5. *Let μ be a compactly supported probability measure with nontrivial absolutely continuous part. Let $\{\mu_n\}$ be a sequence of probability measures such that $\mathrm{Kol}(\mu_n, \mu) \le \frac{C}{\log n}$ for some $C > 0$. Then for all n, for all $j \le \frac{1}{2C}\log n$ we have that μ_n is (δ, ω, j)-equidistributed for some $\delta, \omega > 0$.*

Remark 5.3.6. 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 5.3.3.

Example 5.3.7. 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 5.3.8. 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 provided in Section 1.5 are now trivial.

Proof of Example 1.5.2 and Example 1.5.8. It is enough to put together Proposition 5.3.4 and Example 5.3.7. □

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 5.3.9. 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 5.3.10. A particular case of Observation 5.3.9 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 1.5.3 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 1.5.4 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.

5.3.2 Jacobi Coefficients

We now have the necessary tools to prove concentration for the entries of the Jacobi matrix.

Proposition 5.3.11 (Jacobi coefficients are globally Lipschitz)**.** *Suppose the spectrum of A 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 5.2.6 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 5.3.1 says that because the spectrum of A 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 5.2.12 says that $I_n(\delta, \varepsilon)$ is $\sqrt{2/\varepsilon}$-connected in the larger set $I_n(4\delta, \varepsilon/\sqrt{2})$, so Lemma 5.2.4 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.

Theorem 5.3.12 (Restatement of Theorem 1.5.4)**.** *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*

$$2\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + 2\exp\left\{-\frac{1}{64}\left(\frac{\omega}{4\|A\|}\right)^{2i}\delta^2t^2n\right\}. \tag{5.11}$$

Proof. The local Lévy lemma (Lemma 1.5.11) 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 5.3.11. 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 5.2.10 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. □

Combining the previous theorem with Corollary 5.3.5 we get convergence in probability of the Jacobi matrices in the regime $k = O(\log n)$.

Proposition 5.3.13. *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 $k_n \geq c_2 \log n$ 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 5.3.5, 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 1.5.4, 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. □

As mentioned in the introduction, convergence for *fixed* k to the infinite Jacobi matrix J of μ for deterministic μ_n (with no hypothesis on the rate of convergence of μ_n) is proven in [72, Theorem 4]. In Proposition 5.3.13 we leave it open to prove that the limit is actually J, 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 1.5.9, proven in Section 5.

5.3.3 Ritz Values

Theorem 1.5.4 yields concentration of the entries of the random matrix $J_k(u)$. So to control the Ritz values (which are the eigenvalues of $J_k(u)$) it is enough to apply Weyl's inequality (Lemma 1.1.5).

Following the notation in Theorem 1.5.4, 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 $\tilde{r}_1 \geq \cdots \geq \tilde{r}_k$.

Proposition 5.3.14 (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 the probability $\mathbb{P}[\|\vec{r}(u) - \vec{r}\|_\infty \geq t\|A\|]$ is bounded above by*

$$4k\left[\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32} n\right\} + \exp\left\{-\frac{1}{192}\left(\frac{\omega}{4\|A\|}\right)^{2k} \delta^2 t^2 n\right\}\right].$$

Proof. 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

$$\|J_k(u) - \tilde{J}_k\| \leq \max_{0\leq i\leq k-1}\{|\alpha_i(u) - \tilde{\alpha}_i|\} + 2\max_{0\leq i\leq k-2}\{|\beta_i(u) - \tilde{\beta}_i|\}. \tag{5.12}$$

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 1.1.5 and the second inequality from (5.12). Now observe that the event $\{\max_{0\leq i\leq k-1}\{|\alpha_i(u) - \tilde{\alpha}_i|\} + 2\max_{0\leq i\leq k-2}\{|\beta_i(u) - \tilde{\beta}_i|\} \geq t\}$ is contained in the event

$$\left\{\max_{0\leq i\leq k-1}\{|\alpha_i(u) - \tilde{\alpha}_i|\} \geq \frac{t}{3}\right\} \cup \left\{\max_{0\leq i\leq k-2}\{|\beta_i(u) - \tilde{\beta}_i|\} \geq \frac{t}{3}\right\},$$

which in turn is contained in the event

$$\bigcup_{i=1}^{k}\left\{|\alpha_i(u) - \tilde{\alpha}_i| \geq \frac{t}{3}\right\} \bigcup \left\{|\beta_i(u) - \tilde{\beta}_i| \geq \frac{t}{3}\right\}.$$

Using a union bound and applying Theorem 1.5.4, we obtain the desired result. □

5.3.4 Ritz Vectors

Here we will use the same notation as in Section 5.3.3. 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 $r_i(u)$. We will use the fact that $J_k(u)$ concentrates around $\tilde{J}_k$, together with the Davis-Kahan theorem (Lemma 1.1.6) to establish the concentration of the vectors $w_i(u)$.

Under the assumption that $\tilde{r}_i(u)$ is not close to the other Ritz values, we get the following result.

Proposition 5.3.15 (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 $0 \leq c < 1/2$, the probability $\mathbb{P}[\sin\theta \geq 2\|A\|/\varepsilon n^c]$ is bounded above by*

$$4k\left[\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + \exp\left\{-\frac{1}{192}\left(\frac{\omega}{4\|A\|}\right)^{2k}\delta^2 n^{1-2c}\right\}\right].$$

Note. The same result holds for the Ritz vectors, since these are obtained by applying an isometry to the $w_i(u)$.

Proof. From Theorem 1.1.6 we have that

$$\sin\theta \leq \frac{2\|\tilde{J}_k(u) - \tilde{J}_k(u)\|}{\varepsilon}$$

and hence

$$\begin{aligned}\mathbb{P}[\sin\theta \geq t] &\leq \mathbb{P}[\|J_k(u) - \tilde{J}_k\| \geq t] \\ &\leq \mathbb{P}\left[\max_{0\leq i\leq k-1}\{|\alpha_i(u) - \tilde{\alpha}_i|\} + 2\max_{0\leq i\leq k-2}\{|\beta_i(u) - \tilde{\beta}_i|\} \geq t\right],\end{aligned}$$

where the latter inequality was established in the proof of Proposition 5.3.14. Using the bound obtained in the aforementioned proof and substituting $t = \frac{2}{\varepsilon n^c}$ we obtain the desired result. □

5.4 Location of the Output

5.4.1 Undetected Outliers

We now prove our theorem about the Lanczos algorithm missing outliers in the spectrum. First we start by showing an asymtptotic version of this result.

Proposition 5.4.1. *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 $\mathrm{Kol}(\mu_n, \mu) = O(1/\log n)$[2]. 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.)

Proof. By Proposition 5.3.4, 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 5.2.9. Then by Proposition 5.3.1, 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 2, it is a classical fact that the monic orthogonal polynomial

[2]Here and in Chapter 5 we will use $\mathrm{Kol}(\cdot,\cdot)$ to denote the Kolmogorov-Smirnov distance between two measures.

of any given degree has minimal L^2 norm over all monic polynomials of that degree. Thus, we in fact have

$$\int q(x)^2 \, d\mu^u(x) \geq \varepsilon\omega^{2j} \tag{5.13}$$

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 $\rho(u) \leq R + \kappa$ it suffices to show that $\|P_t\|_{L^2(\mu^u)}$ is strictly increasing in t for $t > R + \kappa$. We have

$$\|P_t\|_{L^2(\mu^u)}^2 = \int \left(\frac{\pi_k^u(x)}{x - \rho(u)}(x - t) \right)^2 d\mu^u(x) = \sum_{i=1}^{k} u_i^2(\lambda_i - t)^2 \prod_{j=2}^{k} (\lambda_i - r_j)^2,$$

where we let $r_2, \ldots, r_k$ denote the roots of $p_k^u(x)$ besides the maximum root $\rho(u)$, and we omit the argument u for brevity. We calculate the derivative

$$\frac{d}{dt}\|P_t\|_{L^2(\mu^u)}^2 = -2\sum_{i=1}^{m} u_i^2(\lambda_i - t) \prod_{j=1}^{k-1} (\lambda_i - r_j)^2 - 2\sum_{i=m+1}^{n} u_i^2(\lambda_i - t) \prod_{j=2}^{k} (\lambda_i - r_j)^2.$$

We wish to show that this quantity is positive whenever $t \geq R + \kappa$. We have assumed that there are only m outliers, so assume $\lambda_i \leq R$ for all $i > m$. Then $t - \lambda_i \geq \kappa$ for every $m < i \leq n$.

Thus,

$$\begin{aligned}
\frac{d}{dt}\|P_t\|_{L^2(\mu^u)}^2 &\geq -2\sum_{i=1}^{m} u_i^2(\lambda_i - t) \prod_{j=1}^{k-1} (\lambda_i - r_j)^2 + 2\sum_{i=m+1}^{n} u_i^2 \kappa \prod_{j=2}^{k} (\lambda_i - r_j)^2 \\
&= -2\sum_{i=1}^{m} u_i^2(\lambda_i - t) \prod_{j=2}^{k} (\lambda_i - r_j)^2 \\
&\quad + \left[2\kappa \int \left(\frac{p_k^u(x)}{x - \rho(u)} \right)^2 d\mu^u(x) - 2\sum_{i=1}^{m} u_i^2 \kappa \prod_{j=2}^{k} (\lambda_i - r_j)^2 \right] \\
&\geq -2\sum_{i=1}^{m} u_i^2(\lambda_i - t) \prod_{j=2}^{k} (\lambda_i - r_j)^2 + 2\kappa\varepsilon\omega^{2(k-1)} - 2\sum_{i=1}^{m} u_i^2 \kappa \prod_{j=2}^{k} (\lambda_i - r_j)^2,
\end{aligned}$$

where in the last step we used the inequality (5.13) on the degree $k-1$ polynomial $p_k^u(x)/(x - \rho(u))$. Simplifying, we have

$$\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} \geq 2\kappa\varepsilon\omega^{2(k-1)} - 2\sum_{i=1}^{m} u_i^2(\lambda_i + \kappa - t)\prod_{j=2}^{k}(\lambda_i - r_j)^2.$$

By uniform boundedness of the spectra, there exists M large such that $\lambda_i - r_j \leq M$ for all $1 \leq i \leq m$. Let g be the maximum of the outlier gaps $\lambda_i - R$ over all $1 \leq i \leq m$. Recall that $t \geq R + \kappa$, so $\lambda_i + \kappa - t \leq \lambda_i - R \leq g$ for all $1 \leq i \leq m$. 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 5.4.3 below. Putting this all together, we have

$$\frac{d}{dt}\|P_t\|^2_{L^2(\mu^u)} \geq 2\kappa\varepsilon\omega^{2k-2} - 2n^{-c}M^{2k-2}mg.$$

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

$$2k - 2 < \frac{1}{\log\frac{M}{\omega}}\left(c\log n + \log\frac{\kappa\varepsilon}{mg}\right). \tag{5.14}$$

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 5.4.2. 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.

We now turn our attention towards proving Theorem 1.5.7. First we will need the following mass concentration lemma.

Lemma 5.4.3. *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 5.2.9. Define X_i as in that proof. Then

$$\mathbb{P}\left[\sum_{i=1}^m u_i^2 > n^{-c}\right] \le \mathbb{P}\left[\sum_{i=1}^m X_i^2 > \frac{1}{2}n^{-c}\right] + \mathbb{P}\left[\sum_{i=1}^m X_i^2 < -\frac{1}{2}n^{-c} + \sum_{i=1}^m u_i^2\right].$$

Using Lemma 5.2.11, 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

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^m X_i^2 > \frac{1}{2}n^{-c}\right] &\le \exp\left\{-\left(\frac{-2\sqrt{m}+\sqrt{2}n^{\frac{1}{2}-\frac{c}{2}}}{4}\right)^2\right\} \\ &= \exp\left\{-\frac{1}{16}\left(4n^{\alpha} - 4\sqrt{2}n^{\frac{1}{2}-\frac{c}{2}+\frac{\alpha}{2}} + 2n^{1-c}\right)\right\},\end{aligned}$$

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 5.2.9 and using Lemma 5.2.11 again, we get

$$\mathbb{P}\left[\sum_{i=1}^m X_i^2 < -\frac{1}{2}n^{-c} + \sum_{i=1}^m u_i^2\right] \le \exp\left\{-\frac{1}{16}n^{1-2c}\right\}.$$

□

We can now show Theorem 1.5.7.

Proof of Theorem 1.5.7. From the proof of Proposition 5.4.1, setting $\varepsilon = \delta/2$ we have that the Ritz values are contained in the desired interval for

$$k \le \frac{1}{2\log\frac{M}{\omega}}\left(c\log n + \log\frac{\kappa\delta}{2mg}\right)$$

as long as $k \le j$, $u \in I_n(\delta, \delta/2)$ and $\sum_{i=1}^m u_i^2 > n^{-c}$. Applying Corollary 5.2.10, the probability that u violates either condition is at most

$$\begin{aligned}&\mathbf{P}[u \notin I_n(\delta,\delta/2)] + \mathbb{P}\left[\sum_{i=1}^m u_i^2 > n^{-c}\right] \\ &\le 2\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + \mathbb{P}\left[\sum_{i=1}^m u_i^2 > n^{-c}\right] \\ &\le 2\exp\left\{-\frac{\min\{\delta, 1/50\}^2}{32}n\right\} + 2\exp\left\{-\frac{1}{16}n^{1-2c}\right\},\end{aligned}$$

where in the last step, we apply Lemma 5.4.3 and note that for $n \ge e^{\frac{1}{1-c-\alpha}}$ we have $4\sqrt{2}n^{\frac{1-c+\alpha}{2}} \le n^{1-c}$. □

5.4.2 Asymptotic Locations of Ritz Values and Jacobi Coefficients

For $C > 0$ let $\mathcal{P}_C$ denote the space of Borel probability measures supported on $[-C, C]$. In order to prove Theorem 1.5.9 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 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 5.2.6. It remains an open question if a better rate can be achieved at this level of generality; see the concluding remarks for some natural directions to pursue.

We will use the following well known result which, for convenience of the reader, we restate as it appears in Lemma 1.1 in [76].

Lemma 5.4.4. *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 5.4.5. *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*

$$\|A^{(i)} - B^{(i)}\|_2 \leq \varepsilon \quad \text{and} \quad \max\{\|A^{(i)}\|_2, \|B^{(i)}\|_2\} \leq C. \tag{5.15}$$

Then

$$|\det(A) - \det(B)| \leq \varepsilon k (C+\varepsilon)^{k-1}.$$

Proof. By the assumption in (5.15) we can write $B = A + E$, where E is a matrix with columns of norm less than or equal to ε. Then, using Lemma 5.4.4, the inequalities in (5.15), 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

$$m_k(\mu) = k\int_0^\infty x^{k-1}\mu(x,\infty)dx. \tag{5.16}$$

This identity is enough to obtain the following bound.

Lemma 5.4.6. *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 (5.16) 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 2.1 will be used.

Proposition 5.4.7. *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 $j=1,\dots,k$. 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 5.4.6 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 5.4.5 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 (5.3) we have that

$$\begin{aligned} |\beta_k^\mu - \beta_k^\nu| &= \left| \frac{\sqrt{x_{k-1}x_{k+1}}}{x_k} - \frac{\sqrt{y_{k-1}y_{k+1}}}{y_k} \right| \\ &\leq \frac{1}{x_k} |\sqrt{x_{k-1}x_{k+1}} - \sqrt{y_{k-1}y_{k+1}}| + \sqrt{y_{k-1}y_{k+1}} \left| \frac{1}{x_k} - \frac{1}{y_k} \right|. \end{aligned} \tag{5.17}$$

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

$$\frac{1}{x_k} |\sqrt{x_{k-1}x_{k+1}} - \sqrt{y_{k-1}y_{k+1}}| \leq \frac{\exp\{gk^2\}\mathrm{Kol}(\mu,\nu)}{2s_k^2}. \tag{5.18}$$

On the other hand,

$$\sqrt{y_{k-1}y_{k+1}} \left| \frac{1}{x_k} - \frac{1}{y_k} \right| = \sqrt{y_{k-1}y_{k+1}} \frac{|x_k - y_k|}{x_k y_k} \leq \frac{\exp\{gk^2\}\mathrm{Kol}(\mu,\nu)}{2s_k^2}. \tag{5.19}$$

The result then follows from combining the previous inequalities (5.17), (5.18), and (5.19). □

Remark 5.4.8. The constants s_k have already been studied with sophisticated techniques for some families of measures; see [148] for an example. However, using results only from Section 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 1.5.9.

In a similar fashion we can show that the coefficients of $p_k^\mu(x)$ are locally Lipschitz.

Proposition 5.4.9. *Fix a positive integer k. Let μ, ν and s_k be as in Proposition 5.4.7. 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 (5.5) 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 5.4.7 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 (5.18). The result follows. □

Corollary 5.4.10. *Let μ, ν, s_k be as in Proposition 5.4.7. 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 5.4.9 and Lemma 5.4.6 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 1.5.9 and Proposition 1.5.10 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 5.4.11. *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 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 5.2.11 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 are the solutions to

$$n^{-\frac{1}{4}} = \frac{2\sqrt{kt_1}}{n} \quad \text{and} \quad n^{-\frac{1}{4}} = \frac{2\sqrt{kt_2} + 2t_2}{n}, \tag{5.20}$$

respectively. Since $k \leq n$ it is clear from (5.20) 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 (5.2.9) we can show by a standard coupling argument that if we take $u_i = X_i/\sqrt{Z_n}$, we will have that

$$\mathbb{P}\left[\max_{1 \leq k \leq n} \left| Z_k - \sum_{i=1}^{k} u_i^2 \right| > n^{-\frac{1}{4}} \right] \leq \frac{1}{2} \exp\{-n^{\frac{1}{4}}/8\}$$

and the result follows. □

Proof of Theorem 1.5.9. From Lemma 5.4.11, 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 5.4.7 and Corollary 5.4.10 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 5.3.3 and Corollary 5.3.5 that $|\log(\gamma_k^\mu)| = O(k)$. Hence, from (5.4) 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 1.5.10. As mentioned in Section 2, this proposition is a direct consequence of Theorem 1.5.9 and Lemma 1.1.5. □

Remark 5.4.12. 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 that we pay for by getting a rate of $O(\sqrt{\log n})$ instead of $\Theta(\log n)$. At first glance, it may seem that the results in Section 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.

Bibliography

[1] Kensuke Aishima, Takayasu Matsuo, Kazuo Murota, and Masaaki Sugihara. "A Wilkinson-like multishift QR algorithm for symmetric eigenvalue problems and its global convergence". In: *Journal of Computational and Applied Mathematics* 236.15 (2012), pp. 3556–3560.

[2] Michael Aizenman, Ron Peled, Jeffrey Schenker, Mira Shamis, and Sasha Sodin. "Matrix regularizing effects of Gaussian perturbations". In: *Communications in Contemporary Mathematics* 19.03 (2017), p. 1750028.

[3] Diego Armentano, Carlos Beltrán, Peter Bürgisser, Felipe Cucker, and Michael Shub. "A stable, polynomial-time algorithm for the eigenpair problem". In: *Journal of the European Mathematical Society* 20.6 (2018), pp. 1375–1437.

[4] Diego Armentano and Felipe Cucker. "A randomized homotopy for the Hermitian eigenpair problem". In: *Foundations of Computational Mathematics* 15.1 (2015), pp. 281–312.

[5] Zhaojun Bai and James Demmel. "On a block implementation of Hessenberg multishift QR iteration". In: *International Journal of High Speed Computing* 1.01 (1989), pp. 97–112.

[6] 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.

[7] 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.

[8] Grey Ballard, James Demmel, and Ioana Dumitriu. "Minimizing communication for eigenproblems and the singular value decomposition". In: *arXiv preprint arXiv:1011.3077* (2010).

[9] Grey Ballard, James Demmel, Ioana Dumitriu, and Alexander Rusciano. "A Generalized Randomized Rank-Revealing Factorization". In: *arXiv preprint arXiv:1909.06524* (2019).

[10] Jess Banks, Jorge Garza-Vargas, Archit Kulkarni, and Nikhil Srivastava. "Overlaps, eigenvalue gaps, and pseudospectrum under real Ginibre and absolutely continuous perturbations". In: *arXiv preprint arXiv:2005.08930* (2020).

[11] Jess Banks, Jorge Garza-Vargas, Archit Kulkarni, and Nikhil Srivastava. "Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time". In: *Foundations of Computational Mathematics* (2022), pp. 1–89.

[12] Jess Banks, Jorge Garza-Vargas, and Nikhil Srivastava. "Global Convergence of Hessenberg Shifted QR I: Dynamics". In: *arXiv preprint arXiv:2111.07976* (2021).

[13] Jess Banks, Jorge Garza-Vargas, and Nikhil Srivastava. "Global Convergence of Hessenberg Shifted QR II: Numerical Stability". In: *arXiv preprint arXiv:2205.06810* (2022).

[14] Jess Banks, Jorge Garza-Vargas, and Nikhil Srivastava. "Global Convergence of Hessenberg Shifted QR III: Approximate Ritz Values via Shifted Inverse Iteration". In: *arXiv preprint arXiv:2205.06804* (2022).

[15] Jess Banks, Archit Kulkarni, Satyaki Mukherjee, and Nikhil Srivastava. "Gaussian regularization of the pseudospectrum and Davies' conjecture". In: *Communications on Pure and Applied Mathematics* 74.10 (2021), pp. 2114–2131.

[16] Steve Batterson. "Convergence of the Francis shifted QR algorithm on normal matrices". In: *Linear algebra and its applications* 207 (1994), pp. 181–195.

[17] Steve Batterson. "Convergence of the shifted QR algorithm on 3× 3 normal matrices". In: *Numerische Mathematik* 58.1 (1990), pp. 341–352.

[18] Steve Batterson. "Dynamical analysis of numerical systems". In: *Numerical linear algebra with applications* 2.3 (1995), pp. 297–310.

[19] Steve Batterson and David Day. "Linear convergence in the shifted QR Algorithm". In: *mathematics of computation* 59.199 (1992), pp. 141–151.

[20] Steve Batterson and John Smillie. "Rayleigh quotient iteration for nonsymmetric matrices". In: *mathematics of computation* 55.191 (1990), pp. 169–178.

[21] Steve Batterson and John Smillie. "The dynamics of Rayleigh quotient iteration". In: *SIAM journal on numerical analysis* 26.3 (1989), pp. 624–636.

[22] 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.

[23] 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.

[24] Mohammed Bellalij, Yousef Saad, and Hassane Sadok. "Further analysis of the Arnoldi process for eigenvalue problems". In: *SIAM J. Numer. Anal.* 48.2 (2010), pp. 393–407.

[25] 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.

[26] Rajendra Bhatia. *Perturbation bounds for matrix eigenvalues*. SIAM, 2007.

[27] David Bindel, Shivkumar Chandresekaran, James Demmel, David Garmire, and Ming Gu. *A fast and stable nonsymmetric eigensolver for certain structured matrices*. Tech. rep. Technical report, University of California, Berkeley, CA, 2005.

[28] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. *Complexity and real computation*. Springer Science & Business Media, 1998.

[29] Charles Bordenave, Djalil Chafaı, et al. "Around the circular law". In: *Probability surveys* 9 (2012), pp. 1–89.

[30] Paul Bourgade and Guillaume Dubach. "The distribution of overlaps between eigenvectors of Ginibre matrices". In: *Probability Theory and Related Fields* (2019), pp. 1–68.

[31] Karen Braman, Ralph Byers, and Roy Mathias. "The multishift QR algorithm. Part I: Maintaining well-focused shifts and level 3 performance". In: *SIAM Journal on Matrix Analysis and Applications* 23.4 (2002), pp. 929–947.

[32] Karen Braman, Ralph Byers, and Roy Mathias. "The multishift QR algorithm. Part II: Aggressive early deflation". In: *SIAM Journal on Matrix Analysis and Applications* 23.4 (2002), pp. 948–973.

[33] Hendrik Jan Buurema. *A geometric proof of convergence for the QR method*. Vol. 62. 1958.

[34] Ralph Byers. "Numerical stability and instability in matrix sign function based algorithms". In: *Computational and Combinatorial Methods in Systems Theory*. Citeseer. 1986.

[35] 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.

[36] 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.

[37] 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.

[38] John T. Chalker and Bernhard Mehlig. "Eigenvector statistics in non-Hermitian random matrix ensembles". In: *Physical review letters* 81.16 (1998), p. 3367.

[39] Moody T Chu. "Linear algebra algorithms as dynamical systems". In: *Acta Numerica* 17 (2008), pp. 1–86.

[40] Giorgio Cipolloni, László Erdos, and Dominik Schröder. "On the condition number of the shifted real Ginibre ensemble". In: *SIAM Journal on Matrix Analysis and Applications* 43.3 (2022), pp. 1469–1487.

[41] Giorgio Cipolloni, László ErdHos, and Dominik Schröder. "Fluctuation around the circular law for random matrices with real entries". In: *arXiv preprint arXiv:2002.02438* (2020).

[42] Giorgio Cipolloni, László ErdHos, and Dominik Schröder. "Optimal Lower Bound on the Least Singular Value of the Shifted Ginibre Ensemble". In: *arXiv preprint arXiv:1908.01653* (2019).

[43] Robert M. Corless, Gaston H. Gonnet, David E. G. Hare, David J. Jeffrey, and Donald E. Knuth. "On the Lambert W function". In: *Advances in Computational mathematics* 5.1 (1996), pp. 329–359.

[44] E Brian Davies. "Approximate diagonalization". In: *SIAM Journal on Matrix Analysis and Applications* 29.4 (2007), pp. 1051–1064.

[45] E. Brian Davies. "Approximate diagonalization". In: *SIAM Journal on Matrix Analysis and Applications* 29.4 (2007), pp. 1051–1064.

[46] Chandler Davis and William M Kahan. "Some new bounds on perturbation of subspaces". In: *Bull. Amer. Math. Soc.* 75.4 (1969), pp. 863–868.

[47] David Day. "How the QR algorithm fails to converge and how to fix it". In: (1996).

[48] Percy Deift. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*. Vol. 3. American Mathematical Soc., 1999.

[49] Percy Deift, Tara Nanda, and Carlos Tomei. "Ordinary differential equations and the symmetric eigenvalue problem". In: *SIAM Journal on Numerical Analysis* 20.1 (1983), pp. 1–22.

[50] Percy Deift and Thomas Trogdon. "Universality in numerical computation with random data: Case studies and analytical results". In: *Journal of Mathematical Physics* 60.10 (2019), p. 103306.

[51] Percy A Deift, Govind Menon, Sheehan Olver, and Thomas Trogdon. "Universality in numerical computations with random data". In: *Proceedings of the National Academy of Sciences* 111.42 (2014), pp. 14973–14978.

[52] TJ Dekker and JF Traub. "The shifted QR algorithm for Hermitian matrices". In: *Linear Algebra Appl* 4 (1971), pp. 137–154.

[53] James Demmel, Ioana Dumitriu, and Olga Holtz. "Fast linear algebra is stable". In: *Numerische Mathematik* 108.1 (2007), pp. 59–91.

[54] James Demmel, Ioana Dumitriu, Olga Holtz, and Robert Kleinberg. "Fast matrix multiplication is stable". In: *Numerische Mathematik* 106.2 (2007), pp. 199–224.

[55] James W Demmel. *Applied numerical linear algebra*. SIAM, 1997.

[56] James W. Demmel. "The probability that a numerical analysis problem is difficult". In: *Mathematics of Computation* 50.182 (1988), pp. 449–480.

[57] James Weldon Demmel. "On condition numbers and the distance to the nearest ill-posed problem". In: *Numerische Mathematik* 51.3 (1987), pp. 251–289.

[58] 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.

[59] Jack Dongarra and Francis Sullivan. "Guest editors' introduction: The top 10 algorithms". In: *IEEE Computer Architecture Letters* 2.01 (2000), pp. 22–23.

[60] Ioana Dumitriu. "Smallest eigenvalue distributions for two classes of β-Jacobi ensembles". In: *Journal of Mathematical Physics* 53.10 (2012), p. 103301.

[61] PJ Eberlein and CP Huang. "Global convergence of the QR algorithm for unitary matrices with some results for normal matrices". In: *SIAM Journal on Numerical Analysis* 12.1 (1975), pp. 97–104.

[62] Alan Edelman. "Eigenvalues and condition numbers of random matrices". In: *SIAM Journal on Matrix Analysis and Applications* 9.4 (1988), pp. 543–560.

[63] Alan Edelman, Eric Kostlan, and Michael Shub. "How many eigenvalues of a random matrix are real?" In: *Journal of the American Mathematical Society* 7.1 (1994), pp. 247–267.

[64] Alan Edelman and N. Raj Rao. "Random matrix theory". In: *Acta Numerica* 14 (2005), pp. 233–297.

[65] 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.

[66] Jiang Erxiong. "A note on the double-shift QL algorithm". In: *Linear algebra and its applications* 171 (1992), pp. 121–132.

[67] Peter J. Forrester. *Log-gases and random matrices (LMS-34)*. Princeton University Press, 2010.

[68] John GF Francis. "The QR transformation a unitary analogue to the LR transformation—Part 1". In: *The Computer Journal* 4.3 (1961), pp. 265–271.

[69] John GF Francis. "The QR transformation—Part 2". In: *The Computer Journal* 4.4 (1962), pp. 332–345.

[70] 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.

[71] Jorge Garza-Vargas and Archit Kulkarni. "The Lanczos algorithm under few iterations: Concentration and location of the output". In: *SIAM Journal on Matrix Analysis and Applications* 41.3 (2020), pp. 1312–1346.

[72] Walter Gautschi. "Construction of Gauss-Christoffel quadrature formulas". In: *Math. Comp.* 22.102 (1968), pp. 251–270.

[73] Stephen Ge. "The Eigenvalue Spacing of IID Random Matrices and Related Least Singular Value Results". PhD thesis. UCLA, 2017.

[74] Viacheslav Leonidovich Girko. *Theory of random determinants*. Vol. 45. Springer Science & Business Media, 2012.

[75] E Gluskin and A Olevskii. "Invertibility of sub-matrices and the octahedron width theorem". In: *Israel Journal of Mathematics* 186.1 (2011), pp. 61–68.

[76] Chris Godsil. *Algebraic combinatorics*. Routledge, 2017.

[77] Gene Golub and Frank Uhlig. "The QR algorithm: 50 years later its genesis by John Francis and Vera Kublanovskaya and subsequent developments". In: *IMA Journal of Numerical Analysis* 29.3 (2009), pp. 467–485.

[78] Gene Golub and Richard Underwood. "The block Lanczos method for computing eigenvalues". In: *Mathematical software*. Elsevier, 1977, pp. 361–377.

[79] Gene Golub and Charles Van Loan. *Matrix computations. Johns Hopkins studies in the mathematical sciences*. 1996.

[80] 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.

[81] 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.

[82] 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.

[83] Roger Haydock. "The recursive solution of the Schrödinger equation". In: *Comput. Phys. Commun.* 20.1 (1980), pp. 11–16.

[84] Nicholas J Higham, Mark R Dennis, Paul Glendinning, Paul A Martin, Fadil Santosa, and Jared Tanner. *The Princeton companion to applied mathematics*. Princeton University Press Princeton, NJ, USA: 2015.

[85] Nicholas J. Higham. *Accuracy and stability of numerical algorithms*. Vol. 80. SIAM, 2002.

[86] Nicholas J. Higham. *Functions of matrices: theory and computation*. Vol. 104. SIAM, 2008.

[87] 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.

[88] 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.

[89] Roger A. Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[90] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

[91] Ilse CF Ipsen. "Computing an eigenvector with inverse iteration". In: *SIAM review* 39.2 (1997), pp. 254–291.

[92] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. "On the Real Davies' Conjecture". In: *arxiv preprint arXiv:2005.08908* (2020).

[93] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. "On the real Davies' conjecture". In: *The Annals of Probability* 49.6 (2021), pp. 3011–3031.

[94] Shmuel Kaniel. "Estimates for some computational techniques in linear algebra". In: *Math. Comp.* 20.95 (1966), pp. 369–378.

[95] Charles S. Kenney and Alan J Laub. "The matrix sign function". In: *IEEE Transactions on Automatic Control* 40.8 (1995), pp. 1330–1348.

[96] Daniel Kressner. In: *Personal communication* (2021).

[97] Daniel Kressner. "On the use of larger bulges in the QR algorithm". In: *Electronic Transactions on Numerical Analysis* 20.ARTICLE (2005), pp. 50–63.

[98] Daniel Kressner. "The effect of aggressive early deflation on the convergence of the QR algorithm". In: *SIAM journal on matrix analysis and applications* 30.2 (2008), pp. 805–821.

[99] Vera N Kublanovskaya. "On some algorithms for the solution of the complete eigenvalue problem". In: *USSR Computational Mathematics and Mathematical Physics* 1.3 (1962), pp. 637–657.

[100] Jacek Kuczyński and Henryk Woźniakowski. "Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm". In: *SIAM J. Matrix Anal. Appl.* 15.2 (1994), pp. 672–691.

[101] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Ann. Statist.* (2000), pp. 1302–1338.

[102] Ricardo S Leite, Nicolau C Saldanha, and Carlos Tomei. "Dynamics of the symmetric eigenvalue problem with shift strategies". In: *International Mathematics Research Notices* 2013.19 (2013), pp. 4382–4412.

[103] Lin Lin, Yousef Saad, and Chao Yang. "Approximating spectral densities of large matrices". In: *SIAM Rev.* 58.1 (2016), pp. 34–65.

[104] Galyna Livshyts, Grigoris Paouris, and Peter Pivovarov. "On sharp bounds for marginal densities of product measures". In: *Israel Journal of Mathematics* 216.2 (2016), pp. 877–889.

[105] 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.

[106] Kyle Luh and Sean O'Rourke. *Eigenvectors and controllability of non-Hermitian random matrices and directed graphs*. 2020. arXiv: 2004.10543 [math.PR].

[107] Alexander N. Malyshev. "Parallel algorithm for solving some spectral problems of linear algebra". In: *Linear algebra and its applications* 188 (1993), pp. 489–520.

[108] Francesco Mezzadri. "How to generate random matrices from the classical compact groups". In: *arXiv preprint math-ph/0609050* (2006).

[109] Cleve Moler. "Variants of the QR Algorithm". In: *Cleve's Corner, Mathworks Technical Articles* (2014).

[110] Cleve B Moler. "Three research problems in numerical linear algebra". In: *AMS Proceedings of Symposia in Applied Math, vol. 22* (1978), pp. 1–18.

[111] Mervin E Muller. "A note on a method for generating points uniformly on n-dimensional spheres". In: *Commun. ACM* 2.4 (1959), pp. 19–20.

[112] 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.

[113] Yuji Nakatsukasa and Nicholas J Higham. "Backward stability of iterations for computing the polar decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 33.2 (2012), pp. 460–479.

[114] Assaf Naor and Pierre Youssef. "Restricted invertibility revisited". In: *A journey through discrete mathematics*. Springer, 2017, pp. 657–691.

[115] 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.

[116] Hoi H. Nguyen. "Random matrices: Overcrowding estimates for the spectrum". In: *Journal of functional analysis* 275.8 (2018), pp. 2197–2224.

[117] Alexander M Ostrowski. "On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. I". In: *Archive for Rational Mechanics and Analysis* 1.1 (1957), pp. 233–241.

[118] Christopher Conway Paige. "The computation of eigenvalues and eigenvectors of very large sparse matrices". PhD thesis. University of London, 1971.

[119] Victor Y Pan. "Univariate polynomials: nearly optimal algorithms for numerical factorization and root-finding". In: *Journal of Symbolic Computation* 33.5 (2002), pp. 701–733.

[120] 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.

[121] Beresford Parlett. "Normal Hessenberg and moment matrices". In: *Linear Algebra and its Applications* 6 (1973), pp. 37–43.

[122] Beresford Parlett. "Singular and invariant matrices under the QR transformation". In: *Mathematics of Computation* 20.96 (1966), pp. 611–615.

[123] Beresford N Parlett. "The QR algorithm". In: *Computing in Science & Engineering* 2.1 (2000), pp. 38–42.

[124] Beresford N Parlett. "The Rayleigh quotient iteration and some generalizations for nonnormal matrices". In: *Mathematics of Computation* 28.127 (1974), pp. 679–693.

[125] Beresford N Parlett. *The symmetric eigenvalue problem*. SIAM, 1998.

[126] Beresford N Parlett and Jian Le. "Forward instability of tridiagonal QR". In: *SIAM Journal on Matrix Analysis and Applications* 14.1 (1993), pp. 279–316.

[127] Beresford N. Parlett and William Kahan. "On the convergence of a practical QR algorithm." In: *IFIP Congress (1)*. 1968, pp. 114–118.

[128] G Peters and James H Wilkinson. "Inverse iteration, ill-conditioned equations and Newton's method". In: *SIAM review* 21.3 (1979), pp. 339–360.

[129] Gwendoline Peters and James H Wilkinson. "The calculation of specified eigenvectors by inverse iteration". In: *Handbook for Automatic Computation*. Springer, 1971, pp. 418–439.

[130] Christian W. Pfrang, Percy Deift, and Govind Menon. "How long does it take to compute the eigenvalues of a random symmetric matrix". In: *Random Matrix Theory, Interacting Particle Systems, and Integrable Systems, Math. Sci. Res. Inst. Publ* 65 (2013), pp. 411–442.

[131] 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.

[132] Mark Rudelson and Roman Vershynin. "Small ball probabilities for linear images of high-dimensional distributions". In: *International Mathematics Research Notices* 2015.19 (2015), pp. 9594–9617.

[133] Mark Rudelson and Roman Vershynin. "The Littlewood–Offord problem and invertibility of random matrices". In: *Advances in Mathematics* 218.2 (2008), pp. 600–633.

[134] Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. Vol. 66. SIAM, 2011.

[135] 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.

[136] James George Sanderson. *A proof of convergence for the tridiagonal ql algorithm in floating-point arithmetic.* The University of New Mexico, 1976.

[137] 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.

[138] Meiyue Shao, Felipe H da Jornada, Lin Lin, Chao Yang, Jack Deslippe, and Steven G Louie. "A structure preserving Lanczos algorithm for computing the optical absorption spectrum". In: *SIAM J. Matrix Anal. Appl.* 39.2 (2018), pp. 683–711.

[139] 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).

[140] 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.

[141] Steve Smale. "Complexity theory and numerical analysis". In: *Acta Numerica* 6 (1997), pp. 523–551.

[142] Steve Smale. "On the efficiency of algorithms of analysis". In: *Bulletin (New Series) of The American Mathematical Society* 13.2 (1985), pp. 87–121.

[143] Piotr Śniady. "Random regularization of Brown spectral measure". In: *Journal of Functional Analysis* 193.2 (2002), pp. 291–313.

[144] 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.

[145] Gilbert W Stewart. "A Krylov–Schur algorithm for large eigenproblems". In: *SIAM Journal on Matrix Analysis and Applications* 23.3 (2002), pp. 601–614.

[146] Ji-Guang Sun. "Perturbation bounds for the Cholesky and QR factorizations". In: *BIT Numerical Mathematics* 31.2 (1991), pp. 341–352.

[147] Stanislaw J. Szarek. "Condition numbers of random matrices". In: *Journal of Complexity* 7.2 (1991), pp. 131–149.

[148] Gabor Szegö. "Hankel forms". In: *Amer. Math. Soc. Transl.* 108 (1977).

[149] Gabor Szegö. *Orthogonal polynomials*. Vol. 23. American Mathematical Society, 1939.

[150] Terence Tao and Van Vu. "Random matrices: The distribution of the smallest singular values". In: *Geometric And Functional Analysis* 20.1 (2010), pp. 260–297.

[151] Terence Tao, Van Vu, Manjunath Krishnapur, et al. "Random matrices: Universality of ESDs and the circular law". In: *The Annals of Probability* 38.5 (2010), pp. 2023–2065.

[152] R. C. Thompson. "The behavior of eigenvalues and singular values under perturbations of restricted rank". In: *Linear Algebra and its Applications* 13.1-2 (1976), pp. 69–78.

[153] Konstantin Tikhomirov. "Invertibility via distance for non-centered random matrices with continuous distributions". In: *arXiv preprint arXiv:1707.09656* (2017).

[154] Konstantin Tikhomirov. "Quantitative invertibility of non-Hermitian random matrices". In: *arXiv preprint arXiv:2206.00601* (2022).

[155] Francoise Tisseur. *Backward stability of the QR algorithm*. Tech. rep. Technical Report 239, Equipe d'Analyse Numerique, Universit e Jean Monnet de ..., 1996.

[156] Francoise Tisseur. *Backward stability of the QR algorithm*. Tech. rep. 239, UMR 5585, Lyon Saint-Etienne, 1996.

[157] Lloyd N. Trefethen and David Bau III. *Numerical linear algebra*. Vol. 50. SIAM, 1997.

[158] Lloyd N. Trefethen and Mark Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.

[159] Richard Ray Underwood. *An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems.* Stanford University, 1975.

[160] Walter Van Assche. "Padé and Hermite-Padé approximation and orthogonality". In: *Surv. Approx. Theory* 2 (2006), pp. 61–91.

[161] 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 J. Matrix Anal. Appl.* 22.3 (2001), pp. 837–852.

[162] Charles Van Loan. "On estimating the condition of eigenvalues and eigenvectors". In: *Linear Algebra and Its Applications* 88 (1987), pp. 715–732.

[163] JM Varah. "The calculation of the eigenvectors of a general complex matrix by inverse iteration". In: *Mathematics of Computation* 22.104 (1968), 785–s13.

[164] 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.

[165] 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.

[166] Tai-Lin Wang. "Convergence of the tridiagonal QR algorithm". In: *Linear algebra and its applications* 322.1-3 (2001), pp. 1–17.

[167] Tai-Lin Wang and William Gragg. "Convergence of the shifted QR Algorithm for unitary Hessenberg matrices". In: *Mathematics of computation* 71.240 (2002), pp. 1473–1496.

[168] Tai-Lin Wang and William Gragg. "Convergence of the unitary QR algorithm with a unimodular Wilkinson shift". In: *Mathematics of computation* 72.241 (2003), pp. 375–385.

[169] David S Watkins. "Forward stability and transmission of shifts in the QR algorithm". In: *SIAM Journal on Matrix Analysis and Applications* 16.2 (1995), pp. 469–487.

[170] David S Watkins. *The matrix eigenvalue problem: GR and Krylov subspace methods*. SIAM, 2007.

[171] David S Watkins. "The QR algorithm revisited". In: *SIAM review* 50.1 (2008), pp. 133–145.

[172] David S Watkins. "The transmission of shifts and shift blurring in the QR algorithm". In: *Linear algebra and its applications* 241 (1996), pp. 877–896.

[173] Helmut Wielandt. "Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben". In: *Mathematische Zeitschrift* 50.1 (1944), pp. 93–143.

[174] James H Wilkinson. "Global convergence of tridiagonal QR algorithm with origin shifts". In: *Linear Algebra and its Applications* 1.3 (1968), pp. 409–420.

[175] JH Wilkinson. "The algebraic eigenvalue problem". In: *Handbook for Automatic Computation, Volume II, Linear Algebra*. Springer-Verlag New York, 1971.

[176] Thomas G. Wright and Lloyd N. Trefethen. "Eigtool". In: *Software available at http://www.comlab. ox.ac.uk/pseudospectra/eigtool* (2002).

[177] Qiaochu Yuan, Ming Gu, and Bo Li. "Superlinear convergence of randomized block Lanczos algorithm". In: *2018 IEEE International Conference on Data Mining*. IEEE. 2018, pp. 1404–1409.

Appendix A

Spectral Stability Under Complex Ginibre Perturbations

Throughout this section we will use G_n to denote a complex normalized $n \times n$ Ginibre matrix.

A.1 Approach of Armentano et al.

There is an essentially different route (which can only be applied to the complex Gaussian case) to the smoothed analysis of minimum eigenvalue gap and eigenvector condition number discussed in Section 1.2. We'll begin by recalling some notation from [3], and direct the reader to their paper for an expanded treatment.

For any n let $\mathbb{P}(\mathbb{C}^n)$ denote the projective space associated to $\mathbb{C}^n$, and given $A \in \mathbb{C}^{n\times n}$, $\lambda \in \mathbb{C}$ and $v \in \mathbb{P}(\mathbb{C})$, define $A_{\lambda,v} : v^\perp \to v^\perp$ by

$$A_{\lambda,v} := P_v^\perp \circ (A - \lambda)|_{v^\perp}$$

where $v^\perp = \{x \in \mathbb{C}^n \mid \langle x, v\rangle = 0\}$ and $P_{v^\perp} : \mathbb{C}^n \to v^\perp$ denotes the orthogonal projection. With this in hand, [3] defines the condition number of a triple $(A, \lambda, v) \in \mathbb{C}^{n\times n} \times \mathbb{C} \times \mathbb{P}(\mathbb{C}^n)$ as

$$\mu(A, \lambda, v) := \begin{cases} \|A\|_F \|A_{\lambda,v}^{-1}\| & \text{if } A_{\lambda,v} \text{ is invertible,} \\ \infty & \text{otherwise.} \end{cases}$$

They similarly define the mean square condition number of a matrix as

$$\mu_{F,\mathsf{av}}(A) := \left(\frac{1}{n} \sum_{j=1}^{n} \|A\|_F^2 \|A_{\lambda_j,v_j}^{-1}\|_F^2 \right)^{\frac{1}{2}},$$

where (λ_j, v_j) are the eigenpairs of A. In particular, note that $\mu_{F,\mathsf{av}}(A) < \infty$ only when A has simple eigenvalues, and therefore $\mu_{F,\mathsf{av}}(A) < \infty$ implies that A is diagonalizable.

A.1.1 Controlling κ_V

To compare the notions of eigenvalue condition number and the condition number of a triple we recall the following theorem from [3]:

Theorem A.1.1 (Part of Proposition 2.7 of [3]). *Let $\mathcal{V}$ denote the* solution variety *for the eigenpair problem, defined as*

$$\mathcal{V} = \mathcal{V}_n := \{(A, \lambda, v) \in \mathbb{C}^{n\times n} \times \mathbb{C} \times \mathbb{P}(\mathbb{C}) \mid (A - \lambda)v = 0\},$$

and let $\Gamma : [0, 1] \to \mathcal{V}$, $\Gamma(t) = (A_t, \lambda_t, v_t)$ be a smooth curve such that A_t lies in the unit sphere of $\mathbb{C}^{n\times n}$ for all t. Then for all $t \in [0, 1]$,

$$|\dot{\lambda}_t| \leq \sqrt{1 + \mu(A_t, \lambda_t, v_t)^2}\|\dot{A}_t\|.$$

Now recall that $\kappa(\lambda)$ has the following variational description (see [80, Theorem 1], or deduce from (1.7)) for any a simple eigenpair (λ, v) of A, in terms of the derivatives of smooth curves going through the point (A, λ, v). Namely

$$\kappa(\lambda) = \sup_{\Gamma:[0,1]\to\mathcal{V},\, \Gamma(0)=(A,\lambda,v)} \frac{|\dot{\lambda}_0|}{\|\dot{A}_0\|}.$$

Hence, Theorem A.1.1 implies

$$\kappa(\lambda) \leq \sqrt{1 + \mu(A, \lambda, v)^2}. \tag{A.1}$$

It is then clear that $\mu_{F,\mathsf{av}}(A)$ can also be used to upper bound $\kappa_V(A)$. In view of this, we remind the reader of the following result from [3].

Theorem A.1.2 (Theorem 2.14 of [3]). *Let $G_n \in \mathbb{C}^{n\times n}$ denote a complex Ginibre matrix with $\mathcal{N}(0, 1_{\mathbb{C}}/n)$ entries. For any $A \in \mathbb{C}^{n\times n}$ and $\gamma > 0$, we have*

$$\mathbb{E}\left[\frac{\mu_{F,\mathsf{av}}(A + \gamma G_n)^2}{\|A + \gamma G_n\|_F^2}\right] \leq \frac{n^2}{\gamma^2}.$$

One can use this to derive results of the eigenvector condition number of perturbations of an arbitrary matrix, for example, the above directly implies Davies' conjecture [44] (for comparison, Theorem 1.1 of [15] is the same result but the exponent of n is $3/2$ instead of $5/2$.)

Proposition A.1.3. *Suppose $A \in \mathbb{C}^{n\times n}$ and $\gamma \in (0, 1)$. Then there is a matrix $E \in \mathbb{C}^{n\times n}$ such that $\|E\| \leq \gamma\|A\|$ and*

$$\kappa_V(A + E) \leq C\frac{n^{5/2}}{\gamma}$$

where C is an absolute constant.

Proof. Let λ_i, v_i denote the (random) eigenvalues and eigenvectors of $A+\gamma G_n$. Let B_r denote the event $\|A+\gamma G_n\|_F < r$. Because $\|G_n\|_F < 2\sqrt{n}$ with probability at least some absolute positive constant, for $r = \|A\| + 2\sqrt{n}$ the event B_r holds with that probability as well. Now note that

$$\begin{aligned}\mathbb{E}\left[\sum_i \kappa(\lambda_i)^2 \mid B_r\right] &\le \mathbb{E}\left[n + \sum_i \mu(A+\gamma G_n, \lambda_i, v_i)^2 \mid B_r\right] \qquad \text{by (A.1)}\\ &\le \mathbb{E}\left[n + n\mu_{F,\mathsf{av}}(A+\gamma G_n)^2 \mid B_r\right]\\ &\le n + \frac{n^3 r^2}{\gamma^2 \mathbb{P}[B_r]}, \end{aligned} \tag{A.2}$$

where in the last line we use Theorem A.1.2 and

$$\mathbb{E}\left[\frac{\mu_{F,\mathsf{av}}(A+\gamma G_n)^2}{r^2}\,\middle|\, B_r\right] \le \mathbb{E}\left[\frac{\mu_{F,\mathsf{av}}(A+\gamma G_n)^2}{\|A+\gamma G_n\|_F^2}\,\middle|\, B_r\right] \le \frac{\mathbb{E}\left[\frac{\mu_{F,\mathsf{av}}(A+\gamma G_n)^2}{\|A+\gamma G_n\|_F^2}\right]}{\mathbb{P}[B_r]}.$$

Using (1.8) we get

$$\mathbb{E}[\kappa_V(A+\gamma G_n)^2 | B_r] \le n\mathbb{E}\left[\sum_i \kappa(\lambda_i)^2 | B_r\right].$$

So, when $\|A\| = 1$ and $\gamma < 1$, if we set $r = \|A\| + 2\sqrt{n}$ as discussed above, the event B_r occurs with positive probability, and by (A.2) we know that $n\mathbb{E}\left[\sum \kappa(\lambda_i)^2 \mid B_r\right] \le \frac{Cn^5}{\gamma^2}$ for some constant C. It follows that there is some realization of G_n for which $\kappa_V(A+\gamma G_n)^2 \le \frac{Cn^5}{\gamma^2}$, as we wanted to show. □

One can also obtain tail bounds for $\kappa_V(A+\gamma G_n)$ of the sort discussed in Section 1.2, but we will not pursue this here since (at least naively) this method does not yield a tail bound as strong as that of Section A.2. However this method does yield the strongest known (left) tail bound for $\mathrm{gap}(A+\gamma G_n)$, as explained below.

A.1.2 Controlling gap

Let $A \in \mathbb{C}^{n\times n}$ be any matrix, and let $\lambda_1, \ldots, \lambda_n$ be its eigenvalues. Recall the notation

$$\mathrm{gap}_i(A) := \min_{j\neq i} |\lambda_i - \lambda_j|.$$

We begin by comparing these quantities to the condition number of the corresponding triple.

Lemma A.1.4. *Let A be a matrix with distinct eigenvalues and spectral decomposition $A = \sum_{i=1}^n \lambda_i v_i w_i^*$. Then, for every $i = 1, \ldots, n$ it holds that*

$$\frac{\mu(A, \lambda_i, v_i)}{\|A\|_F} \ge \frac{1}{\mathrm{gap}_i(A)}.$$

Proof. First we show that $\Lambda(A_{\lambda_i,v_i}) = \Lambda(A-\lambda_i)\setminus\{0\}$. To see this, take any $j \neq i$ and note that

$$w_j^* P_{v_i^\perp} \circ (A-\lambda_i)|_{v_i^\perp} = (\lambda_j - \lambda_i) w_i^*,$$

and hence $\lambda_j - \lambda_i$ is an eigenvalue of A_{λ_i,v_i}.

Now, using that the norm of a matrix is bigger than its spectral radius we get

$$\begin{aligned}\|A_{\lambda_i,v_i}^{-1}\| &\geq \sup_{\lambda\in\Lambda(A_{\lambda_i,v_i})} \frac{1}{|\lambda|} \\ &= \frac{1}{\mathrm{gap}_i(A)} \qquad\qquad \text{because } \Lambda(A_{\lambda_i,v_i}) = \Lambda(A-\lambda_i)\setminus\{0\}.\end{aligned}$$

The claim then follows from the definition of $\mu(A,\lambda_i,v_i)$. □

Using Theorem A.1.2 we get the following.

Proposition A.1.5. *Let $A \in \mathbb{C}^{n\times n}$ be an arbitrary matrix and let G_n be a normalized complex Ginibre matrix. Then for any $t,\gamma > 0$*

$$\mathbb{P}[\mathrm{gap}(A+\gamma G_n) < t\gamma] \leq n^3 t^2.$$

Thus, $\mathrm{gap}(A+\gamma G_n) = O(\gamma/n^{3/2})$ with probability bounded away from zero.

Proof. Using Lemma A.1.4 we get

$$\frac{1}{\mathrm{gap}(A+\gamma G_n)^2} = \max_i \frac{1}{\mathrm{gap}_i(A+\gamma G_n)^2} \leq \max_i \frac{\mu(A+\gamma G_n,\lambda_i,v_i)^2}{\|A+\gamma G_n\|_F^2} \leq n\frac{\mu_{F,\mathsf{av}}(A+\gamma G_n)^2}{\|A+\gamma G_n\|_F^2}.$$

Combining this with Theorem A.1.2 we obtain

$$\mathbb{E}\left[\frac{1}{\mathrm{gap}(A+\gamma G_n)^2}\right] \leq \frac{n^3}{\gamma^2}.$$

The proof is then concluded using Markov's inequality. □

Remarkably, the γ dependence in the bound of Proposition A.1.5 is optimal, and stronger than what can be proven using the techniques from Chapter 2. That said, this technique heavily exploit that the random perturbation has a complex Gaussian distribution, and it does not seem possible to extend these result to other distributions.

A.2 Tail Bounds for κ_V

Here we use a result from [15] to obtain the strongest known tail bound on $\kappa_V(A+\gamma G_n)$.

Lemma A.2.1 (Eigenvector condition number)**.** *For any $A \in \mathbb{C}^{n\times n}, \gamma \in (0, \|A\|)$ and $t > 0$ satisfying*

$$t < \frac{\gamma}{\|A\| n^{3/2}},$$

we have

$$\mathbb{P}\Big[\kappa_V(A+\gamma G_n) \geq \frac{1}{t}\Big] \leq 2\left(2\sqrt{2} + \frac{\|A\|}{\gamma} + \sqrt{\frac{4\log(1/t)}{n}}\right)^2 n^3 t^2.$$

Proof. To simplify notation put $M := A + \gamma G_n$ and let $\lambda_1, \ldots, \lambda_n$ be its random eigenvalues. Then for any $s, t > 0$

$$\begin{aligned}
\mathbb{P}\left[\kappa_V(M) \geq \frac{1}{t}\right] &= \mathbb{P}\left[\kappa_V(M)^2 \geq \frac{1}{t^2}\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^n \kappa(\lambda_i)^2 \geq \frac{1}{nt^2}\right] && \text{by (1.8)} \\
&\leq \mathbb{P}[\|G_n\| \geq s] + \mathbb{P}\left[\|G_n\| \leq s \text{ and } \sum_{i=1}^n \kappa(\lambda_i)^2 \geq \frac{1}{nt^2}\right].
\end{aligned}$$

Moreover, from (3.7) we have $\mathbb{P}[\|G_n\| \geq s] \leq 2\exp\big(-n(s-2\sqrt{2})^2\big)$. On the other hand

$$\begin{aligned}
\mathbb{P}\left[\|G_n\| \leq s \text{ and } \sum_{i=1}^n \kappa(\lambda_i)^2 \geq \frac{1}{nt^2}\right] &\leq \mathbb{P}\left[\sum_{\lambda_i \in D(0,\|A\|+s\gamma)} \kappa(\lambda_i)^2 \geq \frac{1}{nt^2}\right] \\
&\leq \left(\frac{\|A\|}{\gamma} + s\right)^2 n^3 t^2,
\end{aligned}$$

where the last inequality follows from Lemma 1.2.9 and Markov's inequality. Putting everything together we get that

$$\mathbb{P}\left[\kappa_V(M) \geq \frac{1}{t}\right] \leq 2\exp\big(-n(s-2\sqrt{2})^2\big) + \left(\frac{\|A\|}{\gamma} + s\right)^2 n^3 t^2.$$

Now, to simplify notation define $\mathrm{P} := \frac{\|A\|}{\gamma} n^{3/2} t$. Then choose s to be the solution of the equation $2\exp\big(-n(s-2\sqrt{2})^2\big) = \mathrm{P}^2$, and plug it into the above inequality to obtain

$$\begin{aligned}
\mathbb{P}\left[\kappa_V(M) \geq \frac{1}{t}\right] &\leq \mathrm{P}^2 + \left(\frac{\|A\|}{\gamma} + 2\sqrt{2} + \frac{1}{\sqrt{n}}\log(2/\mathrm{P}^2)\right)^2 n^3 t^2 \\
&\leq 2\left(\frac{\|A\|}{\gamma} + 2\sqrt{2} + \frac{1}{\sqrt{n}}\log(2/\mathrm{P}^2)\right)^2 n^3 t^2 \\
&\leq 2\left(\frac{\|A\|}{\gamma} + 2\sqrt{2} + \frac{2}{\sqrt{n}}\log(1/t)\right)^2 n^3 t^2 && 2\mathrm{P}^{-2} \leq t^{-2}.
\end{aligned}$$

□

Appendix B

Appendix for Chapter 3

B.1 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 B.1.1. *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 $\max\{n/5, 1\}$ 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 $1 \leq i < s_2$ such that $r_1 + \cdots + r_i \geq n/5$ and $r_{i+1} + \cdots + r_{s_1} \geq n/5$—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 $r_i \geq 3/5$. 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 $n \leq 5$, 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. The main observation is that, given any matrix X, we can determine how many eigenvalues are on either side of any horizontal or vertical line by approximating the sign function of a shift of the matrix. To be precise, in exact arithmetic $\operatorname{Tr}\operatorname{sgn}(X - h) = n_+ - n_-$, where $n_\pm$ are the eigenvalue counts for X on either side of the line $\operatorname{Re} z = h$. We will now show that under the shattered pseudospectrum assumption, one can exactly compute $n_+ - n_-$ using the advertised precision.

Running SGN to a final accuracy of β,

$$|\operatorname{Tr}\mathsf{SGN}(M) + e_4 - \operatorname{Tr}\operatorname{sgn}(M)|$$

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 $\beta \le 0.05/n$.
Algorithm: $(\widetilde{P}_+, \widetilde{P}_-, \mathsf{g}_+, \mathsf{g}_-) = \mathsf{SPLIT}(A, \mathsf{g}, \epsilon, \beta)$

1. $h \leftarrow \operatorname{Re} z_0 + \omega s_1/2$
2. $M \leftarrow A - h + E_2$
3. $\alpha_0 \leftarrow 1 - \frac{\epsilon}{2\operatorname{diam}(\mathsf{g})^2}$
4. $\phi \leftarrow \operatorname{round}\left(\operatorname{Tr}\mathsf{SGN}(M, \epsilon/4, \alpha_0, \beta) + e_4\right)$
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 $\operatorname{Tr}\mathsf{SGN}(A-h, \epsilon/4, \alpha_0, \beta) \le \frac{3n}{5}$, at which point output $\mathsf{g}_\pm$, the subgrids on either side of the shift h, and set $\widetilde{P}_\pm \leftarrow \frac{1}{2}\left(\mathsf{SGN}(h-A, \epsilon/4, \alpha_0, \beta)\right)$.
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)$\operatorname{rank}(P_\pm) = n_\pm \ge n/5$, (iii) $\|P_\pm - \widetilde{P}_\pm\| \le \beta$, and (iv) $P_\pm$ are the spectral projectors onto the interiors of $\mathsf{g}_\pm$.

$$
\begin{aligned}
&\le |\operatorname{Tr}\mathsf{SGN}(M) - \operatorname{Tr}\operatorname{sgn}(M)| + |e_4| \\
&\le n\big(\|\mathsf{SGN}(M) - \operatorname{sgn}(M)\| + \|\mathsf{SGN}(M)\|\mathbf{u}\big) \qquad \text{Using (3.2) to bound } |e_4| \\
&\le n\big(\beta + (\beta + \|\operatorname{sgn}(M)\|)\mathbf{u}\big).
\end{aligned}
$$

It remains to control $\|\operatorname{sgn}(M)\|$ and quantify the distance between $\operatorname{sgn}(M) = \operatorname{sgn}(A - h + E_2)$ and $\operatorname{sgn}(A-h)$. We first do the latter. Since we need only to modify the diagonal entries of A when creating M, the incurred *diagonal* error matrix E_2 has norm at most $\mathbf{u}\max_i |A_{i,i} - h|$. Using $|A_{i,i}| \le \|A\| \le 4$ and $|h| \le 4$, the fact that $\mathbf{u} \le \epsilon/100n \le \epsilon/16$ ensures that the $\epsilon/2$-pseudospectrum of M will still be shattered with respect to g. We can then form $\operatorname{sgn}(A-h)$ and $\operatorname{sgn}(M)$ by integrating around the boundary of the portions of g on either side of the line $\operatorname{Re} z = h$, then using the resolvent identity as in Section 3.4, and the fact that $\Lambda_\epsilon(A)$ and

$\Lambda_{\epsilon/2}(M)$ are shattered we get

$$\|\mathrm{sgn}(A) - \mathrm{sgn}(M)\| \leq \frac{\|E_2\|}{2\pi} \cdot \frac{1}{\epsilon} \cdot \frac{2}{\epsilon}\omega(2s_1 + 4s_2) \leq \frac{128\mathbf{u}}{\epsilon^2}$$

where in the last inequality we have used that g has side lengths of at most 8 and $\|E_2\| \leq 8\mathbf{u}$.

Now, using the contour integral again and the shattered pseudospectrum assumption

$$\|\mathrm{sgn}(A - h)\| \leq \frac{1}{2\pi}\frac{1}{\epsilon}\omega(2s_1 + 4s_2) \leq 8/\epsilon.$$

Combining the above bounds we get a a total additive error of $n(\beta + \beta\mathbf{u} + 8\mathbf{u}/\epsilon) + \frac{128\mathbf{u}}{\epsilon^2}$ in computing the trace of the sign function. If $\beta \leq 0.1/n$ and $\mathbf{u} \leq \min\{\epsilon/100n, \frac{\epsilon^2}{512}$, this error will strictly be 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.

Now that we have shown that it is possible to compute $n_+ - n_-$ exactly, recall that from the above discussion, the $\epsilon/2$-pseudospectrum of M will still be shattered with respect to the translation of the original grid g. 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

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

□

B.2 Analysis of DEFLATE

The algorithm DEFLATE, defined in Section 3.5, can be viewed as a small variation of the randomized rank revealing algorithm introduced in [53] and revisited subsequently in [9]. 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

$$R = \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix}, \tag{B.1}$$

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.

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.

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 B.2.12 and Proposition B.2.18 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 to the span of the top r eigenvectors of A. Our implementation of DEFLATE is as follows.

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 $1 - \frac{(20n)^3\sqrt{\beta}}{\eta^2\sigma_r(A)}$.

Throughout this section we use rurv$(\cdot)$ and deflate$(\cdot,\cdot)$ to denote the exact arithmetic versions of RURV and DEFLATE respectively. In Subsection B.2.1 we present a random matrix result that will be needed in the analysis of DEFLATE. In Subsection B.2.3 we state the properties of RURV that will be needed. Finally in Subsections B.2.4 and B.2.5 we prove the main guarantees of deflate and DEFLATE, respectively, that are used throughout this paper.

B.2.1 Smallest Singular Value of the Corner of a Haar Unitary

We recall the defining property of the Haar measure on the unitary group:

Definition B.2.1. 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 $\sigma_r(X)$. We begin by showing a fact that allows us to reduce our analysis to the case when $r \leq n/2$.

Observation B.2.2. 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_{11} are equal to 1, while the remaining $n - r$ are equal to those of V_{22}.

Proof. Decompose V as follows

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

Since V is unitary $VV^* = I_n$, and looking at the upper-left corner of this equation we get $V_{11}V_{11}^* + V_{12}V_{12}^* = I_r$. Then, since $V_{11}V_{11}^* = I_r - V_{12}V_{12}^*$, we have $\Lambda(V_{11}V_{11}^*) = 1 - \Lambda(V_{12}V_{12}^*)$.

Now, looking at the lower-right corner of the equation $V^*V = I_n$ we get $V_{12}^*V_{12} + V_{22}^*V_{22} = I_{n-r}$ and hence $\Lambda(V_{22}^*V_{22}) = 1 - \Lambda(V_{12}^*V_{12})$.

Now recall that for any two matrices X and Y, the symmetric difference of the sets $\Lambda(XY)$ and $\Lambda(YX)$ is $\{0\}$, with multiplicity equal to the difference between the dimensions. Hence $\Lambda(V_{12}V_{12}^*) = \Lambda(V_{12}^*V_{12}) \cup \{0\}$ where the multiplicity of 0 is $r-(n-r) = 2r-n$. Combining this with $\Lambda(V_{11}V_{11}^*) = 1 - \Lambda(V_{12}V_{12}^*)$ and $\Lambda(V_{22}^*V_{22}) = 1 - \Lambda(V_{12}^*V_{12})$ we get the desired result. □

Proposition B.2.3 ($\sigma_{\min}$ 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 $\theta \in (0, 1]$*

$$\mathbb{P}\left[\frac{1}{\sigma_r(X)} \leq \frac{1}{\theta}\right] = (1-\theta^2)^{r(n-r)}. \tag{B.2}$$

In particular, for every $\theta \in (0, 1]$ we have

$$\mathbb{P}\left[\frac{1}{\sigma_r(X)} \leq \frac{\sqrt{r(n-r)}}{\theta}\right] \geq 1 - \theta^2. \tag{B.3}$$

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 [60], from which one can extract and simplify the density of $\sigma_r(X)$. We will begin by introducing the relevant pieces of [60], 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

$${}_2F_1^\beta(0, b; c, (x_1, \ldots, x_m)) = 1, \tag{B.4}$$

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 [67] or Definition 2.2 in [60]).

The generic results in [60] concern the *β-Jacobi* random matrices, which we have no cause here to define in full. Of particular use to us will be [60, Theorem 3.1], which expresses the density of the smallest singular value of such a matrix in terms of the generalized hypergeometric function:

Theorem B.2.4 ([60])**.** *The density of the probability distribution of the smallest eigenvalue λ, of the β-Jacobi ensembles of parameters a, b and size m, which we denote by $f_{\lambda_{\min}}(\lambda)$, is given by*

$$C_{\beta,a,b,m}\lambda^{\frac{\beta}{2}(a+1)-1}(1-\lambda)^{\frac{\beta}{2}m(b+m)-1}$$

$$\cdot\, {}_2F_1^{2/\beta}\left(1-\frac{\beta(a+1)}{2}, \frac{\beta(b+m-1)}{2}; \frac{\beta(b+2m-1)}{2}+1; (1-\lambda)^{m-1}\right), \tag{B.5}$$

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 $\sigma_r^2(X)$. The connection between singular values of corners of Haar unitary matrices and β-Jacobi ensembles is the content of [65, Theorem 1.5], which we rephrase below to match our context.

Theorem B.2.5 ([65]). *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 B.2.4 gives a formula for the density of $\sigma_r^2(X)$.

Corollary B.2.6 (Density of $\sigma_r^2(X)$). *Let V be an $n \times n$ Haar unitary and X be its upper-left $r \times r$ corner with $r < n$, then $\sigma_r^2(X)$ has the following density*

$$f_{\sigma_r^2}(x) := \begin{cases} r(n-r)\,(1-x)^{r(n-r)-1} & \text{if} \quad 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{B.6}$$

Proof. If $r > n/2$, since we care only about the smallest singular value of X, we can use Observation B.2.2 to analyse the $(n-r)\times(n-r)$ lower right corner of V instead. Hence, we can assume without loss of generality that $r \leq n/2$. Now, substitute $\beta = 2, a = 0, b = n-2r, m = r$ in Theorem B.2.4 and observe that in this case

$$f_{\lambda_{\min}}(x) = C(1-x)^{r(n-r)-1}\, {}_2F_1^1(0, n-r-1; n; (1-x)^{r-1}) = C(1-x)^{r(n-r)-1} \tag{B.7}$$

where the last equality follows from (B.4). Using the relation between the distribution of $\sigma_r^2(X)$ and the distribution of the minimum eigenvalue of the respective β-Jacobi ensemble described in Theorem B.2.5 we have $f_{\sigma_r^2}(x) = f_{\lambda_{\min}}(x)$. By integrating on $[0,1]$ the right side of (B.7) we find $C = r(n-r)$. □

Proof of Proposition B.2.3. From (B.6) we have that

$$\mathbb{P}\left[\sigma_r^2(X) \leq \theta\right] = r(n-r)\int_0^\theta (1-x)^{r(n-r)-1}dx = 1-(1-\theta)^{r(n-r)},$$

from where (B.2) follows. To prove (B.3) 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 $t \in [0,1]$. □

B.2.2 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 [64, Section 4.6] and [108] for a detailed discussion. In this subsection we carefully analyze this process in finite arithmetic.

The following fact (see [108, Section 5]) is the starting point of our discussion.

Lemma B.2.7 (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 3.2.1 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 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 [146, Theorem 1.6]:

Lemma B.2.8 (Condition number for the QR decomposition [146])**.** *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

$$\mathbf{u}\mu_{\mathsf{QR}}(n) \leq 1. \tag{B.8}$$

Proposition B.2.9 (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*

[1] G_n is almost surely invertible and under this event U and R are uniquely determined by these conditions.

[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

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 (B.8), that

$$\|\widetilde{\widetilde{G}}_n - G_n\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}\left(\|G_n\| + \sqrt{n}c_{\mathsf{N}}\mathbf{u}\right) + \sqrt{n}c_{\mathsf{N}}\mathbf{u} \leq \mu_{\mathsf{QR}}(n)\mathbf{u}\|G_n\| + 2\sqrt{n}c_{\mathsf{N}}\mathbf{u}. \tag{B.9}$$

Let $(U, R') := \mathrm{QR}(G_n)$. From Lemma B.2.7 we know that U is Haar distributed on the unitary group, so using (B.9) and Lemma B.2.8, and the fact that $\|M\| \leq \|M\|_F \leq \sqrt{n}\|M\|$ for any $n \times n$ matrix M, we know that

$$\begin{aligned}
\|U - V\| - \mu_{\mathsf{QR}}(n)\mathbf{u} &\leq \|U - V\| - \|\widetilde{V} - V\| \\
&\leq \|U - \widetilde{V}\| \\
&\leq 4\sqrt{n}c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u}\|G_n\|\|G_n^{-1}\| + 10nc_{\mathsf{N}}\mathbf{u}\|G_n^{-1}\|.
\end{aligned} \tag{B.10}$$

Now, from $\|G_n^{-1}\| = 1/\sigma_n(G_n)$ and from (1.17) 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 [15] 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 (B.10) 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. □

B.2.3 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 B.2.11 below, which is the complex analog to the upper bound given in equation (4) of [9, Theorem 5.1]. Actually, Corollary B.2.11 is a direct consequence of Lemma 4.1 in the aforementioned paper and Proposition B.2.3 proved above. We elaborate below.

Lemma B.2.10 ([9]). *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 $X = Q^*V^*$,*

$$\|R_{22}\| \leq \frac{\sigma_{r+1}(A)}{\sigma_r(X_{11})},$$

where X_{11} is the upper left $r\times r$ block of X.

This lemma reduces the problem to obtaining a lower bound on $\sigma_r(X_{11})$. 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 B.2.10 and Proposition B.2.3 gives the following result.

Corollary B.2.11. *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 $\theta > 0$*

$$\mathbb{P}\left[\|R_{22}\| \leq \frac{\sqrt{r(n-r)}}{\theta}\sigma_{r+1}(A)\right] \geq 1-\theta^2.$$

B.2.4 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 $\mathrm{deflate}(A,r)$ 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 $(U,R) = \mathrm{rurv}(A)$ 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 that $\mathrm{rank}(AV^*) = r$ and that the first r columns of AV^* are linearly independent, so since UR is the QR decomposition of AV^*, almost surely, $R_{22} = 0$ and $R_{11} \in \mathbb{C}^{r\times r}$, where R_{11} and R_{22} are as in (B.1). 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

$$AV^* = UR = \begin{pmatrix} U_{11}R_{11} & U_{11}R_{12}+U_{12}R_{22} \\ U_{21}R_{11} & U_{21}R_{12}+U_{22}R_{22} \end{pmatrix}. \tag{B.11}$$

On the other hand, almost surely the first r columns of AV^* span the range of A. Using the right side of equation (B.11) 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 B.2.12 (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*

$$\|S - TW^*\| \leq \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\beta}{\theta}}. \tag{B.12}$$

Remark B.2.13 (The projector case)**.** In the case in which the matrix A of Proposition B.2.12 is a (not necessarily orthogonal) projector, $T^*AT = I_r$, and the σ_r term in the denominator of (B.12) 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 [89, Theorem 3.3.16] .

Lemma B.2.14 (Stability of singular values)**.** *Let $X, E \in \mathbb{C}^{n\times n}$. Then, for any $k = 1, \ldots, n$ we have*

$$|\sigma_k(X + E) - \sigma_k(X)| \leq \|E\|.$$

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 B.2.15. *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*

$$\|(SS^* - I_n)A\| \leq \|R_{22}\| + \beta, \tag{B.13}$$

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

$$(SS^* - I_n)\widetilde{A} = (SS^* - I)\begin{pmatrix} S & U' \end{pmatrix} RV = \begin{pmatrix} 0 & -U' \end{pmatrix} RV = \begin{pmatrix} 0 & -U'R_{2,2} \end{pmatrix} V.$$

[3] For example, diagonalizable matrices satisfy this criterion.

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 (B.13) can be applied to quantify the distance between the ranges of deflate$(\widetilde{A}, r)$ and deflate(A, r) in terms of $\|R_{22}\|$, as the following result shows.

Lemma B.2.16 (Bound in terms of $\|R_{22}\|$)**.** *Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that* $\operatorname{rank}(A) = \operatorname{rank}(A^2) = r$ *and $\|A - \widetilde{A}\| \leq \beta$. Denote by $(U, R) :=$* rurv$(\widetilde{A})$*, $S :=$* deflate$(\widetilde{A}, r)$ *and $T :=$* deflate(A, r)*. Then, almost surely there exists a unitary $W \in \mathbb{C}^{r\times r}$ such that*

$$\|S - TW^*\| \leq 2\sqrt{\frac{\|R_{22}\| + \beta}{\sigma_r(T^*AT)}}, \tag{B.14}$$

where R_{22} is the lower right $(n - r) \times (n - r)$ block of R.

Proof. From Lemma B.2.15 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

$$\|T^*SS^*T - I_r\| = \sup_{w\in\mathbb{C}^r, \|w\|=1} \|T^*(SS^* - I_r)Tw\| = \sup_{w\in\operatorname{range}(A), \|w\|=1} \|T^*(SS^* - I_r)w\|. \tag{B.15}$$

Now, since $\operatorname{rank}(A) = \operatorname{rank}(A^2)$, if $w \in \operatorname{range}(A)$ then $w = Av$ for some $v \in \operatorname{range}(A)$. So by the Courant-Fischer formula

$$\frac{\|w\|}{\|v\|} = \frac{\|Av\|}{\|v\|} \geq \inf_{u\in\operatorname{range}(A)} \frac{\|Au\|}{\|u\|} = \sigma_r(T^*AT).$$

We can then revisit (B.15) and get

$$\sup_{w\in\operatorname{range}(A), \|w\|=1} \|T^*(SS^* - I_r)w\| = \sup_{v\in\operatorname{range}(A), \|v\|\leq 1} \frac{\|T^*(SS^* - I_r)Av\|}{\sigma_r(T^*AT)} \leq \frac{\|T^*(SS^* - I_r)AT\|}{\sigma_r(T^*AT)}. \tag{B.16}$$

On the other hand $\|T^*(SS^* - I_r)AT\| \leq \|(SS^* - I_r)A\| \leq \|R_{22}\| + \beta$, so combining this fact with (B.15) and (B.16) 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 B.2.12 comes from the non-deterministic bound on $\|R_{22}\|$.

Proof of Proposition B.2.12. From Lemma B.2.14 we have $\sigma_{r+1}(\widetilde{A}) \leq \beta$. Now combine Lemma B.2.16 with Corollary B.2.11. □

B.2.5 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

$$4\|A\| \cdot \max\{c_{\mathsf{N}}\mu_{\mathsf{MM}}(n)\mathbf{u}, c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u}\} \leq \beta \leq \frac{1}{4} \leq \|A\| \quad \text{and} \quad 1 \leq \min\{\mu_{\mathsf{MM}}(n), \mu_{\mathsf{QR}}(n), c_{\mathsf{N}}\}. \tag{B.17}$$

We begin by analyzing the subroutine RURV in finite arithmetic. This was done in [53, 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 B.2.17 (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 (B.17) 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. $\|U - \widetilde{U}\| \leq \mu_{\mathsf{QR}}(n)\boldsymbol{u}$.
2. $\widetilde{V}$ *is Haar distributed in the unitary group.*
3. *For every* $1 > \alpha > 0$ *and* $t > 2\sqrt{2} + 1$*, the event:*

$$\left\{\|\widetilde{V} - V\| < \frac{8tn^{\frac{3}{2}}}{\alpha}c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha}\mathbf{u} \quad \text{and} \right.$$
$$\left. \|A - \widetilde{A}\| < \|A\|\left(\frac{9tn^{\frac{3}{2}}}{\alpha}c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha}c_{\mathsf{N}}\mathbf{u}\right)\right\} \tag{B.18}$$

occurs with probability at least $1 - 2e\alpha^2 - 2e^{-t^2n}$.

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 B.2.9, we know that there is a Haar unitary $\widetilde{V}$ and a random matrix E_0, such that $V = \widetilde{V} + E_0$ and

$$\mathbb{P}\left[\|E_0\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}} \mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u}\right] \geq 1 - 2e\alpha^2 - 2e^{-t^2 n}. \tag{B.19}$$

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 (B.17). 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 (B.17) again, we get

$$\|E_1\| \leq \frac{5}{4}\|A\|\mu_{\mathsf{MM}}(n)\mathbf{u} \quad \text{and} \quad B \leq \frac{5}{4}\|A\|(1 + \mu_{\mathsf{MM}}(n)\mathbf{u}) < 2\|A\|. \tag{B.20}$$

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 (B.20) we obtain

$$\|E_3\| \leq \|B\|\mu_{\mathsf{QR}}(n)\mathbf{u} < 2\|A\|\mu_{\mathsf{QR}}(n)\mathbf{u}. \tag{B.21}$$

4. Finally, define $\widetilde{A} := \widetilde{B}\widetilde{V}$. Note that $\widetilde{A} = \widetilde{U}R\widetilde{V}$ and

$$\widetilde{A} = \widetilde{B}\widetilde{V} = (B - E_3)\widetilde{V} = (AV^* + E_1 - E_3)\widetilde{V} = (A(\widetilde{V} + E_0)^* + E_1 - E_3)\widetilde{V},$$

   and the latter is equal to $A + (AE_0^* + E_1 - E_3)\widetilde{V}$, which translates into

$$\|A - \widetilde{A}\| \leq \|A\|\|E_0\| + \|E_1\| + \|E_3\|.$$

   Hence, on the event described in the left side of (B.19), 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 B.2.12.

Proposition B.2.18 (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 (B.17) holds, then, for every $t > 2\sqrt{2} + 1$ there exist a unitary $W \in \mathbb{C}^{r\times r}$ such that*

$$\|S - TW^*\| \leq \mu_{\mathsf{QR}}(n)\boldsymbol{u} + 12\sqrt{\frac{tn^2\sqrt{r(n-r)}}{\sigma_r(T^*AT)}}\cdot\sqrt{\frac{\beta}{\theta^2}}, \tag{B.22}$$

with probability at least $1 - 7\theta^2 - 2e^{-t^2 n}$.

Proof. Let $(U, R) = \mathsf{RURV}(\widetilde{A})$. From Lemma B.2.17 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 (B.18), for every $1 > \alpha > 0$ and $t > 2\sqrt{2} + 1$ we have

$$\begin{aligned}\left\|A - \widetilde{\widetilde{A}}\right\| \leq& \left\|\widetilde{\widetilde{A}} - \widetilde{A}\right\| + \|\widetilde{A} - A\| \\ \leq& (\|A\| + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha}\mu_{\mathsf{QR}}(n)c_{\mathsf{N}}\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha}c_{\mathsf{N}}\mathbf{u}\right) + \beta,\end{aligned} \tag{B.23}$$

with probability $1 - 2e\alpha^2 - 2e^{-t^2 n}$.

Now, from (B.17) 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 (B.23) by β:

$$\begin{aligned}&(\|A\| + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha}c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha}c_{\mathsf{N}}\mathbf{u}\right) + \beta \\ \leq& (1 + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha}\beta + 2\beta + \frac{10n^2}{\alpha}\beta\right) + \beta \\ \leq& \frac{(12t + 16)}{\alpha}n^2\beta,\end{aligned} \tag{B.24}$$

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 $\widetilde{S} = \operatorname{deflate}(\widetilde{\widetilde{A}}, r)$ is the matrix formed by the first r columns of $\widetilde{U}$, and that by Proposition B.2.12 we know that for every $\theta > 0$, with probability $1 - \theta^2$ there exists a

unitary W such that

$$\|\widetilde{S} - TW^*\| \leq \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\left\|A - \widetilde{\widetilde{A}}\right\|}{\theta}}. \tag{B.25}$$

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

$$\|S - TW^*\| \leq \|S - \widetilde{S}\| + \|\widetilde{S} - TW^*\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u} + \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\left\|A - \widetilde{\widetilde{A}}\right\|}{\theta}}. \tag{B.26}$$

Now, taking $\alpha = \theta$, we note that both events in (B.23) and (B.25) 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 (B.23), (B.24) and (B.26). □

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 B.2.13 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 B.2.19. 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).

Appendix C

Appendix for Chapter 4

C.1 Deferred Proofs from Section 4.2.4

Proof of Lemma 4.2.8. For the purpose of the analysis, let us define $\widetilde{H_0} := H - s$ and for each $i = 1, ..., n-1$, denote by $\widetilde{H}_i$ the matrix $\widetilde{R}$ as it stands at the end of line 2(c) on the the ith step of the loop. Additionally, write G_i for the unitary matrix which applies $\mathsf{giv}(X_{1:2,i})$ to the span of e_i and e_{i+1} and is the identity elsewhere. We will show that the unitary $\widetilde{Q} := \widetilde{Q}_{n-1}$ satisfies the guarantees of IQR. We then have

$$\widetilde{H}_i = G_i^* \widetilde{H}_{i-1} + E_{2,i},$$

where $E_{2,i}$ is the structured error matrix which in rows $(i : i+1)$ is equal to

$$\left(\begin{array}{c|c} E_{2,i,c} & E_{2,i,b} \\ 0 & \end{array}\right)$$

and is zero otherwise. From the discussion at the beginning of this appendix, we know that each entry of $E_{2,i,b}$ has size at most $8\|\widetilde{H}_{i-1}\|\mathbf{u}$ and similarly that $|E_{2,i,c}| \leq 2\|X_{1:2,i}\|\mathbf{u} \leq 8\|\widetilde{H}_{i-1}\|\mathbf{u}$. Thus $\|E_{2,i}\| \leq 8\sqrt{n}\|\widetilde{H}_{i-1}\|\mathbf{u}$, and inductively we have

$$\begin{aligned} \|\widetilde{H}_i\| &\leq \|\widetilde{H}_{i-1}\| + \|E_{2,i}\| \\ &\leq \|\widetilde{H}_{i-1}\| \left(1 + 8\sqrt{n}\mathbf{u}\right) \\ &\leq \|\widetilde{H}_0\| \left(1 + 8\sqrt{n}\mathbf{u}\right)^n \\ &\leq \|\widetilde{H}_0\| \exp\left(8n^{3/2}\mathbf{u}\right) \\ &\leq 2\|H - s\| \qquad i = 1, ..., n-1. \end{aligned}$$

Since $\widetilde{Q}$ and every G_i is unitary, this gives

$$\|H - s - \widetilde{Q}\widetilde{R}\| = \|\widetilde{Q}^*\widetilde{H_0} - \widetilde{R}\| \leq \sum_{i \in [n-1]} \|E_{2,i}\| \leq 16n^{3/2}\mathbf{u} \cdot \|H - s\|.$$

A similar inductive argument applied to line 4 gives that $\|E_{4,i}\| \leq 16\sqrt{n}\mathbf{u} \cdot \|H - s\|$ for every $i \in [n-1]$, and thus that the $\widetilde{H}$ output by $\mathsf{IQR}(H, s)$ satisfies

$$\begin{aligned}\widetilde{H} - s &= \widetilde{R}\widetilde{Q} + E_{4,n-1}(G_1 \cdots G_{n-2}) + \cdots + E_{4,2}G_1 + E_{4,1} \\ &= \widetilde{Q}^*(H - s)\widetilde{Q} + E_{4,n-1}(G_1 \cdots G_{n-2}) + \cdots + E_{4,2}G_1 + E_{4,1} \\ &\quad + (G_{n-2}^* \cdots G_1^*)E_{2,1}\widetilde{Q} + (G_{n-3}^* \cdots G_1^*)E_{2,2}\widetilde{Q} + \cdots + G_1^* E_{2,n-1}\widetilde{Q},\end{aligned}$$

meaning

$$\|\widetilde{H} - \widetilde{Q}^* H \widetilde{Q}\| \leq 32n^{3/2}\mathbf{u} \cdot \|H - s\|$$

and

$$\|\widetilde{H}\| \leq \|H\| + 32n^{3/2}\|H - s\|\mathbf{u},$$

as desired.

In terms of arithmetic operations, it costs n to compute $\widetilde{R}$ from H in line 1. In line 2(b), computing $\|X_{1:2,i}\|$ costs 4, computing $\mathsf{giv}(X_{1:2,i})$ given this norm costs another 2, zeroing out $\widetilde{R}_{i+1,i}$ costs 1, replacing $\widetilde{R}_{i,i}$ with $\|X_{1:2,i}\|$ costs one, and applying the rotation to $\widetilde{R}_{i:i+1,i+1:n}$ costs $4(n-i+1)$. We do this for each of $i = 1, 2, \ldots n-1$, giving $6(n-1)+2(n-1)+2n(n-1)$. In line 4, assuming we have stored each Givens rotation, applying them again requires $2n(n+1)-4$. Finally, in line 5 we pay another n to re-apply the shift. Thus in total we have

$$n + 6(n-1) + 2(n-1) + 2n(n-1) + 2n(n+1) - 4 + n = 4n^2 + 12n - 12 \leq 7n^2 \qquad n \geq 2.$$

□

Proof of Lemma 4.2.9. Let $\widetilde{H}_1 = H$, and for each $\ell \in [m-1]$, let $[\widetilde{H}_{\ell+1}, \widetilde{R}_\ell] = \mathsf{IQR}(\widetilde{H}_\ell, r_\ell)$ and $\widetilde{Q}_\ell$ be as guaranteed by Definition 4.1.2. We have

$$\|\widetilde{H}_2 - \widetilde{Q}_1 * \widetilde{H_1}\widetilde{Q}_1\| \leq \|\widetilde{H}_1 - s_1\|\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq (1+C)\|H\|\nu_{\mathsf{IQR}}(n)\mathbf{u},$$

and inductively, assuming that

$$\|\widetilde{H}_\ell - \widetilde{Q}_{\ell-1}^* \widetilde{H}_{\ell-1}\widetilde{Q}_{\ell-1}\| \leq (1+C)\|H\|(\nu_{\mathsf{IQR}}(n)\mathbf{u} + \cdots + (\nu_{\mathsf{IQR}}(n)\mathbf{u})^\ell),$$

we have

$$\begin{aligned}\|\widetilde{H}_{\ell+1} - \widetilde{Q}_\ell^* \widetilde{H}_\ell \widetilde{Q}_\ell\| &\leq \|\widetilde{H}_\ell - s_\ell\|\nu_{\mathsf{IQR}}(n)\mathbf{u} \\ &\leq \|H\|(1 + (1+C)(\nu_{\mathsf{IQR}}(n)\mathbf{u} + \cdots + (\nu_{\mathsf{IQR}}(n)\mathbf{u})^\ell) + C)\nu_{\mathsf{IQR}}(n)\mathbf{u} \\ &\leq (1+C)\|H\|(\nu_{\mathsf{IQR}}(n)\mathbf{u} + \cdots + (\nu_{\mathsf{IQR}}(n)\mathbf{u})^{\ell+1}).\end{aligned}$$

This gives the first asserted bound, since

$$\|\widetilde{H} - \widetilde{Q}^* \widetilde{H} \widetilde{Q}\| \leq \sum_{\ell \in [m-1]} \|\widetilde{H}_{\ell+1} - \widetilde{Q}_\ell^* \widetilde{H}_\ell \widetilde{Q}_\ell\| \leq (1+C)\|H\| \frac{m\nu_{\mathsf{IQR}}(n)\mathbf{u}}{1 - \nu_{\mathsf{IQR}}(n)\mathbf{u}}$$

and $\frac{1}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}} \leq 4/3 \leq 1.4$.

For the second assertion, we will mirror the proof of Lemma 4.2.5, using backward stability guarantees on a single IQR step from Definition 4.1.2. In particular, in view of the definition and the above bound, we can write

$$\widetilde{H}_\ell - s_\ell = \widetilde{Q}_\ell \widetilde{R}_\ell + E_\ell \qquad \|E_\ell\| \leq (1+C)\|H\| \frac{\nu_{\mathsf{IQR}}(n)\mathbf{u}}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}$$

$$\widetilde{H}_1 \widetilde{Q}_\ell \cdots \widetilde{Q}_1 = \widetilde{Q}_\ell \cdots \widetilde{Q}_1 \widetilde{H}_{\ell+1} + \|\Delta_{\ell+1}\| \qquad \Delta_{\ell+1} \leq (1+C)\|H\| \frac{\nu_{\mathsf{IQR}}(n)\mathbf{u}}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}$$

so that

$$\begin{aligned}
p(H) &= p(\widetilde{H}_1) \\
&= (\widetilde{H}_1 - s_m)\cdots(\widetilde{H}_1 - s_1) \\
&= (\widetilde{H}_1 - s_m)\cdots(\widetilde{Q}_1\widetilde{R}_1 + \widetilde{Q}_1^* E_1) \\
&= (\widetilde{H}_1 - s_m)\cdots(\widetilde{H_1} - s_2)\widetilde{Q}_1(\widetilde{R}_1 + \widetilde{Q}_1^* E_1) \\
&= (\widetilde{H}_1 - s_m)\cdots\widetilde{Q}_1(\widetilde{H}_2 - s_2 + \Delta_2)(\widetilde{R}_1 + \widetilde{Q}_1^* E_1) \\
&= (\widetilde{H}_1 - s_m)\cdots(\widetilde{H}_1 - s_3)\widetilde{Q}_1\widetilde{Q_2}(\widetilde{R}_2 + \widetilde{Q}_2^* E_2 + \widetilde{Q}_2^*\Delta_2)(\widetilde{R}_1 + \widetilde{Q}_1^* E_1) \\
&= \widetilde{Q}_1\cdots\widetilde{Q}_m(\widetilde{R}_m + \widetilde{Q}_m^* E_m + \widetilde{Q}_m^*\Delta_m)\cdots(\widetilde{R}_2 + \widetilde{Q}_2^* E_2 + \widetilde{Q}_2^*\Delta_2)(\widetilde{R}_1 + \widetilde{Q}_1^* E_1).
\end{aligned}$$

Thus, using the bounds on E_ℓ and Δ_ℓ, and the fact that $\|\widetilde{R}_\ell\| = \|\widetilde{H}_\ell - s_\ell\| \leq \frac{(1+C)\|H\|}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}$,

$$\begin{aligned}
&\|p(H) - \widetilde{Q}_1\cdots\widetilde{Q}_m\widetilde{R}_m\cdots\widetilde{R}_1\| \\
=&\|\widetilde{R}_m\cdots\widetilde{R}_1 - (\widetilde{R}_m + \widetilde{Q}_m^* E_m + \widetilde{Q}_m^*\Delta_m)\cdots(\widetilde{R}_2 + \widetilde{Q}_2^* E_2 + \widetilde{Q}_2^*\Delta_2)(\widetilde{R}_1 + \widetilde{Q}_1^* E_1)\| \\
\leq& \prod_{\ell\in[m]}\left(\|\widetilde{R}_\ell\| + \tfrac{2(1+C)\|H\|}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}\right) - \prod_{\ell\in[m]}\|\widetilde{R}_\ell\| \\
\leq& \left(\tfrac{(1+C)\|H\|}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}\right)^m ((1+2\nu_{\mathsf{IQR}}(n)\mathbf{u})^m - 1) \\
\leq& 4\big(2(1+C)\|H\|\big)^m \nu_{\mathsf{IQR}}(n)\mathbf{u};
\end{aligned}$$

in the final line we are using again that $\nu_{\mathsf{IQR}}(n)\mathbf{u} \leq 1/4$ and thus that $((1+2\nu_{\mathsf{IQR}}(n)\mathbf{u})^m - 1) \leq (3/2)^m \nu_{\mathsf{IQR}}(n)\mathbf{u}/4$, whereas $(1-\nu_{\mathsf{IQR}}(n)\mathbf{u})^{-m} \leq (4/3)^m$ □

C.2 Proof of Lemma 4.3.30

Proof of Lemma 4.3.30. First, if we take $t_1 := \frac{\varphi^{1/2}\gamma}{\sqrt{2}n^{3/2}}$ and apply Proposition A.1.5 we get that

$$\mathbb{P}[\mathrm{gap}(A + \gamma G_n) \geq t_1] \geq 1 - \varphi/2.$$

Then, taking $t_2 = \frac{\gamma\varphi^{1/2}}{60\|A\|\log(1/\varphi)n^{3/2}}$ and applying Lemma A.2.1 we get

$$\begin{aligned}\mathbb{P}[\kappa_V(A+\gamma G_n) \geq 1/t_2] &\leq 2\left(2\sqrt{2} + \frac{\|A\|}{\gamma} + \frac{2}{\sqrt{n}}\log(1/t_2)^{1/2}\right)^2 n^3 t_2^2 \\ &\leq 6\left(8 + \frac{\|A\|^2}{\gamma^2} + \frac{4}{n}\log(1/t_2)\right) n^3 t_2^2 \qquad \text{AM-QM} \\ &\leq \varphi/6 + \varphi/6 + \varphi/6\end{aligned}$$

yielding

$$\mathbb{P}[\kappa_V(A+\gamma G_n) \leq 1/t_2] \geq 1 - \varphi/2.$$

Now define $\zeta = t_1/3$ and $\epsilon = t_1 t_2/3$. By the tail bounds obtained above we have the event $\{\text{gap}(A+\gamma G_n) \geq t_1 \text{ and } \kappa_V(A+\gamma G_n) \leq 1/t_2\}$ occurs with probability $1-\varphi$, and, by Lemma 1.1.9, under this event we have that $\Lambda_\epsilon(A+\gamma G_n)$ is ζ-shattered, as we wanted to show. □