Random Matrices and Provable Algorithms for Approximate Diagonalization

by

Jesse M. Banks

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nikhil Srivastava, Chair
Professor James Demmel
Professor Prasad Raghavenddra

Spring 2022

Random Matrices and Provable Algorithms for Approximate Diagonalization

Abstract

Random Matrices and Provable Algorithms for Approximate Diagonalization

by

Jesse M. Banks

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Nikhil Srivastava, Chair

In this thesis we study the computational problem of *approximately diagonalizing* an arbitrary complex, $n \times n$ matrix A in floating point arithmetic — that is, finding $\widetilde{V}$ invertible and $\widetilde{D}$ diagonal with $\|A - \widetilde{V}\widetilde{D}\widetilde{V}^{-1}\| \leq \delta\|A\|$ for some desired accuracy $\delta > 0$. Our major contributions are the following.

- We prove that a randomized variant of the Spectral Bisection algorithm [30] can approximately diagonalize any $n \times n$ matrix using $O(T_{\mathsf{MM}}(n)\log(n/\delta)\log n)$ arithmetic operations, on a floating point machine with $O(\log^4(n/\delta)\log(n))$ bits of precision, with probability $1 - O(1/n)$ — where $T_{\mathsf{MM}}(n)$ is the arithmetic cost of numerically stable $n \times n$ matrix multiplication.

- We prove that for every $k = 2, 4, 8, \ldots$ and $B \geq 1$ there is a shifting strategy for the Shifted QR algorithm, $\mathsf{Sh}_{k,B}$, which converges rapidly in *exact* arithmetic on every Hessenberg matrix H_0 with eigenvector condition number less than B (e.g. of the form $H_0 = VDV^{-1}$ with D diagonal and $\|V\|\|V^{-1}\| \leq B$). The resulting algorithm gives a sequence $H_0, H_1, H_2, \ldots$ of Hesssenberg matrices unitarily similar to H_0, and rapid convergence means: each iteration costs $O(n^2 \cdot k(\log k + B^{16k^{-1}\log k}))$ arithmetic operations, and for any $\omega > 0$, it takes $O(\log(1/\omega))$ iterations to drive the smallest subdiagonal entry below $\omega\|H_0\|$.

- We prove that the shifting strategy $\mathsf{Sh}_{k,B}$ can be implemented in floating point as a randomized algorithm, using $O(k\log\frac{nB\|H_0\|}{\omega\,\mathrm{gap}(H_0)})$ bits of precision (where $\mathrm{gap}(H)$ is the minimum eigenvalue gap, or the smallest distance between two eigenvalues of H), and succeeding with probability $1 - O(1/n)$. For any $n \times n$ matrix A, this can be used to give a randomized algorithm for computing the matrix $\widetilde{D}$ in the notion of approximate diagonalization above (e.g., the eigenvalues of a matrix $\delta\|A\|$-close to A), using $O(n^3 \cdot \log^2(n/\delta)\,\mathrm{poly}(\log\log(n/\delta)))$ arithmetic operations, on a floating point machine with $O(\log^2(n/\delta)\log\log(n/\delta))$ bits of precision, again succeeding with probability $1 - O(1/n)$.

A crucial step along the way is a new result in random matrix theory: for any $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$, if $\boldsymbol{G_n}$ is an $n \times n$ matrix with independent, standard *complex* Gaussian entries, then with probability at least $1 - O(1/n)$, $A + \delta G_n$ has eigenvector condition number at most n^2/δ and minimum eigenvalue gap at least δ/n^2. We show similar results for random *real* perturbations of real matrices as well.

For Maya, Judi, Matthew, and Molly.

Contents

Acknowledgments

Above all I thank Nikhil Srivastava for lending me his singular mathematical style and sensibility; for treating me as a peer and collaborator from the moment of my arrival at Berkeley; for his generosity with long afternoons at the whiteboard and late nights in the Simons Institute before a conference deadline; and for teaching me both the value of sticking with hard problems and the good sense, sometimes, to put them aside. I also owe a great mathematical debt to Cris Moore who has shaped my taste in problems and proofs more than anyone, and has been an invaluable mentor and guide for nearly the last decade.

Thank you to my coauthors and collaborators: my first graduate advisor Luca, Prasad (who was also kind enough to sit on my committee), Sidhanth, Afonso, Tim, Alex, Bobby, and most of all Jorge, Archit, and Satyaki, without whose without whose hard work the results in this thesis might well have died on the vine. Thank you also to my mathematical companions at Berkeley and elsewhere: Melissa, Paula, Milind, and Sarah.

My union siblings in UAW deserve special mention in this thesis (although your primary impact was to distract me from its completion). Nevertheless, you have my deepest gratitude teaching me how to organize, and for sharing with me your boundless dedication to building a public university of, by, and for the people — and with it, a better and more just world.

And a final thank you to those dear ones near and far not already mentioned; a few among you are Riley, Loey, Grace, Quinn, B, Miles, Taia, Reah, Stella, Andy, Shelby, Shelby, Hannah, Hanna, Nathan, Kaela, Megan, Sarah, and Neder.

Chapter 1

Introduction

Eigenvectors and eigenvalues of matrices are central objects of study throughout pure and applied mathematics, and their computation is of practical importance across all stripes of quantitative scientific research. At this very minute, countless practitioners are calling a function like MATLAB's `eig`, to cluster or compress data; probe the structure of graphs and networks; solve a differential equation; or any of countless other applications. Modern algorithms for this problem are some of the crown jewels of numerical linear algebra and applied computer science, and in practice they are observed to be fast and accurate on virtually every input. And yet, despite this problem's ubiquity and decades of rich and detailed study, critical *theoretical* questions remain open, perhaps most notably being that *there is to date no rigorously proven guarantee that the algorithm employed by MATLAB's* `eig` *quickly and accurately approximates the eigenvectors and eigenvalues of every matrix.*

The purpose of this thesis is to present three main theoretical contributions to this area. Chapter 5 contains a rigorous, front-to-end finite precision analysis of a randomized variant of the well-known Spectral Bisection algorithm [30] for rapidly approximating the eigenvectors and eigenvalues of every matrix. The crucial use of randomness is in an initial step *regularizing* the input matrix with a small Gaussian perturbation, which we will show in Chapters 3-4 results in well-spaced eigenvalues and an eigenvector matrix with moderate condition number; the proofs require several tools from random matrix theory, some classic and others novel. Finally, in Chapters 6-7, we turn to the Shifted QR algorithm, whose ubiquity and observed practical efficacy in solving eigenvalue problems as yet lacks theoretical explanation. Bridging this gap in theory, we introduce a new 'shifting strategy' for the Shifted QR algorithm, and show that it can be implemented in floating point arithmetic to approximate the eigenvalues of any matrix. The results in Chapters 5 and 7 are the among the first provable algorithms for the eigenvalue problem on non-normal matrices.

1.1 Approximate Diagonalization

A matrix $A \in \mathbb{C}^{n\times n}$ is *diagonalizable* if it can be written as

$$A = VDV^{-1} \tag{1.1}$$

for some invertible V and diagonal D. The matrix D contains as its diagonal elements the eigenvalues of A, which we will denote as $\operatorname{Spec} A = \{\lambda_1, ..., \lambda_n\}$ throughout this work; the columns $v_1, ..., v_n$ of V and rows $w_1^*, ..., w_n^*$ of V^{-1} are the right and left eigenvectors of A, respectively, and invertibility of V means that the right and left eigenvectors constitute a pair of *dual bases* for $\mathbb{C}^n$. Note that V and V^{-1} are not uniquely defined, as one can freely rescale the columns of the former and adjust accordingly the rows of the latter. If A is Hermitian (or, more generally, *normal*, meaning that it satisfies $AA^* = A^*A$), then it is well-known to be diagonalizable by a unitary matrix, meaning that we can scale V to have orthonormal columns. Conversely, there exist non-normal matrices which are not diagonalizable at all, although fortunately these constitute a set of Lebesgue measure zero in $\mathbb{C}^{n\times n}$.

Example 1.1 (Non-diagonalizable Matrix)**.** Consider the $n \times n$ *Jordan block*

$$J_n \triangleq \begin{pmatrix} & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \\ & & & & \end{pmatrix},$$

which has ones on its superdiagonal and zeros elsewhere. One can quickly verify that $\operatorname{Spec} J_n = \{0\}$, but that that the standard basis vectors e_n and e_1^* are the only right and left eigenvectors (respectively) of J, each with eigenvalue 0.

Definition of diagonalizability in hand, it is natural to ask a corresponding *algorithmic* question: given A, can V and D be computed or approximated? Since the eigenvalues of A are the roots of its characteristic polynomial, when $n \geq 5$ we cannot expect a closed form expression for V and D. The following computational task will therefore be the focus of this thesis.

Problem 1.2 (Approximate Diagonalization)**.** *Let $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$; find $\widetilde{V}$ invertible and $\widetilde{D}$ diagonal so that*

$$\|A - \widetilde{V}\widetilde{D}\widetilde{V}^{-1}\| \leq \delta\|A\|.$$

Problem 1.2 asks for approximate diagonalization in the sense of *backward approximation*, i.e. that we exactly diagonalize a matrix close to A. This is a natural notion of approximation in scientific applications, when A is a data matrix measured with finite resolution and stored to some finite precision. Contrast this with *forward approximate diagonalization*, where one tries to approximate the exact eigenvectors and eigenvalues of A.

Problem 1.3 (Forward Approximate Diagonalization)**.** *Let $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$. Find $\widetilde{V}$ invertible and $\widetilde{D}$ diagonal so that, for some V and D for which $A = VDV^{-1}$,*

$$\|D - \widetilde{D}\| \le \delta\|D\|$$
$$\|V - \widetilde{V}\| \le \delta\|V\|.$$

Practical scientific settings require that we solve Problem 1.2 in *floating point arithmetic*, meaning that our algorithms store and manipulate numbers using $\lg(1/\mathbf{u})$ *bits of precision* for some $\mathbf{u} > 0$. This means we can assume, as is customary, that our algorithms can add, subtract, multiply, and divide scalars with relative error at most the *machine precision* $\mathbf{u}$. This computational setting further necessitates the backward approximate formulation of Problem 1.2: since every arithmetic operation introduces small and possibly adversarial errors, our ability to compute forward approximations to the eigenvalues and eigenvectors of a matrix is constrained by the sensitivity of these quantities under small perturbations. The following example shows that the precision necessary to solve Problem 1.3 can depend *exponentially* on the accuracy δ.

Example 1.4. Consider the family of matrices

$$J_n(\epsilon) \triangleq \begin{pmatrix} & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \\ \epsilon & & & & \end{pmatrix}$$

obtained by adding an ϵ-perturbation to the lower left corner of an $n \times n$ Jordan block. The characteristic polynomial of $J_n(\epsilon)$ is $z^n - \epsilon$, meaning that its eigenvalues are the nth roots of ϵ, and are thus depart the origin at infinite velocity. For instance, $\|J_{10}(10^{-10}) - J_{10}(0)\| = 10^{-10}$ but the eigenvalues have moved by 0.1, meaning that miniscule backward error has produced a macroscopic change in the spectrum. It is the content of Theorem 2.6 that this $\Omega(\epsilon^{1/n})$ behavior is the worst possible for an $n \times n$ matrix.

We will see in Chapter 2 that the cost of passing between forward and backward approximate diagonalization for a matrix $A = VDV^{-1}$ is controlled by the *eigenvector condition number*

$$\kappa_V(A) \triangleq \min_{V : A = VDV^{-1}} \|V\|\|V^{-1}\| \tag{1.2}$$

and *minimum eigenvalue gap*

$$\mathrm{gap}(A) \triangleq \min_{i \neq j} |\lambda_i - \lambda_j|; \tag{1.3}$$

both of these measure the non-normality of A, since as one approaches the set of non-diagonalizable matrices, former tends to infinity and the latter to zero. Let us define the *condition number of the eigenproblem* as

$$\kappa_{\mathrm{eig}}(A) \triangleq \frac{2\kappa_V(A)}{\mathrm{gap}(A)};$$

the proposition below follows from Theorem 1.8 and Lemma 2.9.

Proposition 1.5. *Let $A \in \mathbb{C}^{n\times n}$ and $\delta \le \frac{1}{2\kappa_{\mathrm{eig}}(A)}$. If $\widetilde{V}\widetilde{D}\widetilde{V}^{-1}$ is a δ-backward approximate diagonalization of A in the sense of Problem 1.2, then*

$$\|D - \widetilde{D}\| \le \kappa_{\mathrm{eig}}(A)\delta\|D\| \qquad \|V - \widetilde{V}\| \le 2n^2\kappa_{\mathrm{eig}}(A)\delta\|V\|.$$

It is known that the best constant for which conclusion of Proposition 1.5 holds for all small enough δ is $\kappa_{\mathrm{eig}}(A)$ times some polynomial in n. Since $\kappa_{\mathrm{eig}}(A)$ is unbounded over $\mathbb{C}^{n\times n}$, we have no choice but to settle for backward approximation if we seek an an approximate diagonalization algorithm that works on *every* input. The fundamental question of this thesis is therefore:

Question 1.6. *How many arithmetic operations and bits of precision are required to solve Problem 1.2 on a floating point machine, as a function of the accuracy δ and dimension n?*

For Hermitian matrices, Question 1.6 is widely considered settled by the work of Wilkinsin [164], Dekker and Traub [60], and Hoffman and Parlett [96] on the Shifted QR algorithm. However, their results are in exact arithmetic only, and as we will discuss in Section 1.3, *there is to our knowledge no published rigorous analysis of the finite arithmetic case.* Ben-Or and Eldar in [31] give an algorithm requiring $O(nT_{\mathsf{MM}}(n)\,\mathrm{poly}\log(n))$ bit operations (where $T_{\mathsf{MM}}(n)$ gives the arithmetic operations required to numerically stably multiply two $n \times n$ matrices) which solves Problem 1.2 for $\|A\| \le 1$ and $\delta = O(1/\mathrm{poly}(n))$. On the other hand, for non-Hermitian matrices Question 1.6 remained wide open until the recent work of Armentano et al., who in [4] gave an algorithm which, in $O(n^{10}/\delta^2)$ arithmetic operations, produces an approximate diagonalization of $A + \delta G_n$, where G_n is a *complex Ginibre matrix*, whose entries are independent centered complex Gaussians of variance $1/n$.

In Chapter 5, we will present the following answer to Question 1.6 which, as this thesis goes to press, stands as the (asymptotically) fastest provable approximate diagonalization algorithm in the case of non-Hermitian matrices. The algorithm in question is a variant of the Spectral Bisection algorithm of [30], and for simplicity we state it for matrices of norm at most one.[1]

Theorem 1.7. *There is a randomized algorithm,* EIG, *which on input an $n \times n$ complex matrix A with $\|A\| \le 1$, and an accuracy parameter δ, outputs V invertible and D diagonal such that*

$$\|A - VDV^{-1}\| \le \delta \qquad \|V\|\|V^{-1}\| \le 32n^{2.5}/\delta, \tag{1.4}$$

in $O(T_{\mathsf{MM}}(n)\log^2(n/\delta))$ arithmetic operations, on a floating point machine with $O(\log^4(n/\delta)\log n)$ bits of precision, with probability $1 - O(1/n)$

At a high level, the Spectral Bisection algorithm first *splits* Spec A into two roughly equal parts, $\Lambda_\pm$, and then computes matrices $Q_\pm$ whose orthogonal columns span the subspaces spanned by

[1] A solution to Problem 1.2 can be found by computing upper and lower bounds on $\|A\|$ and rescaling appropriately.

the right eigenvectors associated to each of $\Lambda_\pm$. It is not hard to see that $Q_\pm$ can be used to reduce to two subproblems of a smaller size: defining $A_\pm = Q_\pm^* A Q_\pm$, we have $\operatorname{Spec} A = \operatorname{Spec} A_+ \sqcup \operatorname{Spec} A_-$, and moreover every eigenvalue of A is of the form $Q_\pm^* v_\pm$, where $v_\pm$ is an eigenvector of $A_\pm$. Of course, without prior knowledge of the spectrum or eigenvectors, it remains mysterious how one could bisect the former or reason about subspaces spanned by the latter.

The key ingredient for both of these tasks is the *matrix sign function*, usually attributed to Zolotarev and dating back to the late nineteenth century. Recall that for $z \in \mathbb{C}$, $\operatorname{sgn}(z) = \pm 1$ according to whether $\Re z$ is positive or negative, and is undefined for z on the imaginary axis. Analogously, if $A = VDV^{-1}$ is diagonalizable and has no eigenvalues on the imaginary axis, then

$$\operatorname{sgn}(A) \triangleq V \operatorname{sgn}(D) V^{-1},$$

where the middle term indicates that we are to apply sgn to each of the diagonal elements of D. Pleasantly, $\operatorname{tr}\operatorname{sgn}(A)$ encodes the number of eigenvalues in the right vs. left halfplane of $\mathbb{C}$, and the column spaces of the matrices $\operatorname{sgn}(A) \pm 1$ are equal to the spans of the eigenvectors in each of these halfplanes. We can therefore use it to split the spectrum (e.g. by computing $\operatorname{tr}\operatorname{sgn}(z - wA)$ for several $z, w \in C$ until a suitable bisection is found) and compute the requisite projections $Q_\pm$ (e.g. by running Gram-Schmidt on $1 \pm \operatorname{sgn}(z - wA)$).

This discussion perhaps raises more questions than it answers — after all, how does one compute $\operatorname{sgn}(A)$ without V or D? Finding a stable and provable method in finite arithmetic has been a long-standing open problem, but in exact arithmetic there is a simple iterative method due to Roberts [133]:

$$\begin{aligned} A_0 &= A \\ A_{t+1} &= \tfrac{1}{2}(A_t + A_t^{-1}). \end{aligned}$$

(The reader may recognize the above as the Newton iteration for computing roots of $z^2 - 1$). In Theorem 5.1 we will show that Roberts' iteration converges in finite arithmetic after $O(\log(1/\epsilon))$ iterations, so long as A_0 is ϵ-far from every matrix with a purely imaginary eigenvalue. This criterion was identified as a natural measure of conditioning for sgn in [10].

Chapter 5 checks the many remaining details, and on the whole our strategy is to show that the Spectral Bisection algorithm is convergent and numerically stable on matrices for which κ_{eig} is a modest polynomial in n. To upgrade this fact to an algorithm for approximately diagonalizing *every* matrix A, we proceed in the footsteps of [6, 31] and preprocess A with a small amount of additive Gaussian noise (say, $(\delta/8)G_n$ if we are hoping for accuracy δ), whose purpose is to effectuate any needed reduction in $\kappa_{\mathrm{eig}}(A)$. This *smoothed analysis* approach — where we study the performance of algorithms on inputs that are adversarially chosen and then randomly perturbed — was pioneered in [143] in the context of linear programming, building off of earlier work by Demmel and Edelman on the condition number of Gaussian matrices [63, 70], and used to great effect in [136] to study linear system solvers. The hypothesis of bounded κ_{eig} is critical to our analysis of the Spectral Bisection algorithm, and since we are in the business of backward approximation in the first place, it is well worth spending some of our 'budget' of accuracy passing to a well-conditioned matrix.

1.2 Gaussian Regularization

A critical tool in the study of eigenvalue perturbation theory generally, and in this thesis specifically, is the *ϵ-pseudospectrum*, a subset of the complex plane parametrized by $\epsilon \geq 0$ which contains the eigenvalues of all ϵ-close matrices:

$$\begin{aligned} \Lambda_\epsilon(A) &= \bigcup_{\widetilde{A} : \|A - \widetilde{A}\| \leq \epsilon} \operatorname{Spec} \widetilde{A} && (1.5) \\ &= \left\{ z \in \mathbb{C} \ : \ \|(z - A)^{-1}\| \geq 1/\epsilon \right\} && (1.6) \end{aligned}$$

The proof of the equivalent definition in (1.6), along with countless other results, can be found in the essential book [153]. By definition, pseudospectrum is stable under small perturbations — in the sense that if $\|A - \widetilde{A}\| \leq \epsilon$, then $\Lambda_{\epsilon - \|A - \widetilde{A}\|}(\widetilde{A}) \subset \Lambda_\epsilon(A) \subset \Lambda_{\epsilon + \|A - \widetilde{A}\|}(\widetilde{A})$ — and is thus particularly well-suited to the study of backward approximate diagonalization. On the other hand, $\Lambda_\epsilon(A)$ always contains a disk of radius ϵ about every eigenvalue (by considering the perturbation $\widetilde{A} = A + \zeta$ for some $|\zeta| \leq \epsilon$ in $\mathbb{C}$), and its size and shape in excess of these disks captures geometrically the cost when translating from backward to forward approximation. The Bauer-Fike theorem [28, Theorem III] is a classic result bounding the pseudospectrum in terms of the eigenvector condition number.

Theorem 1.8 (Bauer-Fike)**.** *If $A \in \mathbb{C}^{n \times n}$ is any matrix, then*

$$\bigcup_{\lambda \in \operatorname{Spec} A} \mathbb{D}(\lambda, \epsilon) \subset \Lambda_\epsilon(A) \subset \bigcup_{\lambda \in \operatorname{Spec} A} \mathbb{D}(\lambda, \kappa_V(A) \cdot \epsilon).$$

When A is normal, $\kappa_V(A) = 1$, so the containments above become equalities; at the other extreme, Example 1.4 can be adapted to give an example where $\Lambda_\epsilon(A)$ contains a ball of radius $\epsilon^{1/n}$ about each eigenvalue. The virtue of Theorem 1.8 is that it bounds the displacement of each eigenvalue *non-asymptotically*; but it is not always tight. When A has distinct eigenvalues $\lambda_1, ..., \lambda_n$, each with left and right eigenvectors w_i^* and v_i, the *instantaneous* rate of change of λ_i upon perturbation is at most the the *eigenvalue condition number*

$$\kappa_i(A) \triangleq \frac{\|w_i\|\|v_i\|}{|w_i^* v_i|}. \tag{1.7}$$

The perturbation theory for eigenvectors is somewhat more delicate, since small perturbations of a matrix with eigenvalue multiplicity can generate severely ill-conditioned eigenvectors, as the following example illustrates.

Example 1.9. Consider the matrices $1 + \epsilon J_n$. When $\epsilon = 0$, we have the $n \times n$ identity matrix, with eigenvector condition number $\kappa_V(1) = 1$. On the other hand, for any $\epsilon > 0$, $1 + \epsilon J_n$ is non-diagonalizable!

If A has distinct eigenvalues $\lambda_1, ..., \lambda_n$, Theorem 1.8 ensures the eigenvalues will *remain* distinct after perturbations of scale $\kappa_{\text{eig}}(A)$.

A critical step in Theorem 1.7 is to show that one can tame the pseudospectrum, gap, and eigenvector and eigenvalue condition numbers of *any* matrix by adding a small, complex Gaussian perturbation. The following result is proved in Chapter 3.

Theorem 1.10. *Let A be any $n \times n$ matrix with $\|A\| \le 1$, G_n a complex Ginibre matrix, and $\delta \le 1/2$. Then, with probability at least $1 - 10/n$,*

$$\begin{aligned}\sum_i \kappa_i^2(A + \delta G_n) &\le n^3/\delta^2 \\ \kappa_V(A + \delta G_n) &\le n^2/\delta \\ \text{gap}(A + \delta G_n) &\ge \delta/n^2\end{aligned}$$

and $\Lambda_\epsilon(A + \delta G_n)$ has n disjoint connected components for every $\epsilon \le \delta^2/2n^4$.

The above additionally resolves a conjecture of E. B. Davies [54]: *every $n \times n$ matrix is δ-close to a marix with eigenvector condition number* $O(\text{poly}(n)/\delta)$. Theorem 1.10 is illustrated in Figure 1.2, which depicts the pesudospectrum of a non-diagonalizable matrix before and after a random Gaussian perturbation.

In Chapter 4, we show that an anologue of Theorem 1.10 holds in the case when $A \in \mathbb{R}^{n\times n}$ is an arbitrary *real* (but still not necessarily symmetric) matrix, and the complex Ginibre perturbation G_n is replaced by a *real* Ginibre matrix H_n, whose entries are independent, centered real Gaussians of variance $1/n$. This sort of result is desirable, for instance, in practical settings where one seeks to regularize a real matrix without resorting to complex arithmetic.

Theorem 1.11. *Let $n \ge 7$, $A \in \mathbb{R}^{n\times n}$ be a matrix with $\|A\| \le 1$, H_n a real Ginibre matrix, and $\delta \le 1$. Then, with probability at least $1 - O(1/n)$,*

$$\begin{aligned}\sum_{i:\lambda_i\in\mathbb{R}} \kappa_i(A + \delta H_n) &= O(n^2/\delta) \\ \sum_{i:\lambda_i\in\mathbb{C}\setminus\mathbb{R}} \kappa_i^2(A + \delta H_n) &= O(n^6/\delta^3 \cdot \log(n/\delta)) \\ \kappa_V(A + \delta H_n) &= O(n^4/\delta^{3/2}\sqrt{\log(n/\delta)}) \\ \text{gap}(A + \delta H_n) &= \Omega(\delta^9/n^{14}).\end{aligned}$$

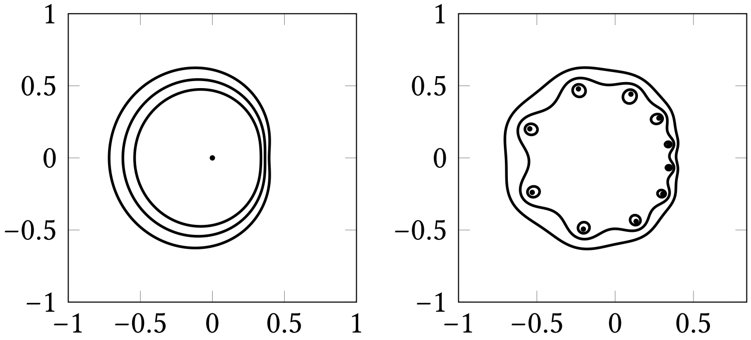

Figure 1.1: T is a sample of an upper triangular 10×10 Toeplitz matrix with zeros on the diagonal and independent (modulo the Toeplitz structure) standard real Gaussian entries above the diagonal. Pictured is the boundary of the ϵ-pseudospectrum of T (left) and $T + 10^{-6}G_n$ (right) for $\epsilon = 10^{-5}$, $\epsilon = 10^{-5.5}$, and $\epsilon = 10^{-6}$, along with the spectra. These plots were generated with the MATLAB package EigTool [165].

In fact, Theorem 1.11 extends even to non-Gaussian perturbations whose entries are independent (but not necessarily distributed) real random variables whose densities are absolutely continuous with respect to the Lebesgue measure. The corresponding statements may be found in Chapter 4.

The key ingredient to prove Theorem 1.10 is to furnish appropriate *singular value tail bounds* for the matrices $z - A - \delta G_n$, for $z \in \mathbb{C}$. Since $z \in \Lambda_\epsilon(A + \delta G_n)$ if and only if $\sigma_n(z - A - \delta G_n) \le \epsilon$, by applying Fubini one can for instance compute that, for any Lebesgue measurable set $\Omega \subset \mathbb{C}$,

$$\mathbb{E}\,\mathrm{Leb}\,\Lambda_\epsilon(A + \delta G_n) \cap \Omega = \mathbb{E}\int_\Omega \mathbf{1}\{\sigma_n(z - A - \delta G_n) \le \epsilon\}\,\mathrm{d}z = \int_\Omega \mathbb{P}[\sigma_n(A + \delta G_n) \le \epsilon]\,\mathrm{d}z,$$

where Leb denotes the Lebesgue measure on $\mathbb{C}$, and $\mathbf{1}\{\cdot\}$ is the indicator function. In Chapter 3 below, we will see that the volume of the ϵ-pseudospectrum scales like $\pi\epsilon^2 \sum_i \kappa_i^2$ as $\epsilon \longrightarrow 0$, so an $O(\epsilon^2)$ bound for the singular value tail event in the integrand gives us control over the eigenvalue condition numbers, which can in turn be used to control $\kappa_V(A + \delta G_n)$.

On the other hand, by the log-majorization property of singular values and eigenvalues, if $\lambda_1, ..., \lambda_n$ are the (random) eigenvalues of $A + \delta G_n$, then $\sigma_n(z - A - \delta G_n)\sigma_{n-1}(z - A - \delta G_n) \le r^2$ if and only if two eigenvalues of $A + \delta G_n$ lie in $\mathbb{D}(z, r)$. By taking an appropriate net of z, one can thus once again control the probability that $\mathrm{gap}(A + \delta G_n)$ is small by way of tail bounds on the smallest two singular values of $A + \delta G_n$. This approach gives a quick bound with a simple proof; the improved gap bound in Theorem 1.10 comes via an alternate approach using a result in [6].

The tail bounds themselves are proved via a comparison result of P. Śniady from the context of free probability, which allows one to transfer classical tail bounds on $\sigma_i(G_n)$ [70] to ones on $\sigma_i(A + G_n)$ for an arbitrary matrix A.

The real Ginibre result in Theorem 1.11 requires several additional innovations. In Chapter 4 we will prove an analogue of Śniady's result for real Ginibre matrices, but this allows us only to obtain shifted singular value tail bounds for real z, with $O(\epsilon)$ as opposed to $O(\epsilon^2)$ behavior. Luckily, a variant of the limiting area formula discussed above maybe used to relate the limiting *length* of the pseudospectrum on the real line with the eigenvalue condition numbers of the real eigenvalues. For nonreal z, we are forced to produce the tail bounds by other means, and find behavior of type $O(\epsilon^2/|\Im z|)$. The complex Ginibre argument can thus be repeated — so long as we verify that there are only rarely eigevalues with small imaginary part. Since the complex eigenvalues of real matrices come in complex conjugate pairs, such a result follows from a lower bound on the minimum eigenvalue gap, which we obtain similarly to the complex Ginibre case. Finally, in the case of generic absolutely continuous real perturbations we lose Śniady-type theorem's entirely, and must develop further machinery.

1.3 The Shifted QR Algorithm

Much of the practical and theoretical study of approximate diagonalization has centered on the *shifted QR algorithm*, which was introduced independently by Francis and Kublanovskaya [78, 106] in the 1950s and remains the state of the art for solving Problem 1.2 in floating point arithmetic on dense matrices. Recalling that each (nonsingular) matrix $A \in \mathbb{C}^{n\times n}$ has a unique *QR decomposition*

$$A = QR$$

where $Q \in \mathbb{C}^{n\times n}$ is unitary and $R \in \mathbb{C}^{n\times n}$ is upper triangular with nonnegative diagonal, the elementary QR algorithm is the iterative procedure

$$\begin{aligned} A_0 &= A \\ A_{t+1} &= R_t Q_t, \qquad \text{where } A_t = Q_t R_t \text{ is (of course) the QR decomposition.} \end{aligned}$$

As $A_{t+1} = R_t Q_t = Q_t^* Q_t R_t Q_t = Q_t^* A_t Q_t$, this procedure produces a sequence $A = A_0, A_1, \ldots$ of matrices unitarily similar to A, and generically the sequence (A_t) converges to an upper triangular matrix containing $\operatorname{Spec} A$ on its diagonal [78, 106]. However, each iteration is expensive (as QR decomposition is as hard as matrix multiplication), and the kth subdiagonal entry tends to zero with rate depending on $|\lambda_{k+1}(A)/\lambda_k(A)|$ — meaning that the convergence can be arbitrarily slow.

To address these limitations, the modern QR algorithm contains three key innovations. The first is the observation that, at a cost of $O(n^3)$ arithmetic operations, every matrix can be converted to a unitarily equivalent *upper Hessenberg matrix*, namely an "almost-triangular" matrix $H \in \mathbb{C}^{n\times n}$ with $H_{i,j} = 0$ for all $i > j$. This is significant in that each QR iteration can be executed numerically stably on Hessenberg matrices in time $O(n^2)$, and moreover preserves the Hessenberg structure. The second innovation is the use of a *shifting strategy* to speed convergence. In Chapters 6-7 we will use the following notation. Let

$$\mathsf{Sh} : \mathbb{H}^{n\times n} \longrightarrow \mathcal{P}_k$$

be a function which assigns to each Hessenberg matrix a polynomial p of degree at most k, and iterate as

$$\begin{aligned} H_0 &= \mathsf{Hessenberg}(A) \\ H_{t+1} &= Q_t^* H_t Q_t, \qquad \text{where } p_t(H_t) = Q_t R_t \text{ is the QR decomposition, and } p_t = \mathsf{Sh}(H_t). \end{aligned}$$

The role of the shifting polynomial can be interpreted as accentuating the ratios between eigenvalues of $p_t(H_t)$, thus skirting the source of slow convergence in the elementary algorithm.

The final modern innovation is to further exploit the Hessenberg structure by way of *decoupling and deflation.* Specifically, we say that an iterate H_t is *ω-decoupled* if one of its subdiagonal elements is smaller than $\omega\|H\|$. Equivalently, this means H_t is $\omega\|H_t\|$-close in operator norm to a block upper triangular matrix. Since we are in the business of backward approximation, one can now approximate $\operatorname{Spec} H_t = \operatorname{Spec} H$ by finding the eigenvalues of the resulting diagonal blocks. Crucially, these blocks are themselves upper Hessenberg, and we can recursively apply the same Shifted QR iteration; taking $\omega = O(\delta/n)$ will result a δ-backward approximation of all eigenvalues of H_0. This passage to a nearby block upper triangular matrix is known as *deflation.*

Thus the following question is natural; it has been the subject of numerous papers in linear algebra and dynamical systems over the past half century, and is our focus in Chapters 6-7.

Question 1.12. *Is there an efficiently computable shifting strategy* Sh *which* ***provably***

(i) achieves rapid decoupling on non-Hermitian Hessenberg matrices, and

(ii) can be implemented numerically stably in finite arithmetic?

Exact Arithmetic

In the Hermitian case, it is Question 1.12(i) which was solved by the aforementioned work of Wilkinson, Dekker-Traub and Hoffman-Parlett [164, 60, 96]. The *Wilkinson shift* considered there is the linear shift $z - r_W$, where r_W is whichever eigenvalue of H_t's lower right 2×2 corner is closer to the entry $(H_t)_{n,n}$. Dekker-Traub and Hoffman-Parlett show, amazingly, that Wilkinson's shift achieves ω-decoupling in $O(\log(1/\omega))$ iterations on Hermitian matrices in exact arithmetic, building on Wilkinson's initial proof of global convergence (without a rate). In the non-Hermitian case, a shifting strategy with global convergence on unitary Hessenberg matrices is evinced by Wang and Gragg in [157, 158], and otherwise no strategy was known to converge globally (let alone quickly) on any appreciable category of non-Hermitian matrices. Moreover, to our knowledge Question 1.12(ii) has remained entirely unanswered: it has not previously been proven that Wilkinson's (or, indeed, any other) shift converges rapidly when implemented in finite arithmetic, even in the case of real symmetric matrices.

The final contribution of this thesis is to answer both parts of Question 1.12 by producing a rapidly decoupling and stably implementable shifting strategy for matrices with bounded κ_{eig}.

Like Wilkinson's shift and many that have been invented since, our strategy uses the *Ritz values* of H_t (in other words the eigenvalues of the lower right hand $k \times k$ corner of H_t) to produce the shift polynomial p_t. Ritz values (being eigenvalues) cannot in general be computed exactly, and the shifting strategy below asks only that we approximate them in a certain sense to be made specific in the sequel (see Definition 6.3). For now let us call an algorithm which executes this approximation a *Ritz value finder.* In Chapter 6, we will prove the following exact arithmetic guarantee for our strategy.

Theorem 1.13. *Let $\mathbb{H}_B^{n\times n}$ denote the set of $n \times n$ Hessenberg matrices with eigenvector condition number at most B. For each $k = 2, 4, 8, \ldots$ and $B \geq 1$, there is a shifting strategy $\mathsf{Sh}_{k,B} : \mathbb{H}_B^{n\times n} \longrightarrow \mathcal{P}_k$ which, in exact arithmetic,*

(i) achieves ω-decoupling in $4\log_2(1/\omega)$ iterations for every matrix in $\mathbb{H}_B^{n\times n}$, and

(ii) costs $O((\log k + B^{\frac{16 \log k}{k}})n^2 + \mathrm{poly}(k)\log(1/\omega))$ arithmetic operations, plus one call to a Ritz value finder, per iteration.

A few comments are in order. Given the presence of the term $B^{O(\log k/k)}$ in the runtime, we will need to take $k = O(\log B \log\log B)$ to obtain a reasonably efficient algorithm. One should therefore regard our result as a reduction: given an order-k Ritz value finder, we can decouple arbitrarily large matrices with condition number nearly exponential in k. Moreover, by Theorem 1.10, every $H \in \mathbb{H}^{n\times n}$ is $\delta\|H\|$-close to a matrix in $\mathbb{H}^{n\times n}_{n^2/\delta}$, so Theorem 1.13 asserts that most Hessenberg matrices can be ω-decoupled in $O(\log n \log\log n \cdot n^2 \cdot \log(1/\omega))$ arithmetic operations.

Much of the intuition for our shifting strategy is captured by $\mathsf{Sh}_{2,1}$ which, as we will now sketch, gives a globally and rapidly convergent shifting strategy for normal matrices — itself an open problem. We will use, without proof, a few standard facts about QR iteration, developed fully in Chapter 6. At a high level $\mathsf{Sh}_{2,1}$ first tries a *main shift* defined, like Wilkinson's, in terms of the order-2 Ritz values $\mathcal{R} = \mathrm{Spec}(H_t)_{(2)}$, this time taking the shifting polynomial to be

$$(z - r_*)^2, \qquad \text{where } r_* = \operatorname*{argmax}_{r\in\mathcal{R}} \|e_n^*(H_t - r)^{-2}\|.$$

In the event that the main shift fails to make progress towards convergence, $\mathsf{Sh}_{2,1}$ uses this information to quickly produce an *exceptional shift* which succeeds. Such an approach was used in [69, 157] in the case of unitary Hessenberg matrices.

It is standard and sensible to measure the progress of Shifted QR step by way of the *potential function* $\psi(H) = (H_{n-1,n-2}H_{n,n-1})^{1/2}$, since we can guarantee ω-decoupling by driving down ψ. The motivation for our main shift is that, if p is a monic polynomial and $\widehat{H}$ is the result of one QR step with shift p, the potential of $\widehat{H}$ may be bounded as

$$\psi(\widehat{H}) \leq \|e_n^* p(H)^{-1}\|^{-1/\deg p}. \tag{1.8}$$

In other words, we are choosing the Ritz value for which this upper bound is minimal. One can show that this shift always causes the potential to decrase or remain fixed, by reasoning in terms of a certain spectral measure.

Proof Sketch of Monotonicity. Let H be a normal Hessenberg matrix, with unitary diagonalization $H = UDU^*$. Since $e_n^* U$ is a unit vector, define Z_H to be a random variable supported on $\operatorname{Spec} H$, with distribution $\mathbb{P}[Z_H = \lambda_i] = |U_{n,i}|^2$; then for any function $f : \operatorname{Spec} H \longrightarrow \mathbb{C}$ we can write

$$\|e_n^* f(H)\|^2 = e_n^* U f(D) f(D)^* U^* e_n = \sum_i |U_{n,i}|^2 |f(\lambda_i)|^2 = \mathbb{E}|f(Z_H)|^2. \tag{1.9}$$

The potential $\psi(H)$ has an equivalent variational definition via Z_H. Write $\chi_2(z)$ for the characteristic polynomial of $H_{(2)}$; then

$$\psi(H) = \left(\mathbb{E}|\chi_2(Z_H)|\right)^{1/4} = \min_{p \in \mathcal{P}_2} \left(\mathbb{E}|p(Z_H)|^2\right)^{1/4}. \tag{1.10}$$

By the definition of r_* above as the maximizer of $\|e_n^*(H - r)^{-1}\| = \mathbb{E}|Z_H - r|^{-2}$ over the two roots r_1, r_2 of χ_2, the inequality of arithmetic and geometric means gives

$$\mathbb{E}|Z_H - r_*|^{-2} \geq \mathbb{E}\frac{1}{2}\left(|Z_H - r_1|^{-2} + |Z_H - r_2|^{-2}\right) \geq \mathbb{E}|\chi_2(Z_H)|^{-1}. \tag{1.11}$$

Finally, let $\widehat{H}$ be the result of a QR step with the shift $(z - r_*)^2$. Combining the preceding discussion with two applications of Jensen's inequality, we have

$$\begin{aligned}
\psi_k(\widehat{H}) &\leq \left(\mathbb{E}|Z_H - r_*|^{-4}\right)^{-1/4} && \text{(1.8)-(1.9)} \\
&\leq \left(\mathbb{E}|Z_H - r_*|^{-2}\right)^{-1/2} && \text{Jensen} \\
&\leq \left(\mathbb{E}|\chi_2(Z_H)|^{-1}\right)^{-1/2} && \text{(1.11)} \\
&\leq (\mathbb{E}|\chi_2(Z_H)|)^{1/4} && \text{Jensen} \\
&= \psi_k(H) && \text{(1.10).}
\end{aligned}$$

□

The above is *not* sufficient to give global convergence of the shift $(z - r_*)^2$, and indeed it can happen that this shift fails to reduce the potential. However, this *stagnation* can only occur for matrices with a particular structure: in the event that $\psi_k(\widehat{H}) = \psi_k(H)$, the first application of Jensen's inequality holds with equality, which means that $|Z_H - r_*|$ is identically equal to $\psi(H)$, or in other words that Z_H — and thus $\operatorname{Spec} H$ — is supported on a circle of radius $\psi(H)$ about r_*![2] This can be made quantitative, in the sense that in the sense that if $\psi(\widehat{H}) \approx \psi(H)$, then the mass

[2] Compare to Parlett's result in [125] scale multiples of unitaries are the only Hessenberg matrices fixed under an unshifted QR step.

of Z_H is concentrated on a disk Ω of radius roughly $\psi(H)$ centered at the stagnated shift r_*. It is this concentration that facilitates the computation of an exceptional shift: by searching over a deterministic lattice of points in Ω, one can quickly find a point s close enough to $\operatorname{Spec} H$ to ensure potential reduction with the shift $(z-s)^2$. The crucial feature here is that failure of the main shift to make progress tells you the correct scale to on which to search for an exceptional one.

Boosting this argument to matrices with bounded eigenvalue condition number requires a few further ideas, which are the subject of Chapter 6. Specifically, we lose in the non-normal case the ability to express quantities like $\|e_n^* f(H)\|$ in terms of a spectral measure Z_H as in (1.9), since H is no longer unitarily diagonalizable. Instead, we will design an analogue of Z_H for which the equalities in (1.9) hold only up to a factor of $\kappa_V(H)$, which in turn necessitates using higher-degree shifts and higher-order Ritz values.

Floating Point Arithmetic

In the final Chapter 7, we will show that if k is suitably chosen, the strategy $\mathsf{Sh}_{k,B}$ can be implemented in finite arithmetic, along with a suitable protocol for deflation and recursion, to find every eigenvalue of a Hessenberg matrix with bounded eigenvector condition number and minimum eigenvalue gap. As in Theorem 1.13, the result below is a reduction: we show that the strategy can be implemented with polynomially many calls to a *forward* approximate diagonalization algorithm that works on $k \times k$ or smaller matrices. Specifically, assume that we have access to an algorithm, $\mathsf{SmallEig}(A, \delta, \phi)$ on input a matrix A of dimension at most k, with probability at least $1-\phi$ approximates each eigenvalue of A with additive error δ.[3] The role of $\mathsf{SmallEig}$ is twofold: we use it (i) produce the approximate Ritz values used by $\mathsf{Sh}_{k,B}$; and (ii) as a 'base case' once we have decoupled to matrices of size at most $k \times k$.

Theorem 1.14. *Let $H \in \mathbb{H}^{n\times n}_{B/2}$, and assume further that $\|H\| \le 1$ and $\operatorname{gap}(H)/2 \le \Gamma$. For some $k = O(\log B \log\log B)$, the shifting strategy $\mathsf{Sh}_{k,B}$ can be implemented in finite arithmetic to give a randomized shifted QR algorithm, $\mathsf{ShiftedQR}$, with the following guarantee: for any $\delta > 0$ $\mathsf{ShiftedQR}(H, \delta, \phi)$ produces the eigenvalues of a matrix $\widetilde{H}$ with $\|H - \widetilde{H}\| \le \delta$, with probability at least $1-\phi$, using*

(i) $O\left((\log \frac{nB}{\delta\Gamma} k \log k + k^2)n^3\right)$ arithmetic operations on a floating point machine with $O(k \log \frac{nB}{\delta\Gamma\phi})$ bits of precision; and

(ii) $O(n \log \frac{nB}{\delta\Gamma})$ calls to $\mathsf{SmallEig}$ with accuracy $\Omega(\frac{\delta^2\Gamma^2}{n^3B^4})$ and failure probability tolerance $\Omega(\frac{\phi}{n^2 \log \frac{nB}{\delta\Gamma}})$.

Remark 1.15. Theorem 1.14 can easily be adapted to find backward approximations to the eigenvalues of any matrix $A \in \mathbb{C}^{n\times n}$. First, pass to $\widetilde{A} = A + (\delta/8)G_n$, which with probability $1 - O(1/n)$ has $\kappa_V(\widetilde{A}) = O(n^2/\delta)$, $\operatorname{gap}(\widetilde{A}) = \Omega(\delta/n^2)$, and $\|A - \widetilde{A}\| \le \delta$. Second, compute $H =$

[3] In practice it is common to solve these smaller eigenvalue problems recursively, but such an approach is suboptimal in this context.

$\mathsf{Hessenberg}(\widetilde{A})$ and renormalize so that $\|H\| \le 1$. Finally, run $\mathsf{ShiftedQR}(H, \delta/2, \Theta(1/n))$ with $k = O(\log(n/\delta)\log\log(n/\delta))$, giving

$$\begin{aligned} O\left(\log^2(n/\delta)\,\mathrm{poly}(\log\log(n/\delta))\cdot n^3\right) &\qquad \text{arithmetic operations, and} \\ O\left(\log^2(n/\delta)\log\log(n/\delta)\right) &\qquad \text{bits of precision,} \end{aligned}$$

plus the $O(n\log(n/\delta))$ calls to $\mathsf{SmallEig}$. Implementing $\mathsf{SmallEig}$ by calling EIG with accuracy $\Omega((\delta/\mathrm{poly}(n))^k)$ uses $n\,\mathrm{poly}(\log(n/\delta))$ bit operations, for a large but not preposterous polynomial. Compared to Theorem 1.7, this approach has improved precision when $\delta = \Omega(ne^{-n})$, but suffers on the other hand from inflated arithmetic (and boolean) operations.

Remark 1.16. Omitted from Theorem 1.14 is any discussion of the eigenvectors, but these can be found in a variety of ways. One option is to use shifted inverse iteration with the computed eigenvalues guiding the choice of shifts.

One major obstacle to proving Theorem 1.14 — and indeed to proving rapid convergence of Shifted QR in finite arithmetic for any shifting scheme — is shift instability. To be precise, even though QR iteration steps are backward stable (meaning that the computed $\widetilde{H}_{t+1}$ is a unitary conjugate of H_t by some $\widetilde{Q}_t$ appearing in the QR decomposition of a Hessenberg matrix close to H_t), they are only forward stable when $p_t(H_t)$ is far from singular. On the other hand, the roots of p_t are meant to approximate eigenvalues of H_t in order to speed convergence, so we should not expect $p_t(H_t)$ to remain far from singular! And without forward stability, it is not clear how to translate a proof of rapid convergence in exact arithmetic to the finite arithmetic case.

In particular, one can show (see Lemma 7.12) that the computed H_{t+1} has absolute entrywise error of

$$O\left(\kappa_V(H_t)\left(\frac{\|H_t\|}{\mathrm{dist}(\mathrm{Roots}\,p_t, \mathrm{Spec}\,H)}\right)^k \mathbf{u}\right).$$

Since we are hoping to prove ω-decoupling for $\omega = O(\delta/n)$, we need to reason about changes in the subdiagonal entries of the iterates, which seems to require forward stability of the shifts. Thus we are forced to set the machine precision $\mathbf{u}$ on the order of $\mathrm{dist}(\mathrm{Roots}\,p_t, \mathrm{Spec}\,H)^k$. We address this by randomly perturbing each shift at scale $\Omega(\delta^2)$, where δ is the desired accuracy. For $\mathrm{gap}(H)$ suitably large, such a perturbation with constant probability is at least $\Omega(\delta^2)$-far from $\mathrm{Spec}\,H$, so we are able to get a away with $O(k\log 1/\delta)$ bits of precision (hiding the dependence on other variables). Reducing this would require a fundamentally new approach, and one intriguing starting point is the unproven observation in [159] that QR steps with ill-conditioned shifts often induce immediate decoupling.

The other obstacle is one that is rarely remarked on in the Shifted QR literature: how to compute or approximate the Ritz values of each iterate that are used to produce the shift, and, given the discussion above, what to do in the event a Ritz value lands too close to an eigenvalue. We will show that after computing $\Omega(\delta^2)$-*forward* approximations to the Ritz values and randomly perturbing at scale $\Omega(\delta^2)$, the resulting points are either (i) a set of approximate Ritz values in the sense required for $\mathsf{Sh}_{k,B}$ and or (ii) can be used to rapidly decouple the current matrix — and in either case are $\Omega(\delta^2)$ separated from the eigenvalues, to ensure forward stability when they used as shifts. This is similar in spirit to the phenomenon of 'premature deflation' discussed in [130].

1.4 Bibliographic Note

The main results in this thesis are all joint work with various of J. Garza-Vargas, A. Kulkarni, S. Mukherjee, and N. Srivatava, and have either appeared already in print, are currently in review, or are in preparation. Chapter 2 contains standard preliminary material; some of the presentation is drawn from [19, 15]. Chapters 3-4 are adapted from [19], [15, Section 3 and Appendix D], and [16] in order to unify the presentation and reduce redundancies. Chapter 5 contains the remainder of [15], Chapters 6 and 7 contain [17] and [18], respectively, and all three have been lightly edited and rearranged to suit this setting. Finally, Figure 1.2 from the current chapter appeared originally in [15]

Chapter 2

Preliminary Material

Vectors, Matrices, and Norms

We will work with finite dimensional complex matrices and vectors throughout this thesis. Let $A \in \mathbb{C}^{n\times n}$ denote a matrix and $v \in \mathbb{C}^n$ a vector; we will denote by A^* and v^* the conjugate transpose and $A^\intercal$ and $v^\intercal$ the transpose of A and v, respectively, so that w^*v gives the usual inner product between $v, w \in \mathbb{C}^n$. Unless otherwise specified, $\|v\| = \sqrt{v^*v}$ will denote the ℓ^2 vector norm, and

$$\|A\| \triangleq \sup_{v\in\mathbb{C}^n} \frac{\|Av\|}{\|v\|}$$

the induced *operator norm.* We will write range A for the span of the columns of A, tr A for its trace, and det A for its determinant, as usual. Because there is rarely chance of confusion, we will write the identity matrix as 1, and shorten $z \triangleq z1$ for scalars $z \in \mathbb{C}$.

Eigenvalues, Eigenvectors, and Singular Values

It is standard that every $n \times n$ complex matrix A has a *spectrum* Spec A consisting of n eigenvalues $\lambda_1(A), ..., \lambda_n(A)$ (possibly with multiplicity), equal to the roots of its characteristic polynomial $\chi_A(z) \triangleq \det(z - A)$; we will usually drop from the notation the dependence of the eigenvalues on the matrix. The *singular values* of A are the square roots of the n nonnegative eigenvalues of AA^*, which we will write as

$$\|A^{-1}\|^{-1} = \sigma_n(A) \le \cdots \le \sigma_1(A) = \|A\|.$$

Finally, we will denote the *Frobenius* (or entrywise ℓ^2) norm by

$$\|A\|_F = \sqrt{\sum_{i,j} |A_{i,j}|^2} = \sqrt{\operatorname{tr} AA^*} = \sqrt{\sum_i \sigma_i^2(A)}.$$

Vectors v_i and w_i^* are *right and eigenvectors* associated to an eigenvalue $\lambda_i \in \operatorname{Spec} A$ if $Av_i = \lambda_i v_i$ and $w_i^* A = \lambda_i w_i^*$, respectively. Additionally a subspace $\mathcal{V} \subset \mathbb{C}^n$ is an *invariant subspace* of A if

$Av \in \mathcal{V}$ for every $v \in \mathcal{V}$. We say that A is *normal* if $AA^* = A^*A$ (in which it has an orthonormal basis of right eigenvectors that are additionally left eigenvectors), *Hermitian* if moreover $A = A^*$ (in which case $\operatorname{Spec} A \subset \mathbb{R}$), and *unitary* if $AA^* = 1$.

Recall from the introduction that A is *diagonalizable* if for some invertible $V \in \mathbb{C}^{n\times n}$ it can be written as VDV^{-1}. We will write $v_1, ..., v_n$ for the columns of V and $w_1^*, ..., w_n^*$ for the rows of V^{-1}, which constitute a pair of biorthogonal bases of right and left eigenvectors for A. Unless we specify otherwise, we will always assume the normalization $w_i^* v_i = 1$. In this case we can alternatively write A in terms of its *spectral expansion*

$$A = \sum_i \lambda_i P_i \triangleq \sum_i \lambda_i v_i w_i^*.$$

Spectral Projectors, Resolvent, and Holomorphic Functional Calculus

A *projector* is a matrix $P \in \mathbb{C}^{n\times n}$ satisfying $P^2 = P$; true to its name, for any vector v, Pv is the projection of v to range P. A projector is *orthogonal* if $P^*P = 1$. The rank-one matrices $P_i \triangleq v_i w_i^*$ appearing in the sum above are examples of *spectral projectors* for A, meaning that they are projectors that furthermore satisfy $AP = PA$. (They need not be orthogonal projectors.) Each spectral projector for A is a projection onto a invariant subspace of A, and in fact the two are in one-to-one correspondence. These projectors — and many other properties of A — can be studied by way of the *resolvent* $(z - A)^{-1}$ of A, a matrix-valued function of the complex variable z which is holomorphic on $\mathbb{C} \setminus \operatorname{Spec} A$. The following *resolvent identity* is standard

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

If $\operatorname{Spec} A$ is contained in a region $\Omega \subset \mathbb{C}$ whose simply connected components Ω_i have rectifiable boundary, and $f : \Omega \longrightarrow \mathbb{C}$ is a holomorphic function, one can use the resolvent to define

$$f(A) \triangleq \frac{1}{2\pi i} \sum_i \oint_{\partial\Omega_i} (z - A)^{-1} f(z)\, \mathrm{d}z,$$

where the above indicates a sum of positively oriented contour integrals about the boundaries of the Ω_i This definition is known as the *holomorphic functional calculus*, and it gives an algebra homomorphism from the space of holomorphic functions on Ω to the algebra of matrices commuting with A, in the sense that when f and g are two such functions, $(fg)(A) = f(A)g(A)$. We moreover have the *spectral mapping property*: $\operatorname{Spec} f(A) = f(\operatorname{Spec} A)$.

For each simply connected $\Omega \subset \mathbb{C}$, there is a spectral projector for A corresponding to the invariant subspace spanned by all eigenvectors with eigenvalues in Ω, and one can compute this projector by integrating $(z - A)^{-1}$ about the boundary of Ω. One special case will be particularly useful: if such an Ω contains exactly one eigenvalue λ_i of A, then

$$P_i = v_i w_i^* = \frac{1}{2\pi i} \oint_{\partial\Omega} (z - A)^{-1}\, \mathrm{d}z.$$

Pseudospectrum, Eigenvector and Eigenvalue Condition Numbers

In this subsection we record some important results relating the pseudospectrum, minimum eigenvalue gap and eigenvector/eigenvalue condition numbers, defined already in Chapter 1. First, since Spec A is exactly the set of poles of the resolvent, the following elementary result follows from the level set definition in (1.5).

Lemma 2.1. *Each connected component of $\Lambda_\epsilon(A)$ contains at least one eigenvalue of A.*

Note that the eigenvalue condition numbers are exactly the norms of the spectral projectors for each eigenvalue, since

$$\kappa_i(A) = \|v_i\|\|w_i^*\| = \|P_i\|.$$

Upper bounds on $\kappa_V(A)$ are hard to come by, but we can compare it to the eigenvalue condition numbers as follows.

Lemma 2.2. *Let $A \in \mathbb{C}^{n\times n}$ be any diagonalizable matrix with distinct eigenvalues. Then*

$$\max_i \kappa_i(A) \leq \kappa_V(A) \leq \sum_i \kappa_i(A).$$

Proof. Let us first choose V diagonalizing A so that $\|V\|\|V^{-1}\| = \kappa_V(A)$. Then if $v_1, ..., v_n$ and $w_1^*, ..., w_n^*$ are the columns and rows V and V^{-1}, respectively, for each $i \in [n]$

$$\kappa_i(A) = \|v_i\|\|w_i\| \leq \|V\|\|V^{-1}\| = \kappa_V(A).$$

For the upper bound, instead normalize V so that $\|v_i\|^2 = \|w_i^*\|^2 = \kappa_i(A)$ for every i. Then

$$\kappa_V(A) \leq \|V\|\|V^{-1}\| \leq \|V\|_F\|V^{-1}\|_F = \sum_i \kappa_i(A).$$

□

By the standard comparison of ℓ^2 an ℓ^1 norms, Lemma 2.2 easily implies

$$\kappa_V(A) \leq \sqrt{n\sum_i \kappa_i^2(A)}. \tag{2.1}$$

When the A approaches a normal matrix, $\kappa_V(A)$ and every $\kappa_i(A)$ approach 1, which means that the upper bound in Lemma 2.2 is loose by a factor of n. The following alternate bound, which we have not seen in the literature, is an improvement in the regime where every $\kappa_i(A) = O(1 + 1/n)$.

Lemma 2.3. *For every diagonalizable A with distinct eigenvalues,*

$$\kappa_V(A) \leq 1 + \sum_i \kappa_i(A) - n + \sqrt{\left(\sum_i \kappa_i(A) - n\right)\left(\sum_i \kappa_i(A) - n + 2\right)}.$$

Proof. Again scale so that $\|w_i^*\| = \|v_i\| = \sqrt{\kappa_i(A)}$. This gives $\|w_i - v_i\|^2 = 2\kappa_i(A) - 2$, so that $\|V - V^{-*}\| \le \sqrt{2\sum_i(\kappa_i(A) - 1)}$, and

$$\begin{aligned}\|V\|^2 &= \|V^*V\| \\ &= \|1 + V^*(V - V^{-*})\| \\ &\le 1 + \|V\|\sqrt{2\sum_i \kappa_i(A) - 2n},\end{aligned}$$

or $\|V\| \le \frac{1}{\sqrt{2}}\left(\sqrt{\sum_i \kappa_i(A) - n} + \sqrt{\sum_i \kappa_i(A) - n + 2}\right)$. Multiplying by the corresponding bound for $\|V^{-1}\|$ gives the result. □

We have already seen in the Bauer-Fike Theorem 1.8 that $\kappa_V(A)$ can be used to control the size of the pseudospectrum. On the other hand, we can bound the eigenvalue condition numbers in terms of a certain scaling limit of the area of $\Lambda_\epsilon(A)$ as $\epsilon \longrightarrow 0$, a relationship that will be crucial in Chapter 3. The following can be extracted from [36].

Lemma 2.4. *If $A \in \mathbb{C}^{n\times n}$ has distinct eigenvalues, and $\Omega \subset \mathbb{C}$ is a measurable open set, then*

$$\sum_{\lambda_i \in \Omega} \kappa_i^2(A) = \lim_{\epsilon \longrightarrow 0} \frac{\mathrm{Leb}_{\mathbb{C}}\, \Lambda_\epsilon(A)}{\pi\epsilon^2}.$$

Proof. Since Ω is open, A has distinct eigenvalues, and $\Lambda_\epsilon(A)$ is contained in a union of disks of radius $\epsilon\kappa_V(A)$ about $\operatorname{Spec} A$, for $\epsilon \ll \frac{\mathrm{gap}(A)}{2\kappa_V(A)}$. Thus (i) each eigenvalue of A inside Ω lies in a unique connected component of $\Lambda_\epsilon(A)$ completely contained in Ω, and (ii) for any eigenvalue outside Ω, its connected component is outside as well. Now choose some $z \in \partial\Lambda_\epsilon(A)$ on the boundary of the component containing an eigenvalue λ_i. We have

$$\frac{|z - \lambda_i|}{\epsilon} \le \frac{\kappa_i(A)}{1 - \epsilon\sum_{i\neq j}\frac{\kappa_j(A)}{|z-\lambda_j|}} \le \frac{\kappa_i(A)}{1 - \epsilon\sum_{j\neq i}\frac{\kappa_j(A)}{\mathrm{gap}(A) - \epsilon\kappa_V(A)}} \le \frac{\kappa_i(A)}{1 - \epsilon\sum_{j\neq i}\frac{2\kappa_j(A)}{\mathrm{gap}(A)}},$$

along with an analogous lower bound. Since these bounds are uniform over z on the boundary of the given component, we find that the area of the component shrinks as $\pi\epsilon^2\kappa_i^2(A)$, and repeating this argument for each component in turn finishes the proof. □

Perturbation Theory

In the course of this thesis, we will repeatedly need to understand how the various aspects and attributes of a matrix change after a small perturbation. It is routine that the eigenvalues of a matrix are continuous functions of its entries (for instance by the result that the roots of a polynomial are continuous in its coefficients), and we begin this section by briefly stating some quantitative results on the first order perturbation theory of both eigenvalues and eigenvectors.

The following facts are verified carefully in [88, Theorems 1-2] and the surrounding discussion, and we refer the reader there for more detail and a thorough survey of the literature.

Assume that $A(t)$ is a smooth curve in $\mathbb{C}^{n\times n}$, passing through a diagonalizable matrix $A(0) = A = VDV^{-1}$. Then for t on some neighborhood of zero there are known to exist smooth curves $V(t)$ invertible and $D(t)$ diagonal, so that $V(0) = V$, $D(0) = D$, and $A(t) = V(t)D(t)V(t)^{-1}$. In this setup, the derivative of each eigenvalue is well-understood. Denote as usual λ_i for the eigenvalues of A, and v_i and w_i^* for the columns of V and the rows of V^{-1}; let us write as well $\dot{A}(t)$ for the derivative of A, and similarly for $\dot{\lambda}_i$ and $\dot{v}_i$. Then

$$\dot{\lambda}_i(t) = w_i^*(t)\dot{A}(t)v_i(t),$$

which implies that the magnitude of $\dot{lambda_i}(t)$ can be bounded in terms of its eigenvalue condition number:

$$|\dot{\lambda}_i(t)| \leq \kappa_i(A(t))\|\dot{A}(t)\|.$$

In fact, κ_i is exactly the maximal derivative of the eigenvalue λ_i along any smooth curve passing through A, which we can see by considering a curve with $\dot{A}(0) = v_i w_i^*$. A bound on the derivatives of the right eigenvectors is also known [88, p. 468], and will be useful to us in the sequel. It reads

$$\|\dot{v}_i(t)\| \leq \frac{\kappa_V(A(t))}{\mathrm{gap}(A(t))}\|\dot{A}(t)\|\|v_i(t)\|. \tag{2.2}$$

Such first order perturbation results are a useful starting point, but we shall ultimately require non-asymptotic counterparts which control the spectral properties of a matrix after a macroscopic perturbation. A simple result, which follows from the fact that $\sigma_{n-j+1}(A)$ is the distance from $A \in \mathbb{C}^{n\times n}$ in operator norm to the set of rank-j matrices, is that for any matrix $\widetilde{A} \in \mathbb{C}^{n\times n}$

$$|\sigma_i(A) - \sigma_i(\widetilde{A})| \leq \|A - \widetilde{A}\| \qquad \forall i \in [n].$$

As a consequence, we can easily control the pseudospectrum after a perturbation, as mentioned in the introductory material already: for any $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ and $\epsilon \geq \|A - \widetilde{A}\|$,

$$\Lambda_{\epsilon - \|A-\widetilde{A}\|}(\widetilde{A}) \subset \Lambda_\epsilon(A).$$

For eigenvalues, the Bauer-Fike Theorem and the definition of pseudospectrum imply that if $A \in \mathbb{C}^{n\times n}$ is diagonalizable and $\widetilde{A} \in \mathbb{C}^{n\times n}$ is any matrix, then

$$\mathrm{dist}(\widetilde{\lambda}_i, \mathrm{Spec}\, A) \leq \kappa_V(A)\|A - \widetilde{A}\|$$

for every $\widetilde{\lambda}_i \in \mathrm{Spec}\, \widetilde{A}$. More granular study of the eigenvalues is possible as well, such as the following theorem collating two corollaries of Theorem 1.8 (see [33, Exercise VIII.3.2]). Recall from the introduction the *condition number of the eigenproblem*, which we defined as

$$\kappa_{\mathrm{eig}}(A) \triangleq \frac{2\kappa_V(A)}{\mathrm{gap}(A)},$$

and which will appear throughout this section.

Theorem 2.5. *Let $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be diagonalizable. Then the eigenvalues of A and $\widetilde{A}$ may be matched so that*

$$|\lambda_i - \tilde{\lambda}_i| \leq (2n-1)\kappa_V(A)\|A - \widetilde{A}\| \qquad \forall i \in [n].$$

If furthermore $\|A - \widetilde{A}\| \leq \delta \leq \kappa_{\mathrm{eig}}^{-1}(A)$, then the eigenvalues of A and $\widetilde{A}$ may be matched so that

$$|\lambda_i - \tilde{\lambda}_i| \leq \kappa_V(A)\|A - \widetilde{A}\| \qquad \forall i \in [n].$$

We also have the following bound independent of $\kappa_V(A)$, [33, Theorem VIII.1.5], which will be useful in Chapter 7

Theorem 2.6. *Let $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be any matrices. Then the eigenvalues of A and $\widetilde{A}$ may be matched so that*

$$|\lambda_i - \tilde{\lambda}_i| \leq 2(\|A\| + \|\widetilde{A}\|)^{1-1/n}\|A - \widetilde{A}\|^{1/n} \qquad \forall i \in [n].$$

The contour integral formulae discussed earlier for the spectral projectors of A above allow us to easily bound their sensitivity to perturbation — a technique which will be of repeated use throughout this thesis.

Lemma 2.7. *Let $A \in \mathbb{C}^{n\times n}$ have distinct eigenvalues and spectral projectors $P_1, ..., P_n$, and assume that $\|A - \widetilde{A}\| \leq \frac{1}{2\kappa_{\mathrm{eig}}(A)}$. Then $\widetilde{A}$ has distinct eigenvalues as well, and its spectral projectors $\tilde{P}_1, ..., \tilde{P}_n$ may be matched with those of A so as to satisfy*

$$\|P_i - \tilde{P}_i\| \leq 2\kappa_V(A)\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|$$

Proof. From the Bauer-Fike theorem, if we set $\epsilon = \kappa_{\mathrm{eig}}^{-1}(A) = \frac{\mathrm{gap}(A)}{2\kappa_V(A)}$, then $\Lambda_\epsilon(A)$ has n connected components, each contained in a disk of radius $\mathrm{gap}(A)/2$ about some eigenvalue of A. The eigenvalues of $\widetilde{A}$ lie in these components and are thus distinct as well, meaning that we can compute both P_i and $\tilde{P}_i$ by integrating the resolvent around the boundary of some such disk D_i. Thus using the resolvent identity and the triangle inequality

$$\begin{aligned} \|P_i - \tilde{P}_i\| &\leq \left\| \frac{1}{2\pi i} \oint_{\partial D_i} (z - A)^{-1} - (z - \widetilde{A})^{-1}\, \mathrm{d}z \right\| \\ &\leq \frac{1}{2\pi} \oint_{\partial D_i} \|(z - A)^{-1}\|\|A - \widetilde{A}\|\|(z - \widetilde{A})^{-1}\|\, \mathrm{d}z, \end{aligned}$$

and the result follows if we use the definition of $\Lambda_\epsilon(A)$ and the singular value perturbation bounds above to control the two resolvent norms on the boundary of the disk. □

More generally, given any simple, closed, rectifiable contour encircling a connected component of $\Lambda_\epsilon(A)$ with a single eigenvalue λ_i, if $\|A - \widetilde{A}\| \leq \epsilon$ then we can control $\|P_i - \tilde{P}_i\|$ in terms of ϵ, $\|A - \widetilde{A}\|$, and the length of the contour. This observation will be useful in Chapter 5. In the course of the proof above, we have also furnished a perturbation bound for $\mathrm{gap}(A)$.

Lemma 2.8. *Let $A \in \mathbb{C}^{n\times n}$ have distinct eigenvalues, and assume that $\|A - \widetilde{A}\| \le \kappa_{\mathrm{eig}}^{-1}(A)$. Then*

$$\mathrm{gap}(\widetilde{A}) \ge \mathrm{gap}(A) - 2\kappa_V(A)\|A - \widetilde{A}\|.$$

A further corollary of Lemma 2.7 controls the sensitivity of right eigenvectors themselves to perturbation, and thus on the sensitivity of $\kappa_V(A)$ itself.

Lemma 2.9. *Let $A \in \mathbb{C}^{n\times n}$ have distinct eigenvalues, and assume that $A = VDV^{-1}$ for some matrix $V = [v_1, ..., v_n]$. If $\|A - \widetilde{A}\| < \frac{1}{2\kappa_{\mathrm{eig}}(A)}$, then $\widetilde{A} = \widetilde{V}\widetilde{D}\widetilde{V}^{-1}$ for some $\widetilde{V} = [\tilde{v}_1, ..., \tilde{v}_n]$ satisfying*

$$\|v_i - \tilde{v}_i\| \le 2n\kappa_{\mathrm{eig}}(A)\|v_i\|\|A - \widetilde{A}\| \qquad \forall i \in [n],$$

which implies $\|V - \widetilde{V}\| \le 2n^2\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|\|V\|$.

Proof. Let us write $A(t) = (1-t)A + t\widetilde{A}$; for all $t \in (0,1)$, Bauer-Fike implies that $A(t)$ has distinct eigenvalues. From the earlier discussion of first order perturbation theory, there are smooth curves $v_i(t)$ which are right eigenvectors of $A(t)$ for every t and satisfy $v_i(0) = v_i$. Let us write $\hat{v}_i(t) = \|v_i(t)\|^{-1}v_i(t)$, and set $\tilde{v}_i = \|v_i\|\hat{v}_i(1)$. Writing $\dot{\hat{v}}_i(t)$ for the derivative of $\hat{v}_i(t)$, the chain rule gives us

$$\dot{\hat{v}}_i(t) = \|v_i(t)\|^{-1}\dot{v}_i'(t) - \|v_i(t)\|^{-2}v_i^*(t)\dot{v}_i(t)\,v_i(t).$$

Since

$$\|\dot{\hat{v}}_i(t)\|^2 \le \|v_i(t)\|^{-2}\left(\|\dot{v}_i(t)\| - \frac{|v_i^*(t)\dot{v}_i(t)|^2}{\|v_i(t)\|}\right) \le \|\dot{v}_i(t)\|^2,$$

the bound (2.2) implies

$$\|\dot{\hat{v}}_i(t)\| \le \frac{\|\dot{v}_i(t)\|}{\|v_i(t)\|} \le \frac{\kappa_V(A(t))}{\mathrm{gap}(A(t))}\|\dot{A}(t)\| = \frac{\kappa_V(A(t))}{\mathrm{gap}(A(t))}\|A - \widetilde{A}\|.$$

On the other hand, $\mathrm{gap}(A(t)) \ge \mathrm{gap}(A) - 2\kappa_V(A)\|A - \widetilde{A}\| \ge \mathrm{gap}(A)/2$ and

$$\kappa_V(A(t)) \le \sum_i \kappa_i(A(t)) \le \sum_i \left(\kappa_i(A) + \|P_i - \widetilde{P}_i\|\right) \le 2n\kappa_V(A)$$

by Lemma 2.7. Thus

$$\|\hat{v}_i(t) - \hat{v}_i(0)\| \le \int_0^1 \|\dot{\hat{v}}_i(t)\|\,\mathrm{d}t \le 4n\frac{\kappa_V(A)}{\mathrm{gap}(A)}\|A - \widetilde{A}\| = 2n\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|,$$

and the promised bound follows by scaling up $\hat{v}_i(t)$ and $\hat{v}_i(0)$ accordingly. □

In the course of the proof above we have given the loose bound $\kappa_V(\widetilde{A}) \le 2n\kappa_V(A)$, but we will see that this can be boosted to a stronger result.

Lemma 2.10. *Let $A \in \mathbb{C}^{n\times n}$ have distinct eigenvalues, and $\|A - \widetilde{A}\| \leq \frac{1}{4n^2\kappa_V(A)\kappa_{\mathrm{eig}}(A)}$. Then*

$$\kappa_V(\widetilde{A}) \leq \kappa_V(A) + 8n^2\kappa_V(A)\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|.$$

Proof. Let V diagonalize A and be scaled so that $\|V\| = \|V^{-1}\| = \kappa_V(A)$. Using lemma 2.9, A may be diagonalized by $\widetilde{V}$ satisfying

$$\|V - \widetilde{V}\| \leq \|V - \widetilde{V}\|_F \leq 4n^2\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|\|V\|$$

Using stability of singular values, we then have

$$\kappa_V(\widetilde{A}) \leq \|\widetilde{V}\|\|\widetilde{V}^{-1}\| \leq \frac{\|V\| + \|V - \widetilde{V}\|}{\|V^{-1}\|^{-1} - \|V - \widetilde{V}\|} \leq \kappa_V(A)\frac{1 + 2n^2\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|}{1 - 2n^2\kappa_V(A)\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|}.$$

Since $\|A - \widetilde{A}\| \leq \frac{1}{4n^2\kappa_V(A)\kappa_{\mathrm{eig}}(A)}$, we can use convexity of the function $f(x) = \frac{1-x/\kappa_V(A)}{1-x}$ to bound it by the line interpolating between $f(0) = 1$ and $f(1/2) = 4 + 2\kappa_V^{-1}(A)$. This gives

$$\kappa_V(\widetilde{A}) \leq \kappa_V(A)\left(1 + 8n^2\kappa_V(A)\kappa_{\mathrm{eig}}(A)\|A - \widetilde{A}\|\right).$$

□

It will also be useful in Chapters 5 and 7 to understand how κ_V and gap change when passing to a submatrix.

Lemma 2.11. *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. Aoreover, 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 QR Decomposition

Every nonsingular matrix $A \in \mathbb{C}^{n\times n}$ has a unique *QR decomposition* $A = QR$, where Q is unitary, an R is upper triangular with nonnegative entries. We will write

$$[Q, R] = \mathsf{qr}(A)$$

to denote that Q and R are the promised factorizing matrices. It is easy to check that Q can be obtained by running Gram-Schmidt to orthonormalize the columns of A, left to right. In Chapters 5 and 7 we will need the following result of J. Sun on the condition number of the QR decomposition [145, Theorem 1.6].

Lemma 2.12 (Condition Number of the QR Decomposition)**.** *Let $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ with A invertible and $\|A - \widetilde{A}\|\|A^{-1}\| \le 1/2$. Then If $[Q, R] = \mathsf{qr}(A)$ and $[\widetilde{Q}, \tilde{R}] = \mathsf{qr}(\widetilde{A})$, then*

$$\|\widetilde{Q} - Q\|_F \le 4\|A^{-1}\|\|A - \widetilde{A}\|_F \quad \text{and} \quad \|\tilde{R} - R\| \le 3\|A^{-1}\|\|R\|\|A - \widetilde{A}\|.$$

Finite Precision Arithmetic

As discussed in the introductory material, the algorithms in Chapters 5 and 7 are analyzed in the computational model of *floating point arithmetic* with machine precision (sometimes called *unit roundoff*) $\mathbf{u}$. This model is axiomatized as follows [94]. Whenever our algorithm performs an operation $\square \in \{+, -, \times, \div\}$ on two complex numbers x and y, we assume

$$\mathsf{fl}(x \square y) = (x \square y)(1 + \Delta) \qquad |\Delta| \le \mathbf{u},$$

where $\mathsf{fl}(x \square y)$ denotes the outcome of the operation when performed in floating point arithmetic, and Δ is an adversarially chosen complex number. We assume that the same guarantee holds when computing the square root of a positive real number. A number of further assumptions and results regarding finite precision arithmetic — including matrix multiplication and inversion, QR decomposition, random sampling, and QR iteration — are remanded to Chapters 5 and 7.

As discussed at length in [94] and elsewhere, such a system can be implemented by allowing our machine to access complex numbers whose real and imaginary parts are of the form

$$\pm s \cdot 2^{\log 1/\mathbf{u} \pm e},$$

where the *significand* s lies in the range $0 \le s \le 1/\mathbf{u}$, and the *exponent* lies in some range $0 \le e \le e_{\max}$. This corresponds allocating $\log 1/\mathbf{u}$ *bits of precision* to store the significand and another $\log e_{\max}$ for the exponent. For point of reference, IEEE extended precision (real) arithmetic uses 64 bits of precision, plus 8 for the exponent, giving $\mathbf{u} = 2^{-64}$. The implementation of floating point arithmetic has numerous subtleties, some of which which (as is customary in the literature) we will elect to ignore. One such issue is that of *overflow* and *underflow* when we encounter a number whose exponent is too small, or too large. The tenacious reader could verify that none of the numbers encountered in the course of our algorithms are problematic in this regard.

Over the course of this thesis and depending on the context, we use a few different ways to distinguish between exact arithmetic algorithms and finite arithmetic implementations. Sometimes we will write, e.g., $\mathsf{alg}(\,)$ for the former and $\mathsf{ALG}(\,)$ for the latter. Other times we will write the outcome of a finite arithmetic calculation as $\mathsf{alg}(\,) + E$, where E is some adversarial error, or alternatively as $\widetilde{\mathsf{alg}(}\,)$.

Bibliographic Note

The proof of Lemma 2.9 is adapted from [15, Proposition 1.1]; similarly with Lemma 2.10 and [18, Lemma 6.1].

Chapter 3

Regularization by Complex Gaussian Perturbations

We will write $\boldsymbol{G}_n$ for the *$n \times n$ complex Ginibre matrix*, a random matrix whose entries are independent and distributed according to the complex Gaussian distribution of variance $1/n$; in other words each entry $\boldsymbol{G}_{i,j}$ satisfies $\mathbb{E}\boldsymbol{G}_{i,j} = 0$ and $\mathbb{E}|\boldsymbol{G}_{i,j}|^2 = 1/n$. (For the probibalistic Chapters 3-4 only, we will write all random variables in **bold face font**.) In this chapter we will study the eigenvector and eigenvalue condition numbers, minimum eigenvalue gap, and pseudospectrum of random matrices of the form

$$A + \delta \boldsymbol{G}_n,$$

where $A \in \mathbb{C}^{n\times n}$ is arbitrary and deterministic, and δ is a small parameter controlling the size of the random perturbation. Our goal is to prove Theorem 1.10, which we restate here.

Restatement of Theorem 1.10. *Let A be any $n \times n$ matrix with $\|A\| \le 1$, $\boldsymbol{G}_n$ a complex Ginibre matrix, and $\delta \le 1/2$. Then, with probability at least $1 - 10/n$,*

$$\sum_i \kappa_i^2(A + \delta \boldsymbol{G}_n) \le n^3/\delta^2$$
$$\kappa_V(A + \delta \boldsymbol{G}_n) \le n^2/\delta$$
$$\mathrm{gap}(A + \delta \boldsymbol{G}_n) \ge \delta/n^2$$

and $\Lambda_\epsilon(A + \delta \boldsymbol{G}_n)$ has n disjoint connected components for every $\epsilon \le \delta^2/2n^4$.

3.1 Shifted Singular Value Bounds

As discussed in Chapter 1, our approach in studying the aforementioned properties of $A + \delta\boldsymbol{G}_n$ will be to prove tail bounds on the singular values of the scalar shifts $z - A - \delta\boldsymbol{G}_n$, for $z \in \mathbb{C}$. The distribution of the smallest singular value of $\boldsymbol{G}_n$ itself was computed by Edelman in [70, Chapter

5] as

$$\mathbb{P}[\sigma_n(\boldsymbol{G}_n) < t] = 1 - e^{-t^2 n^2} \le \epsilon^2 n^2. \tag{3.1}$$

and the resulting bound was generalized to the remaining singular values by Szarek [146, Theorem 1.2].

Theorem 3.1. *For $\boldsymbol{G}_n$ an $n \times n$ complex Ginibre matrix and for any $t \ge 0$ it holds that*

$$\mathbb{P}\left[\sigma_j(\boldsymbol{G}_n) < t\right] \le \left(\sqrt{2e}\, t \frac{n}{n-j+1}\right)^{2(n-j+1)^2}.$$

In order to translate Edelman and Szarek's bounds to the shifted case, we will import a powerful comparison result of Śniady from the context of free probability [142]. The proof is quite beautiful, and we include a sketch for the readers edification.

Theorem 3.2 (Śniady's Comparison Theorem). *Let $A^{(1)}$ and $A^{(2)}$ be $n \times n$ complex matrices such that $\sigma_i(A^{(1)}) \le \sigma_i(A^{(2)})$ for all $1 \le i \le n$. Assume further that $\sigma_i(A^{(1)}) \ne \sigma_j(A^{(1)})$ and $\sigma_i(A^{(2)}) \ne \sigma_j(A^{(2)})$ for all $i \ne j$. Then for every $t \ge 0$, there exists a joint distribution on pairs of $n \times n$ complex matrices $(\boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)})$ such that*

(i) the marginals $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ are distributed as complex Ginibre matrices, and

(ii) almost surely $\sigma_i(A^{(1)} + t\boldsymbol{G}^{(1)}) \le \sigma_i(A^{(2)} + t\boldsymbol{G}^{(2)})$ for every i.

Sketch of proof. The key insight of the proof is that it is possible to couple the distributions of $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ through their singular values. To do so, one first derives a stochastic differential equation satisfied by the singular values $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ of a matrix Brownian motion (i.e., a matrix whose entries are independent complex Brownian motions): for all $i \in [n]$,

$$\mathrm{d}\boldsymbol{s}_i = \frac{1}{\sqrt{2n}} \mathrm{d}\boldsymbol{B}_i + \frac{\mathrm{d}t}{2\boldsymbol{s}_i}\left(1 - \frac{1}{2n} + \sum_{j \ne i} \frac{\boldsymbol{s}_i^2 + \boldsymbol{s}_j^2}{n(\boldsymbol{s}_i^2 - \boldsymbol{s}_j^2)}\right), \tag{3.2}$$

where the $\boldsymbol{B}_i$ are independent standard real Brownian motions. Next, one uses a single n-tuple of real Brownian motions $\boldsymbol{B}_1, ..., \boldsymbol{B}_n$ to drive two processes $(\boldsymbol{s}_1^{(1)}, \dots, \boldsymbol{s}_n^{(1)})$ and $(\boldsymbol{s}_1^{(2)}, \dots, \boldsymbol{s}_n^{(2)})$ according to (3.2), with initial conditions $\boldsymbol{s}_i^{(1)}(0) = \sigma_i(A^{(1)})$ and $\boldsymbol{s}_i^{(2)}(0) = \sigma_i(A^{(2)})$ for all i. (To do this rigorously, one needs existence and uniqueness of strong solutions to the above SDE; this is shown in [101] under the hypothesis $\boldsymbol{s}_i(0) \ne \boldsymbol{s}_j(0)$ for all $i \ne j$.)

Things have been arranged so that the joint distribution of $(\boldsymbol{s}_1^{(j)}, \dots, \boldsymbol{s}_n^{(j)})$ at time t^2 matches the joint distribution of the singular values of $A^{(j)} + t\boldsymbol{G}^{(j)}$ for each $j = 1, 2$. One can then sample unitaries $\boldsymbol{U}^{(j)}$ and $\boldsymbol{V}^{(j)}$ from the distribution arising from the singular value decomposition $A^{(j)} + t\boldsymbol{G}^{(j)} = \boldsymbol{U}^{(j)}\boldsymbol{D}^{(j)}(\boldsymbol{V}^{(j)})^*$, conditioned on $\boldsymbol{D}^{(j)} = \mathrm{diag}(\boldsymbol{s}_1^{(j)}, \dots, \boldsymbol{s}_n^{(j)})$. Thus each $\boldsymbol{G}^{(j)}$ is separately Ginibre-distributed. However, $A^{(1)} + t\boldsymbol{G}^{(1)}$ and $A^{(2)} + t\boldsymbol{G}^{(2)}$ are coupled through the shared randomness driving the evolution of their singular values. In particular, since the same $\boldsymbol{B}_i$ were used for both processes, from (3.2) one can verify that the n differences $\boldsymbol{s}_i^{(2)} - \boldsymbol{s}_i^{(1)}$ are C^1 in t. By taking derivatives, one can then show the desired monotonicity property: if $\boldsymbol{s}_i^{(2)} - \boldsymbol{s}_i^{(1)} \geq 0$ holds for all i at $t = 0$, it must hold for all $t \geq 0$. □

From Theorem 3.2 immediately follows a stochastic dominance result relating the singular values of pairs of matrices after a small, complex Ginibre perturbation. We will repeatedly apply this corollary in combination with Edelman and Szarek's results, in the case when $A^{(1)} = 0$ and $A^{(2)} = z - A$ for some $A \in \mathbb{C}^{n\times n}$ and $z \in \mathbb{C}$.

Corollary 3.3. *Let $A^{(1)}$ and $A^{(2)}$ be $n \times n$ complex matrices such that $\sigma_i(A^{(1)}) \leq \sigma_i(A^{(2)})$ for all $1 \leq i \leq n$. Then for any $0 \leq t, s_1, ..., s_n \in \mathbb{R}$,*

$$\mathbb{P}\left[\sigma_i(A^{(1)} + t\boldsymbol{G}_n) \leq s_i \, \forall i \in [n]\right] \geq \mathbb{P}\left[\sigma_i(A^{(2)} + t\boldsymbol{G}_n) \leq s_i \, \forall i \in [n]\right].$$

Proof. When $A^{(1)}$ and $A^{(2)}$ both have distinct singular values, let $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ be the coupled matrices promised by 3.2. Then we have

$$\begin{aligned}
\mathbb{P}\left[\sigma_i(A^{(1)} + t\boldsymbol{G}_n) \leq s_i \, \forall i \in [n]\right] &= \mathbb{P}\left[\sigma_i(A^{(1)} + t\boldsymbol{G}^{(1)}) \leq s_i \, \forall i \in [n]\right] \\
&\geq \mathbb{P}\left[\sigma_i(A^{(2)} + t\boldsymbol{G}^{(2)}) \leq s_i \, \forall i \in [n]\right] \\
&= \mathbb{P}\left[\sigma_i(A^{(2)} + t\boldsymbol{G}_n) \leq s_i \, \forall i \in [n]\right].
\end{aligned}$$

The result for general $A^{(1)}$ and $A^{(2)}$ follows if we approach each by matrices with distinct singular values. □

3.2 Eigenvector and Eigenvalue Condition Numbers

Our first regularization result concerns the eigenvalue condition numbers of the perturbed matrix $A + \delta\boldsymbol{G}_n$ corresponding to those eigenvalues of $A + \delta\boldsymbol{G}$ in any deterministic region of $\mathbb{C}$. Interestingly, we can control these condition numbers (or, rather, the sum of their squares) without any knowledge of the locations of the eigenvalues.

Theorem 3.4. *Let $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$, and denote by $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n \in \mathbb{C}$ the (random) eigenvalues of $A + \delta\boldsymbol{G}_n$). Then for every measurable open set $\Omega \subset \mathbb{C}$,*

$$\mathbb{E} \sum_{i : \boldsymbol{\lambda}_i \in \Omega} \kappa_i(A + \delta\boldsymbol{G}_n) \leq \frac{n^2}{\pi\delta^2} \operatorname{Leb}_{\mathbb{C}} \Omega.$$

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

$$\mathbb{P}[z \in \Lambda_\epsilon(A + \delta\boldsymbol{G}_n)] = \mathbb{P}[\sigma_n(z - A - \delta\boldsymbol{G}_n) < \epsilon] \leq (n\epsilon/\delta)^2, \tag{3.3}$$

by combining Corollary 3.3 and Edelman's singular value tail bound (3.1), and noting that G_n and $-G_n$ have the same distribution. Now, fix a measurable open set $\Omega \subset \mathbb{C}$. Then

$$\begin{aligned}
\mathbb{E}\ \mathrm{Leb}_{\mathbb{C}}(\Lambda_\epsilon(A+\delta \boldsymbol{G}_n)\cap\Omega) &= \mathbb{E}\int_\Omega \mathbf{1}\{z \in \Lambda_\epsilon(A+\delta \boldsymbol{G}_n)\}\, dz \\
&= \int_\Omega \mathbb{E}\,\mathbf{1}\{z \in \Lambda_\epsilon(A+\delta \boldsymbol{G}_n)\}\, dz && \text{by Fubini} \\
&\le \int_\Omega (n\epsilon/\delta)^2\, dz && \text{by (3.3)} \\
&= (n\epsilon/\delta)^2\, \mathrm{Leb}_{\mathbb{C}}(\Omega) && (3.4)
\end{aligned}$$

where the integrals are with respect to Lebesgue measure on $\mathbb{C}$.

Finally, using Lemma 2.4 and taking a limit as $\epsilon \longrightarrow 0$ yields the desired bound:

$$\begin{aligned}
\mathbb{E}\sum_{i:\lambda_i\in\Omega}\kappa_i^2(A+\delta \boldsymbol{G}_n) &= \mathbb{E}\liminf_{\epsilon\longrightarrow 0}\frac{\mathrm{Leb}_{\mathbb{C}}(\Lambda_\epsilon(A+\delta \boldsymbol{G}_n)\cap\Omega)}{\pi\epsilon^2} && \text{by Lemma 2.4} \\
&\le \liminf_{\epsilon\longrightarrow 0}\mathbb{E}\,\frac{\mathrm{Leb}_{\mathbb{C}}(\Lambda_\epsilon(A+\delta \boldsymbol{G}_n)\cap\Omega)}{\pi\epsilon^2} && \text{by Fatou's Lemma} \\
&\le \frac{n^2\,\mathrm{Leb}_{\mathbb{C}}(\Omega)}{\pi\delta^2} && \text{by (3.4).}
\end{aligned}$$

□

We can convert Theorem 3.4 into a tail bound for $\kappa_V(A+\delta \boldsymbol{G}_n)$ by using the upper bound $\kappa_V \le \sqrt{n\sum_i \kappa_i^2}$ from (2.1). Doing do requires that we control every eigenvalue condition number, whereas Theorem 3.4 addresses only those κ_i's lying in some deterministic set. Fortunately, it is the case with high probability that $\|\boldsymbol{G}_n\| \le 4$ (say), in which case every eigenvalue of $A+\delta \boldsymbol{G}_n$ is contained in a disk of radius $\|A\| + 4\delta$, and we can proceed by truncating to this set. To make this precise, recall the following coarse (but simple) tail bound for $\|\boldsymbol{G}_n\|$.

Lemma 3.5. *For a complex Ginibre matrix G_n,*

$$\mathbb{P}[\sigma_1(\boldsymbol{G}_n) > 2\sqrt{2} + t] \le 2\exp(-nt^2).$$

Proof. We can write $\boldsymbol{G}_n = \frac{1}{\sqrt{2}}(\boldsymbol{X} + i\boldsymbol{Y})$ where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent with i.i.d. *real* $\mathcal{N}(0, 1/n)$ entries. It is well-known that

$$\mathbb{E}\,\sigma_1(\boldsymbol{G}_n) \le \sqrt{2}\,\mathbb{E}\,\|\boldsymbol{X}\| \le 2\sqrt{2};$$

see e.g. the argument by way of Slepian's inequality in [52, Theorem II.11]. Lipschitz concentration of functions of real Gaussian random variables yields the result. □

Theorem 3.6. *Let $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$. Then*

$$\mathbb{P}\left[\kappa_V(A+\delta \boldsymbol{G}_n) \ge t\right] \le \frac{n^3(\|A\| + 4\delta)^2}{\delta^2 t^2} + 2e^{-n}.$$

Proof. Again write $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n$ for the random eigenvalues of $A + \delta \boldsymbol{G}_n$. The event $\|\boldsymbol{G}_n\| \le 4$ ensures that $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n \in \Omega \triangleq \mathbb{D}(0, \|A\| + 4\delta)$. Using (2.1), Lemma 3.5, and Markov's inequality, we have

$$\begin{aligned}\mathbb{P}\left[\kappa_V(A + \delta \boldsymbol{G}_n) \ge t\right] &\le \mathbb{P}\left[\sum_i \kappa_i^2(A + \delta \boldsymbol{G}_n) \ge \frac{t^2}{n}, \|\boldsymbol{G}_n\| \le 4\right] + \mathbb{P}\left[\|\boldsymbol{G}_n\| > 4\right] \\ &\le \mathbb{P}\left[\sum_{i : \boldsymbol{\lambda}_i \in \Omega} \kappa_i^2(A + \delta \boldsymbol{G}_n) \ge \frac{t^2}{n}\right] + 2e^{-n} \\ &\le \frac{n^3(\|A\| + 4\delta)^2}{\delta^2 t^2} + 2e^{-n}\end{aligned}$$

as promised. Some improvement may be possible by instead considering the event $\|\boldsymbol{G}_n\| \ge u$ for a different $u > 2\sqrt{2}$. □

We pause to show that the dimension-dependence in Theorem 3.4 cannot be improved.

Proposition 3.7. *There exists $c > 0$ such that for all n,*

$$\mathbb{E} \sum_{i \in [n]} \kappa_i^2\left(\boldsymbol{G}_n\right) \ge cn^2.$$

Proof. Bourgade and Dubach [36, Theorem 1.1, Equation 1.8] show that eigenvalue condition numbers in the bulk of the spectrum of complex Ginibre matrices are of order $\sqrt{n}$. Precisely, if $\boldsymbol{G}_n$ has eigenvalues $\boldsymbol{\lambda}_1, \ldots \boldsymbol{\lambda}_n$, then for any $r < 1$,

$$\lim_{n \to \infty} \frac{\mathbb{E}[\kappa_i^2(\boldsymbol{G}_n) | \boldsymbol{\lambda}_i = z]}{n} = 1 - |z|^2$$

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

$$\lim_{n \to \infty} \frac{\mathbb{E}\left|\operatorname{Spec} \boldsymbol{G}_n \cap \mathbb{D}(0, r)\right|}{n} = \frac{\operatorname{Leb}_{\mathbb{C}} \mathbb{D}(0, r)}{\operatorname{Leb}_{\mathbb{C}} \mathbb{D}(0, 1)} = r^2 m$$

meaning that

$$\liminf_{n \to \infty} \frac{\mathbb{E} \sum_{i \in [n]} \kappa_i^2(\boldsymbol{G}_n)}{n^2} \ge r^2(1 - r^2) > 0.$$

□

3.3 Davies' Conjecture

In this section we will use Theorem 3.4 to study the following question posed by E. B. Davies in [53]:

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

Our main theorem is as follows.

Theorem 3.8. *Suppose $A \in \mathbb{C}^{n\times n}$ and $\delta \in (0,1)$. Then there is a matrix $\widetilde{A} \in \mathbb{C}^{n\times n}$ such that $\|A - \widetilde{A}\| \le \delta\|A\|$ and*

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

In other words, every matrix is at most inverse polynomially close to a matrix whose eigenvectors have condition number at most polynomial in the dimension. The previously best known general bound in such a result was [53, Theorem 3.8]:

$$\kappa_V(\widetilde{A}) \le \left(\frac{n}{\delta}\right)^{(n-1)/2}, \tag{3.5}$$

so Theorem 3.8 constitutes an exponential improvement in the dependence on both δ and n. We show in Proposition 3.10 that the $1/\delta$-dependence in Theorem 3.8 cannot be improved beyond $1/\delta^{1-1/n}$, so our bound is essentially optimal in δ for large n.

Theorem 3.8 implies a positive resolution to a conjecture of Davies [53].

Conjecture 3.9. *For every positive integer n there is a constant c_n such that for every $A \in \mathbb{C}^{n\times n}$ with $\|A\| \le 1$ and any $\mathbf{u} \in (0,1)$:*

$$\inf_{\widetilde{A}\in\mathbb{C}^{n\times n}} \left(\kappa_V(\widetilde{A})\mathbf{u} + \|\widetilde{A}\|\right) \le c_n\sqrt{\mathbf{u}}. \tag{3.6}$$

Proof of Conjecture 3.9. Given $\epsilon > 0$, set $\delta = d_n\sqrt{\mathbf{u}}$ for some $d_n > 0$ and apply Theorem 3.8. This yields $c_n = 4n^{3/2} + 4n^{3/2}/d_n + d_n$. This is minimized at $d_n = 2n^{3/4}$, which yields $c_n = 4n^{3/2} + 4n^{3/4} \le 8n^{3/2}$. □

The phrasing of Conjecture 3.9 is motivated by a particular application in numerical analysis. Suppose one wants to evaluate analytic functions $f(A)$ of a given matrix A, which may be non-normal. If A is diagonalizable, one can use the formula $f(A) = Vf(D)V^{-1}$, where $f(D)$ means the function is applied to the scalar diagonal entries of D. However, this may be numerically infeasible if $\kappa_V(A)$ is very large: if all computations are carried to precision $\mathbf{u}$, the result may be off by an error of $\kappa_V(A)\mathbf{u}$. Davies' idea was to replace A by a perturbation $\widetilde{A}$ with a much smaller $\kappa_V(\widetilde{A})$, and compute $f(\widetilde{A})$ instead. In [53, Theorem 2.4], he showed that the net error incurred by this scheme for a given $\mathbf{u} > 0$ and sufficiently regular f is controlled by:

$$\kappa_V(\widetilde{A})\mathbf{u} + \|A - \widetilde{A}\|,$$

which is the quantity appearing in (3.6). The key desirable feature of (3.6) is the dimension-independent fractional power of $\mathbf{u}$ on the right-hand side, which shows that the total error scales slowly.

Davies proved his conjecture in the special case of upper triangular Toeplitz matrices, in dimension $n = 3$ with the constant $c_n = 2$, as well as in the general case with the weaker dimension-dependent and nonconstructive bound $(n+1)\mathbf{u}^{2/(n+1)}$. This last result corresponds to (3.5) above. He also speculated that a *random* regularizing perturbation suffices to prove Conjecture 3.9, and presented empirical evidence to that effect. Our proof of Theorem 3.8 below indeed follows this strategy, by way of Theorem 3.4.

Proof of Theorem 3.8. We proceed similarly to the proof of Theorem 3.6. Let $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n$ be the eigenvalues of the random matrix $A + \delta \boldsymbol{G}_n$, and $t > 2\sqrt{2}$ and $s > 1$ be parameters to be optimized later. Davies' original bound (3.5) implies our bound for $n \leq 3$, so assume $n \geq 4$. Then Lemma 3.5 tells us that

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

Letting $\Omega = \mathbb{D}(0, \|A\| + t\delta)$, we have

$$\mathbb{P}\left[\sum_{i:\boldsymbol{\lambda}_i \in \Omega} \kappa_i^2(A + \delta \boldsymbol{G}_n) \neq \sum_i \kappa_i^2(A + \delta \boldsymbol{G}_n)\right] \leq \mathbb{P}[\|\delta \boldsymbol{G}_n\| \geq t\delta] \leq 2\exp\left(-4(t - 2\sqrt{2})^2\right). \tag{3.7}$$

On the other hand, by Theorem 3.4 applied to Ω and Markov,

$$\mathbb{P}\left[\sum_{i:\boldsymbol{\lambda}_i \in \Omega} \kappa_i^2(A + \delta \boldsymbol{G}_n) \geq s\frac{n^2 \operatorname{Leb}_{\mathbb{C}} \Omega}{\delta^2 \pi}\right] \leq \frac{1}{s}. \tag{3.8}$$

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

$$2\exp\left(-4(t - 2\sqrt{2})^2\right) + \frac{1}{s} < 1 \tag{3.9}$$

then with nonzero probability over $\boldsymbol{G}_n$, neither the events (3.7), (3.8) occurs, and thus there is some matrix $\widetilde{A}$ with $\|A - \widetilde{A}\| \leq t\delta$ and eigenvalues $\tilde{\lambda}_1, \dots, \tilde{\lambda}_n$, for which

$$\sum_{i=1}^n \kappa_i^2(\widetilde{A}) = \sum_{i:\tilde{\lambda}_i \in \Omega} \kappa_i^2(\widetilde{A}) \leq s\frac{n^2 \operatorname{Leb}_{\mathbb{C}} \Omega}{\pi \delta^2}.$$

Taking a square root and applying (2.1), we have

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

Because $\|A - \widetilde{A}\| \leq t\delta$ and not δ, replacing δ by δ/t yields the bound

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

To get the best bound, we must minimize $t\sqrt{s}$ subject to the constraints (3.9), $t > 2\sqrt{2}$ and $s > 1$. Solving for s this becomes a univariate optimization problem, and one can check numerically that the optimum is achieved at $t \approx 3.7487$ and $t\sqrt{s} \approx 3.8822 < 4$, as advertised. □

We will close out this section by showing that 3.8 has essentially the optimal dependence on δ, at least when n is large. The example which requires this dependence is simply a Jordan block J_n, for which Davies [53] established the upper bound $\kappa_V(J_n + \delta E) \le 2/\delta^{1-1/n}$, for some E with $||E|| < 1$.

Proposition 3.10. *Fix $n > 0$ and let $J_n \in \mathbb{C}^{n\times n}$ be the upper triangular Jordan block with ones on the superdiagonal and zeros everywhere else. Then there exist $c_n > 0$ and $\delta_n > 0$ such that for all $E \in \mathbb{C}^{n\times n}$ with $\|E\| \le 1$ and all $\delta < \delta_n$, we have*

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

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

$$|\lambda|^n \le \|(J_n + \delta E)^n\| \le (1+\delta)^n - 1 = O(\delta) \tag{3.10}$$

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

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

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

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

Let $0 \le k \le n$. Then $\|(J_n + \delta E)^k v\| = |\lambda|^k$, and analogously to (3.10),

$$\|(J_n + \delta E)^k - J_n^k\| \le (1+\delta)^k - 1 = O(\delta).$$

Since J_n acts on the left as a left shift,

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

where the final line follows from (3.10). Similarly,

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

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

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

□

3.4 Minimum Eigenvalue Gap

We turn now to the minimum eigenvalue gap of $A + \delta \boldsymbol{G}_n$. As discussed in Chapter 1, our initial strategy for controlling this quantity will be to first study the probability that $A + \delta \boldsymbol{G}_n$ has two eigenvalues in an small disk $\mathbb{D}(z, r) \subset \mathbb{C}$, which can in turn be controlled by the two smallest singular values of $z - (A + \delta \boldsymbol{G}_n)$. The necessary connection between eigenvalues and singular values comes from the classical *log-majorization property*; see, for example, [97, Theorem 3.3.4].

Lemma 3.11. *For any complex $n \times n$ matrix with eigenvalues labelled $|\lambda_n| \le \cdots \le |\lambda_1|$ and singular values $\sigma_n \le \cdots \le \sigma_1$, and any $k \in [n]$,*

$$\sigma_n \cdots \sigma_{n-k+1} \le |\lambda_n| \cdots |\lambda_{n-k+1}|.$$

As an immediate corollary of Lemma 3.11, if $A + \delta \boldsymbol{G}_n$ has two eigenvalues in $\mathbb{D}(z, r)$ then $\sigma_n(z - A - \delta \boldsymbol{G}_n)\sigma_{n-1}(z - A - \delta \boldsymbol{G}_n) \le r^2$. As in the preceding material, control on this event comes via a tail bound on the product of $\sigma_n(\boldsymbol{G}_n)\sigma_{n-1}(\boldsymbol{G}_n)$ in the centered case; the following bound is simple, but may not be optimal.

Lemma 3.12. *The complex Ginibre matrix $\boldsymbol{G}_n$ satisfies*

$$\mathbb{P}\left[\sigma_n(\boldsymbol{G}_n)\sigma_{n-1}(\boldsymbol{G}_n) \le r^2\right] \le \frac{4e^{4/5}}{2^{12/5}}(nr)^{16/5} \le 2.2 \cdot (nr)^{3.2}.$$

Proof. Using Edelman's bound (3.1) and Szarek's Theorem 3.1, for any $t > 0$ we have

$$\begin{aligned} \mathbb{P}\left[\sigma_n(\boldsymbol{G}_n)\sigma_{n-1}(\boldsymbol{G}_n) \le r^2\right] &\le \mathbb{P}\left[\sigma_n(\boldsymbol{G}_n) \le tr\right] + \mathbb{P}\left[\sigma_{n-1}(\boldsymbol{G}_n) \le r/t\delta\right] \\ &\le (trn)^2 + (e/2)^4 (r/t)^8. \end{aligned}$$

Instantiating the optimal $t = (e^2/2)^{1/5}(nr)^{3/5}$ and bounding the resulting constant gives the result. □

We now combine Śniady and Szarek's results with Lemma 3.12 to bound the the probability that $A + \delta \boldsymbol{G}_n$ has two eigenvaues in a disk.

Corollary 3.13. *Let $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$. Then for any $z \in \mathbb{C}$ and $r > 0$,*

$$\mathbb{P}\left[|\operatorname{Spec}(A + \delta \boldsymbol{G}_n) \cap \mathbb{D}(z, r)| \geq 2\right] \leq 2.2(nr/\delta)^{3.2}$$

Finally, we can state and prove our main result.

Theorem 3.14. *For any $A \in \mathbb{C}^{n\times n}$, $\delta > 0$, and $t \leq 1$,*

$$\mathbb{P}[\operatorname{gap}(A + \delta \boldsymbol{G}_n) \leq t] \leq 200(\|A\| + 4\delta)^2 t^{1.2}(n/\delta)^{3/2}.$$

Proof. We proceed in the same spirit as the proof of Theorem 3.6, by truncating to the event that $\|\boldsymbol{G}_n\| \leq 4$. On this event, $\operatorname{Spec}(A + \delta \boldsymbol{G}_n) \subset \mathbb{D}(0, \|A\| + 4\delta)$, and it is standard that this large disk by smaller ones of radius t, centered at fewer than

$$\left((\|A\| + 4\delta)\frac{2+t}{t}\right)^2$$

points. If $\operatorname{gap}(A + \delta \boldsymbol{G}_n) \leq t$, then there must be two eigenvalues in the disk of radius $2t$ centered at some such point. Using Lemma 3.13 to bound the probability that this occurs at each point and taking a union bound, we find that

$$\mathbb{P}\left[\operatorname{gap}(A + \delta \boldsymbol{G}_n) \leq t\right] \leq \left((\|A\| + 4\delta)\frac{2+t}{t}\right)^2 \cdot 2.2(2nt/\delta)^{3.2}.$$

Using $t \leq 1$ to upper bound the constant finishes the proof. □

Improving the Bound

Theorem 3.14 may be improved using an alternate and essentially different technique, via an auxiliary result from [5]. We begin by recalling some notation from that paper. 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 $v \in \mathbb{P}\mathbb{C}^n$, define $A_{\lambda,v} : v^\perp \longrightarrow v^\perp$ by

$$A_{\lambda,v} \triangleq P_v^\perp \circ (A - \lambda)\|_{v^\perp}$$

where $v^\perp = \{x \in \mathbb{C}^n \mid x^* v = 0\}$ and $P_{v^\perp} : \mathbb{C}^n \longrightarrow v^\perp$ denotes the orthogonal projection. With this in hand, [5] 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

$$\mu(A, \lambda, v) \triangleq \begin{cases} \|A\|_F \|A_{\lambda,v}^{-1}\| & \text{if } A_{\lambda,v} \text{ is invertible,} \\ \infty & \text{otherwise.} \end{cases}$$

They similarly define the *mean square condition number* of a matrix as

$$\mu_{F,\mathrm{av}}(A) \triangleq \left(\frac{1}{n}\sum_{j=1}^{n} \|A\|_F^2 \|A_{\lambda_j,v_j}^{-1}\|_F^2\right)^{\frac{1}{2}},$$

where (λ_j, v_j) are the eigenpairs of A. In particular, note that $\mu_{F,\mathrm{av}}(A) < \infty$ only when A has simple eigenvalues, and therefore $\mu_{F,\mathrm{av}}(A) < \infty$ implies that A is diagonalizable.

If $\lambda_1, \dots, \lambda_n$ are the eigenvalues of a matrix $A \in \mathbb{C}^{n\times n}$, we will denote

$$\mathrm{gap}_i(A) \triangleq \min_{j\neq i} |\lambda_i - \lambda_j|,$$

and these quantities may be bounded in terms of the condition number of the corresponding triple.

Lemma 3.15. *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 $i = 1, \dots, n$ it holds that*

$$\frac{\mu(A, \lambda_i, v_i)}{\|A\|_F} \geq \frac{1}{\mathrm{gap}_i(A)}.$$

Proof. First we show that $\mathrm{Spec}\, A_{\lambda_i, v_i} = \mathrm{Spec}(A - \lambda_i) \setminus \{0\}$. 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

$$\begin{aligned}\|A_{\lambda_i, v_i}^{-1}\| &\geq \sup_{\lambda \in \Lambda(A_{\lambda_i, v_i})} \frac{1}{|\lambda|} \\ &= \frac{1}{\mathrm{gap}_i(A)}.\end{aligned}$$

The claim then follows from the definition of $\mu(A, \lambda_i, v_i)$. □

The key device needed to bound $\mathrm{gap}(A + \delta \boldsymbol{G}_n)$ is [5, Theorem 2.14]:

Theorem 3.16. *For any $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$, we have*

$$\mathbb{E}\left[\frac{\mu_{F,\mathrm{av}}(A + \delta \boldsymbol{G}_n)^2}{\|A + \delta \boldsymbol{G}_n\|_F^2}\right] \leq \frac{n^2}{\delta^2}.$$

Theorem in hand, we can furnish the promised improvement to Theorem 3.14.

Proposition 3.17. *Let $A \in \mathbb{C}^{n\times n}$ be an arbitrary matrix and let $\boldsymbol{G}_n$ be a complex Ginibre matrix. Then for any $t, \delta > 0$*

$$\mathbb{P}[\mathrm{gap}(A + \delta \boldsymbol{G}_n) < t] \leq n^3 (t/\delta)^2.$$

Thus, $\mathrm{gap}(A + \delta \boldsymbol{G}_n) = O(\delta / n^{3/2})$ with probability bounded away from zero.

Proof. Using Lemma 3.15 and again writing $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ and $\boldsymbol{v}_1, ..., \boldsymbol{v}_n$ for the eigenalues and right eigenvectors of $A + \delta \boldsymbol{G}_n$, we find

$$\frac{1}{\operatorname{gap}(A + \delta \boldsymbol{G}_n)^2} = \max_i \frac{1}{\operatorname{gap}_i(A + \delta \boldsymbol{G}_n)^2} \leq \max_i \frac{\mu(A + \delta \boldsymbol{G}_n, \boldsymbol{\lambda}_i, \boldsymbol{v}_i)^2}{\|A + \delta \boldsymbol{G}_n\|_F^2} \leq n \frac{\mu_{F,\mathrm{av}}(A + \delta \boldsymbol{G}_n)^2}{\|A + \delta \boldsymbol{G}_n\|_F^2}.$$

Combining this with Theorem 3.16 we can bound

$$\mathbb{E}\left[\frac{1}{\operatorname{gap}(A + \delta \boldsymbol{G}_n)^2}\right] \leq \frac{n^3}{\delta^2}$$

and conclude the proof by applying Markov's inequality. □

3.5 Discussion

Our main regularization result, Theorem 1.10, quickly follows from Theorem 3.6 and Proposition 3.17. After pausing to prove it, we will briefly situate thee techniques and results of this chapter within the broader random matrix theory literature.

Proof of Theorem 1.7. Instantiating the argument of Theorem 3.6 with $t = n^2/\delta$, Proposition 3.17 with $t = \delta/n^2$, and using $\|A\| \leq 1$ an $\delta \leq 1/2$, we find

$$\mathbb{P}\left[\sum_i \kappa_i^2(A + \delta \boldsymbol{G}_n) \geq n^3/\delta^2,\ \operatorname{gap}(A + \boldsymbol{G}_n) \leq \delta/n^2,\ \|\boldsymbol{G}_n\| \leq 4\right] \leq 9/n + 1/n \leq 10/n.$$

On this event, we clearly have $\kappa_V(A + \delta \boldsymbol{G}_n) \leq n^2/\delta$, and moreover by Bauer-Fike, $\Lambda_\epsilon(A + \delta \boldsymbol{G}_n)$ is contained in disks of radius $n^2\epsilon/\delta$ about the eigenvalues, and thus has disjoint connected components for every $\epsilon \leq \delta^2/2n^4$. □

Eigenvector and Eigenvalue Condition Numbers of Random Matrices. There have been numerous studies of the eigenvalue condition numbers $\kappa(\lambda_i)^2$, sometimes called eigenvector *overlaps* in the random matrix theory and mathematical physics literature, for non-Hermitian random matrix models of type $A + \delta G_n$. In the centered case $A = 0$ and $\delta = 1$ of a standard complex Ginibre matrix, the seminal work of Chalker and Mehlig [46] calculated the large-n limit of the conditional expectations

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

whenever $|z| < 1$. Recent works by Bourgade and Dubach [36] and Fyodorov [80] improved on this substantially by giving exact nonasymptotic formulas for the distribution of $\kappa(\lambda)^2$ conditional on the location of the eigenvalue λ, as well as concise descriptions of the scaling limits for these formulas. The paper [32] proved (in the more general setup of invariant ensembles) that the angles between the right eigenvectors $(v_i^* v_j)/\|v_i\|\|v_j\|$ have subgaussian tails, which has some bearing on κ_V (for instance, a small angle between unit eigenvectors causes $\|V^{-1}\|$ and therefore κ_V to blow up.)

In the non-centered case, Davies and Hager [55] showed that if A is a Jordan block and $\delta = n^{-\alpha}$ for some appropriate α, then almost all of the eigenvalues of $A + \delta G_n$ lie near a circle of radius $\delta^{1/n}$ with probability $1 - o_n(1)$. Basak, Paquette, and Zeitouni [20, 21] showed that for a sequence of banded Toeplitz matrices A_n with a finite symbol, the spectral measures of $A_n + n^{-\alpha} G_n$ converge weakly in probability, as $n \longrightarrow \infty$, to a predictable density determined by the symbol. Both of the above results were recently and substantially improved by Sjöstrand and Vogel [138, 139] who proved that for any Toeplitz A, almost all of the eigenvalues of $A + n^{-\alpha} G_n$ are close to the symbol curve of A with exponentially good probability in n. Note that none of the results mentioned in this paragraph explicitly discuss the $\kappa(\lambda_i)$; however, they do deal qualitatively with related phenomena surrounding spectral instability of non-Hermitian matrices.

The idea of managing spectral instability by adding a random perturbation can be traced back to the influential papers of Haagerup and Larsen [92] and Śniady [142] (see also [91, 76]), who used it to study convergence of the eigenvalues of certain non-Hermitian random matrices to a limiting Brown measure, in the context of free probability theory.

There are three notable differences between Theorems 3.4 and 3.6 and the results mentioned above. Our result is much coarser, and only guarantees an upper bound on the $\mathbb{E}\,\kappa(\lambda_i)^2$, rather than a precise description of any distribution, limiting or not; applies to any $A \in \mathbb{C}^{n\times n}$ and $\delta \in (0, 1)$; and is completely nonasymptotic in n.

Minimum Eigenvalue Gap of Random Matrices The minimum eigenvalue gap of random matrices has been studied in the case of Hermitian and unitary matrices, beginning with the work of Vinson [154], who proved an $\Omega(n^{-4/3})$ lower bound on this gap in the case of the Gaussian Unitary Ensemble (GUE) and the Circular Unitary Ensemble (CUE). Bourgade and Ben Arous [7] derived exact limiting formulas for the distributions of all the gaps for the same ensembles. Nguyen, Tao, and Vu [121] obtained non-asymptotic inverse polynomial bounds for a large class of non-integrable Hermitian models with i.i.d. entries (including Bernoulli matrices). In a different direction, Aizenman et al. proved an inverse-polynomial bound [2] in the case of an arbitrary Hermitian matrix plus a GUE matrix or a Gaussian Orthogonal Ensemble (GOE) matrix, which may be viewed as a smoothed analysis of the minimum gap. Theorem 3.14 and Proposition 3.17 may be viewed as non-Hermitian analogues of this last result.

In the non-Hermitian case, Ge [81] obtained an inverse polynomial bound for i.i.d. matrices with real entries satisfying some mild moment conditions, and [137] proved an inverse polynomial lower bound for the complex Ginibre ensemble. Theorem 3.14 thus generalizes these results to non-centered complex Gaussian matrices.

Bibliographic Note

This chapter interleaves material [19] and [15], and much of the material appears verbatim here as it did there. Sections 3.1-3.2 draw from the presentation in both [19, 15], and Section 3.3 contains the bulk of the former's main text. Section 3.4 is new, but follows the approach in [16]. The alternative gap bounds in 3.4 are drawn from Appendix D of [15]. Finally, Section 3.5 collates the discussion of related work in [19] and [15].

Chapter 4

Regularization by Real Perturbations

In Chapter 3, we showed that entrywise complex Gaussian perturbation of any (real or complex) matrix tames the eigenvector and eigenvalue condition numbers, as well as the eigenvalue gap. Here, we show an analogous result for entrywise *real* Gaussian perturbations of arbitrary *real* matrices — and moreover extend to the case of arbitrary real random perturbation matrix with independent, absolutely continuous entries. Beyond their theoretical value within the random matrix literature, these results may be employed in as preconditioners for algorithms that solve the eigenproblem on *real* matrices, where the complex regularization approach of Chapter 3 may be undesirable. However, the provable algorithms for the eigenproblem discussed in Chapters 5-7 are stated in terms of arbitrary complex matrices, and will not use the results in the present chapter — as such, the reader primarily interested in the forthcoming eigenvalue algorithms may skip to Chapter 5 and return to the present one at her leisure.

4.1 Introduction

Throughout this chapter, we will write $\boldsymbol{H}_n$ to denote a *normalized real Ginibre matrix*; in other words, the entries of $\boldsymbol{H}_n$ are independent real random variables, each distributed as $\mathcal{N}_{\mathbb{R}}(0, 1/n)$. More generally, we will write $\boldsymbol{M}_n$ for an $n \times n$ real random matrix satisfying the following assumption:

Assumption 4.1. The matrix $\boldsymbol{M}_n$ has independent entries, each with density on $\mathbb{R}$ bounded almost everywhere by $\sqrt{n}K > 0$. (Equivalently, $\boldsymbol{M}_n = n^{-1/2}\widehat{\boldsymbol{M}}_n$ where $\widehat{\boldsymbol{M}}_n$ has independent real entries with density bounded by K.)

Of course, $\boldsymbol{H}_n$ satisfies Assumption 4.1 with $K = 1/\sqrt{2\pi}$. We do not require that $\boldsymbol{M}_n$ have mean zero, nor will we make any explicit moment assumptions on its entries. Instead, our results will often be stated in terms of the L_p norm of its operator norm, which we denote by

$$B_{\boldsymbol{M}_n,p} \triangleq \mathbb{E}\left[\|\boldsymbol{M}_n\|^p\right]^{1/p}. \tag{4.1}$$

Remark 4.2. The main result in [166] implies that if the entries of $\boldsymbol{M}_n$ have finite fourth moment then $\mathbb{E}[\|\boldsymbol{M}_n\|] = O(1)$, and in particular in the Gaussian case the numbers $B_{\boldsymbol{H}_n,p}$ are constants (see Lemma 4.14 below). On the other hand, the complementary result in [12] shows that without the fourth moment assumption $\lim_n \mathbb{E}[\|\boldsymbol{M}_n\|] = \infty$. It is important to remark that even in the latter case we obtain meaningful results, since a finite second moment assumption of the entries of $\boldsymbol{M}_n$ is enough to obtain a bound on $B_{\boldsymbol{M}_n,p}$ with polynomial dependence on n.

In the above notation, we will be interested in the minimum gap and eigenvector/eigenvalue condition numbers of random matrices with the form

$$A + \delta \boldsymbol{H}_n \qquad \text{or} \qquad A + \delta M_n,$$

where $A \in \mathbb{R}^{n\times n}$ is deterministic, and $\delta > 0$ is a fixed small parameter. As in Chapter 3, we will study the eigenvalue condition numbers by way of the ϵ-pseudospectrum and the limiting area formula in Lemma 2.4. However, the case of real matrices with real random perturbations presents one additional complication: the possibility of purely real eigenvalues, whose behavior can be substantively different from their complex counterparts. For instance, Real Ginibre matrices alone are known to have $\Theta(\sqrt{n})$ real eigenvalues on average [71], with eigenvalue condition numbers satisfying

$$\mathbb{E} \sum_{i\,:\,\lambda_i \in \mathbb{R}} \kappa_i^2(\boldsymbol{H}_n) = \infty.$$

(See the discussion following [80, Remark 2.2].) We will address this issue via a complementary result to Lemma 2.4, whose proof is analogous.

Lemma 4.3 (Limiting Length of Pseudospectrum on Real Line)**.** *Let $A \in \mathbb{R}^{n\times n}$ have n distinct eigenvalues $\lambda_1, ..., \lambda_n$. Let $\mathrm{Leb}_{\mathbb{R}}$ denote the Lebesgue measure on $\mathbb{R}$, and let $\Omega \subset \mathbb{R}$ be an open, measurable set. Then*

$$\sum_{i\,:\,\lambda_i \in \Omega} \kappa_i(A) \le \liminf_{\epsilon \longrightarrow 0} \frac{\mathrm{Leb}_{\mathbb{R}}\left(\Lambda_\epsilon(A) \cap \Omega\right)}{2\epsilon}.$$

Lemmas 2.4 and 4.3 in hand, we can control eigenvalue condition numbers of $A + \delta\boldsymbol{H}_n$ with upper bounds on the probabilities

$$\mathbb{P}[z \in \Lambda_\epsilon(A + \delta \boldsymbol{H}_n)] = \mathbb{P}[\sigma_n(z - A - \delta \boldsymbol{H}_n) \le \epsilon] \tag{4.2}$$

(and similarly for M_n), *provided that one obtains the correct exponents* ϵ^1 for $z \in \mathbb{R}$ and ϵ^2 for $z \in \mathbb{C}\backslash\mathbb{R}$. The same singular value tail bounds will also allow us to control the minimum eigenvalue gap, using the approach in Section 3.4.

The pursuit of tail bounds with this sharp ϵ-dependence will be a main technical theme of this chapter, in contrast to much of the rest of random matrix theory where the emphasis is instead on obtaining sharp dependence on n. Our main probabilistic results below show that the probability in (4.2) is $O(\epsilon)$ for $z \in \mathbb{R}$ and $O(\epsilon^2/|\Im(z)|)$ for $z \notin \mathbb{R}$, which is good enough to take the limit as

$\epsilon \longrightarrow 0$ after establishing that there are unlikely to be eigenvalues of $A + \delta \boldsymbol{H}_n$ of $A = \delta M_n$ near the real line but not on it. Because the nonreal eigenvalues of real matrices come in complex conjugate pairs, this property charmingly follows from a lower bound on the minimum eigenvalue gap.

Let us now state the shifted singular value bounds precisely. Here (and throughout the paper) we will give separate statements for general matrices satisfying Assumption 4.1 vs. for real Ginibre perturbations, as in the latter case we are frequently able to obtain improvements by exploiting specific properties of Gaussians.

Theorem 4.4 (Singular Values of $\boldsymbol{M}_n$)**.** *Let $\boldsymbol{M}_n \in \mathbb{R}^{n\times n}$ be a random matrix satisfying Assumption 4.1 with parameter $K > 0$. Then*

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

Note that Theorem 4.4 includes as a special case matrices of type $z - A - \delta\boldsymbol{M}_n$ for real z and A, as such matrices themselves satisfy Assumption 1.

Theorem 4.4G (Singular Values of Real Shifts: Gaussian)**.** *Let $z \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{H}_n$ be a normalized Ginibre matrix. For every $\delta > 0$,*

$$\mathbb{P}[\sigma_{n-k+1}(z - A - \delta\boldsymbol{H}_n) \le \epsilon] \le \left(\frac{\sqrt{2}en\epsilon}{k\delta}\right)^{k^2}.$$

In the case $k = 1$, one has a better constant:

$$\mathbb{P}[\sigma_n(z - A - \delta\boldsymbol{H}_n) \le \epsilon] \le \frac{n\epsilon}{\delta}.$$

The key improvement we obtain the case of nonreal complex z is an extra factor of 2 in the exponent.

Theorem 4.5 (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 $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. For every $k \le \sqrt{n} - 2$,*

$$\mathbb{P}\left[\sigma_{n-k+1}\left(z - A - \boldsymbol{M}_n\right) \le \epsilon\right] \le (1+k^2)\binom{n}{k}^2\left(C_{4.5}k^2(nK)^3\left((B_{\boldsymbol{M}_n,2k^2} + \|A\| + |\Re z|)^2 + |\Im z|^2\right)\frac{\epsilon^2}{|\Im z|}\right)^{k^2},$$

where $C_{4.5} = 8\sqrt{3}(e\pi)^{3/2}$.

Theorem 4.5G (Singular Values of Complex Shifts: Gaussian)**.** *Let $z \in \mathbb{C}\setminus\mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{H}_n$ be a normalized $n\times n$ real Ginibre matrix. For every $\delta > 0$, and every $k \le n/7$,*

$$\mathbb{P}\left[\sigma_{n-k+1}(z - A - \delta\boldsymbol{H}_n) \le \epsilon\right] \le \binom{n}{k}^2\left(\frac{\sqrt{7e}k^2n^3}{2\delta^3}\left((9\delta + \|A\| + |\Re z|)^2 + |\Im z|^2\right)\frac{\epsilon^2}{|\Im z|}\right)^{k^2}.$$

The proofs of Theorems 4.4–4.5G appear in Sections 4.4 and 4.5, and rely on anticoncentration bounds for quadratic polynomials in independent, absolutely continuous random variables as well as Gaussians, which may be of independent interest and are developed in Section 4.3.

Theorems 4.4–4.5G are valuable because they apply to $O(n)$ singular values and have the correct dependence on ϵ in the limit $\epsilon \longrightarrow 0$. In the large-$n$ context where one hopes to prove convergence of the empirical spectral distribution of $A + \delta\boldsymbol{M}$ to the Brown measure of an appropriate limiting object, such tail bounds give one crucial control over the log potential — we refer the reader to [35] for a survey and to [142, 148, 90] for examples of specific applications. In a different direction, by the log majorization property of singular values and eigenvalues in Lemma 3.11, they also imply bounds on the probability that many eigenvalues of $A + \delta\boldsymbol{M}$ lie in a small region of the complex plane — such bounds are variously referred to as Wegner- or Minami-type estimates and have been studied primarily for Hermitian random matrices and the Anderson model on $\mathbb{Z}^d$, see for example [74, 2, 163, 115, 50] and references within.

Along these lines, we will obtain the following minimum gap bounds in Section 4.6 by controlling the bottom two singular values of complex shifts and employing a simple net argument over the complex plane.

Theorem 4.6 (Minimum Eigenvalue Gap). *Let $n \geq 16$, $A \in \mathbb{R}^{n\times n}$ be deterministic, and $\boldsymbol{M}_n$ be a random matrix satisfying Assumption 4.1 with parameter $K > 0$. For any $0 < \delta < K$ and $s < 1 < R$:*

$$\mathbb{P}\left[\operatorname{gap}(A + \delta\boldsymbol{M}_n) \leq s\right] \leq C_{4.6} R^2 \left(\delta B_{\boldsymbol{M}_n,8} + \|A\| + R\right)(K/\delta)^{5/2} n^4 s^{2/7} + \mathbb{P}\left[\|A + \boldsymbol{M}_n\| \geq R\right], \tag{4.3}$$

where $C_{4.6}$ is a universal constant defined in equation (4.51). Moreover, if $\boldsymbol{H}_n$ is an $n \times n$ real Ginibre and $0 < \delta < 1$ then

$$\mathbb{P}\left[\operatorname{gap}(A + \delta\boldsymbol{H}_n) \leq s\right] \leq 15\left(\|A\| + 7\right)^3 n^3 \delta^{-5/2} s^{2/7} + e^{-2n}. \tag{4.4}$$

The novelty of this result in comparison to existing minimum gap bounds (such as [81, 112]) is that it works for *heterogeneous non-centered* random matrices X, as opposed to only matrices with i.i.d. entries.

Finally, by combining the above results and using the limiting area/length approach (employing the minimum gap bound to rule out nonreal eigenvalues with small imaginary part, we can finally control the eigenvector condition number. In the following theorem, a typical setting has $\|A\|, \|M_n\|, K$, and R all of order $\Theta(1)$, so one may obtain upper bounds of order $\operatorname{poly}(n, 1/\delta)$ with high probability by setting ϵ_1, ϵ_2 appropriately.

Theorem 4.7 (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 4.1 with parameter $K > 0$. Let $0 < \delta < K\min\{1, \|A\| + R\}$, and write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \delta\boldsymbol{M}_n$. Let $R > \mathbb{E}\|\delta\boldsymbol{M}_n\|$. Then for any $\epsilon_1, \epsilon_2 > 0$, with probability at least*

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

we have

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{R}} \kappa_i(A+\delta \boldsymbol{M}_n) \le \epsilon_1^{-1} C_{4.7} K n^2 \frac{\|A\|+R}{\delta},$$

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}} \kappa_i^2(A+\delta \boldsymbol{M}_n) \le \epsilon_1^{-1}\log(1/\epsilon_2) C_{4.7} K^3 n^5 \cdot \frac{(\|A\|+R)^3}{\delta^3}, \qquad \text{and}$$

$$\kappa_V(A+\delta \boldsymbol{M}_n) \le \epsilon_1^{-1}\sqrt{\log(1/\epsilon_2)} C_{4.7} K^{3/2} n^3 \cdot \frac{(\|A\|+R)^{3/2}}{\delta^{3/2}},$$

Theorem 4.7G (Eigenvalue and Eigenvector Condition Numbers: Gaussian)**.** *Let $n \ge 7$. Let $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{H}_n$ be a real Ginibre matrix. Let $0 < \delta < \min\{1, \|A\|\}$, and write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \delta\boldsymbol{H}_n$. Then for any $\epsilon_1, \epsilon_2 > 0$, with probability at least $1 - 2\epsilon_1 - \frac{30\|A\|^{8/5} n^{8/5}}{\delta^{8/5}}\epsilon_2^{3/5} - 2e^{-2n}$ we have*

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{R}} \kappa_i(A+\delta \boldsymbol{H}_n) \le 5\epsilon_1^{-1} n \frac{\|A\|}{\delta},$$

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}} \kappa_i^2(A+\delta \boldsymbol{H}_n) \le 1000\epsilon_1^{-1}\log(1/\epsilon_2) \frac{n^5\|A\|^3}{\delta^3}, \qquad \text{and}$$

$$\kappa_V(A+\delta \boldsymbol{M}_n) \le 1000\epsilon_1^{-1}\sqrt{\log(1/\epsilon_2)} \frac{n^3\|A\|^{3/2}}{\delta^{3/2}}.$$

By assuming a smaller upper bound on δ, one can make order of magnitude improvements in the constants, so we have made no effort to optimize them. To prove Theorem 1.11, simply take a union bound over the events in Theorems 4.6 and 4.7G, and set s, ϵ_1, and ϵ_2 to ensure that each probability term is $O(1/n)$. The proofs of Theorems 4.7 and 4.7G appear in Section 4.7, and we conclude with a discussion of open questions in Section 4.8.

Related Work

Eigenvalue Condition Numbers and Overlaps. We surveyed the literature on eigenvalue condition numbers in the complex Ginibre ensemble in Chapter 3; for the real Ginibre ensemble, results are more limited. The paper [80] gives a formula for the joint density of a *real* eigenvalue and its (squared) eigenvalue condition number. Compared to such a joint density formula, our Theorem 4.7 (a polynomial upper bound with high probability) is rather coarse, but our theorem holds for general continuous matrices. Besides our result, we are not aware of any results in the literature regarding diagonal overlaps for nonreal eigenvalues of the real Ginibre ensemble, or any other non-Hermitian random matrix model with real entries.

Singular Values of Real Matrices with Complex Shifts. In the course of our proof, it will be of particular importance to quantify the behavior of the small singular values of $z - A - \delta \boldsymbol{H}_n$ or $z - A - \delta \boldsymbol{M}_n$ as a function of the imaginary part of the complex scalar $z \in \mathbb{C}$. There have already been a number of recent results in this direction, which we summarize below.

In the thesis of Ge [81] 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

$$\mathbb{P}\left[\sigma_n(z - \boldsymbol{M}_n) \le \epsilon \text{ and } \|\boldsymbol{M}_n\| \le M\right] \le \frac{Cn^2\epsilon^2}{|\Im(z)|} + e^{-cn} \tag{4.5}$$

for all z, where C and c are universal constants, independent of n. 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, Erdős and Schröder in [48]:

$$\mathbb{P}\left[\sigma_n(\boldsymbol{H}_n - z) \le \epsilon\right] \le C(n^2(1 + |\log \epsilon|)\epsilon^2 + n\epsilon e^{-\frac{1}{2}n(\Im z)^2}) \tag{4.6}$$

for $|z| \le 1 + O(1/\sqrt{n})$, with an improved n-dependence at the edge $|z - 1| = O(1/\sqrt{n})$. In later work [49], the same authors showed that when $\boldsymbol{M}_n$ has real i.i.d. entries with unit variance and $|\Im z| \sim 1$, the statistics of the small singular values $z - \boldsymbol{M}_n$ agree with those of the complex Ginibre ensemble.[1]

As remarked in the introduction, 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 4.23) between submatrices of the real and imaginary parts of the resolvent.

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 [122], 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 4.1 with parameter $K > 0$,

$$\mathbf{P}\left[\sigma_{n-k+1}(\boldsymbol{M}_n) \le \epsilon\right] \le n^{k(k-1)}(CkK\epsilon)^{(k-1)^2}, \tag{4.7}$$

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

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

Result	Bound	Setting
[70]	$\mathbb{P}[\sigma_n(\boldsymbol{M}_n) < \epsilon] \leq n\epsilon$	real Ginibre
[134]	$\mathbb{P}[\sigma_n(\boldsymbol{M}_n) < \epsilon] \leq Cn\epsilon + e^{-cn}$	real i.i.d. subgaussian
[147]	$\mathbb{P}[\sigma_n(\boldsymbol{M}_n) < \epsilon] \leq n\epsilon + O(n^{-c})$	real i.i.d., finite moment assumption
[136]	$\mathbb{P}[\sigma_n(A + \boldsymbol{M}_n) < \epsilon] \leq Cn\epsilon$	real Ginibre, A real
[150]	$\mathbb{P}[\sigma_n(A + \boldsymbol{M}_n) < \epsilon] \leq Cn\epsilon$	real ind. rows with log-concave law, A real
Theorem 4.4G	$\mathbb{P}[\sigma_n(A + \boldsymbol{M}_n) < \epsilon] \leq n\epsilon$	real Ginibre, A real

Table 4.1: Some bounds on σ_n for real $\boldsymbol{M}_n$ and A. Entries of $\boldsymbol{M}_n$ have variance $1/n$.

of ϵ in (4.7) is suboptimal, which renders (4.7) incompatible with our approach. In Theorem 4.4 we 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 4.19) tailored to our setting. Finally for bounds on the least singular value alone, there is a substantial literature; see Table 4.1 for a non-exhaustive summary.

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 [81] that when $\boldsymbol{M}_n$ has i.i.d. entries with zero mean and variance $1/n$, satisfying a standard anticoncentration condition,

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

for every $C > 0$ and every $\zeta > s > n^{-C}$. In recent work, Luh and O'Rourke [112] build on Ge's result, dropping the mean zero assumption and extending the range of s all the way down to 0:

$$\mathbb{P}[\mathrm{gap}(\boldsymbol{M}_n) \leq s \text{ and } \|\boldsymbol{M}_n\| \leq M] \leq Cs^{2/3}n^{16/15} + Ce^{-cn} + \mathbb{P}[\|\boldsymbol{M}_n\| \geq M]. \tag{4.8}$$

However, (4.8) still requires the entries of $\boldsymbol{M}_n$ to be identically distributed, so it does not imply a gap bound for the noncentered Ginibre ensemble $A + \boldsymbol{H}_n$ unless A is a scalar multiple of the all-ones matrix. The only other work we are aware of proving gap bounds for the case of matrices with i.i.d. entries is [137], which proves an inverse polynomial lower bound for the complex Ginibre ensemble.

Alternative Condition Number and Gap Bounds. Independent of (and concurrent to) the results in this chapter, [99] obtained some similar results to ours. Their bound on κ_V improves Theorem 4.7 by a factor of $O(n/(\sqrt{\delta}\log(n/\delta)))$, thus almost matching the dependence on δ in Davies' conjecture, discussed in Section 3.3; their bound on the minimum eigenvalue gap is also better than that supplied by Theorem 4.6 by a poly(n/δ) factor. They do not obtain specific control on the eigenvalue condition numbers for real and complex eigenvalues separately, and our bound for the sum of condition numbers of the real eigenvalues in Theorem 4.7 implies a bound for the maximum which is slightly better than their κ_V bound alone.

The techniques used by [99], and those used here, 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 4.4–4.5G) 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. Finally, they do not take the limit as $\epsilon \longrightarrow 0$ to derive κ_V bounds, relying instead on a bootstrapping scheme, while we do.

4.2 Probabilistic Tools

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. If the distribution $f_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, we denote by

$$\delta_\infty(\boldsymbol{x}) \triangleq \|f_{\boldsymbol{x}}\|_\infty \tag{4.9}$$

the infinity norm of its density. We will use, ad nauseam, two basic observations about the quantity δ_∞. First, for any $v \in \mathbb{R}^d$,

$$\mathbb{P}\left[\|\boldsymbol{x} - v\| \leq \epsilon\right] \leq \frac{\pi^{d/2}}{\Gamma(d/2+1)}\delta_\infty(\boldsymbol{x}) \leq \frac{1}{\sqrt{\pi d}}\left(\frac{2e\pi}{d}\right)^{d/2}\delta_\infty(\boldsymbol{x}), \tag{4.10}$$

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 4.8 (Convolution Bound)**.** Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ be independent random vectors. 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, due first to Rudelson and Vershynin in [135] and improved in [110], quantifying the deterioration of δ_∞ after orthogonal projection.[2]

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

Theorem 4.9 (Projection Bound)**.** *Let $\boldsymbol{x} \in \mathbb{R}^d$ have independent entries, each with density pointwise 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*

$$\delta_\infty(P\boldsymbol{x}) \leq (\sqrt{2}K)^k.$$

If $\boldsymbol{x}$ has independent $\mathcal{N}(0,1)$ entries, the lemma of course holds with the constant 1 instead of $\sqrt{2}$, and we may take $K = (2\pi)^{-1/2}$.

Many of our results on real random matrices whose independent entries have bounded density—in other words, matrices satisfying Assumption 4.1—can be strengthened for real Ginibre matrices. In the complex case treated in Chapter 3, we used the comparison Theorem 3.2 of Śniady to translate classical centered bounds of Szarek and Edelman, Theorem 3.1 and (3.1), to the non-centered case. For real Ginibre matrices, a novel real analogue of Theorem 3.2 facilitates the same approach.

Theorem 4.10 (Real Śniady Theorem)**.** *Let $k \leq n$, and let $A^{(1)}$ and $A^{(2)}$ be $n \times k$ real matrices, each with k distinct singular values, such that $\sigma_i(A^{(1)}) \leq \sigma_i(A^{(2)})$ for all $i \in [k]$. Then for every $t \geq 0$, there exists a joint distribution on pairs of real $n \times k$ random matrices $(\boldsymbol{H}^{(1)}, \boldsymbol{H}^{(2)})$ such that*

(i) Each marginal $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ has independent $\mathcal{N}(0,1)$ entries, and

(ii) Almost surely $\sigma_i(A^{(1)} + t\boldsymbol{H}^{(1)}) \leq \sigma_i(A^{(2)} + t\boldsymbol{H}^{(2)})$ for all $1 \leq i \leq k$.

Proof. For simplicity, we will prove the case $k = n$; adaptation to the general case is straightforward. As in the complex case, the squared singular values $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_n$ of a real matrix Brownian motion satisfy a stochastic differential equation, which was derived by Bru in her work on Wishart processes [39, 40] and independently by Le in her work on shape theory [107, 108]. The equation reads as follows:

$$\mathrm{d}\boldsymbol{\eta}_i = \frac{2\sqrt{\boldsymbol{\eta}_i}}{n}\,\mathrm{d}\boldsymbol{B}_i + \left(1 + \sum_{j \neq i} \frac{\boldsymbol{\eta}_i + \boldsymbol{\eta}_j}{\boldsymbol{\eta}_i - \boldsymbol{\eta}_j}\right)\mathrm{d}t, \qquad 1 \leq i \leq n. \tag{4.11}$$

Śniady's strategy in proving Theorem 3.2, sketched in Chapter 3, crucially relies on the existence and uniqueness of strong solutions to the singular value SDE. This is needed in order to obtain two solutions $\boldsymbol{\eta}_1^{(1)}, ..., \boldsymbol{\eta}_n^{(1)}$ and $\boldsymbol{\eta}_1^{(2)}, ..., \boldsymbol{\eta}_n^{(2)}$ driven by the same Brownian motion but with initial conditions given by the squares singular values of $A^{(1)}$ and $A^{(2)}$, respectively. In particular, we need to assert (i) that the law of each solution indeed matches the law of the singular values of a noncentered Ginibre matrix, and (ii) that they preserve the monotonicity property of the initial singular values. (See [3] for a definition of strong solution and a rigorous proof of existence and uniqueness of strong solutions for Dyson Brownian motion, the Hermitian analogue of the Ginibre singular values process.)

Fortunately, such results are known for the SDE (4.11). Let Λ denote the domain

$$\Lambda \in \mathbb{R}^n \triangleq \{\boldsymbol{\eta} : 0 \leq \boldsymbol{\eta}_n < \cdots < \boldsymbol{\eta}_1\}.$$

For any initial data $\boldsymbol{\eta}(0)$ lying in the closure $\overline{\Lambda}$, it is known that strong solutions to (4.11) exist, are unique, and lie in Λ for all $t > 0$, almost surely [87, Corollary 6.5]. Combining this with [39, Theorem 1], we have that for initial data $\boldsymbol{\eta}(0)$ lying in Λ, the law of the strong solutions to (4.11) matches the law of the squared singular values process of $A + n^{-1/2}\boldsymbol{W}$, where $\boldsymbol{W}$ is a matrix of i.i.d. standard real Brownian motions and A has squared singular values $\boldsymbol{\eta}(0)$. (It should be possible to extend this last statement for initial data in $\overline{\Lambda}$, but the proof may be somewhat involved; a starting point is again [3], which contains a proof of the corresponding extension for Dyson Brownian motion.)

Let $a_i(\boldsymbol{\eta}) = 1 + \sum_{j \neq i} \frac{\eta_i + \eta_j}{\eta_i - \eta_j}$ denote the drift coefficient in (4.11). As in Śniady's proof for the complex Ginibre case, the key property of a allowing for the comparison theorem is the so-called *quasi-monotonicity* (see [66]) or *Kamke–Ważewski condition* [116, §XI.13] from differential inequalities, which is simply that

$$\text{for all } i,\ a_i(\boldsymbol{\eta}^{(1)}) \leq a_i(\boldsymbol{\eta}^{(2)}) \text{ whenever } \boldsymbol{\eta}_i^{(1)} = \boldsymbol{\eta}_i^{(2)} \text{ and } \boldsymbol{\eta}_j^{(1)} \leq \boldsymbol{\eta}_j^{(2)} \text{ for all } j \neq i. \tag{4.12}$$

One easily checks that a satisfies this condition on the domain Λ.

The nonconstant (indeed, non-Lipschitz) diffusion coefficient $2\sqrt{\boldsymbol{\eta}_i}/n$ in (4.11) is a technical obstacle which does not appear in the SDE (3.2) for the complex case. Consequently, the final step of Śniady's proof as sketched below Theorem 3.2 cannot be repeated naively, because taking the difference of two solutions no longer cancels out the diffusion terms. Fortunately, theory has been developed to handle Hölder-1/2 diffusion coefficients; see [132, §IX.3] for exposition of the one-dimensional case and see [102] for a survey of comparison theorems for SDEs in general.

Quasi-monotonicity and the one-dimensional Hölder-1/2 comparison theory are combined in a rather general multidimensional comparison theorem of Geiß and Manthey [82, Theorem 1.2]. Applied to the SDE (4.11), this theorem provides exactly the right conclusion to replace the final step of Śniady's proof. We state the relevant special case of their theorem below:

Theorem 4.11 (Geiß-Manthey). *Consider the SDE*

$$\mathrm{d}\boldsymbol{X}_i = \sigma_i(\boldsymbol{X})\,\mathrm{d}\boldsymbol{B}_i + a_i(\boldsymbol{X})\,\mathrm{d}t, \qquad 1 \leq i \leq n,$$

where the $\boldsymbol{B}_i$ are independent standard real Brownian motions, and $\sigma_i, a_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ are continuous. Suppose the following conditions are satisfied:

(i) the drift coefficient a satisfies the, quasi-monotonicity condition (4.12);

(ii) there exists $\rho : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ increasing with $\int_0^\epsilon \rho^{-2}(u)\,du = \infty$ for some $\epsilon > 0$, such that $|\sigma_i(x) - \sigma_i(y)| \leq \rho(|x_i - y_i|)$ for all i and all $x, y \in \mathbb{R}^n$; and

(iii) strong solutions for the SDE exist for all time and are unique.

Further, assume initial conditions $\boldsymbol{X}^{(1)}(0)$ and $\boldsymbol{X}^{(2)}(0)$ satisfy the inequality $\boldsymbol{X}_i^{(1)}(0) \leq \boldsymbol{X}_i^{(2)}(0)$ for all i. Then almost surely, $\boldsymbol{X}_i^{(1)}(t) \leq \boldsymbol{X}_i^{(2)}(t)$ for all i and for all $t > 0$.

Setting $\rho(u) \triangleq \sqrt{u}$, the SDE (4.11) satisfies the conditions of the Geiß-Manthey theorem, except that our domain for both a_i and σ_i is Λ, not $\mathbb{R}^n$. We address these two coefficients in turn.

First we deal with the drift coefficient a_i, using a standard localization argument already implicit in the proof of Geiß and Manthey. They (implicitly) define the stopping time $\boldsymbol{\vartheta}_N$ to be the first time $\|\boldsymbol{X}^{(1)}\| \geq N$ or $\|\boldsymbol{X}^{(2)}\| \geq N$, and use the fact that a is Lipschitz on the restricted domain $\|\boldsymbol{X}\| \leq N$ to show that

$$\mathbb{P}\left[\boldsymbol{X}_i^{(1)}(t) \leq \boldsymbol{X}_i^{(2)}(t) \text{ for all } 0 \leq t \leq \boldsymbol{\vartheta}_N\right] = 1.$$

Since strong solutions exist for all time, we have $\boldsymbol{\vartheta}_N \longrightarrow \infty$ as $N \longrightarrow \infty$ almost surely, which proves the theorem. We modify this strategy for our SDE (4.11) in the standard way: Define the stopping time $\boldsymbol{\tau}_{1/m}$ to be the first time either $\boldsymbol{\eta}^{(1)}$ or $\boldsymbol{\eta}^{(2)}$ leaves the set

$$\Lambda_{1/m} \triangleq \{\boldsymbol{\eta} \in \Lambda \, : \, |\boldsymbol{\eta}_i - \boldsymbol{\eta}_{i+1}| > 1/m \text{ for all } 1 \leq i \leq n-1.\}.$$

Since strong solutions starting in Λ stay in Λ for all $t \geq 0$ and are continuous, we have $\boldsymbol{\tau}_{1/m} \longrightarrow \infty$ as $m \longrightarrow \infty$ almost surely. Since our a is Lipschitz on $\Lambda_{1/m}$, the proof of Theorem 4.11 shows that

$$\mathbb{P}\left[\boldsymbol{\eta}_i^{(1)}(t) \leq \boldsymbol{\eta}_i^{(2)}(t) \text{ for all } 0 \leq t \leq \boldsymbol{\tau}_{1/m}\right] = 1$$

for all m. Taking $m \longrightarrow \infty$, the result follows.

Finally, we address the diffusion coefficient $\sigma_i(\boldsymbol{\eta}) = 2\sqrt{\boldsymbol{\eta}_i}/n$. The standard fix is to first modify the SDE to have diffusion coefficients $2\sqrt{|\boldsymbol{\eta}_i|}/n$ for all i, so that the domain of σ_i is enlarged to $\mathbb{R}^n$ and Theorem 4.11 may be applied. For this modified SDE, note that the constant zero function $\boldsymbol{\eta}^{(1)}(t) = 0$ is a strong solution. Now let $\boldsymbol{\eta}^{(2)}$ be any solution with $\boldsymbol{\eta}_i^{(2)}(0) \geq 0$ for all i. Applying Theorem 4.11 to $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\eta}^{(2)}$, we conclude that in fact, $\boldsymbol{\eta}^{(2)}(t) \geq 0$ for all $t \geq 0$. Thus, the absolute value bars in the modified SDE can be removed a posteriori. This argument is used, for example, when setting up the SDE for the so-called *Bessel process*, which shares this square-root diffusion coefficient—see [132, §XI.1] for details. □

As from Theorem 3.2, we obtain from Theorem 4.10 a stochastic dominance result relating the singular value distributions of non-centered real Gaussian matrices.

Corollary 4.12. *Let $k \leq n$, and let $A^{(1)}$ and $A^{(2)}$ be $n \times k$ real matrices satisfying $\sigma_i(A^{(1)}) \leq \sigma_i(A^{(2)})$ for all $i \in [k]$. Then, for any $t, s_1, \ldots, s_k \in \mathbb{R}_{\geq 0}$,*

$$\mathbb{P}\left[\sigma_i(A^{(1)} + t\boldsymbol{H}_n) \leq s_i, \forall i \in [k]\right] \geq \mathbb{P}\left[\sigma_i(A^{(2)} + t\boldsymbol{H}_n) \leq s_i, \forall i \in [k]\right].$$

Finally, we will need classical singular value tail bounds for centered real Ginibre matrices, due (as in the complex case, Theorem 3.1), to Szarek in [146].

Theorem 4.13 (Szarek)**.** *Let $\boldsymbol{H}_n$ be a normalized real Ginibre matrix. There exists a universal constant $c > 0$ so that*

$$(c\epsilon)^{k^2} \leq \mathbb{P}\left[\sigma_{n-k+1}(\boldsymbol{H}_n) \leq \frac{k\epsilon}{n}\right] \leq (\sqrt{2e}\epsilon)^{k^2}.$$

To end this section, we will bound the quantities $B_{\boldsymbol{H}_n,p} = \mathbb{E}\left[\|\boldsymbol{H}_n\|^p\right]^{1/p}$ explicitly in the Gaussian case:

Lemma 4.14. *Let $\boldsymbol{H}_n$ be an $n \times n$ real Ginibre matrix and assume that $1 \leq p \leq 2n$. Then $B_{\boldsymbol{H}_n,p} \leq 9$.*

Proof. The proof proceeds by integrating well-known tail bounds on th operator norm of a real Ginibre matrix. Begin by observing that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{H}_n\|^p] &= p\int_0^2 t^{p-1}\mathbb{P}[\|\boldsymbol{H}_n\| \geq t]\,\mathrm{d}t + p\int_2^\infty t^{p-1}\mathbb{P}[\|\boldsymbol{H}_n\| \geq t]\,\mathrm{d}t \\ &\leq 2^p + p\int_2^\infty t^{p-1}\exp\left(-n(t-2)^2/2\right)\,\mathrm{d}t \qquad (4.13)\end{aligned}$$

where the last inequality used a standard tail bound on $\|\boldsymbol{H}_n\|$ (see for example [52]). Now, by Jensen's inequality, for $t \geq 2$ we have

$$t^{p-1} = (t-2+2)^{p-1} \leq \frac{1}{2}\left(2^{p-1}(t-2)^{p-1} + 4^{p-1}\right).$$

Then, use this inequality and the formula for the absolute moments of the Gaussian distribution to bound the last integral in (4.13). That is,

$$\int_2^\infty t^{p-1}\exp\left(-n(t-2)^2/2\right)\,\mathrm{d}t \leq 2^{p-2}\cdot\frac{2^{\frac{p-1}{2}}\Gamma(p/2)}{2n^{\frac{p-1}{2}}\sqrt{\pi}} + 4^{p-2}.$$

Hence

$$\mathbb{E}[\|\boldsymbol{H}_n\|^p] \leq 2^p + \frac{p2^{\frac{p-1}{2}}\Gamma(p/2)}{2n^{\frac{p-1}{2}}\sqrt{\pi}} + p4^{p-2} = 2^p + \frac{2^{\frac{p-1}{2}}\Gamma(p/2+1)}{n^{\frac{p-1}{2}}\sqrt{\pi}} + p4^{p-2} \leq 2^p + \left(\frac{\sqrt{p}}{n^{\frac{p-1}{2p}}}\right)^p + 5^p$$

Now, since $p \leq \sqrt{n}$ and using the fact that all the terms in the above inequality are positive

$$\mathbb{E}[\|\boldsymbol{H}_n\|^p]^{\frac{1}{p}} \leq 2 + \frac{\sqrt{p}}{n^{\frac{p-1}{2p}}} + 5.$$

Since for $x > 1$ the function $x^{\frac{x}{x-1}}$ is increasing, and we are assuming that $p \leq 2n$ we have $p^{\frac{p}{p-1}} \leq (2n)^{\frac{2n}{2n-1}} \leq 4n$. Thus $p \leq 4^{\frac{p-1}{p}} n^{\frac{p-1}{p}}$, which implies $\sqrt{p} \leq 2n^{\frac{p-1}{2p}}$ and concludes the proof. □

4.3 Anticoncentration

In this section we study the anticoncentration properties of certain quadratic functions of rectangular matrices with independent entries. These will be necessary in Section 4.5 to extract singular value tail bounds.

Theorem 4.15 (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 $Z \in \mathbb{R}^{n\times n}$, $U, V \in \mathbb{R}^{n\times k}$, and $W \in \mathbb{R}^{k\times k}$ be deterministic, and write $q(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \boldsymbol{X}^\top Z\boldsymbol{Y} + \boldsymbol{X}^\top U + V^\top \boldsymbol{Y} + W$. Then*

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

Whenever $\sigma_j(Z)$ is zero, we interpret $1/\sigma_j(Z) = \infty$; thus the above theorem has content only when $\mathrm{Rank}(Z) > k^2 + k + 1$. After presenting the proof, we will comment on some improvements when X, Y are Gaussian or $k = 1$. Let us begin with a small observation that we will use in the proof to come.

Lemma 4.16. *Consider measurable functions $f : \mathbb{R}^p \times \mathbb{R}^q \longrightarrow \mathbb{R}^r$ and $c : \mathbb{R}^q \longrightarrow \mathbb{R}_{\ge 0}$. 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 $\delta_\infty\left(f(\boldsymbol{x}, y)\right) \le c(y)$. Then*

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

Proof. Let $\mathrm{Leb}^r_{\mathbb{R}}$ 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

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

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

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

as we wanted to show. □

Second, we will require the following left tail bound on the smallest singular value of certain rectangular random matrices, which is a direct consequence of Theorem 4.9.

Lemma 4.17. *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 $k \le j \le n$ let V be a $j \times n$ projector. Then*

$$\mathbb{P}[\sigma_k(V\boldsymbol{Y}) \le s] \le k\frac{(\sqrt{2}K\sqrt{\pi k}s)^{j-k+1}}{\Gamma((j-k+3)/2)} \triangleq C_{j,k}s^{j-k+1} \tag{4.14}$$

Proof. Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_k$ be the columns of $\boldsymbol{Y}$ and for every $i = 1, \dots, k$ 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" [148], 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 4.9 and Observation 4.16 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 proceed with the proof.

Proof of Theorem 4.15. 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 an orthogonal projection and $S \succeq 0$. By Theorem 4.9, the density of the random matrix $\boldsymbol{X}^\mathsf{T} V$ in $\mathbb{R}^{k \times k}$ is at most $(\sqrt{2}K)^{k^2}$, and thus the density of $\boldsymbol{X}^\mathsf{T} VS$ 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 4.16,

$$\delta_\infty(q(\boldsymbol{X}, \boldsymbol{Y})) \leq (\sqrt{2}K)^{k^2} \mathbb{E}\left[\prod_{i \in k} \sigma_i(Z\boldsymbol{Y} + U)^{-k} \right]. \tag{4.15}$$

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,

$$Z = P^T \Sigma Q = \begin{pmatrix} P_1^\mathsf{T} & P_2^\mathsf{T} \end{pmatrix} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}, \tag{4.16}$$

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

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

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

$$\mathbb{E}\left[\prod_{i\in[k]} \sigma_i(\Sigma_1 Q_1 Y + P_1 U)^{-k}\right] \leq \sigma_j(\Sigma_1)^{-k^2} \mathbb{E}\left[\prod_{i\in[k]} \sigma_i(Q_1 Y + \Sigma_1^{-1} P_1 U)^{-k}\right]. \tag{4.17}$$

Now, since Q_1 is a orthogonal projection, we can select a matrix $\check{U}$ so that $Q_1\check{U} = \Sigma_1^{-1} P_1 U$, and observe that

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

The random matrix $Y + \check{U}$ satisfies the conditions of Lemma 4.17, so we can apply the tail formula for expectation to obtain

$$\begin{aligned}
\mathbb{E}\left[\sigma_k(Q_1(Y + \check{U}))^{-k^2}\right] &= \int_0^\infty \mathbb{P}\left[\sigma_k(Q_1(Y + \check{U}))^{-k^2} \geq t\right] \mathrm{d}t \\
&\leq \lambda + C_{j,k} \int_\lambda^\infty t^{-\frac{j-k+1}{k^2}} \,\mathrm{d}t && C_{j,k} \text{ from (4.14)} \\
&= \lambda + C_{j,k} \frac{k^2}{j - k^2 - k + 1} \lambda^{\frac{k^2+k-j-1}{k^2}} && \text{if } j - k + 1 > k^2.
\end{aligned}$$

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

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

where we have used that $j - k + 1 > k^2$ in the second and fourth lines, as well as Stirling's approximation — $\Gamma(z + 1) \geq \sqrt{2\pi z}(z/e)^z$, valid for real $z \geq 2$ — in the third. To complete the proof, we combine the above with equation (4.15). □

To end this section, we offer some improvements of the above results when k is small or $\boldsymbol{X}$ and Y are Gaussian.

Corollary 4.18. *In the case $k = 1$, the conclusion of Theorem 4.15 may be improved to*

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

Moreover, in the Gaussian case, we may replace $(\sqrt{2}K)^2$ with $(2\pi)^{-1}$.

Proof. The discussion between equations (4.15) and (4.17) 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 4.15 should hold, for every k, with the jth singular value of Z exchanged for the geometric mean of the top j. The main obstacle seems to be that Theorem 4.9 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.

In a different direction, one can improve the constant in Theorem 4.15 under a Gaussian assumption.

Theorem 4.15G. *If $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times k}$ have independent, standard Gaussian entries, then Theorem 4.15 holds with the stronger conclusion:*

$$\delta_\infty\left(q(\boldsymbol{X}, \boldsymbol{Y})\right) \leq \left(\frac{1}{2}\min_{j>2k}\frac{1}{\sqrt{j-2k+1}\sigma_j(Z)}\right)^{k^2}. \tag{4.18}$$

Proof. Once again we modify the proof beginning at (4.17). Observing that $Q_1\boldsymbol{Y} + \Sigma_1^{-1}P_1 U$ is a $j\times k$, non-centered Gaussian matrix, Theorem 4.10 implies

$$\mathbb{E}\prod_{i=1}^{k}\sigma_i(Q_1\boldsymbol{Y} + \Sigma_1^{-1}P_1 U)^{-k^2} \leq \mathbb{E}\prod_{i=1}^{k}\sigma_i(Q_1\boldsymbol{Y})^{-k} = \mathbb{E}(\det\boldsymbol{Y}^\intercal Q_1^\intercal Q_1 Y)^{-k^2/2}.$$

Now, $\boldsymbol{Y}^\top Q_1^\top Q_1 \boldsymbol{Y}$ is a real Wishart matrix with parameters (j, k), and it is known [86] that the determinant of such a matrix is distributed as a product of independent χ^2 random variables $\boldsymbol{\nu}_j \boldsymbol{\nu}_{j-1} \cdots \boldsymbol{\nu}_{j-k+1}$, where $\boldsymbol{\nu}_l \sim \chi^2(l)$. Computing directly,

$$\mathbb{E}\boldsymbol{\nu}_l^{-k/2} = \int_0^\infty \frac{x^{l/2-k/2-1}\exp(-x/2)}{2^{l/2}\Gamma(l/2)}\,\mathrm{d}x = \frac{2^{-k/2}\Gamma\left((l-k)/2\right)}{\Gamma\left(l/2\right)},$$

whenever $l > k$. For even k, this has the closed form $(l-2)^{-1}(l-4)^{-1}\cdots(l-k)^{-1} \le (l-k)^{-k/2}$. This final bound holds for odd $k \ge 3$, by repeated application of $z\Gamma(z) = \Gamma(z+1)$ and one use of the inequality $\sqrt{2z/\pi}\Gamma(z) \le \Gamma(1/2+z) \le \sqrt{z}\Gamma(z)$, valid for all $z \ge 1/2$. When $k = 1$, this inequality again gives us $\mathbb{E}\nu_l^{-1/2} \le (\pi(l-1)/2)^{-1/2}$. As above, we can apply Theorem 4.9 with the constant 1 instead of $\sqrt{2}$ and $K = (2\pi)^{-1/2}$ in the Gaussian case, so

$$\begin{aligned}
\delta_\infty\left(q(\boldsymbol{X}, \boldsymbol{Y})\right) &\le \left(\frac{1}{\sqrt{2\pi}\sigma_j(Z)}\right)^{k^2} \prod_{l=j-k+1}^{k} \mathbb{E}\nu_l^{-k/2} \\
&\le \left(\frac{1}{2\sigma_j(Z)}\right)^{k^2} \prod_{l=j-k+1}^{j} (l-k)^{-k/2} \\
&\le \left(\frac{1}{2\sqrt{j-2k+1}\sigma_j(Z)}\right)^{k^2}.
\end{aligned}$$

The condition $j > 2k$ ensures that each $\mathbb{E}\nu_l^{-k/2} < \infty$ for $l = j-2k+1, \ldots, j$. □

4.4 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 4.1.

As a warm-up, we obtain as an immediate consequence of Theorem 4.13 and Corollary 4.12—Szarek's singular value bounds for centered real Ginibre matrices, and the stochastic dominance corollary to the real Śniady Comparison Theorem —that

$$\mathbb{P}\left[\sigma_{n-k+1}(z - (A + \delta\boldsymbol{H}_n)) \le \epsilon\right] \le \left(\frac{\sqrt{2}e n\epsilon}{k\delta}\right)^{k^2} \tag{4.19}$$

for every $\delta > 0$ and $k \in [n]$. This ϵ^{k^2} behavior will be a useful benchmark by which to assess our results below.

For matrices with i.i.d. subgaussian entries, results similar to Szarek's theorem 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 $\epsilon \longrightarrow 0$. The closest result to ours is due to Nguyen in [122]; 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.

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.

Lemma 4.19 (Principal Submatrix with Large λ_k)**.** *Let $X \in \mathbb{C}^{n\times n} \setminus \{0\}$ be positive semidefinite with eigenvalues $\lambda_n(X) \leq \cdots \leq \lambda_1(X)$. Then for every $1 \leq k \leq n$, there exists an $k \times k$ principal submatrix $X_{S,S}$, with eigenvalues $\lambda_k(X_{S,S}) \leq \cdots \leq \lambda_1(X_{S,S})$, such that*

$$\lambda_k(X_{S,S}) \geq \frac{\operatorname{tr}(X)}{\sum_{i=1}^k \lambda_i(X)} \cdot \frac{\lambda_k(X)}{k(n-k+1)}. \tag{4.20}$$

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

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

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 $|T| = k - 1$,

$$\begin{aligned}\sum_{j\in[k]} \lambda_j(X) \sum_{\ell\notin T} \lambda_\ell(X) &= \sum_{j\in[k]} \lambda_j(X) \left(e_1(X) - \sum_{j\in T} \lambda_j(X) \right) \\ &= \lambda_k(X)e_1(X) + \left(\sum_{j\in[k-1]} \lambda_j(X) \right) e_1(X) - \left(\sum_{j\in T} \lambda_j(X) \right) \left(\sum_{j\in[k]} \lambda_j(X) \right) \\ &\geq \lambda_k(X)e_1(X),\end{aligned}$$

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 $e_1(X) = \operatorname{tr} X$,

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

as desired. □

We will employ Lemma 4.19 in the form of the corollary below.

Corollary 4.20. *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*

$$\sigma_k(Q) \geq \frac{\sigma_k(R)}{\sqrt{k(n-k+1)}}. \tag{4.21}$$

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

$$\sigma_k(A_{S,T}) \geq \frac{\|A\|_F}{\sqrt{\sum_{i\in[k]} \sigma_i(A)^2}} \frac{\sigma_k(A)}{k(n-k+1)} \geq \frac{\sigma_k(A)}{k(n-k+1)} \tag{4.22}$$

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 4.20 additionally sits within a much larger literature on *restricted invertibility*; see [120] for a comprehensive introduction. Most notably, the main result in [83] 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

$$\frac{1}{\sum_{i=1}^k \sigma_i(Q)^{-2}} \geq \frac{1}{(n-k+1)\sum_{i=1}^k \sigma_i(R)^{-2}}. \tag{4.23}$$

Note that neither (4.21) implies (4.23) nor (4.23) implies (4.21). However, from (4.21) one can derive an inequality very similar to (4.23) that has a slightly weaker dependence on k, and vice versa. The proof in [83] shares some features with our proof of Lemma 4.17, 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 (4.22), which is not clear how to obtain from (4.23).

Proof of the Tail Bound

Restatement of Theorem 4.4. *Let $\boldsymbol{M}_n \in \mathbb{R}^{n\times n}$ be a random matrix satisfying Assumption 4.1 with parameter $K > 0$. Then*

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

Proof of Theorem 4.4. We repeat the argument of Nguyen [122], but using Corollary 4.20 where Nguyen uses the restricted invertibility theorem of [120].

Suppose $\sigma_{n-k+1}(\boldsymbol{M}_n) \le \epsilon$. By the minimax formula for singular values, there exist (random) orthogonal unit vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_k \in \mathbb{R}^n$ such that $\|\boldsymbol{M}_n\boldsymbol{z}_i\| \le \epsilon$. 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 $\|\boldsymbol{M}_n\boldsymbol{Z}\|_F \le \epsilon\sqrt{k}$. Since $\sigma_k(\boldsymbol{Z}) = 1$, by Corollary 4.20, there is a $k \times k$ submatrix $\boldsymbol{Z}_1$ of $\boldsymbol{Z}$ for which

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

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

$$\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}_1^{-1} = \begin{pmatrix}\boldsymbol{M}_1 & \boldsymbol{M}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\end{pmatrix}\boldsymbol{Z}_1^{-1} = \boldsymbol{M}_1 + \boldsymbol{M}_2\boldsymbol{Z}_2\boldsymbol{Z}_1^{-1}. \tag{4.24}$$

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 $\{\boldsymbol{m}_i\}_{i\notin S}$, so that $\boldsymbol{H}\boldsymbol{M}_2 = 0$. Thus we have

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

Since the entries of $\boldsymbol{M}_n$ are independent, with densities on $\mathbb{R}$ bounded by $\sqrt{n}K$, by Theorem 4.9 the above event occurs with probability at most

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

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

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

□

Comparing with Szarek's centered singular value bounds (Theorem 4.13 above), we conclude that the exponent of ϵ in Theorem 4.4 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 4.21. *Let $z \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. Then*

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

We record our initial observation regarding real Ginibre matrices, equation (4.19), as the following theorem.

Restatement of Theorem 4.4G. *Let $z \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and $\boldsymbol{H}_n$ be a normalized Ginibre matrix. For every $\delta > 0$,*

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

In the case $k = 1$, one has a better constant:

$$\mathbb{P}[\sigma_n(z - (A + \delta \boldsymbol{M}_n)) \leq \epsilon] \leq \frac{n\epsilon}{\delta}.$$

Proof. When $A = 0$, this is Theorem 4.13, and the better constant for $k = 1$ is a result of Edelman [70]. The conclusion for general A then follows from Corollary 4.12. □

4.5 Singular Value Bounds for Real Matrices with Complex Shifts

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

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

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

Restatement of Theorem 4.5. *Let $z \in \mathbb{C} \setminus \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. For every $k \leq \sqrt{n} - 2$,*

$$\mathbb{P}\left[\sigma_{n-k+1}\left(z - A - \boldsymbol{M}_n\right) \leq \epsilon\right] \leq (1+k^2)\binom{n}{k}^2 \left(C_{4.5} k^2 (nK)^3 \left((B_{\boldsymbol{M}_n,2k^2} + \|A\| + |\Re z|)^2 + |\Im z|^2\right)\frac{\epsilon^2}{|\Im z|}\right)^{k^2},$$

where $C_{4.5}$ is a universal constant defined in (4.27).

In the Gaussian case, we can excise this factor of $(1+k^2)$ and extend the range of k.

Restatement of Theorem 4.5G. *Let $z \in \mathbb{C} \setminus \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{H}_n$ be a normalized $n \times n$ real Ginibre matrix. For every $\delta > 0$, and every $k \leq n/7$,*

$$\mathbb{P}\left[\sigma_{n-k+1}(z - (A + \delta \boldsymbol{H}_n)) \leq \epsilon\right] \leq \binom{n}{k}^2 \left(\frac{\sqrt{7e}k^2 n^3}{2\delta^3}\left((\delta B_{\boldsymbol{H}_n,2k^2} + \|A\| + |\Re z|)^2 + |\Im z|^2\right)\frac{\epsilon^2}{|\Im z|}\right)^{k^2}.$$

Proof of Theorem 4.5

In view of Corollary 4.20, 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 the main technical work in proving Theorem 4.5 occurs in proving the following lemma, which we shall do later on in this section. Theorem 4.5G requires only a few small modifications to the arguments of the general case, and we defer the proof until the end of the section.

Lemma 4.22 (Tail Bound for Corner of the Resolvent)**.** *Let $\zeta \in \mathbb{R}$, let U be a permutation matrix, and let $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. Denote the upper-left $k \times k$ corner of $(\zeta i U - \boldsymbol{M}_n)^{-1}$ by $\boldsymbol{N}_k$. If $n \geq (k+2)^2$,*

$$\mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \geq 1/\epsilon\right] \leq (1+k^2)\left(8\sqrt{3}(e\pi)^{3/2}K^3 n \frac{\epsilon^2}{|\zeta|}\right)^{k^2} \mathbb{E}\left[\left(\|\boldsymbol{M}_n\|^2 + \zeta^2\right)^{k^2}\right]. \tag{4.25}$$

Proof of Theorem 4.5 assuming Lemma 4.22. Applying Corollary 4.20 and a union bound,

$$\begin{aligned}\mathbb{P}\left[\sigma_{n-k+1}(z - A - \boldsymbol{M}_n) \leq \epsilon\right] &= \mathbb{P}\left[\sigma_k\left((z - A - \boldsymbol{M}_n)^{-1}\right) \geq 1/\epsilon\right] \\ &\leq \mathbb{P}\left[\max_{S,T\subset[n],|S|=|T|=k} \sigma_k\left((z - A - \boldsymbol{M}_n)^{-1}_{S,T}\right) \geq \frac{1}{k(n-k+1)\epsilon}\right] \\ &\leq \sum_{S,T\subset[n],|S|=|T|=k} \mathbb{P}\left[\sigma_k\left((z - A - \boldsymbol{M}_n)^{-1}_{S,T}\right) \geq \frac{1}{k(n-k+1)\epsilon}\right]. \end{aligned} \tag{4.26}$$

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

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

As $PQ^\intercal$ is a permutation matrix and $P(\Re z - (A + \boldsymbol{M}_n))Q^\intercal$ satisfies Assumption 4.1 with parameter $K > 0$, we can apply Lemma 4.22. Defining

$$C_{4.5} \triangleq 8\sqrt{3}(e\pi)^{3/2}, \tag{4.27}$$

this gives

$$\begin{aligned}&\mathbb{P}\left[\sigma_k\left((z - A - \boldsymbol{M}_n)^{-1}_{S,T}\right) \geq \frac{1}{k(n-k+1)\epsilon}\right] \\ =&\mathbb{P}\left[\sigma_k\left(i\Im z PQ^\intercal - P(\Re z - (A + \boldsymbol{M}_n))Q^\intercal\right)^{-1}_{[k],[k]} \geq \frac{1}{k(n-k+1)\epsilon}\right] \\ \leq&(1+k^2)\left(C_{4.5}K^3 n \frac{k^2(n-k+1)^2\epsilon^2}{|\Im z|}\right)^{k^2} \mathbb{E}\left[\left(\|P(\Re z - A + M_n)Q^\intercal\|^2 + |\Im z|^2\right)^{k^2}\right] \\ \leq&(1+k^2)\left(C_{4.5}k^2 n^3 K^3 \frac{\epsilon^2}{|\Im z|}\right)^{k^2} \mathbb{E}\left[\left(\|P(\Re z - A - \boldsymbol{M}_n)Q^\intercal\|^2 + |\Im z|^2\right)^{k^2}\right],\end{aligned}$$

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,

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

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. □

Proof of Lemma 4.22

In what follows we use the notation and assumptions of Lemma 4.22. In particular, $\boldsymbol{M}_n$ satisfies Assumption 4.1 with parameter $K > 0$, U is a permutation matrix, and $\zeta \in \mathbb{R}$. Once again writing

$\boldsymbol{N}_k$ for the upper left $k \times k$ block of $(\zeta iU + \boldsymbol{M}_n)^{-1}$, we need to show that $\mathbb{P}[\|\boldsymbol{N}_k^{-1}\| \leq \epsilon] = O(\epsilon^{2k^2})$. 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 ζU in the following block form:

$$\boldsymbol{M}_n = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{pmatrix} \quad \text{and} \quad \zeta U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \tag{4.28}$$

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

$$\boldsymbol{X} \triangleq \Re(\boldsymbol{M}_{22} + iU_{22})^{-1} \quad \text{and} \quad \boldsymbol{Y} \triangleq \Im(\boldsymbol{M}_{22} + iU_{22})^{-1}. \tag{4.29}$$

Applying the Schur complement formula to the block decomposition in (4.28), 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

$$\Re\boldsymbol{N}_k^{-1} = \boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{X}\boldsymbol{M}_{21} + U_{12}\boldsymbol{Y}\boldsymbol{M}_{21} - \boldsymbol{M}_{12}\boldsymbol{Y}U_{21} + U_{12}\boldsymbol{X}U_{21} \tag{4.30}$$

$$\Im\boldsymbol{N}_k^{-1} = U_{11} - \boldsymbol{M}_{12}\boldsymbol{Y}\boldsymbol{M}_{21} - \boldsymbol{M}_{12}\boldsymbol{X}U_{21} - U_{12}\boldsymbol{X}\boldsymbol{M}_{21} + U_{12}\boldsymbol{Y}U_{21}. \tag{4.31}$$

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

Writing this conditioning explicitly,

$$\begin{aligned} \mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \geq 1/\epsilon\right] &= \mathbb{P}\left[\|\boldsymbol{N}_k^{-1}\| \leq \epsilon\right] \\ &\leq \mathbb{P}\left[\|\Re\boldsymbol{N}_k^{-1} + i\Im\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}\right] \\ &\leq \mathbb{P}\left[\|\Re\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}, \|\Im\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}\right] \\ &= \mathbb{E}\,\mathbb{E}\left[\mathbf{1}\{\|\Re\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}\}\mathbf{1}\{\|\Im\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}\} \,\middle|\, \boldsymbol{M}_{12}, \boldsymbol{M}_{21}, \boldsymbol{M}_{22}\right] \\ &= \mathbb{E}\left[\mathbf{1}\{\|\Im\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}\}\mathbb{E}\left[\mathbf{1}\{\|\Re\boldsymbol{N}_k^{-1}\|_F \leq \epsilon\sqrt{k}\} \,\middle|\, \boldsymbol{M}_{12}, \boldsymbol{M}_{21}, \boldsymbol{M}_{22}\right]\right]. \end{aligned} \tag{4.32}$$

We can bound the inner conditional expectation using Observation 4.23:

$$\mathbb{E}\left[\mathbf{1}\{\|\Re \boldsymbol{N}_k^{-1}\|_F \le \epsilon\sqrt{k}\}\,\middle|\,\boldsymbol{M}_{12},\boldsymbol{M}_{21},\boldsymbol{M}_{22}\right] \le \frac{(\sqrt{\pi k n}K\epsilon)^{k^2}}{\Gamma(k^2/2+1)} \le \left(\frac{\sqrt{2e\pi n}K\epsilon}{\sqrt{k}}\right)^{k^2} \tag{4.33}$$

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 (4.32) gives

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

and we now turn to the more serious task of the requisite small-ball probability estimate for $\Im\boldsymbol{N}_k^{-1}$. This calculation is facilitated by a second key observation, which is an immediate consequence of the full expression (4.31) for $\Im\boldsymbol{N}_k^{-1}$.

Observation 4.24. Conditional on $\boldsymbol{M}_{22}$, the imaginary part $\Im\boldsymbol{N}_k^{-1}$ is a quadratic function in $\boldsymbol{M}_{12}$ and $\boldsymbol{M}_{21}$, of the type studied in Section 4.3.

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 4.15 implies

$$\begin{aligned}\mathbb{P}&\left[\|U_{12} - \boldsymbol{M}_{12}Y\boldsymbol{M}_{21} - \boldsymbol{M}_{12}XU_{21} - U_{12}X\boldsymbol{M}_{21} + U_{12}YU_{21}\|_F \le \epsilon\sqrt{k}\right]\\ &\le (1+k^2)\left(\frac{2K^2n\sqrt{2e\pi k}}{\sqrt{j-k+1}\sigma_j(Y)}\right)^{k^2}\left(\frac{\sqrt{2e\pi}\epsilon}{\sqrt{k}}\right)^{k^2}\\ &= (1+k^2)\left(\frac{4K^2n\cdot e\pi\cdot\varepsilon}{\sqrt{j-k+1}\sigma_j(Y)}\right)^{k^2},\end{aligned} \tag{4.34}$$

(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

$$\mathbb{P}\left[\|\Im\boldsymbol{N}_k^{-1}\| \le \epsilon\right] \le (1+k^2)\left(\frac{4K^2n\cdot e\pi\cdot\varepsilon}{\sqrt{j-k+1}}\right)^{k^2}\mathbb{E}\left[\sigma_j(\boldsymbol{Y})^{-k^2}\right]. \tag{4.35}$$

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 (4.29), we defined $\boldsymbol{Y} = \Im(\boldsymbol{M}_{22}+iU_{22})^{-1}$, 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

$$Y = -(\boldsymbol{M}_{22} + U_{22}\boldsymbol{M}_{22}^{-1}U_{22})^{-1}U_{22}\boldsymbol{M}_{22}^{-1}. \tag{4.36}$$

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

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

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 ζU, by the usual interlacing of singular values for submatrices [98, Corollary 7.3.6], we can be sure that $\sigma_1(U_{22}) = \cdots = \sigma_{n-2k}(U_{22}) = |\zeta|$. Hence, there exists a matrix E of rank at most $2k$ such that $\widehat{U}_{22} \triangleq U_{22} + E$ is invertible, with all singular values equal to $|\zeta|$. We can therefore write

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

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 $\mathrm{Rank}(\boldsymbol{E}_2) \le \mathrm{Rank}(E) \le 2k$, interlacing of singular values upon low-rank updates [149, Theorem 1] ensures

$$\sigma_j(\boldsymbol{Y}) \ge \sigma_{j+2k}\left((\boldsymbol{M}_{22} + \widehat{U}_{22}\boldsymbol{M}_{22}^{-1}U_{22} + \boldsymbol{E}_1)^{-1}\widehat{U}_{22}\boldsymbol{M}_{22}^{-1}\right). \tag{4.37}$$

On the other hand

$$(\boldsymbol{M}_{22} + \widehat{U}_{22}\boldsymbol{M}_{22}^{-1}U_{22} + \boldsymbol{E}_1)^{-1}\widehat{U}_{22}\boldsymbol{M}_{22}^{-1} = (\boldsymbol{M}_{22}\widehat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22} + M_{22}\widehat{U}_{22}^{-1}\boldsymbol{E}_1)^{-1}, \tag{4.38}$$

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

$$\sigma_{j+2k}\left((\boldsymbol{M}_{22}\widehat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22} + \boldsymbol{M}_{22}\widehat{U}_{22}^{-1}\boldsymbol{E}_1)^{-1}\right) \ge \sigma_{j+4k}\left((\boldsymbol{M}_{22}\widehat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22})^{-1}\right). \tag{4.39}$$

Putting together (4.37), (4.38), and (4.39), we get

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

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

$$\sigma_{n-5k}(\boldsymbol{Y})^{-k^2} \le \left\|\boldsymbol{M}_{22}\widehat{U}_{22}^{-1}\boldsymbol{M}_{22} + U_{22}\right\|^{k^2} \le \left(|\zeta|^{-1}\|\boldsymbol{M}_n\|^2 + |\zeta|\right)^{k^2} \tag{4.40}$$

We now assemble our work so far. For every k satisfying $n - k \geq j \geq k^2 + k + 1$,

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

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.

Proof of Theorem 4.5G

We will first modify the proof of Lemma 4.22, referring back to the argument in the prior section. In order to perform these modifications, set $K = 1/\delta$, and think of $\boldsymbol{M}_n = K^{-1}\boldsymbol{H}_n$. As above, $\zeta \in \mathbb{R}$ is a real number, U is a permutation, and we write $\mathbf{N}_k$ for the upper left $k \times k$ block of $(\zeta i U - \boldsymbol{M}_n)^{-1}$. In (4.33), using that the density of each entry of $\boldsymbol{M}_n$ is bounded by $(2\pi)^{-1/2}K\sqrt{n}$, we find

$$\mathbb{E}\left[\mathbf{1}\{\|\Re N_k^{-1}\| \leq \epsilon\} \,\middle|\, M_{12}, M_{21}M_{22}\right] \leq \left(\frac{\sqrt{en}K\epsilon}{\sqrt{k}}\right)^{k^2}.$$

In (4.34) and (4.35), swapping Theorem 4.15G for Theorem 4.15, we have that for any $n - k \geq j > 2k$,

$$\mathbb{P}\left[\|\Im \boldsymbol{N}_k^{-1}\| \leq \epsilon\right] \leq \left(\frac{K^2 n}{2\sqrt{j-2k+1}}\right)^{k^2} \mathbb{E}\left[\sigma_j(\boldsymbol{Y})^{-k^2}\right];$$

finally, in (4.40) if we now set $j = n - 5k$, we have

$$\mathbb{E}\left[\sigma_{n-5k}(\boldsymbol{Y})^{-k^2}\right] \leq \mathbb{E}\left[\left(|\zeta|^{-1}\|M\|^2 + |\zeta|\right)^{k^2}\right].$$

Putting all this together, for any k satisfying $n \geq 7k$,

$$\mathbb{P}\left[\sigma_k(\boldsymbol{N}_k) \geq 1/\epsilon\right] \leq \left(\frac{\sqrt{7e}K^3 n}{2}\frac{\epsilon^2}{|\zeta|}\right)^{k^2} \mathbb{E}\left[\left(\|\boldsymbol{M}_n\|^2 + \zeta^2\right)^{k^2}\right]. \tag{4.41}$$

Now, let $z \in \mathbb{C}$, and continue as in the proof of Theorem 4.5 from Lemma 4.22. Recalling $K = 1/\delta$, and substituting (4.41) in place of (4.25), we obtain

$$\mathbb{P}\left[\sigma_k(z - A - \delta\boldsymbol{H}_n) \leq \epsilon\right] \leq \binom{n}{k}^2 \left(\frac{\sqrt{7e}k^2 n^3}{2\delta^3}\left((\delta B_{\boldsymbol{H}_n,2k^2} + \|A\| + |\Re z|)^2 + |\Im z|^2\right)\frac{\epsilon^2}{|\Im z|}\right)^{k^2}.$$

4.6 Minimum Eigenvalue Gap

This section is devoted to several results regarding eigenvalue gaps of real random matrices with independent entries. Below we state the main result of this section.

Restatement of Theorem 4.6. *Let $n \geq 16$, $A \in \mathbb{R}^{n\times n}$ be deterministic, and $\boldsymbol{M}_n$ be a random matrix satisfying Assumption 4.1 with parameter $K > 0$. For any $0 < \delta < K$ and $s < 1 < R$:*

$$\mathbb{P}\left[\mathrm{gap}(A + \delta \boldsymbol{M}_n) \leq s\right] \leq C_{4.6} R^2 \left(\delta B_{\boldsymbol{M}_n,8} + \|A\| + R\right) (K/\delta)^{5/2} n^4 s^{2/7} + \mathbb{P}\left[\|A + \boldsymbol{M}_n\| \geq R\right],$$

where $C_{4.6}$ is a universal constant defined in equation (4.51). Moreover, if $\boldsymbol{H}_n$ is an $n \times n$ real Ginibre and $0 < \delta < 1$ then

$$\mathbb{P}\left[\mathrm{gap}(A + \delta \boldsymbol{H}_n) \leq s\right] \leq 15\left(\|A\| + 7\right)^3 n^3 \delta^{-5/2} s^{2/7} + e^{-2n}.$$

As discussed in the introduction to this chapter, our proof will mirror the one in Section 3.4, and hinge on the log majorization relationship between eigenvalues and singular values in lemma 3.11. As we did above, we will use a union bound over a well-chosen net. However, the proof here contains as well a new complication, namelly 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 4.4 , and away from it by Theorem 4.5. 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 $\zeta > 0$,

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

Proof of Theorem 4.6 . For most of the proof, let us absorb δ into the constant K—the condition $\delta < 1/K$ will not be relevant until the end.

First observe that if $\zeta > 0$,

$$\{\mathrm{gap}(A + \boldsymbol{M}_n) \leq s\} = \{\mathrm{gap}_{\mathbb{R}}(A + \boldsymbol{M}_n) \leq s\} \cup \{\Im_{\min}(A + \boldsymbol{M}_n) \leq \zeta\} \cup \{\mathrm{gap}_{\Im\geq\zeta}(A + \boldsymbol{M}_n) \leq s\}. \quad (4.42)$$

Now choose a covering of the region $\mathbb{D}(0, R) \subset \mathbb{C}$ 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 (4.42), we will set up a separate net to union bound each of the events appearing on the right-hand side: let

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

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

$$\begin{aligned}\mathbb{P}\left[\operatorname{gap}(A + \boldsymbol{M}_n) \le s\right] \le \sum_{z \in \mathcal{N}_{2s}^{\mathbb{R}}} &\mathbb{P}\left[|\Lambda(A + \boldsymbol{M}_n) \cap \mathbb{D}(z, 3s/2)| \ge 2\right] \\ &+ \sum_{z \in \mathcal{N}_{\zeta}^{\mathbb{R}}} \mathbb{P}\left[|\Lambda(A + \boldsymbol{M}_n) \cap \mathbb{D}(z, \sqrt{2}\zeta)| \ge 2\right] \\ &+ \sum_{z \in \mathcal{N}_{\zeta,s}^{\mathbb{C}}} \mathbb{P}\left[\Lambda(A + \boldsymbol{M}_n) \cap \mathbb{D}(z, \sqrt{5/4}s)| \ge 2\right] \\ &+ \mathbb{P}\left[\|A + \boldsymbol{M}_n\| \ge R\right].\end{aligned} \tag{4.43}$$

The first line controls $\operatorname{gap}_{\mathbb{R}}$, the second one $\Im_{\min}$, the third one $\operatorname{gap}_{\Im \ge \zeta}$, 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 4.4 and 4.5—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. The Gaussian case is quite similar, and we will treat it at the end of the proof.

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

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

for $z \in \mathbb{R}$. As we did in Section 3.4, we will first produce a tail bound on the product of the two smallest singular values of $z - A - \boldsymbol{M}_n$. For every $z \in \mathbb{R}$ and $x > 0$,

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

Optimizing in x, we have

$$\mathbb{P}\left[|\Lambda(A + \boldsymbol{M}_n) \cap \mathbb{D}(z, r)| \ge 2\right] \le \left(4^{1/5} + 4^{-4/5}\right)(2Kr)^{8/5} n^{14/5} \le 3n^{14/5}(\sqrt{2}Kr)^{8/5}. \tag{4.44}$$

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

$$\begin{aligned}\sum_{z \in \mathcal{N}_{s}^{\mathbb{R}}} \mathbb{P}\left[|\Lambda(A + \boldsymbol{M}_n) \cap \mathbb{D}(z, 3s/2)| \ge 2\right] &\le \left|\mathcal{N}_{s}^{\mathbb{R}}\right| \cdot 3n^{14/5}(3\sqrt{2}Ks/2)^{8/5} \\ &\le 9R(\sqrt{2}K)^{8/5} n^{14/5} s^{3/5}\end{aligned} \tag{4.45}$$

Step 2: Eigenvalues Near the Real Line. Using (4.44) and imitating the remainder of Step 1,

$$\sum_{z \in \mathcal{N}_{\zeta}^{\mathbb{R}}} P\left[|\Lambda(A + \boldsymbol{M}_n) \cap \mathbb{D}(z, \sqrt{2}\zeta)| \ge 2\right] \le 8R(\sqrt{2}K)^{8/5} n^{14/5} \zeta^{3/5} \tag{4.46}$$

This directly implies a stand-alone tail bound on $\Im_{\min}$, which we record for use in Section 4.7,:

$$\mathbb{P}\left[\Im_{\min}(A+\boldsymbol{M}_n) \le \zeta\right] \le 8R(\sqrt{2}K)^{8/5} n^{14/5} \zeta^{3/5} + \mathbb{P}[\|\boldsymbol{M}_n\| \ge R]. \tag{4.47}$$

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 $n \ge 16$, Theorem 4.5 implies

$$\begin{aligned}
\mathbb{P}\left[|\|\Lambda(A+\boldsymbol{M}_n) \cap \mathbb{D}(z,r)| \ge 2\right] &\le \min_{x>0} \left\{\mathbb{P}\left[\sigma_n(A+\boldsymbol{M}_n) \le rx\right] + \mathbb{P}\left[\sigma_{n-1}(A+\boldsymbol{M}_n) \le r/x\right]\right\} \\
&\le \min_{x>0} \left\{ 2C_{4.5}K^3 n^5 \left((B_{\boldsymbol{M}_n,2} + \|A\| + |\Re z|)^2 + |\Im z|^2\right) \frac{(rx)^2}{|\Im z|} \right. \\
&\qquad \left. +640 C_{4.5}^4 K^{12} n^{14} \left((B_{\boldsymbol{M}_n,8} + \|A\| + |\Re z|)^2 + |\Im z|^2\right)^4 \frac{r^8}{x^8 |\Im z|^4} \right\} \\
&\le C_{(4.48)} \left(\frac{(B_{\boldsymbol{M}_n,8} + \|A\| + |\Re z|)^2 + |\Im z|^2}{|\Im z|} \right)^{8/5} K^{24/5} r^{16/5} n^{34/5} \qquad (4.48)
\end{aligned}$$

where we have used $B_{\boldsymbol{M}_n,1} \le B_{\boldsymbol{M}_n,8}$ and defined $C_{(4.48)} = 11C_{4.5} = 88\sqrt{3}(e\pi)^{3/2}$.

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

$$\begin{aligned}
&\sum_{z \in \mathcal{N}^{\mathbb{C}}_{\zeta,s}} \mathbb{P}\left[|\Lambda(A+\boldsymbol{M}_n) \cap \mathbb{D}(z, \sqrt{5/4}s)| \ge 2\right] \\
&\qquad \le 6(R/s)^2 C_{(4.48)} \left(\frac{(B_{\boldsymbol{M}_n,8} + \|A\| + R)^2}{\zeta} \right)^{8/5} K^{24/5} (\sqrt{5}s/2)^{16/5} n^{34/5} \\
&\qquad \le C_{(4.49)} R^2 (B_{\boldsymbol{M}_n,8} + \|A\| + R)^{16/5} \frac{K^{24/5} s^{6/5} n^{34/5}}{\zeta^{8/5}} \qquad (4.49)
\end{aligned}$$

where $C_{(4.49)} \triangleq 6(5/4)^{8/5} C_{(4.48)} = 528(5/4)^{8/5}\sqrt{3}(e\pi)^{3/2}$.

Step 4: Conclusion. We now put together the three steps above, substituting (4.45), (4.46), and (4.49) into (4.43), and adding back in the δ scaling. Using the fact that $\psi\zeta^{3/5} + \phi s^{6/5}\zeta^{-8/5} \le 2\psi^{8/11}\phi^{3/11}s^{18/55}$, we obtain

$$\begin{aligned}
\mathbb{P}\left[\operatorname{gap}(A+\boldsymbol{M}_n) \le s\right] &\le 9R(\sqrt{2}K/\delta)^{8/5} n^{14/5} s^{3/5} \\
&\quad + 2\left(C_{(4.49)} R^2 (\delta B_{\boldsymbol{M}_n,8} + \|A\| + R)^2 (K/\delta)^{24/5} n^{34/5}\right)^{3/11} \left(8(\sqrt{2}K/\delta)^{8/5} n^{14/5}\right)^{8/11} s^{18/55} \\
&\quad + \mathbb{P}\left[\|A + \delta\boldsymbol{M}_n\| \ge R\right] \\
&\le C_{4.6} R^{14/11} \left(\delta B_{\boldsymbol{M}_n,8} + \|A\| + R\right)^{6/11} (K/\delta)^{136/55} n^{214/55} s^{18/55} + \mathbb{P}\left[\|A+\boldsymbol{M}_n\| \ge R\right] \\
&\le C_{4.6} R^2 \left(\delta B_{\boldsymbol{M}_n,8} + \|A\| + R\right) (K/\delta)^{5/2} n^4 s^{2/7} + \mathbb{P}\left[\|A+\boldsymbol{M}_n\| \ge R\right], \qquad (4.50)
\end{aligned}$$

where

$$C_{4.6} \triangleq 2C_{(4.49)}^{3/11} \cdot 8^{8/11} \sqrt{2}^{64/55} + 9\sqrt{2}^{8/5} < 250. \tag{4.51}$$

Gaussian Case. Finally, we tackle the Gaussian case; we will be terse, as the structure of the proof is identical. When $z \in R$, Theorem 4.4G gives

$$\begin{aligned} \mathbb{P}\left[|\Lambda(A + \delta \boldsymbol{H}_n) \cap \mathbb{D}(z, r)| \geq 2\right] &\leq \min_{x>0} \left\{ \frac{nrx}{\delta} + 4e^2 \left(\frac{nr}{2\delta x} \right)^4 \right\} \\ &= \frac{5e^{2/5}}{4} (nr/\delta)^{8/5} \leq 2(nr/\delta)^{8/5}. \end{aligned} \tag{4.52}$$

Similarly, using Theorem 4.5G for $z \notin \mathbb{R}$,

$$\begin{aligned} \mathbb{P}\left[|\Lambda(A + \delta \boldsymbol{H}_n) \cap \mathbb{D}(z, r)| \geq 2\right] &\leq \min_{x>0} \left\{ \frac{\sqrt{7e}n^4}{2\delta^3} \frac{(9\delta + \|A\| + |\Re z|)^2 + |\Im z|^2}{|\Im z} (rx)^2 \right. \\ &\qquad \left. + \frac{4 \cdot 7^2 e^2 n^{14}}{8\delta^{12}} \left(\frac{(9\delta + \|A\| + |\Re z|)^2 + |\Im z|^2}{|\Im z|} \right)^4 (r/x)^8 \right\} \\ &= \frac{5(7e)^{4/5}}{4 \cdot 2^{3/5}} \left(\frac{(9\delta + \|A\| + |\Re z|)^2 + |\Im z|^2}{|\Im z|} \right)^{8/5} n^6 r^{16/5} \delta^{-24/5} \\ &\leq 9 \left(\frac{(9\delta + \|A\| + |\Re z|)^2 + |\Im z|^2}{|\Im z|} \right)^{8/5} n^6 r^{16/5} \delta^{-24/5} \end{aligned}$$

Using the same net as in the general proof above, and taking $R \triangleq \|A\| + 4\delta$,

$$\begin{aligned} \mathbb{P}\left[\operatorname{gap}(A + \delta \boldsymbol{H}_n) \leq s\right] &\leq \sum_{z \in \mathcal{N}_{2s}^{\mathbb{R}}} \mathbb{P}\left[|\Lambda(A + \delta \boldsymbol{H}_n) \cap \mathbb{D}(z, 3s/2)| \geq 2\right] \\ &\quad + \sum_{z \in \mathcal{N}_{\zeta}^{\mathbb{R}}} P\left[|\Lambda(A + \delta \boldsymbol{H}_n) \cap \mathbb{D}(z, \sqrt{2}\zeta)| \geq 2\right] \\ &\quad + \sum_{z \in \mathcal{N}_{\zeta,s}^{\mathbb{C}}} \mathbb{P}\left[\Lambda(A + \delta \boldsymbol{H}_n) \cap \mathbb{D}(z, \sqrt{5/4}s)| \geq 2\right] + \mathbb{P}\left[\|A + \delta \boldsymbol{H}_n\| \geq R\right] \\ &\leq \frac{3(\|A\| + 4\delta)}{2s} \cdot 2(3ns/2\delta)^{8/5} + \frac{3(\|A\| + 4\delta)}{2\zeta} \cdot 2(\sqrt{2}n\zeta/\delta)^{8/5} \\ &\quad + 6\left(\frac{\|A\| + 4\delta}{s} \right)^2 \cdot 9 \left(\frac{4(\|A\| + 6.5\delta)^2}{\zeta} \right)^{8/5} n^6 (\sqrt{5/4}s)^{16/5} \delta^{-24/5} + e^{-2n} \\ &\leq 6\left(\|A\| + 4\delta\right)(n/\delta)^{8/5} s^{3/5} + 6\left(\|A\| + 4\delta\right)(n/\delta)^{8/5} \zeta^{3/5} \\ &\quad + 800\left(\|A\| + 6.5\delta\right)^{26/5} n^6 s^{6/5} \zeta^{-8/5} \delta^{-24/5} + e^{-2n}. \end{aligned}$$

Optimizing in ζ using the same argument as the main proof, and $\delta < 1$,

$$\begin{aligned}\mathbb{P}\left[\operatorname{gap}(A + \delta \boldsymbol{H}_n) \leq s\right] &\leq 6\left(\|A\| + 4\delta\right)(n/\delta)^{8/5} s^{3/5} \\ &\quad + 2\left(6\left(\|A\| + 4\delta\right)(n/\delta)^{8/5}\right)^{8/11}\left(800\left(\|A\| + 6.5\delta\right)^{26/5} n^6 \delta^{-24/5}\right)^{3/11} s^{18/55} + e^{-2n} \\ &\leq 6(\|A\| + 4\delta)(n/\delta)^{8/5} s^{3/5} + 7(\|A\| + 6.5\delta)^{118/55} n^{64/55} n^{18/11} \delta^{-136/55} s^{18/55} + e^{-2n} \\ &\leq 15\left(\|A\| + 7\right)^3 n^3 \delta^{-5/2} s^{2/7} + e^{-2n}.\end{aligned}$$

As in the non-Gaussian case, we separately state a tail bound for $\Im_{\min}$:

$$\mathbb{P}\left[\Im_{\min}(A + \delta \boldsymbol{H}_n) \leq \zeta\right] \leq 6\left(\|A\| + 4\delta\right)(n/\delta)^{8/5} \zeta^{3/5}. \tag{4.53}$$

□

4.7 Eigenvalue and Eigenvector Condition Numbers

In this section, we convert our probabilistic lower bounds on the least singular value into upper bounds on the mean eigenvalue condition numbers, following Section 3.2 and using Lemmas 2.4 and 4.3.

Bounds in Expectation

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

Proposition 4.25 (Condition Numbers of Real Eigenvalues). *Let $A \in \mathbb{R}^{n \times n}$ be deterministic, and let $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. Write $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \delta \boldsymbol{M}_n$. Then for every measurable open set $\Omega \subset \mathbb{R}$,*

$$\mathbb{E} \sum_{i : \boldsymbol{\lambda}_i \in \Omega} \kappa_i(A + \delta \boldsymbol{M}_n) \leq \frac{K n^2}{\sqrt{2}\delta} \cdot \operatorname{Leb}_{\mathbb{R}}(\Omega).$$

In the real Ginibre case, one has the improvement

$$\mathbb{E} \sum_{i : \boldsymbol{\lambda}_i \in \Omega} \kappa_i(A + \delta \boldsymbol{H}_n) \leq \frac{n}{2\delta} \cdot \operatorname{Leb}_{\mathbb{R}}(\Omega).$$

Proof. When z is real, $z - A$ is also real, so we may apply the tail bound in Corollary 4.21. In particular, setting $k = 1$, we obtain the following tail bound for real z:

$$\mathbb{P}[\sigma_n((z - A) + \delta(-\boldsymbol{M}_n)) \leq \epsilon] < \frac{\sqrt{2} K n^2 \epsilon}{\delta}.$$

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

$$\begin{aligned}
2\mathbb{E} \sum_{i : \lambda_i \in \Omega} \kappa_i(A + \delta \boldsymbol{M}_n) &\leq \mathbb{E} \liminf_{\epsilon \to 0} \epsilon^{-1} \operatorname{Leb}_{\mathbb{R}} \left(\Lambda_\epsilon(A + \delta \boldsymbol{M}_n) \cap \Omega \right) && \text{Lemma 4.3} \\
&\leq \liminf_{\epsilon \to 0} \epsilon^{-1} \mathbb{E} \int_\Omega \mathbf{1}_{\{z \in \Lambda_\epsilon(A + \delta \boldsymbol{M}_n)\}} \, dz && \text{Fatou's lemma} \\
&= \liminf_{\epsilon \to 0} \epsilon^{-1} \int_\Omega \mathbb{P}[z \in \Lambda_\epsilon(A + \delta \boldsymbol{M}_n)] \, dz && \text{Fubini's theorem} \\
&= \liminf_{\epsilon \to 0} \epsilon^{-1} \int_\Omega \mathbb{P}[\sigma_n(z - (A + \delta \boldsymbol{M}_n)) < \epsilon] \, dz \\
&\leq \frac{\sqrt{2} K n^2}{\delta} \operatorname{Leb}_{\mathbb{R}}(\Omega). && \text{Corollary 4.21}
\end{aligned}$$

To obtain the improvement in the Ginibre case, in the final inequality we use the bound

$$\mathbb{P}[\sigma_n(z - (A + \delta \boldsymbol{H}_n)) \leq \epsilon] \leq \frac{n\epsilon}{\delta}$$

instead, from Theorem 4.4G. □

We now give the analogous proposition for the nonreal eigenvalues.

Proposition 4.26 (Condition Numbers of Eigenvalues off the Real Line)**.** *Let $n \geq 9$. Let $A \in \mathbb{R}^{n \times n}$ be deterministic. Let $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. Let $\delta > 0$, and write $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \delta \boldsymbol{M}_n$. Then for every open set $\Omega \subseteq \mathbb{C} \setminus \mathbb{R}$,*

$$\mathbb{E} \sum_{i : \boldsymbol{\lambda}_i \in \Omega} \kappa_i^2(A + \delta \boldsymbol{M}_n)^2 \leq \frac{C_{4.5} K^3 n^5}{\delta^3} \int_\Omega \frac{(\delta \mathbb{E}\|\boldsymbol{M}_n\| + \|A\| + |\Re z|)^2 + |\Im z|^2}{|\Im z|} \, dz.$$

In the real Ginibre case, one may take $n \geq 7$ and replace the term $C_{4.5} K^3$ with $\frac{\sqrt{7e}}{4\pi}$.

Proof. In the proof of Theorem 4.25, since $\Omega \subseteq \mathbb{C} \setminus \mathbb{R}$ we replace Lemma 4.3 with Lemma 2.4. Since z is no longer real we must also replace the singular value tail bound in Corollary 4.21 with the one in Theorem 4.5 (or the one in Theorem 4.5G, for the Ginibre case). □

Bounds with High Probability: Proofs of Theorems 4.7 and 4.7G

We now prove the main theorem of this section, which implies that all eigenvalue condition numbers are bounded by $\operatorname{poly}(n/\delta)$ with probability $1 - 1/\operatorname{poly}(n)$. In the notation of the theorem below, $R, \|A\|, K$, and δ will be $\Theta(1)$ in most applications, so ϵ_1 and ϵ_2 may be set to $1/n^D$ for sufficiently high D.

Restatement of Theorem 4.7. *Let $n \geq 9$. Let $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{M}_n$ satisfy Assumption 4.1 with parameter $K > 0$. Let $0 < \delta < K\min\{1, \|A\| + R\}$, and write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \delta\boldsymbol{M}_n$. Let $R > \mathbb{E}\|\delta\boldsymbol{M}_n\|$. Then for any $\epsilon_1, \epsilon_2 > 0$, with probability at least $1 - 2\epsilon_1 - O\left(\frac{R(R+\|A\|)^{3/5}K^{8/5}n^{14/5}\epsilon_2^{3/5}}{\delta^{8/5}}\right) - 2\mathbb{P}[\delta\|\boldsymbol{M}_n\| > R]$ we have*

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{R}} \kappa_i(A+\delta\boldsymbol{M}_n) \leq \epsilon_1^{-1}C_{4.7}Kn^2\frac{\|A\|+R}{\delta},$$

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}} \kappa_i^2(A+\delta\boldsymbol{M}_n) \leq \epsilon_1^{-1}\log(1/\epsilon_2)C_{4.7}K^3n^5\cdot\frac{(\|A\|+R)^3}{\delta^3}, \qquad \textit{and}$$

$$\kappa_V(A+\delta\boldsymbol{M}_n) \leq \epsilon_1^{-1}\sqrt{\log(1/\epsilon_2)}C_{4.7}K^{3/2}n^3\cdot\frac{(\|A\|+R)^{3/2}}{\delta^{3/2}},$$

for some universal constant $C_{4.7} > 0$.

Proof. Going forward, assume that each of $\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{R}}\kappa_i(A+\delta\boldsymbol{M}_n)$ and $\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}}\kappa_i^2(A+\delta\boldsymbol{M}_n)$ 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 $\zeta \in (0, R)$ be a small parameter to be optimized later. Let $L \triangleq \|A\| + R$, and define the regions $\Omega_\mathbb{R}$ and $\Omega_\mathbb{C}$ as follows:

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

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

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{R}}\kappa_i(A+\delta\boldsymbol{M}_n) = \sum_{i:\boldsymbol{\lambda}_i\in\Omega_\mathbb{R}}\kappa_i(A+\delta\boldsymbol{M}_n) \leq \frac{Kn^2}{\sqrt{2}\delta}\operatorname{Leb}_\mathbb{R}(\Omega_\mathbb{R}) \leq \frac{\sqrt{2}Kn^2L}{\delta}$$

and

$$\begin{aligned}\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}}\kappa_i^2(A+\delta\boldsymbol{M}_n) &= \sum_{i:\boldsymbol{\lambda}_i\in\Omega_\mathbb{C}}\kappa_i^2(A+\delta\boldsymbol{M}_n)\\ &\leq \frac{C_{4.5}K^3n^5}{\delta^3}\int_{\Omega_\mathbb{C}}\frac{(\delta\mathbb{E}\|M_n\|+\|A\|+|\Re z|)^2+|\Im z|^2}{|\Im z|}\,\mathrm{d}z\\ &\leq 2\frac{CK^3n^5}{\delta^3}\int_\zeta^L\int_{-L}^L\frac{(\delta\mathbb{E}\|M_n\|+\|A\|+|x|)^2+y^2}{y}\,\mathrm{d}x\,\mathrm{d}y\\ &\leq 2\frac{C_{4.5}K^3n^5}{\delta^3}\int_\zeta^L 2L\frac{(2L)^2+L^2}{y}\,\mathrm{d}y\\ &= 20\frac{C_{4.5}K^3n^5}{\delta^3}L^3(\log L+\log(1/\zeta)).\end{aligned}$$

Finally recall from (4.47) that

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

so setting $\zeta = L\epsilon_2$ yields the result.

To obtain the bound on κ_V, use Lemma 2.2. □

In the special case of Ginibre matrices, we will endeavor to give an explicit bound on the constant factors appearing in the proof of Theorem 4.7 without being too wasteful. We also save one factor of n in the bound for real eigenvalues in comparison to Theorem 4.7.

Restatement of Theorem 4.7G. *Let $n \geq 7$. Let $A \in \mathbb{R}^{n\times n}$ be deterministic, and let $\boldsymbol{H}_n$ be a real Ginibre matrix. Let $0 < \delta < \min\{1, \|A\|\}$, and write $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ for the eigenvalues of $A + \delta\boldsymbol{H}_n$. Then for any $\epsilon_1, \epsilon_2 > 0$, with probability at least $1 - 2\epsilon_1 - \frac{30\|A\|^{8/5}n^{8/5}}{\delta^{8/5}}\epsilon_2^{3/5} - 2e^{-2n}$ we have*

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{R}} \kappa_i(A + \delta\boldsymbol{H}_n) \leq 5\epsilon_1^{-1}n\frac{\|A\|}{\delta},$$

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}} \kappa_i^2(A + \delta\boldsymbol{H}_n) \leq 1000\epsilon_1^{-1}\log(1/\epsilon_2)\frac{n^5\|A\|^3}{\delta^3}, \qquad \textit{and}$$

$$\kappa_V(A + \delta\boldsymbol{M}_n) \leq 1000\epsilon_1^{-1}\sqrt{\log(1/\epsilon_2)}\frac{n^3\|A\|^{3/2}}{\delta^{3/2}}.$$

Proof. We identify the necessary modifications to the proof of Theorem 4.7. First, set $R = 4\delta$, so that $\mathbb{P}[\delta\|\boldsymbol{H}_n\| > R] < e^{-2n}$. The statement for real eigenvalues is then immediate, using the improvement for Ginibre matrices in Proposition 4.25.

Now we proceed to the bound for the nonreal eigenvalues. Take $\zeta = \epsilon_2\|A\|$, so that by (4.53) we have

$$\mathbb{P}[E_{\text{strip}}] \leq 6(\|A\| + 4\delta)\frac{n^{8/5}\|A\|^{3/5}\epsilon_2^{3/5}}{\delta^{8/5}} \leq \frac{30n^{8/5}\|A\|^{8/5}\epsilon_2^{3/5}}{\delta^{8/5}},$$

where we use $\delta < \|A\|$. Recall $\mathbb{E}\|\boldsymbol{H}_n\| \leq 2$ (see [8]). Replacing $C_{4.5}K^3$ with $\frac{\sqrt{7e}}{4\pi}$ as indicated in Proposition 4.26, and computing the integral

$$\begin{aligned}\int_\zeta^L \int_{-L}^L \frac{(L+|x|)^2 + y^2}{|y|}\,\mathrm{d}x\,\mathrm{d}y &= \frac{14}{3}L^3(\log L + \log(1/\zeta)) + L^3 - L\zeta^2 \\ &\leq \frac{14}{3}L^3(\log L + \log(1/\epsilon_2) - \log\|A\|) + L^3,\end{aligned}$$

one obtains

$$\sum_{i:\boldsymbol{\lambda}_i\in\mathbb{C}\setminus\mathbb{R}} \kappa_i^2(A + \delta\boldsymbol{H}_n) \leq \frac{7\sqrt{7e}}{6\pi\delta^3}n^5(\|A\| + 4\delta)^3(\log(\|A\| + 4\delta) + \log(1/\epsilon_2) - \log\|A\| + 3/14).$$

Using $\delta < \|A\|$ and cleaning up the constants, we arrive at the form in the theorem statement. □

4.8 Further Questions

There are a few natural directions to pursue. For instance, what can be said about the eigenvalue condition numbers for random matrices without continuous entries? Solving this question would require essentially different ideas from those presented in this chapter. More concretely, our proof technique requires

$$\lim_{\epsilon \longrightarrow 0} \mathbb{P}[\sigma_n(z - (A + \boldsymbol{M}_n)) \leq \epsilon] = 0,$$

and this may no longer hold if the distributions of the entries of $\boldsymbol{M}_n$ are allowed to be discrete. A natural starting point is the case of i.i.d. ±1 entries:

Problem 4.27. *Let $\boldsymbol{M}_n$ be a matrix with independent Rademacher entries. For which deterministic matrices A and which $\delta > 0$ does it hold, with high probability, that $\kappa_V(A + \delta \boldsymbol{M}_n) = O(n^C)$ for some $C > 0$?*

With regards to the least singular value of complex shifts of real ensembles, we posit the following possible improvement to Theorem 4.5G in the dependence on n:

Conjecture 4.28. *Let $\boldsymbol{H}_n$ be an $n \times n$ real Ginibre matrix. Then, for any constant $C > 0$ there exists a constant C' (depending on C only) such that for any $\epsilon > 0$ and $z \in \mathbb{C} \setminus \mathbb{R}$ with $|z| \leq C$ it holds that*

$$\mathbb{P}\left[\sigma_n(z - \boldsymbol{H}_n) \leq \epsilon\right] \leq \frac{C' n^2 \epsilon^2}{|\Im z|}. \tag{4.54}$$

Actually, we believe that a stronger conjecture is true. Namely, the bound in (4.54) should hold even when $\boldsymbol{H}_n$ is substituted by $A + \boldsymbol{H}_n$, where $A \in \mathbb{R}^{n \times n}$ is deterministic. In this case C' is also allowed to depend on $\|A\|$.

Our next conjecture is that Szarek's bound for singular values of real Ginibre matrices in Theorem 4.13 holds, up to the value of the universal constant C, in the more general setting of matrices satisfying Assumption 4.1. This would constitute an improvement of Theorem 4.4 in the dependence on k and n.

Conjecture 4.29. *Let $\boldsymbol{M}_n$ be a real random matrix satisfying Assumption 4.1 with parameter $K > 0$ and perhaps with some moment assumptions on its entries. Then, there is a universal constant C such that for any deterministic $A \in \mathbb{R}^{n \times n}$, it holds that*

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

It is worth noting that in some sense Conjecture 4.29 is known to be true when $k = 1$. This was proven by Tikhomirov in [150] under weaker assumptions on the independence of the entries of $\boldsymbol{M}_n$ and with a mild technical assumption about the decay of their densities.

Bibliographic Note

This chapter is primarily drawn from [16], and the material appears largely as it did there, other than some minor adaptation of the introductory section. The only exception is Theorem 4.10, the proof and statement of which appeared originally as [19, Theorem 2.4].

Chapter 5

Diagonalization by Spectral Bisection

5.1 Introduction

We now begin the second main focus of this thesis: studying the algorithmic problem of approximately finding all of the eigenvalues and eigenvectors of a given arbitrary $n \times n$ complex matrix. While this problem is quite well-understood in the special case of Hermitian matrices (see, e.g., [127]), the general non-Hermitian case has remained mysterious from a theoretical standpoint even after several decades of research. In particular, the previously best known *provable* algorithms for this problem run in time $O(n^{10}/\delta^2)$ [5] or $O(n^c \log(1/\delta))$ [45] with $c \geq 12$ where $\delta > 0$ is the desired accuracy, depending on the model of computation and notion of approximation considered.[1] To be sure, the non-Hermitian case is well-motivated: coupled systems of differential equations, linear dynamical systems in control theory, transfer operators in mathematical physics, and the nonbacktracking matrix in spectral graph theory are but a few situations where finding the eigenvalues *and eigenvectors* of a non-Hermitian matrix is important.

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

In this chapter, we overcome these difficulties by way of the Gaussian regularization results in Chapter 3: adding a small complex Gaussian perturbation to any matrix typically yields a matrix with well-conditioned eigenvectors and a large minimum gap between the eigenvalues, implying spectral stability. We complement the above by proving that a variant of the well-known spectral

[1] A detailed discussion of these and other related results appears in Section 5.2.

bisection algorithm in numerical linear algebra [30] is both fast and numerically stable when run on a matrix whose eigenvector condition number and minimum gap are suitably controlled — we call an iterative algorithm *numerically stable* if it can be implemented using finite precision arithmetic with polylogarithmically many bits, corresponding to a dynamical system whose trajectory to the approximate solution is robust to adversarial noise (see, e.g. [141]).

The main result of this chapter, Theorem 1.7 from Chapter 1, is that the spectral bisection algorithm in finite arithmetic can be reduced to a *polylogarithmic* (in the desired accuracy and dimension n) number of invocations of standard numerical linear algebra routines (multiplication, inversion, and QR factorization), each of which is reducible to matrix multiplication [61], yielding a nearly matrix multiplication runtime for the whole algorithm. This improves on the previously best known running time in the Hermitian case (which is $O(n^{\omega+1}\text{polylog}(n))$ bit operations in the setting $\delta = 1/\operatorname{poly}(n)$ [31]) and yields the same improvement for the related problem of computing the singular value decomposition of a matrix.

Matrix Sign Function

The key step in the bisection algorithm is computing the *sign function* of a matrix, a problem of independent interest in many areas such including control theory and approximation theory [100]. Recall from Chapter 1 that the sign function of a matrix $A \in \mathbb{C}^{n\times n}$ is

$$\operatorname{sgn}(A) \triangleq P_+ - P_-,$$

where $P_\pm$ are the spectral projectors onto the invariant subspaces associated to the eigenvalues in the right and left halfplanes of $\mathbb{C}$, respectively. The sign function is undefined for matrices with eigenvalues on the imaginary axis. Quantifying this discontinuity, Bai and Demmel [10] defined the following condition number for the sign function:

$$\kappa_{\text{sgn}}(M) \triangleq \inf \left\{ 1/\epsilon^2 \;:\; \Lambda_\epsilon(M) \text{ does not intersect the imaginary axis} \right\}, \tag{5.1}$$

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

Roberts [133] showed that the simple iteration

$$A_{k+1} = \frac{1}{2}\left(A_k + A_k^{-1}\right) \tag{5.2}$$

converges globally and quadratically to $\operatorname{sgn}(A)$ in exact arithmetic, but his proof relies on the fact that all iterates of the algorithm are simultaneously diagonalizable, a property which is destroyed in finite arithmetic since inversions can only be done approximately.[2] In Section 5.6 we show that this iteration is indeed convergent when implemented in finite arithmetic for matrices with small κ_{sgn}, given a numerically stable matrix inversion algorithm. This leads to the following result:

[2] 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 5.1 (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 $K \geq \kappa_{\mathrm{sgn}}(A)$, 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

$$O((\log K + \log\log(1/\beta))T_{\mathsf{INV}}(n)) \tag{5.3}$$

arithmetic operations on a floating point machine with

$$O(\log n \log^3 K(\log K + \log(1/\beta)))$$

bits of precision, where $T_{\mathsf{INV}}(n)$ denotes the number of arithmetic operations used by a numerically stable matrix inversion algorithm (satisfying Definition 5.5).

The key idea in the proof of Theorem 5.1 is to control the evolution of the pseudospectra $\Lambda_{\epsilon_k}(A_k)$ of the iterates with appropriately decreasing (in k) parameters ϵ_k, using a sequence of carefully chosen shrinking contour integrals in the complex plane. The pseudospectrum provides a richer induction hypothesis than scalar quantities such as condition numbers, and allows one to control all quantities of interest using the holomorphic functional calculus. This technique is introduced in Section 5.6, yielding Theorem 5.1.

Diagonalization by Spectral Bisection

Given an algorithm for computing the sign function, there is a natural and well-known approach to the eigenproblem pioneered in [30]. The idea is that the matrices $(1 \pm \mathrm{sgn}(A))/2$ are spectral projectors onto the invariant subspaces corresponding to the eigenvalues of A in the left and right open half planes, so if some shifted matrix $z + A$ or $z + iA$ has roughly half its eigenvalues in each half plane, the problem can be reduced to smaller subproblems appropriate for recursion.

The two difficulties in carrying out the above approach are: (a) efficiently computing the sign function (b) finding a balanced splitting along an axis that is well-separated from the spectrum. These are nontrivial even in exact arithmetic, since the iteration (5.2) converges slowly if (b) is not satisfied, even without roundoff error. We use Theorem 1.10 to ensure that a good splitting always exists after a small Gaussian perturbation of order δ, and Theorem 5.1 to compute splittings efficiently in finite precision. Combining this with well-understood techniques such as rank-revealing QR factorization, we obtain the our main theorem, whose proof appears in Section 5.5.

Restatement of Theorem 1.7. *There is a randomized algorithm* EIG *which on input any matrix $A \in \mathbb{C}^{n\times n}$ with $\|A\| \leq 1$ and a desired accuracy parameter $\delta > 0$ outputs a diagonal D and invertible V such that*

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

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 $1 - 12/n$. *Here* $T_{\mathsf{MM}}(n)$ *refers to the running time of a numerically stable matrix multiplication algorithm (detailed in Section 5.3).*

Remark 5.2 (Accuracy vs. Precision)**.** The gold standard of "backward stability" in numerical analysis postulates that

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

i.e., the number of bits of precision is linear in the number of bits of accuracy. The relaxed notion of "logarithmic stability" introduced in [62] requires

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

for some constant c, where κ is an appropriate condition number. In comparison, Theorem 1.7 obtains the weaker relationship

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

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

5.2 Related Work

Smoothed Analysis and Free Probability. The study of numerical algorithms on Gaussian random matrices (i.e., the case $A = 0$ of smoothed analysis) dates back to [155, 140, 63, 70]. 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 [143], in the context of the simplex algorithm. Sankar, Spielman, and Teng [136] showed that adding real Gaussian noise to any matrix yields a matrix with polynomially-bounded condition number. The main difference between our results and most of the results on smoothed analysis (including [5]) 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 [142] and Haagerup and Larsen [92] 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 [29, 30, 65] and Roberts [133], who used it first to solve the algebraic Ricatti and Lyapunov equations and then as an approach to the eigenproblem; see [100] for a broad survey of its early history. The numerical stability of Roberts' Newton iteration was investigated by Byers [42], who identified some cases where it is and isn't stable. Malyshev [113], Byers, He, and Mehrmann [43], Bai, Demmel, and Gu [11], and Bai and Demmel [10] 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,[3] but did not prove convergence in finite arithmetic and left this as an open question.[4] 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. −[95, Section 5.7]

This is precisely the problem solved by Theorem 5.1, 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 [93] established efficient reductions between the sign function and the polar decomposition. Byers and Xu [44] 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 [119] (see also the improvement [118]) 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. The eigenproblem has been thoroughly studied in the numerical analysis community, in the floating point model of computation. While there are provably fast and accurate algorithms in the Hermitian case (see the next subsection) and a large body of work for various structured matrices (see, e.g., [34]), the general case is not nearly as well-understood. As recently as 1997, J. Demmel remarked in his well-known textbook [64]: "... the problem of

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

[4] [43] states: "A priori backward and forward error bounds for evaluation of the matrix sign function remain elusive."

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 [5]. They exhibited an algorithm (see their Theorem 2.28) which given any $A \in \mathbb{C}^{n\times n}$ with $\|A\| \le 1$ and $\sigma > 0$ produces in $O(n^9/\sigma^2)$ expected arithmetic operations the diagonalization of the nearby random perturbation $A + \sigma G$ where G is a matrix with standard complex Gaussian entries. By setting σ sufficiently small, this may be viewed as a backward approximation algorithm for diagonalization, in that it solves a nearby problem essentially exactly[5] – in particular, by setting $\sigma = \delta/\sqrt{n}$ and noting that $\|G\| = O(\sqrt{n})$ with very high probability, their result implies a running time of $O(n^{10}/\delta^2)$ in our setting. 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.

Result	Error	Arithmetic Ops	Boolean Ops	Restrictions
[5][a]	Backward	n^{10}/δ^2	$n^{10}/\delta^2 \cdot \mathrm{polylog}(n/\delta)$	
[31][b]	Backward	$n^{\omega+1}\mathrm{polylog}(n)\log(1/\delta)$	$n^{\omega+1}\mathrm{polylog}(n)\log(1/\delta)$	Hermitian
Theorem 1.7 [c]	Backward	$T_{\mathsf{MM}}(n)\log^2(n/\delta)$	$T_{\mathsf{MM}}(n)\log^6(n/\delta)\log(n)$	

[a] Does not specify a particular bound on precision.

[b] Bounds circuit complexity in exact arithmetic and argues informally that the algorithm is stable if all operations are rounded to $O(\log(1/\delta))$ bits.

[b] $T_{\mathsf{MM}}(n) = O(n^{\omega+\eta})$ for every $\eta > 0$, see Definition 5.4 for details.

Table 5.1: Results for finite-precision floating-point arithmetic

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

[5] The output of their algorithm is n vectors on each of which Newton's method converges quadratically to an eigenvector, which they refer to as "approximation à la Smale".

Result	Model	Error	Arithmetic Ops	Boolean Ops	Restrictions
[45]	Rational	Forward[a]	$\mathrm{poly}(a, n, \log(1/\delta))$[b]	$\mathrm{poly}(a, n, \log(1/\delta))$	
[124]	Rational	Forward	$n^\omega + n \log\log(1/\delta)$	$n^{\omega+1}a + n^2 \log(1/\delta) \log\log(1/\delta)$	Eigs only[c]
[111]	Finite[c]	Forward	$n^\omega \log(n) \log(1/\delta)$	$n^\omega \log^4(n) \log^2(n/\delta)$	λ_1 of Herm.

[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 5.2: Results for other models of arithmetic

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 [124] 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.[6]

Finally, the important work of Demmel, Dumitriu, and Holtz [61] (see also the followup [13]), 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 better understood. We cannot give a complete bibliography of this huge area, but mention one relevant landmark result: the work of Wilkinson [164], who exhibited a globally convergent shifting strategy for the QR algorithm, and the work of Dekker and Traub [60] who quantified the rate of convergence of Wilkinson's shift. We refer the reader to [127, §8.10] for the simplest and most insightful proof of this result, due to Hoffman and Parlett [96]. However, the above convergence proofs are all in exact arithmetic, and we are not aware of rigorous analysis of Wilkinson's shift (or any other) in finite arithmetic.

[6] "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." – [124, Section 12].

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 [111] show how to find a δ−approximation of just the largest eigenvalue in $O(n^\omega \log^4(n) \log^2(1/\delta))$ bit operations, and Ben-Or and Eldar [31] give an $O(n^{\omega+1}\mathrm{polylog}(n))$-bit-operation algorithm for finding a $1/\operatorname{poly}(n)$-approximate diagonalization of an $n \times n$ Hermitian matrix normalized to have $\|A\| \leq 1$.

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

5.3 Finite Arithmetic Assumptions

We begin by briefly elaborating on the axioms for floating-point arithmetic given in Chapter 2. Similar guarantees to the ones appearing in that section for scalar-scalar operations also hold for operations such as matrix-matrix addition and matrix-scalar multiplication. In particular, if A is an $n \times n$ complex matrix,

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

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

$$\|A \circ \Delta\| \leq \sqrt{n} \max_i \|(A \circ \Delta)e_i\| \leq \sqrt{n} \max_{i,j} |\Delta_{i,j}| \max_i \|Ae_i\| \leq \mathbf{u}\sqrt{n}\|A\|. \tag{5.4}$$

For more complicated operations such as matrix-matrix multiplication and matrix inversion, we use existing error guarantees from the literature; this is discussed further below.

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

$$|\mathsf{fl}\,(\operatorname{tr} A) - \operatorname{tr} A| \leq n\|A\|\mathbf{u} \tag{5.5}$$

$$\|\mathsf{fl}(\hat{x}) - \hat{x}\| \leq n\mathbf{u}. \tag{5.6}$$

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

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

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 5.3 (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 this thesis, we will omit issues of underflow or overflow. To simplify some of our bounds, we will also assume that $c_{\mathsf{N}} \geq 1$.

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

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

Our definitions are inspired by the definition of *logarithmic stability* introduced in [61]. 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 5.4. 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 5.5. 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 5.6. 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 (i) R is exactly upper triangular, and (ii) there is a unitary $\widetilde{Q}$ and a matrix $\widetilde{A}$ such that $\widetilde{Q}\widetilde{A} = R$ 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 5.7. 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 $c_{\mathsf{INV}} = 1$ [94]. 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 [61]. Specifically, the following parameters are known to be achievable.

Theorem 5.8 (Fast and Stable Instantiations of $\mathsf{MM}, \mathsf{INV}, \mathsf{QR}$).

(i) 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.

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

(iii) *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 $O(T_{\mathsf{MM}}(n))$ *for an* $n \times n$ *matrix.*

Proof. The first assertion is Theorem 3.3 of [62]. The second is Theorem 3.3 (see also equation (9) above its statement) of [61]. 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 [61]. □

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

5.4 Pseudospectral Shattering

Our main Gaussian regularization result, Theorem 1.10, ensures that a small random perturbation with high probability *shatters* the ϵ-pseudospectrum of any $n \times n$ matrix into disjoint connected components, for some modest ϵ. The virtue of the shattering property is after any *further* perturbation of size at most ϵ, each eigenvalue of the perturbed matrix will remain in one of these connected components. The following key definitions make this phenomenon quantitative in a sense which is useful for our analysis of spectral bisection.

Definition 5.9 (Grid)**.** A *grid* in the complex plane consists of the boundaries of a lattice of squares with lower edges parallel to the real axis. We will write

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

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

Definition 5.10 (Shattering)**.** A pseudospectrum $\Lambda_\epsilon(A)$ is *shattered* with respect to a grid g if every square of g has at most one eigenvalue of A and $\Lambda_\epsilon(A) \cap \mathsf{g} = \emptyset$.

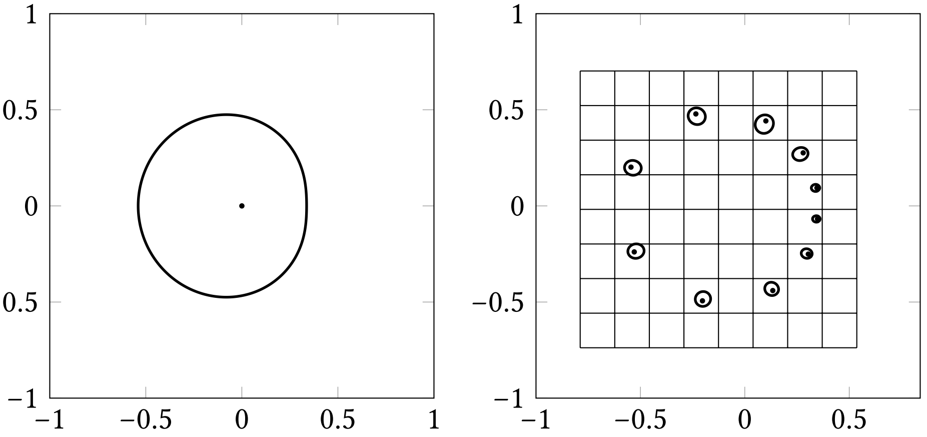

Figure 5.1: The numerical example from Figure 1.2 is pictured again; on the left is the 10^{-6}-pseudospectrum of a non-diagonalizable Toeplitz matrix T, and on the right is the same pseudospectrum of $T + 10^{-6}G_n$, shattered with respect to the shown grid.

Observation 5.11. As $\Lambda_\epsilon(A)$ contains a ball of radius ϵ about each eigenvalue of A, shattering of the ϵ-pseudospectrum with respect to a grid with side length ω implies $\epsilon \leq \omega/2$.

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

Lemma 5.12. *If $\lambda_1, \dots, \lambda_n$ 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 $\kappa_i(A) \leq \frac{2\omega}{\pi\epsilon}$.*

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

$$\kappa_i(A) = \|vw^*\| = \left\| \frac{1}{2\pi i} \oint_\Gamma (z - A)^{-1}\, \mathrm{d}z \right\| \leq \frac{1}{2\pi} \oint_\Gamma \left\|(z - A)^{-1}\right\| \,\mathrm{d}z \leq \frac{2\omega}{\pi\epsilon}. \tag{5.7}$$

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

The theorem below quantifies the extent to which perturbing by a Ginibre matrix results in a shattered pseudospectrum. See Figure 5.1 for an illustration in the case where the initial matrix is poorly conditioned. In general, not all eigenvalues need move so far upon such a perturbation, in particular if the respective κ_i are small.

Theorem 5.13 (Exact Arithmetic Shattering)**.** *Let $A \in \mathbb{C}^{n\times n}$ and $A + \delta G_n$ for G_n a complex Ginibre matrix. Assume $\|A\| \le 1$ and $0 < \delta < 1/2$. Let $\mathsf{g} \triangleq \mathsf{grid}(z, \omega, \lceil 8/\omega\rceil, \lceil 8/\omega\rceil)$ with $\omega \triangleq \frac{\delta}{4n^2}$, and z chosen uniformly at random from the square of side ω cornered at $-4-4i$. Then, $\kappa_V(A+\delta G_n) \le n^2/\delta$, $\|\delta G_n\| \le 4\delta$, and $\Lambda_\epsilon(A+\delta G_n)$ is shattered with respect to g for*

$$\epsilon \triangleq \frac{\delta^2}{16n^6},$$

with probability at least $1 - 11/n$.

Proof. Condition on the event in Theorem 1.10, so that

$$\kappa_V(A+\delta G_n) \le \frac{n^2}{\delta}, \quad \|\delta G_n\| \le 4\delta, \quad \text{and } \operatorname{gap}(A+\delta G_n) \ge \frac{\delta}{n^2} = 4\omega.$$

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

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

Let $\lambda_1, \ldots \lambda_n$ be the eigenvalues of $A+\delta G_n$). We now have for each λ_i and every $s < \frac{\omega}{2}$:

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

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

$$\mathbb{P}[\min_{i\le n} \operatorname{dist}_{\mathsf{g}}(\lambda_i) > \omega/4n^2] > 1 - 1/n, \tag{5.8}$$

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

We now recall the Bauer-Fike theorem that

$$\Lambda_\epsilon(A+\delta G_n) \subset \bigcup_{i\le n} \mathbb{D}(\lambda_i, \epsilon\kappa_V(A+\delta G_n)),$$

Thus, if both (5.8) and the event from Theorem 1.10 hold, we see that $\Lambda_\epsilon(A+\delta G_n)$ is shattered with respect to g as long as

$$\kappa_V(A+\delta G_n)\epsilon < \frac{\omega}{4n^2},$$

which is implied by

$$\epsilon < \frac{\delta}{4n^2} \cdot \frac{1}{4n^2} \cdot \frac{\delta}{n^2} = \frac{\delta^2}{16n^6}.$$

Thus, the advertised claim holds with probability at least $1 - 1/n - 10/n = 1 - 11/n$ as desired. as desired. □

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

```
SHATTER

Input: Matrix $A \in \mathbb{C}^{n\times n}$, Gaussian perturbation size $\delta \in (0, 1/2)$.
Requires: $\|A\| \le 1$.
Output: Matrix $X \in \mathbb{C}^{n\times n}$, grid $\mathsf{g}$, shattering parameter $\epsilon > 0$.
Ensures: $\|X - A\| \le 4\delta$, $\kappa_V(X) \le n^2/\delta$, and $\Lambda_\epsilon(X)$ is shattered with respect to $\mathsf{g}$, with probability at least $1 - 11/n$.

    1. $G_{ij} \longleftarrow \mathsf{N}(1/n)$ for $i, j = 1, \ldots, n$.
    2. $X \longleftarrow A + \delta G + E$.
    3. Let $\mathsf{g}$ be a random grid with $\omega = \frac{\delta}{4n^2}$ and bottom left corner $z$ chosen as in Theorem 5.13.
    4. $\epsilon \longleftarrow \frac{1}{2} \cdot \frac{\delta^2}{16n^6}$
```

Theorem 5.14 (Finite Arithmetic Shattering)**.** *Assume there is a c_{N}-stable Gaussian sampling algorithm N satisfying the requirements of Definition 5.3. Then* SHATTER *has the advertised guarantees as long as the machine precision satisfies*

$$\mathbf{u} \le \frac{1}{2}\frac{\delta^2}{16n^6} \cdot \frac{1}{(3 + c_{\mathsf{N}})\sqrt{n}}, \tag{5.9}$$

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 $n \cdot c_{\mathsf{N}} \cdot (1/\sqrt{n}) \cdot \mathbf{u} \le c_{\mathsf{N}}\sqrt{n} \cdot \mathbf{u}$ from N, by Definition 5.3.
2. An additive error of norm at most $\sqrt{n} \cdot \|X\| \cdot \mathbf{u} \le 3\sqrt{n}\mathbf{u}$, with probability at least $1 - 1/n$, from the roundoff E in step 2.

Thus, as long as the precision satisfies (5.9), we have

$$\|\mathsf{SHATTER}(A, \delta) - \mathsf{shatter}(A, \delta)\| \le \frac{1}{2}\frac{\delta^2}{16n^6},$$

where $\mathsf{shatter}(\cdot)$ refers to the (exact arithmetic) outcome of Theorem 5.13. The correctness of SHATTER now follows from the stability of pseudospectrum under perturbations. Its running time is bounded by

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

arithmetic operations, as advertised. □

5.5 The Spectral Bisection Algorithm

In this section we will prove Theorem 1.7, deferring analysis of its subroutines to the later sections. As discussed in Section 5.1, our algorithm is not new, and in its idealized form it reduces to the two following tasks:

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

Span: 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 *deflate* A to the two matrices $A_\pm \triangleq Q_\pm^* A Q_\pm$, and continue recursively. The observation below verifies that this recursion is sound.

Observation 5.15. The spectrum of A is exactly $\Lambda(A_+) \sqcup \Lambda(A_-)$, and every eigenvector of A is of the form $Q_\pm v$ for some eigenvector v of one of $A_\pm$.

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

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

Let us now sketch the implementations and state carefully the guarantees for SPLIT and SPAN; the analysis of these will be deferred to Sections 5.7 and 5.8. Our splitting algorithm is presented a matrix A whose ϵ-pseudospectrum is shattered with respect to a grid g. For any vertical grid line with real part h, $\operatorname{tr}\ \operatorname{sgn}(A-h)$ gives the difference between the number of eigenvalues lying to its left and right. As

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

we can determine these eigenvalue counts *exactly* by running SGN to accuracy $O(1/n)$ and rounding $\operatorname{tr}\ \mathsf{SGN}(A-h)$ to the nearest integer. We will show in Section 5.7 that, by mounting a binary search over horizontal and vertical lines of g, we will always arrive at a partition of the eigenvalues into two parts with size at least $\min\{n/5, 1\}$. Having found it, we run SGN one final time at the desired precision to find the approximate spectral projectors.

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 $\beta \le 0.05/n$
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.

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

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

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

3. Once a shift is found,

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

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

Theorem 5.16 (Guarantees for SPLIT). *Assume* INV *is a* $(\mu_{\mathsf{INV}}, c_{\mathsf{INV}})$*-stable matrix inversion algorithm satisfying Definition 5.5. Let* $\epsilon \le 0.5$*,* $\beta \le 0.05/n$*, and* $\|A\| \le 4$ *and* g *have side lengths of at most*

8, and define

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

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

$$\mathbf{u} \le \mathbf{u}_{\mathsf{SPLIT}} \triangleq \min\left\{ \frac{\left(1-\frac{\epsilon}{256}\right)^{2^{N_{\mathsf{SPLIT}}+1}(c_{\mathsf{INV}}\log n+3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{SPLIT}}}, \frac{\epsilon}{100n}, \frac{\epsilon^2}{512} \right\},$$

using at most

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

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 [89], or probabilistically in matrix-multiplication time with a variant of the method of [61]; we will use the latter.

SPAN

Input: Matrix $\widetilde{P} \in \mathbb{C}^{n\times n}$, desired rank k, input precision β, and desired accuracy η
Requires: $\|\widetilde{P} - P\| \le \beta \le \frac{1}{4}$ for some rank-k projector P.
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 $\|\widetilde{Q} - Q\| \le \eta$, with probability at least $1 - \frac{(20n)^3\sqrt{\beta}}{\eta^2}$.

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

Theorem 5.17 (Guarantees for SPAN)**.** *Assume* MM *and* QR *are matrix multiplication and QR factorization algorithms satisfying Definitions 5.4 and 5.6. Then* SPAN *has the advertised guarantees when run on a machine with precision:*

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

The number of arithmetic operations is at most:

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

Remark 5.18. The proof of the above theorem, which is deferred to Section 5.8, closely follows and builds on the analysis of the randomized rank revealing factorization algorithm (RURV) introduced in [61] and further studied in [14]. 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 [61] and [14] 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 5.55) 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 [14]. 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, span, and deflate framework that began this section.

Theorem 5.19 (EIG: Finite Arithmetic Guarantee)**.** *Assume* MM, QR, *and* INV *are numerically stable algorithms for matrix multiplication, QR factorization, and inversion satisfying Definitions 5.4, 5.6, and 5.5. 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*

$$N_{\mathsf{EIG}} \triangleq \lg\frac{256n}{\epsilon} + 3\lg\lg\frac{256n}{\epsilon} + \lg\lg\frac{(5n)^{26}}{\theta^2\delta^4\epsilon^9} + 7.59.$$

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

$$\begin{aligned}\mathbf{u} &\leq 2^{-\max\left\{\lg^3 \frac{n}{\epsilon}\lg\left(\frac{(5n)^{26}}{\theta^2\delta^4\epsilon^8}\right)2^{9.59}(c_{\mathsf{INV}}\log n+3)+\lg N_{\mathsf{EIG}},\lg\frac{(5n)^{30}}{\theta^2\delta^4\epsilon^8}+\lg\max\{\mu_{\mathsf{MM}}(n),\mu_{\mathsf{QR}}(n),n\}\right\}}\\ &\leq 2^{-O\left(\log^3\frac{n}{\epsilon}\log\frac{n}{\theta\delta\epsilon}\log n\right)}.\end{aligned}$$

The number of arithmetic operations is at most

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

EIG

Input: Matrix $A \in \mathbb{C}^{m\times m}$, desired eigenvector accuracy δ, grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$, pseudospectral guarantee ϵ, acceptable failure probability θ, and global instance size n
Requires: $\Lambda_\epsilon(A)$ is shattered with respect to g, and $m \le n$.
Output: Eigenvectors and eigenvalues $(\widetilde{V}, \widetilde{D})$
Ensures: With probability at least $1-\theta$, each entry $\widetilde{\lambda_i} = \widetilde{D}_{i,i}$ lies in the same square as exactly one eigenvalue $\lambda_i \in \Lambda(A)$, and each column $\tilde{v}_i$ of $\widetilde{V}$ has norm $1 \pm n\mathbf{u}$, and satisfies $\|\tilde{v}_i - v_i\| \le \delta$ for some exact unit right eigenvector $Av_i = \lambda_i v_i$.

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

Remark 5.20. We have not fully optimized the large constant $2^{9.59}$ appearing in the bit length above.

Theorem 5.19 easily implies Theorem 1.7 when combined with $\mathsf{SHATTER}$.

Restatement of Theorem 1.7. *There is a randomized algorithm* EIG *which on input any matrix* $A \in \mathbb{C}^{n\times n}$ *with* $\|A\| \le 1$ *and a desired accuracy parameter* $\delta \in (0,1)$ *outputs a diagonal* D *and invertible* V *such that*

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

in

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

arithmetic operations on a floating point machine with

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

bits of precision, with probability at least $1 - 12/n$*. Here* $T_{\mathsf{MM}}(n)$ *refers to the running time of a numerically stable matrix multiplication algorithm (detailed in Section 5.3).*

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

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

$$\delta' \triangleq \frac{\delta^3}{n^{4.5} \cdot 6 \cdot 128 \cdot 2}. \tag{5.10}$$

The X, g, ϵ output by $\mathsf{SHATTER}(A, \delta/8)$ easily satisfy the assumptions in Theorem 5.19, since $\delta' \leq \delta < 1$, $\epsilon = \frac{(\delta/8)^2}{32n^6} \leq 1/2$, g is defined by $\mathsf{SHATTER}$ to have side length 8, $\|X\| \leq \|A\| + \|X - A\| \leq 1 + 4(\delta/8) \leq 3.5$, and X has ϵ-pseudospectrum shattered with respect to g.

We will show that the choice of δ' in (5.10) guarantees

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

Theorem 5.14 implies that $X = WCW^{-1}$ is diagonalizable with probability one, and moreover

$$\|W\|\|W^{-1}\| \leq 8n^2/\delta$$

when W is normalized to have unit columns, by (2.1) (where we are using the proof of Theorem 1.10), with probability at least $1 - 11/n$.

Since $\|X\| \leq \|A\| + \|A - X\| \leq 1 + 4\delta \leq 3$ from Theorem 5.14, the hypotheses of Theorem 5.19 are satisfied. Thus EIG succeeds with probability at least $1 - 1/n$, and both EIG and $\mathsf{SHATTER}$ succeed with probability at least $1 - 12/n$ by a union bound. 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 $\omega = \frac{\delta^2}{4*8^2*n^6} \leq \frac{\delta}{\sqrt{2}}$ provided $\delta < 1$. Thus the guarantees on EIG in Theorem 5.19 tell us each eigenvalue of $X = WCW^{-1}$ 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

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

with probability $1 - 12/n$ as desired.

To compute the running time and precision, we observe that $\mathsf{SHATTER}$ outputs a grid with parameters

$$\omega = \Omega\left(\frac{\delta}{n^2}\right), \quad \epsilon = \Omega\left(\frac{\delta^2}{n^6}\right).$$

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. □

Proof of Theorem 5.19

A key stepping-stone in our proof will be the following elementary result controlling the spectrum, pseudospectrum, and eigenvectors after perturbing a shattered matrix — similar to Lemma 2.9.

Lemma 5.21 (Eigenvector Perturbation for a Shattered Matrix)**.** *Let $\Lambda_\epsilon(A)$ be shattered with respect to a grid whose squares have side length ω, and assume that $\|\widetilde{A} - A\| \le \eta < \epsilon$. Then, (i) each eigenvalue of $\widetilde{A}$ lies in the same grid square as exactly one eigenvalue of A, (ii) $\Lambda_{\epsilon-\eta}(\widetilde{A})$ is shattered with respect to the same grid, and (iii) for any right unit eigenvector $\tilde{v}$ of $\widetilde{A}$, there exists a right unit eigenvector of A corresponding to the same grid square, and for which*

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

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

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

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

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

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

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

Multiplying v by a phase we can assume without loss of generality that $\tilde{v}^*v \ge 0$ which implies that

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

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

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

Finally, note that $\|v - \tilde{v}\| = \sqrt{2 - 2\tilde{v}^*v} \le \frac{\sqrt{8}\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}$ 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. The following is similar in spirit to Lemma 2.11.

Lemma 5.22 (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 $\|(z - Q^*AQ)v\| \le \epsilon\|v\|$. Since $I_k = Q^*I_nQ$ we have

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

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 $(z - A)Qv \in \text{range}(Q)$, and therefore $\|Q^*(z - A)Qv\| = \|(z - A)Qv\|$. From where we obtain

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

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 $\|v^*Q^*(z - A)Q\| \le \epsilon\|v\|$. □

Observation 5.23. Since $\delta, \omega(\mathsf{g}), \epsilon \le 1$, our assumption on η in Line 2 of the pseudocode of EIG implies the following bounds on η which we will use below:

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

Initial lemmas in hand, let us begin to analyze the algorithm. At several points we will make an assumption on the machine precision in the margin. These will be collected at the end of the proof, where we will verify that they follow from the precision hypothesis of Theorem 5.19.

Correctness.

Lemma 5.24 (Accuracy of $\widetilde{\lambda}_i$)**.** *When* SPAN *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 SPLIT(A, β), $Q_\pm$ be the matrices whose orthogonal columns span their ranges, and $A_\pm \triangleq Q_\pm^* A Q_\pm$. From Theorem 5.17, on the event that SPAN succeeds, the approximation $\widetilde{Q}_\pm$ that it outputs satisfies $\|\widetilde{Q}_\pm - Q_\pm\| \le \eta$, so in particular $\|\widetilde{Q}_\pm\| \le 2$ as $\eta \le 1$. The error $E_{6,\pm}$ from performing the matrix multiplications necessary to compute $\widetilde{A}_\pm$ admits the bound

$$\begin{aligned}
\|E_{6,\pm}\| &\le \mu_{\mathsf{MM}}(n)\|\widetilde{Q}_\pm\|\|A\widetilde{Q}_\pm\|\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\|\widetilde{Q}_\pm A\|\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\|\widetilde{Q}_\pm\|^2\|A\|\mathbf{u} \\
&\le 16\left(\mu_{\mathsf{MM}}(n)\mathbf{u} + \mu_{\mathsf{MM}}(n)^2\mathbf{u}^2\right) && \|A\| \le 4, \|\widetilde{Q}_\pm\| \le 1 + \eta \le 1.02 \\
&\le 3\eta && \mathbf{u} \le \frac{\eta}{10\mu_{\mathsf{MM}}(n)^2}.
\end{aligned}$$

Iterating the triangle inequality, we obtain

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

We can now apply Lemma 5.21. □

Everything is now in place to show that, if every call to SPAN 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 5.21, 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 $\tilde{\tilde{v}}_\pm$ be some approximate right unit eigenvector of one of $\widetilde{A}_\pm$ output by EIG (with norm $1 \pm n\mathbf{u}$), $\tilde{v}_\pm$ 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

$$\tilde{v} \triangleq \frac{\widetilde{Q}_\pm \tilde{\tilde{v}}_\pm + e}{\|\widetilde{Q}_\pm \tilde{\tilde{v}}_\pm + e\|} + e',$$

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

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

and e', a column E_9 incurred by performing the normalization in floating point; in our initial discussion of floating point arithmetic we assumed in (5.6) that $\|e'\| \le n\mathbf{u}$.

First, since $\tilde{v} - e'$ and $\widetilde{Q}_\pm \tilde{\tilde{v}}_\pm + e$ are parallel, the distance between them is just the difference in their norms:

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

Inductively $\|\tilde{\tilde{v}}_\pm - \tilde{\tilde{v}}_\pm\| \le 4\delta/5$, and since $\|A_\pm - \widetilde{A}_\pm\| \le \epsilon/5$ and $A_\pm$ has shattered ϵ-pseudospectrum from Lemma 5.22, Lemma 5.21 ensures

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

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

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

This concludes the proof of correctness of EIG.

Running Time and Failure Probability. Let's begin with a simple lemma bounding the depth of EIG's recursion tree.

Lemma 5.25 (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 5.16, 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

$$d(n) = \begin{cases} n & n \leq 5 \\ 1 + \max_{\theta \in [1/5, 4/5]} d(\theta n) & n > 5 \end{cases} \tag{5.11}$$

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 8, and $\|A\| \leq 3.5$ in Theorem 5.19 ensure that every call to SPLIT throughout the algorithm satisfies the hypotheses of Theorem 5.16, 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 5.16 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 5.24 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 5.26 (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* SPAN *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 $\epsilon \longleftarrow 4\epsilon/5$ and $\delta \longleftarrow 4\delta/5$ 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 5.27 (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 SPAN. 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

$$\begin{aligned} T_{\mathsf{EIG}}(n, \delta, \mathsf{g}, \epsilon, \theta, n) \leq {} & T_{\mathsf{SPLIT}}(n, \epsilon, \beta) + T_{\mathsf{SPAN}}(n, \beta, \eta) + 2T_{\mathsf{MM}}(n) \\ & + T_{\mathsf{EIG}}(n_+, 4\delta/5, \mathsf{g}_+, 4\epsilon/5, \theta, n) + T_{\mathsf{EIG}}(n_-, 4\delta/5, \mathsf{g}_-, 4\epsilon/5, \theta, n) \\ & + 2T_{\mathsf{MM}}(n) + O(n^2). \end{aligned}$$

All of T_{SPLIT}, T_{SPAN}, and T_{MM} are of the form polylog(n) poly(n), with all coefficients nonnegative and exponents in the poly(n) no smaller than 2. So, for any $n_+ + n_- = n$ and $n_\pm \geq 4n/5$, holding all other parameters fixed, $T_{\mathsf{SPLIT}}(n_+, \ldots) + T_{\mathsf{SPLIT}}(n_-, \ldots) \leq \left((4/5)^2 + (1/5)^2\right) T_{\mathsf{SPLIT}}(n, \ldots) = (17/25) T_{\mathsf{SPLIT}}(n, \ldots)$ and the same holds for T_{SPAN} and T_{MM}. Applying this recursively, with all parameters other than n set to their lower bounds from Lemma 5.26, we then have

$$\begin{aligned} T_{\mathsf{EIG}}(n, \delta, \mathsf{g}, \epsilon, \theta, n) &\leq \frac{1}{1 - 17/25}\left(T_{\mathsf{SPLIT}}\left(n, \epsilon/n, \mathsf{g}, \frac{\delta^4\epsilon^8\theta^2}{(5n)^{26}}\right)\right. \\ &\qquad \left. + T_{\mathsf{SPAN}}\left(n, \beta/n, \epsilon/n, \frac{\delta^4\epsilon^8\theta^2}{(5n)^{26}}\right) + 4T_{\mathsf{MM}}(n) + O(n^2)\right) \\ &= \frac{25}{8}\left(12N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})}\left(T_{\mathsf{INV}}(n) + O(n^2)\right) + 2T_{\mathsf{QR}}(n)\right. \\ &\qquad \left. + 5T_{\mathsf{MM}}(n) + n^2 T_{\mathsf{N}} + O(n^2)\right) \\ &\leq 60N_{\mathsf{EIG}} \lg \frac{1}{\omega(\mathsf{g})}\left(T_{\mathsf{INV}}(n) + O(n^2)\right) + 10T_{\mathsf{QR}}(n) + 25T_{\mathsf{MM}}(n), \end{aligned}$$

where

$$N_{\mathsf{EIG}} \triangleq \lg \frac{256n}{\epsilon} + 3 \lg\lg \frac{256n}{\epsilon} + \lg\lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^9} + 7.59.$$

In the above inequalities, we've substituted in the expressions for T_{SPLIT} and T_{SPAN} from Theorems 5.16 and 5.17, 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 5.8, we have

$$T_{\mathsf{EIG}}(n, \delta, \mathbf{g}, \epsilon, \theta, n) = O\left(\log \frac{1}{\omega(\mathbf{g})}\left(\log \frac{n}{\epsilon} + \log\log \frac{1}{\theta\delta}\right) T_{\mathsf{MM}}(n, \mathbf{u})\right); \tag{5.12}$$

exact constants can be extracted by analyzing N_{EIG} and opening Theorem 5.8.

Required Bits of Precision. We will need the following bound on the norms of all spectral projectors.

Lemma 5.28 (Sizes of Spectral Projectors)**.** *Throughout the algorithm, every approximate spectral projector $\widetilde{P}$ given to* SPAN *satisfies $\|\widetilde{P}\| \le 10n/\epsilon$.*

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

$$\|\widetilde{P}\| \le 2 + \|P\| \le 2 + \frac{32}{2\pi}\frac{n}{\epsilon} \le 10n/\epsilon,$$

with the last inequality following from $\epsilon < 1$. □

Collecting the machine precision requirements $\mathbf{u} \le \mathbf{u}_{\mathsf{SPLIT}}, \mathbf{u}_{\mathsf{SPAN}}$ from Theorems 5.16 and 5.17, as well as those we used in the course of our proof so far, and substituting in the parameter lower bounds from Lemma 5.26, we need $\mathbf{u}$ to satisfy

$$\begin{aligned}\mathbf{u} \le \min\Bigg\{ &\frac{\left(1 - \frac{\epsilon}{256n}\right)^{2^{N_{\mathsf{EIG}}+1}(c_{\mathsf{INV}}\log n + 3)}}{\mu_{\mathsf{INV}}(n)\sqrt{n}N_{\mathsf{EIG}}}, \\ &\frac{\epsilon}{100n^2}, \frac{\theta^2\delta^4\epsilon^8}{(5n)^{26}}\frac{1}{4\|\widetilde{P}\|\max\{\mu_{\mathsf{QR}}(n), \mu_{\mathsf{MM}}(n)\}}, \\ &\frac{\delta\epsilon^2}{100n^3 \cdot 2\mu_{\mathsf{QR}}(n)}, \frac{\delta\epsilon^2}{100n^3\max\{4\mu_{\mathsf{MM}}(n), n, 2\mu_{\mathsf{QR}}(n)\}}\Bigg\}\end{aligned}$$

From Lemma 5.28, $\|\widetilde{P}\| \le 10n/\epsilon$, so the conditions in the second two lines are all satisfied if we make the crass upper bound

$$\mathbf{u} \le \frac{\theta^2\delta^4\epsilon^8}{(5n)^{30}}\frac{1}{\max\{\mu_{\mathsf{QR}}(n), \mu_{MM}(n), n\}}, \tag{5.13}$$

i.e. if $\lg 1/\mathbf{u} \geq O\left(\lg \frac{n}{\theta\delta\epsilon}\right)$. Unpacking the first requirement, using the definition $N_{\mathsf{EIG}} \triangleq \lg \frac{256n}{\epsilon} + 3\lg\lg \frac{256n}{\epsilon} + \lg\lg \frac{(5n)^{26}}{\theta^2\delta^4\epsilon^9} + 7.59$ from Theorem 5.19, 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 (5.13), and completes the proof of Theorem 5.19.

Remark 5.29. 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}.

5.6 Approximating the Matrix Sign Function

The algortithmic centerpiece of this chapter is the analysis, in finite arithmetic, of Roberts' iterative method for approximating to the matrix sign function. Recall from Section 5.1 that if A is a matrix whose spectrum avoids the imaginary axis, then

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

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

$$g(z) \triangleq \frac{1}{2}(z + z^{-1}). \tag{5.14}$$

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 $+1$.

```
SGN
Input: Matrix $A \in \mathbb{C}^{n\times n}$, pseudospectral guarantee $\epsilon$, circle parameter $\alpha$, and desired accuracy $\delta$
Requires: $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$.
Output: Approximate matrix sign function $S$
Ensures: $\|S - \mathrm{sgn}(A)\| \le \delta$

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

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

In Subsection 5.6 we briefly discuss the specific preliminaries that will be used throughout this section. In Subsection 5.6 we give a *pseudospectral* proof of the rapid global convergence of this iteration when implemented in exact arithmetic. In Subsection 5.6 we show that the proof provided in Subsection 5.6 is robust enough to handle the finite arithmetic case; a formal statement of this main result is the content of Theorem 5.38.

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 we will write $m(z) = \frac{1-z}{1+z}$ for the Möbius transformation taking the right half-plane to the unit disk, and for each $\alpha \in (0,1)$ we denote by

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

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 5.2 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

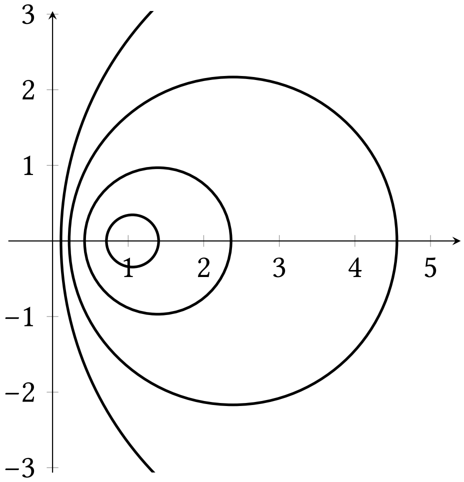

Figure 5.2: Apollonian circles appearing in the analysis of the Newton iteration. Depicted are $\partial\mathsf{C}^+_{\alpha^{2^k}}$ for $\alpha = 0.8$ and $k = 0, 1, 2, 3$, with smaller circles corresponding to larger k.

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 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 $\mathsf{A}_{\alpha,\beta} = \mathsf{A}^+_{\alpha,\beta} \cup \mathsf{A}^-_{\alpha,\beta}$. The following observation is due to Roberts [133].

Observation 5.30. 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 5.30 that under repeated application of the Newton map g, any point in the right or left half-plane converges to +1 or −1, respectively.

Exact Arithmetic

In this section, we set $A_0 \triangleq A$ and $A_{k+1} \triangleq g(A_k)$ for all $k \geq 0$. In the case of exact arithmetic, Observation 5.30 implies global convergence of the Newton iteration when A is diagonalizable. For the convenience of the reader we provide this argument (due to [133]) below.

Proposition 5.31. *Let A be a diagonalizable $n \times n$ matrix and assume that $\Lambda(A) \subset \mathsf{C}_\alpha$ for some $\alpha \in (0,1)$. Then for every $N \in \mathbb{N}$ we have the guarantee*

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

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

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

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

$$\|A_N - \operatorname{sgn}(A)\| \leq \left\| \sum_{\Re(\lambda_i)>0} (\lambda_i^{(N)} - 1) v_i w_i^* \right\| + \left\| \sum_{\Re(\lambda_i)<0} (\lambda_i^{(N)} + 1) v_i w_i^* \right\|. \tag{5.15}$$

Now we will use that the operator norm of any matrix is at most the spectral radius times the eigenvector condition number. Observe that the spectral radii of the two matrices appearing on the right hand side of (5.15) 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 [133], 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 of the above proof is needed, tolerant to perturbations. 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 $\|A_N - \mathrm{sgn}(A)\|$ solely using information from the pseudospectrum of A_N.

Lemma 5.32 (Pseudospectral Error Bound)**.** *Let A be any $n \times n$ matrix and let 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*

$$\|A_N - \mathrm{sgn}(A)\| \le \frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\epsilon_N}. \tag{5.16}$$

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

$$\begin{aligned}
\|A_N - \mathrm{sgn}(A_N)\| &= \left\| \frac{1}{2\pi i}\oint_{\partial \mathsf{C}_{\alpha_N}} z(z-A_N)^{-1}\,\mathrm{d}z - \frac{1}{2\pi i}\left(\oint_{\partial \mathsf{C}^+_{\alpha_N}} (z-A_N)^{-1}\,\mathrm{d}z - \oint_{\partial \mathsf{C}^-_{\alpha_N}} (z-A_N)^{-1}\,\mathrm{d}z\right)\right\| \\
&= \left\| \frac{1}{2\pi i}\oint_{\partial \mathsf{C}^+_{\alpha_N}} z(z-A_N)^{-1} - (z-A_N)^{-1}\,\mathrm{d}z + \frac{1}{2\pi i}\oint_{\partial \mathsf{C}^-_{\alpha_N}} z(z-A_N)^{-1} + (z-A_N)^{-1}\,\mathrm{d}z\right\| \\
&\le \frac{1}{2\pi}\left\|\oint_{\partial \mathsf{C}^+_{\alpha_N}} (z-1)(z-A_N)^{-1}\,\mathrm{d}z\right\| + \frac{1}{2\pi}\left\|\oint_{\partial \mathsf{C}^-_{\alpha_N}} (z+1)(z-A_N)^{-1}\,\mathrm{d}z\right\| \\
&\le 2\cdot\frac{1}{2\pi}\ell(\partial \mathsf{C}^+_{\alpha_N})\sup\{|z-1| \,:\, z \in \mathsf{C}^+_{\alpha_N}\}\frac{1}{\epsilon_N} \\
&= \frac{4\alpha_N}{1-\alpha_N^2}\left(\frac{1+\alpha_N}{1-\alpha_N} - 1\right)\frac{1}{\epsilon_N} \\
&= \frac{8\alpha_N^2}{(1-\alpha_N)^2(1+\alpha_N)\epsilon_N}.
\end{aligned}$$

□

In view of Lemma 5.32, 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 5.6.

The lemma below is instrumental in determining the sequences α_k, ϵ_k.

Lemma 5.33 (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*

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

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 5.3 for an illustration. We must show that $w \notin \Lambda_{\epsilon'}(g(A))$. Since $w \notin \mathsf{C}_{\alpha^2}$, Observation 5.30 ensures no $z \in \mathsf{C}_\alpha$ satisfies $g(z) = w$; in other words, the function $(w - g(z))^{-1}$ is holomorphic in z on C_α. As $\Lambda(A) \subset \Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, Observation 5.30 also guarantees that $\Lambda(g(A)) \subset \mathsf{C}_{\alpha^2}$. Thus for w in the union of the two Appolonian annuli in question, we can calculate the resolvent of $g(A)$ at w using the holomorphic functional calculus:

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

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 5.30 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 5.34. Thus, with ϵ' defined as in the theorem statement, $\mathsf{A}_{\alpha,\alpha'}$ contains none of the ϵ'-pseudospectrum of $g(A)$. Since $\Lambda(g(A)) \subset \mathsf{C}_{\alpha^2}$, Lemma 2.1 tells us that there can be no ϵ'-pseudospectrum in the remainder of $\mathbb{C} \setminus \mathsf{C}_{\alpha'}$, as such a connected component would need to contain an eigenvalue of $g(A)$. □

Lemma 5.34. *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 $|x - y| \geq (\alpha - \beta)/2$.*

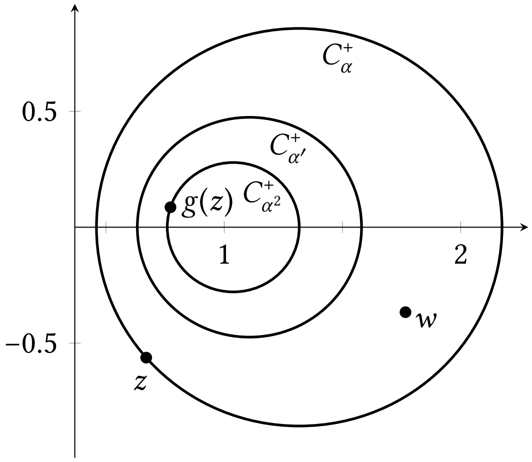

Figure 5.3: Illustration of the proof of Lemma 5.33

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

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

□

Lemma 5.33 will also be useful in bounding the condition numbers of the A_k, which is necessary for the finite arithmetic analysis.

Corollary 5.35 (Condition Number Bound). *Using the notation of Lemma 5.33, if* $\Lambda_\epsilon(A) \subset \mathsf{C}_\alpha$, *then*

$$\|A^{-1}\| \le \frac{1}{\epsilon} \quad \textit{and} \quad \|A\| \le \frac{4\alpha}{(1-\alpha)^2\epsilon}.$$

Proof. The bound $\|A^{-1}\| \le 1/\epsilon$ 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

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

□

Another direct application of Lemma 5.33 yields the following.

Lemma 5.36. *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 5.33 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 5.31.

Proposition 5.37. *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 5.36 and the bound (5.16), 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. □

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 $\mathsf{G}(A) \triangleq g(A) + E_A$ where E_A is an adversarial perturbation coming from the round-off error. Hence, the sequence of interest is given by $\widetilde{A}_0 \triangleq A$ and $\widetilde{A}_{k+1} \triangleq \mathsf{G}(\widetilde{A}_k)$.

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 5.38 (Main guarantees for SGN)**.** *Assume* INV *is a* $(\mu_{\mathsf{INV}}(n), c_{\mathsf{INV}})$*-stable matrix inversion algorithm satisfying Definition 5.5. 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*

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

when run with machine precision satisfying

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

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 5.38.

Lemma 5.39. *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*

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

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 5.31, it suffices to take

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

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

The proof of Theorem 5.38 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 5.40 (One-step additive error)**.** *Assume the matrix inverse is computed by an algorithm* INV *satisfying the guarantee in Definition 5.5. Then* $\mathsf{G}(A) = g(A) + E$ *for some error matrix* E *with norm*

$$\|E\| \le \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|\right) 2\sqrt{n}\mathbf{u}. \tag{5.17}$$

The proof of this lemma is deferred to Section 5.10.

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 $s \triangleq 1 - \alpha_0$, which should be thought of as a small parameter. As in the exact arithmetic case, for $k \ge 1$, we will recursively define decreasing sequences α_k and ϵ_k maintaining the property

$$\Lambda_{\epsilon_k}(\widetilde{A}_k) \subset \mathsf{C}_{\alpha_k} \qquad \text{for all } k \ge 0 \tag{5.18}$$

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

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

In the notation of Subsection 5.6, this corresponds to setting $D = 1 + s/4$. This definition ensures that $\alpha_k^2 \le \alpha_{k+1} \le \alpha_k$ for all k, and also gives us the bound $(1 + s/4)\alpha_0 \le 1 - s/2$. We also have the closed form

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

which implies the useful bound

$$\alpha_k \le (1 - s/2)^{2^k}. \tag{5.19}$$

3. Here we recursively define ϵ_{k+1}. Combining Lemma 5.33, the recursive definition of α_{k+1}, and the fact that $1 - \alpha_k^2 \ge 1 - \alpha_0^2 \ge 1 - \alpha_0 = s$, we find that $\Lambda_{\epsilon'}\left(g(\widetilde{A}_k)\right) \subset \mathsf{C}_{\alpha_{k+1}}$, where

$$\epsilon' = \epsilon_k \frac{\left(\alpha_{k+1} - \alpha_k^2\right)(1 - \alpha_k^2)}{8\alpha_k} = \epsilon_k \frac{s\alpha_k(1 - \alpha_k^2)}{32} \ge \epsilon_k \frac{\alpha_k s^2}{32}.$$

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, stability of pseudospectrum ensures that if we set

$$\epsilon_{k+1} \triangleq \epsilon_k \frac{s^2\alpha_k}{32} - \|E_k\| \tag{5.20}$$

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 (5.20), 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 (5.17) on $\|E_k\|$ from Lemma 5.40, as functions of ϵ_k. By Corollary 5.35, 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 (5.17) as

$$K_{\epsilon_k} \triangleq \left[\frac{4}{s^2\epsilon_k} + \frac{1}{\epsilon_k} + \mu_{\mathrm{INV}}(n)\left(\frac{4}{s^2\epsilon_k^2}\right)^{c_{\mathrm{INV}}\log n}\frac{1}{\epsilon_k}\right] 2\sqrt{n}$$

so that Lemma 5.40 reads

$$\|E_k\| \leq K_{\epsilon_k}\mathbf{u}. \tag{5.21}$$

Plugging this into the definition (5.20) of ϵ_{k+1},we have

$$\epsilon_{k+1} \geq \epsilon_k \frac{s^2\alpha_k}{32} - K_{\epsilon_k}\mathbf{u}. \tag{5.22}$$

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

$$K_{\epsilon_k}\mathbf{u} \leq \frac{1}{3}\epsilon_k\frac{s^2\alpha_k}{32}. \tag{5.23}$$

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

$$\epsilon_{k+1} \geq \frac{2}{3}\epsilon_k\frac{s^2\alpha_k}{32} = \frac{1}{48}\epsilon_k s^2\alpha_k, \tag{5.24}$$

which implies

$$\|E_k\| \leq \frac{1}{2}\epsilon_{k+1}; \tag{5.25}$$

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 5.41 (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 (5.23):*

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

Then we have

$$\epsilon_k \geq e_k \triangleq \epsilon_0\left(\frac{s^2}{50}\right)^k \alpha_k.$$

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 (5.23) implies (5.24), we may apply (5.24) 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 5.41 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 5.41. 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 5.42 (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*

$$K_{e_i} \mathbf{u} \leq \frac{1}{3} e_i \frac{s^2 \alpha_i}{32} \tag{5.26}$$

where e_i is defined in Lemma 5.41. 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 5.41 to obtain $\epsilon_{i+1} \geq e_{i+1}$, as desired. □

Lemma 5.43 (Main Accuracy Bound). *Suppose $\mathbf{u}$ satisfies the requirement (5.23) for all $0 \leq k \leq N$. Then*

$$\|\widetilde{A}_N - \operatorname{sgn}(A)\| \leq \frac{8}{s} \sum_{k=0}^{N-1} \frac{\|E_k\|}{\epsilon_{k+1}^2} + \frac{8 \cdot 50^N}{s^{2N+2} \epsilon_0} (1 - s/2)^{2^N}. \tag{5.27}$$

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

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

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

$$\begin{aligned}
&\frac{1}{2\pi}\left\| \oint_{\partial \mathsf{C}^+_{\alpha_{k+1}}} \left((z - \widetilde{A_{k+1}})^{-1} - (z - (\widetilde{A_{k+1}} - E_k))^{-1}\right)\, \mathrm{d}z - \oint_{\partial \mathsf{C}^-_{\alpha_{k+1}}} \left((z - \widetilde{A_{k+1}})^{-1} - (z - (\widetilde{A_{k+1}} - E_k))^{-1}\right)\, \mathrm{d}z \right\| \\
&\qquad \le \frac{1}{\pi} \oint_{\partial \mathsf{C}^+_{\alpha_{k+1}}} \|(z - (\widetilde{A_{k+1}} - E_k))^{-1}\| \|E_k\| \|(z - \widetilde{A_{k+1}})^{-1}\|\, \mathrm{d}z \\
&\qquad \le \frac{1}{\pi} \ell(\partial \mathsf{C}^+_{\alpha_{k+1}}) \|E_k\| \frac{1}{\epsilon_{k+1} - \|E_k\|} \frac{1}{\epsilon_{k+1}} \\
&\qquad = \frac{4\alpha_{k+1}}{1 - \alpha_{k+1}^2} \|E_k\| \frac{1}{\epsilon_{k+1} - \|E_k\|} \frac{1}{\epsilon_{k+1}},
\end{aligned}$$

where we use the definition and stability of pseudospectrum, together with the property (5.18). Ultimately, this chain of inequalities implies

$$\|\mathrm{sgn}(\widetilde{A_{k+1}}) - \mathrm{sgn}(\widetilde{A_k})\| \le \frac{4\alpha_{k+1}}{1 - \alpha_{k+1}^2} \|E_k\| \frac{1}{\epsilon_{k+1} - \|E_k\|} \frac{1}{\epsilon_{k+1}}.$$

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

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

where in the last step we use $\alpha_k \le 1$ and $1 - \alpha_{k+1}^2 \ge s$, as well as (5.25).

By Lemma 5.32 (to be precise, by repeating the proof of that lemma with $\widetilde{A_N}$ substituted for A_N), we have

$$\begin{aligned}
\|\widetilde{A_N} - \mathrm{sgn}(\widetilde{A_N})\| &\le \frac{8\alpha_N^2}{(1 - \alpha_N)^2(1 + \alpha_N)\epsilon_N} \\
&\le \frac{8}{s^2} \alpha_N \frac{\alpha_N}{\epsilon_N} \\
&\le \frac{8}{s^2} \alpha_N \frac{1}{\epsilon_0} \left(\frac{50}{s^2}\right)^N \\
&\le \frac{8}{s^2 \epsilon_0} (1 - s/2)^{2^N} \left(\frac{50}{s^2}\right)^N \\
&\le \frac{8 \cdot 50^N}{s^{2N+2}\epsilon_0} (1 - s/2)^{2^N}.
\end{aligned}$$

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 5.43 to ensure $\|\widetilde{A_N} - \operatorname{sgn}(A)\|$ is at most β, the desired accuracy parameter. The upper bound (5.27) in Lemma 5.43 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 5.45 and 5.46.

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

Lemma 5.44. *Let* $1/800 > t > 0$ *and* $1/2 > c > 0$ *be given. Then for*

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

we have

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

The proof is deferred to Section 5.10.

Lemma 5.45 (Bound on Second Term of (5.27))**.** *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 5.44 with $c = \beta s^2 \epsilon_0/16$ and $t = s/8$.

□

Now we move to the first term in the bound of Lemma 5.43.

Lemma 5.46 (Bound on First Term of (5.27))**.** *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*

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

Then we have

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

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

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

In view of (5.21), which says $\|E_k\| \le K_{\epsilon_k}\mathbf{u}$, it is sufficient to have for all $0 \le k \le N-1$

$$\mathbf{u} \le \frac{1}{K_{\epsilon_k}}\frac{\beta\epsilon_{k+1}^2 s}{16N}. \tag{5.28}$$

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

$$\mathbf{u} \le \frac{1}{K_{e_k}}\frac{\beta e_{k+1}^2 s}{16N}. \tag{5.29}$$

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 (5.29) implies $\mathbf{u} \le \frac{1}{3K_{e_k}}\frac{s^2 e_k}{32}$, which means that the assumption in Corollary 5.42 is satisfied. On the other hand Corollary 5.42 yields $e_k \le \epsilon_k$ for all $0 \le k \le N$, which in turn, combined with (5.29) would give (5.28) and conclude the proof.

We now show that (5.29) holds for all $0 \le k \le N-1$. Because $1/K_{e_k}$ and e_k are decreasing in k, it is sufficient to have the single condition

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

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

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

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

Substituting the value of e_N as defined in Lemma 5.41, we get the sufficient condition

$$\mathbf{u} \le \frac{\beta}{192\mu_{\mathsf{INV}}(n)\sqrt{n}N}\left(\frac{\epsilon_0(s^2/50)^N \alpha_N s^2}{4}\right)^{c_{\mathsf{INV}}\log n+3},$$

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

$$\mathbf{u} \le \frac{\beta}{192\mu_{\mathsf{INV}}(n)\sqrt{n}N}\left(\frac{\epsilon_0(s^2/50)^N (1-s)^{2^N} s^2}{4}\right)^{c_{\mathsf{INV}}\log n+3}.$$

Now we finally will use our hypothesis on the size of N to simplify this expression. Applying Lemma 5.45, we have

$$\epsilon_0(s^2/50)^N/4 \ge \frac{4(1-s)^{2^N}}{s^2\beta}.$$

Thus, our sufficient condition becomes

$$\mathbf{u} \le \frac{\beta}{192\mu_{\mathsf{INV}}(n)\sqrt{n}N}\left(\frac{4(1-s)^{2^{N+1}}}{\beta}\right)^{c_{\mathsf{INV}}\log n+3}.$$

To make the expression simpler, since $c_{\mathsf{INV}}\log n + 3 \ge 4$ we may pull out a factor of $4^4 > 192$ and remove the occurrences of β to yield the sufficient condition

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

□

Matching the statement of Theorem 5.38, we give a slightly cleaner sufficient condition on N that implies the hypothesis on N appearing in the above lemmas. The proof is deferred to Section 5.10.

Lemma 5.47 (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

$$N \ge \lg(8/s) + 2\lg\lg(8/s) + \lg\lg(16/(\beta s^2\epsilon_0)) + 1.62.$$

Taking the logarithm of the machine precision yields the number of bits required:

Lemma 5.48 (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 5.38), at the cost of making the computation somewhat messier.

Immediately we have

$$\lg(1/\mathbf{u}_{\mathsf{SGN}}) \le \lg\mu_{\mathsf{INV}}(n) + \frac{1}{2}\lg n + \lg N + (c_{\mathsf{INV}}\log n + 3)2^{N+1}\log(1/(1-s)).$$

Note that $\log(1/(1-s)) < s$ for $s < 1/2$. Also, $2^{N+1} \le (1/s)\lg(1/s)^3(\lg(1/\beta)+\lg(1/\epsilon_0))2^{9.59}$. Putting this together, we have

$$\lg(1/\mathbf{u}_{\mathsf{SGN}}) \le \lg\mu_{\mathsf{INV}}(n) + \frac{1}{2}\lg n + \lg N + 1000(c_{\mathsf{INV}}\log n + 3)\lg(1/s)^3(\lg(1/\beta)+\lg(1/\epsilon_0)).$$

We now crudely bound $\lg N$. Note that for $s < 1/100$ we have $\lg(1/s) + 3\lg\lg(1/s) + 7.59 \le 1/s$. Thus,

$$\begin{aligned} \lg N &\le \lg(1/s + \lg\lg(1/(\beta\epsilon_0))) \\ &\le \lg(1/s + \lg(1/(\beta\epsilon_0))) \\ &\le \lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) && \lg(a+b) \le \lg a + \lg b \text{ for } a, b > 2 \\ &\le \lg(1/s)^3\lg(1/(\beta\epsilon_0)). \end{aligned}$$

Combining the above, we may fold the $\lg N$ and $\lg n$ terms into the final term to obtain

$$\lg(1/\mathbf{u}_{\mathsf{SGN}}) \le \lg\mu_{\mathsf{INV}}(n) + 5000c_{\mathsf{INV}}\log n\lg(1/s)^3(\lg(1/\beta)+\lg(1/\epsilon_0)) \tag{5.30}$$

where we use that $c_{\mathsf{INV}}\log n > 1$ and therefore $c_{\mathsf{INV}}\log n + 3 < 4c_{\mathsf{INV}}\log n$. 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 5.38. Finally, we may prove the theorem advertised in Section 5.1.

Proof of Theorem 5.1. Set $\epsilon \triangleq \min\{\frac{1}{K}, 1\}$. Then $\Lambda_\epsilon(A)$ does not intersect the imaginary axis, and furthermore $\Lambda_\epsilon(A) \subseteq \mathbb{D}(0,2)$ because $\|A\| \le 1$. Thus, we may apply Lemma 5.39 with $\mathrm{diag}(\mathbf{g}) = 4\sqrt{2}$ 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 5.38 now yields the desired conclusion. □

5.7 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 5.49. *Let A have ϵ-pseudospectrum shattered with respect to some grid g. Then there exists a horizontal or vertical grid line of g partitioning g into two grids $\mathsf{g}_\pm$, each containing at least* $\min\{n/5, 1\}$ *eigenvalues.*

Proof. We will view g as a $s_1 \times s_2$ array of squares. Write $r_1, r_2, ..., r_{s_1}$ for the number of eigenvalues in each row of the grid. Either there exists $1 \le i < s_2$ such that $r_1 + \cdots + r_i \ge n/5$ and $r_{i+1} + \cdots + r_{s_1} \ge n/5$—in which case we can bisect at the grid line dividing the ith from $(i+1)$st rows—or there exists some i for which $r_i \ge 3/5$. In the latter case, we can always find a vertical grid line so that at least $n/5$ of the eigenvalues in the ith row are on each of the left and right sides. Finally, if $n \le 5$, we may trivially pick a grid line to bisect along so that both sides contain at least one eigenvalue. □

Proof of Theorem 5.16. The main observation is that, given any matrix A, we can determine how many eigenvalues are on either side of any horizontal or vertical line by approximating the sign function of a shift of the matrix. To be precise, in exact arithmetic $\mathrm{tr}\ \mathrm{sgn}(A - h) = n_+ - n_-$, where $n_\pm$ are the eigenvalue counts for A on either side of the line $\Re z = h$. 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,

$$\begin{aligned} |\,\mathrm{tr}\ \mathsf{SGN}(M) + e_4 - \mathrm{tr}\ \mathrm{sgn}(M)| &\le |\,\mathrm{tr}\ \mathsf{SGN}(M) - \mathrm{tr}\ \mathrm{sgn}(M)| + |e_4| \\ &\le n\big(\|\mathsf{SGN}(M) - \mathrm{sgn}(M)\| + \|\mathsf{SGN}(M)\|\mathbf{u}\big) \quad \text{Using (5.5) to bound } |e_4| \\ &\le n\big(\beta + (\beta + \|\mathrm{sgn}(M)\|)\mathbf{u}\big). \end{aligned}$$

It remains to control $\|\mathrm{sgn}(M)\|$ and quantify the distance between $\mathrm{sgn}(M) = \mathrm{sgn}(A - h + E_2)$ and $\mathrm{sgn}(A - h)$. 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 $\mathrm{sgn}(A - h)$ and $\mathrm{sgn}(M)$ by integrating around the boundary of the portions of g on either side of the line $\Re z = h$, then using the resolvent identity as in Section 5.6, and the fact that $\Lambda_\epsilon(A)$ and $\Lambda_{\epsilon/2}(M)$ are shattered we get

$$\|\mathrm{sgn}(A) - \mathrm{sgn}(M)\| \le \frac{\|E_2\|}{2\pi} \cdot \frac{1}{\epsilon} \cdot \frac{2}{\epsilon}\omega(2s_1 + 4s_2) \le \frac{128\mathbf{u}}{\epsilon^2}$$

SPLIT

Input: Matrix $A \in \mathbb{C}^{n\times n}$, grid $\mathsf{g} = \mathsf{grid}(z_0, \omega, s_1, s_2)$ pseudospectral guarantee ϵ, and a desired accuracy ν.
Requires: $\Lambda_\epsilon(A)$ is shattered with respect to g, and $\beta \leq 0.05/n$.
Output: Sub-grids $\mathsf{g}_\pm$, approximate spectral projectors $\widetilde{P}_\pm$, and ranks $n_\pm$.
Ensures: There exist true spectral projectors $P_\pm$ satisfying (i) $P_+ + P_- = 1$, (ii) $\mathrm{Rank}(P_\pm) = n_\pm \geq n/5$, (iii) $\|P_\pm - \widetilde{P}_\pm\| \leq \beta$, and (iv) $P_\pm$ are the spectral projectors onto the interiors of $\mathsf{g}_\pm$.

```
1. $h \longleftarrow \Re z_0 + \omega s_1/2$
2. $M \longleftarrow A - h + E_2$
3. $\alpha_0 \longleftarrow 1 - \frac{\epsilon}{2\,\mathrm{diag}(\mathsf{g})^2}$
4. $\phi \longleftarrow \mathrm{round}\,(\mathrm{tr}\ \mathsf{SGN}(M, \epsilon/4, \alpha_0, \beta) + e_4)$
5. If $|\phi| < \min(3n/5, n-1)$
    a) $\mathsf{g}_- = \mathsf{grid}(z_0, \omega, s_1/2, s_2)$
    b) $z_0 \longleftarrow z_0 + h$
    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 $\mathrm{tr}\,\mathsf{SGN}(A - h, \epsilon/4, \alpha_0, \beta) \leq \frac{3n}{5}$, at which point output $\mathsf{g}_\pm$, the subgrids on either side of the shift $h$, and set
   $\widetilde{P}_\pm \longleftarrow \frac{1}{2}\,(\mathsf{SGN}(h - A, \epsilon/4, \alpha_0, \beta))$.
7. If this fails, set $A \longleftarrow iA$, and execute a binary search among vertical shifts from the original grid.
```

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 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 5.39 and the fact that $\mathrm{diag}(\mathsf{g})^2 = 128$, we can safely call SGN with parameters $\epsilon_0 = \epsilon/4$ and

$$\alpha_0 = 1 - \frac{\epsilon}{256}.$$

Plugging these in to the Theorem 5.38 ($\epsilon < 1/2$ 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

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

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

$$\lg s_1 s_2 \, T_{\mathsf{SGN}} \left(n, \epsilon/2, 1 - \frac{\epsilon}{2\,\mathrm{diag}(\mathsf{g})^2}, \beta \right)$$

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 5.38, we obtain

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

□

5.8 Analysis of SPAN

The algorithm SPAN, defined in Section 5.5, can be viewed as a small variation of the randomized rank revealing algorithm introduced in [61] and revisited subsequently in [14]. 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

$$R = \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix}, \tag{5.31}$$

where $R_{11} \in \mathbb{C}^{r \times r}$ has smallest singular value close to $\sigma_r(A)$, and R_{22} has largest singular value roughly $\sigma_{r+1}(A)$. We will use and analyze the following implementation of RURV.

As discussed in Section 5.5, we hope to use SPAN 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

RURV

Input: Matrix $A \in \mathbb{C}^{n\times n}$
Output: A pair of matrices (U, R).
Ensures: $\|R_{22}\| \le \frac{\sqrt{r(n-r)}}{\theta}\sigma_{r+1}(A)$ with probability $1-\theta^2$, for every $1 \le r \le n-1$ and $\theta > 0$, where R_{22} is the $(n-r)\times(n-r)$ lower-right corner of R.

```
1. $G \longleftarrow n \times n$ complex Ginibre matrix $+E_1$
2. $(V, R) \longleftarrow \mathsf{QR}(G)$
3. $B \longleftarrow AV^* + E_3$
4. $(U, R) \longleftarrow \mathsf{QR}(B)$
```

from the output of $\mathsf{RURV}(\widetilde{P})$ 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 5.60 and Proposition 5.65 below.

Since it may be of broader use, we will work in somewhat greater generality, and define the subroutine SPAN 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 SPAN is as follows.

SPAN

Input: Matrix $\widetilde{A} \in \mathbb{C}^{n\times n}$ and parameter $r \le n$
Requires: $1/3 \le \|A\|$, and $\|\widetilde{A} - A\| \le \beta$ for some $A \in \mathbb{C}^{n\times n}$ with $\mathrm{Rank}(A) = \mathrm{Rank}(A^2) = r$, as well as $\beta \le 1/4 \le \|\widetilde{A}\|$ and $1 \le \mu_{\mathsf{MM}}(n), \mu_{\mathsf{QR}}(n), c_{\mathsf{N}}$.
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 $\|\widetilde{S} - S\| \le \eta$, with probability at least $1 - \frac{(20n)^3\sqrt{\beta}}{\eta^2\sigma_r(A)}$.

```
1. $(U, R) \longleftarrow \mathsf{RURV}(A)$
2. $\widetilde{S} \longleftarrow$ first $r$ columns of $U$.
3. Output $\widetilde{S}$
```

Throughout this section we use $\mathsf{rurv}(\cdot)$ and $\mathsf{span}(\cdot,\cdot)$ to denote the exact arithmetic versions of RURV and SPAN respectively. In Subsection 5.8 we present a random matrix result that will be needed in the analysis of SPAN. In Subsection 5.8 we state the properties of RURV that will be needed. Finally in Subsections 5.8 and 5.8 we prove the main guarantees of span and SPAN, respectively, that are used throughout this chapter.

Smallest Singular Value of the Corner of a Haar Unitary

We recall the defining property of the Haar measure on the unitary group:

Definition 5.50. A random $n \times n$ unitary matrix V_n is *Haar-distributed* if, for any other unitary matrix W, $V_n W$ and $W V_n$ 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_n 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 5.51. 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 $\mathrm{Spec}(V_{11}V_{11}^*) = 1 - \mathrm{Spec}(V_{12}V_{12}^*)$.

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 $\mathrm{Spec}(V_{22}^*V_{22}) = 1 - \mathrm{Spec}(V_{12}^*V_{12})$.

Now recall that for any two matrices X and Y, the symmetric difference of the sets $\mathrm{Spec}(XY)$ and $\mathrm{Spec}(YX)$ is $\{0\}$, with multiplicity equal to the difference between the dimensions. Hence $\mathrm{Spec}(V_{12}V_{12}^*) = \mathrm{Spec}(V_{12}^*V_{12}) \cup \{0\}$ where the multiplicity of 0 is $r - (n-r) = 2r - n$. Combining this with $\mathrm{Spec}(V_{11}V_{11}^*) = 1 - \mathrm{Spec}(V_{12}V_{12}^*)$ and $\mathrm{Spec}(V_{22}^*V_{22}) = 1 - \mathrm{Spec}(V_{12}^*V_{12})$ we get the desired result. □

Proposition 5.52 ($\sigma_{\min}$ of a Submatrix of a Haar Unitary)**.** *Let $n > r > 0$ and let V_n be an $n \times n$ Haar unitary. Let X be the upper left $r \times r$ corner of V_n. Then, for all $\theta \in (0, 1]$,*

$$\mathbb{P}\left[\frac{1}{\sigma_r(X)} \leq \frac{1}{\theta}\right] = (1 - \theta^2)^{r(n-r)}. \tag{5.32}$$

In particular, for every $\theta \in (0, 1]$ we have

$$\mathbb{P}\left[\frac{1}{\sigma_r(X)} \leq \frac{\sqrt{r(n-r)}}{\theta}\right] \geq 1 - \theta^2. \tag{5.33}$$

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 [68], from which one can extract and simplify the density of $\sigma_r(X)$. We will begin by introducing the relevant pieces of [68], 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 ${}_2F_1^{\beta}(a, b; c; (x_1, \dots, x_m))$. For our application it is sufficient to know that

$$ {}_2F_1^{\beta}(0, b; c, (x_1, \dots, x_m)) = 1, \tag{5.34}$$

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 [77] or Definition 2.2 in [68]).

The generic results in [68] concern the *β-Jacobi* random matrices, which we have no cause here to define in full. Of particular use to us will be [68, Theorem 3.1], which expresses the density of the smallest singular value of such a matrix in terms of the generalized hypergeometric function:

Theorem 5.53. *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*

$$C_{\beta,a,b,m}\lambda^{\frac{\beta}{2}(a+1)-1}(1-\lambda)^{\frac{\beta}{2}m(b+m)-1}{}_2F_1^{2/\beta}\left(1-\frac{\beta(a+1)}{2}, \frac{\beta(b+m-1)}{2}; \frac{\beta(b+2m-1)}{2}+1; (1-\lambda)^{m-1}\right), \tag{5.35}$$

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 [73, Theorem 1.5], which we rephrase now to match our context.

Theorem 5.54. *Let V_n 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_n. Then, the eigenvalues of XX^* distribute as the eigenvalues of a β–Jacobi matrix of size r with parameters $\beta = 2$, $a = 0$ and $b = n - 2r$.*

In view of the above result, Theorem 5.53 gives a formula for the density of $\sigma_r^2(X)$.

Corollary 5.55 (Density of $\sigma_r^2(X)$)**.** *Let V_n 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*

$$f_{\sigma_r^2}(x) \triangleq \begin{cases} r(n-r)(1-x)^{r(n-r)-1} & \textit{if} \quad 0 \leq x \leq 1, \\ 0 & \textit{otherwise.} \end{cases} \tag{5.36}$$

Proof. If $r > n/2$, since we care only about the smallest singular value of X, we can use Observation 5.51 to analyse the $(n-r)\times(n-r)$ lower right corner of V_n 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 5.53 and observe that in this case

$$f_{\lambda_{\min}}(x) = C(1-x)^{r(n-r)-1}{}_2F_1^1(0, n-r-1; n; (1-x)^{r-1}) = C(1-x)^{r(n-r)-1} \tag{5.37}$$

where the last equality follows from (5.34). 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 5.54 we have $f_{\sigma_r^2}(x) = f_{\lambda_{\min}}(x)$. By integrating on $[0,1]$ the right side of (5.37) we find $C = r(n-r)$. □

Proof of Proposition 5.52. From (5.36) 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 (5.32) follows. To prove (5.33) note that $g(t) \triangleq (1-t)^{r(n-r)}$ is convex in $[0,1]$, and hence $g(t) \geq g(0) + tg'(0)$ for every $t \in [0,1]$. □

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 [72, Section 4.6] and [114] for a detailed discussion. In this subsection we carefully analyze this process in finite arithmetic. The following fact (see [114, Section 5]) is the starting point of our discussion.

Lemma 5.56 (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*[7]*. 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 5.3 we can assume that we have access to a random matrix $\widetilde{G_n}$, with

$$\widetilde{G_n} = G_n + E,$$

[7] G_n is almost surely invertible and under this event U and R are uniquely determined by these conditions.

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 $\|E\| \le \|E\|_F \le \sqrt{n} c_{\mathsf{N}}\mathbf{u}$.

In what follows we use $\mathsf{qr}(\cdot)$ to denote the exact arithmetic version of $\mathsf{QR}(\cdot)$. Furthermore, we assume that for any $A \in \mathbb{C}^{n\times n}$, $\mathsf{qr}(A)$ returns a pair (U, R) with the property that R has nonnegative entries on the diagonal. Since we want to compare $\mathsf{qr}(G_n)$ with $\mathsf{QR}(\widetilde{G_n})$ it is necessary to have a bound on the condition number of the QR decomposition; recall from Lemma 2.12 that if U and $\widetilde{U}$ are the unitaries produced by $\mathsf{qr}(A)$ and $\mathsf{qr}(A + E)$, where A is invertible and $\|E\|\|A^{-1}\| \le \frac{1}{2}$, then

$$\|\widetilde{U} - U\|_F \le 4\|A^{-1}\|\|E\|_F.$$

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

$$\mathbf{u}\mu_{\mathsf{QR}}(n) \le 1. \tag{5.38}$$

Proposition 5.57 (Guarantees for Finite-Arithmetic Haar Unitaries)**.** *Suppose that* QR *satisfies the assumptions in Definition 5.6 and that it is designed to output upper triangular matrices with nonnegative entries on the diagonal.*[8] *If* $(V, R) = \mathsf{QR}(\widetilde{G_n})$*, then there is a Haar unitary matrix* U *and a random matrix* $\mathbf{E}$ *such that* $\widetilde{V} = U + E$*. Moreover, for every* $1 > \alpha > 0$ *and* $t > 2\sqrt{2} + 1$ *we have*

$$\mathbb{P}\left[\|\boldsymbol{E}\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u}\right] \ge 1 - 2e\alpha^2 - 2e^{-t^2 n}.$$

Proof. From our Gaussian sampling assumption, $\widetilde{G_n} = G_n + E$ where $\|E\| \le \sqrt{n} c_{\mathsf{N}}\mathbf{u}$. Also, by the assumptions on QR from Definition 5.6, there are matrices $\widetilde{\widetilde{G_n}}$ and $\widetilde{V}$ such that $(\widetilde{V}, R) = \mathsf{qr}(\widetilde{\widetilde{G_n}})$, and

$$\begin{aligned}
&\|\widetilde{V} - V\| < \mu_{\mathsf{QR}}(n)\mathbf{u} \\
&\|\widetilde{\widetilde{G_n}} - \widetilde{G_n}\| \le \mu_{\mathsf{QR}}(n)\mathbf{u}\|\widetilde{G_n}\| \le \mu_{\mathsf{QR}}(n)\mathbf{u}\left(\|G_n\| + \sqrt{n} c_{\mathsf{N}}\mathbf{u}\right).
\end{aligned}$$

The latter inequality implies, using (5.38), that

$$\|\widetilde{\widetilde{G_n}} - G_n\| \le \mu_{\mathsf{QR}}(n)\mathbf{u}\left(\|G_n\| + \sqrt{n} c_{\mathsf{N}}\mathbf{u}\right) + \sqrt{n} c_{\mathsf{N}}\mathbf{u} \le \mu_{\mathsf{QR}}(n)\mathbf{u}\|G_n\| + 2\sqrt{n} c_{\mathsf{N}}\mathbf{u}. \tag{5.39}$$

Let $(U, R') \triangleq \mathsf{qr}(G_n)$. From Lemma 5.56 we know that U is Haar distributed on the unitary group, so using (5.39) and Lemma 2.12, and the fact that $\|M\| \le \|M\|_F \le \sqrt{n}\|M\|$ for any $n \times n$ matrix M, we know that

$$\|U - V\| - \mu_{\mathsf{QR}}(n)\mathbf{u} \le \|U - V\| - \|\widetilde{V} - V\| \le \|U - \widetilde{V}\| \le 4\sqrt{n} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u}\|G_n\|\|G_n^{-1}\| + 10n c_{\mathsf{N}}\mathbf{u}\|G_n^{-1}\|. \tag{5.40}$$

[8] Any algorithm that yields the QR decomposition can be modified in a stable way to satisfy this last condition at the cost of $O^*(n\log(1/\mathbf{u}))$ operations

Now, from $\|G_n^{-1}\| = 1/\sigma_n(G_n)$ and from Theorem 3.1 we have that

$$P\left[\|G_n^{-1}\| \geq \frac{n}{\alpha}\right] \leq (\sqrt{2e}\alpha)^2 = 2e\alpha^2.$$

On the other hand, from Lemma 3.5, 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 (5.40) 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. □

Preliminaries of RURV

Let $A \in \mathbb{C}^{n\times n}$ and $(U, R) = \mathsf{rurv}(A)$. As will become clear later, in order to analyze $\mathsf{SPAN}(A, r)$ 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 5.59 below, which is the complex analog to the upper bound given in equation (4) of [14, Theorem 5.1]. Actually, Corollary 5.59 is a direct consequence of [13, Lemma 4.1], stated below, and Proposition 5.52 proved above.

Lemma 5.58. *Let $n > r > 0$, $A \in \mathbb{C}^{n\times n}$ and $A = P\Sigma Q^*$ be its singular value decomposition. Let $(U, R) = \mathsf{rurv}(A)$, 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 5.58 and Proposition 5.52 gives the following result.

Corollary 5.59. *Let $n > r > 0$, $A \in \mathbb{C}^{n\times n}$, $(U, R) = \mathsf{rurv}(A)$ and R_{22} be the lower right $(n - r) \times (n - r)$ corner of R. Then for any $\theta > 0$*

$$\mathbb{P}\left[\|R_{22}\| \leq \frac{\sqrt{r(n - r)}}{\theta}\sigma_{r+1}(A)\right] \geq 1 - \theta^2.$$

Exact Arithmetic Analysis of SPAN

It is a standard consequence of the properties of the QR decomposition that if A is a matrix of rank r, then almost surely $\mathsf{span}(A, r)$ 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 $(U, R) = \mathsf{rurv}(A)$ 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 (5.31). 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

$$AV^* = UR = \begin{pmatrix} U_{11}R_{11} & U_{11}R_{12} + U_{12}R_{22} \\ U_{21}R_{11} & U_{21}R_{12} + U_{22}R_{22} \end{pmatrix}. \tag{5.41}$$

On the other hand, almost surely the first r columns of AV^* span the range of A. Using the right side of equation (5.41) 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 $\mathrm{Rank}(A) = \mathrm{Rank}(A^2)$.[9] We make this precise below and defer the proof to the end of the subsection.

Proposition 5.60 (Main Guarantee for span)**.** *Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that $\|A - \widetilde{A}\| \le \beta$ and $\mathrm{Rank}(A) = \mathrm{Rank}(A^2) = r$. Denote $S \triangleq \mathsf{span}(\widetilde{A}, r)$ and $T \triangleq \mathsf{span}(A, r)$. 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*

$$\|S - TW^*\| \le \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\beta}{\theta}}. \tag{5.42}$$

Remark 5.61 (Projectors)**.** In the case in which the matrix A of Proposition 5.60 is a (not necessarily orthogonal) projector, $T^*AT = I_r$, and the σ_r term in the denominator of (5.42) becomes a 1.

We now show that the orthogonal projection $P \triangleq \mathsf{span}(\widetilde{A}, r)\mathsf{span}(\widetilde{A}, r)^*$ is close to a projection onto the range of A, in the sense that $PA \approx A$.

[9] For example, diagonalizable matrices satisfy this criterion.

Lemma 5.62. *Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that* $\mathrm{Rank}(A) = r$ *and* $\|A - \widetilde{A}\| \leq \beta$. *Let* $(U, R) \triangleq \mathsf{rurv}(\widetilde{A})$ *and* $S \triangleq \mathsf{span}(\widetilde{A}, r)$. *Then, almost surely*

$$\|(SS^* - I_n)A\| \leq \|R_{22}\| + \beta, \tag{5.43}$$

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 $\mathsf{span}(\cdot,\cdot)$ outputs the first r columns of U, we have the block decomposition $U = \begin{pmatrix} S & U' \end{pmatrix}$, where $S \in \mathbb{C}^{n\times r}$ and $U' \in \mathbb{C}^{n\times(n-r)}$.

On the other hand we have $\widetilde{A} = URV$, so

$$(SS^* - I_n)\widetilde{A} = (SS^* - I)\begin{pmatrix} S & U' \end{pmatrix} RV = \begin{pmatrix} 0 & -U' \end{pmatrix} RV = \begin{pmatrix} 0 & -U'R_{2,2} \end{pmatrix} V.$$

Since $\|U'\| = \|V\| = 1$ from the above equation we get $\|(SS^* - I_n)\widetilde{A}\| \leq \|R_{22}\|$. Now we can conclude that

$$\|(SS^* - I_n)A\| \leq \|(SS^* - I_n)\widetilde{A}\| + \|(SS^* - I_n)(A - \widetilde{A})\| \leq \|R_{22}\| + \beta.$$

□

The inequality (5.43) can be applied to quantify the distance between the ranges of $\mathsf{span}(\widetilde{A}, r)$ and $\mathsf{span}(A, r)$ in terms of $\|R_{22}\|$, as the following result shows.

Lemma 5.63 (Bound in Terms of $\|R_{22}\|$)**.** *Let $\beta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that* $\mathrm{Rank}(A) = \mathrm{Rank}(A^2) = r$ *and* $\|A - \widetilde{A}\| \leq \beta$. *Denote by* $(U, R) \triangleq \mathsf{rurv}(\widetilde{A})$, $S \triangleq \mathsf{span}(\widetilde{A}, r)$ *and* $T \triangleq \mathsf{span}(A, r)$. *Then, almost surely there exists a unitary $W \in \mathbb{C}^{r\times r}$ such that*

$$\|S - TW^*\| \leq 2\sqrt{\frac{\|R_{22}\| + \beta}{\sigma_r(T^*AT)}}, \tag{5.44}$$

where R_{22} is the lower right $(n-r)\times(n-r)$ block of R.

Proof. From Lemma 5.62 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

$$\|T^*SS^*T - I_r\| = \sup_{w\in\mathbb{C}^r, \|w\|=1} \|T^*(SS^* - I_r)Tw\| = \sup_{w\in\mathrm{range}(A), \|w\|=1} \|T^*(SS^* - I_r)w\|. \tag{5.45}$$

Now, since $\mathrm{Rank}(A) = \mathrm{Rank}(A^2)$, if $w \in \mathrm{range}(A)$ then $w = Av$ for some $v \in \mathrm{range}(A)$. So by the Courant-Fischer formula

$$\frac{\|w\|}{\|v\|} = \frac{\|Av\|}{\|v\|} \geq \inf_{u\in\mathrm{range}(A)} \frac{\|Au\|}{\|u\|} = \sigma_r(T^*AT).$$

We can then revisit (5.45) and get

$$\sup_{w\in\text{range}(A),\|w\|=1} \|T^*(SS^*-I_r)w\| = \sup_{v\in\text{range}(A),\|v\|\le 1} \frac{\|T^*(SS^*-I_r)Av\|}{\sigma_r(T^*AT)} \le \frac{\|T^*(SS^*-I_r)AT\|}{\sigma_r(T^*AT)}. \tag{5.46}$$

On the other hand $\|T^*(SS^*-I_r)AT\| \le \|(SS^*-I_r)A\| \le \|R_{22}\| + \beta$, so combining this fact with (5.45) and (5.46) we obtain

$$\|T^*SS^*T - I_r\| \le \frac{\|R_{22}\|+\beta}{\sigma_r(T^*AT)}.$$

Now define $X \triangleq S^*T$, $\beta' \triangleq \frac{\|R_{22}\|+\beta}{\sigma_r(T^*AT)}$ and let $X = W|X|$ be the polar decomposition of X. Observe that

$$\||X| - I_r\| \le \sigma_1(X) - 1 \le |\sigma_1(X)^2 - 1| = \|X^*X - I_r\| \le \beta'.$$

Thus $\|S^*T - W\| = \|X - W\| = \|(|X| - I_n)W\| \le \beta'$. Finally note that

$$\begin{aligned}
\|S - TW^*\|^2 &= \|(S^* - WT^*)(S - TW^*)\| \\
&= \|2I_r - S^*TW^* - WT^*S\| \\
&= \|2I_r - S^*T(T^*S + W^* - T^*S) - (S^*T + W - S^*T)T^*S\| \\
&\le 2\|I_r - S^*TT^*S\| + \|S^*T(W^* - T^*S)\| + \|(W - S^*T)T^*S\| \le 4\beta',
\end{aligned}$$

which concludes the proof. □

Note that so far our results have been deterministic. The possibility of failure of the guarantee given in Proposition 5.60 comes from the non-deterministic bound on $\|R_{22}\|$.

Proof of Proposition 5.60. From the stability of singular values we have $\sigma_{r+1}(\widetilde{A}) \le \beta$. Now combine Lemma 5.63 with Corollary 5.59. □

Finite Arithmetic Analysis of SPAN

In what follows we will have an approximation $\widetilde{A}$ of a matrix A of rank r with the guarantee that $\|A - \widetilde{A}\| \le \beta$.

For the sake of readability we will not present optimal bounds for the error induced by roundoff, and we will assume that

$$4\|A\| \cdot \max\{c_{\mathsf{N}}\mu_{\mathsf{MM}}(n)\mathbf{u}, c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u}\} \le \beta \le \frac{1}{4} \le \|A\| \quad \text{and} \quad 1 \le \min\{\mu_{\mathsf{MM}}(n), \mu_{\mathsf{QR}}(n), c_{\mathsf{N}}\}. \tag{5.47}$$

We begin by analyzing the subroutine RURV in finite arithmetic. This was done in [61, 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 5.64 (Analysis of RURV). *Assume that* QR *and* MM *satisfy the guarantees in Definitions 5.4 and 5.6. Also suppose that the assumptions in (5.47) hold. Then, if* $(U, R) \triangleq \mathsf{RURV}(A)$ *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:*

$$\|\widetilde{V} - V\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha}\mathbf{u} \text{ and } \|A - \widetilde{A}\| < \|A\| \left(\frac{9tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u} \right) \tag{5.48}$$

occurs with probability at least $1 - 2e\alpha^2 - 2e^{-t^2 n}$.

Proof. By definition $V = \mathsf{QR}(\widetilde{G}_n)$ with $\widetilde{G}_n = G_n + E$, where G_n is an $n \times n$ Ginibre matrix and $\|E\| \leq \sqrt{n}\mathbf{u}$. A direct application of the guarantees on each step yields the following:

1. From Proposition 5.57, we know that there is a Haar unitary $\widetilde{V}$ and a random matrix E_0, such that $V = \widetilde{V} + E_0$ and

$$\mathbb{P}\left[\|E_0\| < \frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u}\right] \geq 1 - 2e\alpha^2 - 2e^{-t^2 n}. \tag{5.49}$$

2. If $B \triangleq \mathsf{MM}(A, V^*) = AV^* + E_1$, 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 (5.47). 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 (5.47) again, we get

$$\|E_1\| \leq \frac{5}{4}\|A\|\mu_{\mathsf{MM}}(n)\mathbf{u} \quad \text{and} \quad B \leq \frac{5}{4}\|A\|(1 + \mu_{\mathsf{MM}}(n)\mathbf{u}) < 2\|A\|. \tag{5.50}$$

3. Let $(U, R) = \mathsf{QR}(B)$. Then there is a unitary $\widetilde{U}$ and a matrix $\tilde{B}$ such that $U = \widetilde{U} + E_2$, $B = \tilde{B} + E_3$, and $\tilde{B} = \widetilde{U}R$, with error bounds $\|E_2\| \leq \mu_{\mathsf{QR}}(n)\mathbf{u}$ and $\|E_3\| \leq \|B\|\mu_{\mathsf{QR}}(n)\mathbf{u}$. Using (5.50) we obtain

$$\|E_3\| \leq \|B\|\mu_{\mathsf{QR}}(n)\mathbf{u} < 2\|A\|\mu_{\mathsf{QR}}(n)\mathbf{u}. \tag{5.51}$$

4. Finally, define $\widetilde{A} \triangleq \widetilde{B}\widetilde{V}$. Note that $\widetilde{A} = \widetilde{U}R\widetilde{V}$ and

$$\widetilde{A} = \widetilde{B}\widetilde{V} = (B - E_3)\widetilde{V} = (AV^* + E_1 - E_3)\widetilde{V} = (A(\widetilde{V} + E_0)^* + E_1 - E_3)\widetilde{V} = A + (AE_0^* + E_1 - E_3)\widetilde{V},$$

which translates into

$$\|A - \widetilde{A}\| \le \|A\|\|E_0\| + \|E_1\| + \|E_3\|.$$

Hence, on the event described in the left side of (5.49), we have

$$\|A - \widetilde{A}\| \le \|A\| \left(\frac{8tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{10n^2}{\alpha} c_{\mathsf{N}}\mathbf{u} + \frac{5}{4}\mu_{\mathsf{MM}}(n)\mathbf{u} + 2\mu_{\mathsf{QR}}(n)\mathbf{u} \right),$$

and using some crude bounds, the above inequality yields the advertised bound.

□

We can now prove a finite arithmetic version of Proposition 5.60.

Proposition 5.65 (Main Guarantee for SPAN)**.** *Let $n > r$ be positive integers, and let $\beta, \theta > 0$ and $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ be such that $\|A - \widetilde{A}\| \le \beta$ and* $\mathrm{Rank}(A) = \mathrm{Rank}(A^2) = r$*. Let $S \triangleq \mathsf{SPAN}(\widetilde{A}, r)$ and $T \triangleq \mathsf{span}(A, r)$. If* QR *and* MM *satisfy the guarantees in Definitions 5.4 and 5.6, and (5.47) holds, then, for every $t > 2\sqrt{2} + 1$ there exist a unitary $W \in \mathbb{C}^{r\times r}$ such that*

$$\|S - TW^*\| \le \mu_{\mathsf{QR}}(n)\boldsymbol{u} + 12\sqrt{\frac{tn^2\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\beta}{\theta^2}}, \tag{5.52}$$

with probability at least $1 - 7\theta^2 - 2e^{-t^2 n}$.

Proof. Let $(U, R) = \mathsf{RURV}(\widetilde{A})$. From Lemma 5.64 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 $(\widetilde{U}, R) = \mathsf{rurv}(\widetilde{\widetilde{A}})$ for the respective realization of an exact Haar unitary matrix. Then, from $\|\widetilde{A}\| \le \|A\| + \beta$ and (5.48), for every $1 > \alpha > 0$ and $t > 2\sqrt{2} + 1$ we have

$$\left\|A - \widetilde{\widetilde{A}}\right\| \le \left\|\widetilde{\widetilde{A}} - \widetilde{A}\right\| + \|\widetilde{A} - A\| \le (\|A\| + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha}\mu_{\mathsf{QR}}(n)c_{\mathsf{N}}\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha}c_{\mathsf{N}}\mathbf{u} \right) + \beta, \tag{5.53}$$

with probability $1 - 2e\alpha^2 - 2e^{-t^2 n}$.

Now, from (5.47) we have $\mathbf{u} \le \beta \le \frac{1}{4}$ and $c_{\mathsf{N}}\|A\|\mu\mathbf{u} \le \beta$ for $\mu = \mu_{\mathsf{QR}}(n), \mu_{\mathsf{MM}}(n)$, so we can bound the respective terms in (5.53) by β:

$$\begin{aligned}(\|A\| + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha} c_{\mathsf{N}}\mu_{\mathsf{QR}}(n)\mathbf{u} + 2\mu_{\mathsf{MM}}(n)\mathbf{u} + \frac{10n^2}{\alpha}c_{\mathsf{N}}\mathbf{u} \right) + \beta &\le (1 + \beta)\left(\frac{9tn^{\frac{3}{2}}}{\alpha}\beta + 2\beta + \frac{10n^2}{\alpha}\beta \right) + \beta \\ &\le \frac{(12t + 16)}{\alpha} n^2\beta, \end{aligned} \tag{5.54}$$

where the last crude bound uses $1 \le n^{\frac{3}{2}} \le n^2, 1 + \beta \le \frac{5}{4}$ and $t > 2$.

Observe that $\tilde{S} = \mathsf{span}(\widetilde{\widetilde{A}}, r)$ is the matrix formed by the first r columns of $\widetilde{U}$, and that by Proposition 5.60 we know that for every $\theta > 0$, with probability $1 - \theta^2$ there exists a unitary W such that

$$\|\tilde{S} - TW^*\| \le \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\left\|A - \widetilde{\widetilde{A}}\right\|}{\theta}}. \tag{5.55}$$

On the other hand, S is the matrix formed by the first r columns of U. Hence

$$\|S - \tilde{S}\| \le \|U - \widetilde{U}\| \le \mu_{\mathsf{QR}}(n)\mathbf{u}.$$

Putting the above together we get that under this event

$$\|S - TW^*\| \le \|S - \tilde{S}\| + \|\tilde{S} - TW^*\| \le \mu_{\mathsf{QR}}(n)\mathbf{u} + \sqrt{\frac{8\sqrt{r(n-r)}}{\sigma_r(T^*AT)}} \cdot \sqrt{\frac{\left\|A - \widetilde{\widetilde{A}}\right\|}{\theta}}. \tag{5.56}$$

Now, taking $\alpha = \theta$, we note that both events in (5.53) and (5.55) 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 (5.53), (5.54) and (5.56). □

We end by proving Theorem 5.17, the guarantees on SPAN that we will use when analyzing the main algorithm.

Proof of Theorem 5.17. As Remark 5.61 points out, in the context of this theorem we are passing to SPAN an approximate projector $\tilde{P}$, and the above result simplifies. Using this fact, as well as the upper bound $r(n-r) \le n^2/4$, we get that

$$\|S - TW^*\| \le \mu_{\mathsf{QR}}(n)\mathbf{u} + \frac{12\sqrt{tn^3\beta}}{\theta}.$$

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 $\beta \le 1/4$, 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

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

□

Remark 5.66. Note that the exact same proof of Theorem 5.17 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 SPAN that appears in Section 5.8.

5.9 Discussion

In this chapter, we reduced the approximate diagonalization problem to a polylogarithmic number of matrix multiplications, inversions, and QR factorizations on a floating point machine with precision depending only polylogarithmically on n and $1/\delta$. The key phenomena enabling this were: (a) every matrix is δ-close to a matrix with well-behaved pseudospectrum, and such a matrix can be found by a complex Gaussian perturbation; and (b) the spectral bisection algorithm can be shown to converge rapidly to a forward approximate solution on such a well-behaved matrix, using a polylogarithmic in n and $1/\delta$ amount of precision and number of iterations. The combination of these facts yields a δ-backward approximate solution for the original problem.

Using fast matrix multiplication, we obtain algorithms with nearly optimal asymptotic computational complexity (as a function of n, compared to matrix multiplication), for general complex matrices with no assumptions. Using naïve matrix multiplication, we get easily implementable algorithms with $O(n^3)$ type complexity and much better constants which are likely faster in practice. The constants in our bit complexity and precision estimates (see Theorem 5.19 and equations

(5.30) and (5.10)), while not huge, are likely suboptimal. The reasonable practical performance of spectral bisection based algorithms is witnessed by the many empirical papers (see e.g. [11]) which have studied it. The more recent of these works further show that such algorithms are communication-avoiding and have good parallelizability properties.

Remark 5.67 (Hermitian Matrices)**.** A curious feature of our algorithm is that even when the input matrix is Hermitian or real symmetric, it begins by adding a complex non-Hermitian perturbation to regularize the spectrum. If one is only interested in this special case, one can replace this first step by a Hermitian GUE or symmetric GOE perturbation and appeal to the result of [2] instead of Theorem 1.10, which also yields a polynomial lower bound on the minimum gap of the perturbed matrix. It is also possible to obtain a much stronger analysis of the Newton iteration in the Hermitian case, since the iterates are all Hermitian and $\kappa_V = 1$ for such matrices. By combining these observations, one can obtain a running time for Hermitian matrices which is significantly better (in logarithmic factors) than our main theorem. We do not pursue this further since our main goal was to address the more difficult non-Hermitian case.

We conclude by listing several problems which merit further study.

Open Problem 5.68. *Devise a deterministic algorithm with similar guarantees to* EIG*.*

The main bottleneck to doing this is deterministically finding a regularizing perturbation, which seems quite mysterious. Another obstacle is computing a rank-revealing QR factorization in near matrix multiplication time deterministically, as all of the currently known deterministic algorithms require $\Omega(n^3)$ time.

Open Problem 5.69. *Reduce the dependence of the running time and precision to a smaller power of* $\log(1/\delta)$*.*

The Shifted QR algorithm we analyze in Chapter 7 improves the precision somewhat, still short of the holy grail of $O(\log(1/\delta))$ bits of precision. The bottleneck in the current algorithm is the precision required for stable convergence of the Newton iteration for computing the sign function. Other, "inverse-free" iterative schemes have been proposed for this, which conceivably require lower precision.

Open Problem 5.70. *Study the convergence of "scaled Newton iteration" and other rational approximation methods (see [95, 118]) for computing the sign function on non-Hermitian matrices.*

More broadly, we hope that the techniques introduced in this chapter—pseudospectral shattering and pseudospectral analysis of matrix iterations using contour integrals—are useful in attacking other problems in numerical linear algebra.

5.10 Deferred Proofs from Section 5.6

Restatement of Lemma 5.40. *Assume the matrix inverse is computed by an algorithm* INV *satisfying the guarantee in Definition 5.5. Then* $\mathsf{G}(A) = g(A) + E$ *for some error matrix* E *with norm*

$$\|E\| \le \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|\right) 2\sqrt{n}\mathbf{u}. \tag{5.57}$$

Proof. The computation of $\mathsf{G}(A)$ consists of three steps:

1. Form A^{-1} according to Definition 5.5. This incurs an additive error of $E_{\mathsf{INV}} = \mu_{\mathsf{INV}}(n)\cdot\mathbf{u}\cdot\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|$. The result is $\mathsf{INV}(A) = A^{-1} + E_{\mathsf{INV}}$.
2. Add A to $\mathsf{INV}(A)$. This incurs an entry-wise relative error of size $\mathbf{u}$: The result is
$$(A + A^{-1} + E_{\mathsf{INV}}) \circ (J + E_{add})$$
where J denotes the all-ones matrix, $\|E_{add}\|_{max} \le \mathbf{u}$, and where $\circ$ denotes the entrywise (Hadamard) product of matrices.
3. Divide the resulting matrix by 2, 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 (5.4)

$$\|M \circ E\| \le \|M\|\|E\|_{max}\sqrt{n}.$$

Putting it all together, we have

$$\begin{aligned}\|\mathsf{G}(A) - g(A)\| &\le \frac{1}{2}\left(\|A\| + \|A^{-1}\|\right)\mathbf{u}\sqrt{n} + \|E_{\mathsf{INV}}\|(1+\mathbf{u})\sqrt{n} \\ &\le \frac{1}{2}\left(\|A\| + \|A^{-1}\|\right)\mathbf{u}\sqrt{n} + \mu_{\mathsf{INV}}(n)\cdot\mathbf{u}\cdot\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|(1+\mathbf{u})\sqrt{n} \\ &\le \left(\|A\| + \|A^{-1}\| + \mu_{\mathsf{INV}}(n)\kappa(A)^{c_{\mathsf{INV}}\log n}\|A^{-1}\|\right) 2\sqrt{n}\mathbf{u}\end{aligned}$$

where we use $\mathbf{u} < 1$ in the last line. □

In what remains of this section we will repeatedly use the following simple calculus fact.

Lemma 5.71. *Let* $x, y > 0$*, then*

$$\log(x+y) \le \log(x) + \frac{y}{x} \quad \text{and} \quad \lg(x+y) \le \lg(x) + \frac{1}{\log 2}\frac{y}{x}.$$

Proof. This follows directly from the concavity of the logarithm. □

Restatement of Lemma 5.44. *Let* $1/800 > t > 0$ *and* $1/2 > c > 0$ *be given. Then for*

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

we have

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

Proof. An exact solution for j can be written in terms of the *Lambert W-function*; see [51] 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 5.71 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. □

Restatement of Lemma 5.47. *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 5.71, 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

$$N = \lceil \lg(1/s) + 3\lg\lg(1/s) + \lg\lg(1/(\beta\epsilon_0)) + 7.59 \rceil.$$

□

Bibliographic Note

This chapter (including the two figures) is lightly adapted Sections 1-2, 4-6, and Appendices A-D of [15].

Chapter 6

The Shifted QR Algorithm in Exact Arithmetic

6.1 Introduction

The Hessenberg Shifted QR Algorithm, discovered in the late 1950s independently by Francis [78, 79] and Kublanovskaya [106], has been for several decades the most widely used method for approximately computing all of the eigenvalues of a dense matrix. 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 [67, 128]. As discussed in Chapter 1, the algorithm is specified by a *shifting strategy*, which is an efficiently computable function

$$\mathsf{Sh} : \mathbb{H}^{n\times n} \longrightarrow \mathcal{P}_k,$$

where $\mathbb{H}^{n\times n}$ is the set of $n \times n$ complex Hessenberg 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:

$$H_{t+1} = Q_t^* H_t Q_t \qquad\qquad \text{Where } p_t(H_t) = Q_t R_t, \text{ for } p_t = \mathsf{Sh}(H_t) \tag{6.1}$$

The expression $p_t(H_t) = Q_t R_t$ is (6.1) is a QR decomposition so that Q_t is unitary, and it is not hard to see that each iteration preserves the Hessenberg structure.

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 [78, 106] 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 both power and inverse iteration (the paper [162] articulates this connection and is an essential reference, see also [152]). 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 these ratios and thereby accelerate convergence. The challenge is that this must be done efficiently without prior knowledge of the eigenvalues.

As discussed in Chapter 1, when looking for a backward approximation to the eigenvalues of H_0, is not advisable to wait around for the *entire* subdiagonal of H_t to become small. Instead, the relevant notion of convergence is the time required for *some* subdiagonal entry to become so: once this happens, we can *deflate* H_t by deleting the small entry (or entries), and continue the iteration separately on the diagonal blocks of the resulting block upper triangular matrix. We therefore quantify the rate of convergence of (6.1) in terms of the *ω-decoupling time*, or in other words the number of iterations required to drive at least one subdiagonal entry below $\omega\|H_t\|$. Each deflation introduces a backward error of $\omega\|H_t\|$ into the calculation, so relative accuracy δ is ensured by taking $\omega = O(\delta/n)$. In this setup "rapid" convergence means that the ω-decoupling time is a very slowly growing function of n and $1/\omega$, ideally logarithmic or polylogarithmic.

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. These strategies typically have attractive fixed points and cycles which are not upper triangular, leading to slow convergence or nonconvergence (e.g. see [125, 22, 56]). 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.

[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.

Example 6.1. Both pathologies are seen in the instructive family of $n \times n$ examples

$$M = \begin{pmatrix} & & & & m_n \\ m_1 & & & & \\ & m_2 & & & \\ & & \ddots & & \\ & & & m_{n-1} & \end{pmatrix}$$

where $m_1, \dots, m_n \in (0,1)$. Observe that, for $k \le n-1$, 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 m_i, while leaving the zero pattern of M intact. This means that for adversarially chosen $m_1, ..., m_n$, the bottom few subdiagonal entries of M — the traditional place to look for monotonicity in order to prove convergence — exhibit arbitrary behavior. At very long time scales of n steps, the behavior becomes periodic and predictable, but there is still no convergence.

Previous approaches this problem 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. Building off of the sketched shifting strategy for normal matrices in the introduction, the key idea is that when the eigenvector condition number $\kappa_V(H_0)$ is bounded, the dynamics of shifted QR can be understood in terms of certain measures, similar in spirit to the notion of spectral measure of a normal matrix used in Chapter 1, associated with the (not necessarily normal) iterates H_t. On short time scales, these measures lack any semblance of the monotonicity exhibited by the shift from the introduction, but over time scales of $k \gg \lg \kappa_V(H_0)$ they evolve in a predictable way, much like in the normal case. This is explained in detail in Section 6.4. Moreover, the behavior of these measures can be related to the *geometric mean* of the bottom k subdiagonal entries of the current iterate, a quantity we will write as

$$\psi_k(H) \triangleq (H_{n-k,n-k+1} \cdots H_{n-1,n})^{1/k} \tag{6.2}$$

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 6.1, it can be seen that the ratios of the m_i cannot be arbitrary and the geometric mean of the bottom $\lg \kappa_V(M)$ subdiagonal entries of M behaves predictably rather than chaotically on intervals of $k \gg \lg \kappa_V(M)$ time steps.

Guided by this insight, we carefully design a shifting strategy which satisfies the following dichotomy: either (i) a QR step of degree k significantly decreases the potential definedin (6.2), or, (ii) the measure associated to the current iterate must have a special structure. 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, giving linear convergence in either case. Thus, our proof articulates that transience and symmetry are the only obstacles to rapid convergence of the shifted QR iteration on nonsymmetric matrices.

As discussed in Chapter 1, we define $\mathsf{Sh}_{k,B}(H_t)$ in terms of the *Ritz values* of the current iterate H_t. Recall that the Ritz values of order k of a Hessenberg matrix H are the eigenvalues of its bottom right $k \times k$ corner $H_{(k)}$; they are related to the potential $\psi_k(H)$ via the variational characterization of the latter which we discussed in the introduction and pause here to restate and prove [152, Theorem 34.1]. Let us write $\chi_k(z) = \det(z - H_{(k)})$.

Lemma 6.2 (Variational Formula for ψ_k)**.** *For any Hessenberg $H \in \boldsymbol{H}^{n\times n}$ and any $k \in \mathbb{N}$,*

$$\psi_k(H) = \min_{p\in\mathcal{P}_k} \|e_n^* p(H)\|^{1/k},$$

with the minimum attained for $p = \chi_k$.

Proof. Since H is upper Hessenberg, for any polynomial $p \in \mathcal{P}_k$ we have

$$p(H)_{n,n-j} = \begin{cases} p(H_{(k)})_{k,k-j+1} & j = 0,\dots,k-1, \\ H_{n-k,n-k-1}\cdots H_{n,n-1} & j = k, \\ 0 & j \geq k+1. \end{cases}$$

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. □

Since computing eigenvalues exactly is impossible when $k \geq 5$, we assume access to a method for computing approximate Ritz values, in a specific sense motivated by Lemma 6.2 and encapsulated in the following definition.

Definition 6.3 (θ-Optimal Ritz Values and Ritz Value Finders)**.** Let $\theta \geq 0$. We call $\mathcal{R} = \{r_1,\dots,r_k\} \subset \mathbb{C}$ a set of *θ-optimal Ritz values* of a Hessenberg matrix H if

$$\Big\|e_n^* \prod_{i\leq k}(H - r_i)\Big\|^{1/k} \leq \theta\psi_k(H) = \theta \min_{p\in\mathcal{P}_k} \|e_n^* p(H)\|^{1/k}. \tag{6.3}$$

A *Ritz value finder* is an algorithm $\mathsf{OptRitz}(H,k,\theta)$ that takes as inputs a Hessenberg matrix $H \in \boldsymbol{C}^{n\times n}$, a positive integer k and an accuracy parameter $\theta > 1$, and outputs a set $\mathcal{R} = \{r_1,\dots,r_k\}$ of θ-optimal Ritz values of H whenever the right hand side of (6.3) is nonzero. Let $T_{\mathsf{OptRitz}}(k,\theta,\omega)$

be the maximum number of arithmetic operations used by $\mathsf{OptRitz}(H, k, \theta)$ over all inputs H such that the $\psi_k(H) \geq \omega\|H\|$.[2]

With some additional insights, a Ritz value finder satisfying Definition 6.3 can be efficiently instantiated using polynomial root finders or other provable eigenvalue computation algorithms (e.g. Theorem 1.7's EIG) with guarantees of type $T_{\mathsf{OptRitz}}(\theta, k, \omega) = O(k^c \lg(\frac{1}{\omega(\theta-1)}))$. We defer a detailed discussion of numerical precision issues surrounding this implementation to Chapter 7. The subtlety of not being able to compute Ritz values exactly is secondary to the dynamical phenomena which are the focus of the current chapter, so on first reading it is recommended to assume $\theta = 1$ (i.e., Ritz values are computed exactly), even though this is unrealistic when $k > 4$. Here, as in Chapter 7, we will take $\theta = 2$. We now restate Theorem 1.13 precisely; it will be proved in Section 6.4.

Restatement of Theorem 1.13. *For each $k = 2, 4, 8, \ldots$ and $B \geq 1$, there is a shifting strategy $\mathsf{Sh}_{k,B} : \mathbb{H}_B^{n\times n} \longrightarrow \mathcal{P}_k$ which, in exact arithmetic,*

(i) achieves ω-decoupling in at most $4\lg(1/\omega)$ iterations, and

(ii) costs at most

$$\left(\lg k + N_{\mathsf{net}}\left(0.002\, B^{-\frac{8\lg k+4}{k-1}}\right)\right) \cdot 7kn^2 + T_{\mathsf{OptRitz}}(k, 2, \omega) + \lg k \tag{6.4}$$

arithmetic operations per iteration before ω-decoupling occurs, where $N_{\mathsf{net}}(\varepsilon) = O(\varepsilon^{-2})$ is the number of points in an efficiently computable ε-net of the unit disk.

The term involving N_{net} captures the the cost of performing certain "exceptional shifts" (see Section 5.2) used in the strategy and the number $7kn^2$ corresponds to an upper bound on the arithmetic cost of a degree k implicit QR step (see Section 6.4).

The tradeoff between the nonnormality of the input matrix and the efficiency of the shifting strategy appears in the cost of the exceptional shift, and we can see that by setting

$$k = \Omega(\lg B \lg\lg B) \tag{6.5}$$

yields a total running time of $O(n^2 k \lg k)$ 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, which causes the arithmetic complexity of each iteration to depend polynomially on B, rather than logarithmically.

[2] This lower bound is no issue since we are using the $\mathsf{OptRitz}$ to ω-decouple H, but is necessary since otherwise we could use $\mathsf{OptRitz}$ to compute the eigenvalues of $H_{(k)}$ to arbitrary accuracy in finite time, which is unrealistic.

Remark 6.4. For normal matrices, we will show for good measure that $\mathsf{Sh}_{4,1}$ achieves rapid decoupling at a cost of at most $1500n^2 + O(1)$ arithmetic operations per iteration; the reason is that we may take $B = 1$ and (since the $k = 4$ Ritz values are the roots of a quartic, and can thus be computed exactly using $O(1)$ arithmetic operations) $\theta = 1$ as well.

Remark 6.5 (Higher Degree Shifts)**.** It is well-known, and verified in the sequel, that a QR step with a degree k shift $p(z) = (z - r_1) \dots (z - r_k)$ is identical to a sequence of k steps with degree 1 shifts $(z - r_1), (z - r_2), \dots, (z - r_k)$, so any degree k strategy can be simulated by a degree 1 strategy while increasing the iteration count by a factor of k.[3] 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 [37, Section 3] and $k = 50$ is often used in practice [105].

6.2 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 [24, 141, 47] or [128, 162, 84] for a dynamical or numerical viewpoint, respectively, or to the books [85, 152, 64, 161] 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. [161]. 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 [69, 156, 157, 158]).

Shifting strategies which combine more than one of these through some kind of case analysis are called "mixed" strategies.

Symmetric Matrices. In a celebrated work, Wilkinson [164] proved global convergence of shifted QR on all *real symmetric* tridiagonal matrices using the shifting strategy that now carries his name. The linear convergence bound of ω-decoupling in $O(\lg(1/\omega))$ iterations for this shifting strategy was then obtained by Dekker and Traub [60] (in the more general setting of Hermitian matrices), and reproven by Hoffman and Parlett [96] using different arguments. Other than these results for Hermitian matrices, there is no known bound on the worst-case decoupling time of shifted QR

[3] This also has some important advantages with regards to numerical stability, which are discussed in Chapter 7.

for any large class of matrices or any other shifting strategy. Shifted QR is nonetheless the most commonly used algorithm for the nonsymmetric eigenproblem on dense matrices, though there is little theoretical foundation for this practice. The strategies implemented in standard software libraries heuristically converge very rapidly on "typical" inputs, but occasionally examples of nonconvergence are found [56, 117] and dealt with in ad hoc ways.

In the realm of higher order shifts, Jiang [75] 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. [152]). In particular, the Rayleigh shifting strategy corresponds to a vector iteration process known as Rayleigh Quotient Iteration (RQI). Parlett [126] (building on [123, 41, 129]) 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 [22] 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 [23] 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 [125] 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 [26, 27] 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 [25] 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 [56], 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 library EISPACK fail to converge rapidly; such an example was independently discovered by Moler [117]. 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 $\mathsf{Sh}_{k,B}$ with modest parameters $k = B = 2$ is guaranteed to converge rapidly on them.

Using topological considerations, Leite et al. [109] proved that no single shifting strategy which is continuous in the entries of the matrix can cause decoupling on every symmetric matrix. Accordingly (in retrospect), the most successful shifting strategy for symmetric matrices, Wilkinson's, is discontinuous and explicitly breaks symmetry when the latter occurs. Our strategy $\mathsf{Sh}_{k,B}$ is also discontinuous in the entries of the matrix.

Mixed and Exceptional Shifts. Eberlein and Huang [69] showed global convergence (without a bound on the rate) of a certain mixed strategy for unitary Hessenberg matrices; more recently, the works [156, 157, 158] 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 $\mathsf{Sh}_{k,B}$ 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 [69, 157] — 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 [78, 60], but was popularized in by Bai and Demmel in [9], who observed that higher order shifts can sometimes be implemented more efficiently than a sequence of lower order ones; see [9, 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 [57]; such a correspondence is not known for any nontrivial shifting scheme or for nonnormal matrices. See [47] for a detailed survey of this connection. More recently, the line of work [131, 59, 58] 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.

6.3 Preliminaries

We begin with some preliminaries on QR steps in exact arithmetic. Recall our notation $[Q, R] = \mathsf{qr}(A)$ for the *QR decomposition* of an matrix A, where Q is unitary and R is upper triangular with nonnegative diagonal entries. Given a polynomial $p(z)$ and a Hessenberg matrix H, $\mathsf{iqr}(H, p(z))$ will denote the matrix $\widehat{H} = Q^* H Q$ where $[Q, R] = p(H)$. When $p(z) = z - s$ we will use $\mathsf{iqr}(H, s)$ as a shorthand notation for $\mathsf{iqr}(H, z - s)$. The *QR* iteration iqr has a fundamental composition property articulated in the lemma below; the proof is standard, but we will need to adapt it in Chapter 7 so we include it for the reader's convenience.

Lemma 6.6. *For any invertible H and polynomial $p(z) = (z - r_1) \cdots (z - r_k)$,*

$$\mathsf{iqr}(H, p(z)) = \mathsf{iqr}(\cdots \mathsf{iqr}(\mathsf{iqr}(H, r_1), r_2), \ldots, r_k). \tag{6.6}$$

Moreover, if $p(H) = QR$, $H_1 = H$, and for each $\ell \in [k]$ we set $[Q_\ell, R_\ell] \triangleq \mathsf{qr}(H_\ell - r_\ell)$ and $H_{\ell+1} \triangleq Q_\ell^* H_\ell Q_\ell$, then

$$Q = Q_1 \cdots Q_k \quad \textit{and} \quad R = R_k R_{k-1} \cdots R_1. \tag{6.7}$$

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 repeatedly pass the product $Q_1 \cdots Q_\ell$ across the term $H_1 - r_\ell$ and then use the fact that $H_\ell - r_\ell = Q_\ell R_\ell$. 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 (6.6) is then immediate. □

The following corollary will be repeatedly useful.

Lemma 6.7. *Under the hypotheses of Lemma 6.6,*

$$\|e_n^* p(H)^{-1}\|^{-1} = R_{n,n} = (R_1)_{n,n} \cdots (R_k)_{n,n} \tag{6.8}$$

Proof. Maintaining the notation of Lemma 6.6, 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 (6.7) implies $R_{n,n} = (R_1)_{n,n} \cdots (R_k)_{n,n}$. □

The "i" in iqr is for "implicit," since one of the many virtues of Hessenberg matrices is that one can execute a shifted QR iteration step in $O(n^2)$ arithmetic operations, without fully computing a QR decomposition [162, Section 3, e.g.]. In Chapter 7 we will implement a degree-1 implicit QR step using $7n^2$ arithmetic operations by way of 2 × 2 *Givens rotations*, which are the unitary matrices mapping

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \sqrt{|x|^2 + |y|^2} \\ 0 \end{pmatrix}.$$

Since for any $s \in \mathbb{C}$ the shifted matrix $H - s$ is still upper Hessenberg, we can bring it to upper triangular form by applying $n - 1$ sequential 2×2 Givens rotations on the left, each zeroing out one of the subdiagonal entries in $O(n)$ arithmetic operations. The resulting matrix is upper triangular with nonnegative entries, so it must be the R in the QR decomposition of $H - s$. To obtain $\widehat{H} = \mathsf{iqr}(H, s)$, it suffices to apply the same Givens rotations to R on the right, each again requiring $O(n)$ arithmetic operations. Using Lemma 6.6, we can therefore assume black box access to an *implicit QR algorithm* routine for efficiently performing a degree-k QR step with the guarantees below.

Definition 6.8 (Implicit QR Algorithm)**.** For $k \le n$, an efficient implicit QR algorithm $\mathsf{iqr}(H, p(z))$ takes as inputs a Hessenberg matrix $H \in \boldsymbol{C}^{n\times n}$ and a polynomial $p(z) = (z - s_1)\cdots(z - s_k)$ and, whenever $p(H)$ is invertible, outputs a Hessenberg matrix $\widehat{H}$ satisfying

$$\widehat{H} = Q^* H Q,$$

where $[Q, R] = \mathsf{qr}(p(H))$, as well as the number $R^{-1}_{n,n} = \|e_n^* p^{-1}(H)\|$. It runs in at most $7kn^2$ operations.

We have proposed to use the potential

$$\psi_k(H) \triangleq (H_{n-k,n-k+1} \cdots H_{n-1,n})^{1/k}$$

to track convergence of the QR iterates, and it will be accordingly useful to have a mechanism for proving upper bounds on the potential of $\widehat{H} = \mathsf{IQR}(H, p(z))$. To this end, for let $p \in \mathcal{P}_k$ and define

$$\tau_p(H) \triangleq \|e_n^* p(H)^{-1}\|^{-1/k}, \tag{6.9}$$

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[96, e.g.], 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 6.9 (Upper Bounds on $\psi_k(\widehat{H})$)**.** *Let $H \in \boldsymbol{C}^{n\times n}$ be a Hessenberg matrix, $p(z)$ a monic polynomial of degree k and $\widehat{H} = \mathsf{iqr}(H, p(z))$. Then*

$$\psi_k(\widehat{H}) \le \tau_p(H).$$

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 $\psi_k(\widehat{H}) = |p(\widehat{H})_{1,k+1}|^{1/k} = 0 = \tau_p(H)$. When $p(H)$ is invertible, applying Lemma 6.2 and using repeatedly that Q is unitary, R is triangular, and $p(H) = QR$,

$$\psi_k(\widehat{H})^k \le \|e_n^* p(\widehat{H})\| = \|e_n^* Q^* p(H)\| = \|e_n^* R\| = \|e_n^* R^{-1} Q^*\|^{-1} = \|e_n^* p(H)^{-1}\|^{-1} = \tau_p(H)^k.$$

□

Lemma 6.9 ensures that given H, we can reduce the potential with an implicit QR step by producing a polynomial p with $\|e_n^* p(H)^{-1}\|^{1/k} \leq \gamma \psi_k(H)$. 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.

The following notation will be used extensively throughout the this chapter and the next. Assume that $H = VDV^{-1}$ is diagonalizable, with V chosen so that $\|V\| = \|V^{-1}\| = \sqrt{\kappa_V(H)}$, D a diagonal matrix with $D_{i,i} = \lambda_i$, the eigenvalues of H. Write Z_H for the random variable[4] supported on the eigenvalues of 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. 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.

When H is normal, the distribution of Z_H is the spectral measure of H associated to e_n^* that we considered in the introduction. We checked there that by the functional calculus we have $\|e_n^* p(H)^{-1}\| = \mathbb{E}[|p(Z_H)|^{-2}]^{1/2}$, meaning that the (inverse) moments of Z_H are observable to us even without knowing the true eigenvectors and eigenvalues of H. The following lemma generalizes this fact to the nonnormal case, at a multiplicative cost of $\kappa_V(H)$.

Lemma 6.10 (Approximate Functional Calculus). *For any upper Hessenberg H and complex function f whose domain includes the eigenvalues of H,*

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

Proof. By the definition of Z_H above,

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

and the left hand inequality is analogous. □

Using this lemma with some carefully chosen rational functions f of degree k, we are able to probe the distribution of Z_H for each iterate H of the algorithm by examining the observable quantities $\|e_n^* f(H)\|^{1/k}$ — for appropriately large k, these are related to $(\mathbb{E}|f(Z_H)|^2)^{1/k}$ by a multiplicative factor of $\kappa_V(H)^{1/k} \approx 1$, 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] Our shifting strategy is deterministic, but we use random variables rather than measures for notational convenience.

6.4 Finite Arithmetic Analysis of $\mathsf{Sh}_{k,B}$

We will prove Theorem 1.13 by showing that, for some $\gamma \in (0,1)$, our shifting strategy $\mathsf{Sh}_{k,B}$ guarantees *potential reduction*, namely the efficient computation of a Hessenberg $\widehat{H}$, unitarily equivalent to H, with

$$\psi_k(\widehat{H}) \leq \gamma \psi_k(H).$$

It follows immediately that we can achieve ω-decoupling in $\frac{\lg \gamma}{\lg wacc}$ iterations.

The table below collates several constants which will appear throughout the chapter.

Symbol	Meaning	Typical Scale
H	Upper Hessenberg matrix	
B	Eigenvector condition bound	$B \geq \kappa_V(H)$
k	Shift degree	$O(\lg B \lg \lg B)$
ω	Decoupling parameter	
γ	Decoupling rate	0.8
θ	Ritz Value Optimality	2
α	Promising Ritz value parameter	$B^{4k^{-1}\lg k} = 1 + o(1)$

Table 6.1: Constants used by $\mathsf{Sh}_{k,B}$

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 6.11 (α-Promising Ritz value)**.** Let $\alpha \geq 1$, $\mathcal{R} = \{r_1, ..., r_k\}$ be a set of θ-optimal 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

$$\mathbb{E}\frac{1}{|Z_H - r|^k} \geq \frac{1}{\alpha^k}\mathbb{E}\frac{1}{|p(Z_H)|}. \tag{6.10}$$

Note that there is at least one 1-promising Ritz value in every set of θ-optimal Ritz values, since

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\frac{1}{|Z_H - r_i|^k} = \mathbb{E}\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|Z_H - r_i|^k} \geq \mathbb{E}\frac{1}{|p(Z_H)|} \tag{6.11}$$

by linearity of expectation and the inequality of arithmetic and geometric means. 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 6.4, 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 $k \gg \lg \kappa_V(H)$ and r is a promising Ritz value, the potential is *almost monotone* under the shift $(z-r)^k$. This justifies the intuition from Section 6.1 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 (6.12) established below.

Lemma 6.12 (Almost-Monotonicity and Moment Comparison). *Let $\mathcal{R} = \{r_1, \ldots, r_k\}$ 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

$$\mathbb{E}\left[|Z_H - r|^{-2k}\right] \geq \mathbb{E}\left[|Z_H - r|^{-k}\right]^2 \geq \frac{1}{\kappa_V(H)^2(\alpha\theta\psi_k(H))^{2k}}. \tag{6.12}$$

Proof. Let $p(z) = \prod_{i=1}^k (z - r_i)$. The claim follows from the following chain of inequalities:

$$\begin{aligned}
\sqrt{\mathbb{E}\left[|Z_H - r|^{-2k}\right]} &\geq \mathbb{E}\left[|Z_H - r|^{-k}\right] && \text{Jensen, } x \mapsto x^2 \\
&\geq \frac{1}{\alpha^k}\mathbb{E}[|p(Z_H)|^{-1}] && r \text{ is } \alpha\text{-promising} \\
&\geq \frac{1}{\alpha^k}\frac{1}{\sqrt{\mathbb{E}[|p(Z_H)|^2]}} && \text{Jensen, } x \mapsto x^2 \\
&\geq \frac{1}{\alpha^k}\frac{1}{\|e_n^* p(H)\|\kappa_V(H)} && \text{Lemma 6.10} \\
&\geq \frac{1}{\alpha^k}\frac{1}{\theta^k\|e_n^*\chi_k(H)\|\kappa_V(H)} && \text{Definition 6.3 of } \theta\text{-optimal} \\
&= \frac{1}{\alpha^k}\frac{1}{\theta^k\psi_k(H)^k\kappa_V(H)} && \text{Lemma 6.2.}
\end{aligned}$$

This already shows (6.12). For the other claim, rearrange both extremes of the above inequality to get

$$\begin{aligned}
\alpha\theta\kappa_V(H)^{1/k}\psi_k(H) &\geq \mathbb{E}\left[|Z_H - r|^{-2k}\right]^{-\frac{1}{2k}} \\
&\geq \frac{\tau_{(z-r)^k}(H)}{\kappa_V(H)^{1/k}} && \text{Lemma 6.10} \\
&\geq \frac{\psi_k(\widehat{H})}{\kappa_V(H)^{1/k}} && \text{Lemma 6.9}
\end{aligned}$$

which concludes the proof. □

In Section 6.4, we will see that when the shift associated with a promising Ritz value does not reduce the potential, Lemma 6.12 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.

The Shifting Strategy

An important component of our shifting scheme, discussed in detail below, is a simple subroutine, "find," guaranteed to produce an α-promising Ritz value with $\alpha = \kappa_V(H)^{\frac{4 \lg k}{4}}$. Guarantees for this subroutine are stated in the lemma below and proved in Section 6.4.

Lemma 6.13 (Guarantees for find)**.** *The subroutine* find *specified in Section 6.4 produces a* $\kappa_V(H)^{4\frac{\lg k}{k}}$*-promising Ritz value, using at most* $7k \lg kn^2 + \lg k$ *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 6.4. The overall strategy is specified below.

```
                              $\mathsf{Sh}_{k,B}$
Input: Hessenberg $H$ and a set $\mathcal{R}$ of $\theta$-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. $r \longleftarrow \mathsf{find}(H, \mathcal{R})$
    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, $\mathcal{S} \longleftarrow \mathsf{exc}(H, r, B)$
    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 $\mathsf{Sh}_{k,B}$ 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 6.14 (Stagnation Implies Support)**.** *Let* $\gamma \in (0, 1)$ *and* $\theta \ge 1$*, and let* $\mathcal{R} = \{r_1, \dots, r_k\}$ *be a set of* θ*-approximate Ritz values of* H*. Suppose* $r \in \mathcal{R}$ *is* α*-promising and assume*

$$\psi_k\left(\mathsf{iqr}(H,(z-r)^k)\right) \geq \gamma\,\psi_k(H) > 0. \tag{6.13}$$

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 $t \in (0,1)$:

$$\mathbb{P}\left[|Z_H - r| \leq \theta\alpha\left(\frac{\kappa_V(H)}{t}\right)^{1/k}\psi_k(H)\right] \geq (1-t)^2\frac{\gamma^{2k}}{\alpha^{2k}\theta^{2k}\kappa_V(H)^4}. \tag{6.14}$$

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 6.9. Our assumption implies that that:

$$\begin{aligned}
\gamma\,\psi_k(H) &\leq \psi_k(\widehat{H}) && \text{hypothesis}\\
&\leq \tau_{(z-r)^k}(H) && \text{Lemma 6.9}\\
&= \|e_n^*(H-r)^{-k}\|^{-1/k} && \text{definition}\\
&\leq \left(\frac{\kappa_V(H)}{\mathbb{E}\left[|Z_H - r|^{-2k}\right]^{\frac{1}{2}}}\right)^{1/k} && \text{Lemma 6.10.}
\end{aligned}$$

Rearranging and using (6.12) from Lemma 6.12 we get

$$\frac{\kappa_V(H)^2}{(1-\gamma)^{2k}\psi_k(H)^{2k}} \geq \mathbb{E}\left[|Z_H - r|^{-2k}\right] \geq \mathbb{E}\left[|Z_H - r|^{-k}\right]^2 \geq \frac{1}{\alpha^{2k}\theta^{2k}\psi_k(H)^{2k}\kappa_V(H)^2}, \tag{6.15}$$

which upon further rearrangement yields the “reverse Jensen” type bound:

$$\frac{\mathbb{E}[|Z_H - r|^{-2k}]}{\mathbb{E}[|Z_H - r|^{-k}]^2} \leq \left(\frac{\alpha\theta}{\gamma}\right)^{2k}\kappa_V(H)^4. \tag{6.16}$$

We now have

$$\begin{aligned}
\mathbb{P}\left[|Z_H - r| \leq \frac{\alpha}{t^{1/k}}\theta\psi_k(H)\kappa_V^{1/k}\right] &= \mathbb{P}\left[|Z_H - r|^{-k} \geq t\frac{1}{\alpha^k\theta^k\psi_k(H)^k\kappa_V}\right]\\
&\geq \mathbb{P}\left[|Z_H - r|^{-k} \geq t\mathbb{E}[|Z_H - r|^{-k}]\right] && \text{by (6.15)}\\
&\geq (1-t)^2\frac{\mathbb{E}[|Z_H - r|^{-k}]^2}{\mathbb{E}[|Z_H - r|^{-2k}]} && \text{Paley-Zygmund}\\
&\geq (1-t)^2\frac{\gamma^{2k}}{\alpha^{2k}\theta^{2k}\kappa_V(H)^4} && \text{by (6.16),}
\end{aligned}$$

establishing (6.14), as desired. □

Note that the conclusion of Lemma 6.14 in fact follows from the weaker condition $\tau_{(z-r)^k}(H) \geq \gamma\psi_k(H)$; this fact will be used in Chapter 7. We will shortly use Lemma 6.14 to prove the following guarantee on exc.

Lemma 6.15 (Guarantees for exc)**.** *The subroutine* exc *specified in Section 6.4 produces a set $\mathcal{S}$ of exceptional shifts, one of which achieves potential reduction. If $\theta \leq 2$, $\gamma = 0.8$, and $\alpha = B^{4\lg k/k}$, then both the arithmetic operations required for* exc*, and the size of $\mathcal{S}$, are at most*

$$N_{\mathsf{net}}\left(0.002B^{-\frac{8\lg k+4}{k}}\right),$$

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 6.16 (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 (6.16) on the inverse moment $\mathbb{E}[|Z_H - r|^{-2k}]$. Then, for $k \gg \lg \kappa_V(H)$, 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). 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 decided to not pursue it any further.

We are now ready to prove Theorem 1.13.

Proof of Theorem 1.13. 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 6.15 then guarantees we achieve potential reduction in (3). Thus each iteration decreases the potential by a factor of at least $\gamma = 0.8$, and since $\psi_k(H_0) \leq \|H\|$ we need at most

$$\frac{\lg(1/\omega)}{\lg(1/\gamma)} \leq 4\lg(1/\omega)$$

iterations before $\psi_k(H_t) \leq \omega\|H_0\|$, which in particular implies ω-decoupling.

Arithmetic Complexity. Computing a full set $\mathcal{R}$ of θ-approximate Ritz values of H has a cost $T_{\mathsf{OptRitz}}(k, \theta, \omega)$. Then, using an efficient implicit QR algorithm (cf. Definition 6.8) each computation of $\mathsf{iqr}(H, (z - r_i)^k)$ has a cost of $7kn^2$. By Lemma 6.13, we can produce a promising Ritz value in at most $12k\lg kn^2 + \lg k$ arithmetic operations. Then, in the event that the promising shift fails to reduce the potential the algorithm calls exc, which takes $N_{\mathsf{net}}(0.002B^{-\frac{8\lg k+4}{k-1}})$ 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.

Using the normal case of Lemma 6.15, with $k = 4$, the coefficient on n^2 in the arithmetic operations is $7 \cdot 4 \cdot 2 + 7 \cdot 4 \cdot 50 \leq 1500$, as promised in Remark 6.4. □

Efficiently Finding a Promising Ritz Value

In this section we show how to efficiently find a promising Ritz value, in $O(n^2 k \lg k)$ 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 $i = 1, \dots, k$ with k calls to $\mathsf{iqr}(H, (z - r_i)^{k/2})$, choosing the maximizing index i, and appealing to Lemma 6.10. The content of Lemma 6.13 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 $\mathcal{R} = \{r_1, \dots, r_k\}$ of $\theta$-optimal Ritz values of $H$.
Output: A complex number $r \in \mathcal{R}$
Requires: $\psi_k(H) > 0$
Ensures: $r$ is $\alpha$-promising for $\alpha = \kappa_V(H)^{\frac{4 \lg k}{k}}$.

  1. For $j = 1, ..., \lg k$
       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) $\mathcal{R} \longleftarrow \mathcal{R}_b$, where $b$ maximizes $\|e_n^* p_{j,b}(H)^{-2^{j-1}}\|$
  2. Output $\mathcal{R} = \{r\}$
```

Proof of Lemma 6.13 (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

$$
\begin{aligned}
\max_b \|e_n^* p_{1,b}(H)^{-1}\| &\geq \frac{1}{\kappa_V(H)^2}\mathbb{E}\left[\frac{1}{2}\left(|p_{1,0}(Z_H)|^{-2} + |p_{1,1}(Z_H)|^{-2}\right)\right] && \text{Lemma 6.10} \\
&\geq \frac{1}{\kappa_V(H)^2}\mathbb{E}[|p(Z_H)|^{-1}] && \text{AM/GM.}
\end{aligned}
$$

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

$$\begin{aligned}
\max_b \|e_n^* p_{j+1,b}(H)^{-2^j}\|^2 &\geq \frac{1}{\kappa_V(H)^2}\mathbb{E}\left[\frac{1}{2}\left(|p_{j+1,0}(Z_H)|^{-2^{j+1}} + |p_{j+1,1}(Z_H)|^{-2^{j+1}}\right)\right] && \text{Lemma 6.10} \\
&\geq \frac{1}{\kappa_V(H)^2}\mathbb{E}\left[|p_{j+1,0}(Z_H)p_{j+1,1}(Z_H)|^{-2^j}\right] && \text{AM/GM} \\
&\geq \frac{1}{\kappa_V(H)^4}\|e_n^*(p_{j+1,0}(H)p_{j+1,1}(H))^{-2^{j-1}}\| && \text{Lemma 6.10} \\
&= \frac{1}{\kappa_V(H)^4}\max_b \|e_n^* p_{j,b}(H)^{-2^{j-1}}\|.
\end{aligned}$$

Paying a further $\kappa_V(H)^2$ on the final step to convert the norm into an expectation, we get

$$\mathbb{E}\left[|Z_H - r|^{-k}\right] \geq \frac{1}{\kappa_V(H)^{4\lg k}}\mathbb{E}\left[|p(Z_H)|^{-1}\right]$$

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 $2\lg k$ 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 $\lg k$ steps. □

Remark 6.17 (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.

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 6.14 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 $R \triangleq 2^{1/k}\theta\psi_k(H)$.

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

$$\tau_{(z-s)^k}(H)^{-2k} = \|e_n^*(H-s)^{-k}\|^2 = \mathbb{E}[|Z_H - s|^{-2k}] \geq \frac{\mathbb{P}[|Z_H - s| \leq \psi_k(H)]}{|\mathcal{S}|(R\epsilon)^{2k}} \approx \frac{1}{4}\left(\frac{\gamma}{\theta}\right)^{2k}\frac{1}{R^{2k}\epsilon^{2k-2}},$$

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 $\kappa_V(H)^{4 \lg k/k}$-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 $\theta$-approximate Ritz value $r$, a condition number bound $B$, promising parameter $\alpha$
Output: A set $\mathcal{S} \subset \mathbb{C}$
Requires: $\kappa_V(H) \leq B$, $r$ is $\alpha$-promising, and $\psi_k(\mathsf{iqr}(H, (z-r)^k) \geq \gamma \psi_k(H)$
Ensures: For some $s \in \mathcal{S}$, $\psi_k(\mathsf{iqr}(H, (z-s)^k) \leq \gamma \psi_k(H)$

    1. $R \longleftarrow 2^{1/k} \theta \alpha B^{1/k} \psi_k(H)$
    2. $\epsilon \longleftarrow \left( \frac{\gamma^2}{(12B^4)^{1/k} \alpha^2 \theta^2} \right)^{\frac{k}{k-1}}$
    3. $\mathcal{S} \longleftarrow \epsilon R$-net of $R\psi_k(H)$.
```

Proof of Lemma 6.15: Guarantees for exc. Instantiating $t = 1/2$ in equation (6.14), we find that for the setting of R in line (1) of exc,

$$\mathbb{P}\left[|Z_H - r| \leq \mathbb{D}(r, R)\right] \geq \frac{1}{4B^4} \left(\frac{\gamma}{\alpha \theta} \right)^{2k}.$$

Let $\mathcal{S}$ be an ϵR-net of $\mathbb{D}(r, R)$; it is routine that such a net has at most $(1 + 2/\epsilon)^2 \leq 9/\epsilon^2$ points. By Lemma 6.9, to show that some $s \in \mathcal{S}$ achieves potential reduction, it suffices to find one for which

$$\|e_n^*(H - s)^{-k}\|^2 \geq \frac{1}{\gamma^{2k} \psi_k(H)^{2k}}.$$

We thus compute

$$\begin{aligned}
\max_{s\in\mathcal{S}} \|e_n^*(H-s)^{-k}\|^2 &\geq \frac{1}{\kappa_V(H)^2|\mathcal{S}|}\sum_{s\in\mathcal{S}} \mathbb{E}\left[|Z_H - s|^{-2k}\right] \\
&\geq \frac{\epsilon^2}{9B^2}\mathbb{E}\left[\sum_{s\in\mathcal{S}} |Z_H - s|^{-2k}\cdot \mathbf{1}\{Z_H \in \mathbb{D}(r,R)\}\right] && \text{Fubini and } \kappa_V(H) \leq B \\
&\geq \frac{\epsilon^2}{9B^2}\mathbb{E}\left[\max_{s\in\mathcal{S}} |Z_H - s|^{-2k}\cdot \mathbf{1}\{Z_H \in \mathbb{D}(r,R)\}\right] \\
&\geq \frac{\epsilon^2}{9B^2}\mathbb{E}\left[\frac{\mathbf{1}\{Z_H \in \mathbb{D}(r,R)\}}{(\epsilon R)^{2k}}\right] && \mathcal{S} \text{ is an } \epsilon R\text{-net} \\
&\geq \frac{\mathbb{P}[Z_H \in \mathbb{D}(r,R)]}{9B^2R^{2k}\epsilon^{2k-2}} \\
&\geq \frac{1}{\gamma^{2k}\psi(H)^{2k}}
\end{aligned}$$

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 $\mathbb{D}(r,R)$, 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 $\alpha = B^{4\lg k/k}$, and using $k \geq 2$, we obtain the expression appearing in $N_{\mathsf{net}}(\cdot)$ in the statement of Lemma 6.15.

A practical choice of net, which we will return to in Chapter 7's forthcoming finite arithmetic analysis, is an equilateral triangular lattice with spacing $\sqrt{3}\epsilon$, intersected with the $\mathbb{D}(r,(1+\epsilon)R)$. Such a construction is optimal as $\epsilon \longrightarrow 0$, 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 [4, 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

$$|\mathcal{S}| \leq N_{\mathsf{net}}\left(\left(\frac{0.8^2}{12^{1/4}}\right)^{4/3}\right) \leq 49.9.$$

□

Bibliographic Note

This chapter is lightly adapted from its original presentation in [17], and includes some of the preliminary material on QR decomposition and iteration from the forthcoming [18].

Chapter 7

The Shifted QR Algorithm in Finite Arithmetic

7.1 Introduction

In Chapter 6 we gave a family of shifting strategies, $\mathsf{Sh}_{k,B}$, for which the Hessenberg shifted QR algorithm converges globally and rapidly on nonsymmetric matrices whose eigenvector condition number is bounded, *in exact arithmetic*. Our final task in this thesis is to that both the correctness and rapid convergence of these strategies continue to hold in floating point arithmetic with an appropriate implementation, and prove a bound on the number of bits of precision needed, for matrices with controlled eigenvector condition number and minimum eigenvalue gap.

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 $\mathsf{Sh}_{k,B}$ is a special case. We specifically address the following two issues.

Issue 7.1 (Forward Stability of QR Steps)**.** Consider a degree k shifted QR step:

$$p(H) = QR \qquad \widehat{H} = Q^{*}HQ,$$

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 $\widehat{H}$ which is the unitary conjugation of a matrix near H [151]. 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 shifted QR algorithm with any shifting strategy in finite arithmetic, even on symmetric matrices.*

Issue 7.2 (Computation of Ritz Values)**.** All 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. In practice, and in the current version of LAPACK, the prescription is just to run the shifted QR algorithm itself on $H_{(k)}$, but there are no proven guarantees for this approach.

The two obstacles above are closely related. A natural strategy for obtaining forward stability is to perturb the zeros $r_1, \dots, r_k$ of the shift polynomial $p(z)$ so that they avoid the eigenvalues of H. Such a perturbation 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" thus determines how constrained we are in choosing our shifts while maintaining good convergence properties.

This chapter contains the following three contributions, which together provide a solution to Issues 7.1-7.2 for a wide class of shifting strategies, on matrices with nonzero eigenvalue gap (and thus finite eigenvector condition number).

(i) Forward Stability by Regularization. We handle Issue 7.1 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 w_k are independent random numbers of an appropriate size (which depends on $\kappa_V(H)$ and $\mathrm{gap}(H)$). We refer to this technique as *shift regularization* and show in Section 7.4 (Lemma 7.14) that it yields forward stability of an implicit QR step with high probability, for any Hessenberg matrix H with an upper bound on $\kappa_V(H)$ and a lower bound on $\mathrm{gap}(H)$, and any shifts $r_1, \dots, r_k$.

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 Sections 7.4 and 7.4 and may be of independent interest.

(ii) Optimal Ritz Value/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 Definition 6.3 that $\{r_1, \dots, r_k\} \subset \mathbb{C}$ is called a set of *θ-optimal Ritz values* of a Hessenberg matrix H if:

$$\|e_n^*(H - r_1)\dots(H - r_k)\|^{1/k} \leq \theta \min_p \|e_n^* p(H)\|^{1/k}, \tag{7.1}$$

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. Our key result (Theorem 7.15) 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 under small perturbations $r_i \longrightarrow r_i + w_i$ 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.

(iii) Approximating the Eigenvalues of Small Matrices. In order to carry out (ii), we require an efficient way to obtain forward approximations to the eigenvalues of the small $k \times k$ matrix $H_{(k)}$. Since k is very small, it is acceptable to use an algorithm with worse than $O(k^3)$ complexity. On the other hand, our shifting strategy breaks down on matrices of size $k \times k$ or smaller, so we also need an algorithm to compute approximations to the eigenvalues of small matrices, to use once we have deflated to a sufficiently small matrix. We will assume access to a black box algorithm $\mathsf{SmallEig}$ for use in these two situations, with the following guarantee on a matrix M of dimension k or smaller.

Definition 7.3. A *small eigenvalue solver* $\mathsf{SmallEig}(M, \beta, \phi)$ takes as input a matrix M of size at most $k \times k$, and with probability at least $1 - \phi$, outputs $\widetilde{\lambda}_1, ..., \widetilde{\lambda}_k \in \mathbb{C}$ such that $|\widetilde{\lambda}_i - \lambda_i| \leq \beta$ for each of $\lambda_1, ..., \lambda_k \in \operatorname{Spec} M$.

Remark 7.4. The notion of forward error here is absolute, instead of relative — this will simplify some of the analysis later on.

Remark 7.5. Using a forward error algorithm for the base case calls is overkill, since of course we only require backward error in that case, but we will see that the computational cost of these base case calls is dominated by that of the forward errors along the way, and the choice will thus not impact the asymptotics of the runtime. For concreteness, $\mathsf{SmallEig}$ may be instantiated with EIG from Chapter 5, using Theorem 2.6 to verify that backward error $\Omega(\beta^k)$ ensures forward error β in the sense above.

Finally, we combine (i-iii) above in Section 7.7 to prove Theorem 1.14, which we restate here precisely.

Restatement of Theorem 1.14. *Let $H \in \mathbb{H}^{n \times n}_{B/2}$, and assume further that $\|H\| \leq 2\Sigma$ and $\Gamma \leq \operatorname{gap}(H)/2$. For a certain $k = O(\lg B \lg\lg B)$, the shifting strategy $\mathsf{Sh}_{k,B}$ can be implemented in finite arithmetic* *to give a randomized shifted QR algorithm,* ShiftedQR*, with the following guarantee: for any* $\delta > 0$ ShiftedQR(H, δ, ϕ) *produces the eigenvalues of a matrix* H' *with* $\|H - H'\| \leq \delta$*, with probability at least* $1 - \phi$*, using*

(i) $O\left((\lg \frac{nB\Sigma}{\delta\Gamma} \cdot k \lg k + k^2)n^3\right)$ *arithmetic operations on a floating point machine with* $O(k \lg \frac{nB\Sigma}{\delta\Gamma\phi})$ *bits of precision, and*

(ii) $O(n \lg \frac{nB\Sigma}{\delta\Gamma})$ *calls to* SmallEig *with accuracy* $\Omega(\frac{\delta^2\Gamma^2}{n^4B^4\Sigma})$ *and failure probability tolerance* $\Omega(\frac{\phi}{n^3 \lg \frac{nB\Sigma}{\delta\Gamma}})$.

Note that the bounds B, Γ, and Σ must be known to the algorithm in advance. The above restatement differs from the statement in the introduction only in that the former set $\Sigma = 1/2$, hiding the resulting factor in the asymptotic notation.

Let us now collect the many algorithm inputs, constants, parameters, and subroutines that the reader will encounter in this chapter, along with their 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 global constant is the 'non-normality' or condition number bound B, from which we define the shift degree k to be the smallest power of 2 for which

$$B^{\frac{8 \lg k + 3}{k-1}} \cdot (2B^4)^{\frac{2}{k-1}} \leq 3, \tag{7.2}$$

which makes $k = O(\lg B \lg \lg B)$. We further define the auxiliary constants

$$\alpha \triangleq (1.01B)^{4 \log k/k} \in [1, 2], \qquad \theta \triangleq \frac{1.01}{0.998^{1/k}}(2B^4)^{1/2k} \in [1, 2] \qquad \gamma \triangleq 0.8, \tag{7.3}$$

which depend only on B.

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	0.8

Table 7.1: Global Data for ShiftedQR

Table 7.2 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

$$\kappa_V(H') \le 2\kappa_V(H) \le B \qquad \|H'\| \le 2\|H\| \le \Sigma \qquad \mathrm{gap}(H') \ge \mathrm{gap}(H)/2 \ge \Gamma.$$

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 7.1 contains the main subroutines (note that we will write $\mathsf{SH}_{k,B}$ for the finite aritmetic implementation of $\mathsf{Sh}_{k,B}$).

Input Parameter	Meaning	Typical Setting
H	Upper Hessenberg Matrix	
δ	Accuracy	
ϕ	Failure Probability Tolerance	
Internal Parameter		
ω	Working Accuracy	$\Omega\left(\min\{\delta/n, \Gamma/n^2B^2\}\right)$
φ	Working Failure Probability Tolerance	$\Omega\left(\frac{\phi}{\log(\omega/\Sigma)}\right)$
η_1, η_2	Regularization Parameters	$\Omega(\omega^2), \Omega(\omega^2\phi^{-1/2}\Sigma^{-1})$
β	Forward Accuracy for Ritz Values	$\Omega(\omega^2\Sigma^{-1})$
$\mathcal{R}$	Approximate Ritz Values	
$\mathcal{S}$	Exceptional Shifts	

Table 7.2: Input and Internal Parameters for ShiftedQR

7.2 Related Work

Forward Stability of Shifted QR. An important step towards understanding and addressing the two issues mentioned above was taken by Parlett and Le [130], 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.

Subroutine	Action	Output	Input	Global Data
IQR	Implicit QR Step	$\widetilde{\widehat{H}}, \tilde{R}$	$H, p(z)$	
TAU^m	Approximate $\tau^m_{p(z)}(H) = \|e_n^* p(H)^{-1}\|$	$\widetilde{\tau^m}$	$H, p(z)$	
OPTIMAL	Check Ritz Value Optimality	$\texttt{opt}$	$H, \mathcal{R}$	θ
RITZ-OR-DEC	Compute θ-Optimal Ritz Values	$\widehat{H}, \mathcal{R}, \texttt{dec}$	H, ω, ϕ	Σ, Γ, θ
FIND	Find a α-Promising Ritz Value	r	$H, \mathcal{R}$	α
EXC	Compute Exceptional Shifts	$\mathcal{S}$	H, r, ω, ϕ	$B, \Sigma, \theta, \alpha$
SH$_{k,B}$	Shifting Strategy to Reduce $\psi_k(H)$	$\widehat{H}$	$H, \mathcal{R}, \omega, \phi$	$B, \Sigma, \gamma, \theta, \alpha$
DEFLATE	Deflate a Decoupled Matrix	$H_1, H_2, \ldots$	H, ω	

Table 7.3: Subroutines of ShiftedQR

Watkins [159] 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 in this paper.

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 [37, 38] introduced an additional criterion called aggressive early deflation which yields significant improvements in practice. Kressner [104] 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 [144].

The early decoupling procedure introduced in this paper 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 this chapter use shift polynomials $p(z) = (z - r_1) \ldots (z - r_k)$ of degree k where k is roughly proportional to $\lg \kappa_V(H)$. It was initially proposed [9] 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 [160] and further by Kressner [103] 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 paper 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 δ–decoupling increases as a function of k, roughly as $O(k \lg(1/\delta))$ 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.

7.3 Preliminaries

During this chapter, we will find it useful to define $\mathsf{ShiftedQR}$ in terms of an *absolute* notion of decoupling, instead of the relative one used in Chapter 6. The reason is that our algorithm does not have access to the norm of any matrices it encounters, only to the upper bound Σ.

Definition 7.6. We will say that an upper Hessenberg matrix H is *absolutely ω-decoupled* if some subdiagonal entry of H is smaller than ω (as opposed to $\omega\|H\|$).

We will require some finite arithmetic assumptions and results beyond the floating point axioms from Chapter 2. First, as mentioned in Chapter 6, 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 [94, 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. These can be computed in the following sense, for instance by running Newton's method with a starting point found by bisection.

Lemma 7.7 (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 $a \in \boldsymbol{R}^+$, there exists an algorithm that computes $a^{\frac{1}{k}}$ with relative error ϵ in at most*

$$T_{\mathsf{root}}(k, \epsilon) \triangleq C_{\mathsf{root}} k \lg(k \lg(1/\epsilon))$$

arithmetic operations.

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 7.8 (Efficient Perturbation Algorithm)**.** An efficient random perturbation algorithm takes as input $r \in \mathbb{C}$ an $R > 0$, and generates a random $w \in \mathbb{C}$ distributed uniformly in the disk $\mathbb{D}(r, R)$ using C_{D} arithmetic operations.

7.4 Implicit QR: Implementation, Forward Stability, and Regularization

In this section we present a standard implementation (called "IQR") of a degree 1 (i.e., single shift) implicit QR step using Givens rotations and provide an analysis of its backward stability which is slightly stronger than the guarantees of [151] for an implementation with Householder reflectors. 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. Finally, we prove bounds on the forward error of a degree k IQR step in terms of the eigenvector condition number of the matrix and the distance of the roots of the shifting polynomial to its spectrum. The former is controlled by assumption, and the latter can be tamed by a random perturbation of the shifts; we check this in Section 7.4.

Degree-1 IQR

We will use the following notion of backward stability for a single implicit QR step; the distinction with [151] is the additional second equation below.

Definition 7.9 (Backward-Stable Degree-1 Implicit QR Algorithm)**.** A $\nu_{\mathsf{IQR}}(n)$-stable single-shift implicit QR algorithm takes as inputs a Hessenberg matrix $H \in \boldsymbol{C}^{n \times n}$ and a shift $s \in \boldsymbol{C}$ and outputs a Hessenberg matrix $\widetilde{\widehat{H}}$ and an exactly triangular matrix $\tilde{R}$, for which there exists a unitary $\widetilde{Q}$ satisfying

$$\left\|\widetilde{\widehat{H}} - \widetilde{Q}^* H \widetilde{Q}\right\| \leq \|H - s\| \nu_{\mathsf{IQR}}(n) \mathbf{u} \tag{7.4}$$

$$\left\|H - s - \widetilde{Q}\tilde{R}\right\| \leq \|H - s\| \nu_{\mathsf{IQR}}(n) \mathbf{u} \tag{7.5}$$

We now verify that there is a suitable backward-stable implicit QR algorithm. The pseucodode of our IQR is given below.

```
IQR
Input: Hessenberg $H$, shift $s \in \mathbb{C}$
Output: Hessenberg $\widetilde{\widetilde{H}}$ and triangular $\widetilde{R}$
Ensures: $\|\widetilde{\widetilde{H}}\| \leq \|H\| + 32n^{3/2}\mathbf{u} \cdot \|H - s\|$, and there exists unitary $\widetilde{Q}$ for which $\|\widetilde{\widetilde{H}} - \widetilde{Q}^* H\widetilde{(Q)}\| \leq 32n^{3/2}\mathbf{u} \cdot \|H - s\|$ and $\|H - s - \widetilde{Q}\widetilde{R}\| \leq 16n^{3/2}\mathbf{u} \cdot \|H - s\|$

  1. $\widetilde{R} \longleftarrow H - s$
  2. For $i = 1, 2, ..., n - 1$
       (a) $X_{1:2,i} \longleftarrow \widetilde{R}_{i:i+1,i}$
       (b) $\widetilde{R}_{i:i+1,i+1,n} \longleftarrow \mathsf{giv}(X_{1:2,i})^* \widetilde{R}_{i:i+1,i+1:n} + E_{2,i,b}$
       (c) $\widetilde{R}_{i:i+1,i} \longleftarrow \begin{pmatrix} \|X_{1:2,i}\| + E_{2,i,c} \\ 0 \end{pmatrix}$
  3. $\widetilde{\widetilde{H}} \longleftarrow \widetilde{R}$
  4. For $i = 1, 2, ...n - 1$
       (a) $\widetilde{\widetilde{H}}_{1:n,i:i+1} \longleftarrow \widetilde{\widetilde{H}}_{1:n,i:i+1}\mathsf{giv}(X_{1:2,i}) + E_{4,i}$
  5. $\widetilde{\widetilde{H}} \longleftarrow \widetilde{\widetilde{H}} + s$
```

Lemma 7.10 (Backward Stability of Degree 1 IQR). *Assuming*

$$\mathbf{u} \leq \min\left\{\frac{1}{24}, \frac{\lg 2}{8n^{5/2}}\right\} = 2^{-O(\lg n)}, \tag{7.6}$$

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* $\nu_{\mathsf{IQR}}(n) = 32n^{3/2}$.

Proof. For the purpose of the analysis, let us define $\widetilde{H}_0 \triangleq H - s$ and for each $i = 1, ..., n - 1$, denote by $\widetilde{H}_i$ the matrix $\widetilde{R}$ as it stands at the end of line 2c 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 $\widetilde{Q} \triangleq \widetilde{Q}_{n-1}$ 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

$$\left(\begin{matrix} E_{2,i,c} \\ 0 \end{matrix} \middle| \ E_{2,i,b} \right)$$

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

$$\begin{aligned}
\|\widetilde{H}_i\| &\leq \|\widetilde{H}_{i-1}\| + \|E_{2,i}\| \\
&\leq \|\widetilde{H}_{i-1}\| \left(1 + 8\sqrt{n}\mathbf{u}\right) \\
&\leq \|\widetilde{H}_0\| \left(1 + 8\sqrt{n}\mathbf{u}\right)^n \\
&\leq \|\widetilde{H}_0\| \exp\left(8n^{3/2}\mathbf{u}\right) \\
&\leq 2\|H - s\| \qquad i = 1, \ldots, n-1.
\end{aligned}$$

Since $\widetilde{Q}$ and every G_i is unitary, this gives

$$\|H - s - \widetilde{Q}\tilde{R}\| = \|\widetilde{Q}^*\widetilde{H}_0 - \tilde{R}\| \leq \sum_{i \in [n-1]} \|E_{2,i}\| \leq 16n^{3/2}\mathbf{u} \cdot \|H - s\|.$$

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 $\widetilde{\widetilde{H}}$ output by $\mathsf{IQR}(H, s)$ satisfies

$$\begin{aligned}
\widetilde{\widetilde{H}} - s &= \tilde{R}\widetilde{Q} + E_{4,n-1}(G_1 \cdots G_{n-2}) + \cdots + E_{4,2}G_1 + E_{4,1} \\
&= \widetilde{Q}^*(H - s)\widetilde{Q} + E_{4,n-1}(G_1 \cdots G_{n-2}) + \cdots + E_{4,2}G_1 + E_{4,1} \\
&\qquad + (G_{n-2}^* \cdots G_1^*)E_{2,1}\widetilde{Q} + (G_{n-3}^* \cdots G_1^*)E_{2,2}\widetilde{Q} + \cdots + G_1^*E_{2,n-1}\widetilde{Q},
\end{aligned}$$

meaning

$$\|\widetilde{\widetilde{H}} - \widetilde{Q}^*H\widetilde{Q}\| \leq 32n^{3/2}\mathbf{u} \cdot \|H - s\|$$

and

$$\|\widetilde{\widetilde{H}}\| \leq \|H\| + 32n^{3/2}\|H - s\|\mathbf{u},$$

as desired.

In terms of arithmetic operations, it costs n to compute $\tilde{R}$ from H in line 1. In line 2b, computing $\|X_{1:2,i}\|$ costs 4, computing $\mathsf{giv}(X_{1:2,i})$ given this norm costs another 2, zeroing out $\tilde{R}_{i+1,i}$ costs 1, replacing $\tilde{R}_{i,i}$ with $\|X_{1:2,i}\|$ costs 1, and applying the rotation to $\tilde{R}_{i:i+1,i+1:n}$ 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

$$n + 6(n-1) + 2(n-1) + 2n(n-1) + 2n(n+1) - 4 + n = 4n^2 + 12n - 12 \leq 7n^2 \qquad n \geq 2.$$

□

Backward Stability of Higher-Degree IQR

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 7.9, and shifts $s_1, \dots, s_k$, we will define

$$\mathsf{IQR}(H, \{s_1, \dots, s_k\}) \triangleq \mathsf{IQR}(\mathsf{IQR}(\cdots \mathsf{IQR}(\mathsf{IQR}(H, s_1), s_2), \cdots), s_k), \tag{7.7}$$

which can be executed in $T_{\mathsf{IQR}}(n, k) = 7kn^2$ arithmetic operations. We will sometimes use the notation

$$\mathsf{IQR}(H, p(z)) = \mathsf{IQR}(H, \{s_1, \dots, s_k\})$$

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 7.10 is readily adapted to give backward stability guarantees for $\mathsf{IQR}(H, p(z))$.

Lemma 7.11 (Backward Error Guarantees for Higher Degree IQR). *Fix $C > 0$ and let $p(z) = \prod_{\ell \in [k]}(z - s_\ell)$, where $\mathcal{S} = \{s_1, ..., s_k\} \subset \mathbb{D}(0, C\|H\|)$. Write $\left[\widetilde{\widehat{H}}, \widetilde{R}_1, ..., \widetilde{R}_k\right] = \mathsf{IQR}(H, p(z))$, and let $\widetilde{Q}_\ell$ be the unitary guaranteed by Definition 7.9 to the ℓth internal call to IQR. Assuming*

$$\nu_{\mathsf{IQR}}(n)\mathbf{u} \le 1/4,$$

the outputs $\widetilde{R} = \widetilde{R}_k \cdots \widetilde{R}_1$ and $\widetilde{Q} = \widetilde{Q}_1 \cdots \widetilde{Q}_k$ satisfy

$$\left\|\widetilde{\widehat{H}} - \widetilde{Q}^* H \widetilde{Q}\right\| \le 1.4k(1 + C)\|H\| \nu_{\mathsf{IQR}}(n)\mathbf{u} \tag{7.8}$$

$$\left\|p(H) - \widetilde{Q}\widetilde{R}\right\| \le 4\Big(2(1 + C)\|H\|\Big)^k \nu_{\mathsf{IQR}}(n)\mathbf{u}. \tag{7.9}$$

Proof of Lemma 7.11. 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 7.9. We have

$$\|\widetilde{H}_2 - \widetilde{Q}_1^* \widetilde{H}_1 \widetilde{Q}_1\| \le \|\widetilde{H}_1 - s_1\| \nu_{\mathsf{IQR}}(n)\mathbf{u} \le (1 + C)\|H\| \nu_{\mathsf{IQR}}(n)\mathbf{u},$$

and inductively, assuming that

$$\|\widetilde{H}_\ell - \widetilde{Q}_{\ell-1}^* \widetilde{H}_{\ell-1} \widetilde{Q}_{\ell-1}\| \le (1 + C)\|H\|(\nu_{\mathsf{IQR}}(n)\mathbf{u} + \cdots + (\nu_{\mathsf{IQR}}(n)\mathbf{u})^\ell),$$

we have

$$\begin{aligned} \|\widetilde{H}_{\ell+1} - \widetilde{Q}_\ell^* \widetilde{H}_\ell \widetilde{Q}_\ell\| &\le \|\widetilde{H}_\ell - s_\ell\| \nu_{\mathsf{IQR}}(n)\mathbf{u} \\ &\le \|H\|(1 + (1 + C)(\nu_{\mathsf{IQR}}(n)\mathbf{u} + \cdots + (\nu_{\mathsf{IQR}}(n)\mathbf{u})^\ell) + C)\nu_{\mathsf{IQR}}(n)\mathbf{u} \\ &\le (1 + C)\|H\|(\nu_{\mathsf{IQR}}(n)\mathbf{u} + \cdots + (\nu_{\mathsf{IQR}}(n)\mathbf{u})^{\ell+1}). \end{aligned}$$

This gives the first asserted bound, since

$$\|\widetilde{H} - \widetilde{Q}^*\widetilde{H}\widetilde{Q}\| \le \sum_{\ell\in[m-1]} \|\widetilde{H}_{\ell+1} - \widetilde{Q}_\ell^*\widetilde{H}_\ell\widetilde{Q}_\ell\| \le (1+C)\|H\|\frac{m\nu_{\mathsf{IQR}}(n)\mathbf{u}}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}$$

and $\frac{1}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}} \le 4/3 \le 1.4$.

For the second assertion, we will mirror the proof of Lemma 6.6, using backward stability guarantees on a single IQR step from Definition 7.9. In particular, in view of the definition and the above bound, we can write

$$\begin{aligned} \widetilde{H}_\ell - s_\ell &= \widetilde{Q}_\ell\tilde{R}_\ell + E_\ell & \qquad \|E_\ell\| &\le (1+C)\|H\|\frac{\nu_{\mathsf{IQR}}(n)\mathbf{u}}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}} \\ \widetilde{H}_1\widetilde{Q}_\ell\cdots\widetilde{Q}_1 &= \widetilde{Q}_\ell\cdots\widetilde{Q}_1\widetilde{H}_{\ell+1} + \|\Delta_{\ell+1}\| & \qquad \Delta_{\ell+1} &\le (1+C)\|H\|\frac{\nu_{\mathsf{IQR}}(n)\mathbf{u}}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}} \end{aligned}$$

so that

$$\begin{aligned} p(H) = p(\widetilde{H}_1) &= (\widetilde{H}_1 - s_m)\cdots(\widetilde{H}_1 - s_1) \\ &= (\widetilde{H}_1 - s_m)\cdots(\widetilde{Q}_1\tilde{R}_1 + \widetilde{Q}_1^*E_1) \\ &= (\widetilde{H}_1 - s_m)\cdots(\widetilde{H}_1 - s_2)\widetilde{Q}_1(\tilde{R}_1 + \widetilde{Q}_1^*E_1) \\ &= (\widetilde{H}_1 - s_m)\cdots\widetilde{Q}_1(\widetilde{H}_2 - s_2 + \Delta_2)(\tilde{R}_1 + \widetilde{Q}_1^*E_1) \\ &= (\widetilde{H}_1 - s_m)\cdots(\widetilde{H}_1 - s_3)\widetilde{Q}_1\widetilde{Q}_2(\tilde{R}_2 + \widetilde{Q}_2^*E_2 + \widetilde{Q}_2^*\Delta_2)(\tilde{R}_1 + \widetilde{Q}_1^*E_1) \\ &= \widetilde{Q}_1\cdots\widetilde{Q}_m(\tilde{R}_m + \widetilde{Q}_m^*E_m + \widetilde{Q}_m^*\Delta_m)\cdots(\tilde{R}_2 + \widetilde{Q}_2^*E_2 + \widetilde{Q}_2^*\Delta_2)(\tilde{R}_1 + \widetilde{Q}_1^*E_1) \end{aligned}$$

Thus, using the bounds on E_ℓ and Δ_ℓ, and the fact that $\|\tilde{R}_\ell\| = \|\widetilde{H}_\ell - s_\ell\| \le \frac{(1+C)\|H\|}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}$,

$$\begin{aligned} \|p(H) - \widetilde{Q}_1\cdots\widetilde{Q}_m\tilde{R}_m\cdots\tilde{R}_1\| &= \|\tilde{R}_m\cdots\tilde{R}_1 - (\tilde{R}_m + \widetilde{Q}_m^*E_m + \widetilde{Q}_m^*\Delta_m)\cdots(\tilde{R}_2 + \widetilde{Q}_2^*E_2 + \widetilde{Q}_2^*\Delta_2)(\tilde{R}_1 + \widetilde{Q}_1^*E_1)\| \\ &\le \prod_{\ell\in[m]}\left(\|\tilde{R}_\ell\| + \tfrac{2(1+C)\|H\|}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}\right) - \prod_{\ell\in[m]}\|\tilde{R}_\ell\| \\ &\le \left(\tfrac{(1+C)\|H\|}{1-\nu_{\mathsf{IQR}}(n)\mathbf{u}}\right)^m((1+2\nu_{\mathsf{IQR}}(n)\mathbf{u})^m - 1) \\ &\le 4\big(2(1+C)\|H\|\big)^m\nu_{\mathsf{IQR}}(n)\mathbf{u}; \end{aligned}$$

in the final line we are using again that $\nu_{\mathsf{IQR}}(n)\mathbf{u} \le 1/4$ and thus that $((1+2\nu_{\mathsf{IQR}}(n)\mathbf{u})^m - 1) \le (3/2)^m\nu_{\mathsf{IQR}}(n)\mathbf{u}/4$, whereas $(1-\nu_{\mathsf{IQR}}(n)\mathbf{u})^{-m} \le (4/3)^m$. □

Forward Stablity 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. We pause to restate a lemma from the preliminaries on the forward stability of the QR decomposition itself:

Restatement of Lemma 2.12. *Let $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ with A invertible and $\|A - \widetilde{A}\|\|A^{-1}\| \le 1/2$. Then If $[Q, R] = \mathsf{qr}(A)$ and $[\widetilde{Q}, \tilde{R}] = \mathsf{qr}(\widetilde{A})$, then*

$$\|\widetilde{Q} - Q\|_F \le 4\|A^{-1}\|\|A - \widetilde{A}\|_F \quad \textit{and} \quad \|\tilde{R} - R\| \le 3\|A^{-1}\|\|R\|\|A - \widetilde{A}\|.$$

The main result of this subsection, which will be used throughout the paper, is the following.

Lemma 7.12 (Forward Error Guarantees for IQR)**.** *Under the hypotheses of Lemma 7.11, and assuming further that $[Q, R] = \mathsf{qr}(p(H))$, $\widehat{H} = Q^*HQ$, and*

$$\begin{aligned} \mathbf{u} &\le \mathbf{u}_{\mathsf{IQR}}\left(n, k, \|H\|, \kappa_V(H), \mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)\right) \\ &\triangleq \frac{1}{8\kappa_V(H)\nu_{\mathsf{IQR}}(n)}\left(\frac{\mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)}{\|H\|}\right)^k \\ &= 2^{-O\left(\lg n\kappa_V(H) + k\lg\frac{\|H\|}{\mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)}, \end{aligned} \tag{7.10}$$

we have the forward error guarantees:

$$\|\widetilde{Q} - Q\|_F \le 16\kappa_V(H)\left(\frac{(2+2C)\|H\|}{\mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)^k n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u} \tag{7.11}$$

$$\|\tilde{R} - R\| \le 12\kappa_V(H)\left(\frac{(2+2C)^2\|H\|^2}{\mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)^k \nu_{\mathsf{IQR}}(n)\mathbf{u} \tag{7.12}$$

$$\|\widetilde{\widehat{H}} - \widehat{H}\|_F \le 32\kappa_V(H)\|H\|\left(\frac{(2+2C)\|H\|}{\mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)^k n^{1/2}\nu_{\mathsf{IQR}}(n)\mathbf{u}. \tag{7.13}$$

Proof. The first two assertions are immediate from applying Lemma 2.12 to $M = p(H)$, computing

$$\|M^{-1}\| = \|p(H)^{-1}\| \le \frac{\kappa_V(H)}{\mathrm{dist}(\mathcal{S}, \operatorname{Spec} H)^k},$$

bounding $\|p(H)\| \le (2+2C)^k\|H\|^k$, and using Lemma 7.11 to control $\|E\| \le 2(2+2C)^k\|H\|^k\nu_{\mathsf{IQR}}(n)\mathbf{u}$. For the third, observe that

$$\|\widetilde{Q}^*H\widetilde{Q} - QHQ\|_F \le \|\widetilde{Q}^*H(\widetilde{Q} - Q)\|_F + \|(\widetilde{Q}^* - Q^*)HQ\|_F \le 2\|H\|\|\widetilde{Q} - Q\|_F,$$

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.

```
                         $\mathsf{TAU}^k$
Input: Hessenberg $H \in C^{n\times n}$, 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. $[\widetilde{\hat{H}}, \tilde{R}_1, \dots, \tilde{R}_k] \longleftarrow \mathsf{IQR}(H, p(z))$
  2. $\widetilde{\tau^k} \longleftarrow \mathsf{fl}\left((\tilde{R}_1)_{nn} \cdots (\tilde{R}_k)_{nn}\right)$
```

Lemma 7.13 (Guarantees for TAU^k)**.** *If $\mathcal{S} = \{s_1, ..., s_k\} \in \mathbb{D}(0, C\|H\|)$ and*

$$\begin{aligned}
\mathbf{u} &\leq \mathbf{u}_{\mathsf{TAU}}(n, k, C, \|H\|, \kappa_V(H), \operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)) && (7.14)\\
&\triangleq \frac{1}{6\cdot 10^3 \kappa_V(H) \nu_{\mathsf{IQR}}(n)} \left(\frac{\operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)}{(2+2C)\|H\|}\right)^{2k} && (7.15)\\
&= 2^{-O\left(\lg n\kappa_V(H) + k \lg \frac{\|H\|}{\operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)},
\end{aligned}$$

then TAU^k *satisfies its guarantees, and runs in*

$$T_{\mathsf{TAU}}(n, k) \triangleq T_{\mathsf{IQR}}(k, n) + k = O(kn^2)$$

arithmetic operations.

Proof. Let $[Q, R] = \mathsf{qr}(p(H))$ and recall that (6.8) shows that $\tau_p(H)^k = R_{nn}$. As (7.14) implies $\mathbf{u} \leq \mathbf{u}_{\mathsf{IQR}}(n, k, \|H\|, \kappa_V(H), \operatorname{dist}(\mathcal{S}, \operatorname{Spec} H))$, we can apply Lemma 7.12: the matrix $\tilde{R} = \tilde{R}_k \cdots \tilde{R}_1$ satisfies

$$\begin{aligned}
|\tilde{R}_{n,n} - R_{n,n}| &\leq \|\tilde{R} - R\| \\
&\leq 12\kappa_V(H)\left(\frac{(2+2C)^2\|H\|^2}{\operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)^k \nu_{\mathsf{IQR}}(n)\mathbf{u} && \text{Lemma 7.12}\\
&\leq \frac{0.0005}{\|p(H)^{-1}\|} && \text{(7.14) and } \|p(H)^{-1}\| \leq \frac{\kappa_V(H)}{\operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)^k}\\
&\leq 0.0005\,\sigma_n(R) && p(H) = QR\\
&\leq 0.0005\,R_{n,n}. && \sigma_n(R) \leq \|e_n^* R\| = R_{n,n}
\end{aligned}$$

Now, because $\widetilde{\tau^k}$ is the result of computing the product of the $(\tilde{R}_i)_{n,n}$ in floating point arithmetic, we have $\left|\widetilde{\tau^k} - \tilde{R}_{n,n}\right| \leq k\mathbf{u}\tilde{R}_{n,n}$, whence

$$\begin{aligned}\left|\widetilde{\tau^k} - R_{n,n}\right| &\leq \left|\widetilde{\tau^k} - \tilde{R}_{n,n}\right| + \left|\tilde{R}_{n,n} - R_{n,n}\right| \\ &\leq k\mathbf{u}\tilde{R}_{n,n} + 0.0005\, R_{n,n} \\ &\leq (1.0005k\mathbf{u} + 0.0005)R_{n,n} \\ &\leq 0.001R_{n,n}.\end{aligned}$$

It will also be useful to observe that

$$\left|\frac{1}{\widetilde{\tau^k}} - \frac{1}{R_{n,n}}\right| \leq \frac{0.001}{|\widetilde{\tau^k}|} \leq \frac{0.001}{|\widetilde{\tau^k}|} \leq \frac{0.001}{\left|R_{n,n} - |\widetilde{\tau^k} - R_{n,n}|\right|} \leq \frac{0.001}{0.99R_{n,n}} \leq \frac{0.0011}{R_{n,n}}.$$

□

Shift Regularization

The forward error bounds on our shifts are controlled by the distance to Spec H; to ensure that this is not too large, we *regularize* the shifts $r_1, \ldots, r_k$ by randomly perturbing them.

Lemma 7.14 (Regularization of Shifts). *Let* $\mathcal{R} = \{r_1, ..., r_k\} \in C$ *and* $\eta_2 \geq \eta_1 > 0$. *Assume*

$$\eta_1 + \eta_2 \leq \frac{\mathrm{gap}(H)}{2}.$$

Let $w_1, ..., w_k \sim \mathrm{Unif}(\mathbb{D}(0, \eta_2))$ *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* $\mathrm{dist}(\check{\mathcal{R}}, \mathrm{Spec}\, H) \geq \eta_1$.

Proof. Define the bad region $\mathcal{B} \subset C$ as the union of disks $\mathcal{B} \triangleq \bigcup_{\lambda \in \mathrm{Spec}\, H} \mathbb{D}(\lambda, \eta_1)$. The assumption $\eta_1 + \eta_2 \leq \mathrm{gap}(H)/2$ implies that for each r_i, the disk $\mathbb{D}(r_i, \eta_2)$ intersects at most one disk in $\mathcal{B}$; since $\check{r}_i$ is distributed uniformly in $\mathbb{D}(r_i, \eta_2)$ 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.

□

7.5 Finding Well Conditioned Optimal Ritz Values (or Decoupling Early)

The shifting strategy $\mathsf{Sh}_{k,B}$ from Chapter 6 uses the notion of θ-*optimal* Ritz values from Definition 6.3, and we assumed there the existence of a black box algorithm for computing these. In this section we will show how to compute θ-optimal Ritz values for a Hessenberg matrix H which are moreover not too close to $\operatorname{Spec} H$, or else guarantee rapid decoupling. By Lemma 7.12 that any IQR step using such *well-conditioned* approximate Ritz values as shifts is forward stable.

The procedure consists of two steps, and relies on a black box algorithm, SmallEig, for computing forward approximations of the eigenvalues of a $k \times k$ or smaller matrix. 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 7.17). This is the content of Theorem 7.15, which is established in Section 7.5. We give a finite arithmetic implementation of this dichotomy in Section 7.5.

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 7.15 (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*

$$\theta \geq (2\kappa_V^4(H))^{1/2k} \quad \textit{and} \quad \frac{\beta}{\operatorname{gap}(H)} \leq \frac{1}{2}\left(\frac{\theta}{(2\kappa_V^4(H))^{1/2k}} - 1\right) =: c \tag{7.16}$$

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*

$$\|e_n^*(H - r_i)^{-k}\|^{1/k} \geq \frac{1}{2\kappa_V(H)^{2/k}} \cdot \left(\frac{\psi_k(H)}{\|H\| + \beta}\right) \cdot \left(\frac{1 - \frac{(2\kappa_V^4)^{1/2k}}{\theta}}{\beta}\right). \tag{7.17}$$

The remainder of this subsection is dedicated to the proof of Theorem 7.15. Let $\mathrm{P} = \{\rho_1, \dots, \rho_k\}$ and $\mathcal{R} = \{r_1, \dots, r_k\}$ be as in Lemma 7.15, 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 7.15 will be to show that if (i) does not hold, then (ii) does; assuming the former, we can get that

$$\begin{aligned}
\mathbb{E}[|p(Z_H)|^2] &\geq \frac{\|e_n^* p(H)\|^2}{\kappa_V(H)^2} && \text{Lemma 6.10} \\
&\geq \frac{\theta^{2k}\|e_n^*\chi(H)\|^2}{\kappa_V(H)^2} && \text{Negation of i} \\
&\geq \frac{\theta^{2k}\mathbb{E}[|\chi(Z_H)|^2]}{\kappa_V(H)^4} && \text{Lemma 6.10} \qquad (7.18) \\
&= 2(1+2c)^{2k}\mathbb{E}[|\chi(Z_H)|^2] && (7.16) \qquad (7.19)
\end{aligned}$$

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 7.15, 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 7.16. *Assume that* $\frac{\beta}{c} \leq \mathrm{gap}(H)$ *with* c *defined as in* (7.16), $q(z) \triangleq (z - s_1)\cdots(z - s_k)$ *for some* $\mathcal{S} = \{s_1, \dots, s_k\} \subset \mathbb{D}(0, \|H\|)$, *and let* $\check{q}(z) \triangleq (z - \check{s}_1)\cdots(z - \check{s}_k)$ *with* $\check{s}_1, \dots, \check{s}_k \in \mathbb{C}$ *satisfying*

$$\max_{i\in[k]} |s_i - \check{s}_i| \leq \beta.$$

Then

$$\mathbb{P}\left[\mathrm{dist}(Z_H, \{s_1, \dots, s_k\}) \leq \frac{\beta}{2c}\right] \geq \frac{\mathbb{E}[|\check{q}(Z_H)|^2] - (1+2c)^{2k}\mathbb{E}[|q(Z_H)|^2]}{(2(\|H\| + \beta)(1+2c))^{2k}}.$$

Lemma in hand, we can now complete the proof.

Proof of Theorem 7.15. Using Lemma 7.16 with $q(z) = \chi(z) = (z - \rho_1)\dots(z - \rho_k)$ and $\check{q}(z) = p(z) =$

$(z - r_1)\dots(z - r_k)$, we find that

$$\begin{aligned}
\mathbb{P}\left[\operatorname{dist}(Z_H, \mathrm{P}) \le \frac{\beta}{2c}\right] &\ge \frac{\mathbb{E}[|p(Z_H)|^2] - (1+2c)^{2k}\mathbb{E}[|\chi(Z_H)|^2]}{(2(\|H\| + \beta)(1+2c))^{2k}} \\
&\ge \frac{\mathbb{E}[|\chi(Z_H)|^2]}{2^{2k}(\|H\| + \beta)^{2k}} && (7.19) \\
&\ge \frac{\|e_n^*\chi(H)\|^2}{2^{2k}\kappa_V(H)^2(\|H\| + \beta)^{2k}} && \text{Lemma 6.10} \\
&= \frac{\psi_k^{2k}(H)}{2^{2k}\kappa_V(H)^2(\|H\| + \beta)^{2k}} && \text{Lemma 6.2.}
\end{aligned}$$

Since the right hand side is nonzero and Z_H is supported on the spectrum of H (and since $c \le 1/2$ by assumption) this implies that for some $i \in [k]$ and $\lambda \in \operatorname{Spec} H$

$$|\rho_i - \lambda| \le \frac{\beta}{2c}.$$

On the other hand, as we are assuming $\beta/c \le \operatorname{gap}(H)$, 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 $\beta/c \le \operatorname{gap}(H)$ from one another. Since there are only k of the ρ_i's, at least one eigenvalue, say λ, must be at least $\beta/2c$-close some ρ_i and additionally satisfy

$$\mathbb{P}[Z_H = \lambda] \ge \frac{1}{k}\left(\frac{\psi_k(H)}{2\kappa_V(H)^{1/k}(\|H\| + \beta)}\right)^{2k}. \tag{7.20}$$

By the triangle inequality, we then have

$$|r_i - \lambda| \le |r_i - \rho_i| + |\rho_i - \lambda| \le \beta\left(1 + \frac{1}{2c}\right). \tag{7.21}$$

Finally,

$$\begin{aligned}
\|e_n^*(H - r_i)^{-k}\|^{1/k} &\ge \frac{\mathbb{E}\left[|Z_H - r|^{-2k}\right]^{1/2k}}{\kappa_V(H)^{1/k}} && \text{Lemma 6.10} \\
&\ge \frac{1}{\kappa_V(H)^{1/k}} \cdot \frac{1}{(2k)^{1/2k}}\left(\frac{\psi_k(H)}{\kappa_V(H)^{1/k}(\|H\| + \beta)}\right) \cdot \left(\frac{2c}{(2c+1)\beta}\right),
\end{aligned}$$

where the second inequality uses $\mathbb{E}\left[|Z_H - r_i|^{-2k}\right] \ge \frac{\mathbb{P}[Z_H = \lambda]}{|\lambda - r|^{2k}}$ and (7.20), (7.21). This yields the conclusion by substituting c and noting that $(2k)^{1/2k} \le 2$. □

Remark 7.17. By (7.20) and (7.21), 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 7.16. We begin by partitioning the set $\mathcal{S} = \{s_1, ..., s_k\}$ according to which eigenvalue of H is the closest: relabelling $\operatorname{Spec} H = \{\lambda_1, ..., \lambda_n\}$ 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 $q_0, \dots, q_l$ with $l \leq k$ given by $q_0(z) = q(z)$ and

$$q_{j+1}(z) \triangleq \frac{\prod_{i \in S_{j+1}}(z - \check{s}_i)}{\prod_{i \in S_{j+1}}(z - s_i)} q_j(z);$$

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 7.18. *For each $j = 1, ..., \ell$, we have*

$$\mathbb{E}[|q_j(Z_H)|^2] \leq (1 + 2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] + (2(\|H\| + \beta))^{2k}\mathbb{P}[Z_H = \lambda_j]\mathbf{1}\{\operatorname{dist}(\lambda_j, \mathcal{S}) \leq \tfrac{\beta}{2c}\}.$$

In view of the claim, we can inductively assemble these bounds to compare $\mathbb{E}[|q(Z_H)|^2]$ and $\mathbb{E}[|\check{q}(Z_H)^2]$:

$$\begin{aligned}
\mathbb{E}[|\check{q}(Z_H)|^2] &= \mathbb{E}[|q_\ell(Z_H)|^2] \\
&\leq (1 + 2c)^{2|S_\ell|}\mathbb{E}[|q_{\ell-1}(Z_H)|^2] + (2(\|H\| + \beta))^{2k}\mathbb{P}[Z_H = \lambda_\ell]\mathbf{1}\{\operatorname{dist}(\lambda_\ell, \mathcal{S}) \leq \tfrac{\beta}{2c}\} \\
&\leq (1 + 2c)^{2k}\mathbb{E}[|q_0(Z_H)|^2] \\
&\qquad + \sum_{i \in [\ell]} (2(\|H\| + \beta))^{2k}(1 + 2c)^{2\sum_{j=1}^{i}|S_i|}\mathbb{P}[Z_H = \lambda_i]\mathbf{1}\{\operatorname{dist}(\lambda_i, \mathcal{S}) \leq \tfrac{\beta}{2c}\} \\
&\leq (1 + 2c)^{2k}\left(\mathbb{E}[|q(Z_H)|^2] + (2(\|H\| + \beta))^{2k}\sum_{i \in [\ell]}\mathbb{P}[Z_H = \lambda_i]\mathbf{1}\{\operatorname{dist}(\lambda_i, \mathcal{S}) \leq \tfrac{\beta}{2c}\}\right) \\
&\leq (1 + 2c)^{2k}\left(\mathbb{E}[|q(Z_H)|^2] + (2(\|H\| + \beta))^{2k}\mathbb{P}\left[\operatorname{dist}(Z_H, \mathcal{S}) \leq \tfrac{\beta}{2c}\right]\right).
\end{aligned}$$

Rearranging gives the bound advertised in the lemma.

It remains to prove Claim 7.18. 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

$$\begin{aligned}\mathbb{E}[|q_j(Z_H)|^2] &= \mathbb{P}[Z_H = \lambda_j]|q_j(\lambda_j)|^2 + \sum_{m\in[n]\setminus j} \mathbb{P}[Z_H = \lambda_m]|q_{j-1}(\lambda_m)|^2 \prod_{s\in S_{j+1}} \left|\frac{\lambda_m - \check{s}}{\lambda_m - s}\right|^2 \\ &\leq \mathbb{P}[Z_H = \lambda_j]|q_j(\lambda_j)|^2 + (1+2c)^{2|S_j|} \sum_{m\in[n]\setminus j} \mathbb{P}[Z_H = \lambda_m]|q_{j-1}(\lambda_m)|^2 \\ &\leq \mathbb{P}[Z_H = \lambda_j]\left(|q_j(\lambda_j)|^2 - (1+2c)^{2|S_j|}|q_j(\lambda_{j-1})|^2\right) + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] \qquad (7.22) \\ &\leq \mathbb{P}[Z_H = \lambda_j]|q_{j-1}(\lambda_j)|^2 \left(\prod_{s\in S_j}\left(1 + \left|\frac{s-\check{s}}{\lambda_j - s}\right|\right)^2 - (1+2c)^{2|S_j|}\right) \\ &\qquad + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] \qquad (7.23)\end{aligned}$$

We have defined S_j so that λ_j is the closest eigenvalue to every $s \in S_j$, so $\operatorname{dist}(\lambda_j, \mathcal{S}) = \operatorname{dist}(\lambda_j, S_j)$. Thus when $\operatorname{dist}(\lambda_j, \mathcal{S}) > \frac{\beta}{2c}$, we can rearrange to see that

$$\begin{aligned}0 &\geq \left(1 + \frac{\beta}{\operatorname{dist}(\lambda_j, S_j)}\right)^{2|S_j|} - (1+2c)^{2|S_j|} \\ &\geq \prod_{s\in S_j}\left(1 + \frac{|s-\check{s}|}{|\lambda_j - s|}\right)^2 - (1+2c)^{2|S_j|};\end{aligned}$$

the latter is a factor of the first term on the right hand side of (7.23), so in the event $\operatorname{dist}(\lambda_j, \mathcal{S}) > \frac{\beta}{2c}$ 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, (7.22) implies that independent of $\operatorname{dist}(\lambda_j, \mathcal{S})$ — and thus in particular when $\operatorname{dist}(\lambda_j, \mathcal{S}) \leq \frac{\beta}{2c}$ — we have the inequality

$$\begin{aligned}\mathbb{E}[|q_j(Z_H)|^2] &\leq \mathbb{P}[Z_H = \lambda_j]|q_j(\lambda_j)|^2 + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] \\ &\leq \mathbb{P}[Z_H = \lambda_j](2(\|H\| + \beta))^{2k} + (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2.\end{aligned}$$

For the final line, note that $\lambda_j \in \mathbb{D}(0, \|H\|)$ and, because $\mathcal{S} \subset \mathbb{D}(0, \|H\|)$, and $|\check{s} - s| \leq \beta$ for every $s \in \mathcal{S}$, the roots of each q_j are contained in $\mathbb{D}(0, \|H\| + \beta)$. Combining the bounds on $\mathbb{E}[|q_j(Z_H)|^2]$ in the cases $\operatorname{dist}(\lambda_j, \mathcal{S}) > \frac{\beta}{2c}$ and $\operatorname{dist}(\lambda_j, \mathcal{S}) \leq \frac{\beta}{2c}$, we find that

$$\mathbb{E}[|q_j(Z_H)|^2] \leq (1+2c)^{2|S_j|}\mathbb{E}[|q_{j-1}(Z_H)|^2] + (2(\|H\| + \beta))^{2k}\mathbb{P}[Z_H = \lambda_j]\mathbf{1}\left[\operatorname{dist}(\lambda_j, \mathcal{S}) \leq \tfrac{\beta}{2c}\right],$$

estabilishing the claim. □

Finite Arithmetic Implementation of RITZ-OR-DEC

In this subsection we combine Theorem 7.15 and the regularization procedure of Lemma 7.14 to obtain a finite arithmetic algorithm, RITZ-OR-DEC, for finding θ-optimal Ritz values in the sense of Definition 6.3, for θ set as in 7.1, and with the additional feature that the Ritz values are not too close to Spec H. The first step is testing whether a set of putative approximate Ritz values are θ-optimal.

OPTIMAL

Input: Hessenberg $H \in C^{n\times n}$, $\{s_1, \dots, s_k\} = \mathcal{S} \subset C$
Global Data: Optimality parameter θ
Output: Optimality flag `opt`
Requires:
Ensures: If `opt = true`, then $\mathcal{S}$ are θ-optimal; if `opt = false`, then they are not $(.998^{1/k}\theta)$-optimal.

```
1. $\widetilde{v}_0 \longleftarrow e_n$
2. For $j = 0, \dots, k-1$,
    (a) $\widetilde{v_{j+1}} \longleftarrow \mathsf{fl}\left((H - s_{j+1})^* \widetilde{v_j}\right)$
3. If $\mathsf{fl}(\|\widetilde{v}_k\|) \geq .999\theta^k \psi_k^k(H)$, opt ⟵ false, else opt ⟵ true.
```

Lemma 7.19 (Guarantees for OPTIMAL). *Assume that $s_1, \dots, s_k \in \mathbb{D}(0, C\|H\|)$ and*

$$\mathbf{u} \leq \mathbf{u}_{\mathsf{OPTIMAL}}(n, k, C, \|H\|, \theta) \triangleq \frac{1}{2 \cdot 10^3 n^2} \left(\frac{\psi_k(H)}{\theta(2 + 2C)\|H\|} \right)^k = 2^{-O\left(\lg n + k \lg \frac{\theta\|H\|}{\psi_k(H)}\right)}; \tag{7.24}$$

then OPTIMAL *satisfies its guarantees and runs in at most $T_{\mathsf{OPTIMAL}}(k) \triangleq 4k^2 = O(k^2)$ arithmetic operations.*

Proof. 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 $\mathbf{u} \leq n^{-3/2}$,

$$\begin{aligned} \left|\mathsf{fl}\left(\|\widetilde{v}_k\|\right) - \|e_n^* p(H)\|\right| &\leq n\mathbf{u}\|\widetilde{v}_k\| + \left|\|\widetilde{v}_k\| - \|e_n^* p(H)\|\right| \\ &\leq n\mathbf{u}\left((1 + C)\|H\|(1 + n^{3/2}\mathbf{u})\right)^k + kn^{3/2}\mathbf{u} \cdot \left((1 + C)\|H\|(1 + n^{3/2}\mathbf{u})\right)^k \\ &\leq 2n^2(2 + 2C)^k\|H\|^k\mathbf{u}. \end{aligned}$$

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. □

RITZ-OR-DEC

Input: Hessenberg H, working accuracy ω, failure probability ϕ
Global Data: Norm bound Σ, optimality parameter θ as in Table 7.1
Requires: H is absolutely ω-decoupled, $\|H\| \leq \Sigma$, $\mathrm{gap}(H) \geq \frac{2\omega^2}{\Sigma}$, $k/\phi \geq 2$.
Output: Hessenberg $\widehat{H}$, θ-approximate Ritz values $\check{\mathcal{R}}$, decoupling flag `dec`.
Ensures: With probability at least $1-\phi$, $\mathrm{dist}(\check{\mathcal{R}}, \mathrm{Spec}\, H) \geq \eta_1$ (as defined in line 1) *and* exactly one of the following holds:

- `dec = false`, $\widehat{H} = H$, and $\check{\mathcal{R}}$ is an exact set of θ-optimal Ritz values of H, satisfying $\check{\mathcal{R}} \subset \mathbb{D}(0, 1.1\|H\|)$.
- `dec = true` and for some $\check{r} \in \check{\mathcal{R}}$, $\widehat{H} = \mathsf{IQR}(H, (z-\check{r})^k)$ is absolutely ω-decoupled.

```
1. $\beta \longleftarrow \frac{\omega^2}{16\cdot 101\cdot\Sigma}$, $\eta_2 \longleftarrow \frac{\beta}{2}$, $\eta_1 \longleftarrow \frac{\eta_2}{\sqrt{2k/\phi}} = \frac{\omega^2\sqrt{\phi}}{32\cdot 101\cdot\Sigma\sqrt{2k}}$
2. $\mathcal{R} \longleftarrow \mathsf{SmallEig}\left(H_{(k)}, \beta/2, \texttt{forward}, \phi/2\right)$
3. $\{\check{r}_1, \dots, \check{r}_k\} = \check{\mathcal{R}} \longleftarrow \{r_1 + w_1, \dots, r_k + w_k\}$, where the $w_i$ are i.i.d samples from $\mathrm{Unif}\big(\mathbb{D}(0, \eta_2)\big)$
4. If $\mathsf{OPTIMAL}(\check{\mathcal{R}}, H, \theta) = \texttt{true}$, set $\widehat{H} \longleftarrow H$ and $\texttt{dec} \longleftarrow \texttt{false}$
5. Else if $\mathsf{OPTIMAL}(\check{\mathcal{R}}, H, \theta) = \texttt{false}$, for $i = 1, ..., k$
    (a) $\widehat{H} \longleftarrow \mathsf{IQR}(H, (z - \check{r}_i)^k)$
    (b) If $\widehat{H}_{j+1,j} \leq \omega$ for any $j \in \{n-k, n-k+1, \dots, n-1\}$, set $\texttt{dec} \longleftarrow \texttt{true}$ and halt
```

Lemma 7.20 (Guarantees for $\mathsf{RITZ\text{-}OR\text{-}DEC}$). *Assuming that*

$$\begin{aligned} \mathbf{u} &\leq \mathbf{u}_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n, k, \Sigma, B, \theta, \omega, \phi) \\ &\triangleq \min\left\{\mathbf{u}_{\mathsf{OPTIMAL}}(n, k, 1.1, \Sigma, \theta), \frac{\omega}{8n^{1/2}\Sigma}\mathbf{u}_{\mathsf{IQR}}\left(n, k, 1.1, \Sigma, B, \frac{\omega^2\sqrt{\phi}}{32\cdot 101\cdot\Sigma\sqrt{2k}}\right)\right\} && (7.25) \\ &= 2^{-O\left(\lg nB + k\lg\frac{\theta\|H\|\cdot k\Sigma}{\omega\phi}\right)} && (7.26) \end{aligned}$$

then $\mathsf{RITZ\text{-}OR\text{-}DEC}$ *satisfies its guarantees and its running time depends on the value of the decoupling flag. In either case it makes one call to* $\mathsf{SmallEig}$, *in addition to that call*

(i) if `dec = false`, $\mathsf{RITZ\text{-}OR\text{-}DEC}$ *uses at most*

$$T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n, k, \texttt{false}) \triangleq kC_{\mathsf{D}} + k + T_{\mathsf{OPTIMAL}}(k) = O(k^2)$$

arithmetic operations.

(ii) otherwise, $\mathsf{RITZ\text{-}OR\text{-}DEC}$ *uses at most*

$$T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n, k, \texttt{true}) \triangleq T_{\mathsf{OPTIMAL}}(k) + k(T_{\mathsf{IQR}}(n, k) + k + C_{\mathsf{D}} + 1) = O(k^2 n^2)$$

arithmetic operations.

Proof. First, the assumptions of $\mathsf{RITZ\text{-}OR\text{-}DEC}$ on its input parameters imply that

$$\eta_1 + \eta_2 \leq \beta \leq \frac{\omega^2}{\Sigma} \leq \mathrm{gap}(H)/2$$

so we can apply Lemma 7.14 to find that $\mathrm{dist}(\check{\mathcal{R}}, \mathrm{Spec}\, H) \geq \eta_1$ 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 $\mathsf{SmallEig}$, $\mathcal{R}$ 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

$$\beta \leq 0.1\omega \leq 0.1\|H\|$$

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 $\mathsf{OPTIMAL}(\check{\mathcal{R}}, H, \theta)$ outputs `opt = true` we are guaranteed that $\check{\mathcal{R}}$ is indeed a set of θ-optimal Ritz values for H. On the other hand if `opt = false`, then by Lemma 7.19 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 7.15:

$$c = \frac{1}{2}\left(\frac{0.998^{1/k}\theta}{(2\kappa_V(H)^4)^{1/2k}} - 1\right) \geq \frac{1}{2}\left(\frac{\frac{101}{100}(2B^4)^{1/2k}}{(2B^4)^{1/2k}} - 1\right) = \frac{1}{200} \geq \frac{\beta}{\mathrm{gap}(H)},$$

and conclude that there is some $\check{r} \in \check{\mathcal{R}}$ for which

$$\begin{aligned} \|e_n^*(H-\check{r})^{-k}\|^{1/k} &\geq \frac{1}{2\kappa_V(H)^{2/k}} \cdot \left(\frac{\psi_k(H)}{\|H\|+\beta}\right) \cdot \left(\frac{1-\frac{(2\kappa_V^4)^{1/2k}}{0.998^{1/k}\theta}}{\beta}\right) \\ &\geq \frac{1}{4} \cdot \left(\frac{\omega}{2\Sigma}\right) \cdot \left(\frac{1-\frac{100}{101}}{\beta}\right) \qquad B^{2/k} \leq 2,\ \psi_k(H) \leq \omega,\ \beta \leq \|H\| \\ &\geq \frac{2}{\omega} \end{aligned}$$

by the definition of β in line 1. In the event that $\mathrm{dist}(\check{\mathcal{R}}, \mathrm{Spec}\, H) \geq \eta_1$, our choice of $\mathbf{u}$ in (7.25) means that we can apply Lemma 7.12 to $\widehat{H} = \mathsf{IQR}(H, (z-\check{r})^k)$ with $C = 1.1$, giving

$$\|\widehat{H} - \mathsf{iqr}(H, (z-\check{r})^k)\|_F \leq 32\kappa_V(H)\|H\| \left(\frac{4.2\|H\|}{\mathrm{dist}(\check{r}, \mathrm{Spec}\, H)}\right)^k n^{1/2} \nu_{\mathsf{IQR}}(n)\mathbf{u} \leq \omega/2.$$

Using $\psi_k(\mathsf{iqr}(H, (z-\check{r})^k) \leq \tau_{(z-\check{r})^k}(H) \leq \omega/2$, we find that $\mathsf{iqr}(H, (z-\check{r})^k)$ is absolutely $\omega/2$-decoupled, so $\widehat{H}$ must be ω-decoupled, completing the proof of correctness.

To analyze the running time, note that when `dec = false` other than the call to $\mathsf{SmallEig}$, in line 3 k samples are taken from $\mathrm{Unif}(\mathbb{D}(0, \eta_2))$ 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 `dec = true`, at most k calls to IQR with degree k are made and each time k subdiagonals of $\hat{H}$ are checked, adding $kT_{\mathsf{IQR}}(n,k) + k^2$ operations. □

7.6 Finite Arithmetic Analysis of One Iteration of $\mathsf{SH}_{k,B}$

In this section we provide the finite arithmetic implementation and analysis of a single iteration of the shifting strategy $\mathsf{Sh}_{k,B}$ from Chapter 6. In exact arithmetic, $\mathsf{SH}_{k,B}$ 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 the finite arithmetic counterparts of these routines, FIND and EXC, and and state their guarantees.

Computation of τ and ψ_k.

The shifting strategy $\mathsf{SH}_{k,B}$ needs access to both $\tau_p(H)$ and $\psi_k(H)$. The former can be computed using Lemma 7.13. For the latter, we will assume for simplicity that $\psi_k^k(H)$ 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

$$\widetilde{\psi}_k(H) \triangleq \mathsf{fl}\left(\left(\psi_k^k(H)\right)^{1/k}\right),$$

and assume

$$|\widetilde{\psi}_k(H) - \psi_k(H)| \le (1 - 0.999^{1/k})\psi_k(H) \le 0.001\psi_k(H), \tag{7.27}$$

which as per Lemma 7.7 can be computed in $T_\psi(k) \triangleq k + T_{\mathsf{root}}(k, 1 - 0.999^{1/k})$ arithmetic operations provided that

$$\mathbf{u} \le \mathbf{u}_\psi(k) \triangleq \frac{1 - 0.999^{1/k}}{k(c_{\mathsf{root}} + 1 - 0.999^{1/k})} = 2^{-O(\lg k)}. \tag{7.28}$$

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 $(1.01\kappa_V(H))^{\frac{4\lg k}{k}}$-promising Ritz value — as opposed the $\kappa_V(H)^{\frac{4\lg k}{k}}$-optimality we are guaranteed in the exact case.

```
                              FIND
Input: Hessenberg $H$, a set $\mathcal{R} = \{r_1, \dots, r_k\} \subset \mathbb{C}$
Global Data: Promising parameter $\alpha = (1.01B)^{\frac{4\lg k}{k}}$
Output: A complex number $r \in \mathcal{R}$
Requires: $\psi_k(H) > 0$
Ensures: $r$ is $\alpha$-promising for $\alpha$ as in Table 7.1

  1. For $j = 1, \dots, \lg k$
     (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) $\mathcal{R} \longleftarrow \mathcal{R}_{\widetilde{b}_j}$, 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 7.21 (Guarantees for FIND). *Assume that $\mathcal{R} \subset \mathbb{D}(0, C\|H\|)$ and*

$$\begin{aligned} \mathbf{u} \le \mathbf{u}_{\mathsf{FIND}}\,&(n, k, C, \|H\|, \kappa_V(H), \operatorname{dist}(\mathcal{R}, \operatorname{Spec} H)) \\ &\triangleq \mathbf{u}_{\mathsf{TAU}}(n, k/2, C, \|H\|, \kappa_V(H), \operatorname{dist}(\mathcal{R}, \operatorname{Spec} H)) \\ &= 2^{-O\left(\lg n\kappa_V(H) + k \lg \frac{\|H\|}{\operatorname{dist}(\mathcal{R}, \operatorname{Spec} H)}\right)}. \end{aligned} \tag{7.29}$$

Then FIND satisfies its guarantees, and runs in

$$T_{\mathsf{FIND}}(n, k) \triangleq 2 \lg k T_{\mathsf{TAU}}(n, k/2) + \lg k = O(k \lg k \cdot n^2)$$

arithmetic operations.

Proof. This proof will slightly modify that of Lemma 6.13, where we analyzed find in exact arithmetic. The definition of $\mathbf{u}_{\mathsf{FIND}}$ is sufficient to let us invoke Lemma 7.13 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 7.13, for each $b \in \{0, 1\}$ we have

$$\left| \mathsf{TAU}^{k/2}(H, p_{j,b}) - \|e_n^* p_{j,b}(H)^{-1}\|^{-1} \right| \le 0.0011 \|e_n^* p_{j,b}(H)^{-1}\|^{-1},$$

and thus it always holds that

$$\|e_n^* p_{j,\widetilde{b_j}}(H)^{-1}\|^2 \ge (1 - 0.0022)^2 \|p_{j,b_j}(H)^{-1}\|^2 \ge \frac{1}{2.02} \left(\|p_{j,0}(H)^{-1}\|^2 + \|p_{j,1}(H)^{-1}\|^2 \right).$$

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

$$p_{j,\widetilde{b_j}}(z) = p_{j+1,0}(z) p_{j+1,1}(z). \tag{7.30}$$

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

$$\begin{aligned} \|e_n^* p_{1,\widetilde{b_1}}(H)^{-1}\|^2 &\ge \frac{1}{2.02} \left(\|e_n^* p_{1,0}(H)^{-1}\|^2 + \|e_n^* p_{1,1}(H)^{-1}\|^2 \right) \\ &\ge \frac{1}{1.01 \kappa_V(H)^2} \mathbb{E}\left[\frac{1}{2} \left(|p_{1,0}(Z_H)|^{-2} + |p_{1,1}(Z_H)|^{-2} \right) \right] && \text{Lemma 6.10} \\ &\ge \frac{1}{1.01 \kappa_V(H)^2} \mathbb{E}[|p(Z_H)|^{-1}] && \text{AM/GM and (7.30)} \end{aligned}$$

Applying the same argument to each subsequent step,

$$\begin{aligned}
\|e_n^* p_{j+1,\widetilde{b_{j+1}}}(H)^{-2^j}\|^2 &\ge \frac{1}{1.01\kappa_V(H)^2}\mathbb{E}\left[\frac{1}{2}\left(|p_{j+1,0}(Z_H)|^{-2^{j+1}} + |p_{j+1,1}(Z_H)|^{-2^{j+1}}\right)\right] && \text{Lemma 6.10} \\
&\ge \frac{1}{1.01\kappa_V(H)^2}\mathbb{E}\left[|p_{j+1,0}(Z_H)p_{j+1,1}(Z_H)|^{-2^j}\right] && \text{AM/GM} \\
&\ge \frac{1}{1.01\kappa_V(H)^4}\|e_n^*(p_{j+1,0}(H)p_{j+1,1}(H))^{-2^{j-1}}\| && \text{Lemma 6.10} \\
&= \frac{1}{1.01\kappa_V(H)^4}\|e_n^* p_{j,\widetilde{b_j}}(H)^{-2^{j-1}}\|. && (7.30)
\end{aligned}$$

Paying a further $\kappa_V(H)^2$ on the final step to convert the norm into an expectation, we get

$$\mathbb{E}\left[|Z_H - r|^{-k}\right] \ge \left(\frac{1}{1.01\kappa_V(H)}\right)^{4\lg k} \mathbb{E}\left[|p(Z_H)|^{-1}\right]$$

as promised.

For the runtime, we make $2\lg k$ calls to $\mathsf{TAU}^{k/2}$ and $\lg k$ 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. As with FIND, we will proceed similarly to the exact arithmetic setting. However, we will need to ensure that all of our exceptional shifts are suitably far from Spec H, so that the IQR steps executed with them will be forward stable. To achieve this, we will apply a random perturbation, in the same spirit as Section 7.4. Let us first pause to prove a key lemma ensuring potential reduction in finite arithmetic for sufficiently well-conditioned shifts. In particular, we will use the forward error guarantee of Lemma 7.12 to analyze the potential of $\mathsf{IQR}(H, p(z))$, by directly comparing it to that of $\mathsf{iqr}(H, p(z))$.

Lemma 7.22. *Let $p(z) = (z - s_1)...(z - s_m)$ for some floating point complex numbers $\mathcal{S} = \{s_1, ..., s_m\} \subset \mathbb{D}(0, C\|H\|)$, and assume that for some $\omega > 0$,*

$$\begin{aligned}
\mathbf{u} &\le \mathbf{u}_{7.22}(n, k, C, \|H\|, \kappa_V(H), \omega) \\
&\triangleq \frac{\omega}{10^3 n^{1/2}\|H\|}\mathbf{u}_{\mathsf{IQR}}(n, k, C, \|H\|, \kappa_V(H), \operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)) \qquad (7.31) \\
&= 2^{-O\left(\lg \frac{n\kappa_V(H)}{\omega} + k \lg \frac{\|H\|}{\operatorname{dist}(\mathcal{S}, \operatorname{Spec} H)}\right)}.
\end{aligned}$$

Then at least one of the following holds:

(i) (Decoupling) $\mathsf{IQR}(H, p(z))$ is absolutely ω-decoupled

(ii) (Potential Approximation) $\psi_k(\mathsf{IQR}(H, p(z))) \le 1.0011\psi_k(\mathsf{iqr}(H, p(z)))$.

Proof. Calling $\widetilde{\widehat{H}} = \mathsf{IQR}(H, p(z))$ and $\widehat{H} = \mathsf{iqr}(H, p(z))$, one of two cases are possible. If $\widehat{H}_{i+1,i} < 0.999\omega$ for some $i \in [n-1]$, then applying Lemma 7.12 and our assumption on $\mathbf{u}$,

$$\widetilde{\widehat{H}}_{i+1,i} < \widehat{H}_{i+1,i} + 0.001\omega < \omega.$$

On the other hand, if for every $i \in [n-1]$ we have $\widehat{H}_{i+1,i} \ge 0.999\omega$, then

$$\psi_k\left(\widetilde{\widehat{H}}\right) \le \psi_k\left(\widehat{H}\right)\left(\prod_{i\in[n-1]}\left(1 + \frac{0.001\omega}{\widehat{H}_{i+1,i}}\right)\right)^{1/k} \le 1.0011\psi_k\left(\widehat{H}\right).$$

□

EXC

Input: Hessenberg H, initial shift r, working accuracy ω, stagnation ratio ξ, failure probability tolerance ϕ
Global Data: Condition number bound B, norm bound Σ, optimality parameter θ, promising parameter α
Output: Finite subset $\mathcal{S} \subset \mathbb{C}$.
Requires: $\kappa_V(H) \le B$, $\|H\| \le \Sigma$, H is *not* absolutely ω-decoupled 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

- (Decoupling) $\mathsf{IQR}(H, (z-s)^k)$ is absolutely ω-decoupled
- (Potential Reduction) $\psi_k(\mathsf{IQR}(H, (z-s)^k)) \le 1.0011\gamma\psi_k(H)$.

```
1. $\tilde{R} \longleftarrow 2^{1/k}\alpha B^{1/k}\theta\widetilde{\psi_k}(H)$
2. $\varepsilon \longleftarrow \left(\frac{\xi\gamma}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}$
3. $\mathcal{S}_0 \longleftarrow$ maximal $0.99\varepsilon$-net of $\mathbb{D}\left(0, 1+\varepsilon\right)$
4. $w \sim \mathrm{Unif}\left(\mathbb{D}\left(0, \varepsilon\tilde{R}\right)\right)$
5. $\mathcal{S} \longleftarrow \mathsf{fl}\left((r + w + \tilde{R}\mathcal{S}_0)\right) \cap \mathbb{D}\left(r, \tilde{R}\right)$
```

Lemma 7.23 (Guarantees for EXC). *Assume that* $|r| + 1.001\theta\alpha B^{1/k}\psi_k(H) \leq C\|H\|$ *and*

$$\begin{aligned}\mathbf{u} &\leq \mathbf{u}_{\mathsf{EXC}}(n,k,C,\Sigma,B,\theta,\omega,\phi,\gamma,\xi,\alpha)\\ &\triangleq \min\left\{\mathbf{u}_\psi(k), \frac{0.1\varepsilon\cdot 1.998\theta\alpha B\omega}{4(\varepsilon + 2(1+\varepsilon)C\Sigma)},\right.\\ &\qquad \left.\mathbf{u}_{7.22}\left(n,k,C,\Sigma,B,\left(\frac{\xi\gamma}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\cdot\frac{1.998\,\theta\alpha B^{1/k}\omega\sqrt{\phi}}{\sqrt{3n}},\omega\right)\right\} &(7.32)\\ &= 2^{-O\left(k\log\frac{n\Sigma B\alpha\theta}{\xi\gamma\omega\phi}\right)}. &(7.33)\end{aligned}$$

Then EXC *satisfies its guarantees and runs in at most*

$$T_{\mathsf{EXC}}(n,k,\xi,\gamma,B,\alpha,\theta) \triangleq T_\psi(k) + 2S\left(\left(\frac{\xi\gamma}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\right) + C_{\mathsf{D}} + O(1) = O\left(B^{\frac{8}{k-1}}\left(\frac{\alpha^2\theta^2}{\xi\gamma}\right)^{\frac{2k}{k-1}}\right)$$

arithmetic operations and

$$|\mathcal{S}| \leq S\left(\left(\frac{\xi\gamma}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\right) = O\left(B^{\frac{8}{k-1}}\left(\frac{\alpha^2\theta^2}{\xi\gamma}\right)^{\frac{2k}{k-1}}\right)$$

where the function $S(\varepsilon) = O(\varepsilon^{-2})$ *is defined in* (7.35).

Proof. From (7.27), the fact that $\mathbf{u} \leq \mathbf{u}_\psi(k)$ we can bound

$$1.998\,\theta\alpha B\psi_k(H) \leq (2\cdot .999)^{1/k}\theta\alpha B\psi_k(H) \leq \tilde{R} \leq 1.001\cdot\theta\alpha B^{1/k}\psi_k(H), \qquad (7.34)$$

meaning that (as $\psi_k(H) \leq \|H\|$) 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

$$\begin{aligned}\mathbb{P}\left[Z_H \in \mathbb{D}(r,\tilde{R})\right] &\geq \mathbb{P}\left[|Z_H - r| \leq 1.998\,\theta\alpha\kappa_V^{1/k}(H)\psi_k(H)\right] &&\text{by (7.34)}\\ &\geq \left(1 - \frac{1}{1.998}\right)^2\frac{\xi^{2k}}{\kappa_V(H)^4\alpha^{2k}\theta^{2k}} &&\text{[17, Lemma 2.8] with } t = \frac{1}{1.998}\\ &\geq \frac{0.24\,\xi^{2k}}{B^4\alpha^{2k}\theta^{2k}}\\ &\triangleq P.\end{aligned}$$

When we shift and scale each point $s_0 \in \mathcal{S}_0$ in finite arithmetic,

$$\begin{aligned}|\mathsf{fl}(r + w + \tilde{R}s_0) - r + w + \tilde{R}s_0| &\leq \frac{3\mathbf{u}}{1 - 3\mathbf{u}}|r + w + \tilde{R}s_0|\\ &\leq 4\mathbf{u}\left(|r| + \varepsilon + (1+\varepsilon)1.001\theta\alpha B^{1/k}\psi_k(H)\right)\\ &\leq 4\mathbf{u}\left(\varepsilon + 2(1+\varepsilon)C\Sigma\right)\\ &\leq 0.1\varepsilon\cdot 1.998\,\theta\alpha B\omega\\ &\leq 0.1\varepsilon\tilde{R}\end{aligned}$$

from our assumption on $\mathbf{u}$, which means that the computed $\mathcal{S}$ still contains a $\varepsilon\tilde{R}$-net of $\mathbb{D}(r,\tilde{R})$. 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 $\mathbb{D}(r,\tilde{R})$) —this can be achieved, e.g., by intersecting with a slightly larger set and projecting all points outside $\mathbb{D}(r,\tilde{R})$) to this latter set. Since $\mathcal{S}$ is a maximal ε-net of $\mathbb{D}(r,\tilde{R})$), it has size at most $9/\varepsilon^2$, and we may recycle a calculation from [17],

$$\max_{s\in\mathcal{S}} \tau^{-2k}_{(z-s)^k}(H) \ge \frac{P}{9B^2\varepsilon^{2k-2}\tilde{R}^{2k}} \ge \frac{1}{\gamma^{2k}\psi_k^{2k}(H)}$$

provided that ε is no larger than

$$\begin{aligned}\left(\frac{P\gamma^{2k}\psi_k^{2k}(H)}{9B^2\tilde{R}^{2k}}\right)^{\frac{1}{2k-2}} &\ge \left(\frac{0.24\xi^{2k}\gamma^{2k}}{B^6\alpha^{2k}\theta^{2k}\cdot 9\cdot 2.001^2\theta^{2k}\alpha^{2k}B^2}\right)^{\frac{1}{2k-2}}\\ &\ge \left(\frac{\xi\gamma}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}},\end{aligned}$$

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 $\varepsilon\tilde{R}$ of one another) each eigenvalue $\lambda \in \operatorname{Spec} H$ lies within distance $\varepsilon\tilde{R}$ of at most three points in the net, so the probability that $\operatorname{dist}(\lambda,\mathcal{S}) < \eta$ after the random translation is at most $3\eta^2/\varepsilon^2\tilde{R}^2$. Thus the probability that $\operatorname{dist}(\operatorname{Spec} H,\mathcal{S}) < \eta$ after the random translation is at most $3n\eta^2/\varepsilon^2\tilde{R}^2$. To ensure that this is smaller than the failure probability ϕ, we can safely set

$$\eta = \frac{\varepsilon\tilde{R}\sqrt{\phi}}{\sqrt{3n}} \ge \left(\frac{\xi\gamma}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}} \cdot \frac{1.998\theta\alpha B^{1/k}\omega\sqrt{\phi}}{\sqrt{3n}}.$$

In the event that the shifts are all forward stable, the definition of $\mathbf{u}_{\mathsf{EXC}}$ means that we can invoke Lemma 7.22: either some subdiagonal of $\mathsf{IQR}(H,(z-s)^k)$ is smaller than ω, or $\mathsf{IQR}(H,(z-s)^k)$ satisfies

$$\psi_k(\mathsf{IQR}(H,(z-s)^k)) < 1.0011\psi_k(\mathsf{iqr}(H,(z-s)^k)) \le 1.0011\tau_{(z-s)^k}(H) \le 1.0011\gamma\psi_k(H).$$

As discussed in our analysis of exc in Chapter 6, one can take the initial $.99\varepsilon$-net of $\mathbb{D}(0,(1+\varepsilon))$ is to take an equilateral triangular lattice with spacing $\sqrt{3}\varepsilon$ and intersect it with $\mathbb{D}(0,(1+1.99\varepsilon))$, in which case

$$\begin{aligned}|\mathcal{S}| \le |\mathcal{S}_0| &\le \frac{2\pi}{3\sqrt{3}}\left(1.99+\frac{1}{0.99\varepsilon}\right)^2 + \frac{4\sqrt{2}}{\sqrt{3}}\left(1.99+\frac{1}{0.99\varepsilon}\right)+1\\ &\triangleq S(\varepsilon)\end{aligned} \tag{7.35}$$

We will see below that every time EXC is called in the course of the full algorithm ShiftedQR, the same ε is used, depending only on the global data. Thus the original net of $\mathbb{D}(0, 1+\varepsilon)$ 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 $\tilde{R}$, and the cost of obtaining the single random sample, and we get a total of

$$2|\mathcal{S}| + C_{\mathsf{root}} k \lg(k \lg \tfrac{1}{1-0.999^{1/k}}) + O(1)$$

arithmetic operations. Bounding $|\mathcal{S}_0| \le S(\varepsilon)$ yields the assertion of the lemma. □

Analysis of $\mathsf{Sh}_{k,B}$

We now specify and analyze the complete shifting strategy $\mathsf{SH}_{k,K}$.

```
SH_{k,B}
Input: Hessenberg $H$, $\theta$-optimal Ritz values $\mathcal{R}$ of $H$, working accuracy $\omega$, failure probability tolerance $\phi$.
Global Data: Condition number bound $B$, decoupling rate $\gamma$, norm bound $\Sigma$, optimality parameter $\theta$, promising parameter $\alpha$
Output: Hessenberg $\widehat{H}$.
Requires: $H$ is absolutely $\omega$-unreduced and $\kappa_V(H) \le B$
Ensures: With probability at least $1-\phi$, either $\widehat{H}$ is $\omega$-decoupled or $\psi_k(\widehat{H}) \le 1.002\gamma\psi_k(H)$
  1. $r \longleftarrow \mathsf{FIND}(H, \mathcal{R})$
  2. If $\mathsf{TAU}^k(H, (z-r)^k) \le \gamma^k \psi_k^k(H)$, output $\widehat{H} = \mathsf{IQR}(H, (z-r)^k)$.
  3. Else, $\mathcal{S} \longleftarrow \mathsf{EXC}(H, r, \omega, 0.999\gamma, \phi)$.
  4. For each $s \in \mathcal{S}$, if $\psi_k(\mathsf{IQR}(H, (z-s)^k)) \le 1.002\gamma\psi_k(H)$ or some subdiagonal of $\mathsf{IQR}(H, (z-s)^k)$ is smaller than $\omega$, output $\widehat{H} = \mathsf{iqr}(H, (z-s)^k)$
```

Lemma 7.24 (Guarantees for $\mathsf{Sh}_{k,B}$). *Assume that* $|r| + 1.001\theta\alpha B^{1/k}\psi_k(H) \le C\|H\|$ *and*

$$\begin{aligned}
\mathbf{u} &\le \mathbf{u}_{\mathsf{SH}}(n, k, C, \Sigma, B, \mathrm{dist}(\mathcal{R}, \mathrm{Spec}\, H), \theta, \omega, \phi, \gamma, \alpha) && (7.36)\\
&\triangleq \min\Big\{\mathbf{u}_{\mathsf{FIND}}(n, k, C, \Sigma, B, \mathrm{dist}(\mathcal{R}, \mathrm{Spec}\, H)),\\
&\qquad \mathbf{u}_{\mathsf{EXC}}(n, k, C, \Sigma, B, \theta, \omega, \phi, \gamma, 0.999\gamma, \alpha),\\
&\qquad \mathbf{u}_{7.22}(n, k, C, \Sigma, B, \mathrm{dist}(\mathcal{R}, \mathrm{Spec}\, H), \omega)\Big\} && (7.37)\\
&= 2^{-O\left(k \log \frac{n\Sigma B\theta\alpha}{\gamma\omega\phi\, \mathrm{dist}(\mathcal{R}, \mathrm{Spec}\, H)}\right)}
\end{aligned}$$

Then, $\mathsf{SH}_{k,B}$ *satisfies its guarantees, and runs in at most*

$$\begin{aligned} T_{\mathsf{SH}}(n,k,\gamma,B,\alpha,\theta) &\triangleq T_{\mathsf{FIND}}(n,k) + T_{\mathsf{TAU}}(n,k) + T_{\mathsf{EXC}}(n,k,0.999\gamma,\gamma,B,\alpha,\theta) \\ &\qquad + S\left(\left(\frac{0.999\gamma^2}{(13B^4)^{1/k}\alpha^2\theta^2}\right)^{\frac{k}{k-1}}\right)\left(T_{\mathsf{IQR}}(n,k) + T_{\psi}(n,k)\right) \\ &= O\left(kn^2 B^{\frac{8}{k-1}}\left(\frac{\alpha\theta}{\gamma}\right)^{\frac{4k}{k-1}}\right) \end{aligned}$$

arithmetic operations.

Proof. The definition of $\mathbf{u}_{\mathsf{SH}}$ 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 7.1; in line 2 — because every subdiagonal of H is assumed larger than ω — we know from definition of $\mathbf{u}_{\mathsf{SH}}$ and Lemma 7.22 that if $\mathsf{TAU}^k(H,(z-r)^k) \le \gamma^k\psi_k^k(H)$, then

$$\begin{aligned} \psi_k(\mathsf{IQR}(H,(z-r)^k)) &\le 1.0011\psi_k(\mathsf{iqr}(H,(z-r)^k)) \\ &\le 1.0011\tau_{(z-r)^k}(H) \\ &\le 1.0011 \cdot \left(1.001\mathsf{TAU}^k(H,(z-r)^k)\right)^{1/k} \\ &\le 1.002\gamma\psi_k(H). \end{aligned}$$

On the other hand, if $\mathsf{TAU}^k(H,(z-r)^k) > \gamma^k\psi_k^k(H)$ in line 2, then using the guarantees for TAU^k,

$$\tau_{(z-r)^k}^k(H) > 0.999\mathsf{TAU}^k(H,(z-r)^k) \ge 0.999\gamma^k\psi_k(H).$$

Finally, EXC satisfies its guarantees from Lemma 7.23 when called with $\alpha = (1.01B)^{\frac{4\log k}{k}}$ and $\xi = 0.999^{1/k}\gamma$. Thus with probability at least $1-\phi$ at least one exceptional shift $s \in \mathcal{S}$ satisfies either decoupling (some subdiagonal smaller than ω) or potential reduction ($\psi_k(\mathsf{IQR}(H,(z-s)^k)) \le 1.0011\gamma\psi_k(H) \le 1.002\gamma\psi_k(H)$).

For the arithmetic operations, $\mathsf{SH}_{k,B}$ requires one call to FIND, one to TAU^k, one to EXC with stagnation ratio $\xi = 0.999\gamma$, and finally $|\mathcal{S}|$ calls to degree-k IQR. We can bound $|\mathcal{S}| \le S(\varepsilon)$, where ε is defined in the course of EXC with stagnation ratio parameter $\xi = 0.999\gamma$, and $S(\cdot)$ is defined in (7.35). Since and checking every shift in $\mathcal{S}$ for potential reduction dominates the arithmetic operations, we get that

$$T_{\mathsf{SH}}(n,k,B,\gamma,\alpha,\theta) = O\left(kn^2 \cdot B^{\frac{8}{k-1}}\left(\frac{\alpha\theta}{\gamma}\right)^{\frac{4k}{k-1}}\right).$$

□

7.7 Finite Arithmetic Analysis of ShiftedQR

We are now ready to analyze, in finite arithmetic, how the shifting strategy $\mathsf{Sh}_{k,B}$ introduced in Chapter 6 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: $\mathsf{DEFLATE}(H, \omega, k)$ takes as input a Hessenberg matrix H, deletes any of the bottom $k-1$ subdiagonal entries smaller than ω, and outputs the resulting diagonal blocks $H_1, H_1, \ldots.$ It runs in $T_{\mathsf{DEFLATE}}(H, \omega, k) = k$ arithmetic operations.

```
ShiftedQR
Input: Hessenberg matrix $H$, accuracy $\delta$, failure probability tolerance $\phi$
Global Data: Eigenvector condition number bound $B$, eigenvalue gap bound $\Gamma$, matrix norm bound $\Sigma$, original matrix dimension $n$
Requires: $\Sigma \geq 2\|H\|$, $B \geq 2\kappa_V(H)$, $\Gamma \leq \mathrm{gap}(H)/2$, $\delta \leq \Sigma$
Output: A multiset $\Lambda \subset \boldsymbol{C}$
Ensures: With probability at least $1-\phi$, $\Lambda$ are the eigenvalues of some $\widetilde{H}$ with $\|\widetilde{H} - H\| \leq \delta$

  1. $\omega \longleftarrow \frac{1}{2n} \min\left\{\delta, \frac{\Gamma}{8n^2B^2}\right\}$, $\varphi \longleftarrow \frac{\phi}{3n^2} \frac{\log 1.002\gamma}{\log \frac{\omega}{\Sigma}}$
  2. If $\dim(H) \leq k$, $\Lambda \longleftarrow \mathsf{SmallEig}(H, \delta, \phi)$, output $\Lambda$ and stop.
  3. Else $\Lambda \longleftarrow \emptyset$ and
       a) While $\max_{n-k+1 \leq i \leq n} H_{i,i-1} < \omega$,
            i. $[\mathcal{R}, \widehat{H}, \mathtt{dec}] = \mathsf{RITZ\text{-}OR\text{-}DEC}(H, \omega, \varphi)$
           ii. If $\mathtt{dec} = \mathtt{true}$, $H \longleftarrow \widehat{H}$ and end while
          iii. Else if $\mathtt{dec} = \mathtt{false}$, $H \longleftarrow \mathsf{SH}_{k,B}(H, \mathcal{R}, \omega, \varphi)$
       b) $[H_1, H_2, \ldots H_\ell] = \mathsf{DEFLATE}(H, \omega)$
       c) For each $j \in [\ell]$
            i. If $\dim(H_j) \leq k$, $\Lambda \longleftarrow \Lambda \sqcup \mathsf{SmallEig}(H_j, \delta/n, \phi/3n)$.
           ii. Else, repeat lines 3a-3c on $H_i$
```

Theorem 7.25 (Guarantees for ShiftedQR)**.** *Let k, θ, α, and γ be set in terms of B as in* (7.2)*, N_{dec} be* *defined as in* (7.41), *and* ω *and* φ *be defined as in line* 1 *of* ShiftedQR. *Assuming*

$$\begin{aligned}\mathbf{u} &\le \mathbf{u}_{\mathsf{ShiftedQR}}(n,k,\Sigma,B,\delta) \\ &\triangleq \min\left\{\frac{\omega}{4.5kN_{\mathsf{dec}}\cdot n\nu_{\mathsf{IQR}}(n)\Sigma}, \mathbf{u}_{\mathsf{RITZ\text{-}OR\text{-}DEC}}\left(n,k,\Sigma,B,\theta,\omega,\varphi\right),\right. \\ &\qquad\qquad \left.\mathbf{u}_{\mathsf{Sh}}\left(n,k,3,\Sigma,B,\frac{\omega^2\sqrt{\varphi}}{32\cdot 101\cdot\Sigma\sqrt{2k}},\theta,\omega,\varphi,\gamma,\alpha\right)\right\} \qquad (7.38)\\ &= 2^{-O\left(k\log\frac{n\Sigma B}{\delta\Gamma\phi}\right)},\end{aligned}$$

ShiftedQR *satisfies its guarantees and runs in at most*

$$\begin{aligned}T_{\mathsf{ShiftedQR}}(n,k,\delta,B,\Sigma,\gamma) \le n\Big(&T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n,k,\mathtt{true}) \\ &+ N_{\mathsf{dec}}\Big(T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n,k,\mathtt{false}) + T_{\mathsf{Sh}}(n,k,\gamma,B,\alpha,\theta)\Big) \qquad (7.39)\\ &+ T_{\mathsf{DEFLATE}}(k)\Big) \\ = O&\left(\left(\log\frac{nB\Sigma}{\delta\Gamma}k\log k + k^2\right)n^3\right)\end{aligned}$$

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* $\Omega(\frac{\phi}{n^2\log\frac{nB\Sigma}{\delta\Gamma}})$.

Theorem 1.14 follows immediately from 7.25.

Proof of Theorem 7.25. At a high level, ShiftedQR is given an input matrix H, ω-decouples H to a unitarily similar matrix $\widehat{H}$ via a sequence of applications of RITZ-OR-DEC + $\mathsf{SH}_{k,B}$, deflates $\widehat{H}$ 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 RITZ-OR-DEC 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

$$\begin{pmatrix} L_1 & * & * \\ & L_2 & * \\ & & \ddots \end{pmatrix}, \qquad (7.40)$$

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

$$\bigsqcup_j \mathsf{SmallEig}(L_j, \omega, \varphi) = \bigsqcup_j \operatorname{Spec} \widetilde{L}_j = \operatorname{Spec} \begin{pmatrix} \widetilde{L}_1 & * & * \\ & \widetilde{L}_2 & * \\ & & \ddots \end{pmatrix}$$

where $\widetilde{L}_1, \widetilde{L}_2, \ldots$ are some matrices satisfying $\|L_j - \widetilde{L}_j\| \le \delta/n$, and the remaining entries are identical to those in (7.40). Our goal in the proof will thus be to show that for some unitary $\widetilde{Q}$,

$$\left\| \begin{pmatrix} L_1 & * & * \\ & L_2 & * \\ & & \ddots \end{pmatrix} - \widetilde{Q}^* H \widetilde{Q} \right\| \le \delta - \delta/n,$$

where the left hand matrix is a block upper triangular matrix with the blocks $L_1, L_2, \ldots$ on the diagonal. This will in turn imply that

$$\begin{aligned} \left\| \begin{pmatrix} \widetilde{L}_1 & * & * \\ & \widetilde{L}_2 & * \\ & & \ddots \end{pmatrix} - \widetilde{Q}^* H \widetilde{Q} \right\| &\le \left\| \begin{pmatrix} L_1 - \widetilde{L}_1 & * & * \\ & L_2 - \widetilde{L}_2 & * \\ & & \ddots \end{pmatrix} \right\| + \delta - \delta/n \\ &\le \max_i \|L_i - \widetilde{L}_i\| + \delta - \delta/n \le \delta, \end{aligned}$$

as desired.

We begin by analyzing the while loop in line 3a.

Lemma 7.26. *Assume that in the course of* ShiftedQR*, the while loop in line 3a is initialized with a matrix H' satisfying $\|H'\| \le (1 - 1/2n)\Sigma$, $\kappa_V(H') \le (1 - 1/2n)B$, and $\operatorname{gap}(H') \ge (1 + 1/2n)\Gamma$. Let*

$$N_{\mathsf{dec}} \triangleq \frac{\log \frac{\Sigma}{\omega}}{\log \frac{1}{1.002\gamma}}. \tag{7.41}$$

If

$$\mathbf{u} \le \mathbf{u}_{\mathsf{ShiftedQR}}(n, k, \Sigma, B, \delta),$$

then the loop terminates in at most N_{dec} iterations, having produced a ω-decoupled matrix $\widehat{H'}$ at most ω-far from a unitary conjugate of H'.

Proof. 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 RITZ-OR-DEC 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 RITZ-OR-DEC 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 1.002γ. Since $\omega \le \psi_k(H') \le \|H'\| \le \Sigma$, through lines 3(a)i-3(a)iii can have been executed at most N_{dec} times so far, the result of each of which is application of an IQR step of degree k, meaning that we can think of H'' as being produced H' a *single* IQR step of degree kN_{dec}. Thus by Lemma 7.11, our inductive assumption on the prior shifts, and the hypothesis on $\mathbf{u}$, the distance from H'' to a unitary conjugate of H' is at most $4.5\|H\|kN_{\mathsf{dec}}\nu_{\mathsf{IQR}}(n)\mathbf{u} \le \omega$. If H'' is ω-decoupled, then the while loop terminates, and the proof is complete.

Otherwise H'' is not ω-decoupled. By the definition of ω and the fact that $\omega \le \delta/2n \le \Sigma/2n$, we can apply the triangle inequality and Lemmas 2.8 and 2.10 to find

$$\begin{aligned}
\|H''\| &\le \|H'\| + \omega \le (1 - 1/2n)\Sigma + \Sigma/2n \le \Sigma \\
\kappa_V(H'') &\le \kappa_V(H') + 8n^2\frac{\kappa_V^3(H')}{\mathrm{gap}(H')}\omega \le (1 - 1/2n)B + B/2n \le B \\
\mathrm{gap}(H'') &\ge \mathrm{gap}(H') - 2\kappa_V(H')\omega \ge (1 + 1/2n)\Gamma - \Gamma/2n \ge \Gamma,
\end{aligned}$$

and we furthermore have $2\omega^2/\Sigma \le 2\omega \le \Gamma \le \mathrm{gap}(H'')$ by the above and the definition of ω. This means RITZ-OR-DEC(H'', ω, φ) meets its requirements, and from our assumption on $\mathbf{u}$ we can apply Lemma 7.20 to conclude that it satisfies its guarantees. If this call to RITZ-OR-DEC outputs `dec = true`, then the matrix it outputs is indeed decoupled and the while loop terminates.

If on the other hand `dec = false`, then RITZ-OR-DEC outputs H'' and θ-approximate Ritz values $\mathcal{R}$ contained in in a disk of radius $1.1\|H''\|$, and RITZ-OR-DEC guarantees

$$\mathrm{dist}(\mathcal{R}, H'') \ge \frac{\omega^2\sqrt{\varphi}}{32 \cdot 101 \cdot \Sigma\sqrt{2k}}.$$

The bound on $\kappa_V(H'')$ in the previous paragraph ensures that the requirements of $\mathsf{SH}_{k,B}(H, \mathcal{R}, \omega, \varphi)$ have been met, and the parameter settings in (7.2)-(7.3) give us

$$\begin{aligned}
1.001\theta\alpha B^{1/k}\psi_k(H'') &= 1.001\frac{1.01}{0.998^{1/k}}(2B^4)^{1/2k}(1.01B)^{\frac{4\log k}{k}}B^{1/k}\psi_k(H'') \\
&= 1.04 \cdot 2^{1/2k}B^{\frac{4\log k+3}{k}}\psi_k(H'') \\
&\le 1.04 \cdot \sqrt{2^{\frac{2}{k-1}}B^{\frac{8\log k+11}{k-1}}}\|H''\| \\
&\le 1.04\sqrt{3}\|H''\| \\
&\le 1.9\|H''\|,
\end{aligned}$$

so every exceptional shift has modulus at most $3\|H''\|$ in the complex plane. Our assumption on $\mathbf{u}$ lets us invoke Lemma 7.24 to conclude that $\mathsf{SH}_{k,B}$ achieves potential reduction by a multiplicative factor of 1.002γ. Moreover, the shifts executed by RITZ-OR-DEC and SH in the above run through the while loop had modulus at most

$$3\|H''\| \le 3\|H'\|(1 + 4.5kN_{\mathsf{dec}}\nu_{\mathsf{IQR}}(n)\mathbf{u}) \le 3\|H'\| \cdot (1 + \omega/\Sigma) \le 4.5\|H'\|,$$

again since $\omega \leq \delta/2n \leq \Sigma$.

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 1.002γ, 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 $\widehat{H'}$ is at most ω-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 7.26 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 7.26, 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 2ω. Thus, finally using the full force of the $1/4n$ factor in the definition of ω,

$$\begin{aligned}\|H'\| &\leq \|H\| + 2(n-1)\omega \leq (1-1/2n)\Sigma \\ \kappa_V(H') &\leq \kappa_V(H) + 8n^2\frac{\kappa_V^3(H)}{\mathrm{gap}(H)} \cdot 2(n-1)\omega \leq (1-1/2n)B \\ \mathrm{gap}(H') &\geq \mathrm{gap}(H) - 2\kappa_V(H) \cdot 2(n-1)\omega \geq (1+1/2n)\Gamma\end{aligned}$$

by the definition of ω.

This ensures that *every* execution of the while loop throughout ShiftedQR satisfies the conclusion of Lemma 7.26, 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

$$\left\| \begin{pmatrix} L_1 & * & * \\ & L_2 & * \\ & & \ddots \end{pmatrix} - \widetilde{Q}^* H \widetilde{Q} \right\| \leq 2(n-1)\omega \leq \delta - \delta/n,$$

for some unitary $\widetilde{Q}$, as we had set out to show.

Failure Probability. We have already shown that RITZ-OR-DEC and $\mathsf{SH}_{k,B}$ satisfy their guarantees (including their failure probability) throughout ShiftedQR whenever the hypotheses of Theorem 7.25; 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 RITZ-OR-DEC 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

$$\begin{aligned} T_{\mathsf{ShiftedQR}}(n,k,\delta,B,\Sigma,\Gamma) &= T_3(n,k,\delta,B,\Sigma,\Gamma) \\ &\leq T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n,k,\mathtt{true}) \\ &\qquad + N_{\mathsf{dec}}\Big(T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n,k,\mathtt{false}) + T_{\mathsf{SH}}(n,k,\delta,B,\Sigma,\Gamma)\Big) \\ &\qquad + T_{\mathsf{DEFLATE}}(k) + \max_{\sum_i n_i = n} \sum_i T_3(n_i,k,\delta,B,\Sigma,\Gamma). \end{aligned}$$

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 little in the constant, we can bound as

$$\begin{aligned} T_{\mathsf{ShiftedQR}}(n,k,\delta,B,\Sigma,\gamma) &\leq n\Big(T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n,k,\mathtt{true}) \\ &\qquad + N_{\mathsf{dec}}\Big(T_{\mathsf{RITZ\text{-}OR\text{-}DEC}}(n,k,\mathtt{false}) + T_{\mathsf{SH}}(n,k,\delta,B,\Sigma,\Gamma)\Big) \\ &\qquad + T_{\mathsf{DEFLATE}}(k)\Big) \\ &= O\left(\left(\log\frac{nB\Sigma}{\delta\Gamma} k\log k + k^2\right) n^3\right). \end{aligned}$$

In addition, $\mathsf{ShiftedQR}$ requires at most $O(n\log\frac{nB\Sigma}{\delta\Gamma})$ calls to $\mathsf{SmallEig}$ with accuracy $\Omega(\omega^2/\Sigma)$ and failure probability tolerance φ in the course of the calls to $\mathsf{RITZ\text{-}OR\text{-}DEC}$, plus $O(n)$ 'base case' calls with accuracy δ/n and failure probability tolerance $\phi/3n$; the latter calls to $\mathsf{SmallEig}$ are asymptotically dominated by the former. The estimates in the theorem statement come from bounding ω and φ. □

Bibliographic Note

This chapter is lightly adapted from the forthcoming [18].

Bibliography

[1] K. Aishima, T. Matsuo, K. Murota, and M. Sugihara. A Wilkinson-like multishift QR algorithm for symmetric eigenvalue problems and its global convergence. *Journal of Computational and Applied Mathematics*, 236(15):3556–3560, 2012.

[2] M. Aizenman, R. Peled, J. Schenker, M. Shamis, and S. Sodin. Matrix regularizing effects of Gaussian perturbations. *Communications in Contemporary Mathematics*, 19(03):1750028, 2017.

[3] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010.

[4] D. Armentano, C. Beltrán, P. Bürgisser, F. Cucker, and M. Shub. A stable, polynomial-time algorithm for the eigenpair problem. *Journal of the European Mathematical Society*, 20(6):1375–1437, 2018.

[5] D. Armentano, C. Beltrán, P. Bürgisser, F. Cucker, and M. Shub. A stable, polynomial-time algorithm for the eigenpair problem. *Journal of the European Mathematical Society*, 20(6):1375–1437, 2018.

[6] D. Armentano and F. Cucker. A randomized homotopy for the Hermitian eigenpair problem. *Foundations of Computational Mathematics*, 15(1):281–312, 2015.

[7] G. B. Arous and P. Bourgade. Extreme gaps between eigenvalues of random matrices. *The Annals of Probability*, 41(4):2648–2681, 2013.

[8] G. Aubrun and S. J. Szarek. *Alice and Bob meet Banach*, volume 223. American Mathematical Soc., 2017.

[9] Z. Bai and J. Demmel. On a block implementation of Hessenberg multishift QR iteration. *International Journal of High Speed Computing*, 1(01):97–112, 1989.

[10] Z. Bai and J. Demmel. Using the matrix sign function to compute invariant subspaces. *SIAM Journal on Matrix Analysis and Applications*, 19(1):205–225, 1998.

[11] Z. Bai, J. Demmel, and M. Gu. An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems. *Numerische Mathematik*, 76(3):279–308, 1997.

[12] Z. D. Bai, J. W. Silverstein, and Y. Q. Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26(2):166–168, 1988.

[13] G. Ballard, J. Demmel, and I. Dumitriu. Minimizing communication for eigenproblems and the singular value decomposition. *arXiv preprint arXiv:1011.3077*, 2010.

[14] G. Ballard, J. Demmel, I. Dumitriu, and A. Rusciano. A generalized randomized rank-revealing factorization. *arXiv preprint arXiv:1909.06524*, 2019.

[15] J. Banks, J. Garza Vargas, A. Kulkarni, and N. Srivastava. Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time. *arXiv preprint arXiv:1912.08805, to appear in Comm. Pure Appl. Math*, 2019.

[16] J. Banks, J. Garza-Vargas, A. Kulkarni, and N. Srivastava. Overlaps, eigenvalue gaps, and pseudospectrum under real ginibre and absolutely continuous perturbations. *arXiv preprint arXiv:2005.08930*, 2020.

[17] J. Banks, J. Garza-Vargas, and N. Srivastava. Global convergence of Hessenberg shifted QR I: Dynamics. *arXiv preprint arXiv:2111.07976*, 2021.

[18] J. Banks, J. Garza-Vargas, and N. Srivastava. Global linear convergence of Hessenberg shifted QR II: Numerical stability. *in preparation*, 2021.

[19] J. Banks, A. Kulkarni, S. Mukherjee, and N. Srivastava. Gaussian regularization of the pseudospectrum and Davies' conjecture. *Communications on Pure and Applied Mathematics*, 74(10):2114–2131, 2021.

[20] A. Basak, E. Paquette, and O. Zeitouni. Spectrum of random perturbations of Toeplitz matrices with finite symbols. *arXiv preprint arXiv:1812.06207*, 2018.

[21] A. Basak, E. Paquette, and O. Zeitouni. Regularization of non-normal matrices by Gaussian noise - the banded Toeplitz and twisted Toeplitz cases. In *Forum of Mathematics, Sigma*, volume 7. Cambridge University Press, 2019.

[22] S. Batterson. Convergence of the shifted QR algorithm on 3× 3 normal matrices. *Numerische Mathematik*, 58(1):341–352, 1990.

[23] S. Batterson. Convergence of the Francis shifted QR algorithm on normal matrices. *Linear algebra and its applications*, 207:181–195, 1994.

[24] S. Batterson. Dynamical analysis of numerical systems. *Numerical linear algebra with applications*, 2(3):297–310, 1995.

[25] S. Batterson and D. Day. Linear convergence in the shifted QR algorithm. *mathematics of computation*, 59(199):141–151, 1992.

[26] S. Batterson and J. Smillie. The dynamics of Rayleigh quotient iteration. *SIAM Journal on Numerical Analysis*, 26(3):624–636, 1989.

[27] S. Batterson and J. Smillie. Rayleigh quotient iteration for nonsymmetric matrices. *mathematics of computation*, 55(191):169–178, 1990.

[28] F. L. Bauer and C. T. Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960.

[29] A. N. Beavers and E. D. Denman. A computational method for eigenvalues and eigenvectors of a matrix with real eigenvalues. *Numerische Mathematik*, 21(5):389–396, 1973.

[30] A. N. Beavers Jr. and E. D. Denman. A new similarity transformation method for eigenvalues and eigenvectors. *Mathematical Biosciences*, 21(1-2):143–169, 1974.

[31] M. Ben-Or and L. 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.

[32] F. Benaych-Georges and O. Zeitouni. Eigenvectors of non normal random matrices. *Electronic Communications in Probability*, 23, 2018.

[33] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[34] D. Bindel, S. Chandresekaran, J. Demmel, D. Garmire, and M. Gu. A fast and stable non-symmetric eigensolver for certain structured matrices. Technical report, Technical report, University of California, Berkeley, CA, 2005.

[35] C. Bordenave, D. Chafaï, et al. Around the circular law. *Probability surveys*, 9:1–89, 2012.

[36] P. Bourgade and G. Dubach. The distribution of overlaps between eigenvectors of Ginibre matrices. *arXiv preprint arXiv:1801.01219*, 2018.

[37] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. Part I: Maintaining well-focused shifts and level 3 performance. *SIAM Journal on Matrix Analysis and Applications*, 23(4):929–947, 2002.

[38] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. Part II: Aggressive early deflation. *SIAM Journal on Matrix Analysis and Applications*, 23(4):948–973, 2002.

[39] M.-F. Bru. Diffusions of perturbed principal component analysis. *Journal of multivariate analysis*, 29(1):127–136, 1989.

[40] M.-F. Bru. Wishart processes. *Journal of Theoretical Probability*, 4(4):725–751, 1991.

[41] H. J. Buurema. *A geometric proof of convergence for the QR method*, volume 62. 1958.

[42] R. Byers. Numerical stability and instability in matrix sign function based algorithms. In *Computational and Combinatorial Methods in Systems Theory*. Citeseer, 1986.

[43] R. Byers, C. He, and V. Mehrmann. The matrix sign function method and the computation of invariant subspaces. *SIAM Journal on Matrix Analysis and Applications*, 18(3):615–632, 1997.

[44] R. Byers and H. Xu. A new scaling for Newton's iteration for the polar decomposition and its backward stability. *SIAM Journal on Matrix Analysis and Applications*, 30(2):822–843, 2008.

[45] J.-y. Cai. Computing Jordan normal forms exactly for commuting matrices in polynomial time. *International Journal of Foundations of Computer Science*, 5(03n04):293–302, 1994.

[46] J. T. Chalker and B. Mehlig. Eigenvector statistics in non-Hermitian random matrix ensembles. *Physical review letters*, 81(16):3367, 1998.

[47] M. T. Chu. Linear algebra algorithms as dynamical systems. *Acta Numerica*, 17:1–86, 2008.

[48] G. Cipolloni, L. Erdős, and D. Schröder. Optimal lower bound on the least singular value of the shifted ginibre ensemble. *arXiv preprint arXiv:1908.01653*, 2019.

[49] G. Cipolloni, L. Erdős, and D. Schröder. Fluctuation around the circular law for random matrices with real entries. *arXiv preprint arXiv:2002.02438*, 2020.

[50] J.-M. Combes, F. Germinet, and A. Klein. Generalized eigenvalue-counting estimates for the anderson model. *Journal of Statistical Physics*, 135(2):201, 2009.

[51] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert w function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

[52] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[53] E. B. Davies. Approximate diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1051–1064, 2007.

[54] E. B. Davies. Approximate diagonalization. *SIAM journal on matrix analysis and applications*, 29(4):1051–1064, 2008.

[55] E. B. Davies and M. Hager. Perturbations of Jordan matrices. *Journal of Approximation Theory*, 156(1):82–94, 2009.

[56] D. Day. How the QR algorithm fails to converge and how to fix it. 1996.

[57] P. Deift, T. Nanda, and C. Tomei. Ordinary differential equations and the symmetric eigenvalue problem. *SIAM Journal on Numerical Analysis*, 20(1):1–22, 1983.

[58] P. Deift and T. Trogdon. Universality in numerical computation with random data: Case studies and analytical results. *Journal of Mathematical Physics*, 60(10):103306, 2019.

[59] P. A. Deift, G. Menon, S. Olver, and T. Trogdon. Universality in numerical computations with random data. *Proceedings of the National Academy of Sciences*, 111(42):14973–14978, 2014.

[60] T. Dekker and J. Traub. The shifted QR algorithm for Hermitian matrices. *Linear Algebra Appl*, 4:137–154, 1971.

[61] J. Demmel, I. Dumitriu, and O. Holtz. Fast linear algebra is stable. *Numerische Mathematik*, 108(1):59–91, 2007.

[62] J. Demmel, I. Dumitriu, O. Holtz, and R. Kleinberg. Fast matrix multiplication is stable. *Numerische Mathematik*, 106(2):199–224, 2007.

[63] J. W. Demmel. The probability that a numerical analysis problem is difficult. *Mathematics of Computation*, 50(182):449–480, 1988.

[64] J. W. Demmel. *Applied numerical linear algebra*, volume 56. SIAM, 1997.

[65] E. D. Denman and A. N. Beavers Jr. The matrix sign function and computations in systems. *Applied mathematics and Computation*, 2(1):63–94, 1976.

[66] X. Ding and R. Wu. A new proof for comparison theorems for stochastic differential inequalities with respect to semimartingales. *Stochastic Processes and their applications*, 78(2):155–171, 1998.

[67] J. Dongarra and F. Sullivan. Guest editors' introduction: The top 10 algorithms. *Computing in Science & Engineering*, 2(1):22, 2000.

[68] I. Dumitriu. Smallest eigenvalue distributions for two classes of β-Jacobi ensembles. *Journal of Mathematical Physics*, 53(10):103301, 2012.

[69] P. Eberlein and C. Huang. Global convergence of the QR algorithm for unitary matrices with some results for normal matrices. *SIAM Journal on Numerical Analysis*, 12(1):97–104, 1975.

[70] A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

[71] A. Edelman, E. Kostlan, and M. Shub. How many eigenvalues of a random matrix are real? *Journal of the American Mathematical Society*, 7(1):247–267, 1994.

[72] A. Edelman and N. R. Rao. Random matrix theory. *Acta Numerica*, 14:233–297, 2005.

[73] A. Edelman and B. D. Sutton. The β-Jacobi matrix model, the CS decomposition, and generalized singular value problems. *Foundations of Computational Mathematics*, 8(2):259–285, 2008.

[74] L. Erdős, B. Schlein, and H.-T. Yau. Wegner estimate and level repulsion for wigner random matrices. *International Mathematics Research Notices*, 2010(3):436–479, 2010.

[75] J. Erxiong. A note on the double-shift QL algorithm. *Linear algebra and its applications*, 171:121–132, 1992.

[76] O. N. Feldheim, E. Paquette, and O. Zeitouni. Regularization of non-normal matrices by Gaussian noise. *International Mathematics Research Notices*, 2015(18):8724–8751, 2014.

[77] P. J. Forrester. *Log-gases and random matrices (LMS-34).* Princeton University Press, 2010.

[78] J. G. Francis. The QR transformation a unitary analogue to the LR transformation—Part 1. *The Computer Journal*, 4(3):265–271, 1961.

[79] J. G. Francis. The QR transformation—Part 2. *The Computer Journal*, 4(4):332–345, 1962.

[80] Y. V. Fyodorov. On statistics of bi-orthogonal eigenvectors in real and complex Ginibre ensembles: combining partial Schur decomposition with supersymmetry. *Communications in Mathematical Physics*, 363(2):579–603, 2018.

[81] S. Ge. *The Eigenvalue Spacing of IID Random Matrices and Related Least Singular Value Results.* PhD thesis, UCLA, 2017.

[82] C. Geiß and R. Manthey. Comparison theorems for stochastic differential equations in finite and infinite dimensions. *Stochastic processes and their applications*, 53(1):23–35, 1994.

[83] E. Gluskin and A. Olevskii. Invertibility of sub-matrices and the octahedron width theorem. *Israel Journal of Mathematics*, 186(1):61–68, 2011.

[84] G. Golub and F. Uhlig. The QR algorithm: 50 years later its genesis by John Francis and Vera Kublanovskaya and subsequent developments. *IMA Journal of Numerical Analysis*, 29(3):467–485, 2009.

[85] G. H. Golub and C. F. Van Loan. Matrix computations. Johns Hopkins studies in the mathematical sciences, 1996.

[86] N. Goodman. The distribution of the determinant of a complex wishart distributed matrix. *The Annals of mathematical statistics*, 34(1):178–180, 1963.

[87] P. Graczyk and J. Małecki. Strong solutions of non-colliding particle systems. *Electronic Journal of Probability*, 19, 2014.

[88] A. Greenbaum, R.-c. Li, and M. L. Overton. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM Review*, 62(2):463–482, 2020.

[89] M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[90] A. Guionnet, M. Krishnapur, and O. Zeitouni. The single ring theorem. *Annals of mathematics*, pages 1189–1217, 2011.

[91] A. Guionnet, P. Wood, and O. Zeitouni. Convergence of the spectral measure of non-normal matrices. *Proceedings of the American Mathematical Society*, 142(2):667–679, 2014.

[92] U. Haagerup and F. Larsen. Brown's spectral distribution measure for R-diagonal elements in finite von Neumann algebras. *Journal of Functional Analysis*, 176(2):331–367, 2000.

[93] N. J. Higham. The matrix sign decomposition and its relation to the polar decomposition. *Linear Algebra and its Applications*, 212:3–20, 1994.

[94] N. J. Higham. *Accuracy and stability of numerical algorithms*, volume 80. SIAM, 2002.

[95] N. J. Higham. *Functions of matrices: theory and computation*, volume 104. SIAM, 2008.

[96] W. Hoffmann and B. N. Parlett. A new proof of global convergence for the tridiagonal QL algorithm. *SIAM Journal on Numerical Analysis*, 15(5):929–937, 1978.

[97] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* Cambridge university press, 1994.

[98] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge university press, 2012.

[99] V. Jain, A. Sah, and M. Sawhney. On the real davies' conjecture. *arxiv preprint arXiv:2005.08908*, 2020.

[100] C. S. Kenney and A. J. Laub. The matrix sign function. *IEEE Transactions on Automatic Control*, 40(8):1330–1348, 1995.

[101] W. König, N. O'Connell, et al. Eigenvalues of the Laguerre process as non-colliding squared Bessel processes. *Electronic Communications in Probability*, 6:107–114, 2001.

[102] V. Krasin. *Comparison theorem and its applications to finance.* PhD thesis, University of Alberta, Edmonton, Alberta, 2010.

[103] D. Kressner. On the use of larger bulges in the QR algorithm. *Electronic Transactions on Numerical Analysis*, 20(ARTICLE):50–63, 2005.

[104] D. Kressner. The effect of aggressive early deflation on the convergence of the QR algorithm. *SIAM journal on matrix analysis and applications*, 30(2):805–821, 2008.

[105] D. Kressner. *Personal communication*, 2021.

[106] V. N. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *USSR Computational Mathematics and Mathematical Physics*, 1(3):637–657, 1962.

[107] H. Le. Brownian motions on shape and size-and-shape spaces. *Journal of applied probability*, 31(1):101–113, 1994.

[108] H. Le. Singular-values of matrix-valued Ornstein–Uhlenbeck processes. *Stochastic processes and their applications*, 82(1):53–60, 1999.

[109] R. S. Leite, N. C. Saldanha, and C. Tomei. Dynamics of the symmetric eigenvalue problem with shift strategies. *International Mathematics Research Notices*, 2013(19):4382–4412, 2013.

[110] G. Livshyts, G. Paouris, and P. Pivovarov. On sharp bounds for marginal densities of product measures. *Israel Journal of Mathematics*, 216(2):877–889, 2016.

[111] A. Louis and S. S. Vempala. Accelerated newton iteration for roots of black box polynomials. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 732–740. IEEE, 2016.

[112] K. Luh and S. O'Rourke. Eigenvectors and controllability of non-hermitian random matrices and directed graphs, 2020.

[113] A. N. Malyshev. Parallel algorithm for solving some spectral problems of linear algebra. *Linear algebra and its applications*, 188:489–520, 1993.

[114] F. Mezzadri. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.

[115] N. Minami. Local fluctuation of the spectrum of a multidimensional anderson tight binding model. *Communications in mathematical physics*, 177(3):709–725, 1996.

[116] D. S. Mitrinovic, J. Pecaric, and A. M. Fink. *Inequalities involving functions and their integrals and derivatives*, volume 53. Springer Science & Business Media, 1991.

[117] C. Moler. Variants of the QR algorithm.

[118] Y. Nakatsukasa and R. W. Freund. Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarev's functions. *SIAM Review*, 58(3):461–493, 2016.

[119] Y. Nakatsukasa and N. J. Higham. Backward stability of iterations for computing the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 33(2):460–479, 2012.

[120] A. Naor and P. Youssef. Restricted invertibility revisited. In *A journey through discrete mathematics*, pages 657–691. Springer, 2017.

[121] H. Nguyen, T. Tao, and V. Vu. Random matrices: tail bounds for gaps between eigenvalues. *Probability Theory and Related Fields*, 167(3-4):777–816, 2017.

[122] H. H. Nguyen. Random matrices: Overcrowding estimates for the spectrum. *Journal of functional analysis*, 275(8):2197–2224, 2018.

[123] A. M. Ostrowski. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. i-vi. *Archive for Rational Mechanics and Analysis*, 1(1):233–241, 1957.

[124] V. Y. Pan and Z. Q. Chen. The complexity of the matrix eigenproblem. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 507–516. ACM, 1999.

[125] B. Parlett. Singular and invariant matrices under the QR transformation. *Mathematics of Computation*, 20(96):611–615, 1966.

[126] B. N. Parlett. The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Mathematics of Computation*, 28(127):679–693, 1974.

[127] B. N. Parlett. *The symmetric eigenvalue problem*, volume 20. SIAM, 1998.

[128] B. N. Parlett. The QR algorithm. *Computing in Science & Engineering*, 2(1):38–42, 2000.

[129] B. N. Parlett and W. Kahan. On the convergence of a practical QR algorithm. In *IFIP Congress (1)*, pages 114–118, 1968.

[130] B. N. Parlett and J. Le. Forward instability of tridiagonal QR. *SIAM Journal on Matrix Analysis and Applications*, 14(1):279–316, 1993.

[131] C. W. Pfrang, P. Deift, and G. Menon. How long does it take to compute the eigenvalues of a random symmetric matrix. *Random Matrix Theory, Interacting Particle Systems, and Integrable Systems, Math. Sci. Res. Inst. Publ*, 65:411–442, 2013.

[132] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 1999.

[133] J. D. Roberts. Linear model reduction and solution of the algebraic Riccati equation by use of the sign function. *International Journal of Control*, 32(4):677–687, 1980.

[134] M. Rudelson and R. Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[135] M. Rudelson and R. Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 2015(19):9594–9617, 2015.

[136] A. Sankar, D. A. Spielman, and S.-H. Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(2):446–476, 2006.

[137] D. Shi and Y. Jiang. Smallest gaps between eigenvalues of random matrices with complex Ginibre, Wishart and universal unitary ensembles. *arXiv preprint arXiv:1207.4240*, 2012.

[138] J. Sjoestrand and M. Vogel. General Toeplitz matrices subject to Gaussian perturbations. *arXiv preprint arXiv:1905.10265*, 2019.

[139] J. Sjoestrand and M. Vogel. Toeplitz band matrices with small random perturbations. *arXiv preprint arXiv:1901.08982*, 2019.

[140] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (New Series) of The American Mathematical Society*, 13(2):87–121, 1985.

[141] S. Smale. Complexity theory and numerical analysis. *Acta Numerica*, 6:523–551, 1997.

[142] P. Śniady. Random regularization of Brown spectral measure. *Journal of Functional Analysis*, 193(2):291–313, 2002.

[143] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[144] G. W. Stewart. A Krylov–Schur algorithm for large eigenproblems. *SIAM Journal on Matrix Analysis and Applications*, 23(3):601–614, 2002.

[145] J.-G. Sun. Perturbation bounds for the Cholesky and QR factorizations. *BIT Numerical Mathematics*, 31(2):341–352, 1991.

[146] S. J. Szarek. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.

[147] T. Tao and V. Vu. Random matrices: The distribution of the smallest singular values. *Geometric And Functional Analysis*, 20(1):260–297, 2010.

[148] T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of ESDs and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010.

[149] R. C. Thompson. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra and its Applications*, 13(1-2):69–78, 1976.

[150] K. Tikhomirov. Invertibility via distance for non-centered random matrices with continuous distributions. *arXiv preprint arXiv:1707.09656*, 2017.

[151] F. Tisseur. Backward stability of the QR algorithm. Technical report, Technical Report 239, Equipe d'Analyse Numerique, Universit e Jean Monnet de ..., 1996.

[152] L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. SIAM, 1997.

[153] L. N. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.

[154] J. P. Vinson. Closest spacing of eigenvalues. *arXiv preprint arXiv:1111.2743*, 2011.

[155] J. Von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, 53(11):1021–1099, 1947.

[156] T.-L. Wang. Convergence of the tridiagonal QR algorithm. *Linear algebra and its applications*, 322(1-3):1–17, 2001.

[157] T.-L. Wang and W. Gragg. Convergence of the shifted QR algorithm for unitary Hessenberg matrices. *Mathematics of computation*, 71(240):1473–1496, 2002.

[158] T.-L. Wang and W. Gragg. Convergence of the unitary QR algorithm with a unimodular Wilkinson shift. *Mathematics of computation*, 72(241):375–385, 2003.

[159] D. S. Watkins. Forward stability and transmission of shifts in the QR algorithm. *SIAM Journal on Matrix Analysis and Applications*, 16(2):469–487, 1995.

[160] D. S. Watkins. The transmission of shifts and shift blurring in the QR algorithm. *Linear algebra and its applications*, 241:877–896, 1996.

[161] D. S. Watkins. *The matrix eigenvalue problem: GR and Krylov subspace methods*. SIAM, 2007.

[162] D. S. Watkins. The QR algorithm revisited. *SIAM review*, 50(1):133–145, 2008.

[163] F. Wegner. Bounds on the density of states in disordered systems. *Zeitschrift für Physik B Condensed Matter*, 44(1):9–15, 1981.

[164] J. H. Wilkinson. Global convergence of tridiagonal QR algorithm with origin shifts. *Linear Algebra and its Applications*, 1(3):409–420, 1968.

[165] T. G. Wright and L. N. Trefethen. Eigtool. *Software available at http://www.comlab.ox.ac.uk/pseudospectra/eigtool*, 2002.

[166] Y.-Q. Yin, Z.-D. Bai, and P. R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability theory and related fields*, 78(4):509–521, 1988.