UNIVERSITY OF CALIFORNIA SAN DIEGO

Pseudospectral Divide-and-Conquer for the Generalized Eigenvalue Problem

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics

by

Ryan Schneider

Committee in charge:

Professor Ioana Dumitriu, Chair
Professor Alex Cloninger
Professor Todd Kemp
Professor Piya Pal
Professor Rayan Saab

2024

The Dissertation of Ryan Schneider is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor, Ioana Dumitriu, for her endless support and compassion. Ioana, thank you for trusting me with this work, for the countless meetings (and cups of coffee), and most importantly for reminding me why I wanted to complete a PhD. You always told me to “put the hammer away” when writing, and luckily this requires no qualifiers: it has been an incredible privilege completing this with you and inheriting your perspective on applied math.

There are two other mentors I owe the same debt of gratitude – James Demmel and Barry Schneider. Jim, thank you for suggesting this work to us and for serving as our numerical linear algebra oracle along the way; I so look forward to continuing what we’ve started at Berkeley! Barry, thank you for the wonderful times at NIST and for sharing your quantum chemistry wisdom (and libraries). Thank you also to my first academic mentor John Shareshian for getting me to graduate school in the first place.

I would like to acknowledge the many graduate students, postdocs, and faculty members – including those on my committee – who provided generous feedback on this work. Thanks especially to Anne Greenbaum, Jorge Garza-Vargas, and Yuji Nakatsukasa, whose input strongly shaped Chapters 3 and 5 of this thesis.

To my family: thank you for your unwavering support. Mom and Dad, I hope this thesis is a testament to everything you’ve given me. To Lauren, Devante, Ben, Moises, Andrés, Kate, and many more: thank you for being there and for sharing in the joy (and stress) of this process. And finally to Kevin: thank you for dropping everything when I needed you and for being both a brother and a best friend. I could not have done this without you.

Chapters 2 and 4 in full and Chapters 1, 3, and 6 in part comprise a work that has been submitted for publication and was co-authored with James Demmel and Ioana Dumitriu:

- J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.

Chapters 3 and 6 also contain content from another submitted work:

- R. Schneider. When is fast, implicit squaring of $A^{-1}B$ stable? arXiv:2310.00193, 2023.

In both papers, the dissertation author was the primary investigator and author.

VITA

2018	Bachelor of Arts in Mathematics, Washington University in St. Louis
2018–2024	Teaching Assistant, University of California San Diego
2022–2024	Measurement Science and Engineering Fellow, National Institute of Standards and Technology
2024	Doctor of Philosophy in Mathematics, University of California San Diego

PUBLICATIONS

R. Schneider. When is fast, implicit squaring of $A^{-1}B$ stable? arXiv:2310.00193, 2023.

J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.

R. Schneider, H. Gharibnejad, and B. I. Schneider. ITVOLT: An Iterative Solver for the Time-Dependent Schrödinger Equation. *Computer Physics Communications* 291, 2023.

ABSTRACT OF THE DISSERTATION

Pseudospectral Divide-and-Conquer for the Generalized Eigenvalue Problem

by

Ryan Schneider

Doctor of Philosophy in Mathematics

University of California San Diego, 2024

Professor Ioana Dumitriu, Chair

Over the last two decades, randomization has emerged as a leading tool for pursuing efficiency in numerical linear algebra. Its benefits can be explained in part by smoothed analysis, where an algorithm that fails spectacularly in certain cases may be unlikely to do so on random – or randomly perturbed – inputs. This observation implies a simple framework for developing accurate and efficient randomized algorithms: apply a random perturbation and run an existing method, whose worst-case error (or run-time) can be avoided with high probability.

Recent work of Banks, Garza-Vargas, Kulkarni, and Srivastava applied this framework to the standard eigenvalue problem, developing a randomized algorithm that (with high probability) diagonalizes a matrix in nearly matrix multiplication time [Foundations of Computational Math 2022]. Central to their work is the phenomenon of *pseudospectral shattering*, in which a small Gaussian perturbation regularizes the pseudospectrum of a matrix, with high probability breaking it into disjoint components and allowing classical, divide-and-conquer eigensolvers to run successfully. Prior to their work, no way of accessing the benefits of divide-and-conquer's natural parallelization was known in general.

In this thesis, we extend the work of Banks et al. to the generalized eigenvalue problem – e.g., $Av = \lambda Bv$ for matrices $A, B \in \mathbb{C}^{n \times n}$. Our main contributions can be summarized as follows.

1. First, we show that pseudospectral shattering generalizes directly: randomly perturbing A and B has a similar regularizing effect on the pseudospectra of the corresponding matrix pencil (A, B) with high probability.
2. Building on pseudospectral shattering, we construct and analyze a fast, randomized, divide-and-conquer algorithm for diagonalizing (A, B), which begins by randomly perturbing the inputs.
3. Finally, we demonstrate that both pseudospectral shattering and the corresponding diagonalization algorithm can be adapted to definite pencils, further pursuing efficiency by preserving and exploiting symmetry.

The resulting algorithm, which we call pseudospectral divide-and-conquer, is the first general, sub-$O(n^3)$ solver for the generalized eigenvalue problem. It is not only highly parallel and capable of accommodating structure, but also promotes stability by avoiding matrix inversion. In essence, this thesis is a handbook for understanding, adapting, and implementing the method.

Chapter 1

Introduction and Background

For more than a decade, randomization has revolutionized the building blocks of numerical linear algebra. This effort, cumulatively referred to as Randomized Numerical Linear Algebra or RandNLA, has a remarkably simple ethos: the benefits of randomization – i.e., regularization, dimension reduction, etc. – can be leveraged to develop new algorithms or revitalize existing ones, opening a pathway to fast methods with general accuracy guarantees. The result is a growing collection of cutting-edge randomized algorithms, which (probabilistically) achieve optimal or near-optimal performance on a variety of problems in linear algebra, including trace estimation [52, 80, 99], matrix factoring/approximation [12, 16, 46, 112, 135], least squares [45, 110], linear systems [62, 69, 87, 128], and more.[1] The broad applicability of this work has prompted efforts to develop the first standardized libraries for randomized algorithms (e.g., RandBLAS and RandLAPACK [103]).

In this thesis, we apply the RandNLA framework to the generalized eigenvalue problem, which seeks the (generalized) eigenvalues and eigenvectors of a matrix pencil (A, B). Though somewhat less-well-known than its single-matrix counterpart, the generalized eigenvalue problem is ubiquitous in scientific computing, arising in signal processing [79, 111], binary classification problems [67, 97], linear differential equations [70], quantum chemistry [55, 59, 96], and more. In a world of big data, these applications place increasing pressure on computational resources, necessitating efficient eigensolvers that can handle large inputs.

Since its introduction more than 50 years ago, the QZ algorithm of Moler and Stewart [102] has remained the standard method for solving the generalized eigenvalue problem. Though well-studied, QZ requires $O(n^3)$ arithmetic operations to find the eigenvalues and eigenvectors of an $n \times n$ pencil. Since the theoretical bottleneck for solving the generalized eigenvalue problem is matrix multiplication, which offers a variety of sub-$O(n^3)$ implementations, the door remains open: can an algorithm find the eigenvalues and eigenvectors of an arbitrary pencil in fewer than $O(n^3)$ operations? If such an algorithm exists, can it be implemented stably and in parallel?

Analogous questions for the standard eigenvalue problem were recently resolved in work of Banks, Garza-Vargas, Kulkarni, and Srivastava [16], which demonstrated that a randomized, divide-and-conquer algorithm could beat $O(n^3)$ complexity to find the eigenvalues/eigenvectors of an individual matrix. The algorithm they exhibit runs in nearly matrix multiplication time – i.e., complexity equal to that of matrix multiplication up to logarithmic factors. The key insight of their work is the regularizing effect of random perturbations on the spectrum and pseudospectrum of a matrix. With high probability,

[1] For an exhaustive summary of this work, see the survey of Martinsson and Tropp [98].

regularization guarantees success for classical divide-and-conquer algorithms on perturbed inputs. This opens a pathway to an approximate diagonalization of any matrix: simply apply a random perturbation and run a standard formulation of divide-and-conquer, which now succeeds with high probability and offers improved efficiency by way of natural parallelization. In this approach, the accuracy of the resulting diagonalization is determined by the size of the initial perturbation.

Here, we extend this work to the generalized eigenvalue problem. Applying the same high-level strategy – i.e., randomly perturbing the input matrices and running divide-and-conquer – we obtain a randomized algorithm that (1) with high probability produces a backward diagonalization of any matrix pencil, (2) runs in nearly matrix multiplication time, (3) avoids matrix inversion and (4) is highly parallel and communication-conscious. The result is the first known algorithm that can solve the generalized eigenvalue problem on arbitrary inputs in sub-$O(n^3)$ operations. With an eye toward high-performance implementation, our work demonstrates that this near-optimal complexity can be achieved without requiring inversion or sacrificing communication optimality.

In this chapter, we summarize our primary contributions, provide important context for the main results, and trace the theory our work is built on. We also discuss the content presented in the subsequent chapters at a high level.

Notation Considerations: We make the following notation choices here and in the subsequent chapters:

1. $\mathbb{C}^{n \times n}$ represents the vector space of $n \times n$ complex matrices, which are always denoted with capital Roman or Greek letters – i.e., $A, \Lambda \in \mathbb{C}^{n \times n}$.
2. The superscript $\sim$ is used to denote (randomly) perturbed matrices until Chapter 6, where it is used to denote floating-point quantities.
3. The identity matrix is denoted I, with size implied by context.
4. $\mathbb{1}$ represents a vector of ones (with arbitrary size) while $\mathbb{1}_S(z)$ denotes the indicator

function on $\mathbb{C}$ corresponding to the subset $S \subset \mathbb{C}$.

5. A^H and A^{-H} denote the Hermitian transpose and inverse Hermitian transpose of A.
6. The singular values of $A \in \mathbb{C}^{m\times n}$ are denoted $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_k(A)$ for $k = \min\{m, n\}$. When convenient, $\sigma_{\min}(A)$ may be used to refer to the smallest singular value of A.
7. $||\cdot||_2$ denotes the Euclidean norm on vectors and the spectral norm on matrices. $||\cdot||_F$ denotes the Frobenius norm.
8. $\kappa_2(A)$ is the spectral norm condition number of A.
9. $B_r(z)$ is the ball of radius r centered at z.
10. $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ denote the real and imaginary parts of $z \in \mathbb{C}$, respectively.
11. $\mathrm{poly}(\alpha, \beta)$ denotes an arbitrary polynomial in the quantities α and β. $\mathrm{polylog}(\alpha)$ is similarly used to represent a polynomial in $\log(\alpha)$.
12. Standard big-O and big-Omega notation is used to denote asymptotic upper and lower bounds. That is, $f(n) = O(g(n))$ and $f(n) = \Omega(h(n))$ if there exist constants $C_1, C_2 > 0$ such that $C_1 h(n) \leq f(n) \leq C_2 g(n)$ for all n sufficiently large (here f, g, and h are assumed to be positive functions of $n \in \mathbb{Z}_+$).

Guide to Chapter 1: Section 1.1 introduces the generalized eigenvalue problem alongside the necessary background information from linear algebra. Perturbation theory for the problem is subsequently presented in Section 1.2. Section 1.3 discusses divide-and-conquer eigensolvers, identifying the primary algorithmic challenges and exploring randomization as a means of addressing them. To place divide-and-conquer in the necessary context, relevant notions of efficiency and numerical stability are defined in Sections 1.4 and 1.5. Section 1.6 then presents the main results of the thesis and discusses related open problems. Finally, Section 1.7 collects a handful of results that don't fit neatly anywhere else. Throughout, we place key ideas/questions/results in boxes for easy reference.

1.1 The Generalized Eigenvalue Problem

We focus in this thesis on square matrix pencils, which we write throughout as $(A, B) \in \mathbb{C}^{n \times n} \times \mathbb{C}^{n \times n}$. With the exception of Chapter 5, the matrices A and B are arbitrary, and in particular may be dense and non-symmetric. While non-square pencils are of interest in certain applications, they will not be considered here.

In this section, we define the eigenvalues and eigenvectors of (A, B) and discuss the primary computational problem of interest – matrix pencil diagonalization. Much of the following presents standard background information from linear algebra; for more details, see the references [77, 126].

We begin by defining regularity for matrix pencils.

Definition 1.1.1. The matrix pencil (A, B) is *regular* if its characteristic polynomial $\det(A - xB) \in \mathbb{C}[x]$ is not identically zero. If (A, B) is not regular, it is *singular*.

When (A, B) is regular its eigenvalues and eigenvectors can be defined via a straightforward generalization of the single-matrix eigenvalue problem.

Definition 1.1.2. $\lambda \in \mathbb{C}$ is a *finite eigenvalue* of the regular matrix pencil (A, B) if $Av = \lambda Bv$ for some nonzero $v \in \mathbb{C}^n$, which is a corresponding *right eigenvector*. Similarly, nonzero $w \in \mathbb{C}^n$ is a *left eigenvector* if $w^H A = \lambda w^H B$. In either case, λ is a root of the characteristic polynomial $\det(A - xB)$.

We note from this definition an important distinction between the generalized and standard eigenvalue problems. The leading coefficient of $\det(A - xB)$ is $\det(B)$, meaning the characteristic polynomial of (A, B) may have degree less than n even when the pencil is regular. Hence, a regular pencil may not have a full set of finite eigenvalues. When this occurs, we say that (A, B) has an eigenvalue at infinity, whose corresponding right/left eigenvectors belong to the right/left null spaces of B. Including eigenvalues at infinity guarantees that any regular pencil has a full set of n eigenvalues, counting multiplicity.

Note that when $B = I$, the pencil (A, B) is clearly regular and Definition 1.1.2 reduces to the standard eigenvalue problem.

In the singular case, both A and B are singular themselves[2] and Definition 1.1.2 cannot be used to define eigenvalues and eigenvectors. The potential pitfalls of doing so jump out immediately: if A and B have overlapping null spaces, for example, this definition would imply that every complex number is an eigenvalue of (A, B), possibly all corresponding to the same eigenvector. Intuitively, we need definitions that exclude such spurious "eigenvalues" and "eigenvectors," which appear to satisfy Definition 1.1.2 but are actually expressing the singularity of the pencil. For eigenvalues, this results in the following.

Definition 1.1.3. $\lambda \in \mathbb{C}$ is a *finite eigenvalue* of the singular pencil (A, B) if $\operatorname{rank}(A - \lambda B) < \operatorname{rank}(A - xB)$, where the latter is computed over the field of fractions of $\mathbb{C}[x]$.

To define eigenvectors we follow Dopico and Noferini [44], who recently developed a rigorous and abstract theory of singular matrix pencils (and even higher degree matrix polynomials). We re-state their definition below for completeness; as we will see, the bulk of our analysis rests on the regular case.

Definition 1.1.4. $v \in \mathbb{C}^n$ is a (right) *eigenvector* of the singular pencil (A, B) corresponding to finite eigenvalue λ if $Av = \lambda Bv$ and there exists no $w(x) \in \mathbb{C}[x]^n$ such that $(A - xB)w(x) = 0$ and $w(\lambda) = v$.

Note that these definitions are equivalent to Definition 1.1.2 when (A, B) is regular. Once again, a singular pencil may have eigenvalues at infinity, which arise when $\operatorname{rank}(B) < \operatorname{rank}(A - xB)$. Corresponding right eigenvectors in this case belong to $\operatorname{null}(B) \setminus \operatorname{null}(A)$. Nevertheless, we no longer obtain a full set of n eigenvalues if we include those at infinity. In fact, a singular pencil always has strictly fewer than n eigenvalues.

[2] Note that both $\det(A)$ and $\det(B)$ appear as coefficients in the characteristic polynomial $\det(A - xB)$.

Throughout, we use $\Lambda(A, B)$ to denote the spectrum of any matrix pencil (A, B) (and similarly $\Lambda(A)$ represents the spectrum of A). To give some insight into the way generalized eigenvalue problems arise in applications – and to provide a backdrop for the subsequent numerical discussion – we present a motivating example below.

Example 1.1.5 (Motivation from Machine Learning)**.** Suppose we have data stored in the rows of two matrices $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{k \times n}$. If handed a new data point sampled from one of the populations X and Y are drawn from, how do we decide which population it corresponds to? One option is to use a linear support vector machine (SVM), which locates an affine hyperplane separating X and Y that can be used to classify data, assuming such a hyperplane exists [71].

Alternatively, we could construct the matrices

$$A = \begin{pmatrix} X & -\mathbb{1} \end{pmatrix}^T \begin{pmatrix} X & -\mathbb{1} \end{pmatrix}, \quad B = \begin{pmatrix} Y & -\mathbb{1} \end{pmatrix}^T \begin{pmatrix} Y & -\mathbb{1} \end{pmatrix} \tag{1.1}$$

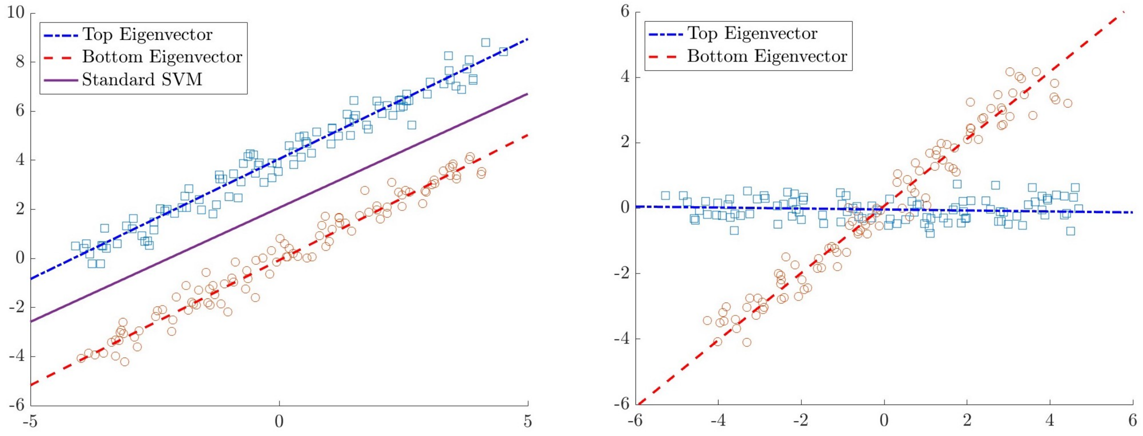
and solve the generalized eigenvalue problem (A, B). Assuming $\begin{pmatrix} Y & -\mathbb{1} \end{pmatrix}$ is full rank, the top and bottom eigenvectors of (A, B) – i.e., corresponding to its largest/smallest eigenvalues in magnitude – define hyperplanes that are close to the data in Y and X, respectively, but maximally far from each other. If the top eigenvector is $v = (w \;\; \gamma)^T$ for $\gamma \in \mathbb{R}$, the corresponding hyperplane $\left\{x \in \mathbb{R}^n : w^T x - \gamma = 0\right\}$ approximates the points in Y, and the same can be said for the bottom eigenvector and X. This is a simple consequence of the observation

$$\max_{(w,\gamma) \neq 0} \left(\frac{||Xw - \gamma \mathbb{1}||_2}{||Yw - \gamma \mathbb{1}||_2} \right)^2 = \max_{v \neq 0} \frac{v^T A v}{v^T B v}, \tag{1.2}$$

where the latter defines the largest eigenvalue of (A, B) when B is positive definite[3] via an extension of Courant-Fischer.

This approach to classification was introduced by Mangasarian and Wild [97], and its advantages over standard SVMs are clear: it can separate overlapping data sets

[3] In this case, (A, B) belongs to the special class of *definite pencils*, which have real eigenvalues.



(a) Separable Data

(b) Overlapping Data

Figure 1.1. Hyperplane approximations for synthetic data obtained via a generalized eigenvalue problem as in Example 1.1.5 (labeled by the eigenvector they are derived from). A standard linear SVM is included for comparison when it is capable of splitting the data.

while also providing (linear) approximations to them. An example in $\mathbb{R}^2$ is presented in Figure 1.1. From a numerical perspective, this method trades an optimization problem for an $(n+1) \times (n+1)$ generalized eigenvalue problem, which will be challenging to work with if the original data is dense and high-dimensional. This is particularly noteworthy since, as with SVMs, the approach outlined here can be combined with a kernel trick to make nonlinear classifications.

1.1.1 Matrix Pencil Diagonalization

As in the single-matrix eigenvalue problem, the spectral information of a pencil can be obtained from a number of factorizations. We summarize the most important of these in this section. First up is a generalized Schur form introduced by Stewart [121].

Definition 1.1.6. $(T_A, T_B) = (U^H AV, U^H BV)$ is a *generalized Schur form* of (A, B) if U, V are unitary and T_A, T_B are upper triangular.

Every pencil has a generalized Schur form.[4] In the regular case, the diagonal

[4] The regular case was the original focus of Stewart [121, Theorem 3.1]. For a detailed discussion of computing a generalized Schur form for singular pencils, see work of Demmel and Kågström [40, 41].

entries $T_A(i,i)/T_B(i,i)$ record the eigenvalues of (A,B) and the columns of U and V span corresponding right/left deflating subspaces.

Definition 1.1.7. Subspaces $\mathcal{X}, \mathcal{Y} \subset \mathbb{C}^n$ are respectively *right and left deflating subspaces* of an $n \times n$ regular pencil (A,B) if $\dim(\mathcal{X}) = \dim(\mathcal{Y})$ and $\operatorname{span}\{Ax, Bx : x \in \mathcal{X}\} = \mathcal{Y}$.

It is not difficult to see that any collection of right eigenvectors of a regular pencil spans a corresponding right deflating subspace. If the leading columns of V form a basis for this space, the corresponding columns of U will span the associated left deflating subspace. In this way, the right and left deflating subspaces of a regular pencil generalize the invariant eigenspaces of an individual matrix. Moreover, the decompositions $A = UT_AV^H$ and $B = UT_BV^H$ imply $A^HU = VT_A^H$ and $B^HU = VT_B^H$ – i.e., trailing columns of U and V span deflating subspaces of (A^H, B^H).

Despite the similarity in naming, the left deflating subspace corresponding to a set of right eigenvectors is not typically spanned by a set of left eigenvectors. In the special case that B is invertible, however, a basis for a left deflating subspace can be constructed from left eigenvectors of the matrix $B^{-H}A^H$ (equivalently right eigenvectors of AB^{-1}). Finally, we note that deflating subspaces are only defined for regular pencils; for a discussion of the singular pencil analog – reducing subspaces – see work of Van Dooren [137].

Next we consider the Kronecker canonical form, which generalizes the Jordan decomposition of an individual matrix. As its name suggests, this canonical form was first introduced by Kronecker [85], though it can also be viewed as an extension of an earlier decomposition for regular pencils derived by Weierstrass [143].

Definition 1.1.8. The *Kronecker canonical form* of the pencil (A,B) is the decomposition

$$S - xT = P^{-1}(A - xB)Q,$$

where P and Q are invertible and $S - xT$ is block diagonal consisting of square blocks

$$J_k(\lambda) = \begin{pmatrix} \lambda - x & 1 & & \\ & \lambda - x & \ddots & \\ & & \ddots & 1 \\ & & & \lambda - x \end{pmatrix}, \quad N_k = \begin{pmatrix} 1 & -x & & \\ & 1 & \ddots & \\ & & \ddots & -x \\ & & & 1 \end{pmatrix} \in \mathbb{C}^{k \times k}$$

and non-square blocks

$$L_k = \begin{pmatrix} -x & 1 & & \\ & \ddots & \ddots & \\ & & -x & 1 \end{pmatrix} \in \mathbb{C}^{k \times (k+1)} \quad \text{or} \quad L_k^T = \begin{pmatrix} -x & & \\ 1 & \ddots & \\ & \ddots & -x \\ & & 1 \end{pmatrix} \in \mathbb{C}^{(k+1) \times k}.$$

The blocks $J_k(\lambda)$ and N_k appearing in the Kronecker canonical form of (A, B) represent the regular structure of the pencil; $J_k(\lambda)$ is a standard Jordan block with finite eigenvalue λ, while N_k corresponds to an infinite eigenvalue with multiplicity k. The remaining blocks L_k and L_k^T constitute the singular structure of (A, B). For any value of x, the block L_k has a one-dimensional (right) null space spanned by the vector $(1 \ x \ \cdots \ x^k)^T$. Note that Definition 1.1.8 technically defines the Kronecker canonical form for both square and non-square pencils. When (A, B) is square, the blocks L_k and L_k^T must be padded by a row or column of zeros, respectively, if $S - xT$ is to be block diagonal.

Example 1.1.9. Consider the pencil (A, B) with

$$A = \begin{pmatrix} 2 & -1 & -5 & -1 \\ 6 & -2 & -11 & -2 \\ 5 & 0 & -2 & 0 \\ 3 & 1 & 3 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & -1 & -4 & -2 \\ 2 & -3 & -12 & -6 \\ -1 & -3 & -11 & -6 \\ -2 & -2 & -7 & -4 \end{pmatrix}. \tag{1.3}$$

The Kronecker canonical form of (A, B) can be obtained by factoring $A - \lambda B$ as

$$A - \lambda B = \begin{pmatrix} -3 & 1 & 1 & 1 \\ -8 & 3 & 2 & 0 \\ -5 & 3 & 0 & 1 \\ -2 & 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1-\lambda & 0 & 0 & 0 \\ 0 & -\lambda & 1 & 0 \\ 0 & 0 & -\lambda & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & 1 & -4 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{pmatrix}^{-1}, \tag{1.4}$$

which reveals that the pencil is singular with only one simple (i.e., multiplicity one) eigenvalue at $\lambda = 1$. Note that the row of zeros in (1.4) indicates that A and B have overlapping left null spaces. We take this example from a paper of Lotz and Noferini [90] and will return to it later in the thesis.

Again taking a cue from the single-matrix eigenvalue problem, we might ask when the Kronecker canonical form simplifies, particularly to something diagonal. This gives rise to the primary factorization we are interested in.

Definition 1.1.10. A pencil (A, B) is *diagonalizable* if there exist invertible S, T such that $(D_1, D_2) = (S^{-1}AT, S^{-1}BT)$ is diagonal. In this case, the matrices S and T *jointly diagonalize* A and B.

Here, our intuition from the standard eigenvalue problem carries over directly: (A, B) is diagonalized by matrices S and T containing its left and right eigenvectors, assuming a full set of independent eigenvectors (including those associated to infinite eigenvalues) exists. If (A, B) is regular, the diagonal pencil (D_1, D_2) records the eigenvalues of (A, B) as $D_1(i, i)/D_2(i, i)$, with zeros on the diagonal of D_2 corresponding to infinity. Note that D_1 and D_2 can never have zero in the same diagonal entry if (A, B) is regular.

As we might expect, an arbitrary pencil (A, B) is not necessarily diagonalizable, even when it is regular. We do, however, note that any regular pencil with a set of distinct eigenvalues (including infinity) admits a diagonalization. This mirrors again the standard eigenvalue problem, where any matrix with a full set of distinct eigenvalues is guaranteed to be diagonalizable.

We are now ready to state the central problem addressed in this thesis.

Approximate Matrix Pencil Diagonalization

Given a pencil (A, B) with $A, B \in \mathbb{C}^{n\times n}$, construct an approximate diagonalization

$$(A, B) \approx \left(SD_1T^{-1},\ SD_2T^{-1}\right)$$

for invertible $S, T \in \mathbb{C}^{n\times n}$ and diagonal (D_1, D_2).

In solving this problem, we seek an exact diagonalization of a nearby pencil, where "nearby" is measured by $||A - SD_1T^{-1}||_2$ and $||B - SD_2T^{-1}||_2$, whose eigenvalues/eigenvectors can stand in as approximations for those of (A, B). This as a backward-error oriented approach to the generalized eigenvalue problem, which acknowledges the reality that not all pencils admit a diagonalization and even those that do cannot be diagonalized exactly in finite-precision arithmetic. Importantly, as we will see, this approach is naturally compatible with the decision to apply regularizing perturbations to A and B.

As mentioned in the preface to this chapter, the standard method for solving the generalized eigenvalue problem – and therefore producing an approximate diagonalization of any matrix pencil – is the QZ algorithm of Moler and Stewart [102], which was first introduced in 1973. QZ finds the eigenvalues and eigenvectors of (A, B) by producing its generalized Schur form, implicitly applying the QR algorithm for the standard eigenvalue problem[5] to AB^{-1} (or $B^{-1}A$). To do this, QZ assumes blindly that B is invertible. While this may seem problematic, it ultimately poses little harm: QZ is capable of identifying infinite eigenvalues and can be used on singular pencils.

Due to the popularity of the QZ algorithm, much research has focused on its performance, including on pencils that have infinite eigenvalues [140] or that are nearly singular [144]. Additional work has sought to refine its numerical details [82, 84, 139]. The bottom line is that modern implementations are regarded as generally reliable and backwards stable, which here means that QZ computes accurately the eigenvalues/eigenvectors of a nearby pencil (see Section 1.5). Indeed, QZ is the default generalized eigensolver called by Matlab's intrinsic function `eig`.

Beyond its $O(n^3)$ complexity, the primary drawback to QZ is its resistance to parallelization. Parallel implementations have been pursued, most notably by Adlerborn,

[5] The QR algorithm was derived independently by Francis [56,57] and Kublanovskaya [86] and computes eigenvalues/eigenvectors of A by (1) reducing the matrix to Hessenberg from and (2) applying (possibly shifted) QR factorizations to obtain a Schur decomposition. For a more recent summary of the method, see [141].

Kågström, and Kressner [1], but are not widely used. As a result, QZ remains somewhat poorly suited to settings where input matrices are prohibitively large – i.e., large enough to exceed available fast memory. An alternative, divide-and-conquer approach to the generalized eigenvalue problem, which we discuss at length in Section 1.3, naturally avoids this issue. For now we note that divide-and-conquer algorithms have remained out-of-reach because, in contrast to QZ, they typically cannot be implemented on arbitrary inputs (that is, in certain situations divide-and-conquer may not only lose accuracy but fail entirely). As we will see, randomization points a way around this problem.

Before moving on, we note that a different set of numerical tools are typically used if only a certain subset of eigenvalues and eigenvectors are desired. For sparse problems in particular, these include the trace minimization algorithm [115], projection methods [114], and extensions of the Lanczos procedure [53].

1.2 Perturbation Theory

Focusing on backward-stable diagonalizations (and allowing random perturbations) begs the question: how sensitive is the generalized eigenvalue problem to changes in the input matrices? In what situations should we expect that the eigenvalues and eigenvectors of an approximate diagonalization are close to those of (A, B)? With these questions in mind, we discuss in this section perturbation theory for the generalized eigenvalue problem. Along the way we define the pseudospectrum of a matrix pencil, a key theoretical tool used throughout the thesis.

To set the stage, we begin by considering a few examples.

Example 1.2.1. Consider first the following 2×2 matrices:

$$A = B = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-10} \end{pmatrix}, \quad \widetilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1.5 \times 10^{-10} \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} 1 & 0 \\ 0 & 5 \times 10^{-11} \end{pmatrix}. \tag{1.5}$$

By construction, $||A - \widetilde{A}||_2 = ||B - \widetilde{B}||_2 = 5 \times 10^{-11}$ while $\Lambda(A, B) = \{1\}$ and $\Lambda(\widetilde{A}, \widetilde{B}) =$

$\{1,3\}$ – i.e., a tiny perturbation in A and B results in macroscopic changes to the eigenvalues, even though (A,B) and $(\widetilde{A},\widetilde{B})$ are both regular and diagonalizable.

Example 1.2.2. Consider next an infamous singular example due to Wilkinson [144]:

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \widetilde{A} = \begin{pmatrix} 2+\epsilon_1 & \epsilon_2 \\ \epsilon_3 & 0 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} 1+\eta_1 & \eta_2 \\ \eta_3 & 0 \end{pmatrix}. \tag{1.6}$$

Here, (A,B) has only one eigenvalue $\lambda = 2$ but $(\widetilde{A},\widetilde{B})$ has eigenvalues ϵ_2/η_2 and ϵ_3/η_3, which can take any two values regardless of perturbation size. In particular, we are not guaranteed to get at least one eigenvalue near the original $\lambda = 2$.

Of course, these perturbations are highly structured and therefore unlikely to occur either randomly or from round-off error in finite-precision arithmetic. Nevertheless, they set our expectations: in certain settings eigenvalue recovery is not possible, so any useful perturbation bound will require some set of assumptions on (A,B).

1.2.1 Pseudospectra and Bauer-Fike

In this thesis, the primary tool for measuring eigenvalue perturbations is the pseudospectrum of the corresponding matrix or matrix pencil. For the single-matrix eigenvalue problem, the definition is standard.

Definition 1.2.3. For any $\epsilon > 0$, the *ϵ-pseudospectrum* of A is

$$\Lambda_\epsilon(A) = \{z : \text{there exists } u \neq 0 \text{ with } (A+E)u = zu \text{ for some } ||E||_2 \leq \epsilon\}.$$

Each pseudospectrum $\Lambda_\epsilon(A)$ consists of connected components in $\mathbb{C}$ containing at least one eigenvalue of A. As $\epsilon \to 0$, these connected components collapse to the true eigenvalues of the matrix. When A is diagonalizable, this is quantified explicitly via the Bauer-Fike Theorem [18].

Theorem 1.2.4 (Bauer-Fike)**.** *If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $A \in \mathbb{C}^{n\times n}$ and V is any invertible matrix that diagonalizes A, then*

$$\bigcup_{i=1}^{n} B_\epsilon(\lambda_i) \subseteq \Lambda_\epsilon(A) \subseteq \bigcup_{i=1}^{n} B_{\epsilon\kappa_2(V)}(\lambda_i).$$

Intuitively, $\Lambda_\epsilon(A)$ contains all eigenvalues that are obtainable from A via a perturbation of size at most ϵ (in the spectral norm). Accordingly, $\Lambda_\epsilon(A)$ can be used either to measure the change in eigenvalues due to a deliberate perturbation or record accumulated error in a numerical method – that is, if an algorithm computes exactly the spectral information of $A + E$ with $||E||_2 \leq \epsilon$, bounds on $\Lambda_\epsilon(A)$ imply a certain accuracy for the resulting eigenvalue approximations. The ubiquity[6] of $\Lambda_\epsilon(A)$ in the analysis of eigenvalue problems and their corresponding algorithms is due to this flexibility (see the standard reference work of Trefethen and Embree [134] for a detailed discussion of applications).

For the generalized eigenvalue problem – i.e., for matrix pencils – there is no standard way to define the pseudospectrum. While Trefethen and Embree suggest a handful of options [134, Chapter 45], we work in this thesis with a definition originally due to Frayssé et al. [58].

Definition 1.2.5. For any $\epsilon > 0$, the *ϵ-pseudospectrum* of (A, B) is

$$\Lambda_\epsilon(A, B) = \{z : (A + \Delta A)u = z(B + \Delta B)u, \; u \neq 0, \text{ and } ||\Delta A||_2, ||\Delta B||_2 \leq \epsilon\}.$$

Our motivation for using Definition 1.2.5 is rooted in the realities of working with a pencil (A, B) numerically. In particular, round-off errors will naturally perturb both A and B, possibly independently, meaning any definition that hopes to capture the impact of finite-precision arithmetic must allow for variation in both A and B, without necessarily restricting to relative perturbation sizes. Once again, $\Lambda_\epsilon(A, B)$ consists of connected components in $\mathbb{C}$ that contain $\Lambda(A, B)$.

Frayssé et al. provide a handful of equivalent characterizations of Definition 1.2.5, which we use throughout when convenient.

Theorem 1.2.6 (Frayssé et al. 1996)**.** *The following are equivalent:*

1. $z \in \Lambda_\epsilon(A, B)$.

[6] While analogs of Definition 1.2.3 have existed in the literature for decades, the use of $\Lambda_\epsilon(A)$ as stated here was broadly popularized by Trefethen [133].

2. There exists a unit vector u such that $||(A - zB)u||_2 \leq \epsilon(1 + |z|)$.

3. $||(A - zB)^{-1}||_2 \geq \frac{1}{\epsilon(1+|z|)}$.

4. $\sigma_n(A - zB) \leq \epsilon(1 + |z|)$.

We also note the impact of infinite eigenvalues on $\Lambda_\epsilon(A, B)$. While $\Lambda_\epsilon(A)$ can never be unbounded, $\Lambda_\epsilon(A, B)$ is *always* unbounded provided ϵ is sufficiently large.

Lemma 1.2.7. *$\Lambda_\epsilon(A, B)$ is bounded if and only if $\epsilon < \sigma_n(B)$.*

Proof. Let $\epsilon < \sigma_n(B)$ and suppose $\Lambda_\epsilon(A, B)$ is unbounded. By Theorem 1.2.6, any nonzero $z \in \Lambda_\epsilon(A, B)$ satisfies

$$\frac{1}{\epsilon(1 + |z|)} \leq ||(A - zB)^{-1}||_2 = \frac{1}{|z|} \left|\left| \left(\frac{1}{z} A - B \right)^{-1} \right|\right|_2 \tag{1.7}$$

and therefore

$$\frac{|z|}{\epsilon(1 + |z|)} \leq \left|\left| \left(\frac{1}{z} A - B \right)^{-1} \right|\right|_2 . \tag{1.8}$$

Since $\Lambda_\epsilon(A, B)$ is unbounded, we can take a limit $z \to \infty$ in this inequality to obtain

$$\frac{1}{\epsilon} \leq ||(-B)^{-1}||_2 = \frac{1}{\sigma_n(B)}, \tag{1.9}$$

which implies $\epsilon \geq \sigma_n(B)$, a contradiction.

To prove the converse, suppose now $\epsilon \geq \sigma_n(B)$ and let $U \Sigma V^H$ be the singular value decomposition of B. If $D \in \mathbb{C}^{n \times n}$ is a diagonal matrix with $D(i, i) = 0$ for $1 \leq i \leq n - 1$ and $D(n, n) = -\sigma_n(B)$, then $||UDV^H||_2 = \sigma_n(B) \leq \epsilon$ and therefore eigenvalues of $(A, B + UDV^H)$ belong to $\Lambda_\epsilon(A, B)$. But by construction $B + UDV^H$ is singular, which means $(A, B + UDV^H)$ is singular and/or has an eigenvalue at infinity. In both cases $\Lambda_\epsilon(A, B)$ must be unbounded. □

Finally, we derive an upper bound on $\Lambda_\epsilon(A, B)$ when it exists.

Lemma 1.2.8. *If $\epsilon < \sigma_n(B)$ then any $z \in \Lambda_\epsilon(A, B)$ satisfies*

$$|z| \leq \frac{\epsilon ||B^{-1}||_2 + ||B^{-1}A||_2}{1 - \epsilon ||B^{-1}||_2}.$$

Proof. Suppose $|z| > ||B^{-1}A||_2$ and consider $\frac{B^{-1}A}{z} - I$. For any vector x,

$$\left|\left|\frac{B^{-1}A}{z}x\right|\right|_2 \leq \frac{||B^{-1}A||_2}{|z|}||x||_2 \tag{1.10}$$

by matrix/vector norm compatibility, so

$$||x||_2 - \left|\left|\frac{B^{-1}A}{z}x\right|\right|_2 \geq ||x||_2 - \frac{||B^{-1}A||_2}{|z|}||x||_2 \geq 0, \tag{1.11}$$

where we know $||x||_2 - (||B^{-1}A||_2/|z|)||x||_2$ is positive since $|z| > ||B^{-1}A||_2$. Thus, applying this result and the reverse triangle inequality,

$$\begin{aligned}\sigma_n\left(\frac{B^{-1}A}{z} - I\right) &= \min_{||x||_2=1}\left|\left|\left(\frac{B^{-1}A}{z} - I\right)x\right|\right|_2 \\ &\geq \min_{||x||_2=1}\left[||x||_2 - \frac{||B^{-1}A||_2}{|z|}||x||_2\right] \\ &= 1 - \frac{||B^{-1}A||_2}{|z|}\end{aligned} \tag{1.12}$$

and therefore

$$\left|\left|\left(\frac{B^{-1}A}{z} - I\right)^{-1}\right|\right|_2 = \frac{1}{\sigma_n\left(\frac{B^{-1}A}{z} - I\right)} \leq \frac{1}{1 - \frac{||B^{-1}A||_2}{|z|}}. \tag{1.13}$$

If $z \in \Lambda_\epsilon(A, B)$, we then have

$$\begin{aligned}\frac{1}{\epsilon(1+|z|)} &\leq \frac{||B^{-1}||_2}{|z|}\left|\left|\left(\frac{B^{-1}A}{z} - I\right)^{-1}\right|\right|_2 \\ &\leq \frac{||B^{-1}||_2}{|z|}\frac{1}{1 - \frac{||B^{-1}A||_2}{|z|}} \\ &= \frac{||B^{-1}||_2}{|z| - ||B^{-1}A||},\end{aligned} \tag{1.14}$$

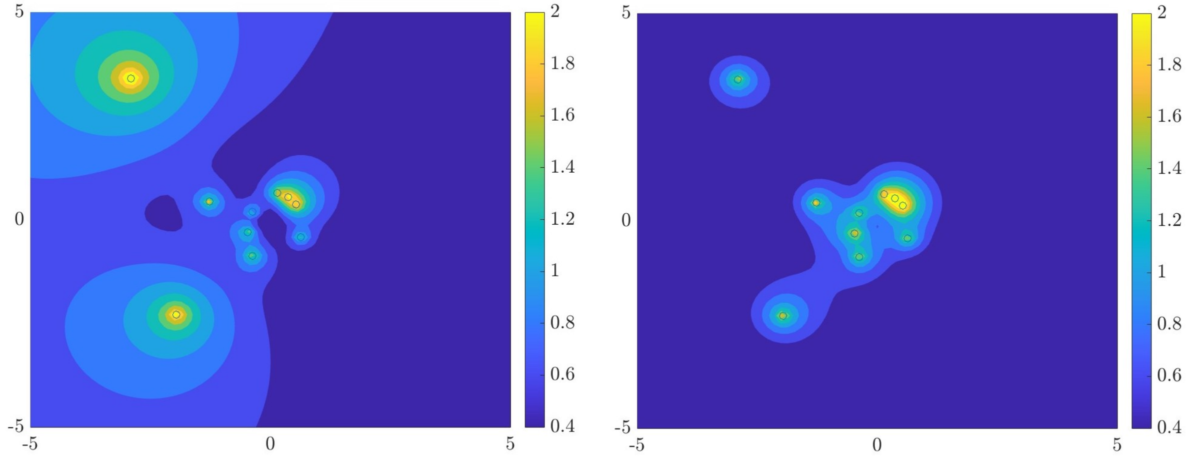
which, rearranging to solve for $|z|$, is equivalent to

$$|z| \leq \frac{\epsilon||B^{-1}||_2 + ||B^{-1}A||_2}{1 - \epsilon||B^{-1}||_2}. \tag{1.15}$$

Since we assumed $|z| > ||B^{-1}A||_2$, we have proved that $z \in \Lambda_\epsilon(A, B)$ for $\epsilon < \sigma_n(B)$ implies

$$|z| \leq \max\left\{||B^{-1}A||_2,\ \frac{\epsilon||B^{-1}||_2 + ||B^{-1}A||_2}{1 - \epsilon||B^{-1}||_2}\right\} = \frac{\epsilon||B^{-1}||_2 + ||B^{-1}A||_2}{1 - \epsilon||B^{-1}||_2} \tag{1.16}$$

which finishes the proof. □



(a) Pseudospectra of (A, B)

(b) Pseudospectra of $B^{-1}A$

Figure 1.2. Pseudospectra of (A, B) and $B^{-1}A$ for Gaussian $A, B \in \mathbb{C}^{10\times 10}$ following Definition 1.2.5 and Definition 1.2.3, respectively. Eigenvalues are plotted with open circles. Pseudospectra are obtained by graphing the level curves of $\log_{10}[(1+|z|)||(A-zB)^{-1}||_2]$ and $\log_{10}[||(B^{-1}A-zI)^{-1}||_2]$ in Matlab R2023a.

Figure 1.2 plots the pseudospectra of (A, B) and $B^{-1}A$ for one (randomly chosen) pair $A, B \in \mathbb{C}^{10\times 10}$. Clearly, the pseudospectra of (A, B) and $B^{-1}A$ differ significantly. Informally, we might say that the pseudospectra of (A, B) are less well-behaved than those of $B^{-1}A$ – particularly around large eigenvalues – despite the fact that they coalesce around the same points.

In an attempt to better understand this comparison, we now extend Theorem 1.2.4 to $\Lambda_\epsilon(A, B)$, obtaining a first perturbation result for the generalized eigenvalue problem. This follows directly from Definition 1.2.5 and Lemma 1.2.8 under the assumption that (A, B) is regular and diagonalizable with no eigenvalues at infinity (i.e., for invertible B).

Theorem 1.2.9 (Bauer-Fike for Matrix Pencils)**.** *Suppose (A, B) is regular and diagonalizable with finite eigenvalues $\lambda_1, \dots, \lambda_n$ and invertible right eigenvector matrix V. For $\epsilon < \sigma_n(B)$ let*

$$r_\epsilon = \epsilon\kappa_2(V)||B^{-1}||_2\left(1 + \frac{\epsilon||B^{-1}||_2 + ||B^{-1}A||_2}{1-\epsilon||B^{-1}||_2}\right)$$

and further set

$$r_i = \begin{cases} \frac{1}{||B||_2}, & \textit{if } A = 0 \\ \max\left\{\frac{1}{||B||_2}, \frac{|\lambda_i|}{||A||_2}\right\}, & \textit{otherwise} \end{cases}$$

for $1 \leq i \leq n$*. Then,*

$$\bigcup_{i=1}^{n} B_{\epsilon r_i}(\lambda_i) \subseteq \Lambda_\epsilon(A,B) \subseteq \bigcup_{i=1}^{n} B_{r_\epsilon}(\lambda_i).$$

Proof. To obtain the first inclusion we note that for any $|\Delta\lambda| \leq \epsilon/||B||_2$ the pencil $(A + \Delta\lambda B, B)$ has eigenvalues $\lambda_i + \Delta\lambda$ while $\Lambda(A + \Delta\lambda B, B) \subseteq \Lambda_\epsilon(A,B)$. Similarly, if $A \neq 0$ then $(A + \Delta\lambda A, B)$ has eigenvalues $\lambda_i(1 + \Delta\lambda)$ and $\Lambda(A + \Delta\lambda A, B) \subseteq \Lambda_\epsilon(A,B)$ as long as $|\Delta\lambda| \leq \epsilon/||A||_2$. For the remaining inclusion we appeal to Theorem 1.2.6: any $z \in \Lambda_\epsilon(A,B)$ satisfies

$$\frac{1}{\epsilon(1+|z|)} \leq ||(A - zB)^{-1}||_2 \leq ||B^{-1}||_2||(B^{-1}A - zI)^{-1}||_2, \tag{1.17}$$

where we note that B is invertible since (A,B) has only finite eigenvalues. Applying the fact that V diagonalizes $B^{-1}A$, meaning $B^{-1}A = V\Lambda V^{-1}$ for a diagonal matrix Λ, this expression becomes

$$\frac{1}{\epsilon(1+|z|)} \leq ||B^{-1}||_2||(V\Lambda V^{-1} - zI)^{-1}||_2 \leq \kappa_2(V)||B^{-1}||_2||(\Lambda - zI)^{-1}||_2. \tag{1.18}$$

Inverting and rearranging, we then have

$$\sigma_n(\Lambda - zI) \leq \epsilon\kappa_2(V)||B^{-1}||_2(1+|z|). \tag{1.19}$$

We complete the proof by noting

$$\sigma_n(\Lambda - zI) = \min_{\lambda_i \in \Lambda(A,B)} |\lambda_i - z| \tag{1.20}$$

and replacing $|z|$ in (1.19) with the upper bound provided by Lemma 1.2.8. □

Remark 1.2.10. Theorem 1.2.9 does not appear to depend on the left eigenvectors of (A,B), which may seem surprising since Theorem 1.2.4 *does* depend on the left eigenvectors of A. Here, we note that if V contains right eigenvectors of (A, B) and B is invertible, then V and $S = BV$ jointly diagonalize A and B, meaning S^{-1} contains left eigenvectors of (A, B) and

$$||V||_2||S^{-1}||_2 = ||V||_2||V^{-1}B^{-1}||_2 \leq \kappa_2(V)||B^{-1}||_2. \tag{1.21}$$

Hence, the bound provided by Theorem 1.2.9 is in fact loosely dependent on the left eigenvectors of (A, B).

Though straightforward, this version of Bauer-Fike is somewhat nonstandard. Classical perturbation results are typically obtained by re-casting each eigenvalue λ as an ordered pair $\langle \alpha, \beta \rangle$, where $\lambda = \alpha/\beta$. In this notation, solutions to the generalized eigenvalue problem satisfy $\beta A v = \alpha B v$ (assuming (A, B) is regular). The main benefit of representing eigenvalues in this way is the natural inclusion of those at infinity, which correspond to $\beta = 0$. Since there are infinitely many choices of α and β that yield the same value of λ, we can think of each eigenvalue $\langle \alpha, \beta \rangle$ as a projective line – i.e., the subspace spanned by $(\alpha\ \beta)^T$.

Under this framework, distance between two eigenvalues $\langle \alpha_1, \beta_1 \rangle$ and $\langle \alpha_2, \beta_2 \rangle$ is measured by the chordal metric:

$$\chi\left(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle\right) = \frac{|\alpha_1 \beta_2 - \beta_1 \alpha_2|}{\sqrt{|\alpha_1|^2 + |\beta_1|^2}\sqrt{|\alpha_2|^2 + |\beta_2|^2}}. \tag{1.22}$$

Dividing both the numerator and denominator of (1.22) by $|\beta_1 \beta_2|$, we observe that if $\lambda_1 = \alpha_1/\beta_1$ and $\lambda_2 = \alpha_2/\beta_2$ then

$$\chi\left(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle\right) = \frac{|\lambda_1 - \lambda_2|}{\sqrt{|\lambda_1|^2 + 1}\sqrt{|\lambda_2|^2 + 1}}. \tag{1.23}$$

In other words, the chordal distance between $\langle \alpha_1, \beta_1 \rangle$ and $\langle \alpha_2, \beta_2 \rangle$ is half the Euclidean distance between the images of λ_1 and λ_2 under the stereographic projection. This ensures that the distance between any two eigenvalues is at most one, including eigenvalues at infinity.

While perturbation results in terms of χ originate with Stewart [122], the standard bound for pencils that are both regular and diagonalizable is due to Elsner and Sun [51]. Theorem 1.2.11 can be interpreted as a chordal metric analog of Bauer-Fike.

Theorem 1.2.11 (Elsner and Sun 1982)**.** *Let (A, B) be a regular, diagonalizable pencil with eigenvalues $\langle \alpha_i, \beta_i \rangle$. If V is any matrix of right eigenvectors of (A, B) and $\langle \widetilde{\alpha}_i, \widetilde{\beta}_i \rangle$ are the eigenvalues of the regular pencil $(\widetilde{A}, \widetilde{B})$, then*

$$\max_i \min_j \chi\left(\langle \alpha_i, \beta_i \rangle, \langle \widetilde{\alpha}_j, \widetilde{\beta}_j \rangle\right) \leq \kappa_2(V) ||(AA^H + BB^H)^{-1/2}||_2 ||(A - \widetilde{A}, B - \widetilde{B})||_2.$$

This bound offers an improvement over Theorem 1.2.9 in that it does not require B to be invertible, as perturbations in infinite eigenvalues can be bounded in χ where they cannot in absolute value. When B is invertible, the two are essentially equivalent. Since the chordal metric is somewhat less intuitive to work with, as $|\lambda - \widetilde{\lambda}|$ may be large even though $\chi(\langle \alpha, \beta \rangle, \langle \widetilde{\alpha}, \widetilde{\beta} \rangle)$ is small, we prefer Theorem 1.2.9. Moreover, for the situation in which we will need Bauer-Fike later on, the assumption that B is invertible will be satisfied (with a bound on $\sigma_n(B)$ known).

Note that the assumptions baked into these results – that (A, B) is regular and diagonalizable – are not included for convenience. Only in this setting can we expect to have perturbation bounds that cover the entire spectrum. Even when we do, these bounds are large if the corresponding eigenvectors are poorly conditioned or the matrices are nearly singular, allowing them to cover a problematic case like Example 1.2.1. As we will explore later in this thesis, there is good empirical evidence to suggest that certain eigenvalues of pencils that are not diagonalizable, or even regular, are stable under perturbation, though the theory to explain these observations is still under development (see for example [90]).

Of course, Theorem 1.2.9 is not the only alternate formulation of Bauer-Fike developed since the work of Elsner and Sun. Minor improvements to Theorem 1.2.11 can be found in subsequent work of Elsner and Lancaster [50]. Chu, meanwhile, stated their own version of Bauer-Fike, which comprises four separate bounds depending on whether the initial and perturbed eigenvalues are finite/infinite [30, 31]. Finally, we note a recent sharp version of Shi and Wei stated in terms of the sign-complex spectral radius [118].

1.2.2 Main Perturbation Bound

To this point, we have not considered how perturbations affect eigenvectors. Classical bounds for eigenspaces (or more precisely deflating subspaces) are discussed at length in the standard work of Stewart and Sun [126, Chapter VI]. Once again, these bounds are somewhat difficult to use numerically, stated in terms of opaque norms on pairs of spaces. For this reason, we again derive our own perturbation result, which bounds the Euclidean distance between specific eigenvectors before and after perturbation. To obtain such a result, we must add the additional assumption that (A, B) has no repeated eigenvalues (and therefore that each eigenvector is unique up to scaling). With this in mind, we introduce a pair of quantities that capture the conditioning of the generalized eigenvalue problem (A, B).

Definition 1.2.12. The (right) *eigenvector condition number* of a diagonalizable pencil (A, B) is

$$\kappa_V(A, B) = \inf_T \kappa(T),$$

where the infimum is taken over all invertible matrices T containing a full set of right eigenvectors of (A, B).

Definition 1.2.13. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of (A, B) repeated according to multiplicity. The *eigenvalue gap* of (A, B) is

$$\operatorname{gap}(A, B) = \min_{i \neq j} |\lambda_i - \lambda_j|.$$

Each of these can be similarly defined for an individual matrix, in which case $\kappa_V(B^{-1}A) = \kappa_V(A, B)$ and $\operatorname{gap}(B^{-1}A) = \operatorname{gap}(A, B)$. Note that Theorem 1.2.9 can be tightened slightly by replacing $\kappa_2(V)$ with $\kappa_V(A, B)$. With this in mind, we now prove a rigorous bound – for both eigenvalues and eigenvectors – in terms of $\kappa_V(A,B)$ and $\mathrm{gap}(A,B)$.

Theorem 1.2.14. *Let $A, B, \widetilde{A}, \widetilde{B} \in \mathbb{C}^{n\times n}$ with $||A||_2 \leq 1$. Assume that B is invertible and that the pencil (A,B) has a full set of distinct, finite eigenpairs $\{(\lambda_i, v_i)\}_{i=1}^n$. If $||A-\widetilde{A}||_2 \leq \varepsilon\sigma_n(B)\mathrm{gap}(A,B)$ and $||B-\widetilde{B}||_2 \leq \varepsilon\sigma_n^2(B)\mathrm{gap}(A,B)$ for*

$$\varepsilon \leq \frac{1}{32\kappa_V(A,B)}$$

then $(\widetilde{A},\widetilde{B})$ has no repeated eigenvalues and any set of corresponding eigenpairs $\{(\widetilde{\lambda}_i,\widetilde{v}_i)\}_{i=1}^n$ satisfies for all $i = 1,\ldots,n$

$$|\lambda_i - \widetilde{\lambda}_i| \leq 4\mathrm{gap}(A,B)\kappa_V(A,B)\varepsilon \quad \textit{and} \quad ||v_i - \widetilde{v}_i||_2 \leq 24n\kappa_V(A,B)\varepsilon$$

after multiplying the v_i by phases and reordering, if necessary.

Proof. We first note that $\varepsilon < \frac{1}{32\kappa_V(A,B)}$ implies

$$||B-\widetilde{B}||_2 \leq \frac{\sigma_n^2(B)\mathrm{gap}(A,B)}{32\kappa_V(A,B)} \leq \frac{\sigma_n(B)}{16\kappa_V(A,B)} \leq \frac{\sigma_n(B)}{2}, \tag{1.24}$$

where the second inequality follows from $\mathrm{gap}(A,B) \leq 2||B^{-1}A||_2 \leq \frac{2}{\sigma_n(B)}$ (since $||A||_2 \leq 1$). Thus, $\sigma_n(\widetilde{B}) \geq \sigma_n(B) - ||B-\widetilde{B}||_2 \geq \frac{\sigma_n(B)}{2}$, so $\widetilde{B}$ is invertible. With this in mind, consider the matrices $B^{-1}A$ and $\widetilde{B}^{-1}\widetilde{A}$, which are diagonalizable with the same eigenpairs as (A,B) and $(\widetilde{A},\widetilde{B})$, respectively. Using again the fact that $||A||_2 \leq 1$, we have

$$\begin{aligned}
||B^{-1}A - \widetilde{B}^{-1}\widetilde{A}||_2 &= ||B^{-1}A - \widetilde{B}^{-1}A + \widetilde{B}^{-1}A - \widetilde{B}^{-1}\widetilde{A}||_2 \\
&\leq ||B^{-1} - \widetilde{B}^{-1}||_2||A||_2 + ||\widetilde{B}^{-1}||_2||A-\widetilde{A}||_2 \\
&\leq ||\widetilde{B}^{-1}||_2||\widetilde{B}-B||_2||B^{-1}||_2 + ||\widetilde{B}^{-1}||_2||A-\widetilde{A}||_2 \\
&\leq \frac{2}{\sigma_n(B)}\varepsilon\sigma_n^2(B)\mathrm{gap}(A,B)\frac{1}{\sigma_n(B)} + \frac{2}{\sigma_n(B)}\varepsilon\sigma_n(B)\mathrm{gap}(A,B) \\
&\leq 4\varepsilon\mathrm{gap}(A,B).
\end{aligned} \tag{1.25}$$

This implies that the eigenvalues of $\widetilde{B}^{-1}\widetilde{A}$ belong to the $4\varepsilon\mathrm{gap}(A,B)$-pseudospectrum of $B^{-1}A$, and moreover we can continuously deform the eigenvalues of $B^{-1}A$ to those of

$\widetilde{B}^{-1}\widetilde{A}$ without leaving this pseudospectrum. Since $\varepsilon < 1/4$ and $B^{-1}A$ is diagonalizable, single matrix Bauer-Fike (Theorem 1.2.4) implies that $\widetilde{B}^{-1}\widetilde{A}$, and by extension $(\widetilde{A}, \widetilde{B})$, has a full set of distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. Bauer-Fike also guarantees that these eigenvalues can be ordered so that each $\widetilde{\lambda}_i$ satisfies

$$|\lambda_i - \widetilde{\lambda}_i| \leq 4\text{gap}(A, B)\kappa_V(B^{-1}A)\varepsilon \tag{1.26}$$

for a corresponding eigenvalue λ_i of $B^{-1}A$. Recalling that $\kappa_V(B^{-1}A) = \kappa_V(A, B)$, we obtain the first guarantee.

The second guarantee follows from the proof of [16, Proposition 1.1] since the requirement $\varepsilon < (32\kappa_V(A, B))^{-1}$ ensures by (1.25) that

$$||B^{-1}A - \widetilde{B}^{-1}\widetilde{A}||_2 \leq \frac{\text{gap}(B^{-1}A)}{8\kappa_V(B^{-1}A)}. \tag{1.27}$$

Note that while [16, Proposition 1.1] comes with norm assumptions on the matrices, these are not used in the proof of the eigenvector guarantee. □

Theorem 1.2.14 parallels a similar bound for the standard eigenvalue problem due to Banks et al. [16, Proposition 1.1], whose proof relies on some basic perturbation results for individual matrices summarized in [64]. Developing corresponding results for the generalized problem may allow for a similar forward error guarantee with looser requirements on $||A - \widetilde{A}||_2$ and $||B - \widetilde{B}||_2$. Absent these improvements, Theorem 1.2.14 is fairly restrictive. Nevertheless, it provides important context for our work: if (A, B) satisfies its requirements, Theorem 1.2.14 indicates how accurately a diagonalization must be computed to yield eigenvalue and eigenvector approximations of a certain quality.

1.3 Divide-and-Conquer

We return now to the numerical question of approximately diagonalizing an arbitrary pencil (A, B). In Section 1.1, we identified a lack of parallel implementations as the primary drawback to the QZ algorithm. In an effort to circumvent this weakness, we focus on divide-and-conquer methods, which recursively split a pencil (A, B) into smaller ones with disjoint spectra and therefore naturally parallelize. In this section, we outline the divide-and-conquer approach to eigenvalue problems and discuss the primary obstacles to implementation.

Suppose we start with a regular input pencil (A, B). Let U_R and U_L be matrices whose orthonormal columns span the right and left deflating subspaces corresponding to a subset $S \subset \Lambda(A, B)$. In this case, it is easy to see that (λ, w) is an eigenpair of $(U_L^H AU_R, U_L^H BU_R)$ if and only if $(\lambda, U_R w)$ is an eigenpair of (A, B). Hence, we can obtain eigenvalues/eigenvectors of (A, B) by diagonalizing the smaller pencil $(U_L^H AU_R, U_L^H AU_R)$, whose spectra is contained in S and whose eigenvectors are in simple correspondence with those of (A, B). While this will only yield eigenpairs associated to S, we can simply repeat this process for $\Lambda(A, B) \setminus S$ to recover the remaining pairs.

Of course to diagonalize $(U_L^H AU_R, U_L^H AU_R)$ we can apply divide-and-conquer again, obtaining two smaller pencils that can be further split themselves. Continuing this recursive division, a full set of eigenpairs for (A, B) can be reconstructed from those of the smallest subproblems, which are either 1×1 and therefore trivial, or small enough to handle with existing techniques (for example QZ). At each step, a subproblem of the form $(U_L^H AU_R, U_L^H AU_R)$ can be passed off to a separate processor and handled completely independently. This is the natural parallelization of divide-and-conquer.

In this framework, the matrices U_R and U_L are constructed by first computing projectors onto corresponding right/left deflating subspaces.

Definition 1.3.1. The linear transformation $P : \mathbb{C}^n \to \mathbb{C}^n$ is a *projector* onto the subspace $\mathcal{X} \subset \mathbb{C}^n$ if $\text{range}(P) = \mathcal{X}$ and $P^2 = P$.

Assuming (A, B) has an invertible (right) eigenvector matrix V whose leading columns correspond to the eigenvalues in S, the matrix

$$P_R = V \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} V^{-1} \tag{1.28}$$

is a projector onto the right deflating subspace spanned by U_R. Similarly, a projector onto the left deflating subspace associated to U_L is

$$P_L = X \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} X^{-1} \tag{1.29}$$

for X a matrix containing a basis for the column space of $\begin{bmatrix} AV & BV \end{bmatrix}$. In practice, we obtain P_L from the pencil (A^H, B^H), recalling that the left eigenvectors of $B^{-H}A^H$ span left deflating subspaces of (A, B). We refer to P_R and P_L as the *spectral projectors* of (A, B) corresponding to the set of eigenvalues $S \subset \Lambda(A, B)$.

Remark 1.3.2. The expressions (1.28) and (1.29) center the case where (A, B) is regular with a full set of right eigenvectors. While we can define P_R and P_L for arbitrary pencils in terms of the Kronecker canonical form of (A, B), this is unnecessary here, as we will ultimately be able to assume our input pencil is both regular and diagonalizable.

Once these projectors are found, U_R and U_L can be obtained by computing rank-revealing factorizations of P_R and P_L. The rank-revealing piece is critical here, as the rank of either projector tells us how many eigenvalues lie in S and therefore how significant the corresponding split is. In practice, a potential split may be rejected if this rank is either too large or too small, in which case it will not significantly reduce the size of the problem. While there are many ways to compute a rank-revealing factorization [29, 66, 125], we are particularly interested in the URV factorization of Stewart [125].

Definition 1.3.3. $A = URV$ is a *URV factorization* of the matrix A if R is upper triangular and U, V are unitary.

When A has effective rank k, meaning there is a significant gap between $\sigma_k(A)$ and $\sigma_{k+1}(A)$, the factorization $A = URV$ with

$$R = \begin{pmatrix} R_{11} & R_{21} \\ 0 & R_{22} \end{pmatrix}, \quad R_{11} \in \mathbb{C}^{k \times k} \tag{1.30}$$

is rank-revealing if $\sigma_k(R_{11})$ is a "good" approximation to $\sigma_k(A)$ and $||R_{22}||_2$ is a "good" approximation of $\sigma_{k+1}(A)$. We will make the meaning of "good" precise in Chapter 4. For now, we note that all of the versions of divide-and-conquer considered in this thesis will make use of a randomized, rank-revealing URV factorization introduced by Demmel, Dumitriu, and Holtz [38], which is simple to implement and especially compatible with inverse-free computations [11].

Remark 1.3.4. Because we use a randomized rank-revealing factorization, there is a chance that U_R and U_L are computed incorrectly, even if P_R and P_L are not. This is our motivation for deriving P_L from (A^H, B^H) instead of U_R; while in principle we could obtain U_L from a rank-revealing factorization of $\begin{bmatrix} AU_R & BU_R \end{bmatrix}$ – or even just BU_R if we know B is invertible – computing P_L and P_R independently allows us to scout potential failures. A situation in which U_R and U_L have different dimensions, for example, would suggest that either a rank-revealing factorization has failed or that P_R and P_L were not computed accurately enough.

In the terminology established above, each step of divide-and-conquer consists of the following:

1. Divide the spectrum into two disjoint pieces.
2. Obtain P_R and P_L for both sets of deflating subspaces.
3. Compute a (random) rank-revealing URV factorization of each projector.
4. Construct the next subproblems and recur.

Note that if we are able to consistently split the eigenvalues into sets of roughly equal size, only $O(\log(n))$ steps of divide-and-conquer are required to find a full set of eigenvalues/eigenvectors for (A, B). In the worst case, divide-and-conquer may require as many as $O(n)$ steps.

Before moving on, we note that the high-level strategy outlined above can also be applied to the standard eigenvalue problem, in which case the left/right projectors are replaced by a single spectral projector onto a corresponding invariant subspace. In fact, divide-and-conquer originates with the standard eigenvalue problem, beginning with work of Beavers and Denman [19], which demonstrated that the matrix sign function of Roberts [109] could be used to compute spectral projectors. Subsequent work for the generalized eigenvalue problem has extended the use of the sign function and considered a number of alternatives [7, 11, 60, 95]. While again this is not our primary focus, we do note that divide-and-conquer has already found wide use in certain structured settings, particularly for symmetric eigenvalue problems [107].

1.3.1 Two Challenges

Practically speaking, there are two main challenges to overcome when implementing a divide-and-conquer eigensolver. In large part, this thesis is dedicated to rigorously addressing both. Since answers will be problem-dependent, we note again that we are primarily interested in the most general setting (though we will explore a specialization in Chapter 5).

Divide-and-Conquer Challenge One

How do we reliably (and significantly) divide the spectrum $\Lambda(A, B)$ at every step? If we do so with a generalized circle Γ in $\mathbb{C}$, can we obtain some guarantee that $\Lambda_\epsilon(A, B) \cap \Gamma = \emptyset$ for ϵ not too small?

This first challenge is in some sense existential; not only are there no known deterministic answers, but simply identifying viable splits will not be good enough. If we hope to outperform an $O(n^3)$ solver like QZ, we will need the splits to be significant, meaning as close to 50/50 as possible, at every step. In short, if at least a fraction of eigenvalues can be separated with each split, only logarithmically many steps of divide-and-conquer will be necessary, and the algorithm will run in nearly matrix multiplication time provided each individual step can be done cheaply (that is, in nearly matrix multiplication time itself).

Historically, this obstacle to divide-and-conquer has been largely ignored. Most work either assumes access to a split as something of a black box or designates a problem that cannot be split by a standard choice of Γ – say the imaginary axis or the unit circle – as ill-posed. The latter is the case in work of Malyshev [95] and Bai, Demmel, and Gu [7], for example. Intuitively, we might expect that a random choice of Γ will work with high probability, an idea explored in a technical report of Ballard, Demmel, and Dumitriu [11], though this is ultimately not rigorous enough to be useful in practice. Put simply, the lack of general answers to this challenge of divide-and-conquer explains its limited use, for both the standard and generalized eigenvalue problems.

Even if a splitting strategy is found, setting this first challenge aside momentarily, we still need to decide how to compute spectral projectors.

Divide-and-Conquer Challenge Two

How do we efficiently and stably compute the spectral projectors P_R and P_L once a subset $S \subset \Lambda(A, B)$ has been identified? Can this be done without matrix inversion?

Unlike the first challenge, this one has many potential answers in the literature. Hence, the task here is to find a method that balances efficiency and stability, where the latter is promoted by avoiding inversion (as we discuss further in Section 1.5).

These two challenges are inherently linked insofar as a pseudospectral guarantee like $\Lambda_\epsilon(A, B) \cap \Gamma = \emptyset$ provides a benchmark for how accurately P_R and P_L must be computed. That is, if the computed subproblems are within ϵ of their exact values, we will be guaranteed that the divide-and-conquer process will not break down.[7] This is a consequence of the following lemma.

[7] Here, a breakdown corresponds to a situation in which one of the computed subproblems has eigenvalues in the wrong region of $\mathbb{C}$.

Lemma 1.3.5. *Let P_R be a spectral projector for a regular matrix pencil (A, B) and let P_L be the projector onto the corresponding left deflating subspace. Let $U_L, U_R \in \mathbb{C}^{n\times k}$ be matrices whose orthonormal columns span the ranges of P_L and P_R respectively. Then*

$$\Lambda_\epsilon(U_L^H AU_R, U_L^H BU_R) \subseteq \Lambda_\epsilon(A, B).$$

Proof. If $z \in \Lambda_\epsilon(U_L^H AU_R, U_L^H BU_R)$ then there exists a unit vector $u \in \mathbb{C}^k$ such that

$$||U_L^H (A - zB)U_R u||_2 = ||(U_L^H AU_R - zU_L^H BU_R)u||_2 \leq \epsilon(1 + |z|). \tag{1.31}$$

Let $y = U_R u \in \mathbb{C}^n$. Since U_R has orthonormal columns $||y||_2 = ||U_R u||_2 = ||u||_2 = 1$. Moreover, y is in the right deflating subspace range(P_R), which means $(A - zB)y$ belongs to the corresponding left deflating subspace range(P_L). Since the columns of U_L are an orthonormal basis for this subspace, we conclude

$$||(A - zB)y||_2 = ||U_L^H (A - zB)y||_2 \leq \epsilon(1 + |z|) \tag{1.32}$$

and therefore $z \in \Lambda_\epsilon(A, B)$. □

Note that the proof of Lemma 1.3.5 requires that U_L and U_R have orthonormal columns, highlighting why it is insufficient to simply compute P_R and P_L.

To overcome the first challenge of divide-and-conquer we might ask: in what situations should spectral bisection be easy? Intuitively, a pencil (A, B) satisfying the assumptions of Theorem 1.2.14 should be amenable to divide-and-conquer, particularly if $\kappa_V(A, B)$ is not too large and gap(A, B) is not too small. In this setting – which we refer to as *well-behaved* – Bauer-Fike implies that the pseudospectra of (A, B) cling tightly to its eigenvalues, which are well-separated, and a randomized splitting strategy is likely to perform well.

Unfortunately, we are not guaranteed to always encounter well-behaved inputs in practice. To work around this, we take a cue from RandNLA, randomly perturbing the initial pencil (A, B) to obtain $(\widetilde{A}, \widetilde{B}) = (A + \gamma G_1, B + \gamma G_2)$, where $0 < \gamma < \frac{1}{2}$ is a tuning parameter and G_1 and G_2 are independent, complex Ginibre random matrices.

Definition 1.3.6. A random matrix $G \in \mathbb{C}^{n \times n}$ is *complex Ginibre* if its entries are independent and identically distributed (i.i.d.) with distribution $\mathcal{N}_{\mathbb{C}}(0, \frac{1}{n})$ – i.e., standard complex Gaussian with mean zero and variance $\frac{1}{n}$.

Given the perturbation results discussed in Section 1.2, this may seem counterintuitive. While it is possible that certain perturbations will dramatically change the spectral information of the input pencil (as in Example 1.2.2), these are ultimately unlikely to occur randomly. Instead, the random perturbation has a regularizing effect on (A, B) with high probability, which we characterize below. We argue that this regularization implies a splitting strategy for divide-and-conquer that will succeed with high probability when applied to $(\widetilde{A}, \widetilde{B})$.

Regularization via Randomization

With high probability, we have the following:

1. $(\widetilde{A}, \widetilde{B})$ is regular and diagonalizable.
2. $\widetilde{A}$ and $\widetilde{B}$ are minimally well-conditioned.
3. $\kappa_V(\widetilde{A}, \widetilde{B}) < \text{poly}(n, \gamma^{-1})$ and $\text{gap}(\widetilde{A}, \widetilde{B}) > \text{poly}(n^{-1}, \gamma)$.

The first two of these follow immediately[8] from a recent probabilistic singular value bound due to Banks et al. [17, Lemma 3.3], which we restate below. The third point is rigorously proved in Chapter 2.

Lemma 1.3.7 (Banks et al. 2021)**.** *Let G be an $n \times n$ complex Ginibre matrix. Then for any $A \in \mathbb{C}^{n \times n}$ and any $\gamma, t > 0$, $\mathbb{P}\left[\sigma_n(A + \gamma G) < t\right] \leq (nt/\gamma)^2$.*

The idea to use a random perturbation to regularize a matrix or pencil can be considered a linear algebra extension of the landmark smoothed analysis work of Spielman and Teng [120]. Lemma 1.3.7 is characteristic of the results in this branch of Random

[8] For point one, we note that $(\widetilde{A}, \widetilde{B})$ almost surely has a full set of distinct eigenvalues.

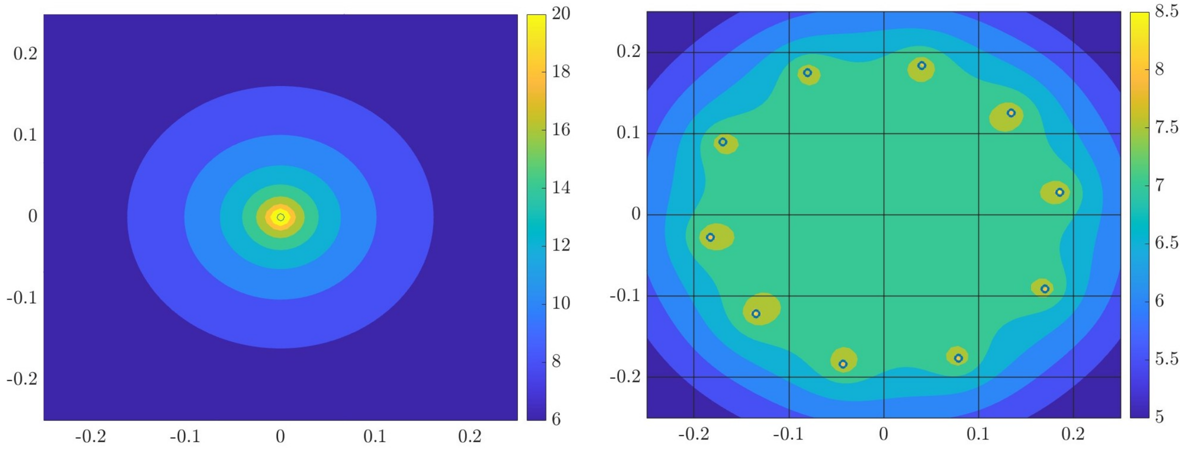
Matrix Theory, which probabilistically bound the norms, singular values, and condition numbers of a matrix under a variety of random perturbations [3, 54, 116]. In general, these bootstrap off of similar results for random matrices themselves, which have been well-studied since their introduction a century ago [147]. Importantly, distributions and probabilistic bounds for the eigenvalues/singular values of a variety random matrix ensembles are known [8, 47–49, 113, 132].

We are of course not the first to consider randomizing the inputs of divide-and-conquer. Again, we take inspiration here from Banks et al. [16], who applied the same high-level framework to the standard eigenvalue problem and, crucially, showed that it could diagonalize any matrix in fewer than $O(n^3)$ operations. They established the regularizing behavior described above for individual matrices and coined the term *pseudospectral shattering* to describe it,[9] a name derived from the fact that a random grid covering the pseudospectrum of the perturbed matrix (or pencil, as we will see) separates its disjoint components and the eigenvalues they contain into separate grid boxes with high probability. While their eigenvector condition number guarantee – i.e., that the perturbed matrix has minimally well-conditioned eigenvectors – is critical to bounding the pseudospectra, it can also be interpreted as a solution to a conjecture of Davies [36].

An example of pseudospectral shattering for matrix pencils is given in Figure 1.3. These plots capture the regularizing behavior described above and imply a straightforward method for splitting the spectrum in divide-and-conquer: simply divide with the grid lines covering $\Lambda_\epsilon(\widetilde{A}, \widetilde{B})$. Importantly, these grid lines are both well-separated from a certain pseudospectrum and easy to search over in pursuit of significant splits.

Of course, these plots also remind us that regularization does not necessarily preserve eigenvalues. In this case, $(\widetilde{A}, \widetilde{B})$ is diagonalizable while (A, B) is not; hence, the perturbation has both changed the Kronecker canonical form of the pencil and significantly

[9] While Banks et al. initially considered Ginibre perturbations [16, 17], they have subsequently generalized this work to other random ensembles [15].



(a) Pseudospectra of (A, B)

(b) Pseudospectra of $(\widetilde{A}, \widetilde{B})$

Figure 1.3. Pseudospectra of a pencil (A, B) before and after perturbation. Here, A is a 10×10 Jordan block, B is the identity matrix, and the perturbed matrices are $\widetilde{A} = A + 10^{-7}G_1$ and $\widetilde{B} = B + 10^{-7}G_2$ for G_1 and G_2 two independent, complex Gaussian matrices. Once again eigenvalues are plotted with open circles. A grid that shatters the 10^{-8}-pseudospectrum of $(\widetilde{A}, \widetilde{B})$ is superimposed over plot (b).

moved its eigenvalues. Nevertheless, a diagonalization of $(\widetilde{A}, \widetilde{B})$ *will* serve as a backward-stable diagonalization of (A, B) here, since $||A - \widetilde{A}||_2$ and $||B - \widetilde{B}||_2$ are small.

In Chapter 2 we rigorously generalize pseudospectral shattering to matrix pencils, thereby codifying theoretically the behavior discussed in this section and depicted in Figure 1.3. The remaining challenge of divide-and-conquer is addressed in Chapter 3, which presents a high-level framework for computing spectral projectors without relying on matrix inversion. In the subsequent chapters, we combine these results to pose and analyze two versions of divide-and-conquer for the generalized eigenvalue problem.

1.4 Efficiency and Fast Linear Algebra

In the remainder of this chapter, we make precise our notions of efficiency and stability, thereby clarifying the advantages of divide-and-conquer suggested by the previous sections. The bottom line (to keep in mind throughout) can be summarized as follows: our algorithm is built primarily on QR and matrix multiplication, both of which have fast and communication-optimal implementations with strong stability guarantees, therefore promoting efficiency and stability simultaneously. We begin here with efficiency and leave numerical stability to Section 1.5.

Throughout this thesis, the efficiency of an algorithm is primarily measured by its asymptotic complexity – i.e., the number of arithmetic operations it performs, expressed as $O(f(n))$ for $f(n)$ a function of the problem size n (here the size of the pencil (A, B)). In this framework, one algorithm is faster than another if its asymptotic complexity is smaller, meaning it requires fewer arithmetic operations in the limit $n \to \infty$. The computational complexity of a basic linear algebra operation – in our case matrix pencil diagonalization – is the asymptotic complexity of the fastest algorithm that performs it.

Despite decades of research in numerical linear algebra, the complexity of many basic matrix operations is not known. Matrix multiplication is the prototypical example here. While the standard multiplication algorithm requires $O(n^3)$ operations, a host of sub-$O(n^3)$ methods have been developed over the years, beginning in 1969 with the pioneering work of Strassen [127] and continuing to the present [5, 32, 117, 145]. At the time of writing, the fastest known algorithm has complexity $O(n^{2.371552})$ and is due to Williams, Xu, Xu, and Zhou [146]. In the meantime, the best known lower bound for matrix multiplication's complexity has remained at $\Omega(n^2)$, a simple consequence of the fact that any matrix multiplication algorithm must read $2n^2$ entries. It remains a major problem in the theory of computing to either show that this lower bound can be achieved or replace it with something tighter.

Cumulatively, algorithms that multiply $n \times n$ matrices in sub-$O(n^3)$ operations are referred to as *fast matrix multiplication*. Algorithms for other linear algebra computations built on top of these routines comprise what is often called *fast linear algebra*. Importantly, many such computations – including QR, LU, least squares, and more – can be implemented to have the same complexity as the matrix multiplication subroutine used [38, 72]. In an abstract sense, these computations are said to have *matrix multiplication time* implementations – i.e., they can be done by an algorithm with complexity equal to that of matrix multiplication, supposing this theoretical complexity was known and an algorithm for achieving it were available.

In our case, we focus on the slightly slower class of algorithms that run in *nearly* matrix multiplication time.

Definition 1.4.1. An algorithm that computes with $n \times n$ matrices runs in *nearly matrix multiplication time* if its asymptotic complexity is $O(\text{polylog}(n)T_{\text{MM}}(n))$, where $T_{\text{MM}}(n)$ is the complexity of $n \times n$ matrix multiplication. A linear algebra computation is doable in nearly matrix multiplication time if it can be executed by such an algorithm.

The allure of performing much of linear algebra in (nearly) matrix multiplication time should be tempered with some caution. For one, we might ask whether fast matrix multiplication exhibits the same stability as the traditional algorithm. While this is essentially the case (as we discuss in the next section), it is still not necessarily clear that reducing asymptotic complexity can offer real benefits in practice. The answer hinges on a classic question in numerical analysis: how large are the constants hidden by the big-O notation? For many of the most recent fast matrix multiplication routines, the constant suppressed by big-O is enormous; as a result, they are much slower than standard matrix multiplication on currently attainable problem sizes and therefore not widely used (if used at all). In contrast, early entries to the fast matrix multiplication library – in particular Strassen's $O(n^{\log_2(7)})$ algorithm – provide actual gains on realistic problems.

Since we are primarily interested in complexity as a measure of efficiency, we present "fast" versions of divide-and-conquer that are compatible with fast matrix multiplication. As a theoretical exercise, this ultimately implies that matrix pencil diagonalization can be done in nearly matrix multiplication time. In practice, however, any implementation of these methods is likely to use, at most, the simpler fast matrix multiplication routines.

Arithmetic operations are of course only one way to measure the efficiency of an algorithm. Given that we intend to apply the highly parallel divide-and-conquer approach, we might instead focus on communication costs, both between levels of memory hierarchy and/or between multiple processors. The former applies when an algorithm is run on a problem that exceeds available fast memory, in which case an input can only be accessed in pieces. The latter arises when information must be passed between processors as part of an explicitly parallel implementation. In either case, associated costs can be quantified in terms of M, the size of the fast (or local) memory.

Efforts to bound communication costs in numerical linear algebra[10] began with dense matrix multiplication in work of Hong and Kung [76] and Irony, Toledo, and Tiskin [81]. Ballard, Demmel, Holtz, and Schwartz [13] subsequently established the general lower bound of $\Omega(\#\text{ arithmetic operations}/\sqrt{M})$ for a variety of algorithms in linear algebra, including general $O(n^3)$ eigensolvers. A variant[11] of the divide-and-conquer approach presented in this thesis was shown to achieve this lower bound in a technical report of Ballard, Demmel, and Dumitriu [11], a consequence of the fact that its primary building blocks are QR and matrix multiplication. We use the same building blocks here, albeit fast linear algebra versions. Nevertheless, these fast alternatives have been shown to exhibit communication optimal implementations themselves [14], meaning our version of divide-and-conquer may achieve $O(\text{polylog}(n)T_{\text{MM}}(n)/\sqrt{M})$ communication cost, though this has not been explored rigorously.

The motivation for considering communication here is intuitive. Divide-and-conquer is both highly parallel and primed for memory-constrained applications, where its splitting power can be used to reduce a problem to fit in fast memory, if necessary. In fact, this is a much more realistic stopping criteria for the divide-and-conquer process than running the algorithm to 1×1 subproblems. Moreover, as arithmetic operations grow

[10] For a detailed discussions of results in this branch of NLA, which seeks communication-optimal or communication-avoidant algorithms, see the thesis of Ballard [9] or the survey paper [10].

[11] Built on classic $O(n^3)$ matrix multiplication and using a different splitting strategy than we develop in the next chapter.

Table 1.1. Best-known efficiency bounds for the standard QZ algorithm and randomized divide-and-conquer. The complexity for the latter is Theorem 1.6.1. Here, $T_{\mathrm{MM}}(n)$ is the complexity of $n \times n$ matrix multiplication and M is the size of available fast/local memory.

Algorithm	Complexity	Communication Cost
QZ	$O(n^3)$ [102]	$\Omega(n^3/\sqrt{M})$ [13]
Divide-and-Conquer	$O(\mathrm{polylog}(n) T_{\mathrm{MM}}(n))$	$O(n^3/\sqrt{M})$ [11]

cheaper, communication costs increasingly dominate run times in practice. In spite of this, communication optimality is not our primary focus, as the work discussed above has already established divide-and-conquer as the best choice in this setting. Our contribution instead is to show that divide-and-conquer is similarly optimal in terms of pure arithmetic operations. In tandem, these results demonstrate the flexibility of the divide-and-conquer strategy.

We summarize the efficiency bounds discussed in this section, for both QZ and divide-and-conquer, in Table 1.1. Note that the complexity and communication bounds for divide-and-conquer are not currently known to be achievable with the same implementation (though this is likely the case).

1.5 Numerical Stability and Inverse-Free Eigensolvers

We turn now to numerical stability. Since QZ is well-known to be backward-stable (as defined below), establishing stability for divide-and-conquer is critical; a fast version of the algorithm is ultimately useless if it is unreliable in a floating-point setting. With this in mind, we consider here classical definitions of numerical stability as applied to fast linear algebra. Once again, stability stems from the decision to avoid matrix inversion. Accordingly, we also dedicate part of this section to commenting on the efficacy of inverse-free algorithms as an approach to eigenvalue problems.

We begin by outlining a model for finite-precision arithmetic, which we use to define our notions of stability and will return to in Chapter 6. For our purposes, we assume a floating-point setting where

$$fl(x \circ y) = (x \circ y)(1+\Delta), \quad |\Delta| \leq \mathbf{u} \tag{1.33}$$

for basic operations $\circ \in \{+, -, \times, \div\}$ and a machine precision $\mathbf{u}(\varepsilon, n)$, which is a function of the desired accuracy ε and problem size n. As is traditional, we also assume that a similar error bound applies to $\sqrt{\cdot}$. This is a standard formulation for finite-precision computations (see for example [74]). Given $\mathbf{u}$, the number of bits of precision required to achieve (1.33) is $\log_2(1/\mathbf{u})$.

If $\text{alg}(x)$ is a finite-precision implementation of an algorithm that computes $f(x)$, we distinguish between two measures of accuracy,[12] which are also standard.

1. **Forward Error:** $||\text{alg}(x) - f(x)||$.
2. **Backward Error:** $||x - \widetilde{x}||$ for $\widetilde{x}$ satisfying $\text{alg}(x) = f(\widetilde{x})$.

Normwise stability for alg can be defined by bounding these errors in terms of $\mathbf{u}$ and $\mu(n)$, the latter a (small) polynomial in n.

Definition 1.5.1. alg is *forward-stable* if $||\text{alg}(x) - f(x)|| \leq \mu(n)\mathbf{u}||f(x)||$.

Definition 1.5.2. alg is *backward-stable* if $\text{alg}(x) = f(\widetilde{x})$ and $||x - \widetilde{x}|| \leq \mu(n)\mathbf{u}||x||$.

As defined here, forward stability is incredibly strict, implying that accuracy is independent of conditioning. For this reason, backwards stability – which suggests alternatively that alg computes f exactly on a nearby input – is typically considered the gold standard in numerical linear algebra. Assuming access to an appropriate condition number $\kappa_f(x)$, classical perturbation theory implies that a backward-stable algorithm satisfies a forward error bound

$$||\text{alg}(x) - f(x)||_2 \leq \kappa_f(x)\mu(n)\mathbf{u}||f(x)||. \tag{1.34}$$

[12] Here, $||\cdot||$ stands for any relevant norm.

Efforts to quantify the stability of fast linear algebra (in the framework outlined above) began with fast matrix multiplication [26, 27, 42, 73]. Work of Demmel, Dumitriu, Holtz, and Kleinberg [39] eventually established that any[13] fast matrix multiplication routine has an implementation (of equal complexity) that is *nearly* forward-stable. If $C = AB$ and C_{comp} is the computed product, this is characterized by the following:

$$||C_{\text{comp}} - C||_2 \leq \mu(n)\mathbf{u}||A||_2||B||_2. \tag{1.35}$$

In this sense, fast matrix multiplication is stable. Note that (1.35) is (possibly much) weaker than Definition 1.5.1, as $||AB||_2 \leq ||A||_2||B||_2$.

In subsequent work, Demmel, Dumitriu, and Holtz extended this stability analysis to fast linear algebra more broadly [38]. Of particular note here, they demonstrated that fast QR satisfies a mixed forward/backward stability bound (to be defined precisely in Chapter 6). For other computations, they defined and proved *logarithmic stability*, which can be interpreted as a relaxation of classical backward stability.

Definition 1.5.3. alg is *logarithmically-stable* if $\text{alg}(x) = f(\widetilde{x})$ and

$$||x - \widetilde{x}|| \leq \kappa_f(x)^{\chi(n)}\mu(n)\mathbf{u}||x||$$

for $\chi(n)$ a polynomial in $\log(n)$.

In place of Definition 1.5.3, Demmel, Dumitriu, and Holtz defined logarithmic stability in terms of the forward error bound

$$||\text{alg}(x) - f(x)|| \leq \kappa_f(x)^{\chi(n)+1}\mu(n)\mathbf{u}||f(x)||. \tag{1.36}$$

We justify borrowing terminology here by noting that any algorithm satisfying Definition 1.5.3 also satisfies (1.36). Comparing these forward-error bounds yields Table 1.2,

[13] Technically, their analysis only covers a certain class of recursive matrix multiplication algorithms, though this includes all of the most popular fast matrix multiplication routines (as well as the current fastest-known algorithm).

Table 1.2. Precision required for an algorithm of a given (normwise) stability to compute $f(x)$ with forward error $\varepsilon||f(x)||$. As in Definition 1.5.3, $\mu(n)$ and $\chi(n)$ are polynomials in n and $\log(n)$, respectively, and $\kappa_f(x)$ is a condition number.

Stability	Required $\mathbf{u}$	Required bits of precision $\log_2(1/\mathbf{u})$
Forward	$\varepsilon/\mu(n)$	$\log_2(1/\varepsilon) + \log_2(\mu(n))$
Backward	$\varepsilon/(\kappa_f(x)\mu(n))$	$\log_2(1/\varepsilon) + \log_2(\mu(n)) + \log_2(\kappa_f(x))$
Logarithmic	$\varepsilon/(\kappa_f(x)^{\chi(n)+1}\mu(n))$	$\log_2(1/\varepsilon) + \log_2(\mu(n)) + (\chi(n)+1)\log_2(\kappa_f(x))$

which shows the correspondence between these classical notions of stability and precision. We see here the origin of the name logarithmic stability: a logarithmically-stable algorithm requires a polylogarithmic increase in precision to obtain the same forward error as a backward-stable alternative.

Recalling that the QZ algorithm is backward-stable, we might hope to show that divide-and-conquer (implemented with fast matrix multiplication) is logarithmically stable, in essence trading an improvement in complexity for a slight, though manageable, loss of stability. We leave a full floating-point analysis to future work, though we anticipate that our approach can be implemented in a logarithmically stable way. Instead, we explore the stability of divide-and-conquer by deriving precision bounds for its main building blocks in Chapter 6 and presenting a handful of numerical examples in Chapter 4.

Our expectation that a logarithmically stable implementation exists is rooted primarily in the design choice to avoid matrix inversion. While it is possible to invert a matrix in $O(T_{\mathrm{MM}}(n))$ operations – meaning the use of inversion would not necessarily increase complexity – such implementations are only logarithmically stable, meaning they require the precision increase captured by Table 1.2 [38, Section 3] if used even once (and divide-and-conquer would require calling an inversion routine $O(\text{polylog}(n))$ times).

To see the drawback of building divide-and-conquer on top of fast inversion, we need only look to the work of Banks et al., whose divide-and-conquer approach to the standard eigenvalue problem uses inversion to compute spectral projectors via the matrix sign function. Accordingly, since they also make use of the fast linear algebra framework to obtain nearly-matrix-multiplication-time complexity, their algorithm is built on a logarithmically-stable black-box inversion algorithm [16, Definition 2.7]. While Banks et al. work hard to bound the error in their divide-and-conquer routine, the lack of stability inherent to fast inversion cannot be overcome; the result of their analysis is a weaker-than-logarithmic stability guarantee for the algorithm as a whole. Because the eigenvalue problem corresponding to the matrix A can be solved via the equivalent generalized eigenvalue problem (A, I), and because we anticipate that our approach admits a logarithmically stable implementation, we position our work as a potentially more stable (but no less efficient) alternative to the algorithm of Banks et al.

There is another dimension to avoiding inversion here that has to this point been overlooked. If B is invertible, the pencil (A, B) and the matrix $B^{-1}A$ have the same set of eigenvalues and (right) eigenvectors. This observation prompts a first, naive approach to the generalized eigenvalue problem: form the product matrix $B^{-1}A$ and apply any algorithm for the standard eigenvalue problem. While it would be convenient to simply dismiss this on the basis that B may very well be singular, the RandNLA approach taken here suggests that it could be viable. In particular, recall from Section 1.3 that randomly perturbing the input matrices guarantees with high probability that both $\widetilde{A}$ and $\widetilde{B}$ are minimally well-conditioned and therefore that the product matrix $\widetilde{B}^{-1}\widetilde{A}$ could be formed. In exact arithmetic, this provides an alternative – and arguably simpler – pathway to a nearly matrix multiplication time algorithm for diagonalizing (A, B).

Nevertheless, we argue that this is inadvisable from a stability perspective. In our approach, higher accuracy diagonalizations of (A, B) require increasingly small perturbations to the matrices. In this limit, the "minimally well-conditioned" guarantee for $\widetilde{B}$ will be poor if B is initially singular, implying that the decision to form $\widetilde{B}^{-1}\widetilde{A}$ will incur significant errors. This is captured by the following example.

Example 1.5.4. Construct a 1000×1000 pencil (A, B) by drawing A and B randomly, computing a singular value decomposition $B = U\Sigma V^H$, and setting $B = B - \sigma_{\min}(B)uv^H$, where u and v are the last columns of U and V, respectively. Let $\widetilde{A} = A + 10^{-10}G_1$ and $\widetilde{B} = B + 10^{-10}G_2$ for G_1, G_2 two independent, Gaussian matrices. Calling the intrinsic function `eig` in Matlab, we can find the eigenvalues of (A, B), $(\widetilde{A}, \widetilde{B})$, and $\widetilde{B}^{-1}\widetilde{A}$ via the QZ and QR algorithms. Ordering by modulus allows us to compute the absolute difference between corresponding eigenvalues of (A, B) and either $(\widetilde{A}, \widetilde{B})$ or $\widetilde{B}^{-1}\widetilde{A}$. Below, we record these errors for the five largest eigenvalues (and one draw of A and B with i.i.d. entries from $\mathcal{N}_{\mathbb{C}}(0, 2)$). Note that, by construction, (A, B) has an eigenvalue at infinity.

Eigenvalue	$\lvert\Lambda(\widetilde{A}, \widetilde{B}) - \Lambda(A, B)\rvert$	$\lvert\Lambda(\widetilde{B}^{-1}\widetilde{A}) - \Lambda(A, B)\rvert$
(largest) 1	Infinity	Infinity
2	4.68×10^{-9}	1.32×10^{-2}
3	3.20×10^{-10}	1.93×10^{-2}
4	1.82×10^{-10}	2.23×10^{-2}
5	2.62×10^{-10}	8.69×10^{-3}

The impact of forming the matrix $\widetilde{B}^{-1}\widetilde{A}$ is clear: because the perturbation to A and B is small, $\widetilde{B}$ is nearly singular and computing $\widetilde{B}^{-1}\widetilde{A}$, in double precision, meaningfully moves the eigenvalues away from those of (A, B). The same cannot be said for $(\widetilde{A}, \widetilde{B})$. Hence, if we hope to recover the original eigenvalues of (A, B), we must work with the latter. Note that this example uses neither divide-and-conquer nor fast inversion, meaning the loss of accuracy here is better than we might expect to see in practice.

In some sense, the decision to pursue an inverse-free approach to the generalized eigenvalue problem is an expression of traditional numerical analysis lore. To quote Nick Higham [74, Chapter 14]:

To most numerical analysts, matrix inversion is a sin.

Some may be skeptical: if high enough precision is available to us, is inverting a matrix really so problematic? We aim to show here that inversion truly is a stability bottleneck for generalized eigenvalue problems, both experimentally and in terms of theoretical precision bounds. To that end, we will return to versions of Example 1.5.4 throughout the thesis. In Chapter 4, we demonstrate that a divide-and-conquer approach based on forming $\widetilde{B}^{-1}\widetilde{A}$ cannot diagonalize (A, B) with high accuracy – or recover its eigenvalues – in standard double precision. Formal bounds are subsequently considered in Chapter 6.

1.6 Main Contributions

We are finally ready to state our main result.

Theorem 1.6.1. *There exists an exact arithmetic, inverse-free, and randomized algorithm that for any pencil (A, B) with $A, B \in \mathbb{C}^{n\times n}$ and any accuracy $\varepsilon < 1$ produces in nearly matrix multiplication time nonsingular matrices S, T and a diagonal matrix D such that*

$$||A - SDT^{-1}||_2 \le \varepsilon \ \text{ and } \ ||B - ST^{-1}||_2 \le \varepsilon$$

with probability at least $1 - O(\frac{1}{n})$.

The remaining chapters can be summarized as follows. Taken together, the first three provide a high-level outline of the proof of Theorem 1.6.1.

- Chapter 2 resolves challenge one of divide-and-conquer by generalizing pseudospectral shattering to matrix pencils.
- Chapter 3 answers the remaining challenge, presenting a strategy for computing spectral projectors that uses only QR and matrix multiplication.
- Chapter 4 combines the results of the previous two to state a provably successful, divide-and-conquer diagonalization algorithm, which we call Randomized Pencil Diagonalization or **RPD**.

- Chapter 5 considers a specialization of pseudospectral shattering and **RPD** to *definite* pencils (which we hold off on defining until then). Importantly, this chapter establishes that our divide-and-conquer approach can be formulated to exploit structure, if applicable.
- Finally, Chapter 6 provides the aforementioned precision bounds for a floating-point implementation.

We hope that this thesis not only establishes the efficacy of randomized divide-and-conquer for the generalized eigenvalue problem but also serves as a guidebook for developing (and analyzing) new variants of our approach.

1.6.1 Open Problems and Future Work

For interested readers, we discuss here a handful of open problems related to the results presented in this thesis.

Full Finite-Precision Analysis: The most immediate goal is to complete the floating-point analysis begun in Chapter 6. While we present precision bounds for the main building blocks of our approach there, a bound for the algorithm as a whole remains to be found. As noted in Section 1.4, a communication-oriented analysis of divide-and-conquer (with fast matrix multiplication) remains similarly open. Can it be shown rigorously, as we anticipate, that our algorithm is both communication-optimal and logarithmically stable?

Universal/Deterministic Pseudospectral Shattering: With the exception of Chapter 5, perturbations here are always complex Ginibre. Given work of Banks et al., which proved the essential building blocks of single-matrix shattering for real perturbations with absolutely continuous entries (under mild moment assumptions) [15], this begs the question: can pseudospectral shattering be established for other random matrix ensembles in the pencil case? Going a step further, can a similar deterministic result be found? Recent work of Bhattacharjee et al. [25] suggests the adjacency matrix of a pseudorandom graph as a candidate for deterministic perturbations.

Improvements to Perturbation Theory: As mentioned in Section 1.2, standard perturbation theory for the generalized eigenvalue problem is somewhat ill-equipped to explain the behavior of algorithms that pursue backward-stable diagonalizations. While interesting in its own right, producing sharper (and easier to use) bounds would provide better user guides for an implementation.

High-Performance Implementation: With the latter point in mind, we note that the experiments presented here were done with a model implementation – i.e., one that did not use fast matrix multiplication and was not explicitly parallelized. Hence, they can be interpreted as primarily a proof of concept. Together with the original pseudospectral shattering paper of Banks et al. (and other subsequent extensions, including by Sobczyk, Mladenović, and Luisier [119]), we believe this work represents a critical mass of sorts for randomized divide-and-conquer, justifying further implementation efforts, particularly from a high-performance computing perspective (see Appendix A). Note that eigensolvers are not included in the current development of RandLAPACK [103].

1.7 Miscellanea

To wrap up this chapter, we state a few general results from linear algebra and complex analysis, which we use throughout.

1.7.1 Singular Value Inequalities

We start with a handful of inequalities for singular values. The first concerns a product of matrices and follows easily from Courant-Fischer.

Lemma 1.7.1. *Let $A, B \in \mathbb{C}^{n \times n}$. Then for any $1 \leq i \leq n$,*

$$\sigma_n(A)\sigma_i(B) \leq \sigma_i(AB) \leq ||A||_2\sigma_i(B).$$

Next is the stability of singular values, a consequence of the stability of the Hermitian eigenvalue problem – i.e., Weyl's inequality [126, Corollary IV.4.9].

Lemma 1.7.2 (Stability of Singular Values)**.** *For any* $A, B \in \mathbb{C}^{n\times n}$

$$|\sigma_i(A) - \sigma_i(B)| \leq ||A - B||_2, \quad 1 \leq i \leq n.$$

Finally, we note that the spectral norm is often generalized to pencils by setting $||(A,B)||_2$ equal to the norm of the $n \times 2n$ matrix obtained by concatenating A and B. Throughout we use the loose upper bound $||(A,B)||_2 \leq ||A||_2 + ||B||_2$.

1.7.2 Möbius Transformations

As we will see, our divide-and-conquer procedure makes frequent use of Möbius transformations.

Definition 1.7.3. A *Möbius transformation* is a map $S : \mathbb{C} \to \mathbb{C}$ of the form.

$$S(z) = \frac{az+b}{cz+d}$$

for $a, b, c, d, \in \mathbb{C}$ satisfying $ad - bc \neq 0$.

Möbius transformations have a number of useful properties: they are conformal and map generalized circles to generalized circles, preserving orientation in the process. Moreover, any Möbius transformation is determined by its action on three points in $\mathbb{C}$ (including possibly a point at ∞). For more background, see the standard reference [2].

Of particular use here, note that the Möbius transformation

$$S(z) = \frac{z-(h-1)}{z-(h+1)} \tag{1.37}$$

maps the line $\text{Re}(z) = h$ to the unit circle, with $\{\text{Re}(z) < h\} \to \{|z| < 1\}$. Applying the Möbius transformation $S(z) = \frac{az+b}{cz+d}$ to (A, B) yields the pencil $(aA + bB, cA + dB)$, whose spectrum is exactly the image of $\Lambda(A, B)$ under S. This provides a cheap way to map the eigenvalues of (A, B), which can be easily undone via $S^{-1}(z) = \frac{dz-b}{-cz+a}$ if necessary.

Content Acknowledgement: Portions of this chapter are repurposed from a submitted work co-authored with James Demmel and Ioana Dumitriu:

- J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.

The dissertation author was the primary investigator and author of this paper.

Chapter 2

Generalized Pseudospectral Shattering

In this chapter, we extend the pseudospectral shattering work of Banks et al. [16] to the generalized eigenvalue problem. In doing so, we establish rigorously the regularizing effect of random (Ginibre) perturbations on the spectrum and pseudospectrum of a matrix pencil. We begin by defining this phenomenon formally.

Definition 2.0.1. The *shattering grid* $g = \mathrm{grid}(z_0, \omega, s_1, s_2)$ is the boundary of the $s_1 \times s_2$ lattice in the complex plane that has lower left corner $z_0 \in \mathbb{C}$ and consists of $\omega \times \omega$ boxes. The grid lines of g are parallel to either the real or the complex axis.

Definition 2.0.2. $\Lambda_\epsilon(A, B)$ is *shattered* with respect to the grid g if (1) $\Lambda_\epsilon(A, B) \cap g = \emptyset$ and (2) each eigenvalue of (A, B) belongs to a unique grid box of g.

Swapping $\Lambda_\epsilon(A, B)$ for $\Lambda_\epsilon(A)$ in Definition 2.0.2 yields the original definition of pseudospectral shattering for the standard eigenvalue problem. Note that the shattering grid g is not simply a net of points but rather the union of the vertical/horizontal lines connecting them. The picture to keep in mind throughout is that of Figure 1.3.

In the terminology established here, the original pseudospectral shattering result of Banks et al. can be stated as follows.

Theorem 2.0.3 (Banks et al. 2022). *Let $A \in \mathbb{C}^{n\times n}$ with $||A||_2 \leq 1$ and set $\widetilde{A} = A + \gamma G$ for G an $n \times n$ Ginibre matrix and $0 < \gamma < \frac{1}{2}$. Set $\omega = \frac{\gamma^4}{4n^5}$ and sample the point z uniformly at random from the $\omega \times \omega$ square in $\mathbb{C}$ with bottom left corner $-4-4i$. Construct the grid $g = grid(z, \omega, \lceil 8/\omega \rceil, \lceil 8/\omega \rceil)$. Then $\Lambda_\epsilon(\widetilde{A})$ is shattered with respect to g for $\epsilon = \frac{\gamma^5}{16n^9}$ with probability at least $1 - O(\frac{1}{n})$.*

The proof of Theorem 2.0.3 boils down to bounding the eigenvalue gap and eigenvector condition number of $\widetilde{A}$, while the norm assumption $||A||_2 \leq 1$ is imposed to provide easy bounds on the spectral radius of $\widetilde{A}$. These ingredients provide our road map: we will seek analogous bounds to prove our counterpart to Theorem 2.0.3.

To do this, we consider throughout the perturbed pencil

$$(\widetilde{A}, \widetilde{B}) = (A + \gamma G_1, B + \gamma G_2) \tag{2.1}$$

for G_1, G_2 independent Ginibre matrices and a tuning parameter $0 < \gamma < \frac{1}{2}$. We apply similar norm assumptions $||A||_2, ||B||_2 \leq 1$ for simplicity. In this case, however, normalizing a pencil does not change its eigenvalues; instead, we require an n^α scaling (with $\alpha > 0$) on $\widetilde{B}$ to obtain bounds on $\Lambda(\widetilde{A}, n^\alpha \widetilde{B})$. Of course, we will eventually have to pay the price for this scaling – i.e., if we compute the eigenvalues of $(\widetilde{A}, n^\alpha \widetilde{B})$ we will have to multiply by n^α to recover the eigenvalues of $(\widetilde{A}, \widetilde{B})$ – though this is essentially equivalent to the cost of normalizing in the single-matrix case.

As we will see, much of the subsequent analysis is done in terms of the product matrix $X = n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$, which has the same eigenvalues and right eigenvectors as $(\widetilde{A}, n^\alpha \widetilde{B})$. This is a theoretical tool only; to again underscore the numerical stability concerns discussed in Chapter 1, we do not recommend forming X in practice.

Guide to Chapter 2: In Section 2.1, we derive a handful of probabilistic singular value bounds and demonstrate how they can be used to control the spectrum and eigenvector conditioning of a perturbed (and scaled) pencil. Section 2.2 subsequently presents our main regularization result: a tail bound on the probability that $\text{gap}(\widetilde{A}, n^\alpha \widetilde{B})$ is not too small and $\kappa_V(\widetilde{A}, n^\alpha \widetilde{B})$ is not too large. We make use of this tail bound in Section 2.3 to prove pseudospectral shattering.

2.1 Singular Value Bounds

As discussed in Section 1.3, pseudospectral shattering, and therefore randomized divide-and-conquer, rests primarily on probabilistic singular value bounds from Random Matrix Theory. In addition to Lemma 1.3.7, which we re-state below for convenience, the single-matrix version makes use of the following. Lemma 2.1.1 is [17, Lemma 2.2] while Lemma 2.1.2 is [16, Corollary 3.3].

Lemma 1.3.7. *Let G be an $n \times n$ complex Ginibre matrix. Then for any $A \in \mathbb{C}^{n\times n}$ and any $\gamma, t > 0$, $\mathbb{P}\left[\sigma_n(A + \gamma G) < t\right] \leq (nt/\gamma)^2$.*

Lemma 2.1.1 (Banks et al. 2021)**.** *Let G be an $n \times n$ complex Ginibre matrix. Then for any $t > 0$, $\mathbb{P}\left[||G||_2 \geq 2\sqrt{2} + t\right] \leq 2e^{-nt^2}$.*

Lemma 2.1.2 (Banks et al. 2022)**.** *Let G be an $n \times n$ complex Ginibre matrix. Then for any $M \in \mathbb{C}^{n\times n}$ and any $\gamma, t > 0$, $\mathbb{P}\left[\sigma_{n-1}(M + \gamma G) < t\right] \leq 4(tn/\gamma)^8$.*

These results are equally useful here. Lemma 1.3.7, for example, implies that with high probability (and an appropriate choice of γ) both $\widetilde{A}$ and $\widetilde{B}$ are nonsingular, meaning the pencil $(\widetilde{A}, \widetilde{B})$ – and therefore also $(\widetilde{A}, n^\alpha \widetilde{B})$ – can be assumed to be regular. Almost surely these pencils have distinct eigenvalues and are therefore also diagonalizable, again with high probability. At the same time, Lemma 2.1.1 guarantees that the norms of $\widetilde{A}$ and $\widetilde{B}$ are not too large, assuming initially $||A||_2, ||B||_2 \leq 1$. Together, and without needing to adapt them in any way, these first two results imply a bound on $\Lambda_\epsilon(\widetilde{A}, n^\alpha \widetilde{B})$.

Lemma 2.1.3. *The eigenvalues of* $(\widetilde{A}, n^\alpha \widetilde{B})$ *are contained in* $B_3(0)$ *– the ball of radius three centered at the origin – with probability at least* $1 - \frac{n^{2-2\alpha}}{\gamma^2} - 2e^{-n}$.

Proof. Consider $X = n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$. Since X and $(\widetilde{A}, n^\alpha \widetilde{B})$ have the same eigenvalues, it is sufficient to bound the probability that $\Lambda(X)$ is not contained in $B_3(0)$. To do this, we note

$$\rho(X) \leq ||X||_2 = n^{-\alpha}||\widetilde{B}^{-1}\widetilde{A}||_2 \leq \frac{1}{n^\alpha \sigma_n(\widetilde{B})}||\widetilde{A}||_2 \leq \frac{1}{n^\alpha \sigma_n(\widetilde{B})}\left(||A||_2 + \gamma||G_1||_2\right). \quad (2.2)$$

Now $||A||_2 \leq 1$ so, conditioning on the event $||G_1||_2 \leq 4$, we have $||A + \gamma G_1||_2 \leq 3$. Thus, if $\Lambda(X) \not\subseteq B_3(0)$ and therefore $\rho(X) > 3$, we obtain $n^\alpha \sigma_n(\widetilde{B}) < 1$. Consequently,

$$\mathbb{P}\left[\Lambda(X) \not\subseteq B_3(0) \mid ||G_1||_2 \leq 4\right] \leq \mathbb{P}\left[n^\alpha \sigma_n(\widetilde{B}) < 1\right] \leq \frac{n^{2-2\alpha}}{\gamma^2}, \quad (2.3)$$

where the last inequality comes from Lemma 1.3.7. By Bayes' Theorem, we therefore have

$$\mathbb{P}\left[\Lambda(X) \not\subseteq B_3(0),\ ||G_1||_2 \leq 4\right] \leq \frac{n^{2-2\alpha}}{\gamma^2}. \quad (2.4)$$

At the same time, applying Lemma 2.1.1, we have

$$\mathbb{P}\left[\Lambda(X) \not\subseteq B_3(0),\ ||G_1||_2 > 4\right] \leq \mathbb{P}\left[||G_1||_2 > 4\right] \leq 2e^{-n(4-2\sqrt{2})^2} \leq 2e^{-n}. \quad (2.5)$$

Putting these two results together, we conclude

$$\mathbb{P}\left[\Lambda(X) \not\subseteq B_3(0)\right] \leq \frac{n^{2-2\alpha}}{\gamma^2} + 2e^{-n} \quad (2.6)$$

which completes the proof. □

With Lemma 2.1.3 in hand, we can now consider building a random grid g over $B_3(0)$, which (with an appropriate choice of α and γ) will contain every eigenvalue of $(\widetilde{A}, n^\alpha \widetilde{B})$ with high probability. Given a grid box size ω, we define the random grid $g = \text{grid}(z, \omega, \lceil 8/\omega \rceil, \lceil 8/\omega \rceil)$ for z a point drawn uniformly at random from the box in $\mathbb{C}$ with bottom left corner $-4-4i$ and side length ω. The construction here follows Banks et al. and is somewhat arbitrary; for convenience, we choose g so that it roughly covers the smallest box with integer side length that contains $B_3(0)$ (with some buffer space allowed).

We have yet to use Lemma 2.1.2. Below, we show that it can be bootstrapped into a bound on the $(n-1)^{\text{st}}$ singular value of $yI - X$ for any $y \in \mathbb{C}$ by way of Lemma 1.7.1. The purpose of the resulting Lemma 2.1.4 is not immediately clear; in the next section, we show that this result implies bounds on the eigenvector condition number of $(\widetilde{A}, n^\alpha \widetilde{B})$.

Lemma 2.1.4. *For any $t > 0$ and any $y \in \mathbb{C}$,*

$$\mathbb{P}\left[\sigma_{n-1}(yI - X) < t \mid ||G_2||_2 \leq 4\right] \leq 4\left(\frac{t(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8.$$

Proof. We begin by rewriting $yI - X$ as

$$yI - X = n^{-\alpha}\widetilde{B}^{-1}\left(n^\alpha y\widetilde{B} - A - \gamma G_1\right). \tag{2.7}$$

Using standard singular value inequalities, we then have

$$\sigma_{n-1}(yI - X) \geq \sigma_n\left(n^{-\alpha}\widetilde{B}^{-1}\right)\sigma_{n-1}\left(n^\alpha y\widetilde{B} - A - \gamma G_1\right). \tag{2.8}$$

Now conditioning on $||G_2||_2 \leq 4$, we have $||\widetilde{B}||_2 \leq 1 + 4\gamma$. Thus, $\sigma_n(n^{-\alpha}\widetilde{B}^{-1}) \geq \frac{1}{n^\alpha(1+4\gamma)}$ and therefore

$$\sigma_{n-1}(yI - X) \geq \frac{1}{n^\alpha(1+4\gamma)}\sigma_{n-1}\left(n^\alpha y\widetilde{B} - A - \gamma G_1\right). \tag{2.9}$$

Consequently, we observe

$$\mathbb{P}\left[\sigma_{n-1}(yI - X) < t \mid ||G_2||_2 \leq 4\right] \leq \mathbb{P}\left[\sigma_{n-1}\left(n^\alpha y\widetilde{B} - A - \gamma G_1\right) < tn^\alpha(1+4\gamma)\right]. \tag{2.10}$$

Setting $M = n^\alpha y\widetilde{B} - A$ and applying Lemma 2.1.2 (noting that M is independent of G_1 and that G_1 and $-G_1$ have the same distribution) we have

$$\mathbb{P}\left[\sigma_{n-1}(yI - X) < t \mid ||G_2||_2 \leq 4\right] \leq 4\left(\frac{t(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8 \tag{2.11}$$

for any $t > 0$. □

2.2 Spectral Regularization

In this section, we use the bounds derived above to prove spectral regularization for $(\widetilde{A}, n^{\alpha}\widetilde{B})$, which here means that with high probability both $\mathrm{gap}(\widetilde{A}, n^{\alpha}\widetilde{B}) > \mathrm{poly}(n^{-1}, \gamma)$ and $\kappa_V(\widetilde{A}, n^{\alpha}\widetilde{B}) < \mathrm{poly}(n, \gamma^{-1})$. Again, we work with $X = n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$. Accordingly, since we may assume that X is diagonalizable with a full set of distinct eigenvalues, we make use of the following condition numbers.

Definition 2.2.1. If λ_i is an eigenvalue of a matrix A with distinct eigenvalues and v_i and w_i are corresponding right/left eigenvectors normalized so that $w_i^H v_i = 1$, then the *condition number* of λ_i is

$$\kappa(\lambda_i) = ||v_i w_i^H||_2 = ||v_i||_2 ||w_i||_2.$$

Importantly, the condition numbers associated to the eigenvalues of X are related to $\kappa_V(\widetilde{A}, n^{\alpha}\widetilde{B})$ in the following way. If V is the eigenvector matrix of X (equivalently the right eigenvector matrix of $(\widetilde{A}, n^{\alpha}\widetilde{B})$) scaled so that each column is a unit vector, we have

$$\kappa_V(\widetilde{A}, n^{\alpha}\widetilde{B}) \leq \kappa_2(V) \leq ||V||_F ||V^{-1}||_F \leq \sqrt{n \sum_{i=1}^{n} \kappa(\lambda_i)^2}. \tag{2.12}$$

Hence, bounding $\kappa(\lambda_i)$ is equivalent to bounding $\kappa_V(\widetilde{A}, n^{\alpha}\widetilde{B})$. Note that (2.12) would *not* hold for a corresponding version of Definition 2.2.1 that uses the left and right eigenvectors of $(\widetilde{A}, n^{\alpha}\widetilde{B})$, as V^{-1} contains the left eigenvectors of X which are not necessarily left eigenvectors of $(\widetilde{A}, n^{\alpha}\widetilde{B})$.

Given Lemma 2.1.3, a bound on $\sum_i^n \kappa(\lambda_i)^2$ can be obtained from the following result (setting $S = B_3(0)$).

Lemma 2.2.2. *Let $\lambda_1, \dots \lambda_n$ be the eigenvalues of X. For any measurable set $S \subset \mathbb{C}$,*

$$\mathbb{E}\left[\sum_{\lambda_i \in S} \kappa(\lambda_i)^2 \,\middle|\, ||G_2||_2 \leq 4\right] \leq \left(\frac{(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^2 \frac{\mathrm{vol}(S)}{\pi}.$$

Proof. By Definition 1.2.3, we know for any $z \in \mathbb{C}$

$$\mathbb{P}\left[z \in \Lambda_\epsilon(X)\right] = \mathbb{P}\left[\sigma_n(zI - X) < \epsilon\right]. \tag{2.13}$$

Following the same argument made in Lemma 2.1.4, swapping Lemma 2.1.2 for Lemma 1.3.7, we therefore have

$$\mathbb{P}\left[z \in \Lambda_\epsilon(X) \mid ||G_2||_2 \leq 4\right] \leq \left(\frac{\epsilon(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^2. \tag{2.14}$$

Consider now the measurable set $S \subset \mathbb{C}$. Using (2.14), we observe

$$\begin{aligned}
\mathbb{E}\left[\mathrm{vol}\left(\Lambda_\epsilon(X) \cap S\right) \mid ||G_2||_2 \leq 4\right] &= \mathbb{E}\left[\int_S \mathbb{1}_{\{\lambda \in \Lambda_\epsilon(X) \mid ||G_2||_2 \leq 4\}}(z)dz\right] \\
&= \int_S \mathbb{P}\left[z \in \Lambda_\epsilon(X) \mid ||G_2||_2 \leq 4\right]dz \qquad (2.15) \\
&\leq \left(\frac{\epsilon(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^2 \mathrm{vol}(S).
\end{aligned}$$

Applying this result alongside [17, Lemma 3.2] and Fatou's Lemma for conditional expectation, we conclude

$$\begin{aligned}
\mathbb{E}\left[\sum_{\lambda_i \in S} \kappa(\lambda_i)^2 \,\middle|\, ||G_2||_2 \leq 4\right] &= \mathbb{E}\left[\liminf_{\epsilon \to 0} \frac{\mathrm{vol}(\Lambda_\epsilon(X) \cap S)}{\pi\epsilon^2} \,\middle|\, ||G_2||_2 \leq 4\right] \\
&\leq \liminf_{\epsilon \to 0} \frac{\mathbb{E}\left[\mathrm{vol}(\Lambda_\epsilon(X) \cap S \mid ||G_2||_2 \leq 4\right]}{\pi\epsilon^2} \qquad (2.16) \\
&\leq \left(\frac{(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^2 \frac{\mathrm{vol}(S)}{\pi},
\end{aligned}$$

which completes the proof. □

With this in mind, we now derive our main tail bound. Its proof generalizes a union bound argument used to obtain an equivalent result of Banks et al. [16, Theorem 3.6].

Theorem 2.2.3. *Define* $P(t,\delta) = \mathbb{P}\left[\kappa_V(\widetilde{A}, n^\alpha \widetilde{B}) < t,\ \mathrm{gap}(\widetilde{A}, n^\alpha \widetilde{B}) > \delta\right]$. *For any* $t, \delta > 0$ *we have*

$$P(t,\delta) \geq \left[1 - \frac{9(1+4\gamma)^2 n^{2\alpha+3}}{t^2\gamma^2} - 1296\delta^6\left(\frac{t(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right].$$

Proof. We condition on the events $||G_2||_2 \leq 4$ and $\Lambda_\epsilon(X) \subseteq B_3(0)$. Combining Lemma 2.1.1 and Lemma 2.1.3, we know these occur with probability at least $1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}$. Noting $\text{gap}(\widetilde{A}, n^\alpha \widetilde{B}) = \text{gap}(X)$ and $\kappa_V(\widetilde{A}, n^\alpha \widetilde{B}) = \kappa_V(X)$ for $X = n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$, define the events

- $E_{\text{gap}} = \{\text{gap}(X) < \delta\}$
- $E_\kappa = \{\kappa_V(X) > t\}$.

We are interested in bounding the probability of $E_{\text{gap}} \cup E_\kappa$. To do this, we construct a minimal $\frac{\delta}{2}$-net $\mathcal{N}$ covering $B_3(0)$, which exists with $|\mathcal{N}| \leq \frac{324}{\delta^2}$ (see for example [138, Corollary 4.2.11]). By construction

$$E_{\text{gap}} \subset \{|D(y,\delta) \cap \Lambda(X)| \geq 2 \text{ for some } y \in \mathcal{N}\} \tag{2.17}$$

where $D(y,\delta)$ is the ball of radius δ centered at y. Now if $D(y,\delta)$ contains two eigenvalues of X, [16, Lemma 3.5] implies $\sigma_{n-1}(yI - X) < \delta\kappa_V(X)$, where we note again that X is almost surely diagonalizable. Thus, if $\text{gap}(X) < \delta$ either $\sigma_{n-1}(yI - X) < \delta t$ for at least one $y \in \mathcal{N}$ or $\kappa_V(X) > t$. In other words, defining the event $E_y = \{\sigma_{n-1}(yI - X) < \delta t\}$ for $y \in \mathcal{N}$,

$$E_{\text{gap}} \subset E_\kappa \cup \bigcup_{y\in\mathcal{N}} E_y \implies E_{\text{gap}} \cup E_\kappa \subset E_\kappa \cup \bigcup_{y\in\mathcal{N}} E_y. \tag{2.18}$$

By a union bound, we then have

$$\mathbb{P}[E_{\text{gap}} \cup E_\kappa] \leq \mathbb{P}[E_\kappa] + |\mathcal{N}| \max_{y\in\mathcal{N}} \mathbb{P}[E_y]. \tag{2.19}$$

To bound $\mathbb{P}[E_\kappa]$, we first use (2.12) to obtain

$$\mathbb{P}[E_\kappa] \leq \mathbb{P}\left[t < \sqrt{n\sum_{i=1}^n \kappa(\lambda_i)^2}\right] = \mathbb{P}\left[\sum_{i=1}^n \kappa(\lambda_i)^2 > \frac{t^2}{n}\right]. \tag{2.20}$$

Applying Markov's inequality and Lemma 2.2.2 (recalling that we've conditioned on $\Lambda(X) \subseteq B_3(0)$) we then have

$$\mathbb{P}[E_\kappa] \leq \frac{n}{t^2}\mathbb{E}\left[\sum_{\lambda_i \in B_3(0)} \kappa(\lambda_i)^2\right] \leq \frac{9(1+4\gamma)^2 n^{2\alpha+3}}{t^2\gamma^2}. \tag{2.21}$$

Similarly, Lemma 2.1.4 implies

$$\mathbb{P}[E_y] \leq 4\left(\frac{\delta t(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8 \tag{2.22}$$

for all $y \in \mathcal{N}$. Putting everything together, we conclude

$$\mathbb{P}[E_{\text{gap}} \cup E_\kappa] \leq \frac{9(1+4\gamma)^2 n^{2\alpha+3}}{t^2\gamma^2} + 4\left(\frac{\delta t(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8 \frac{324}{\delta^2}, \tag{2.23}$$

which implies a lower bound for $\mathbb{P}\left[\kappa_V(X) < t,\ \text{gap}(X) > \delta \mid \Lambda(X) \subseteq B_3(0),\ ||G_2||_2 \leq 4\right]$ of

$$1 - \frac{9(1+4\gamma)^2 n^{2\alpha+3}}{t^2\gamma^2} - 1296\delta^6\left(\frac{t(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8. \tag{2.24}$$

Noting that $\mathbb{P}\left[\kappa_V(X) < t,\ \text{gap}(X) > \delta,\ \Lambda(X) \subseteq B_3(0),\ ||G_2||_2 \leq 4\right]$ bounds $\mathbb{P}[\kappa_V(X) < t,\ \text{gap}(X) > \delta]$ from below, we obtain the final result by Bayes' Theorem. □

Remark 2.2.4. Note that we could replace $\kappa_V(\widetilde{A}, n^\alpha \widetilde{B})$ in Theorem 2.2.3 with $\kappa_2(V)$ for any right eigenvector matrix V satisfying (2.12). In particular, as was used to derive the inequality, this applies to the scaling where each column of V has unit length.

2.3 Shattering

Combining the tail bound provided by Theorem 2.2.3 with Bauer-Fike (Theorem 1.2.9) yields our generalized pseudospectral shattering result.

Theorem 2.3.1. *Let $A, B \in \mathbb{C}^{n\times n}$ with $||A||_2 \leq 1$ and $||B||_2 \leq 1$. Let*

$$(\widetilde{A}, \widetilde{B}) = (A + \gamma G_1, B + \gamma G_2)$$

for G_1, G_2 two independent Ginibre matrices and $0 < \gamma < \frac{1}{2}$. Construct the grid $g = \text{grid}(z, \omega, \lceil 8/\omega \rceil, \lceil 8/\omega \rceil)$ with $\omega = \gamma^4/(4n^{\frac{8\alpha+13}{3}})$, where z is chosen uniformly at random from the square with bottom left corner $-4-4i$ and side length ω. Then $||\widetilde{A}||_2 \leq 3$, $||\widetilde{B}||_2 \leq 3$, and $\Lambda_\epsilon(\widetilde{A}, n^\alpha \widetilde{B})$ is shattered with respect to g for

$$\epsilon = \frac{\gamma^5}{64n^{\frac{11\alpha+25}{3}} + \gamma^5}$$

with probability at least $\left[1-\frac{82}{n}-\frac{531441}{16n^2}\right]\left[1-\frac{n^{2-2\alpha}}{\gamma^2}-4e^{-n}\right]$.

Proof. We condition on the following events: $n^\alpha\sigma_n(\widetilde{B})\geq 1$, $||G_1||_2\leq 4$, and $||G_2||_2\leq 4$. As noted above, these events occur with probability at least $1-\frac{n^{2-2\alpha}}{\gamma^2}-4e^{-n}$ and, following the argument made in Lemma 2.1.3, guarantee that $\Lambda(\widetilde{A},n^\alpha\widetilde{B})\subseteq B_3(0)$. Consequently, we know with probability one that each eigenvalue of $(\widetilde{A},n^\alpha\widetilde{B})$ is contained in a box of g. At the same time,

$$||\widetilde{A}||_2=||A+\gamma G_1||_2\leq ||A||_2+\gamma||G_1||_2\leq 3 \tag{2.25}$$

and similarly $||\widetilde{B}||_2\leq 3$.

Suppose now $\kappa_V(\widetilde{A},n^\alpha\widetilde{B})<\frac{n^{\alpha+2}}{\gamma}$, $\mathrm{gap}(\widetilde{A},n^\alpha\widetilde{B})>\gamma^4 n^{-\frac{8\alpha+13}{3}}$, and

$$\min_{\lambda_i\in\Lambda(X)}\mathrm{dist}_g(\lambda_i)>\frac{\omega}{4n^2}, \tag{2.26}$$

where $\mathrm{dist}_g(\lambda_i)=\min_{y\in g}|\lambda_i-y|$. By (2.24), we know the first two of these occur under our assumptions with probability at least

$$1-\frac{9(1+4\gamma)^2n^{2\alpha+3}}{\left(\frac{n^{\alpha+2}}{\gamma}\right)^2\gamma^2}-1296\left(\frac{\gamma^4}{n^{\frac{8\alpha+13}{3}}}\right)^6\left(\frac{\left(\frac{n^{\alpha+2}}{\gamma}\right)(1+4\gamma)n^{\alpha+1}}{\gamma}\right)^8, \tag{2.27}$$

which, applying $\gamma<\frac{1}{2}$, simplifies to $1-\frac{81}{n}-\frac{531441}{16n^2}$. Similarly, a simple geometric argument (using the fact that the eigenvalues of X are uniformly distributed in their grid boxes) implies

$$\mathbb{P}\left[\min_{\lambda_i\in\Lambda(X)}\mathrm{dist}_g(\lambda_i)>\frac{\omega}{4n^2}\right]\geq 1-\frac{1}{n}. \tag{2.28}$$

Thus, these events occur under our assumptions with probability at least $1-\frac{82}{n}-\frac{531441}{16n^2}$, which means by Bayes' Theorem that we have simultaneously $n^\alpha\sigma_n(\widetilde{B})\geq 1$, $||G_1||_2\leq 4$, $||G_2||_2\leq 4$, $\kappa_V(\widetilde{A},n^\alpha\widetilde{B})<n^{\alpha+2}/\gamma$, $\mathrm{gap}(\widetilde{A},n^\alpha\widetilde{B})>\gamma^4n^{-\frac{8\alpha+13}{3}}$, and (2.26) with probability at least $\left[1-\frac{82}{n}-\frac{531441}{16n^2}\right]\left[1-\frac{n^{2-2\alpha}}{\gamma^2}-4e^{-n}\right]$.

To complete the proof, we now show that these events guarantee shattering. To do this, we first observe that if $\mathrm{gap}(\widetilde{A},n^\alpha\widetilde{B})>\gamma^4n^{-\frac{8\alpha+13}{3}}$ then no two eigenvalues can share a grid box of g. At the same time, (2.26) implies that the ball of radius $\frac{\omega}{4n^2}$ around each eigenvalue does not intersect g. Therefore, it is sufficient to show that $\Lambda_\epsilon(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in these balls, which we can do by appealing to our version of Bauer-Fike for matrix pencils. In particular, recalling that we can replace $\kappa_2(V)$ with $\kappa_V(\widetilde{A}, n^\alpha \widetilde{B})$ by taking an infimum, Theorem 1.2.9 implies that $(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in balls of radius

$$r_\epsilon = \epsilon \kappa_V(\widetilde{A}, n^\alpha \widetilde{B}) ||n^{-\alpha}\widetilde{B}^{-1}||_2 \left(1 + \frac{\epsilon ||n^{-\alpha}\widetilde{B}^{-1}||_2 + ||n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}||_2}{1 - \epsilon ||n^{-\alpha}\widetilde{B}^{-1}||_2}\right). \tag{2.29}$$

Applying the bounds $||n^{-\alpha}\widetilde{B}^{-1}||_2 = (n^\alpha \sigma_n(\widetilde{B}))^{-1} \leq 1$, $||n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}||_2 \leq 3$, and finally $\kappa_V(\widetilde{A}, n^\alpha \widetilde{B}) \leq \frac{n^{\alpha+2}}{\gamma}$, we obtain shattering as long as

$$r_\epsilon \leq \epsilon \left(\frac{n^{\alpha+2}}{\gamma}\right)\left(\frac{4}{1-\epsilon}\right) \leq \frac{\omega}{4n^2} = \frac{\gamma^4}{16n^{\frac{8\alpha+19}{3}}} \tag{2.30}$$

or, equivalently, $\epsilon \leq \frac{\gamma^5}{64n^{\frac{11\alpha+25}{3}} + \gamma^5}$. □

Theorem 2.3.1 clarifies the impact of the n^α scaling. On one hand, increasing α drives the $\frac{n^{2-2\alpha}}{\gamma^2}$ term in the probability bound to zero, assuming γ is fixed. Said another way, a larger choice of α allows us to shrink γ without losing our guarantee of shattering, where a baseline

$$\gamma > n^{1-\alpha} \tag{2.31}$$

is needed to ensure that the probability in Theorem 2.3.1 is not vacuous. This is important, as we'd like to perturb our matrices as little as possible. Nevertheless, we pay a penalty for increasing α in ω and ϵ, both of which shrink as α increases. This trade-off reflects a fundamental geometric reality: the more we scale by, the closer the eigenvalues of $(\widetilde{A}, n^\alpha \widetilde{B})$ are driven to zero (and therefore to each other), meaning we'll need a finer grid and a tighter pseudospectrum to guarantee shattering.

As outlined in Chapter 1, Theorem 2.3.1 also resolves the first challenge of divide-and-conquer by providing the following splitting strategy, which can be applied to any pencil (A, B).

Spectral Bisection via Pseudospectral Shattering

1. Perturb and scale $(A, B) \to (\widetilde{A}, n^{\alpha}\widetilde{B})$.
2. Build a grid g that shatters $\Lambda_{\epsilon}(\widetilde{A}, n^{\alpha}\widetilde{B})$ [Theorem 2.3.1].
3. Split $\Lambda(\widetilde{A}, n^{\alpha}\widetilde{B})$ with the grid lines of g, computing subproblems to accuracy $O(\epsilon)$ in the spectral norm.
4. Once eigenvalues are found, multiply by n^{α}.

The third point in this outline follows from perturbation results discussed below.

2.3.1 Stability under Inversion

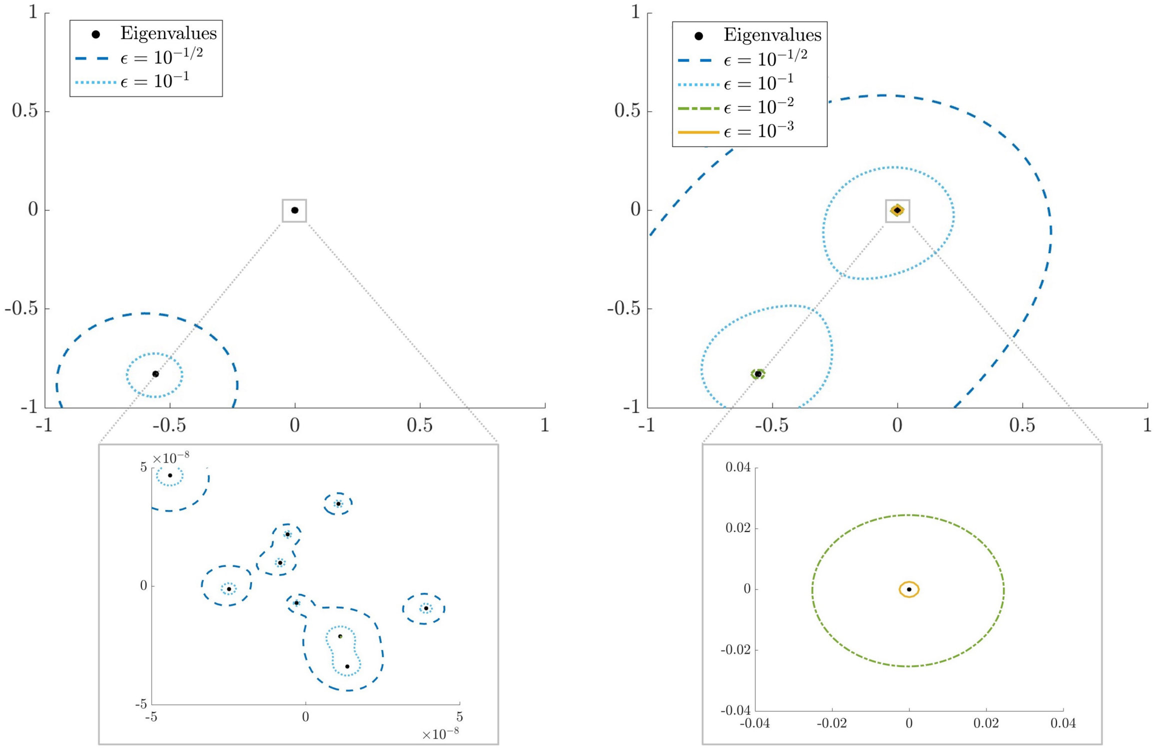
Recall that the tail bound Theorem 2.2.3 was proved for $(\widetilde{A}, n^{\alpha}\widetilde{B})$ by proving the same bound for the product matrix $X = n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$. In some sense, then, our proof of shattering only works with $(\widetilde{A}, n^{\alpha}\widetilde{B})$ as a pencil via the version of Bauer-Fike used. Swapping Theorem 1.2.9 for Theorem 1.2.4, the same proof implies the following pseudospectral shattering result for X.

Proposition 2.3.2. *Let $A, B \in \mathbb{C}^{n \times n}$ and let $0 < \gamma < \frac{1}{2}$. Suppose*

$$(\widetilde{A}, \widetilde{B}) = (A + \gamma G_1, B + \gamma G_2)$$

for G_1, G_2 two independent Ginibre matrices and let $X = n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$ for $\alpha > 0$. Then $\Lambda_{\epsilon}(X)$ is shattered with respect to the grid g (as defined in Theorem 2.3.1) for $\epsilon = \gamma^5/(16n^{\frac{11\alpha+25}{3}})$ with probability at least $\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right]$.

Proposition 2.3.2 implies an alternative strategy for diagonalizing (A, B): form the product matrix X and apply single-matrix, randomized divide-and-conquer as defined by Banks et al. [16, Algorithm EIG], using the grid g and the corresponding pseudospectral guarantee. While this is fairly straightforward – and even implies the same asymptotic complexity for the diagonalization as we derive in Chapter 4 – it is not viable in general



(a) $\Lambda_\epsilon(\widetilde{A}, \widetilde{B})$

(b) $\Lambda_\epsilon(\widetilde{B}^{-1}\widetilde{A})$

Figure 2.1. Pseudospectra of a 10×10 pencil $(\widetilde{A}, \widetilde{B})$ and its corresponding product matrix $\widetilde{B}^{-1}\widetilde{A}$. Here, $\widetilde{A} = A + 10^{-7}G_1$ and $\widetilde{B} = B + 10^{-7}G_2$ for A and B drawn randomly and B modified to be singular (without changing its remaining singular values). In both plots, $\widetilde{B}$ is scaled so that the maximum eigenvalue has modulus one. In addition, we provide a subplot focused around the origin to examine the pseudospectra around the small eigenvalues of $(\widetilde{A}, \widetilde{B})$, which correspond to finite eigenvalues of (A, B). In plot (a), we omit the pseudospectra for $\epsilon = 10^{-2}$ and $\epsilon = 10^{-3}$ since they are too close to the eigenvalues to be visible.

due to potential numerical instability. If B is poorly conditioned and γ is small, inverting $\widetilde{B}$ to form X will incur significant error in finite-precision arithmetic.

This phenomenon is illustrated in Figure 2.1, which plots the pseudospectra of both $(\widetilde{A}, \widetilde{B})$ and $\widetilde{B}^{-1}\widetilde{A}$ for a 10×10 example in which B is initially singular. The difference between the two is striking; while $\Lambda_\epsilon(\widetilde{A}, \widetilde{B})$ easily separates the eigenvalues, $\Lambda_\epsilon(\widetilde{B}^{-1}\widetilde{A})$ covers large regions of the complex plane, even for small ϵ. This indicates that the guarantee provided by Proposition 2.3.2 is more vulnerable in finite precision than Theorem 2.3.1.

Indeed, we prove in Chapter 6 that higher precision is needed to guarantee shattering for the product matrix X. As mentioned in Section 1.5, this increase in precision is our motivation for bypassing this approach to the problem and prioritizing inverse-free computations.

2.3.2 Perturbation Results

In the remainder of this chapter, we develop two perturbation results: one for pseudospectral shattering (Lemma 2.3.3) and one for individual eigenvalues and eigenvectors (Lemma 2.3.4). These follow closely [16, Lemmas 5.8 and 5.9] and provide benchmarks for how accurately a subproblem in divide-and-conquer must be computed (in the spectral norm) to preserve shattering.

In their proofs, we make use of the following consequence of Cauchy's integral formula [2]: if v, w are right/left eigenvectors of a matrix A corresponding to eigenvalue λ (and scaled so that $v^H w = 1$) then the rank-1 spectral projector vw^H can be expressed as

$$vw^H = \frac{1}{2\pi i} \oint_{\Gamma} (z - A)^{-1} dz \tag{2.32}$$

for Γ any closed, rectifiable curve that separates λ from $\Lambda(A)$. We also apply the ML inequality, which says

$$\left| \int_{\Gamma} f(z) dz \right| \leq l(\Gamma) \sup_{z \in \Gamma} |f(z)| \tag{2.33}$$

for any continuous function f and contour Γ of length $l(\Gamma)$.

Lemma 2.3.3. *Suppose (A, B) is regular and $\Lambda_\epsilon(A, B)$ is shattered with respect to a finite grid g. If $||A - A'||_2, ||B - B'||_2 \leq \eta < \epsilon$ then each eigenvalue of (A', B') shares a grid box with exactly one eigenvalue of (A, B) and $\Lambda_{\epsilon-\eta}(A', B')$ is also shattered with respect to g.*

Proof. If $z \in \Lambda_{\epsilon-\eta}(A', B')$ then z is an eigenvalue of a pencil (C, D) with $||A' - C||_2, ||B' - D||_2 \leq \epsilon - \eta$. In this case,

$$||A - C||_2 \leq ||A - A' + A' - C||_2 \leq ||A - A'||_2 + ||A' - C||_2 \leq \eta + \epsilon - \eta = \epsilon \tag{2.34}$$

and similarly $||B - D||_2 \leq \epsilon$, which implies $z \in \Lambda_\epsilon(A, B)$. Thus, $\Lambda_{\epsilon-\eta}(A', B') \subseteq \Lambda_\epsilon(A, B)$, which guarantees that $\Lambda_{\epsilon-\eta}(A', B')$ is also shattered with respect to g. To show that each eigenvalue of (A', B') shares a grid box with exactly one eigenvalue of (A, B), consider $A_t = A + t(A' - A)$ and $B_t = B + t(B' - B)$ for $t \in [0, 1]$. Since (A, B) is regular (and moreover $\epsilon < \sigma_n(B)$ since $\Lambda_\epsilon(A, B)$ is bounded), (A_t, B_t) continuously deforms the eigenvalues of (A, B) to eigenvalues of (A', B') while staying within $\Lambda_\eta(A, B) \subseteq \Lambda_\epsilon(A, B)$. Since $\Lambda_\epsilon(A, B)$ is shattered with respect to g and therefore no two eigenvalues of (A, B) share a grid box, this ensures that each eigenvalue of (A', B') belongs to a grid box with a unique eigenvalue of (A, B) □

Lemma 2.3.4. *Suppose (A, B) is regular and $\Lambda_\epsilon(A, B)$ is shattered with respect to a finite grid g with boxes of side length ω . If $||A - A'||_2, ||B - B'||_2 \leq \eta < \epsilon$ then for any right unit eigenvector v' of (A', B') there exists a right unit eigenvector v of (A, B) such that*

1. *The eigenvalue of (A', B') corresponding to v' shares a grid box of g with the eigenvalue of (A, B) that corresponds to v.*
2. $||v' - v|| \leq \frac{\sqrt{8}\omega}{\pi} \frac{\eta}{\epsilon(\epsilon-\eta)}(1 + ||B^{-1}A||_2)||B||_2$.

Proof. Let λ' be the eigenvalue of (A', B') corresponding to v'. By Lemma 2.3.3, λ' shares a grid box of g with a unique eigenvalue of λ of (A, B). Let v be the right unit eigenvector of (A, B) corresponding to λ. In addition, let w' and w be the left eigenvectors corresponding to v' and v respectively, normalized so that $w^H v = (w')^H v' = 1$. If Γ is the contour of the grid box containing both λ and λ', then by (2.32)

$$\begin{aligned} v'(w')^H - vw^H &= \frac{1}{2\pi i}\oint_\Gamma (z - (B')^{-1}A')^{-1} dz - \frac{1}{2\pi i}\oint_\Gamma (z - B^{-1}A)^{-1} dz \\ &= \frac{1}{2\pi i}\oint_\Gamma (z - (B')^{-1}A')^{-1} - (z - B^{-1}A)^{-1} dz. \end{aligned} \tag{2.35}$$

By the resolvent identity

$$\begin{aligned}(z-(B')^{-1}A')^{-1}-(z-B^{-1}A)^{-1} &= (z-(B')^{-1}A')^{-1}((B')^{-1}A'-B^{-1}A)(z-B^{-1}A)^{-1} \\ &= (zB'-A')^{-1}B'((B')^{-1}A'-B^{-1}A)(zB-A)^{-1}B \\ &= (zB'-A)^{-1}(A'-B'B^{-1}A)(zB-A)^{-1}B,\end{aligned} \tag{2.36}$$

so, applying this along with the ML inequality above, we have

$$||v'(w')^H - vw^H||_2 \leq \frac{2\omega}{\pi}\sup_{z\in\Gamma}||(zB'-A')^{-1}(A'-B'B^{-1}A)(zB-A)^{-1}B||_2. \tag{2.37}$$

Now $\Lambda_\epsilon(A,B)$ is shattered with respect to g and therefore does not intersect Γ, so we know $||(zB-A)^{-1}||_2 \leq \frac{1}{\epsilon(1+|z|)} \leq \epsilon^{-1}$ for all $z\in\Gamma$. Moreover, again by Lemma 2.3.3, $\Lambda_{\epsilon-\eta}(A',B')$ is shattered with respect to g, so the same argument implies $||(zB'-A')^{-1}||_2 \leq (\epsilon-\eta)^{-1}$ for all $z\in\Gamma$. Applying this to the previous inequality, we conclude

$$||v'(w')^H - vw^H||_2 \leq \frac{2\omega}{\pi}\frac{1}{\epsilon(\epsilon-\eta)}||A'-B'B^{-1}A||_2||B||_2. \tag{2.38}$$

Writing $B' = B+E$ for some $||E||_2 \leq \eta$ since $||B-B'||_2 \leq \eta$, we have

$$||A'-B'B^{-1}A||_2 \leq ||A'-A||_2 + ||EB^{-1}A||_2 \leq \eta(1+||B^{-1}A||_2) \tag{2.39}$$

which yields a final upper bound

$$||v'(w')^H - vw^H||_2 \leq \frac{2\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}(1+||B^{-1}A||_2))||B||_2. \tag{2.40}$$

Since without loss of generality we can assume $v^Hv' \geq 0$ (if this is not true we can just rotate v), and using the fact that $||v|| = ||v'|| = 1$, we complete the proof by observing

$$||v'-v||_2 = \sqrt{2-2v^Hv'} \leq \sqrt{2}||v'(w')^H - vw^H||_2. \tag{2.41}$$

This inequality is nontrivial (see the proof of [16, Lemma 5.8] for the details). Combining it with (2.40), we conclude that v is the desired eigenvector of (A,B) with $||v'-v|| \leq \frac{\sqrt{8}\omega}{\pi}\frac{\eta}{\epsilon(\epsilon-\eta)}(1+||B^{-1}A||_2)||B||_2$. □

Content Acknowledgement: This chapter is repurposed from a submitted work co-authored with James Demmel and Ioana Dumitriu:

- J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.

The dissertation author was the primary investigator and author of this paper.

Chapter 3

Inverse-Free Spectral Projectors

We return in this chapter to the question of computing spectral projectors. Our goal here is to address the second challenge of divide-and-conquer identified in Chapter 1: given an arbitrary set $S \subset \mathbb{C}$, what is the best method for computing the projectors P_R and P_L onto the right/left deflating subspaces of (A, B) corresponding to eigenvalues in S? Note that we assume in this chapter that (A, B) is both regular and diagonalizable.

Suppose we could transform (A, B) into a pencil (A_S, B_S) whose (right) eigenvectors are the same as (A, B) but whose eigenvalues are zero or one, depending on whether or not the corresponding eigenvalues of (A, B) belong to S. In this case, the task of finding spectral projectors is easy. Recalling (1.28), we have $P_R = B_S^{-1} A_S$. Similarly, transforming (A^H, B^H) to obtain a new (A_S, B_S), the left projector can be obtained as $P_L = (B_S^{-1} A_S)^H = A_S^H B_S^{-H}$, assuming S is symmetric with respect to the real axis.[1] The latter follows from the observation that left eigenvectors of $B^{-H} A^H$ associated to S, which recall are not necessarily left eigenvectors of either (A, B) or (A^H, B^H), span the corresponding left deflating subspace if B is invertible.

[1] Symmetry is necessary here since the eigenvalues of (A^H, B^H) are the complex conjugates of $\Lambda(A, B)$.

Given the discussion in Section 1.5, the matrix inversions here should raise a red flag. If (A_S, B_S) has only eigenvalues at or near zero and one, B_S *will* be invertible, though we have no guarantee that it will be well-conditioned. For this reason, once again, we do not advocate forming $B_S^{-1}A_S$ (or $A_S^H B_S^{-H}$). Instead, we demonstrate in the next chapter that a rank-revealing factorization of such a product can be computed implicitly via work of Ballard, Demmel and Dumitriu [12]. We can therefore set this concern aside for now.

Taking a step back, our initial question has been subsumed by a new one: how do we obtain the pencil (A_S, B_S) from (A, B)? We demonstrate here that this question naturally reduces to the problem of approximating the indicator function

$$\mathbb{1}_S(z) = \begin{cases} 1 & z \in S \\ 0 & z \in \mathbb{C}\backslash\overline{S} \\ \text{undefined} & z \in \partial S \end{cases} \tag{3.1}$$

with a rational function. Importantly, work of Benner and Byers [21] guarantees that any such approximation can be applied to (A, B) to obtain (A_S, B_S) using only QR and matrix multiplication. This yields a general, inverse-free framework for developing methods that approximate P_R and P_L. In this chapter, we present this framework rigorously and discuss a few examples, which we make use of later on.

In considering these, we should keep in mind the results of Chapter 2. That is, pseudospectral shattering indicates how accurately we need to compute P_R and P_L for divide-and-conquer. If $\Lambda_\epsilon(A, B)$ is shattered with respect to the grid g, Lemma 2.3.3 implies that the next subproblems must be computed to within spectral norm error $O(\epsilon)$ to obtain a similar pseudospectral guarantee for the next step. Since $\epsilon = \text{poly}(n^{-1}, \gamma)$ in Theorem 2.3.1, we therefore anticipate that any method for computing P_R and P_L will require high accuracy.

Guide to Chapter 3: In Section 3.1 we discuss our high-level strategy and the aforementioned work of Benner and Byers. Sections 3.2 and 3.3 then present a handful of example methods, including implicit repeated squaring (**IRS**) and two approaches based on the matrix sign function. The former is the method for computing spectral projectors used in Chapter 4, while the latter is relevant for Chapter 5.

3.1 High-Level Strategy

To derive an approach for computing (A_S, B_S) we might ask: what (inverse-free) operations can transform (A, B) into a pencil with the same eigenvectors but different eigenvalues? An answer to this question can be obtained from work of Benner and Byers [20,21], which defines abstractly an inverse-free arithmetic on matrix pencils. In this section, we present their work and outline the high-level strategy for computing (A_S, B_S) it implies.

The central definition here is the matrix relation $(B\backslash A)$. While $(B\backslash A)$ can be defined for matrices of arbitrary size, we will once again focus on the square case. Throughout, we can think of $(B\backslash A)$ as a representation of $B^{-1}A$ that neither requires B to be invertible nor incurs floating-point errors if B *is* invertible but ill-conditioned. To provide further intuition, we can also draw a comparison to the formal definition of the rational numbers.

Definition 3.1.1. The *matrix relation* on $\mathbb{C}^n$ associated to the pencil (A, B) is

$$(B\backslash A) = \left\{(x, y) \in \mathbb{C}^n \times \mathbb{C}^n : Ax = By\right\}.$$

Like rational numbers, the representation $(B\backslash A)$ is not unique. In fact, we can left multiply A and B by any matrix M whose null space only trivially overlaps with range$\left(\begin{bmatrix} A & -B \end{bmatrix}\right)$ without changing the relation. In particular, $(MB\backslash MA) = (B\backslash A)$ for any invertible matrix M. This indicates, as we might expect, that two pencils associated to the same matrix relation have the same set of right eigenvectors, which correspond to the elements $(v, \lambda v) \in (B\backslash A)$.

The key insight of Benner and Byers is to define an arithmetic on matrix relations, thereby providing a means of computing with $B^{-1}A$ implicitly. While their initial focus in [20] was products of the form $\prod_k B_k^{-1} A_k$, subsequent work [21] extended this framework to include addition. We summarize both operations below.

Definition 3.1.2 (Operations on Matrix Relations)**.** The *sum* and *product* of two matrix relations $(B_1 \backslash A_1)$ and $(B_2 \backslash A_2)$ are subsets of $\mathbb{C}^n \times \mathbb{C}^n$ defined as follows:

$$(B_2 \backslash A_2) + (B_1 \backslash A_1) = \left\{ (x, z) : \exists \;\; y_1, y_2 \;\text{ s.t. } \begin{pmatrix} A_1 & -B_1 & 0 & 0 \\ A_2 & 0 & -B_2 & 0 \\ 0 & I & I & -I \end{pmatrix} \begin{pmatrix} x \\ y_1 \\ y_2 \\ z \end{pmatrix} = 0 \right\}$$

$$(B_2 \backslash A_2)(B_1 \backslash A_1) = \left\{ (x, z) : \exists \;\; y \;\text{ s.t. } \begin{pmatrix} A_1 & -B_1 & 0 \\ 0 & A_2 & -B_2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 0 \right\}.$$

Given the abstract nature of Definition 3.1.2, it is not necessarily clear how to evaluate either addition or multiplication on a pair of matrix relations. Fortunately, Benner and Byers provide an answer to this as well [21, Theorems 2.3 and 2.7].

Theorem 3.1.3 (Benner and Byers 2006)**.** *Let $(B_1 \backslash A_1)$ and $(B_2 \backslash A_2)$ be two matrix relations with $A_1, A_2, B_1, B_2 \in \mathbb{C}^{n \times n}$. Suppose*

$$\operatorname{null}\left(\begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\right) = \operatorname{range}\left(\begin{bmatrix} -B_1 \\ A_2 \end{bmatrix}\right) \quad \text{and} \quad \operatorname{null}\left(\begin{bmatrix} U_1 & U_2 \end{bmatrix}\right)) = \operatorname{range}\left(\begin{bmatrix} -B_1 \\ B_2 \end{bmatrix}\right).$$

Then

$$(B_2 \backslash A_2)(B_1 \backslash A_1) = ((Q_2 B_2) \backslash (Q_1 A_1))$$

and

$$(B_2 \backslash A_2) + (B_1 \backslash A_1) = ((U_2 B_2) \backslash (U_1 A_1 + U_2 A_2)).$$

We can gain some intuition for Theorem 3.1.3 by considering what would happen if the matrices involved were invertible. In that case, $\operatorname{null}\left(\begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\right) = \operatorname{range}\left(\begin{bmatrix} -B_1 \\ A_2 \end{bmatrix}\right)$ implies $-Q_1 B_1 + Q_2 A_2 = 0$, and therefore $A_2 B_1^{-1} = Q_2^{-1} Q_1$. Hence, it is easy to see

$$B_2^{-1} A_2 B_1^{-1} A_1 = B_2^{-1} Q_2^{-1} Q_1 A_1 = (Q_2 B_2)^{-1} Q_1 A_1. \tag{3.2}$$

Similarly, null $\left(\begin{bmatrix} U_1 & U_2 \end{bmatrix}\right) =$ range $\left(\begin{bmatrix} -B_1 \\ B_2 \end{bmatrix}\right)$ implies $B_1^{-1} = B_2^{-1}U_2^{-1}U_1$ and therefore

$$B_2^{-1}A_2 + B_1^{-1}A_1 = B_2^{-1}A_2 + B_2^{-1}U_2^{-1}U_1A_1 = (U_2B_2)^{-1}(U_2A_2 + U_2A_1). \tag{3.3}$$

In essence, the proof of Theorem 3.1.3 generalizes these operations to the setting where invertibility is not guaranteed.

Note that in this arithmetic, the multiplicative and additive identity elements are $(I\backslash I)$ and $(I\backslash 0)$, respectively. The inverse of $(B\backslash A)$ can be defined as $(B\backslash A)^{-1} = (A\backslash B)$, though this is only a true multiplicative inverse satisfying

$$(B\backslash A)^{-1}(B\backslash A) = (I\backslash I) \tag{3.4}$$

if $B^{-1}A$ exists and is invertible. We note a handful of other useful properties, which can be obtained easily from either Definition 3.1.2 or Theorem 3.1.3:

- Standard matrix multiplication/addition can be expressed as $(I\backslash A)(I\backslash B) = (I\backslash(AB))$ and $(I\backslash A) + (I\backslash B) = (I\backslash(A + B))$.
- Scalar multiplication takes the form $\gamma(B\backslash A) = (\beta B\backslash \alpha A)$ for any $\gamma = \alpha/\beta$.

Two questions remain: how do we compute the matrices $\begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ and $\begin{bmatrix} U_1 & U_2 \end{bmatrix}$ in Theorem 3.1.3 and, more importantly, what do we gain by applying these operations to a pencil (A, B)? The latter can be answered with another observation of Benner and Byers [21, Theorems 2.5 and 2.8], which implies that these operations satisfy exactly the properties we outlined at the start of this section.

Theorem 3.1.4 (Benner and Byers 2006)**.** *Let $(B_1\backslash A_1)$ and $(B_2\backslash A_2)$ be matrix relations associated to regular pencils (A_1, B_1) and (A_2, B_2). Suppose v is a shared right eigenvector of (A_1, B_1) and (A_2, B_2) corresponding to finite eigenvalues λ and μ. Then $(\lambda\mu, v)$ and $(\lambda+\mu, v)$ are eigenpairs of the pencils associated to $(B_2\backslash A_2)(B_1\backslash A_1)$ and $(B_2\backslash A_2)+(B_1\backslash A_1)$, respectively.*

Remark 3.1.5. Theorem 3.1.4 extends to infinite eigenvalues in a fairly straightforward way. If either λ or μ is infinite, then (∞, v) is an eigenpair of the pencil associated to $(B_2\backslash A_2)+(B_1\backslash A_1)$. The same can be said for the pencil corresponding to $(B_2\backslash A_2)(B_1\backslash A_1)$ provided we have neither $(\lambda,\mu)=(\infty,0)$ or $(\lambda,\mu)=(0,\infty)$.

If $p(z)$ is any polynomial and (A, B) is regular and diagonalizable, Theorem 3.1.4 implies that the pencil associated to $p(B\backslash A)$ – evaluated according to Theorem 3.1.3 – has the same eigenvectors as (A, B) but transformed eigenvalues $p(\lambda)$. Allowing inversion via $(B\backslash A)^{-1} = (A\backslash B)$, which similarly inverts eigenvalues without changing eigenvectors, the polynomial $p(z)$ can be replaced by any rational function $r(z)$. In this way, the arithmetic on $(B\backslash A)$ implies that any rational function can be applied, inverse-free, to the eigenvalues of (A, B), all while preserving eigenvectors.

Since multiplication on matrix relations – like matrix multiplication itself – is not commutative in general, we need to establish a convention for evaluating $r(B\backslash A)$. If $r(z) = p(z)/q(z)$ for two polynomials p and q, we choose the following:

$$r(B\backslash A) = (q(B\backslash A))^{-1}p(B\backslash A). \tag{3.5}$$

This is somewhat arbitrary and ultimately done for convenience. When (A, B) is diagonalizable, $p(B\backslash A)(q(B\backslash A))^{-1}$ is simply a different representation of the same matrix relation.[2]

Example 3.1.6. Consider a Möbius transformation $r(z) = \frac{az+b}{cz+d}$. In Section 1.7, we noted that such a transformation could be applied to (A, B) as $(aA + bB, cA + dB)$. Using the arithmetic presented above, we can now derive this rigorously. First, we observe

$$a(B\backslash A) + b(I\backslash I) = (B\backslash aA) + (I\backslash bI) = (B\backslash(aA + bB)), \tag{3.6}$$

which follows from Theorem 3.1.3 with $U_1 = B$ and $U_2 = I$. Similarly, $c(B\backslash A) + d(I\backslash I) =$

[2] This is a consequence of the fact that diagonalizable pencils with the same eigenvalues and right eigenvectors are equivalent up to left multiplication by an invertible matrix.

$(B\backslash(cA+dB))$. Combining these yields

$$r(B\backslash A) = ((cA+dB)\backslash B)(B\backslash(aA+bB)) = ((cA+dB)\backslash(aA+bB)), \tag{3.7}$$

where this time we apply Theorem 3.1.3 with $Q_1 = Q_2 = I$.

Remark 3.1.7. Note that in general it is possible that by adding or multiplying matrix relations corresponding to regular pencils we obtain one associated to a singular pencil. Looking to Theorem 3.1.3, this is clearly possible under multiplication if (A_1, B_1) and (A_2, B_2) share a right eigenvector v with corresponding eigenvalues zero and infinity, in which case $A_1 v = B_2 v = 0$.[3] Similarly, the pencil associated to $(B_2\backslash A_2) + (B_1\backslash A_1)$ can be singular if, for example, there exists a vector $w \in \text{null}(B_2)$ such that $\binom{A_1}{A_2}w$ belongs to the range of $\binom{-B_1}{B_2}$. While this appears problematic, Theorem 3.1.4 implies that if (A_1, B_1) and (A_2, B_2) satisfy its assumptions – which will be the case when evaluating $r(B\backslash A)$ provided (A, B) is initially regular and diagonalizable – we are guaranteed that both $(B_2\backslash A_2)(B_1\backslash A_1)$ and $(B_2\backslash A_2) + (B_2\backslash A_1)$ correspond to regular pencils.

To apply these observations, we still need to find the matrices $\begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ and $\begin{bmatrix} U_1 & U_2 \end{bmatrix}$ in Theorem 3.1.3. Perhaps implied by our choice of notation, both can be obtained via (full) QR. This follows from the observation that a factorization

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix} \tag{3.8}$$

implies $\text{null}\left(\begin{bmatrix} Q_{12}^H & Q_{22}^H \end{bmatrix}\right) = \text{range}\left(\begin{bmatrix} A \\ B \end{bmatrix}\right)$. Hence, QR is the computational backbone of the arithmetic on $(B\backslash A)$, implying that it can be performed both efficiently and stably.

We are now ready to return to our original problem. The framework outlined above suggests the following approach for computing (A_S, B_S).

1. Approximate the indicator function $\mathbb{1}_S(z)$ with a rational function $r(z)$.

2. Evaluate $r(B\backslash A)$ using only QR and matrix multiplication [Theorem 3.1.3].

[3] This is the reason Remark 3.1.5 excludes the cases $(\lambda, \mu) = (\infty, 0)$ and $(\lambda, \mu) = (0, \infty)$

Table 3.1. Methods of approximating $\mathbb{1}_S(z)$ for different choices of S. The Newton and Halley iterations are based on the scalar sign function (3.22), as $\mathbb{1}_S(z) = \frac{1}{2}(\text{sign}(z) + 1)$ if S is the right half plane.

Method	S	Approximation $r(z)$
Implicit Repeated Squaring	$\lvert z \rvert < 1$	$(1 + z^{2^k})^{-1}$
Newton Iteration	$\text{Re}(z) > 0$	$f \circ \cdots \circ f(z)$ with $f(z) = \frac{1}{2}(z + z^{-1})$
Halley Iteration	$\text{Re}(z) > 0$	$f \circ \cdots \circ f(z)$ with $f(z) = z\frac{z^2+3}{3z^2+1}$

3. Set $(B_S \backslash A_S) = r(B \backslash A)$.

The resulting pencil (A_S, B_S) will have all of its eigenvalues near zero and one provided $r(z)$ is a good approximation to $\mathbb{1}_S(z)$ and $\Lambda(A, B) \cap \partial S = \emptyset$. This raises an important question: how do we choose the approximation $r(z)$? In weighing our options, we should be mindful of the cost of computing with $(B \backslash A)$. As mentioned above, each addition/multiplication requires a $2n \times n$ full QR factorization. In this sense, the arithmetic on $(B \backslash A)$ is expensive, and a better approximation to $\mathbb{1}_S$ may not result in a more efficient method overall if too many operations are required. There is one exception here: Möbius transformations. As demonstrated in Example 3.1.6, no QR factorizations are required to apply a Möbius transformation to a pencil. Hence, it will be advantageous to write $r(z)$ in terms of Möbius transformations wherever possible.

In this thesis, we focus on a handful of choices for $r(z)$, which are summarized in Table 3.1 and covered rigorously in the remainder of this chapter. Note that while each of these methods is defined for a specific choice of S, they can be used on any appropriate region of $\mathbb{C}$ by first applying a corresponding Möbius transformation (which again can be done cheaply). As we will see, they can also be implemented iteratively, either by iterating the composition of a fixed rational function (in the case of Newton or Halley) or by squaring $(B \backslash A)$ iteratively. In comparing these approaches, we will be interested in quantifying the number of iterations required to obtain an accurate projector.

In some sense, the presentation in this section has flipped the historical narrative.

That is, each of the methods presented in Table 3.1 date back several decades in the literature, while the work of Benner and Byers is a more recent framework for understanding them. The benefit of this framework is its compatibility with algorithmic optimization. In particular, it suggests that the problem of choosing an optimal method for computing spectral projectors reduces to the following.

The Indicator Approximation Problem

Given $S \subset \mathbb{C}$, what is the best rational function approximation to $\mathbb{1}_S$?

Solutions to this problem are only known in a handful of very specialized situations (and under certain notions of "best approximation"). Regardless, any candidate solution $r(z)$ immediately implies a method for computing spectral projectors via the approach outlined above. Moreover, the definition of "best" is flexible, allowing for a prioritization of efficiency or stability. Given the increasing relevance of divide-and-conquer, we suggest the Indicator Approximation Problem as a high-level strategy for deriving variants of the method and refining numerical details.

3.1.1 Perturbation Theory for Full QR

Before moving on, we pause to consider perturbation theory for $2n \times n$ full QR factorizations. Given that these factorizations are the cornerstone of our framework for computing projectors, tight perturbation bounds will be critical for analyzing performance. With this in mind, we derive a spectral norm bound in this subsection.

While perturbation theory for QR originates with Stewart [123], a standard result of Sun [130, Theorem 1.6] is our starting point.

Theorem 3.1.8 (Sun 1991). *Let $A \in \mathbb{C}^{m\times n}$ have rank n and let $A = QR$ be a reduced QR factorization, where $Q \in \mathbb{C}^{m\times n}$ satisfies $Q^H Q = I_n$ and $R \in \mathbb{C}^{n\times n}$ has real, positive diagonal entries. If $E \in \mathbb{C}^{m\times n}$ satisfies $||E||_2 < \sigma_n(A)$ then there exists a unique (reduced)*

QR factorization

$$(A+E) = (Q+W)(R+F)$$

such that

$$||W||_F \leq (1+\sqrt{2}) \cdot \alpha \left(\frac{||E||_2}{\sigma_n(A)} \right) \cdot \kappa_2(A) \cdot \frac{||E||_F}{||A||_2},$$

where $\alpha(\epsilon) = \frac{1}{\epsilon} \ln \left(\frac{1}{1-\epsilon} \right)$ *for* $0 < \epsilon < 1$.

This result has two primary drawbacks. First, it uses the Frobenius norm and is therefore less convenient than a spectral norm bound. We could convert between the two, but doing so naively will incur a factor of $\sqrt{n}$, which – as we will see – is more pessimistic than necessary. On top of this, Theorem 3.1.8 only covers reduced QR factorizations. Luckily, the latter can be addressed with the following lemma.

Lemma 3.1.9. *Let* $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ *and* $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ *be two* $2n \times 2n$ *unitary matrices with* $Q_1, U_1 \in \mathbb{C}^{2n \times n}$. *If* $||Q_1 - U_1||_2 \leq \delta$ *then there exists a unitary* $W \in \mathbb{C}^{n \times n}$ *such that*

$$||Q_2 - U_2 W||_2 \leq 3\delta.$$

Proof. Without loss of generality assume $\delta < \frac{1}{2}$ (the bound is trivial otherwise). We first note that we can use $||Q_1 - U_1||_2$ to control the distance between the orthogonal projectors $Q_1 Q_1^H$ and $U_1 U_1^H$:

$$\begin{aligned} ||Q_1 Q_1^H - U_1 U_1^H||_2 &= ||Q_1 Q_1^H - Q_1 U_1^H + Q_1 U_1^H - U_1 U_1^H||_2 \\ &\leq ||Q_1 (Q_1 - U_1)^H||_2 + ||(Q_1 - U_1) U_1^H||_2 \\ &\leq 2||Q_1 - U_1||_2 \\ &\leq 2\delta. \end{aligned} \tag{3.9}$$

Since $Q_2 Q_2^H = I - Q_1 Q_1^H$ and $U_2 U_2^H = I - U_1 U_1^H$ this similarly implies $||Q_2 Q_2^H - U_2 U_2^H||_2 \leq 2\delta$. With this in mind, let $Q_1 Q_1^H = U_1 U_1^H + E$ for some $E \in \mathbb{C}^{2n \times 2n}$ with $||E||_2 \leq 2\delta$. Noting $Q_2 Q_2^H = U_2 U_2^H - E$ we observe

$$Q_2 - U_2 (U_2^H Q_2) = -E Q_2. \tag{3.10}$$

Consider now $U_2^H Q_2$:

$$(U_2^H Q_2)^H U_2^H Q_2 = Q_2^H (I - U_1 U_1^H) Q_2 = I - Q_2^H (Q_1 Q_1^H - E) Q_2 = I + Q_2^H E Q_2. \quad (3.11)$$

(3.11) implies that each singular value of $U_2^H Q_2$ takes the form $\sqrt{1+\lambda}$ for λ an eigenvalue of $Q_2^H E Q_2$ satisfying $|\lambda| \leq ||Q_2^H E Q_2||_2 \leq 2\delta$. Consequently $\sqrt{1+\lambda} \leq \sqrt{1+2\delta} \leq 1+\delta$ and the SVD of $U_2^H Q_2$ can be written as

$$U_2^H Q_2 = V_1 (I + \Sigma) V_2^H \quad (3.12)$$

for $V_1, V_2 \in \mathbb{C}^{n \times n}$ unitary and Σ diagonal with nonzero entries bounded in magnitude by δ. Letting $W = V_1 V_2^H$ we have

$$Q_2 - U_2 W = -EQ_2 + U_2 V_1 \Sigma V_2^H \quad (3.13)$$

where, by construction, $|| - EQ_2 + U_2 V_1 \Sigma V_2^H ||_2 \leq ||E||_2 + ||\Sigma||_2 \leq 3\delta$. □

Informally, Lemma 3.1.9 says that the trailing columns of two full QR factorizations corresponding to nearby reduced ones are close to rotations/reflections of one another. Put another way, two nearby reduced factorizations can be built into similarly close full factorizations, which allows us to extend any reduced bound to full QR.

With this in mind, we now pursue a (sharper) spectral norm version of Theorem 3.1.8. Ostensibly, Sun's result is written in terms of the Frobenius norm because it makes use of the following inequality: given $L \in \mathbb{C}^{n \times n}$ lower triangular with real diagonal entries,

$$||L||_F \leq \frac{1}{\sqrt{2}} ||L + L^H||_F. \quad (3.14)$$

While no inequality of the form $||L||_2 \leq C||L + L^H||_2$ for C a positive constant can exist (multiply the matrix in [22, Example 3.3] by i for a counterexample, as was pointed out to us by Anne Greenbaum [63]), we can prove the following alternative.

Lemma 3.1.10. *Let $L \in \mathbb{C}^{n \times n}$ be lower triangular with real diagonal entries. Then,*

$$||L||_2 \leq \left(\frac{1}{2} + L_{n+1} \right) ||L + L^H||_2$$

for $L_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} |D_k(\theta)| d\theta$ the k^{th} Lebesgue constant and $D_k(\theta)$ the Dirichlet kernel.

Proof. This follows from Bhatia [22]. Writing

$$L^H = \Delta_U(L + L^H) - \frac{1}{2}\mathcal{D}_0(L + L^H) \tag{3.15}$$

where $\Delta_U(A)$ and $\mathcal{D}_0(A)$ are obtained from A be zeroing out the sub-diagonal and off-diagonal entries respectively, we have

$$||L||_2 = ||L^H||_2 \leq ||\Delta_U(L + L^H)||_2 + \frac{1}{2}||\mathcal{D}_0(L + L^H)||_2 \leq \left(\frac{1}{2} + L_{n+1}\right) ||L + L^H||_2 \tag{3.16}$$

after applying [22, Equations 3 and 15]. □

Remark 3.1.11. It can be shown [23, Section 2.2] that

$$L_k \leq \ln(k) + \ln(\pi) + \frac{2}{\pi}\left(1 + \frac{2}{k}\right). \tag{3.17}$$

In other words, L_k grows (at most) like $\ln(k)$ for large k and Lemma 3.1.10 could be written generally as $||L||_2 \leq O(\ln(n))||L + L^H||_2$. In an effort to keep track of constants, we will use the explicit, though slightly looser,

$$||L||_2 \leq (\ln(n + 1) + 3)||L + L^H||_2 \tag{3.18}$$

going forward. Note that the counterexample mentioned above implies that the dependence on $\ln(n)$ is tight.

Repeating the proof of [130, Theorem 1.6] with (3.18) in place of (3.14) yields our main perturbation bound.

Theorem 3.1.12. *Let $A \in \mathbb{C}^{m\times n}$ have rank n and let $A = QR$ be an reduced QR factorization, where $Q \in \mathbb{C}^{m\times n}$ satisfies $Q^HQ = I_n$ and $R \in \mathbb{C}^{n\times n}$ has real, positive diagonal entries. If $E \in \mathbb{C}^{m\times n}$ satisfies $||E||_2 < \sigma_n(A)$ then there exists a unique (reduced) QR factorization*

$$(A + E) = (Q + W)(R + F)$$

such that

$$||W||_2 \leq (2\ln(n+1)+7)\ln\left(\frac{\sigma_n(A)}{\sigma_n(A)-||E||_2}\right)$$
$$= (2\ln(n+1)+7)\cdot\alpha\left(\frac{||E||_2}{\sigma_n(A)}\right)\cdot\kappa_2(A)\cdot\frac{||E||_2}{||A||_2},$$

for $\alpha(\epsilon)$ as in Theorem 3.1.8.

While this result is new, it not the only QR bound developed since Theorem 3.1.8. Shortly after its publication, for example, Sun derived improvements for real matrices [131]. More general bounds were subsequently found by Bhatia and Mukherjea [24] and Li and Wei [88]. We also note a recent componentwise analysis done by Petkov [108], again for real matrices. Despite their improvements over Theorem 3.1.8, these results are not easily adaptable into a general, spectral norm bound.

3.2 Implicit Repeated Squaring (IRS)

In this section, we consider an approach for computing P_R and P_L based on Implicit Repeated Squaring (**IRS**), a routine for repeatedly squaring a product $A^{-1}B$ without forming it. **IRS** originates with general divide-and-conquer eigensolvers, first in work[4] of Malyshev [91, 92] and later Bai, Demmel, and Gu [7]. It was subsequently stated as it appears in Algorithm 1 under the name **IRS** in a technical report of Ballard, Demmel, and Dumitriu [11].

IRS can be used to compute P_R and P_L by applying the framework from the previous section with $r(z) = (1+z^{2^p})^{-1}$ and $S = \{z : |z| < 1\}$. In these terms, the pseudocode of Algorithm 1 can be viewed as a straightforward application of Theorem 3.1.3 to $(A_p \backslash B_p) = (A\backslash B)^{2^p}$, where squaring naturally drives eigenvalues to zero and infinity (assuming none are on the unit circle). Applying the Möbius transformation $(1+z)^{-1}$, which sends zero to one and infinity to zero, the projector P_R can be obtained from

[4] The paper [92] was translated from Russian in two parts [93,94]. Much of its content was subsequently presented in [95].

Algorithm 1. Implicit Repeated Squaring (**IRS**)
Input: $A, B \in \mathbb{C}^{n\times n}$ and p a positive integer.
Output: $A_p, B_p \in \mathbb{C}^{n\times n}$ satisfying $A_p^{-1}B_p = (A^{-1}B)^{2^p}$.

```
$A_0 = A$
$B_0 = B$
for $j = 0 : p-1$ do
    $$\begin{pmatrix} B_j \\ -A_j \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} R_j \\ 0 \end{pmatrix}$$
    $A_{j+1} = Q_{12}^H A_j$
    $B_{j+1} = Q_{22}^H B_j$
end for
return $A_p, B_p$
```

$((A_p + B_p)\backslash A_p)$ as

$$P_R \approx (A_p + B_p)^{-1}A_p. \tag{3.19}$$

Repeating this process with (A^H, B^H) yields the left projector $P_L \approx \mathcal{A}_p^H(\mathcal{A}_p + \mathcal{B}_p)^{-H}$ for $(\mathcal{A}_p\backslash\mathcal{B}_p) = (A^H\backslash B^H)^{2^p}$.

Note here that **IRS** is applied to $(A\backslash B)$ rather than $(B\backslash A)$. This is done to maintain consistency with the presentation of **IRS** in [7,11], though it also means that P_R and P_L are spectral projectors of (A, B) corresponding to $\{z : |z| > 1\}$ rather than S. To avoid confusion, we label the projectors as $P_{R,|z|>1}$ and $P_{L,|z|>1}$ to make clear the subset of $\Lambda(A, B)$ they depend on.

We turn now to the question of accuracy. Intuitively, **IRS** will fail to compute $P_{R,|z|>1}$ and $P_{L,|z|>1}$ if (A, B) has an eigenvalue on the unit circle, in which case squaring cannot push $\Lambda(A, B)$ to zero and infinity. This observation is at the heart of efforts to derive a condition number for the procedure. Malyshev originally suggested[5] $\omega_{(A,B)}$ – the "criterion of absence of eigenvalues of the pencil $\lambda B - A$ on the unit circle and within a small neighborhood of it," which can be defined as follows.

[5] Malyshev's definition is actually a generalized and scale invariant version of a similar quantity of Bulgakov and Godunov [28].

Definition 3.2.1. For a regular pencil (A,B),

$$\omega_{(A,B)} = \left|\left|\frac{1}{2}\int_0^{2\pi}(B-e^{i\phi}A)^{-1}(AA^H+BB^H)(B-e^{i\phi}A)^{-H}d\phi\right|\right|_2.$$

While $\omega_{(A,B)}$ is only formally defined for regular pencils, Definition 3.2.1 can be easily extended by setting $\omega_{(A,B)}=\infty$ if (A,B) is singular. Regardless, $\omega_{(A,B)}$ is somewhat cumbersome to work with, both computationally and conceptually. Aiming to replace it with something simpler, Bai, Demmel, and Gu [7] subsequently analyzed **IRS** in terms of a distance to the nearest ill-posed problem $d_{(A,B)}$, whose reciprocal is an alternative candidate for a condition number.[6]

Definition 3.2.2. The distance from (A,B) to the nearest ill-posed problem is

$$d_{(A,B)} = \inf\left\{||E||_2+||F||_2 : (A+E)-z(B+F) \text{ is singular for some } |z|=1\right\}.$$

As desired, both $\omega_{(A,B)}$ and $d_{(A,B)}^{-1}$ are infinite if (A,B) is singular or has an eigenvalue on the unit circle. While Malyshev derives rigorous error bounds for **IRS** in terms of $\omega_{(A,B)}$, including in finite-precision arithmetic (see [93]), we prefer $d_{(A,B)}$ here due to its compatibility with bounds on the pseudospectrum of (A,B). In particular, $\Lambda_\epsilon(A,B)\cap\{z:|z|=1\}=\emptyset$ implies $d_{(A,B)}\geq 2\epsilon$. Accordingly, the main error bound we make use of in the next chapter is due to Bai, Demmel, and Gu [7, Theorem 1], which establishes quadratic convergence for exact-arithmetic **IRS** in terms of $||(A,B)||_2/d_{(A,B)}$.

Theorem 3.2.3 (Bai-Demmel-Gu 1994)**.** *Let A_p, B_p be the result of applying* ***IRS*** *to A,B. If*

$$p \geq \log_2\left[\frac{||(A,B)||_2-d_{(A,B)}}{d_{(A,B)}}\right]$$

then

$$||(A_p+B_p)^{-1}A_p - P_{R,|z|>1}||_2 \leq ||P_{R,|z|>1}||_2\frac{2^{p+3}\left(1-\frac{d_{(A,B)}}{||(A,B)||_2}\right)^{2^p}}{\max\left\{0,1-2^{p+2}\left(1-\frac{d_{(A,B)}}{||(A,B)||_2}\right)^{2^p}\right\}}.$$

[6] Note, however, that $d_{(A,B)}$ is not invariant to scaling since $d_{(\alpha A,\alpha B)}=|\alpha|d_{(A,B)}$. For this reason, results of Bai, Demmel, and Gu are stated in terms of $\frac{||(A,B)||_2}{d_{(A,B)}}$.

Note that because $d_{(A^H,B^H)} = d_{(A,B)}$ and $||(A,B)||_2 = ||(A^H,B^H)||_2$, Theorem 3.2.3 can be applied to the left projector by swapping A, B, $(A_p + B_p)^{-1}A_p$, and $P_{R,|z|>1}$ for A^H, B^H, $A_p^H(A_p + B_p)^{-H}$, and $P_{L,|z|>1}$, respectively.

Remark 3.2.4. As we might expect, the spectral projector application considered in this chapter is not the only place **IRS** may be of use in numerical linear algebra. We note here a specific example: the matrix exponential e^A, which for polynomials q_1, q_2 is commonly computed as

$$e^A \approx \left[q_1(A/2^p)^{-1}q_2(A/2^p)\right]^{2^p} \tag{3.20}$$

via the scaling and squaring method [75]. While much effort has gone into evaluating the performance of scaling and squaring, and in particular devising best practices for choosing p and the polynomials q_1, q_2 [4, 6, 101], the stability of the final squaring step of the algorithm – which is typically done explicitly – has been largely overlooked. Of course, applying **IRS** here would necessitate an alternative framework for analyzing the routine; in particular, the tools discussed above may no longer capture performance, as computing e^A is unaffected by the presence of eigenvalues of $(q_2(A/2^p), q_1(A/2^p))$ on the unit circle. This is explored in more detail in Chapter 6.

3.3 Newton and Halley Iterations for the Matrix Sign Function

We close this chapter by considering an alternative family of methods based around the matrix sign function. In the framework of Section 3.1, these use the right half plane for S, in which case

$$\mathbb{1}_S(z) = \frac{1}{2}(\text{sign}(z) + 1), \tag{3.21}$$

for sign(z) the scalar sign function

$$\operatorname{sign}(z) = \begin{cases} +1 & \operatorname{Re}(z) > 0 \\ -1 & \operatorname{Re}(z) < 0 \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{3.22}$$

In this setting, approximations of $\mathbb{1}_S(z)$ can be derived from approximations of sign(z), and moreover computing P_R and P_L reduces to approximating the matrix sign function sign($B^{-1}A$), which can be defined as follows.

Definition 3.3.1. Let A have no eigenvalues on the imaginary axis. Suppose

$$A = P \begin{pmatrix} J_+ & \\ & J_- \end{pmatrix} P^{-1}$$

is the Jordan canonical form of A, where blocks in J_+ and J_- correspond to eigenvalues in the right and left half planes, respectively. Then the *matrix sign function* of A is

$$\operatorname{sign}(A) = P \begin{pmatrix} I & \\ & -I \end{pmatrix} P^{-1}.$$

A generalization of the sign function to matrix pencils was first introduced in work of Gardiner and Laub [60], where

$$\operatorname{sign}(A, B) = B \operatorname{sign}(B^{-1}A), \tag{3.23}$$

assuming B is invertible. In this case, it is not hard to see that $\frac{1}{2}(\operatorname{sign}(A, B) + B)$ is a projector onto the right deflating subspace of (A, B) corresponding to the right half plane.[7] Though interesting in its own right, the high-level strategy of Section 3.1 does not require working with this generalization. Instead, it is sufficient to simply approximate sign($B^{-1}A$) – without assuming or using inversion – via sign($B\backslash A$).

With this in mind, we can now consider the problem of approximating sign(z). Since methods based around the matrix sign function are the most popular choice for computing

[7] This generalizes the single-matrix case, where $\frac{1}{2}(\operatorname{sign}(A) + I)$ is a projector onto the eigenspace corresponding to the right half plane.

Algorithm 2. Inverse-Free Newton Iteration (**IF-Newton**)
Input: $A, B \in \mathbb{C}^{n\times n}$, p a number of iterations.
Requires: (A, B) has no eigenvalues on $\text{Re}(z) = 0$.

```
1: $A_0 = A$
2: $B_0 = B$
3: for $i = 0 : p-1$ do
4:     $\begin{pmatrix} -A_i \\ B_i \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} R_i \\ 0 \end{pmatrix}$
5:     $A_{i+1} = \frac{1}{\sqrt{2}}(Q_{12}^H B_i + Q_{22}^H A_i)$
6:     $B_{i+1} = \sqrt{2} Q_{22}^H B_i$
7: end for
8: return $(A_p, B_p)$
```

spectral projectors in the literature, particularly in the single-matrix case [16, 19, 78], we have a few standard pathways forward. Typically, $\text{sign}(A)$ is approximated via a simple Newton iteration of Roberts [109]. From the viewpoint of function approximation, this iteration approximates $\text{sign}(z)$ by the (rational) function obtained by repeatedly composing $f(z) = \frac{1}{2}(z + z^{-1})$ with itself.

Definition 3.3.2. The *Newton iteration* for computing $\text{sign}(A)$ is given by

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-1}); \quad A_0 = A.$$

Recalling that the inverse of $(B\backslash A)$ is $(A\backslash B)$, the standard Newton iteration can be applied to matrix relations as follows

$$(B_{k+1}\backslash A_{k+1}) = \frac{1}{2}\left[(B_k\backslash A_k) + (A_k\backslash B_k)\right]; \quad (B_0\backslash A_0) = (B\backslash A). \tag{3.24}$$

Algorithm 2 executes p steps of this iteration according to Theorem 3.1.3. Here, the factor of $\frac{1}{2}$ is applied by scaling A_{k+1} by $\frac{1}{\sqrt{2}}$ and B_{k+1} by $\sqrt{2}$, which is necessary to guarantee convergence of the individual matrices as $k \to \infty$ in exact arithmetic (see [21, Theorem 3.6]). As in the approach based on **IRS**, some post-processing is necessary to obtain $P_{R,\text{Re}(z)>0}$, where again the subscript clarifies the corresponding subset of $\Lambda(A, B)$. In this case, $B_p^{-1}A_p$ approximates $\text{sign}(B^{-1}A)$ and therefore

$$P_{R,\text{Re}(z)>0} \approx \frac{1}{2}(B_p^{-1}A_p + I) = \frac{1}{2}B_p^{-1}(A_p + B_p). \tag{3.25}$$

Equivalently, $P_{R,\mathrm{Re}(z)>0}$ corresponds to the matrix relation $(2B_p\backslash(A_p + B_p))$.

As its name suggests, the Newton iteration can be viewed as an extension of classical Newton's method, which finds roots of $z^2 - 1$ according to $z_{k+1} = \frac{1}{2}(z_k + z_k^{-1})$. Indeed, the Newton iteration for $\mathrm{sign}(A)$ applies this version of Newton's method to the eigenvalues of A, and quadratic convergence of classical Newton's method implies quadratic convergence for (3.24).

At the same time, this observation suggests that other methods for approximating $\mathrm{sign}(A)$ can be obtained from alternative root finding iterations. Halley's method, for example, approximates roots of $z^2 - 1$ according to the third-order iteration

$$z_{k+1} = z_k \frac{z_k^2 + 3}{3z_k^2 + 1}. \tag{3.26}$$

Consequently, it implies the other main iteration for $\mathrm{sign}(A)$ we'll consider here.

Definition 3.3.3. The *Halley iteration* for computing $\mathrm{sign}(A)$ is given by

$$A_{k+1} = A_k(3A_k^2 + 1)^{-1}(A_k^2 + 3); \quad A_0 = A.$$

Recalling that any Möbius transformation can be applied to $(B\backslash A)$ for free, only two QR factorizations are required to run the Halley iteration on matrix relations if evaluated as

$$(B_{k+1}\backslash A_{k+1}) = (B_k\backslash A_k)h((B_k\backslash A_k)^2); \quad (B_0\backslash A_0) = (B\backslash A) \tag{3.27}$$

for $h(z) = \frac{3z+1}{z+3}$. As in the Newton iteration, the approximation of $\mathrm{sign}(z)$ corresponding to (3.27) can be obtained by repeated composition, this time with $f(z) = zh(z^2)$. Applying Theorem 3.1.3 yields Algorithm 3, which executes p steps of this Halley iteration on an arbitrary pencil (A, B). The outputs of this routine yield the projector $P_{R,\mathrm{Re}(z)>0}$ according to (3.25). Note that while this approach is somewhat less popular than the Newton iteration, a variant of (3.27) has seen wide use in divide-and-conquer efforts for the symmetric eigenvalue problem [107].

Algorithm 3. Inverse-Free Halley Iteration (**IF-Halley**)
Input: $A, B \in \mathbb{C}^{n \times n}$, p a number of iterations.
Requires: (A, B) has no eigenvalues on $\mathrm{Re}(z) = 0$.

```
$A_0 = A$
$B_0 = B$
**for** $i = 0 : p - 1$ **do**
    $\begin{pmatrix} -B_i \\ A_i \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} R_i \\ 0 \end{pmatrix}$
    $C_i = Q_{12}^H A_i + 3Q_{22}^H B_i$
    $D_i = 3Q_{12}^H A_i + Q_{22}^H B_i$
    $\begin{pmatrix} -D_i \\ A_i \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \begin{pmatrix} \widehat{R}_i \\ 0 \end{pmatrix}$
    $A_{i+1} = U_{12}^H C_i$
    $B_{i+1} = U_{22}^H B_i$
**end for**
**return** $(A_p, B_p)$
```

We could continue from here. The rational function that defines the Halley iteration – i.e., (3.26) – belongs to the family of Padé approximants, so we might next consider other options there, as in [83]. Alternatively, we could interpret (3.26) as the simplest rational function of the form $r(z) = zp(z^2)/q(z^2)$ for p and q polynomials of the same degree, which have been studied as candidates for approximating $\mathrm{sign}(z)$ since work of Zolotarev more than a century ago [106, 148]. In either case, we obtain a new method for approximating $\mathrm{sign}(A)$ from each choice of rational function, where faster convergence can be pursued by increasing the degree of the approximation.

If higher-degree rational functions are too computationally intensive, we might instead consider optimizing the Newton and Halley iterations as they appear in Algorithms 2 and 3. In the case of the Newton iteration, scaling is typically used to promote stability and/or improve convergence (see for example Benner and Byer's version of **IF-Newton** [21, Algorithm 1]). Optimizing Halley's iteration, meanwhile, has been explored by Nakatsukasa, Bai, and Gygi [105] who suggested replacing (3.26) with

$$z_{k+1} = z_k \frac{a_k z_k^2 + b_k}{c_k z_k^2 + d_k} \tag{3.28}$$

for dynamically changing coefficients a_k, b_k, c_k, and d_k. This begs the question: what are the optimal choices for these coefficients to guarantee fast convergence from the starting points $\Lambda(A, B)$? Importantly, Nakatsukasa, Bai, and Gygi provide an answer in the case that $\Lambda(A, B)$ is contained in a union of intervals on the real axis.

With the exception of the latter – which we make use of in Chapter 5 – we will not consider these extensions in detail here. We mention them to again emphasize the flexibility of our high-level strategy, which can accommodate both more complex rational function approximations to $\text{sign}(z)$ and scaling/optimization tricks aimed at improving performance. The remainder of this section is instead devoted to developing a framework for analyzing the performance of methods built on the matrix sign function, including **IF-Newton** and **IF-Halley**

We begin by noting that any method based on the sign function cannot be applied to a pencil (A, B) with eigenvalues on the imaginary axis. As in the case of **IRS** and the unit circle, we can consider such problems ill-posed for this approach to computing spectral projectors. With this in mind, a key theoretical tool for measuring convergence will be the circles of Apollonius, which – as the name suggests – date back to antiquity.

Definition 3.3.4. For $\alpha \in (0, 1)$ let

$$C_\alpha^+ = \left\{ z : \left| \frac{1-z}{1+z} \right| \leq \alpha \right\}, \quad C_\alpha^- = \left\{ z : \left| \frac{1+z}{1-z} \right| \leq \alpha \right\}$$

be sets in the right and left half planes, respectively. The boundaries ∂C_α^+ and ∂C_α^- of these sets are the *circles of Apollonius* corresponding to α.

C_α^+ can be equivalently characterized as the disk with center $\frac{1+\alpha^2}{1-\alpha^2}$ and radius $\frac{2\alpha}{1-\alpha^2}$, with C_α^- its image under a reflection across the imaginary axis. For varying α, ∂C_α^+ and ∂C_α^- define families of non-concentric circles, which collapse to the points ±1 as $\alpha \to 0$. Since this geometric picture will be important to have in mind, Figure 3.1 plots a handful of Apollonian circles. Throughout, we use C_α to denote the region $C_\alpha^+ \cup C_\alpha^-$.

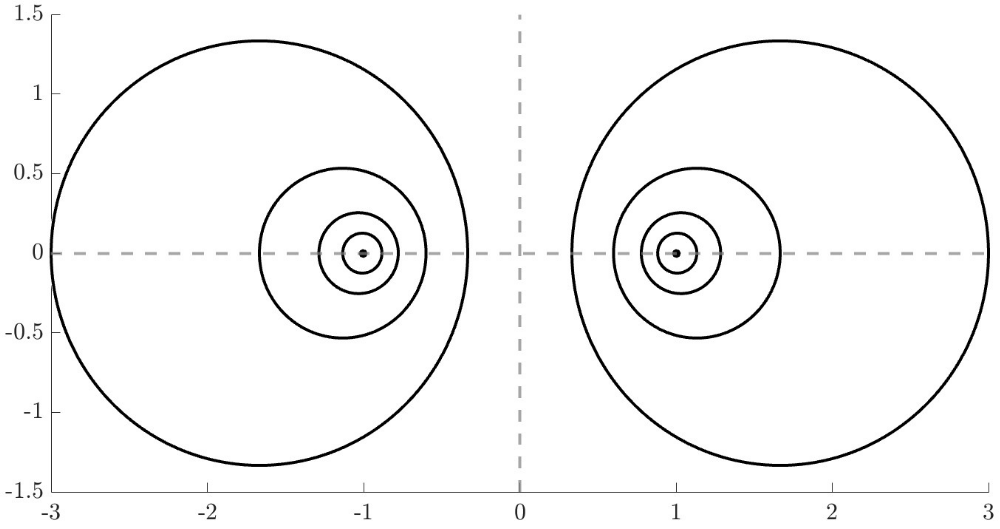


Figure 3.1. The circles of Apollonius corresponding to $\alpha = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$, and $\frac{1}{16}$.

Given their relationship to the points ±1, the circles of Apollonius are naturally equipped to describe convergence to the sign function. Indeed, the Newton iteration can be characterized by the following observation of Roberts [109].

Proposition 3.3.5. *The function $f(z) = \frac{1}{2}(z + z^{-1})$ defining the Newton iteration maps C_α^+ to $C_{\alpha^2}^+$ and C_α^- to $C_{\alpha^2}^-$.*

Extending this to the Halley iteration is straightforward. Lemma 3.3.6 captures the third-order convergence of Halley's method for finding the roots of $z^2 - 1$.

Lemma 3.3.6. *The function $f(z) = zh(z^2) = z\frac{3z^2+1}{z^2+3}$ defining the Halley iteration maps C_α^+ to $C_{\alpha^3}^+$ and C_α^- to $C_{\alpha^3}^-$.*

Proof. Applying the definition of $C_\alpha^\pm$, we have

$$\frac{1 - f(z)}{1 + f(z)} = \frac{1 - \frac{z^3+3z}{3z^2+1}}{1 + \frac{z^3+3z}{3z^2+1}} = \frac{3z^2 + 1 - z^3 - 3z}{3z^2 + 1 + z^3 + 3z} = \frac{(1-z)^3}{(1+z)^3}. \tag{3.29}$$

The result follows immediately. □

Remark 3.3.7. The proof of Lemma 3.3.6 implies yet another strategy for deriving iterative methods for the sign function: work backwards from $\pm\frac{(1-z)^m}{(1+z)^m}$ given a desired order of convergence m. As an example, $-\frac{(1-z)^4}{(1+z)^4}$ can be written as

$$-\frac{(1-z)^4}{(1+z)^4} = \frac{1 - \frac{1+6z^2+z^4}{4z+4z^3}}{1 + \frac{1+6z^2+z^4}{4z+4z^3}}, \tag{3.30}$$

which implies an iterative rational function

$$f(z) = \frac{1+6z^2+z^4}{4z+4z^3} = \frac{z}{4}\left(\frac{1}{z^2} + \frac{z^2+5}{1+z^2}\right), \tag{3.31}$$

where the latter expression indicates that this iteration can be implemented according to Theorem 3.1.3 with only three QR factorizations.

Going a step further, we next pursue an analog of Theorem 3.2.3 under the assumption that $\Lambda_\epsilon(A,B) \subset C_\alpha$. Like the decision to work with $d_{(A,B)}$ to bound error in **IRS**, our goal here is to develop tools for measuring convergence that are compatible with pseudospectral bounds, which we expect to have access to via Chapter 2. To do this, we present a handful of results for an individual matrix X, which we then demonstrate can be applied to our general setting. We begin by restating a key lemma of Banks et al. [16, Lemma 4.3].

Lemma 3.3.8 (Banks et al. 2022)**.** *Suppose $\Lambda_\epsilon(X) \subset C_\alpha$ for some $\epsilon > 0$. Then,*

$$||X - \text{sign}(X)||_2 \leq \frac{8\alpha^2}{\epsilon(1+\alpha)(1-\alpha)^2}.$$

With this result in mind, we next seek insight into how a pseudospectral bound like $\Lambda_\epsilon(X) \subset C_\alpha$ evolves under an iteration for computing sign(X). The resulting Lemma 3.3.9 is a straightforward generalization of [16, Lemma 4.4].

Lemma 3.3.9. *Suppose the rational function f has all of its poles on the imaginary axis and maps $C_\alpha^\pm \to C_{\alpha^m}^\pm$ with $\partial C_\alpha^\pm \to \partial C_{\alpha^m}^\pm$ for any $\alpha \in (0,1)$. Let f define an iteration for* sign(X) *according to*

$$X_{k+1} = f(X_k); \quad X_0 = X.$$

If $\Lambda_\epsilon(X_k) \subset C_\alpha$, then for any $\alpha' \in (\alpha^m, \alpha)$ we have $\Lambda_{\epsilon'}(X_{k+1}) \subset C_{\alpha'}$ with

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

Proof. Let w be any point in the "annulus" between C_α and $C_{\alpha'}$. Since f maps C_α to $C_{\alpha'}$ and $w \notin C_{\alpha'}$, the rational function $\frac{1}{w-f(z)}$ is holomoprhic on C_α. Moreover, C_α contains $\Lambda(X_k)$, meaning we can bound $||(wI - X_{k+1})^{-1}||_2$ as

$$\begin{aligned} ||(wI - X_{k+1})^{-1}||_2 &= \left|\left| \frac{1}{2\pi i} \int_{\partial C_\alpha} \frac{(w-f(z))^{-1}}{zI - X_k} dz \right|\right|_2 \\ &\leq \frac{1}{2\pi} \int_{\partial C_\alpha^+} \frac{||(zI-X_k)^{-1}||_2}{|w - f(z)|} dz + \frac{1}{2\pi} \int_{\partial C_\alpha^-} \frac{||(zI-X_k)^{-1}||_2}{|w - f(z)|} dz. \end{aligned} \tag{3.32}$$

Appealing to the ML-inequality (2.33), the first integral in this expression becomes

$$\begin{aligned} \int_{\partial C_\alpha^+} \frac{||(zI-X_k)^{-1}||_2}{|w - f(z)|} dz &\leq l(\partial C_\alpha^+) \sup_{z \in \partial C_\alpha^+} \frac{||(zI-X_k)^{-1}||_2}{|w - f(z)|} \\ &= \frac{4\pi\alpha}{1-\alpha^2} \sup_{z \in \partial C_\alpha^+} \frac{||(zI-X_k)^{-1}||_2}{|w - f(z)|}. \end{aligned} \tag{3.33}$$

Now $\Lambda_\epsilon(X_k) \cap \partial C_\alpha^+ = \emptyset$, so $||(zI - X_k)^{-1}||_2 \leq \epsilon^{-1}$ for all $z \in \partial C_\alpha^+$. Using the fact that $f(z) \in C_{\alpha^m}^+$ if $z \in C_\alpha^+$, we therefore have

$$\begin{aligned} \int_{\partial C_\alpha^+} \frac{||(zI-X_k)^{-1}||_2}{|w - f(z)|} dz &\leq \frac{4\pi\alpha}{\epsilon(1-\alpha^2)} \sup_{y \in \partial C_{\alpha^m}^+} \frac{1}{|w-y|} \\ &\leq \frac{8\pi\alpha}{\epsilon(1-\alpha^2)(\alpha'-\alpha^m)}, \end{aligned} \tag{3.34}$$

where the last inequality follows from [16, Lemma 4.5]. Since we obtain the same bound on the remaining term of (3.32), we conclude

$$||(wI - X_{k+1}])^{-1}||_2 \leq \frac{8\alpha}{\epsilon(1-\alpha^2)(\alpha'-\alpha^m)}, \tag{3.35}$$

and therefore $w \notin \Lambda_{\epsilon'}(X_{k+1})$ for $\epsilon' = \frac{\epsilon(1-\alpha^2)(\alpha'-\alpha^m)}{8\alpha}$. Since (3.35) applies to any point w between C_α and $C_{\alpha'}$ and $\Lambda(X_{k+1}) \subset C_{\alpha'}$, this suffices to show $\Lambda_{\epsilon'}(X_{k+1}) \subset C_{\alpha'}$. □

In tandem, Lemma 3.3.8 and Lemma 3.3.9 imply the following strategy for bounding error in any method (for computing spectral projectors of (A, B)) based on the matrix sign function.

Table 3.2. Number of full $2n \times n$ QR factorizations and $n \times n$ matrix multiplications required for one iteration of each iterative method discused in Sections 3.2 and 3.3. *Benner and Byers compute one additional $n \times n$ QR factorization to accommodate scaling.

	IRS	**IF-Newton**	[21, Algorithm 1]	**IF-Halley**
QR's per iteration	1	1	1*	2
MM's per iteration	2	3	4	4

1. Start with an initial pseudospectral guarantee $\Lambda_\epsilon(A, B) \subset C_\alpha$ (e.g., coming from Chapter 2). Note that such a bound implies that B is invertible and moreover $\Lambda_{\epsilon/||B||_2}(B^{-1}A) \subset \Lambda_\epsilon(A, B) \subset C_\alpha$.

2. Select a rational function $f(z)$ satisfying the assumptions of Lemma 3.3.9 and define the iteration $(B_{k+1}\backslash A_{k+1}) = f(B_k\backslash A_k)$ with $(B_0\backslash A_0) = (B\backslash A)$. In exact arithmetic, this is equivalent to $B_{k+1}^{-1}A_{k+1} = f(B_k^{-1}A_k)$.

3. For a chosen number of steps p, repeatedly apply Lemma 3.3.9 with $X_k = B_k^{-1}A_k$. The base case here is $\Lambda_{\epsilon/||B||_2}(B^{-1}A) \subset C_\alpha$ for ϵ as in step one.

4. Use Lemma 3.3.8 to bound $||B_p^{-1}A_p - \text{sign}(B^{-1}A)||_2$, as $\text{sign}(B_p^{-1}A_p) = \text{sign}(B^{-1}A)$ since (A, B) and (A_p, B_p) have the same eigenvectors.

5. Bootstrap this bound to one for the projector $P_{R,\text{Re}(z)>0}$ by observing that

$$\left|\left|\frac{1}{2}B_p^{-1}(A_p + B_p) - P_{R,\text{Re}(z)>0}\right|\right|_2 = \frac{1}{2}||B_p^{-1}A_p - \text{sign}(B^{-1}A)||_2. \tag{3.36}$$

Importantly, this framework is general – i.e., it can be applied to both **IF-Newton** and **IF-Halley** but is not specific to either. As in Chapter 2, working with the product matrices $B_k^{-1}A_k$ here is a theoretical exercise only, allowing us to take advantage of results like Lemma 3.3.9 to bound error in an inverse-free approach.

To close, Table 3.2 counts the number of matrix relation operations required for each method of computing P_R and P_L considered in this chapter.

Content Acknowledgement: Portions of this chapter are repurposed from the following submitted works:

- J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.
- R. Schneider. When is fast, implicit squaring of $A^{-1}B$ stable? arXiv:2310.00193, 2023.

The dissertation author was the primary investigator and author of both.

Chapter 4

Pseudospectral Divide-and-Conquer

We are now ready to present our randomized, divide-and-conquer algorithm for the generalized eigenvalue problem, which we refer to as *pseudospectral divide-and-conquer*. In essence, this algorithm is a consequence of the work presented in the previous two chapters; Chapter 2 provides a high-level strategy for spectral bisection while Chapter 3, particularly Section 3.2, clears a pathway for the remaining steps of divide-and-conquer. Accordingly, the contents of this chapter are primarily technical, aimed at piecing together these results to state a provably successful diagonalization routine. The high-level approach can be summarized as follows:

1. Randomly perturb and scale the input pencil (A, B) to obtain $(\widetilde{A}, n^\alpha \widetilde{B})$, thereby gaining access to our main pseudospectral guarantee (Theorem 2.3.1).

2. Diagonalize $(\widetilde{A}, n^\alpha \widetilde{B})$ via divide-and-conquer, where a random shattering grid is used

to split $\Lambda(\widetilde{A}, n^\alpha \widetilde{B})$ and spectral projectors are computed with **IRS** (see Section 3.2).

3. Undoing the n^α scaling, allow the resulting diagonalization of $(\widetilde{A}, \widetilde{B})$ to stand in for a diagonalization of (A, B).

What results is the first-known algorithm that can diagonalize any pencil (A, B), with high probability, in nearly matrix multiplication time.

Guide to Chapter 4: Section 4.1 defines the remaining algorithmic building blocks necessary to state a diagonalization routine, including the randomized rank-revealing factorization we make use of. In Section 4.2, we prove the main result of this thesis, stating our diagonalization algorithm and proving that it both succeeds with high probability on arbitrary inputs and runs in nearly matrix multiplication time. We close the chapter with a handful of numerical examples in Section 4.3.

4.1 Numerical Building Blocks

In this section, we present the remaining numerical stepping stones we'll need – in addition to **IRS** from Section 3.2 – to state a randomized diagonalization algorithm in Section 4.2.

4.1.1 RURV and GRURV

We begin with the building block of divide-and-conquer that has to this point been overlooked: the rank-revealing factorization used to obtain the matrices U_R and U_L from the corresponding sepctral projectors P_R and P_L. As mentioned in Chapters 1 and 3, we make use of a randomized URV factorization **RURV** introduced by Demmel, Dumitriu, and Holtz [38], which is stated below as Algorithm 4. This randomized algorithm is simple to implement, backwards stable [12, Theorem 4.5], and capable of producing strongly rank-revealing factorizations (in the sense of Gu and Eisenstat [66]).

In Chapter 1, we characterized a factorization $A = URV$ with $R = \left(\begin{smallmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{smallmatrix}\right)$ as

Algorithm 4. Randomized Rank-Revealing Factorization (**RURV**)
Input: $A \in \mathbb{C}^{n\times n}$.
Output: U unitary matrix, R upper triangular matrix, and V Haar such that $A = URV$ is a rank-revealing factorization of A.

```
Draw a random matrix $B$ with i.i.d. $\mathcal{N}_{\mathbb{C}}(0,1)$ entries
$[V, \widehat{R}] = \mathbf{QR}(B)$
$\widehat{A} = A \cdot V^H$
$[U, R] = \mathbf{QR}(\widehat{A})$
return $U, R, V$
```

rank-revealing if $\sigma_k(R_{11})$ and $||R_{22}||_2$ were good approximations to $\sigma_k(A)$ and $\sigma_{k+1}(A)$, respectively (assuming A has effective rank k and $R_{11} \in \mathbb{C}^{k\times k}$). The following result of Ballard et al. [12] clarifies what "good" means in this context, demonstrating that **RURV** matches the best-known guarantees for deterministic rank-revealing factorizations.

Theorem 4.1.1 (Ballard et al. 2019)**.** *Let R be the triangular matrix produced by applying exact arithmetic* **RURV** *to $A \in \mathbb{C}^{n\times n}$, where*

$$R = \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix}$$

for $R_{11} \in \mathbb{C}^{k\times k}$. Assume that $k, n-k > 30$. Then with probability $1-\delta$ the following occur:

$$\begin{aligned} \frac{\delta}{2.02}\frac{\sigma_k(A)}{\sqrt{k(n-k)}} &\le \sigma_k(R_{11}) \le \sigma_k(A) \\ \sigma_{k+1}(A) &\le \sigma_1(R_{22}) \le 2.02\frac{\sqrt{k(n-k)}}{\delta}\sigma_{k+1}(A) \\ ||R_{11}^{-1}R_{12}||_2 &\le \frac{6.1\sqrt{k(n-k)}}{\delta} + \frac{\sigma_{k+1}(A)}{\sigma_k(A)}\frac{50\sqrt{k^3(n-k)^3}}{\delta^3}. \end{aligned}$$

The proof of Theorem 4.1.1 boils down to bounding the smallest singular value of a $k \times k$ block of a Haar unitary matrix via [38, Theorem 5.2]. The requirement $k, n-k > 30$ comes from the bound used by Ballard et al. [12, Corollary 3.4]. Banks et al. subsequently demonstrated that this can be relaxed [16, Proposition C.3], although the fundamental guarantees are the same: with high probability **RURV** produces a factorization such that

Algorithm 5. Generalized Randomized Rank-Revealing Factorization (**GRURV**)
Input: k a positive integer, $A_1, A_2, \ldots, A_k \in \mathbb{C}^{n\times n}$, and $m_1, m_2, \ldots, m_k \in \{1, -1\}$.
Output: U unitary, $R_1, R_2, \ldots, R_k$ upper triangular, and V Haar such that $UR_1^{m_1}R_2^{m_2}\cdots R_k^{m_k}V$ is a rank-revealing factorization of $A_1^{m_1}A_2^{m_2}\cdots A_k^{m_k}$.

```
if $m_k = 1$ then
    $[U, R_k, V] = \textbf{RURV}(A_k)$
else
    $[U, L_k, V] = \textbf{RULV}(A_k^H)$
    $R_k = L_k^H$
end if
$U_{\text{current}} = U$
for $i = k-1 : 1$ do
    if $m_i = 1$ then
        $[U, R_i] = \textbf{QR}(A_i \cdot U_{\text{current}})$
        $U_{\text{current}} = U$
    else
        $[U, R_i] = \textbf{RQ}(U_{\text{current}}^H \cdot A_i)$
        $U_{\text{current}} = U^H$
    end if
end for
return $U_{\text{current}}$, optionally $R_1, R_2, \ldots, R_k$, $V$
```

$\sigma_k(R_{11})$ and $||R_{22}||_2$ are at worst a multiplicative factor of $O(\sqrt{k(n-k)})$ away from $\sigma_k(A)$ and $\sigma_{k+1}(A)$, respectively.

Since we use **IRS**, the projectors we would like to apply **RURV** to in divide-and-conquer are not simple matrices but rather products of the form $A^{-1}B$ or AB^{-1} (i.e., the approximate projectors $(A_p + B_p)^{-1}A_p$ and $A_p^H(A_p + B_p)^{-H}$). Given that we avoid inversion, explicitly forming either of these products is not an option. Instead, a generalized version of **RURV** – referred to as **GRURV** and presented here as Algorithm 5 – allows us to apply **RURV** to an arbitrary product of matrices and their inverses. Note that in this routine, **RULV** is a version of **RURV** that replaces the QR factorization in line 4 of Algorithm 4 with QL.

GRURV was first introduced in a technical report of Ballard, Demmel, and Dumitriu [11] specifically with the purpose of applying **RURV** to spectral projectors found by **IRS**. Importantly, exact-arithmetic **GRURV** is essentially equivalent to applying exact arithmetic **RURV** to the corresponding product [12, Theorem 5.2], which allows us to access guarantees like Theorem 4.1.1 for **GRURV** in exact arithmetic without any additional effort.

While Theorem 4.1.1 implies that **RURV** is theoretically optimal, we note that it often underperforms empirically. This has prompted efforts to develop randomized alternatives (see for example the powerURV work of Gopal and Martinsson [61]). Of course, a benefit of **RURV** is its simplicity, which implies that it is both stable and efficient, including from a communication perspective. Nevertheless, using a different rank-revealing factorization may allow for P_R and P_L to be computed to lower accuracy while still correctly estimating their rank.

4.1.2 DEFLATE

We consider next an algorithm that combines **IRS** and **GRURV** to compute the matrices U_L and U_R, whose orthonormal columns span right/left deflating subspaces of (A, B). Such a routine was first stated as **RGNEP** by Ballard, Demmel, and Dumitriu [11, Algorithm 4], albeit in a different form than we present here. In particular, **RGNEP** assumed no knowledge of the number of eigenvalues of (A, B) inside/outside the unit circle (equivalently, the rank of the corresponding spectral projectors), instead multiplying by the full $n \times n$ unitary matrices produced by **GRURV** and deciding where to split the problem to minimize certain matrix norms. Since we will have access to information about the rank of the projectors being computed,[1] we state an alternative **DEFLATE** (Algorithm 6), which simply takes the first k columns of the matrices computed by **GRURV**.

For **DEFLATE** to succeed, we need to know that the first k columns of the U-factor produced by **GRURV** span the range of the rank-k product it is applied to (with high probability). We would also like a guarantee that the result for an approximate

[1] This is how dividing grid lines are selected, where the rank is first computed by a separate, independent application of **IRS** and **GRURV**.

Algorithm 6. Deflating Subspace Finder (**DEFLATE**)
Input: $A, B \in \mathbb{C}^{n\times n}$, positive integers p and k
Requires: $k \leq n$; (A, B) has no eigenvalues on the unit circle and exactly k eigenvalues outside it.
Output: $U_R^{(k)}, U_L^{(k)} \in \mathbb{C}^{k\times n}$ with orthonormal columns that approximately span right and left deflating subspaces of (A, B).

```
$[A_p, B_p] = \mathbf{IRS}(A, B, p)$
$U_R = \mathbf{GRURV}(2, A_p + B_p, A_p, -1, 1)$
$[A_p, B_p] = \mathbf{IRS}(A^H, B^H, p)$
$U_L = \mathbf{GRURV}(2, A_p^H, (A_p + B_p)^H, 1, -1)$
$U_R^{(k)} = U_R(:, 1:k)$
$U_L^{(k)} = U_L(:, 1:k)$
return $U_R^{(k)}$, $U_L^{(k)}$
```

projector is close to that of a true spectral projector. Equivalent results for **RURV** are already known. In that case, the intuition is fairly simple: multiplying by the Haar matrix in line 3 of **RURV** "mixes" the columns of A, distributing information so that the first k columns of A – and therefore the first k columns of U – are likely to span range(A).[2] Additionally, we have the following perturbation result in exact arithmetic due to Banks et al. [16, Poposition C.12].

Theorem 4.1.2 (Banks et al. 2022)**.** *Let $A, A' \in \mathbb{C}^{n\times n}$ with $||A - A'||_2 \leq \delta$ and rank$(A) =$ rank$(A^2) = k$. Let T and S contain the first k columns of the U-factors produced by applying exact arithmetic **RURV** to A and A' respectively. Then for any $\theta \in (0, 1)$ with probability $1 - \theta^2$ there exists a unitary $W \in \mathbb{C}^{k\times k}$ such that*

$$||S - TW^H||_2 \leq \sqrt{\frac{8\sqrt{k(n-k)}}{\sigma_k(T^H A T)}} \cdot \sqrt{\frac{\delta}{\theta}}$$

Letting A be a rank-k spectral projector (in which case $\sigma_k(T^H A T) = 1$), Theorem 4.1.2 says that the first k columns of the U-factor of an approximation A' of A are close to a rotation/reflection of the first k columns of the U-factor of A, provided $||A - A'||_2$ is sufficiently small.

[2] We can identify this as another benefit of randomization, which is independent of the regularizing effect of initial perturbations.

Recalling that exact arithmetic **GRURV** is equivalent to exact arithmetic **RURV** on the corresponding product, these results generalize directly. Not only can we say that almost surely the first k columns of the U-factor produced by **GRURV** span the range of the product, but a perturbation result similar to Theorem 4.1.2 holds. Combining these with our analysis of **IRS** yields the following exact arithmetic guarantee for **DEFLATE**.

Theorem 4.1.3. *Suppose* (A, B)*,* p*, and* k *satisfy the requirements of* **DEFLATE***, where* p *is large enough to ensure error in repeated squaring is at most* δ *in lines 1 and 3. Let* $U_R^{(k)}$ *and* $U_L^{(k)}$ *be the outputs of running this algorithm in exact arithmetic. Then for any* $\nu \in (0, 1)$ *there exist* $U_R, U_L \in \mathbb{C}^{n\times k}$ *with orthonormal columns spanning right and left deflating subspaces of* (A, B) *respectively such that*

$$||U_R^{(k)} - U_R||_2,\ ||U_L^{(k)} - U_L||_2 \leq \sqrt{8\sqrt{k(n-k)}}\sqrt{\frac{\delta}{\nu}}$$

with probability at least $1 - 2\nu^2$*.*

Proof. Consider first $U_R^{(k)}$ and let A_p and B_p be the outputs of applying p steps of exact arithmetic repeated squaring to A and B. We know $(A_p + B_p)^{-1}A_p$ approaches a projector $P_{R,|z|>1}$ onto the right deflating subspace spanned by eigenvectors of (A, B) with eigenvalues outside the unit circle. Let $URV = P_{R,|z|>1}$ be a rank-revealing factorization of $P_{R,|z|>1}$ obtained via exact arithmetic **RURV** and let $U^{(k)}$ contain the first k columns of U. Since k is the number of eigenvalues with modulus greater than one, we know $k = \text{rank}(P_{R,|z|>1})$ and moreover $\text{range}(U^{(k)}) = \text{range}(P_{R,|z|>1})$ almost surely. Thus, since p is large enough to ensure $||(A_p + B_p)^{-1}A_p - P_{R,|z|>1}||_2 \leq \delta$ and exact arithmetic **GRURV** satisfies the same guarantees as exact **RURV**, we have by Theorem 4.1.2 that with probability at least $1 - \nu^2$ there exists a unitary $W \in \mathbb{C}^{k\times k}$ such that

$$||U_R^{(k)} - U^{(k)}W^H||_2 \leq \sqrt{8\sqrt{k(n-k)}}\sqrt{\frac{\delta}{\nu}}. \tag{4.1}$$

Setting $U_R = U^{(k)}W^H$, we have the bound on $||U_R^{(k)} - U_R||_2$ with probability at least $1 - \nu^2$. Repeating the same argument for $U_L^{(k)}$, using this time the fact that $[A_p, B_p] =$

IRS(A^H, B^H, p) implies

$$||A_p^H (A_p + B_P)^{-H} - P_{L,|z|>1}||_2 \leq \delta \tag{4.2}$$

for $P_{L,|z|>1}$ a projector onto the left deflating subspace corresponding to $P_{R,|z|>1}$, we obtain the remaining bound, also with probability at least $1-\nu^2$. Taking a union bound completes the proof. □

4.1.3 EIG

Given **IRS**, **GRURV**, and **DEFLATE**, we can now state our main algorithmic contribution: **EIG** (Algorithm 7), which applies inverse-free divide-and-conquer to a pencil (A, B) under the assumption that $\Lambda_\epsilon(A, B)$ is shattered with respect to a grid g (contained in the square of side length five centered at the origin). In practice, this will be our perturbed and scaled pencil $(\widetilde{A}, n^\alpha \widetilde{B})$ – hence the norm assumption on B in Algorithm 7 has a factor of n^α attached – and the shattering grid guaranteed by Theorem 2.3.1.

Before proving exact arithmetic guarantees for **EIG**, we first provide a high-level overview of the algorithm. Throughout, we rely on the analysis of each of the building blocks from the previous subsections (and Section 3.2).

1. Since **EIG** calls itself recursively, the first three lines check for our stopping criteria. We choose to continue divide-and-conquer until the pencil is 1×1, though as mentioned in Chapter 1 we could choose instead to stop once the pencil is small enough to be handled by another method.

2. The next four lines (4-7) set parameters for the algorithm. Most importantly, they determine how many steps of repeated squaring need to be taken to achieve the desired accuracy (i.e., the value of p in line 7).

3. Lines 8-15 execute a search over g for a grid line that sufficiently splits the spectrum, which here means separating at least a fifth of the eigenvalues on each side. Since g shatters $\Lambda_\epsilon(A, B)$, a grid line that sufficiently splits the spectrum always exists.

Algorithm 7. Divide-and-Conquer Eigensolver (**EIG**)
Input: $n \in \mathbb{N}_+$, $A, B \in \mathbb{C}^{m \times m}$, $\epsilon > 0$, $\alpha > 1$, g an $s_1 \times s_2$ grid with box size ω, $\beta > 0$ a desired eigenvector accuracy, and $\theta \in (0,1)$ a failure probability.
Requires: $m \leq n$, $||A||_2 \leq 3$, $||B||_2 \leq 3n^\alpha$, $g \subset \{z : |\mathrm{Re}(z)|, |\mathrm{Im}(z)| < 5\}$, and $\Lambda_\epsilon(A, B)$ shattered with respect to g.
Output: T an invertible matrix and (D_1, D_2) a diagonal pencil. The eigenvalues of (D_1, D_2) each share a grid box of g with a unique eigenvalue of (A, B) and each column of T is an approximate right unit eigenvector of (A, B).

```
 1: **if** $m = 1$ **then**
 2:     $T = 1;\ D_1 = A;\ D_2 = B$
 3: **else**
 4:     $\zeta = 2\left(\lfloor \log_2(\max\{s_1, s_2\}) + 1 \rfloor\right)$
 5:     $\eta = \min\left\{ \frac{4\pi}{315\sqrt{8}} \frac{\beta\epsilon^2}{\omega n^\alpha},\ \frac{1}{2\log_{5/4}(n)} \right\}$
 6:     $\delta = \min\left\{ \sqrt{\frac{\theta}{10}} \frac{\epsilon^2}{7200 n^{2\alpha+3}},\ \frac{\theta}{2(\theta + 10n^6\zeta)},\ \sqrt{\frac{\theta}{10}} \frac{\eta^2}{288 n^{2\alpha+3}} \right\}$
 7:     $p = \left\lceil \max\left\{ 7,\ -2\log_2\left(-\frac{1}{2}\log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)\right),\ 1 + \log_2\left[\frac{\log_2\left(\frac{\delta\pi\epsilon}{12n^\alpha m\omega + \delta\pi\epsilon}\right)}{\log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)}\right] \right\} \right\rceil$
 8:     Choose a grid line $\mathrm{Re}(z) = h$ of $g$
 9:     $(\mathcal{A}, \mathcal{B}) = (A - (h-1)B, A - (h+1)B)$
10:     $[A_p, B_p] = \mathbf{IRS}(\mathcal{A}, \mathcal{B}, p)$
11:     $[U, R_1, R_2, V] = \mathbf{GRURV}(2, A_p + B_p, A_p, -1, 1)$
12:     $k = \#\left\{ i : \left| \frac{R_2(i,i)}{R_1(i,i)} \right| \geq \sqrt{\frac{\theta}{10\zeta}} \frac{1-\delta}{n^3} \right\}$
13:     **if** $k < \frac{1}{5}m$ or $k > \frac{4}{5}m$ **then**
14:         Return to step 8 and choose a new grid line, executing a binary search if
            necessary. If this fails, search over horizontal grid lines $\mathrm{Im}(z) = h$.
15:     **else**
16:         $[U_R^{(k)}, U_L^{(k)}] = \mathbf{DEFLATE}(\mathcal{A}, \mathcal{B}, p, k)$
17:         $(\mathcal{A}, \mathcal{B}) = (A - (h+1)B, A - (h-1)B)$
18:         $[U_R^{(m-k)}, U_L^{(m-k)}] = \mathbf{DEFLATE}(\mathcal{A}, \mathcal{B}, p, m-k)$
19:
                $(A_{11}, B_{11}) = \left( (U_L^{(k)})^H A U_R^{(k)},\ (U_L^{(k)})^H B U_R^{(k)} \right)$
                $(A_{22}, B_{22}) = \left( (U_L^{(m-k)})^H A U_R^{(m-k)},\ (U_L^{(m-k)})^H B U_R^{(m-k)} \right)$
20:         $g_R = \{z \in g : \mathrm{Re}(z) > h\};\ g_L = \{z \in g : \mathrm{Re}(z) < h\}$
21:         $[\widehat{T}, \widehat{D}_1, \widehat{D}_2] = \mathbf{EIG}(n, A_{11}, B_{11}, \frac{4}{5}\epsilon, \alpha, g_R, \frac{1}{3}\beta, \theta)$
22:         $[\dot{T}, \dot{D}_1, \dot{D}_2] = \mathbf{EIG}(n, A_{22}, B_{22}, \frac{4}{5}\epsilon, \alpha, g_L, \frac{1}{3}\beta, \theta)$
23:
            $T = \begin{pmatrix} U_R^{(k)} & U_R^{(m-k)} \end{pmatrix} \begin{pmatrix} \widehat{T} & 0 \\ 0 & \dot{T} \end{pmatrix} \quad D_1 = \begin{pmatrix} \widehat{D}_1 & 0 \\ 0 & \dot{D}_1 \end{pmatrix} \quad D_2 = \begin{pmatrix} \widehat{D}_2 & 0 \\ 0 & \dot{D}_2 \end{pmatrix}$
24:     **end if**
25: **end if**
26: **return** $T, D_1, D_2$
```

4. We check a line $\mathrm{Re}(z) = h$ of the grid by applying the Möbius transformation $S(z) = \frac{z-(h-1)}{z-(h+1)}$ (line 9). S maps the grid line to the unit circle while sending the half plane $\{z : \mathrm{Re}(z) < h\}$ inside the unit disk. Applying this transformation to (A, B) sends eigenvalues to the left/right of the dividing line inside/outside the unit circle, respectively, without changing eigenvectors.

5. Lines 10 and 11 apply **IRS** and **GRURV** to the transformed pencil $(\mathcal{A}, \mathcal{B})$. This produces a rank-revealing factorization $UR_1^{-1}R_2V$ of the approximate projector onto the right deflating subspace corresponding to eigenvectors of $(\mathcal{A}, \mathcal{B})$ with eigenvalues outside the unit disk (equivalently eigenvectors of (A, B) with eigenvalues to the right of the selected grid line).

6. In line 12, we leverage the rank-revealing guarantees of **RURV**, and by extension **GRURV**, to read off the rank of the approximate projector. Note that we do this without forming $R_1^{-1}R_2$. The grid line is selected if this rank is between $\frac{1}{5}m$ and $\frac{4}{5}m$, where m is the size of the pencil (which shrinks as we recur).

7. In line 8 we assume that the grid line is vertical, however it is possible that only a horizontal grid line sufficiently splits the spectrum. This is covered in line 14. The remainder of the algorithm similarly assumes the split is vertical; the following changes apply if a split is made with the horizontal grid line $\mathrm{Im}(z) = h$.

   - Line 9: $(\mathcal{A}, \mathcal{B}) = (A - i(h-1)B, A - i(h+1)B)$.
   - Line 17: $(\mathcal{A}, \mathcal{B}) = (A - i(h+1)B, A - i(h-1)B)$,
   - Line 20: $g_R = \{z : \mathrm{Im}(z) > h\}$ and $g_L = \{z : \mathrm{Im}(z) < h\}$.

8. Once a dividing line has been identified, **DEFLATE** is called twice to compute orthonormal bases for both sets of deflating subspaces. To recover eigenvectors corresponding to eigenvalues to the left of the line, we apply the alternative Möbius transformation $S(z) = \frac{z-(h+1)}{z-(h-1)}$.

9. In line 19 we compute the next pair of subproblems. We then pass these to **EIG** along with pieces of the grid g and slightly adjusted parameters. In particular, note that the ϵ for which $\Lambda_\epsilon(A,B)$ is shattered shrinks by a factor of $\frac{4}{5}$ at each step. As we will see, this is necessary to guarantee shattering since $U_R^{(k)}$, $U_L^{(k)}$, $U_R^{(m-k)}$ and $U_L^{(m-k)}$ are only approximations of the matrices used in Lemma 1.3.5.

10. Once the recursion finishes, **EIG** reconstructs a diagonal pencil (D_1, D_2) and a set of approximate right eigenvectors T (line 23).

With this outline in mind, we are now ready to state and prove our main guarantee for **EIG** (in exact arithmetic).

Theorem 4.1.4. *Let (A,B) and g be a pencil and grid satisfying the requirements of* **EIG**. *Then for any choice of $\theta \in (0,1)$ and $\beta > 0$, exact-arithmetic* **EIG** *applied to (A,B) and g satisfies the following with probability at least $1-\theta$.*

1. *The recursive procedure converges and each eigenvalue of the diagonal pencil (D_1, D_2) shares a grid box with a unique eigenvalue of (A,B).*

2. *If $\sigma_n(B) \geq 1$, each column t_i of T satisfies $||t_i - v_i||_2 \leq \beta$ for some true unit right eigenvector v_i of (A,B).*

Proof. We start by bounding the probability that the first guarantee does not hold. Since **EIG** calls itself recursively, we do this by bounding the probability of failure for one step of divide and conquer. In this context, success requires two events: first, a dividing line that sufficiently splits the spectrum must be found; second, the subsequent calls to **EIG** must be valid, meaning the inputs satisfy the listed properties.

Computing the probabilities that these occur is fairly lengthy, so to improve readability we number the steps in the proof and provide in bold a description of what each step accomplishes. Throughout, we use the assumptions on the inputs – i.e. $A, B \in \mathbb{C}^{m \times m}$ with $m \leq n$, $||A||_2 \leq 3$, $||B||_2 \leq 3n^\alpha$, and $\Lambda_\epsilon(A,B)$ is shattered with respect to the grid g, which is $s_1 \times s_2$ consisting of boxes of size ω.

Step One: Any transformed pencil $(\mathcal{A}, \mathcal{B})$ in EIG satisfies $d_{(\mathcal{A},\mathcal{B})} \geq \frac{2}{5}\epsilon$.

Consider first a vertical grid line $\mathrm{Re}(z) = h$ and $(\mathcal{A}, \mathcal{B}) = (A - (h-1)B, A - (h+1)B)$ as in line 9. Suppose $z \in \Lambda_{\epsilon'}(\mathcal{A}, \mathcal{B})$ for some $\epsilon' > 0$. In this case, there exist matrices E and F with $||E||_2, ||F||_2 \leq \epsilon'$ such that z is an eigenvalue of $(\mathcal{A} + E,\ \mathcal{B} + F)$. If we apply the Möbius transformation $S(z) = \frac{(h+1)z-(h-1)}{z-1}$ to this pencil, we observe that $S(z)$ is an eigenvalue of

$$((h+1)(\mathcal{A}+E) - (h-1)(\mathcal{B}+F),\ (\mathcal{A}+E) - (\mathcal{B}+F)), \tag{4.3}$$

or, equivalently,

$$(2A + (h+1)E - (h-1)F,\ 2B + E - F). \tag{4.4}$$

Dividing by two, we conclude that $S(z)$ is an eigenvalue of $(A + \frac{h+1}{2}E - \frac{h-1}{2}F,\ B + \frac{1}{2}E - \frac{1}{2}F)$, where

$$\left|\left|\frac{h+1}{2}E - \frac{h-1}{2}F\right|\right|_2 \leq \frac{|h+1|}{2}||E||_2 + \frac{|h-1|}{2}||F||_2 \leq \frac{\epsilon'}{2}\left(|h+1| + |h-1|\right) \tag{4.5}$$

and

$$\left|\left|\frac{1}{2}E - \frac{1}{2}F\right|\right|_2 \leq \frac{1}{2}(||E||_2 + ||F||_2) \leq \epsilon'. \tag{4.6}$$

Thus, $S(z)$ belongs to $\Lambda_{\epsilon''}(A, B)$ for

$$\epsilon'' = \max\left\{\epsilon',\ \frac{\epsilon'}{2}(|h+1| + |h-1|)\right\} \leq 5\epsilon', \tag{4.7}$$

which means the pre-image of $\Lambda_{\epsilon/5}(\mathcal{A}, \mathcal{B})$ under S^{-1} is contained in $\Lambda_\epsilon(A, B)$.

Since $\Lambda_\epsilon(A, B)$ is shattered with respect to g and therefore does not intersect the dividing line $\mathrm{Re}(z) = h$, we conclude that $\Lambda_{\epsilon/5}(\mathcal{A}, \mathcal{B})$ does not intersect the unit circle. By Definition 3.2.2, we obtain $d_{(\mathcal{A},\mathcal{B})} \geq \frac{2}{5}\epsilon$. Making a similar argument for the transformed pencil in line 17 – or in the case of a horizontal dividing line $\mathrm{Im}(z) = h$ – yields $d_{(\mathcal{A},\mathcal{B})} \geq \frac{2}{5}\epsilon$ for any $(\mathcal{A}, \mathcal{B})$ appearing in **EIG**. In the next step, we will use this lower bound to control the error in repeated squaring.

Step Two: The choice of p guarantees that the error in IRS is at most δ. Consider the first call to **IRS**, which applies repeated squaring to the transformed pencil $(\mathcal{A}, \mathcal{B})$. By Theorem 3.2.3, we know that as long as $p \geq \log_2 \left[\frac{||(\mathcal{A},\mathcal{B})||_2 - d_{(\mathcal{A},\mathcal{B})}}{d_{(\mathcal{A},\mathcal{B})}}\right]$ then

$$||(A_p + B_p)^{-1}A_p - P_{R,|z|>1}||_2 \leq ||P_{R,|z|>1}||_2 \frac{2^{p+3}\left(1 - \frac{d_{(\mathcal{A},\mathcal{B})}}{||(\mathcal{A},\mathcal{B})||_2}\right)^{2^p}}{\max\left\{0, 1 - 2^{p+2}\left(1 - \frac{d_{(\mathcal{A},\mathcal{B})}}{||(\mathcal{A},\mathcal{B})||_2}\right)^{2^p}\right\}}. \tag{4.8}$$

We just showed $d_{(\mathcal{A},\mathcal{B})} \geq \frac{2}{5}\epsilon$ and

$$||(\mathcal{A}, \mathcal{B})||_2 \leq ||\mathcal{A}||_2 + ||\mathcal{B}||_2 \leq 2||A||_2 + (|h-1| + |h+1|)||B||_2 \leq 42n^\alpha, \tag{4.9}$$

so to satisfy $p \geq \log_2 \left[\frac{||(\mathcal{A},\mathcal{B})||_2 - d_{(\mathcal{A},\mathcal{B})}}{d_{(\mathcal{A},\mathcal{B})}}\right]$ it is sufficient to take $p \geq \log_2\left(\frac{105n^\alpha}{\epsilon} - 1\right)$. Similarly, to eliminate the maximum from the denominator of (4.8) it is sufficient to take $p \geq -2\log_2\left(\log_2\left(\frac{105n^\alpha}{105n^\alpha - \epsilon}\right)\right)$, provided $p > 6$ (which allows us to simplify the bounds by assuming $\log_2(p+2) < \frac{1}{2}p$).

With this in mind, we can now turn to bounding the right hand side of (4.8). First, we upper bound $||P_{R,|z|>1}||_2$. Recall,

$$P_{R,|z|>1} = V\begin{pmatrix} 0 & 0 \\ 0 & I_r \end{pmatrix}V^{-1} = \sum_{j=m-r+1}^{m} v_j w_j^H \tag{4.10}$$

for $r = \text{rank}(P_{R,|z|>1})$ and V a matrix that diagonalizes $B^{-1}A$. v_i and w_i^H are the columns of V and rows of V^{-1} respectively, scaled so that $w_i^H v_i = 1$ with $v_{m-r+1}, \ldots, v_m$ corresponding to eigenvalues of (A, B) to the right of $\text{Re}(z) = h$. Since $\Lambda_\epsilon(A, B)$ is shattered with respect to g, each of these eigenvalues $\lambda_{m-r+1}, \ldots, \lambda_m$ is contained in a separate grid box of g. If Γ_i is the contour of the grid box containing λ_i, this means

$$v_j w_j^H = \frac{1}{2\pi i}\oint_{\Gamma_j}(z - B^{-1}A)^{-1}dz \tag{4.11}$$

and therefore

$$P_{R,|z|>1} = \frac{1}{2\pi i}\sum_{j=m-r+1}^{m}\oint_{\Gamma_j}(z - B^{-1}A)^{-1}dz = \frac{1}{2\pi i}\sum_{j=m-r+1}^{m}\oint_{\Gamma_j}(zB - A)^{-1}Bdz. \tag{4.12}$$

Thus, by the triangle inequality,

$$||P_{R,|z|>1}||_2 \leq \frac{1}{2\pi} \sum_{j=m-r+1}^{m} \left|\left| \oint_{\Gamma_j} (zB - A)^{-1} B dz \right|\right|_2 . \tag{4.13}$$

Moreover, applying the ML-inequality (2.33) to each term in this sum, we have

$$\begin{aligned} ||P_{R,|z|>1}||_2 &\leq \frac{1}{2\pi} \sum_{j=m-r+1}^{m} 4\omega \sup_{z\in\Gamma_j} ||(zB - A)^{-1} B||_2 \\ &\leq \frac{2\omega ||B||_2}{\pi} \sum_{j=m-r+1}^{m} \sup_{z\in\Gamma_j} ||(A - zB)^{-1}||_2. \end{aligned} \tag{4.14}$$

Since shattering guarantees $\Lambda_\epsilon(A, B) \cap \Gamma_j = \emptyset$ and therefore $||(A - zB)^{-1}||_2 \leq \frac{1}{\epsilon(1+|z|)} \leq \epsilon^{-1}$ for all $z \in \Gamma_j$, we conclude $||P_{R,|z|>1}||_2 \leq \frac{2\omega r ||B||_2}{\pi\epsilon}$. Finally using the fact that $||B||_2 \leq 3n^\alpha$ and $r \leq m$, we have a final upper bound $||P_{R,|z|>1}||_2 \leq \frac{6n^\alpha m\omega}{\pi\epsilon}$.

Combining this bound with $d_{(\mathcal{A},\mathcal{B})} \geq \frac{2}{5}\epsilon$ and $||(\mathcal{A}, \mathcal{B})||_2 \leq 42n^\alpha$, (4.8) becomes

$$||(A_p + B_p)^{-1} A_p - P_{R,|z|>1}||_2 \leq \frac{6n^\alpha m\omega}{\pi\epsilon} \cdot \frac{2^{p+3}\left(1 - \frac{\epsilon}{105n^\alpha}\right)^{2^p}}{1 - 2^{p+2}\left(1 - \frac{\epsilon}{105n^\alpha}\right)^{2^p}} \tag{4.15}$$

for p sufficiently large (i.e., following the bounds derived above). Thus, we obtain $||(A_p + B_p)^{-1} A_p - P_{R,|z|>1}||_2 \leq \delta$ by taking

$$\frac{6n^\alpha m\omega}{\pi\epsilon} \cdot \frac{2^{p+3}\left(1 - \frac{\epsilon}{105n^\alpha}\right)^{2^p}}{1 - 2^{p+2}\left(1 - \frac{\epsilon}{105n^\alpha}\right)^{2^p}} \leq \delta \tag{4.16}$$

which is equivalent to

$$2^p \left[\frac{p+2}{2^p} + \log_2 \left(1 - \frac{\epsilon}{105n^\alpha}\right) \right] \leq \log_2 \left(\frac{\delta\pi\epsilon}{12n^\alpha m\omega + \delta\pi\epsilon} \right). \tag{4.17}$$

Using again the assumption that $p > 6$ and further taking $p \geq -2\log_2\left(-\frac{1}{2}\log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)\right)$ to ensure $\frac{p+2}{2^p} \leq -\frac{1}{2}\log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)$, we get the desired accuracy as long as

$$2^{p-1} \log_2 \left(1 - \frac{\epsilon}{105n^\alpha}\right) \leq \log_2 \left(\frac{\delta\pi\epsilon}{12n^\alpha m\omega + \delta\pi\epsilon} \right) \tag{4.18}$$

which yields a final bound $p \geq 1 + \log_2 \left[\log_2\left(\frac{\delta\pi\epsilon}{12n^\alpha m\omega + \delta\pi\epsilon}\right) / \log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)\right]$.

In the preceding analysis, we derived the following four bounds on p:

- $p \geq \log_2\left(\frac{105n^\alpha}{\epsilon} - 1\right)$ (to allow us to apply Theorem 3.2.3).
- $p \geq -2\log_2\left(\log_2\left(\frac{105n^\alpha}{105n^\alpha - \epsilon}\right)\right)$ (to eliminate the maximum from the error bound).
- $p \geq -2\log_2\left(-\frac{1}{2}\log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)\right)$ (to simplify the upper bound in (4.8)).
- $p \geq 1 + \log_2\left[\log_2\left(\frac{\delta\pi\epsilon}{12n^\alpha m\omega + \delta\pi\epsilon}\right) / \log_2\left(1 - \frac{\epsilon}{105n^\alpha}\right)\right]$ (to ensure an error of at most δ given the other three bounds).

Since the third bound is always at least as large as the first two – and since we also assumed $p > 6$ – we conclude that $||(A_p + B_p)^{-1}A_p - P_{R,|z|>1}||_2 \leq \delta$ for the p chosen in line 7 of **EIG**. Note that since $||\mathcal{A}^H||_2 = ||\mathcal{A}||_2$, $||\mathcal{B}^H||_2 = ||\mathcal{B}||_2$, $d_{(\mathcal{A}^H,\mathcal{B}^H)} = d_{(\mathcal{A},\mathcal{B})}$, and $\Lambda_\epsilon(A^H, B^H)$ is shattered with respect to the grid $g^H = \{\overline{z} : z \in g\}$, the same argument guarantees that running repeated squaring on $\mathcal{A}^H$, and $\mathcal{B}^H$ has error at most δ. Similarly, the same results hold for the other transformed pencils in lines 14 and 17 since in all cases $||(\mathcal{A}, \mathcal{B})||_2 \leq 42n^\alpha$ and $d_{(\mathcal{A},\mathcal{B})} \geq \frac{2}{5}\epsilon$.

Step Three: A dividing line that sufficiently splits the spectrum exists.

A dividing line sufficiently splits the spectrum if it separates at least $\frac{1}{5}m$ of the m eigenvalues of (A, B). Suppose that no vertical line of g does this. In this case, there exists adjacent vertical lines between which more than $\frac{3}{5}m$ eigenvalues lie. Since no eigenvalues share the same grid box, this implies that a horizontal grid line must sufficiently split the spectrum.

Step Four: With probability at least $1 - \frac{\theta}{10n^4}$, EIG finds a dividing line that separates exactly k eigenvalues to the right such that $\frac{1}{5}m \leq k \leq \frac{4}{5}m$.

To obtain a lower bound on this probability, we first compute the probability that for any grid line $\text{Re}(z) = h$ the value of k at line 12 is equal to the number of eigenvalues of (A, B) to the right of the line.

Suppose (A, B) has r eigenvalues to the right of $\text{Re}(z) = h$. In this case, we know in line 10 that $(A_p + B_p)^{-1}A_p$ approaches a rank-r projector $P_{R,|z|>1}$. Now k is obtained by computing a rank-revealing factorization $U_R R_1^{-1} R_2 V = (A_p + B_p)^{-1} A_p$ via **GRURV** and then counting the diagonal entries of $R_1^{-1} R_2$ that have modulus above a certain threshold. With this in mind, write

$$R_1^{-1} R_2 = \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix} \tag{4.19}$$

for R_{11} an $r \times r$ matrix. Since we are working in exact arithmetic, **GRURV** satisfies all of the guarantees of exact arithmetic **RURV**. In particular, extending [38, Theorem 5.2] to complex matrices,

$$\sigma_r(R_{11}) \geq \sigma_r((A_p + B_p)^{-1} A_p) \sigma_r(X_{11}) \tag{4.20}$$

where X_{11} is the upper left $r \times r$ block of $X = Q^H V^H$ for the SVD $(A_p + B_p)^{-1} A_p = P \Sigma Q^H$. Similarly, by [12, Lemma 4.1],

$$||R_{22}||_2 \leq \frac{\sigma_{r+1}((A_p + B_p)^{-1} A_p)}{\sigma_r(X_{11})}. \tag{4.21}$$

Now $||(A_p + B_p)^{-1} A_p - P_{R,|z|>1}||_2 \leq \delta$ as shown in step two above, so since the rank-r projector $P_{R,|z|>1}$ satisfies $\sigma_r(P_{R,|z|>1}) = 1$ and $\sigma_{r+1}(P_{R,|z|>1}) = 0$, we have by Lemma 1.7.2 $\sigma_r((A_p + B_p)^{-1} A_p) \geq 1 - \delta$ and $\sigma_{r+1}((A_p + B_p)^{-1} A_p) \leq \delta$. Moreover, since X is Haar unitary, we have by [16, Proposition C.3]

$$\mathbb{P}\left[\frac{1}{\sigma_r(X_{11})} \leq \frac{\sqrt{r(m-r)}}{\nu}\right] \geq 1 - \nu^2 \tag{4.22}$$

for any $\nu \in (0,1]$. Applying these to (4.20) and (4.21) with $\nu = \sqrt{\frac{\theta}{10\zeta} \frac{1}{n^2}}$ for $\zeta = 2(\lfloor \log_2(s) + 1 \rfloor)$ and $s = \max\{s_1, s_2\}$, we have

$$\sigma_r(R_{11}) \geq (1 - \delta) \sqrt{\frac{\theta}{10\zeta}} \frac{1}{n^2 \sqrt{r(m-r)}} \geq \sqrt{\frac{\theta}{10\zeta}} \frac{1 - \delta}{n^3} \tag{4.23}$$

and

$$||R_{22}||_2 \leq \delta \sqrt{r(m-r)} n^2 \sqrt{\frac{10\zeta}{\theta}} \leq \delta n^3 \sqrt{\frac{10\zeta}{\theta}} \tag{4.24}$$

with probability at least $1 - \frac{\theta}{10\zeta n^4}$. Since the eigenvalues of any matrix are bounded in modulus above and below by its singular values – and the eigenvalues of R_{11} and R_{22}

are their diagonal entries – we conclude $|R_{11}(i,i)| \geq \sqrt{\frac{\theta}{10\zeta}\frac{1-\delta}{n^3}}$ for all $1 \leq i \leq r$ while $|R_{22}(j,j)| \leq \delta n^3\sqrt{\frac{10\zeta}{\theta}}$ for all $1 \leq j \leq m-r$ with probability at least $1 - \frac{\theta}{10\zeta n^4}$. Since the requirement in line 6 that $\delta \leq \frac{\theta}{2(\theta+10n^6\zeta)}$ guarantees $\sqrt{\frac{\theta}{10\zeta}\frac{1-\delta}{n^3}} > \delta n^3\sqrt{\frac{10\zeta}{\theta}}$, this implies that $k = r$ with probability at least $1 - \frac{\theta}{10\zeta n^4}$

We have shown so far that with high probability the value of k at line 12 is equal to the number of eigenvalues to the right of the vertical dividing line selected four lines earlier. Since repeating this argument yields the same probability of success when checking a horizontal grid line and we know a dividing line that separates at least $\frac{1}{5}m$ eigenvalues must exist, we can therefore lower bound the probability that **EIG** finds a suitable line by requiring that the value of k is accurate for all lines that are checked. Since we do at most two binary searches to find a good enough grid line, we check at most ζ lines. By a union bound, we conclude that a suitable line is found, and k is computed accurately for that line, with probability at least $1 - \frac{\theta}{10n^4}$.

Step Five: Assuming k is computed correctly in line 12, there exists matrices $\widehat{U}_R^{(k)}, \widehat{U}_L^{(k)} \in \mathbb{C}^{k\times m}$ and $\widehat{U}_R^{(m-k)}, \widehat{U}_L^{(m-k)} \in \mathbb{C}^{m-k\times m}$ with orthonormal columns spanning corresponding right and left deflating subspaces of (A,B) such that

(a) $||U_R^{(k)} - \widehat{U}_R^{(k)}||_2 \leq \sqrt{\sqrt{\frac{10}{\theta}}8n^3\delta}$

(b) $||U_L^{(k)} - \widehat{U}_L^{(k)}||_2 \leq \sqrt{\sqrt{\frac{10}{\theta}}8n^3\delta}$

(c) $||U_R^{(m-k)} - \widehat{U}_R^{(m-k)}||_2 \leq \sqrt{\sqrt{\frac{10}{\theta}}8n^3\delta}$

(d) $||U_L^{(m-k)} - \widehat{U}_L^{(m-k)}||_2 \leq \sqrt{\sqrt{\frac{10}{\theta}}8n^3\delta}$

With probability at least $1 - \frac{2\theta}{5n^4}$.

This result comes from applying Theorem 4.1.3 twice with $\nu = \sqrt{\frac{\theta}{10n^4}}$ and taking a union bound. In both cases, we use the fact that p is large enough to guarantee error in repeated squaring is a most δ and consequently

$$\sqrt{8\sqrt{k(m-k)}}\sqrt{\frac{\delta}{\nu}} \leq \sqrt{8n\sqrt{\frac{10n^4}{\theta}}\delta} = \sqrt{\sqrt{\frac{10}{\theta}}8n^3\delta}. \tag{4.25}$$

Step Six: Union bound on events that guarantee success.

Let E_a, E_b, E_c, and E_d be events that correspond to the results (a), (b), (c), and (d) from step 5 and let E_k be the event that a sufficient dividing line is found and k computed accurately. We have just shown

$$\mathbb{P}(E_a \cap E_b \cap E_c \cap E_d \mid E_k) \geq 1 - \frac{2\theta}{5n^4}. \tag{4.26}$$

Since we also know $\mathbb{P}(E_k) \geq 1 - \frac{\theta}{10n^4}$, we conclude

$$\begin{aligned} \mathbb{P}(E_a \cap E_b \cap E_c \cap E_d \cap E_k) &= \mathbb{P}(E_a \cap E_b \cap E_c \cap E_d \mid E_k)\mathbb{P}(E_k) \\ &\geq \left(1 - \frac{2\theta}{5n^4}\right)\left(1 - \frac{\theta}{10n^4}\right) \\ &\geq 1 - \frac{\theta}{2n^4}. \end{aligned} \tag{4.27}$$

In the remainder of the proof, we will show that conditioning on these events guarantees success for one step of **EIG**.

Step Seven: If E_a, E_b, E_c, and E_d hold, then $\Lambda_{4\epsilon/5}(A_{11}, B_{11})$ and $\Lambda_{4\epsilon/5}(A_{22}, B_{22})$ are shattered with respect to g.

Let $\widehat{U}_R^{(k)}$, $\widehat{U}_L^{(k)}$, $\widehat{U}_R^{(m-k)}$, and $\widehat{U}_L^{(m-k)}$ be the matrices such that $||U_R^{(k)} - \widehat{U}_R^{(k)}||_2$, $||U_L^{(k)} - \widehat{U}_L^{(k)}||_2$, $||U_R^{(m-k)} - \widehat{U}_R^{(m-k)}||_2$, and $||U_L^{(m-k)} - \widehat{U}_L^{(m-k)}||_2$ are all bounded above by $\sqrt{\sqrt{\frac{10}{\theta}}8n^3\delta}$ as in step five. Since $\delta \leq \sqrt{\frac{\theta}{10}}\frac{\epsilon^2}{7200n^{2\alpha+3}}$, we can replace this upper bound with $\frac{\epsilon}{30n^\alpha}$. With this in mind, let

$$\begin{aligned} (\widehat{A}_{11}, \widehat{B}_{11}) &= \left((\widehat{U}_L^{(k)})^H A \widehat{U}_R^{(k)},\ (\widehat{U}_L^{(k)})^H B \widehat{U}_R^{(k)}\right) \\ (\widehat{A}_{22}, \widehat{B}_{22}) &= \left((\widehat{U}_L^{(m-k)})^H A \widehat{U}_R^{(m-k)},\ (\widehat{U}_L^{(m-k)})^H B \widehat{U}_R^{(m-k)}\right). \end{aligned} \tag{4.28}$$

Note that by Lemma 1.3.5, $\Lambda_\epsilon(\widehat{A}_{11}, \widehat{B}_{11})$ and $\Lambda_\epsilon(\widehat{A}_{22}, \widehat{B}_{22})$ are both contained in $\Lambda_\epsilon(A, B)$

and therefore shattered with respect to g. At the same time,

$$\begin{aligned}
||A_{11} - \widehat{A}_{11}||_2 &= ||(U_L^{(k)})^H A U_R^{(k)} - (\widehat{U}_L^{(k)})^H A \widehat{U}_R^{(k)}||_2 \\
&= ||(U_L^{(k)})^H A U_R^{(k)} - (\widehat{U}_L^{k})^H A U_R^{(k)} + (\widehat{U}_L^{(k)})^H A U_R^{(k)} - (\widehat{U}_L^{(k)})^H A \widehat{U}_R^{(k)}||_2 \\
&\le ||(U_L^{(k)} - \widehat{U}_L^{(k)})^H A U_R^{(k)}||_2 + ||(\widehat{U}_L^{(k)})^H A (U_R^{(k)} - \widehat{U}_R^{(k)}||_2 \\
&\le ||U_L^{(k)} - \widehat{U}_L^{(k)}||_2 ||A||_2 + ||A||_2 ||U_R^{(k)} - \widehat{U}_R^{(k)}||_2 \\
&\le 6n^\alpha \frac{\epsilon}{30n^\alpha} \\
&= \frac{\epsilon}{5}
\end{aligned} \tag{4.29}$$

and similarly $||B_{11} - \widehat{B}_{11}||_1 \le \frac{\epsilon}{5}$. Thus, by Lemma 2.3.3, $\Lambda_{4\epsilon/5}(A_{11}, B_{11})$ is shattered with respect to g. Repeating this argument for (A_{22}, B_{22}) and $(\widehat{A}_{22}, \widehat{B}_{22})$, we conclude $\Lambda_{4\epsilon/5}(A_{22}, B_{22})$ is also shattered with respect to g.

In the preceding analysis, we showed that for one step of recursion in **EIG**, a sufficient dividing line is found, k is computed correctly, and $\Lambda_{4\epsilon/5}(A_{11}, B_{11})$ and $\Lambda_{4\epsilon/5}(A_{22}, B_{22})$ are both shattered with respect to g, and therefore also with respect to the half grids g_R and g_L, with probability at least $1 - \frac{\theta}{2n^4}$. Since multiplying by matrices with orthonormal columns will preserve the norm requirements, we conclude that the subsequent calls to **EIG** in lines 21 and 22 are valid when these events occur. Hence, each recursive step succeeds with probability at least $1 - \frac{\theta}{2n^4}$. Since the recursive tree of **EIG** has depth at most $\log_{5/4}(n)$ and each step calls **EIG** twice, a union bound implies that the first guarantee of **EIG** fails with probability at most

$$2 \cdot 2^{\log_{5/4}(n)} \frac{\theta}{2n^4} \le 2n^4 \frac{\theta}{2n^4} = \theta. \tag{4.30}$$

We turn now to the second guarantee when $\sigma_n(B) \ge 1$. In this case, we will show that conditioning on the same events that ensure the first guarantee also imply the second. Since **EIG** builds the approximate eigenvectors recursively, we do this inductively.

The base case here corresponds to $m = 1$, in which case **EIG** gets the one right unit eigenvector ($v = 1$) exactly correct. Suppose now we are reconstructing the eigenvectors of (A', B') from the two sub-problems (A_{11}, B_{11}) and (A_{22}, B_{22}) it is split into. (A', B') here is any pencil obtained in the divide and conquer process.

Let $\widehat{T}$ and $\dot{T}$ be the invertible matrices obtained from applying **EIG** to (A_{11}, B_{11}) and (A_{22}, B_{22}) as in lines 21 and 22 of the algorithm. Since these calls to **EIG** pass the parameter $\frac{\beta}{3}$, we can assume each column of $\widehat{T}$ or $\dot{T}$ is at most $\frac{\beta}{3}$ away from a true unit right eigenvector of (A_{11}, B_{11}) or (A_{22}, B_{22}). In addition, let

$$T = \begin{pmatrix} U_R^{(k)} & U_R^{(m-k)} \end{pmatrix} \begin{pmatrix} \widehat{T} & 0 \\ 0 & \dot{T} \end{pmatrix} \tag{4.31}$$

be the matrix of approximate eigenvectors of (A', B') computed in line 23. Finally let $\widehat{U}_R^{(k)}$ and $\widehat{U}_R^{(m-k)}$ be the true matrices approximated by $U_R^{(k)}$ and $U_R^{(m-k)}$, as in step 5 above.

Consider now a column t_i of T. It suffices to handle the case where $t_i = U_R^{(k)}\widehat{t}_i$ for a column $\widehat{t}_i$ of $\widehat{T}$, as the same argument applies exactly if $t_i = U_R^{(m-k)}\dot{t}_i$ for $\dot{t}_i$ a column of $\dot{T}$. By our induction hypothesis, we know there exists a true right unit eigenvector $\widehat{v}_i$ of (A_{11}, B_{11}) such that

$$||\widehat{t}_i - \widehat{v}_i||_2 \leq \frac{\beta}{3}. \tag{4.32}$$

Now let $(\widehat{A}_{11}, \widehat{B}_{11})$ be the true problem approximated by (A_{11}, B_{11}). Conditioning on the same events used above, we know (following the same arguments as in steps 5 and 7)

$$||A_{11} - \widehat{A}_{11}||_2,\ ||B_{11} - \widehat{B}_{11}||_2 \leq 6n^\alpha ||U_R^{(k)} - \widehat{U}_R^{(k)}||_2 \leq 6n^\alpha \sqrt{\sqrt{\frac{10}{\theta}} 8n^3 \delta}, \tag{4.33}$$

which, applying the bound $\delta \leq \sqrt{\frac{\theta}{10}} \frac{\eta^2}{288n^{2\alpha+3}}$, becomes

$$||A_{11} - \widehat{A}_{11}||_2,\ ||B_{11} - \widehat{B}_{11}||_2 \leq 6n^\alpha \sqrt{\sqrt{\frac{10}{\theta}} 8n^3 \sqrt{\frac{\theta}{10}} \frac{\eta^2}{288n^{2\alpha+3}}} = \eta. \tag{4.34}$$

Thus, by Lemma 2.3.4, there exists a right unit eigenvector $\bar{v}_i$ of $(\widehat{A}_{11}, \widehat{B}_{11})$ such that

$$||\widehat{v}_i - \bar{v}_i||_2 \leq \frac{\sqrt{8}\omega}{\pi} \frac{\eta}{\epsilon(\epsilon - \eta)} (1 + ||\widehat{B}_{11}^{-1} \widehat{A}_{11}||_2) ||\widehat{B}_{11}||_2. \tag{4.35}$$

To simplify this bound, we first observe that (again by the argument made in step 7 above) we can assume $\epsilon - \eta \geq \frac{4}{5}\epsilon$. In addition, $||\widehat{A}_{11}||_2 \leq 3$ and $||\widehat{B}_{11}||_2 \leq 3n^\alpha$. Finally, since $\sigma_n(B) \geq 1$ and any split of divide and conquer can decrease the smallest singular value by at most η (by Lemma 1.7.2), and since the decision tree of **EIG** has depth at most $\log_{5/4}(n)$, the bound $\eta \leq \frac{1}{2\log_{5/4}(n)}$ ensures

$$\sigma_n(\widehat{B}_{11}) \geq 1 - \log_{5/4}(n)\eta \geq \frac{1}{2}. \tag{4.36}$$

Putting everything together, we have

$$||\widehat{v}_i - \bar{v}_i||_2 \leq \frac{\sqrt{8}\omega}{\pi}\frac{\eta}{\frac{4}{5}\epsilon^2}\left(1 + \frac{||\widehat{A}_{11}||_2}{\sigma_n(\widehat{B}_{11})}\right)||\widehat{B}_{11}||_2 \leq \frac{105\sqrt{8}}{4\pi}\frac{\omega n^\alpha}{\epsilon^2}\eta \leq \frac{\beta}{3}. \tag{4.37}$$

since $\eta \leq \frac{4\pi}{315\sqrt{8}}\frac{\beta\epsilon^2}{\omega n^\alpha}$. Now let $v_i = \widehat{U}_R^{(k)}\bar{v}_i$, which is a true right unit eigenvector of (A', B'). By construction, we have

$$\begin{aligned} ||t_i - v_i||_2 &= ||U_R^{(k)}\widehat{t}_i - \widehat{U}_R^{(k)}\bar{v}_i||_2 \\ &= ||U_R^{(k)}\widehat{t}_i - U_R^{(k)}\widehat{v}_i + U_R^{(k)}\widehat{v}_i - \widehat{U}_R^{(k)}\widehat{v}_i + \widehat{U}_R^{(k)}\widehat{v}_i - \widehat{U}_R^{(k)}\bar{v}_i||_2 \\ &\leq ||U_R^{(k)}(\widehat{t}_i - \widehat{v}_i)||_2 + ||(U_R^{(k)} - \widehat{U}_R^{(k)})\widehat{v}_i||_2 + ||\widehat{U}_R^{(k)}(\widehat{v}_i - \bar{v}_i)||_2. \\ &\leq ||\widehat{t}_i - \widehat{v}_i||_2 + ||U_R^{(k)} - \widehat{U}_R^{(k)}||_2 + ||\widehat{v}_i - \bar{v}_i||_2. \end{aligned} \tag{4.38}$$

Applying (4.32) and (4.37) to this and using the fact that $||U_R^{(k)} - \widehat{U}_R^{(k)}||_2 \leq \frac{\beta}{3}$, we conclude

$$||t_i - v_i||_2 \leq \frac{\beta}{3} + \frac{\beta}{3} + \frac{\beta}{3} = \beta. \tag{4.39}$$

By induction, we obtain the same bound for approximate/true eigenvectors of (A, B). □

The condition on the eigenvector guarantee in Theorem 4.1.4 (i.e., that $\sigma_n(B) \geq 1$) may seem restrictive, but it reflects the use-case in the following section. That is, while it is possible to adjust the parameters of **EIG** to allow for less strict lower bounds on $\sigma_n(B)$, we plan to apply **EIG** to the perturbed and scaled $(\widetilde{A}, n^\alpha\widetilde{B})$, where by construction $\sigma_n(n^\alpha\widetilde{B}) \geq 1$ with high probability.

4.2 Diagonalization in Nearly Matrix Multiplication Time

In this section, we state our diagonalization routine, which is built on **EIG** and the pseudospectral shattering results of Chapter 2. Here, we make use of the following important observation: if B is invertible and T contains right eigenvectors of (A, B), then $S = BT$ and T jointly diagonalize A and B. Since $\widetilde{B}$ is invertible almost surely, we can obtain a diagonalization of our perturbed pencil $(\widetilde{A}, n^{\alpha}\widetilde{B})$ by taking T – the output of **EIG** applied to $(\widetilde{A}, n^{\alpha}\widetilde{B})$ – and setting $S = \widetilde{B}T$. Assuming $||A - \widetilde{A}||_2$ and $||B - \widetilde{B}||_2$ are small, and undoing the n^{α} scaling, this produces an approximate diagonalization of (A, B). We state **RPD**, a routine that wraps around **EIG** to produce this diagonalization, as Algorithm 8 below.

Note that the assumption $||A||_2, ||B||_2 \le 1$ is essentially made for convenience;

Algorithm 8. Randomized Pencil Diagonalization (**RPD**)
Input: $A, B \in \mathbb{C}^{n\times n}$ and $\varepsilon < 1$ a desired accuracy.
Requires: $||A||_2, ||B||_2 \le 1$.
Output: Nonsingular S, T and diagonal D such that $||A - SDT^{-1}||_2, ||B - ST^{-1}||_2 \le \varepsilon$ with high probability.

```
$\gamma = \frac{\varepsilon}{16}$
$\alpha = \frac{\lceil 2\log_n(1/\gamma)+3\rceil}{2}$
$\epsilon = \gamma^5/(64n^{\frac{11\alpha+25}{3}} + \gamma^5)$
$\beta = \frac{\varepsilon\gamma^2}{24(1+4\gamma)}n^{-3\alpha-5}$
$\omega = \frac{\gamma^4}{4}n^{-\frac{8\alpha+13}{3}}$
Draw two independent Ginibre matrices $G_1, G_2 \in \mathbb{C}^{n\times n}$
$(\widetilde{A}, \widetilde{B}) = (A + \gamma G_1, B + \gamma G_2)$
Draw $z$ uniformly from the box of side length $\omega$ cornered at $-4 - 4i$
$g = \mathrm{grid}(z, \omega, \lceil 8/\omega\rceil, \lceil 8/\omega\rceil)$
$[T, D_1, D_2] = \mathbf{EIG}(n, \widetilde{A}, n^{\alpha}\widetilde{B}, \epsilon, \alpha, g, \beta, 1/n)$
for $i = 1 : n$ do
    $D(i,i) = n^{\alpha}\frac{D_1(i,i)}{D_2(i,i)}$
end for
$S = \widetilde{B}T$
return $S, T, D$
```

we can obtain a diagonalization of any pencil (A, B) via **RPD** by first normalizing the matrices accordingly. In Theorem 4.2.1, we show that **RPD** produces an approximate diagonalization with the given accuracy in exact arithmetic, thereby proving the bulk of our main result, Theorem 1.6.1. For reference, Figure 4.1 provides a high level overview of **RPD**, including the details of its call to **EIG**.

Theorem 4.2.1. *For any $A, B \in \mathbb{C}^{n\times n}$ with $||A||_2, ||B||_2 \leq 1$, the outputs of exact arithmetic* **RPD** *satisfy*

$$||A - SDT^{-1}||_2,\ ||B - ST^{-1}||_2 \leq \varepsilon$$

with probability at least

$$\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{1}{n} - 4e^{-n}\right]\left[1 - \frac{1}{n}\right].$$

Proof. Consider the perturbed and scaled pencil $(\widetilde{A}, n^\alpha \widetilde{B})$. By Theorem 2.3.1 and its proof, with probability at least $\left(1 - \frac{82}{n} - \frac{531441}{16n^2}\right)\left(1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right)$ we have the following:

1. $||G_1||_2, ||G_2||_2 \leq 4$,
2. $||\widetilde{A}||_2 \leq 3$,
3. $||n^\alpha \widetilde{B}||_2 \leq 3n^\alpha$,
4. $\sigma_n(n^\alpha \widetilde{B}) \geq 1$,
5. $\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \subseteq B_3(0)$,
6. $\kappa_V(n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}) \leq \frac{n^{\alpha+2}}{\gamma}$,
7. $\Lambda_\epsilon(\widetilde{A}, n^\alpha \widetilde{B})$ is shattered with respect to the grid g (for ϵ as in line 3).

Conditioning on these events, we observe that 2, 3, and 7 ensure that the call to **EIG** in line 10 is valid, meaning with probability at least $1 - \frac{1}{n}$ we can add the guarantees of Theorem 4.1.4 to our list:

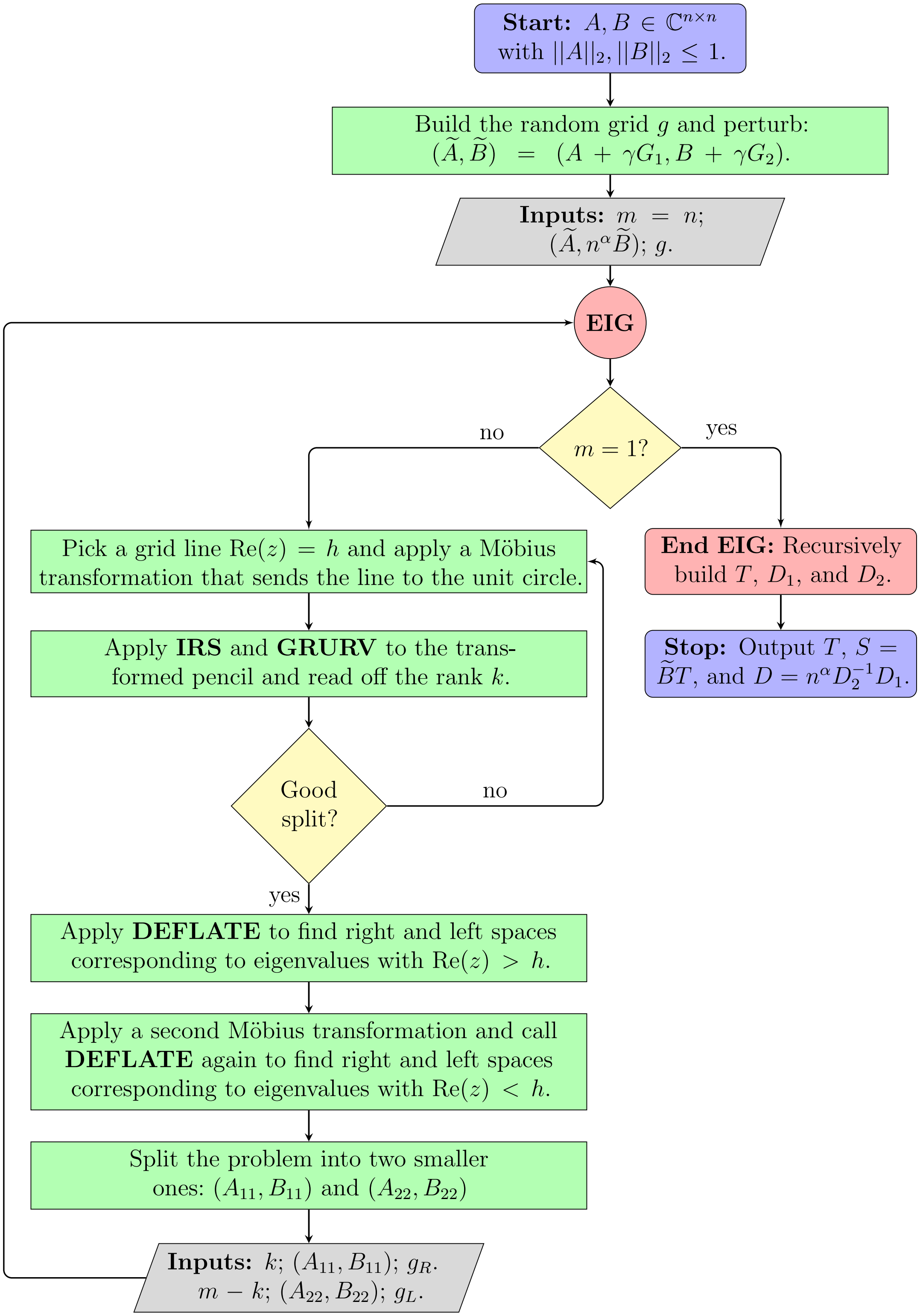

Start: $A, B \in \mathbb{C}^{n\times n}$ with $||A||_2, ||B||_2 \leq 1$.
Build the random grid g and perturb: $(\widetilde{A}, \widetilde{B}) = (A + \gamma G_1, B + \gamma G_2)$.
Inputs: $m = n$; $(\widetilde{A}, n^\alpha \widetilde{B})$; g.
EIG
$m = 1?$
no
yes
Pick a grid line $\mathrm{Re}(z) = h$ and apply a Möbius transformation that sends the line to the unit circle.
Apply **IRS** and **GRURV** to the transformed pencil and read off the rank k.
Good split?
no
yes
Apply **DEFLATE** to find right and left spaces corresponding to eigenvalues with $\mathrm{Re}(z) > h$.
Apply a second Möbius transformation and call **DEFLATE** again to find right and left spaces corresponding to eigenvalues with $\mathrm{Re}(z) < h$.
Split the problem into two smaller ones: (A_{11}, B_{11}) and (A_{22}, B_{22})
Inputs: k; (A_{11}, B_{11}); g_R. $m - k$; (A_{22}, B_{22}); g_L.
End EIG: Recursively build T, D_1, and D_2.
Stop: Output T, $S = \widetilde{B}T$, and $D = n^\alpha D_2^{-1} D_1$.

Figure 4.1. A diagram of **RPD** (Algorithm 8). We assume for simplicity that each split in **EIG** is made by a vertical grid line.

8. Each eigenvalue of (D_1, D_2) shares a grid box of g with a true eigenvalue of $(\widetilde{A}, n^\alpha \widetilde{B})$.

9. Since $\sigma_n(n^\alpha \widetilde{B}) \geq 1$ (item 4) each column t_i of T satisfies $||t_i - v_i||_2 \leq \beta$ for v_i a true unit right eigenvector of $(\widetilde{A}, n^\alpha \widetilde{B})$.

By the definition of conditional probability, all nine of these events occur simultaneously with probability at least

$$\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right]\left[1 - \frac{1}{n}\right]. \tag{4.40}$$

Since the choice of α in line 2 guarantees

$$\frac{n^{2-2\alpha}}{\gamma^2} \leq \frac{1}{n} \tag{4.41}$$

(4.40) can be bounded from below by

$$\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{1}{n} - 4e^{-n}\right]\left[1 - \frac{1}{n}\right]. \tag{4.42}$$

To complete the proof we show that the nine items listed above guarantee both $||A - SDT^{-1}||_2 \leq \varepsilon$ and $||B - ST^{-1}||_2 \leq \varepsilon$. The second of these is trivial: since $\gamma = \frac{\varepsilon}{16}$ and $ST^{-1} = \widetilde{B}$ we have

$$||B - ST^{-1}||_2 = ||B - \widetilde{B}||_2 = ||\gamma G_2||_2 \leq 4\gamma = \frac{\varepsilon}{4} < \varepsilon. \tag{4.43}$$

To show the same for A and SDT^{-1}, we use the following key fact. Let V be the matrix whose columns contain, in order, the true right unit eigenvectors of $(\widetilde{A}, n^\alpha \widetilde{B})$ guaranteed by item 9. Since $\widetilde{B}$ is invertible with probability one and $(\widetilde{A}, \widetilde{B})$ and $(\widetilde{A}, n^\alpha \widetilde{B})$ have the same set of eigenvectors, V diagonalizes $\widetilde{B}^{-1}\widetilde{A}$. Thus, there exists a diagonal matrix Λ such that $\widetilde{B}^{-1}\widetilde{A} = V\Lambda V^{-1}$. Moreover, since the eigenvalue of $(\widetilde{A}, n^\alpha \widetilde{B})$ corresponding to v_i shares a grid box of g with the eigenvalue of (D_1, D_2) corresponding to t_i (this was how v_i was found in the proof of Theorem 4.1.4), each diagonal entry of Λ is at most $\sqrt{2}n^\alpha\omega$ away from the corresponding diagonal entry of $D = n^\alpha D_2^{-1} D_1$. Note that the eigenvalues of (D_1, D_2) are all contained in g, which guarantees that D_2 is invertible. With all of this in mind, expand $||A - SDT^{-1}||_2$ as follows:

$$\begin{aligned} &||A - SDT^{-1}||_2 = ||A - \widetilde{A} + \widetilde{A} - SDT^{-1}||_2 \\ &\leq ||A - \widetilde{A}||_2 + ||\widetilde{A} - SDT^{-1}||_2 \\ &\leq ||A - \widetilde{A}||_2 + ||\widetilde{B}||_2 ||\widetilde{B}^{-1}\widetilde{A} - TDT^{-1}||_2 \\ &\leq 4\gamma + (1+4\gamma)||V\Lambda V^{-1} - TDT^{-1}||_2 \\ &\leq 4\gamma + (1+4\gamma)\left[||(V-T)\Lambda V^{-1}||_2 + ||T(\Lambda - D)V^{-1}||_2 + ||TD(V^{-1} - T^{-1})||_2\right]. \end{aligned} \tag{4.44}$$

To simplify this bound in terms of our parameters, we observe the following.

- Since the columns of V and T are unit vectors, $1 \leq ||V||_2, ||T||_2 \leq \sqrt{n}$.
- Since V diagonalizes $\widetilde{B}^{-1}\widetilde{A}$ – and therefore also $n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$ – and $\kappa_V(n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}) \leq \frac{n^{\alpha+2}}{\gamma}$, Remark 2.2.4 allows us to assume

$$||V||_2||V^{-1}||_2 \leq \frac{n^{\alpha+2}}{\gamma} \tag{4.45}$$

  without changing probabilities. Combined with the previous point, this implies $||V^{-1}||_2 \leq \frac{n^{\alpha+2}}{\gamma}$.
- The columns of T and V satisfy $||t_i - v_i||_2 \leq \beta$ so $||T - V||_2 \leq \sqrt{n}\beta$
- By the Lemma 1.7.2, $||T - V||_2 \leq \sqrt{n}\beta$ implies $\sigma_n(T) \geq \sigma_n(V) - \sqrt{n}\beta$. Combining this with our upper bound on $||V^{-1}||_2$ yields

$$||T^{-1}||_2 \leq \frac{n^{\alpha+2}}{\gamma - n^{\frac{2\alpha+5}{2}}\beta}. \tag{4.46}$$

- Because each diagonal entry of Λ is at most $\sqrt{2}n^\alpha\omega$ from the corresponding diagonal entry of D, $||\Lambda - D||_2 \leq \sqrt{2}n^\alpha\omega$.
- Since both $(\widetilde{A}, n^\alpha\widetilde{B})$ and (D_1, D_2) have eigenvalues in $B_3(0)$, $||\Lambda||_2, ||D||_2 \leq 3n^\alpha$.
- Finally, $||V^{-1} - T^{-1}||_2 = ||T^{-1}(T-V)V^{-1}||_2 \leq ||T^{-1}||_2||T - V||_2||V^{-1}||_2$.

Together, these imply the following bounds:

$$||(V-T)\Lambda V^{-1}||_2 \leq ||V-T||_2||\Lambda||_2||V^{-1}||_2 \leq \sqrt{n}\beta \cdot 3n^{\alpha} \cdot \frac{n^{\alpha+2}}{\gamma} = \frac{3\beta}{\gamma} n^{\frac{4\alpha+5}{2}} \tag{4.47a}$$

$$||T(\Lambda-D)V^{-1}||_2 \leq ||T||_2||\Lambda-D||_2||V^{-1}||_2 \leq \sqrt{n} \cdot \sqrt{2}n^{\alpha}\omega \cdot \frac{n^{\alpha+2}}{\gamma} = \frac{\sqrt{2}\omega}{\gamma} n^{\frac{4\alpha+5}{2}} \tag{4.47b}$$

$$||TD(V^{-1}-T^{-1})||_2 \leq \sqrt{n} \cdot 3n^{\alpha} \cdot \frac{n^{\alpha+2}}{\gamma - n^{\frac{2\alpha+5}{2}}\beta} \cdot \sqrt{n}\beta \cdot \frac{n^{\alpha+2}}{\gamma} = \frac{3\beta n^{3\alpha+5}}{\gamma(\gamma - n^{\frac{2\alpha+5}{2}}\beta)}. \tag{4.47c}$$

Now the choice of β in line 4 ensures $\beta \leq \frac{\varepsilon\gamma}{12(1+4\gamma)} n^{-\frac{4\alpha+5}{2}}$, so (4.47a) simplifies to

$$||(V-T)\Lambda V^{-1}||_2 \leq 3 \cdot \frac{\varepsilon\gamma}{12(1+4\gamma)n^{\frac{4\alpha+5}{2}}} \cdot \frac{n^{\frac{4\alpha+5}{2}}}{\gamma} = \frac{\varepsilon}{4(1+4\gamma)}. \tag{4.48}$$

Similarly, using this time the fact that $\beta < \frac{\gamma}{2} n^{-\frac{2\alpha+5}{2}}$ and therefore $\gamma - n^{\frac{2\alpha+5}{2}}\beta > \frac{\gamma}{2}$, (4.47c) becomes

$$||TD(V^{-1}-T^{-1})||_2 \leq \frac{6}{\gamma^2} n^{3\alpha+5} \cdot \frac{\varepsilon\gamma^2}{24(1+4\gamma)} n^{-3\alpha-5} = \frac{\varepsilon}{4(1+4\gamma)}. \tag{4.49}$$

Finally, $\omega = \frac{\gamma^4}{4} n^{-\frac{8\alpha+13}{3}}$ implies $\omega \leq \frac{\varepsilon\gamma}{4\sqrt{2}(1+4\gamma)} n^{-\frac{4\alpha+5}{2}}$, which allows us to upper bound (4.47b) as

$$||T(\Lambda-D)V^{-1}||_2 \leq \sqrt{2} \cdot \frac{\varepsilon\gamma}{4\sqrt{2}(1+4\gamma)n^{\frac{4\alpha+5}{2}}} \cdot \frac{n^{\frac{4\alpha+5}{2}}}{\gamma} = \frac{\varepsilon}{4(1+4\gamma)}. \tag{4.50}$$

Applying these to (4.44), we obtain

$$||A - SDT^{-1}||_2 \leq 4\gamma + (1+4\gamma)\left[\frac{\varepsilon}{4(1+4\gamma)} + \frac{\varepsilon}{4(1+4\gamma)} + \frac{\varepsilon}{4(1+4\gamma)}\right] = 4\gamma + \frac{3}{4}\varepsilon. \tag{4.51}$$

Since $\gamma = \frac{\varepsilon}{16}$ we conclude $||A - SDT^{-1}||_2 \leq \varepsilon$. □

4.2.1 Asymptotic Complexity

It remains to show that **RPD** runs in nearly matrix multiplication time. We therefore wrap up this section by computing its asymptotic complexity in terms of n, the size of the pencil (A, B), and ε, the accuracy of the approximate diagonalization. Throughout, we assume that we have access to black-box algorithms for multiplying two

$n \times n$ matrices and computing the QR factorization of an $n \times n$ matrix, which require $T_{\mathrm{MM}}(n)$ and $T_{\mathrm{QR}}(n)$ arithmetic operations, respectively. For simplicity, we also assume access to QL/RQ algorithms (used in **GRURV**) that require $T_{\mathrm{QR}}(n)$ operations.

Proposition 4.2.2 shows that **RPD** runs in nearly matrix multiplication time, as desired. Its proof is somewhat more subtle than implied in Chapter 1; rather than simply arguing that each step of divide-and-conquer runs in nearly matrix multiplication time and that only logarithmically many steps are needed, we apply a sharper geometric sum that leverages the shrinking problem size guaranteed by significant eigenvalue splits.

Proposition 4.2.2. *Exact arithmetic* **RPD** *requires at most* $O\left(\log^2\left(\frac{n}{\varepsilon}\right) T_{\mathrm{MM}}(n)\right)$ *arithmetic operations.*

Proof. We track only matrix multiplication and QR, as all other building blocks of **RPD** have smaller complexity. We begin by noting that line 2 of Algorithm 8 implies $\alpha = \mathcal{O}\left(\log_n\left(\frac{1}{\gamma}\right)\right)$ and $\gamma = \Theta(\varepsilon)$, so

$$n^{\alpha} = O\left(n^{\log_n(1/\gamma)}\right) = O\left(\frac{1}{\gamma}\right) = O\left(\frac{1}{\varepsilon}\right). \tag{4.52}$$

Meanwhile in line 3 we set

$$\epsilon > \frac{\gamma^5}{65 n^{\frac{11\alpha+25}{3}}} = \Omega\left(\frac{\varepsilon^{26/3}}{n^{25/3}}\right). \tag{4.53}$$

Together, these imply that the number of steps of repeated squaring required at any point in the recursion can be bounded asymptotically as

$$p = O\left(\log\left(\frac{n^{\alpha}}{\epsilon}\right)\right) = O\left(\log\left(\frac{n^{25/3}}{\varepsilon^{29/3}}\right)\right) = O\left(\log\left(\frac{n}{\varepsilon}\right)\right). \tag{4.54}$$

Consider now working through one step of divide-and-conquer. Lines 8-15 of **EIG** make up the bulk of the work, executing a search over the grid lines for one that sufficiently splits the spectrum. For each line that is checked, we make one call to **IRS** and one call to **GRURV**; each step of repeated squaring consists of one $2m \times m$ QR factorization and two $m \times m$ matrix multiplications, while applying **GRURV** to a product of two $m \times m$ matrices requires $3T_{\mathrm{QR}}(m) + 2T_{\mathrm{MM}}(m)$ operations. Combining these with (4.54), we conclude that each grid line checked results in

$$\begin{aligned} &O\left(\log\left(\frac{n}{\varepsilon}\right)[T_{\mathrm{QR}}(2m) + 2T_{\mathrm{MM}}(m)] + 2T_{\mathrm{QR}}(m) + 2T_{\mathrm{MM}}(m)\right) \\ &= O\left(\log\left(\frac{n}{\varepsilon}\right)[T_{\mathrm{QR}}(m) + T_{\mathrm{MM}}(m)]\right) \end{aligned} \tag{4.55}$$

operations. Since we check at most $O\left(\log\left(\frac{1}{\omega}\right)\right)$ grid lines each time and

$$\omega = \frac{\gamma^4}{4n^{\frac{8\alpha+13}{3}}} = \Omega\left(\frac{\varepsilon^{20/3}}{n^{13/3}}\right), \tag{4.56}$$

lines 8-15 of **EIG** take at most

$$O\left(\log^2\left(\frac{n}{\varepsilon}\right)[T_{\mathrm{QR}}(m) + T_{\mathrm{MM}}(m)]\right) = O\left(\log^2\left(\frac{n}{\varepsilon}\right)T_{\mathrm{MM}}(m)\right) \tag{4.57}$$

operations, where we simplify by noting $T_{\mathrm{QR}}(m) = O(T_{\mathrm{MM}}(n))$ [38, §4.1]. Since the remainder of one step of **EIG** – i.e., the subsequent calls to **DEFLATE** – has complexity equal to that of checking one grid line, we conclude that (4.57) is the asymptotic complexity of one step of divide-and-conquer.

To complete the proof, we sum this expression recursively. Since the $\log^2\left(\frac{n}{\epsilon}\right)$ term of (4.57) is independent of m, this reduces to summing $T_{\mathrm{MM}}(m)$ over all subproblems produced by **EIG**. With this in mind, set $T_{\mathrm{MM}}(n) = O(n^\xi)$ for $\xi \in [2,3]$ and suppose we divide an $m \times m$ pencil into subproblems of size m_1 and m_2. Since we enforce a significant split, we are guaranteed $\frac{1}{5}m \leq m_1, m_2 \leq \frac{4}{5}m$ and therefore

$$m_1^\xi + m_2^\xi = m_1^\xi + (m - m_1)^\xi \leq \left(\frac{4}{5}m\right)^\xi + \left(\frac{1}{5}m\right)^\xi \leq \frac{17}{25}m^\xi, \tag{4.58}$$

where the last inequality is obtained by applying $\xi \geq 2$. Consequently, a sum of m^ξ over all subproblems can be bounded by

$$\sum_{k=0}^{\infty} n^\xi \left(\frac{17}{25}\right)^k = n^\xi \sum_{k=0}^{\infty}\left(\frac{17}{25}\right)^k = \frac{25}{8}n^\xi \tag{4.59}$$

and therefore $\sum_m T_{\mathrm{MM}}(m) = O(T_{\mathrm{MM}}(n))$. Applying this to (4.57) yields the final complexity. □

4.3 Numerical Examples

In this section, we consider several examples to investigate how pseudospectral divide-and-conquer performs in practice. Our first task is to adjust the parameters of **RPD** and **EIG**, as the values listed in the pseudocode of Algorithms 7 and 8 – though necessary in the proof of Theorem 4.2.1 – are prohibitively restrictive for implementation. Here, we make the following relaxations.

- First, we eliminate the n^α scaling, testing examples with eigenvalues exclusively (or predominantly) in $B_3(0)$.
- Extracting the main dependence on γ and n, we set $\epsilon = \beta = \omega = \gamma/n$ and we limit the number of steps of repeated squaring to $p = \lceil \log_2(n/\epsilon) \rceil$.
- Finally, we drop the factor of $1/n^3$ from the criteria used to compute k in **EIG**.

In light of these simplifications, we present the following experiments as simply a proof of concept. Accordingly, we do not consider run times nor do we use an explicitly parallel implementation of the algorithm. Throughout, all results were obtained in Matlab version R2023a.[3]

4.3.1 Model Problems

We start by using **RPD** as stated (i.e., running to subproblems of size 1 × 1) on the following 50 × 50 model problems.

1. **Planted Spectrum:** First, we consider a pencil with equally spaced, real eigenvalues in the interval $[-2, 2]$. To obtain (A, B), we fill a diagonal matrix Λ as

$$\Lambda(j, j) = -2 + \frac{4}{49}(j - 1) \tag{4.60}$$

[3] Check out our implementation: https://github.com/ry-schneider/Randomized_Pencil_Diagonalization.

and set $A = X\Lambda Y^{-1}$ and $B = XY^{-1}$ for X and Y two independent, complex Gaussian matrices. In accordance with **RPD**, we then normalize the pencil so that $||A||_2, ||B||_2 \leq 1$. We can think of this example as the best-case scenario, where $\text{gap}(A, B)$ is large and B is far from singular.

2. **Jordan Block:** Next, we consider a pencil (A, B) with $B = I$ and $A = J_{50}(0)$ for $J_{50}(0)$ a Jordan block with eigenvalue zero. In contrast to the planted spectrum example, this tests a generalized eigenvalue problem with $\text{gap}(A, B) = 0$.

We track the performance of divide-and-conquer on these examples in several ways. First and foremost, we want to verify a finite-preicision counterpart to Theorem 4.2.1: does **RPD** reliably produce accurate diagonalizations of each pencil? With this in mind, we compute diagonalization error as

$$\log_{10}\left(\max\left\{||A - SDT^{-1}||_2, ||B - ST^{-1}||_2\right\}\right), \tag{4.61}$$

for S, D, and T the outputs of **RPD**, and we consider a run to be successful if this error is at most $\log_{10}(\varepsilon)$. Tracking the number of failed runs yields an empirical failure probability for **RPD** (with the simplifications made above). Note that (4.61) is only meaningful if $||A||_2$ and $||B||_2$ are roughly equal and close to one, as is the case in our examples.

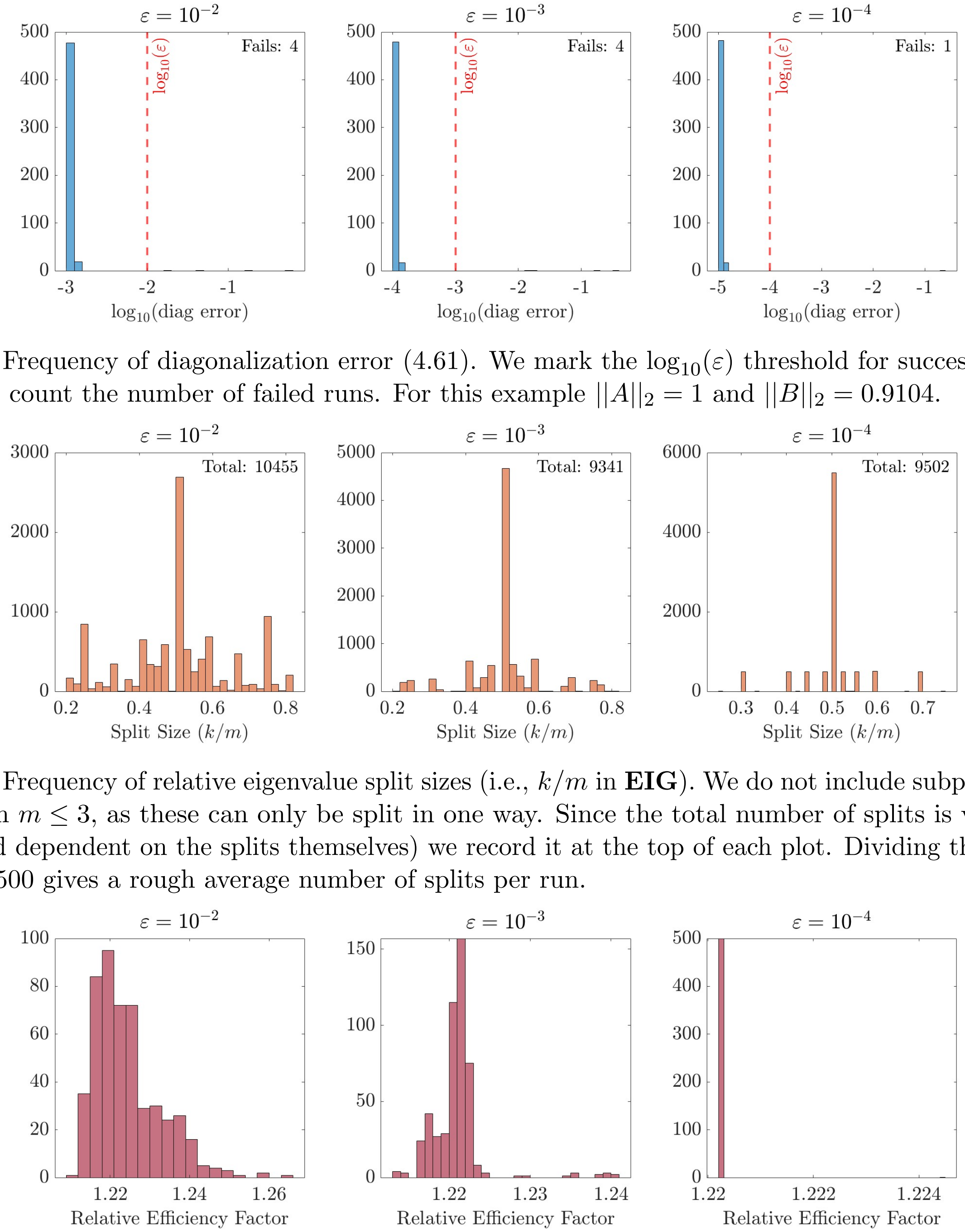
Next, we want to measure the efficiency of the divide-and-conquer process. One way to do this is to catalog the relative split size at each step (i.e., k/m in **EIG**). While we know that **EIG** guarantees that the relative split is at least 0.2 and at most 0.8, divide-and-conquer is most efficient if relative splits are close to 0.5 at each step.

Of course, the split size tells only part of the story. Recalling the proof of Proposition 4.2.2, **EIG** spends most of its time finding a dividing line; thus, even if splits are reliably near 50/50, the algorithm may be slow if too many lines are checked. Assuming access to $O(n^3)$ algorithms for matrix multiplication and QR, one step of our implementation requires $O(\log(\frac{n}{\varepsilon})m^3 l)$ operations, where l is the number of grid lines checked and m

is the size of the current subproblem. Ignoring the $\log(\frac{n}{\varepsilon})$ factor (as it will cancel in our eventual measure of efficiency) we can do a pseudo-flop count by summing $m^3 l$ over all steps with $m > 1$, and we can easily compute an optimal value for this count by requiring at each step that $l = 1$ and that the split is as close to 50/50 as possible. Dividing the actual count by the optimal one produces what we call a *relative efficiency factor* for each run, which tells us roughly how many times more work **RPD** is doing than the best-case scenario.

Histograms of each of these measures of performance are presented for both examples in Figures 4.2 and 4.3. In each test, we run **RPD** 500 times and present results for decreasing values of ε. With only a handful of failed runs on each problem, the results are compelling: **RPD** reliably diagonalizes both pencils, and divide-and-conquer appears to favor near-optimal eigenvalue splits. This carries through to the relative efficiency. Our rough flop count shows that **RPD** executes only slightly more than the optimal amount of work to produce these diagonalizations. Note that the number of failed runs appears to decrease with ε. While this may seem counterintuitive, it is a byproduct of our relaxed parameters, which become more restrictive (or equivalently more sensitive) as ε shrinks.

While these results are promising, we might be more interested in probing the boundaries of Theorem 1.2.14. That is, when **RPD** succeeds, how accurate are the corresponding sets of approximate eigenvalues? With this in mind Figure 4.4 provides eigenvalue approximation data for both model problems, where in each case we consider only approximations produced by successful runs. The results in these plots trace out nicely the challenge of extracting accurate eigenvalues from an accurate diagonalization. In the best case – the planted spectrum example – increasingly accurate diagonalizations provide correspondingly better eigenvalue approximations, as promised by Theorem 1.2.14. The Jordan block example, on the other hand, demonstrates that when $\text{gap}(A, B) = 0$ we cannot hope to recover repeated eigenvalues with any confidence, though this is also the case for classical backwards-stable algorithms like QZ.

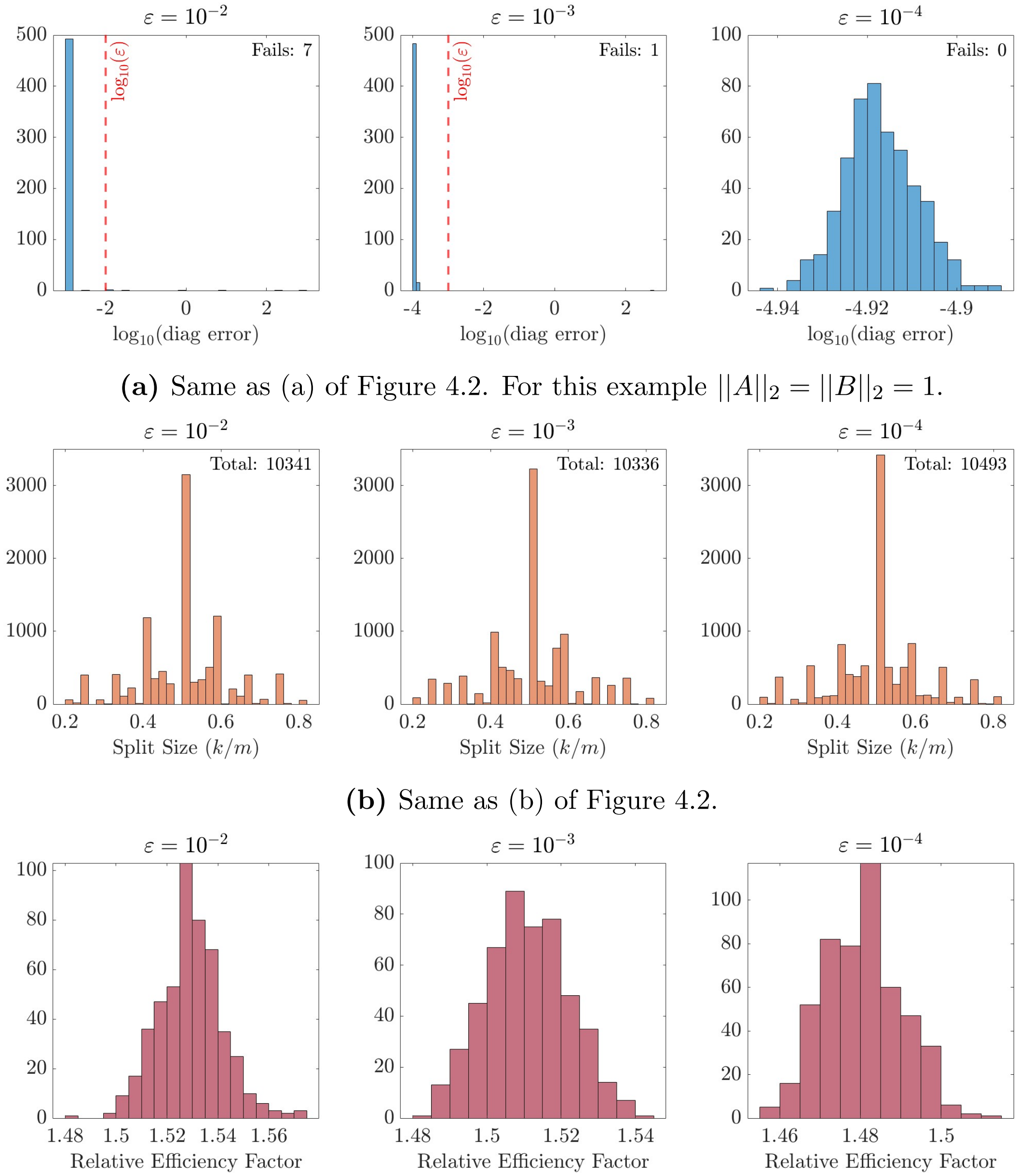


(a) Frequency of diagonalization error (4.61). We mark the $\log_{10}(\varepsilon)$ threshold for success in red and count the number of failed runs. For this example $||A||_2 = 1$ and $||B||_2 = 0.9104$.

(b) Frequency of relative eigenvalue split sizes (i.e., k/m in **EIG**). We do not include subproblems with $m \leq 3$, as these can only be split in one way. Since the total number of splits is variable (and dependent on the splits themselves) we record it at the top of each plot. Dividing this total by 500 gives a rough average number of splits per run.

(c) Frequency of relative efficiency factor (the pseudo-flop count divided by its optimal value).

Figure 4.2. Performance data for **RPD** on the 50×50 planted-spectrum example with decreasing values of ε. Each plot corresponds to 500 runs of **RPD**.

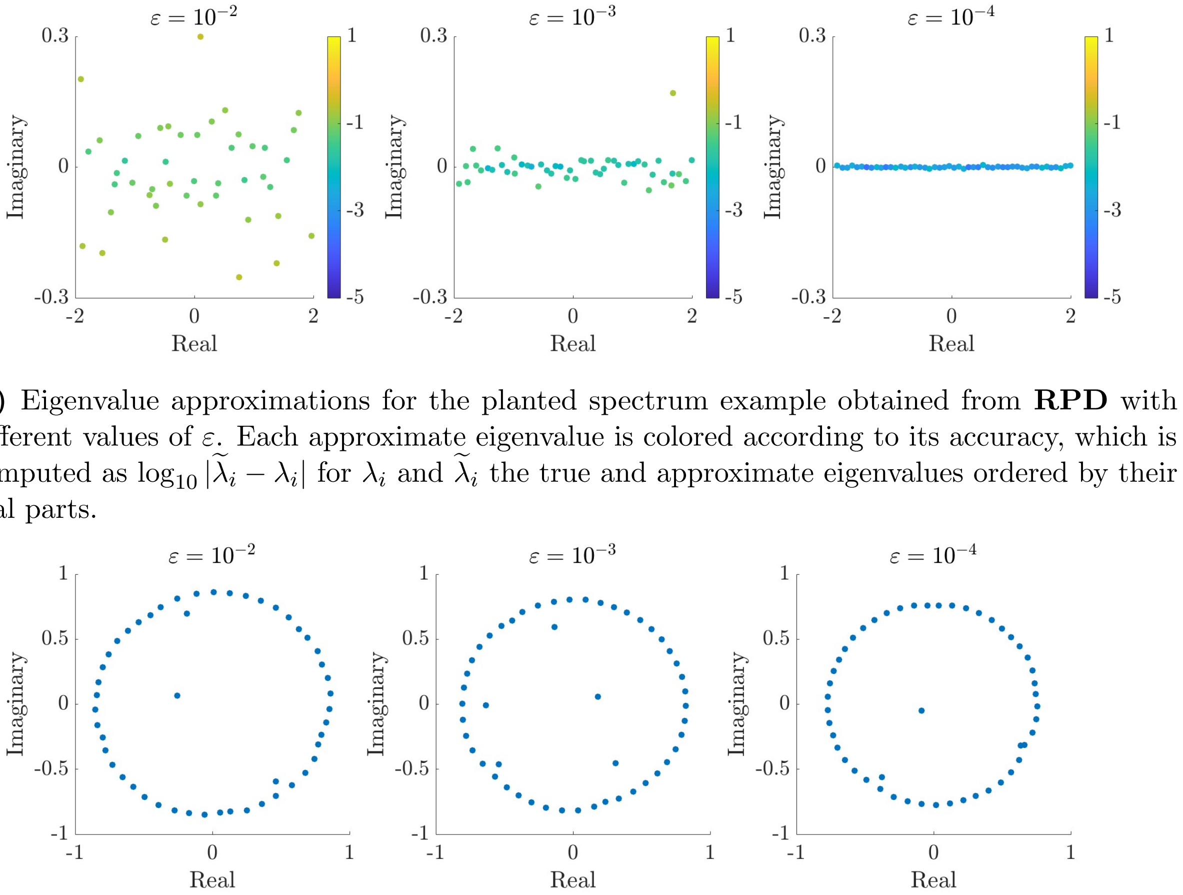


(a) Same as (a) of Figure 4.2. For this example $||A||_2 = ||B||_2 = 1$.

(b) Same as (b) of Figure 4.2.

(c) Same as (c) of Figure 4.2

Figure 4.3. A repeat of Figure 4.2 for the 50 × 50 Jordan block example.



(a) Eigenvalue approximations for the planted spectrum example obtained from **RPD** with different values of ε. Each approximate eigenvalue is colored according to its accuracy, which is computed as $\log_{10}|\widetilde{\lambda}_i - \lambda_i|$ for λ_i and $\widetilde{\lambda}_i$ the true and approximate eigenvalues ordered by their real parts.

(b) Eigenvalue approximations for the Jordan block example obtained from **RPD** with different values of ε. The only true eigenvalue for this problem is zero.

Figure 4.4. Eigenvalue approximation data for **RPD** applied to the planted spectrum and Jordan block examples. We present here only results for successful runs – i.e., each set of approximate eigenvalues corresponds to a diagonalization with $||A - SDT^{-1}||_2 \leq \varepsilon$ and $||B - ST^{-1}||_2 \leq \varepsilon$.

4.3.2 Large n and Infinite Eigenvalues

Running **RPD** down to subproblems of size 1×1 is useful from a theoretical perspective but unlikely to be done in practice. We turn next to a more realistic use case, where n is large and only a few splits are made before passing off to QZ. In this setting, we test the algorithm on pencils with an eigenvalue at infinity (i.e., where B is singular). To do this, we construct a 1000×1000 pencil by drawing A and B randomly, computing a singular value decomposition $B = U\Sigma V^H$, and setting

$$B = B - \sigma_{\min}(B) uv^H \tag{4.62}$$

for u and v the last columns of U and V, respectively. By construction, this forces B to be singular without changing its remaining singular values (and critically ensuring its norm remains comparable to A). As in the previous examples, we normalize A and B before calling **RPD**. Note that (A, B) here is the pencil considered in Example 1.5.4.

Remark 4.3.1. Because we omit the n^α scaling on this example, the perturbed pencil $(\widetilde{A}, \widetilde{B})$ is very likely to have an eigenvalue outside of the shattering grid. While this appears problematic it is ultimately harmless. When only a handful of splits are needed, $(\widetilde{A}, \widetilde{B})$ can have many – even a large fraction – of eigenvalues outside of the grid, and divide-and-conquer will only falter if these fall predominantly in a specific region (for example between two vertical grid lines but above all of the horizontal ones). In part, this justifies our choice to omit the n^α scaling; while it is necessary to state an algorithm that provably runs to scalar subproblems, it is overly restrictive in practice – driving eigenvalues that are initially small closer together and necessitating a finer grid.

In addition to testing larger values of n, we also use this example to further justify our decision to avoid matrix inversion. To that end – and in light of Proposition 2.3.2 – we compare **RPD** to an alternative algorithm that proceeds as follows:

1. Perturb to obtain $(\widetilde{A}, \widetilde{B})$.
2. Form the product $X = \widetilde{B}^{-1}\widetilde{A}$.
3. Apply the divide-and-conquer routine of Banks et al. [16, Algorithm EIG].

To ensure a fair comparison, we run both algorithms with the same perturbations (meaning the same $\widetilde{A}, \widetilde{B}$) and the same grid. We also restrict the number of steps of the Newton iteration for the sign function – the counterpart to **IRS** used by Banks et al. – to $\lceil \log_2(n/\epsilon) \rceil$.

In both cases, we run divide-and-conquer to subproblems with $m \leq 250$, at which point we default to QZ and QR, respectively, by calling Matlab's `eig`. In the tests presented below, both algorithms averaged five splits before defaulting to `eig`.

Once again, we track diagonalization error via (4.61). Importantly, we are interested in producing a diagonalization of (A, B) even if the single-matrix algorithm of Banks et al. is used. In that case, a matrix T containing right eigenvectors (the output of single-matrix divide-and-conquer) gives rise to a corresponding matrix of left eigenvectors $S = \widetilde{B}T$, which together approximately diagonalize (A, B). This mirrors exactly how the diagonalization is obtained in **RPD**.

In addition, we record the eigenvalue error associated with each diagonalization. This is done by ordering the approximate and true eigenvalues by magnitude and computing the average, absolute error over the spectrum, excluding the eigenvalue at infinity. Figure 4.5 records eigenvalue and diagonalization errors for both algorithms and for two choices of ε. Each plot corresponds to 200 trials, derived from twenty random draws of A and B run through each algorithm ten times.

When $\varepsilon = 10^{-5}$ we see little difference between the algorithms; though **RPD** is slightly more accurate, both approaches produce successful diagonalizations and comparably good eigenvalue approximations. Recalling that the size of the perturbation is determined by ε, this appears to be a consequence of the relatively large perturbation applied to B, which ensures that $\widetilde{B}$ is well-conditioned and that the error incurred by forming $\widetilde{B}^{-1}\widetilde{A}$ is small. In this case – or more generally in situations where B is known to be well-conditioned – both algorithms are viable and essentially equivalent.

In contrast, when ε is much smaller **RPD** shows clear advantages. While neither can produce an accurate enough diagonalization – an indication that our relaxed parameters are too loose for this regime – **RPD** is consistently an order of magnitude better than its single-matrix alternative, and its corresponding eigenvalue approximations are remarkably accurate. Again this seems attributable to the conditioning of $\widetilde{B}$; here, forming $\widetilde{B}^{-1}\widetilde{A}$

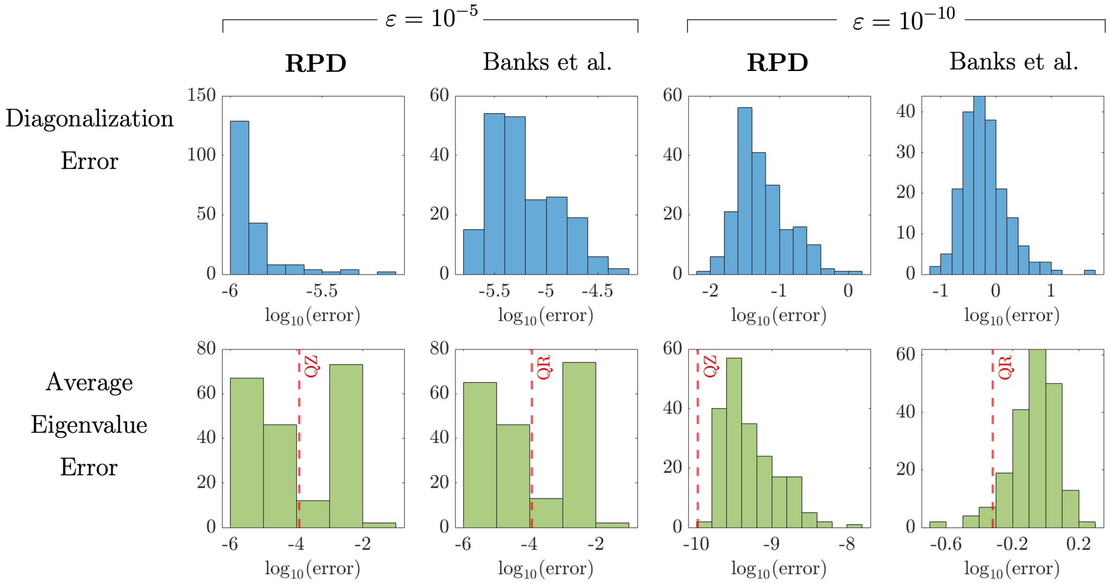


Figure 4.5. Performance data for **RPD** versus the single matrix, divide-and-conquer algorithm of Banks et al. [16]. Here, 1000×1000 pencils are constructed by drawing A and B randomly and subtracting a rank one matrix from B to force it to be singular (without changing its remaining singular values). For twenty random draws of A and B, we present ten runs through both algorithms, with first $\varepsilon = 10^{-5}$ and subsequently $\varepsilon = 10^{-10}$. Diagonalization error is computed according to (4.61) while eigenvalue accuracy is measured by ordering the true and approximate eigenvalues by magnitude and computing the average absolute error, excluding the eigenvalue at infinity. For the latter, we mark the error achieved by QZ/QR (computed in the same way) when applied to $(\widetilde{A}, \widetilde{B})$ and $\widetilde{B}^{-1}\widetilde{A}$, respectively.

not only forces divide-and-conquer to work with a poorly-conditioned matrix but also introduces error that meaningfully shifts the eigenvalues away from those of (A, B).

Indeed, the difference in eigenvalue error present when $\varepsilon = 10^{-10}$ traces a similar gap between QZ and QR, as marked on the histograms. This indicates that the poor eigenvalue recovery of Banks et al. is due primarily to the gap between the eigenvalues of $\widetilde{B}^{-1}\widetilde{A}$ and (A, B), which is essentially what the error in QR represents. Consequently, we cannot hope that by improving the diagonalization produced by Banks et al. – which should be possible by adjusting the parameters – we will see a similar improvement in the eigenvalues. On this example, then, we expect that this approach will break down as ε becomes small; while initially decreasing ε may improve diagonalization and eigenvalue accuracy, $\widetilde{B}$ will eventually become poorly-conditioned enough to guarantee that the eigenvalues of $\widetilde{B}^{-1}\widetilde{A}$ are far from those of (A, B). Contrast this with **RPD**, which does not suffer from the same drawback and can reproduce the finite eigenvalues of (A, B) to high accuracy.

Note that the example considered here is a larger version of the one covered by Figure 2.1. Together, they capture the danger of operating with inversion: when $\widetilde{B}$ is poorly conditioned, not only are the pseudospectra of $\widetilde{B}^{-1}\widetilde{A}$ unwieldy, but the eigenvalues they collapse to may significantly stray from those of the input pencil.

4.3.3 Singular Pencils

To this point we have exclusively tested divide-and-conquer on regular pencils. Since singular pencils are a reality in many applications, we consider as a final test the singular pencil of Lotz and Noferini [90] presented in Example 1.1.9, which recall corresponds to the matrices

$$A = \begin{pmatrix} 2 & -1 & -5 & -1 \\ 6 & -2 & -11 & -2 \\ 5 & 0 & -2 & 0 \\ 3 & 1 & 3 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -1 & -4 & -2 \\ 2 & -3 & -12 & -6 \\ -1 & -3 & -11 & -6 \\ -2 & -2 & -7 & -4 \end{pmatrix}, \tag{4.63}$$

and has only one simple eigenvalue $\lambda = 1$. Practically speaking, recovering this eigenvalue via divide-and-conquer should be difficult due to the initial perturbation made by **RPD**. In fact, Lotz and Noferini show that an arbitrarily small (nonrandom) perturbation to (A, B) can send its eigenvalues to any four points in the complex plane. In spite of this, divide-and-conquer finds the true eigenvalue to roughly five digits of precision when run with $\varepsilon = 10^{-6}$, as shown in Table 4.1.

Of course, QZ also finds $\lambda = 1$ to a remarkable 14 digits of precision. Motivated by this observation, Lotz and Noferini develop a theory of weak condition numbers to help explain the apparent stability of this eigenvalue. The results in Table 4.1 reflect the spirit of these condition numbers; though perturbations exist that produce arbitrary eigenvalues,

Table 4.1. Eigenvalues of the singular pencil (4.63) as computed by QZ and pseudospectral divide-and-conquer. We present three successful runs of **RPD**, each with $\varepsilon = 10^{-6}$.

	Divide-and-Conquer (Algorithm 8)		
QZ [102]	Run 1	Run 2	Run 3
-2.089013	$0.999997 - 0.000007i$	$1.000001 - 0.000008i$	$0.999994 - 0.000032i$
1.000000	$0.463026 - 0.216072i$	$-0.064526 + 0.383530i$	$0.816203 - 0.082329i$
0.445724	$-0.036567 - 0.391359i$	$-0.236744 - 0.298458i$	$0.178283 - 0.299444i$
-0.014976	$-1.987468 - 0.361121i$	$-1.152406 + 0.981613i$	$-2.368016 + 0.144692i$

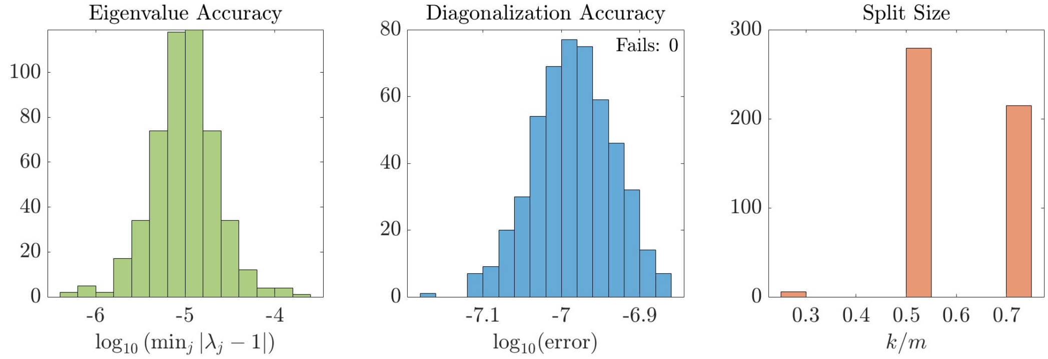


Figure 4.6. Statistics for 500 runs of **RPD** applied to a normalized version of the singular pencil (1.3) with $||A||_2 = 0.6940$, $||B||_2 = 1$, and $\varepsilon = 10^{-6}$. For the first plot, $\{\lambda_i\}_{1 \le i \le 4}$ are the approximate eigenvalues obtained via divide-and-conquer. Once again, no runs defaulted to calling `eig`. As in the previous examples, we consider only splits on subproblems with $m > 3$ for the third histogram; as a result, it records only the first split of each run, and the total number of splits is 500 (the number of runs).

a typical random one is likely to yield an eigenvalue near one.[4] Randomization has another advantage here: while QZ always produces the same three spurious "eigenvalues," divide-and-conquer does not – meaning multiple runs can help distinguish true eigenvalues from fake. Finally, we provide in Figure 4.6 a few empirical statistics for 500 runs of **RPD** on a normalized version of (A, B).

[4] Earlier work of Demmel and Kågström [43] took a different approach to establishing the stability of QZ on singular pencils, demonstrating that QZ will preserve Kronecker canonical form, and therefore successfully recover true eigenvalues, provided round-off errors are small enough. This nevertheless cannot explain the strong performance of **RPD**, as Gaussian perturbations *will* change the Kronecker structure of the pencil with high probability.

Content Acknowledgement: This chapter is repurposed from a submitted work co-authored with James Demmel and Ioana Dumitriu:

- J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.

The dissertation author was the primary investigator and author of this paper.

Chapter 5

Specialization to the Definite Generalized Eigenvalue Problem

In this chapter, we consider a specialization of the generalized eigenvalue problem to *definite* matrix pencils. While we have mentioned definite pencils a few times throughout the thesis, we finally define them precisely here.

Definition 5.0.1. The pencil (A, B) is *definite* if A and B are Hermitian and

$$\gamma(A, B) = \min_{||x||_2=1} |x^H(A + iB)x| = \min_{||x||_2=1} \sqrt{(x^HAx)^2 + (x^HBx)^2} > 0.$$

$\gamma(A, B)$ is the *Crawford number* of (A, B).

In some sense, the definite eigenvalue problem can be considered a generalization of the single-matrix Hermitian eigenvalue problem. A few important properties carry over directly: the eigenvalues of a definite pencil are real (though this is not so easy to see from Definition 5.0.1) and left/right eigenspaces are the same. As we'll explore in this chapter, we also expect better stability for eigenvalues/eigenvectors under perturbation (compared to the generic case of Section 1.2).

From a numerical perspective, any eigensolver hoping to perform optimally on definite pencils should exploit their available structure (and their relatively constrained spectra). With this in mind, we aim to refine the building blocks of divide-and-conquer, specifically those covered in Chapters 2 and 3, to take advantage of the observations discussed above. In particular, we consider the following.

1. Can we obtain pseudospectral shattering without destroying the structure of (A, B)? That is, can we ensure that the perturbed matrices $\widetilde{A}$ and $\widetilde{B}$ are both Hermitian with $(\widetilde{A}, \widetilde{B})$ definite? Note that this cannot be accomplished with the Ginibre perturbations considered so far since these are non-symmetric in general.

2. If $(\widetilde{A}, \widetilde{B})$ is definite – and therefore $\Lambda(\widetilde{A}, \widetilde{B})$ is contained in a union of intervals on the real axis – can an alternative approach compute spectral projectors more efficiently than **IRS**? As hinted at in Chapter 3, we might hope that the Indicator Approximation Problem has better answers for $S \subset \mathbb{R}$.

As we demonstrate in Sections 5.2 and 5.3, the answer to both questions is yes. Throughout, our goal is not only to produce a specialized version **RPD** but also to demonstrate the flexibility of divide-and-conquer as a high-level strategy.

Guide to Chapter Five: In Section 5.1 we present background information on definite pencils and provide motivation for devising a specialized algorithm. Section 5.2 proves a version of pseudospectral shattering for definite pencils under diagonal or GUE perturbations (see Definition 5.2.3). In Section 5.3 we then build a definite version of pseudospectral divide-and-conquer using a weighted Halley iteration of Nakatsukasa, Bai, and Gygi [105].

5.1 Motivation and Background

We begin in this section by reviewing the theory of definite pencils.[1] Our goal here is to unpack Definition 5.0.1. That is, when should we expect that the Crawford number

[1] See [126, Section VI.3] for more.

of an arbitrary pair of Hermitian matrices is nonzero? Perhaps more importantly, what do we gain – besides real eigenvalues – if we can guarantee that the pencil (A, B) is definite? Answers to these questions will inform the modifications of divide-and-conquer discussed in the following sections.

To build intuition, we first note that (A, B) is guaranteed to be definite if either A or B is positive definite. In this case, $\gamma(A, B)$ can be bounded from below by the smallest singular value of A or B (whichever is positive definite) and moreover it is easy to see that (A, B) must have real eigenvalues. If B is positive definite with Cholesky factorization $B = R^H R$, for example, then (A, B) and the Hermitian matrix $R^{-H} A R^{-1}$ have the same spectrum.

While this is arguably the most common way definite pencils appear in application, positive-definiteness is not required for (A, B) to be definite. In general, the Crawford number $\gamma(A, B)$ suggests that (A, B) will be definite provided $x^H A x$ and $x^H B x$ are not simultaneously zero for $x \in \mathbb{C}^n$. Given an arbitrary definite pencil (A, B), it is possible to find a Möbius transformation – with real coefficients – that transforms (A, B) into a pencil with at least one positive-definite matrix (see [126, Theorem VI.1.18]). Hence, every definite pencil has real eigenvalues, and we also have the following.

Proposition 5.1.1. *If the pencil (A, B) is definite then it is regular and diagonalizable. In particular, there exists a nonsingular matrix X such that $X^H A X$ and $X^H B X$ are diagonal.*

Again, this is easy to see if one of A and B is positive definite. Continuing the above example – where $B = R^H R$ is positive definite – the unitary diagonalization $R^{-H} A R^{-1} = U \Lambda U^H$ implies that (A, B) can be diagonalized by $X = R^{-1} U$. As we see here, the matrix X in Proposition 5.1.1, which contains right eigenvectors of (A, B) as in a standard diagonalization, is *not* unitary in general. This marks a first downgrade from the Hermitian eigenvalue problem: the eigenvectors of a definite pencil are not guaranteed to be mutually orthogonal. Instead, the best we can say is that left and right deflating subspaces of (A, B) corresponding to different sets of eigenvectors are orthogonal. To see this, let $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ in Proposition 5.1.1. Since $X^H AX$ and $X^H BX$ are both diagonal, we have

$$X_2^H AX_1 = X_2^H BX_1 = 0. \tag{5.1}$$

In other words, the columns of X_2 – which form a basis for a right deflating subspace of (A, B) – are orthogonal to AX_1 and BX_1, whose columns span the left deflating subspace corresponding to the remaining eigenvectors in X_1. While left/right eigenspaces of (A, B) are the same, note that left/right deflating subspaces are generally not the same.

Throughout this chapter, it will be convenient to work with a particular choice of diagonalizing matrix X satisfying

$$(X^H AX, X^H BX) = (\Lambda_A, \Lambda_B) = (\mathrm{diag}(\alpha_1, \ldots, \alpha_n), \mathrm{diag}(\beta_1, \ldots, \beta_n)) \tag{5.2}$$

for $\alpha_i, \beta_i \in \mathbb{R}$ with $\alpha_i^2 + \beta_i^2 = 1$. Note that X can be obtained from any matrix satisfying Proposition 5.1.1 by scaling its columns accordingly. Importantly, the norms of X and X^{-1} are linked to $\gamma(A, B)$, as demonstrated by the following observation of Elsner and Sun [51, Proof of Theorem 2.3].

Lemma 5.1.2 (Elsner and Sun 1982)**.** *Let* (A, B) *be a definite pencil and let* X *be a nonsingular eigenvector matrix satisfying* (5.2)*. Then*

$$||X||_2^2 \leq \gamma(A, B)^{-1} \quad \textit{and} \quad ||X^{-1}||_2^2 \leq \gamma(A, B)^{-1} ||(A, B)||_2^2$$

and therefore

$$\kappa_2(X) \leq \frac{||(A, B)||_2}{\gamma(A, B)}.$$

While this result leads to clean perturbation bounds, it is possibly quite loose; there may well be diagonalizing matrices with much better conditioning than X. Nevertheless, Lemma 5.1.2 implies that the eigenvector matrix of a definite pencil is guaranteed to be well-conditioned provided its Crawford number is sufficiently far from zero.

Definite pencils were first explored rigorously by Crawford [33,34]. In the numerical linear algebra literature, definite pencils are of interest not just because they appear in applications but because they exhibit stronger stability properties than general pencils [89, 104,124]. As we will see, perturbation bounds typically depend on $\gamma(A,B)$. Hence, it is not enough to simply know that the Crawford number of (A,B) is nonzero; useful perturbation results will depend on tight lower bounds for $\gamma(A,B)$. Unfortunately, bounding $\gamma(A,B)$ is not straightforward in general – e.g., it is not sufficient to have information on $\sigma_n(A)$ and $\sigma_n(B)$, as we might hope (unless, again, one of A and B is positive definite).

Example 5.1.3. Consider the following 2×2 pencils:

1. $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
2. $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$.

In the first example, taken from Stewart and Sun [126, Example VI.1.14], A and B are nonsingular but (A,B) is not definite (it's eigenvalues are $\pm i$). In the second, A and B are both singular but $\gamma(A,B) = 1$. Together these examples imply that $\gamma(A,B)$ cannot be easily bounded from below (or above!) in terms of $\sigma_n(A)$ or $\sigma_n(B)$ in general.

We will not consider the task of estimating the Crawford number of an arbitrary Hermitian pencil in this thesis, instead assuming access to $\gamma(A,B)$, or a lower bound for it, as something of a black box. For a sample of some of the numerical tools available to obtain such a lower bound, see [35,68,136]. While in principle computing $\gamma(A,B)$ requires solving an optimization problem over a field of values, many of these estimations can be done relatively cheaply (i.e., in matrix multiplication time).

With this in mind, we now state a few perturbation bounds for definite pencils. Recalling Lemma 5.1.2, we expect $\gamma(A,B)$ to appear here, implicitly containing information on eigenvector conditioning. Indeed, we have the following standard eigenvalue perturbation bound[2] of Stewart [124, Theorem 3.2].

Theorem 5.1.4 (Stewart 1979)**.** *Let (A, B) be an $n \times n$ definite pencil and suppose that the Hermitian matrices $E, F \in \mathbb{C}^{n\times n}$ satisfy*

$$\frac{\sqrt{||E||_2^2 + ||F||_2^2}}{\gamma(A,B)} < 1.$$

Then the pencil $(\widetilde{A}, \widetilde{B}) = (A+E, B+F)$ is definite. Moreover, if $\lambda_1 \leq \cdots \leq \lambda_n$ and $\widetilde{\lambda}_1 \leq \cdots \leq \widetilde{\lambda}_n$ are the eigenvalues of (A, B) and $(\widetilde{A}, \widetilde{B})$, respectively, then for all $1 \leq i \leq n$,

$$\frac{|\lambda_i - \widetilde{\lambda}_i|}{\sqrt{(|\lambda_i|^2+1)(|\widetilde{\lambda}_i|^2+1)}} \leq \frac{\sqrt{||E||_2^2 + ||F||_2^2}}{\gamma(A,B)}.$$

Representing the eigenvalues in Theorem 5.1.4 as $\lambda_i = \alpha_i/\beta_i$, $\widetilde{\lambda}_i = \widetilde{\alpha}_i/\widetilde{\beta}_i$ and recalling (1.23), we can compare this result to our general perturbation bound from Section 1.2 (Theorem 1.2.11). Stewart's specialization offers a significant improvement, replacing the minmax bound in Theorem 1.2.11 with an explicit eigenvalue pairing. In particular, Theorem 1.2.11 can only guarantee that each perturbed eigenvalue is close to an eigenvalue of (A, B), not necessarily a unique one.

Note that Theorem 5.1.4 also provides a criterion for ensuring that a perturbed pencil is definite. In fact, its proof gives a usable lower bound

$$\begin{aligned}\gamma(\widetilde{A}, \widetilde{B}) &\geq \min_{||x||_2=1} \left\{ \sqrt{(x^H A x)^2 + (x^H B x)^2} - \sqrt{(x^H E x)^2 + (x^H F x)^2} \right\} \\ &\geq \left[1 - \frac{\sqrt{||E||_2^2 + ||F||_2^2}}{\gamma(A,B)} \right] \gamma(A,B).\end{aligned} \tag{5.3}$$

This will be important in the next section, where we hope to apply Hermitian perturbations to (A, B) without sacrificing definiteness.

We turn next to eigenvectors. As in Section 1.2, we forgo results for general eigenspaces/deflating subspaces in favor of bounds on individual eigenvectors. Here, we make use of another result of Stewart [124, Theorem 4.3].

[2] A slight improvement on this bound was subsequently provided by Sun [129], though the difference is unimportant for our purposes.

Theorem 5.1.5 (Stewart 1979)**.** *Let (A, B) be an $n \times n$ definite pencil and suppose $(\widetilde{A}, \widetilde{B}) = (A+E, B+F)$ is also definite, where $E, F \in \mathbb{C}^{n\times n}$ are Hermitian. Let v be a right eigenvector (A, B) corresponding to the eigenvalue λ_1. Suppose that $(\widetilde{A}, \widetilde{B})$ has eigenvalues $\widetilde{\lambda}_1, \dots \widetilde{\lambda}_n$ and let*

$$\delta = \min_{i>1} \left\{ \frac{|\lambda_1 - \widetilde{\lambda}_i|}{\sqrt{(|\lambda_1|^2+1)(|\widetilde{\lambda}_i|^2+1)}} \right\}.$$

If $\frac{\sqrt{||E||_2^2 + ||F||_2^2}}{\delta} < \gamma(\widetilde{A}, \widetilde{B})$, then there exists a right eigenvector $\widetilde{v}$ of $(\widetilde{A}, \widetilde{B})$ corresponding to $\widetilde{\lambda}_1$ such that

$$\frac{||\widetilde{v} - v||_2}{||v||_2} \leq \frac{\sqrt{||E||_2^2 + ||F||_2^2}}{\delta\gamma(\widetilde{A}, \widetilde{B})}.$$

Again, this is much cleaner than our general eigenvector bound Theorem 1.2.14. Importantly, it does not require that either (A, B) or $(\widetilde{A}, \widetilde{B})$ has distinct eigenvalues, though the bound will be poor if $(\widetilde{A}, \widetilde{B})$ has multiple eigenvalues close to λ_1, which we expect to occur if (A, B) has repeated eigenvalues thanks to Theorem 5.1.4. This result also does not require that B or $\widetilde{B}$ is invertible; as in Theorem 5.1.4 and even Theorem 1.2.11, this is a consequence of using the chordal metric (1.22), in this case to define δ.

In total, these perturbation bounds indicate that the near-ideal stability of the Hermitian eigenvalue problem – captured by the classical pair of Weyl's inequality and the Davis-Kahan theorem [37] – does not carry over to the eigenvalues and eigenvectors of definite pencils. Nevertheless, the bounds presented above are much stronger than those available for general pencils, particularly when $\gamma(A, B)$ is far from zero, and they will simplify the analysis for both pseudospectral shattering and divide-and-conquer.

We complete this section with a motivating example,[3] which provides further intuition for the way definite pencils arise in practice.

Example 5.1.6 (Quantum Chemistry)**.** The pioneering 1926 work of Schrödinger demonstrated that stationary states of any quantum system[4] can be described by the time-

[3] Note that Example 1.1.5 also presented a definite pencil.

[4] For background on the quantum mechanics used in this example, see the standard reference [65].

independent Schrödinger equation (TISE)

$$\mathcal{H}(r)\psi(r) = E\psi(r). \tag{5.4}$$

Here, $\mathcal{H}$ is a time-independent Hamiltonian operator and $\psi(r)$ is a stationary wave function corresponding to energy E. For simplicity, we assume that both $\mathcal{H}$ and ψ are radial – i.e., functions of a single spatial variable r taking values in $\mathbb{R}_{\geq 0}$.

Given an arbitrary Hamiltonian $\mathcal{H}(r)$, how do we obtain solutions to the TISE? One option is to choose a set of radial basis functions $\phi_1(r), \ldots, \phi_n(r)$ and expand $\psi(r)$ as

$$\psi(r) = \sum_{j=1}^{n} c_j \phi_j(r) \tag{5.5}$$

for some (unknown) coefficients $c_1, \ldots, c_n$. Inserting this expansion into (5.4) yields

$$\sum_{j=1}^{n} c_j \mathcal{H}(r)\phi_j(r) = E \sum_{j=1}^{n} c_j \phi_j(r) \tag{5.6}$$

and therefore (taking an inner product)

$$\sum_{j=1}^{n} c_j \int \phi_i(r)^H \mathcal{H}(r)\phi_j(r) dr = E \sum_{j=1}^{n} c_j \int \phi_i^H(r)\phi_j(r) dr \quad 1 \leq i \leq n. \tag{5.7}$$

Define now the $n \times n$ matrices H, S as follows:

$$H_{ij} = \int \phi_i^H(r)\mathcal{H}(r)\phi_j(r) dr \quad \text{and} \quad S_{ij} = \int \phi_i^H(r)\phi_j(r) dr. \tag{5.8}$$

It is easy to see from (5.7) that $c = [c_1 \; c_2 \; \cdots \; c_n]^T$ is an eigenvector of the definite pencil (H, S) corresponding to eigenvalue E. Hence, we obtain solutions to the TISE via the generalized eigenvalue problem (H, S). This is a standard approach to the problem in quantum chemistry (see for example [55, 119]).

While the pencil (H, S) is always definite – a consequence of the fact that the *overlap matrix* S is positive definite – the structure of the individual matrices H and S is highly dependent on the choice of basis $\phi_1(r), \ldots, \phi_n(r)$. If this basis is orthonormal with respect to the standard inner product on continuous functions, for example, S is the identity and (H,S) reduces to a standard eigenvalue problem. More importantly, a clever choice of basis may be able to guarantee that the matrices H and S are banded, offering additional structure that can be exploited for efficiency.

5.2 Pseudospectral Shattering under GUE and Diagonal Perturbations

We turn now to a specialized version of pseudospectral shattering, which considers only symmetric or diagonal perturbations. The regularizing effect of such perturbations on the (pseudo)spectra of Hermitian matrices has already been explored (see for example [119]). Our goal is to generalize this work. Given an arbitrary definite pencil (A,B), we aim to prove shattering for a symmetrized version of $\Lambda_\epsilon(A,B)$.

Definition 5.2.1. The *symmetric ϵ-pseudospectrum* of (A,B) is

$$\Lambda_\epsilon^{\mathrm{sym}}(A,B) = \left\{ z : \begin{array}{l} (A+E)u = z(B+F)u \ \text{ for } \ u \neq 0 \ \text{ and} \\ E, F \text{ Hermitian with } \sqrt{||E||_2^2 + ||F||_2^2} \leq \epsilon \end{array} \right\}.$$

We define $\Lambda_\epsilon^{\mathrm{sym}}(A,B)$ here in terms of $\sqrt{||E||_2^2 + ||F||_2^2}$ in an effort to simplify bounds. In particular, Theorem 5.1.4 implies $\Lambda_\epsilon^{\mathrm{sym}}(A,B) \subseteq \mathbb{R}$ provided $\epsilon < \gamma(A,B)$. Hence, for ϵ sufficiently small shattering can be defined relative to a set of equally spaced points on the real axis (as opposed to the two-dimensional grid necessary in the general case). Here, we say that $\Lambda_\epsilon^{\mathrm{sym}}(A,B) \subset \mathbb{R}$ is shattered with respect to a set of points $g = \{g_i\} \subseteq \mathbb{R}$ if each eigenvalue of (A,B) lies in a unique interval (g_i, g_{i+1}) and moreover $\Lambda_\epsilon^{\mathrm{sym}}(A,B) \cap g = \emptyset$. This simplification already suggests efficiency gains for divide-and-conquer: not only will we have fewer potential splits to search over, but we can also guarantee that an optimal one exists.[5]

As in the general case, we'll need a version of Bauer-Fike to prove pseudospectral shattering. While we could simply use $\Lambda_\epsilon^{\mathrm{sym}}(A,B) \subset \Lambda_\epsilon(A,B)$ and therefore Theorem 1.2.9,

[5] Here, an optimal split separates the spectrum into sets of size $\lfloor \frac{n}{2} \rfloor$ and $\lceil \frac{n}{2} \rceil$. Recall that in the general case, the best we could say was that a split separating at least a fifth of the eigenvalues existed.

we prefer the following specialized version, which again restricts to the real axis. Note that R_ϵ here can be bounded via Lemma 1.2.8 (though this is likely sub-optimal).

Theorem 5.2.2. *Let (A, B) be a definite pencil with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. For $\epsilon < \min\{\sigma_n(B), \gamma(A, B)\}$ let $R_\epsilon > 0$ such that $\Lambda_\epsilon^{\mathrm{sym}}(A, B) \subseteq (-R_\epsilon, R_\epsilon)$. Further, set*

$$r_\epsilon = \frac{\epsilon(1 + R_\epsilon^2)}{\gamma(A, B)}$$

and

$$r_i = \begin{cases} \frac{1}{||B||_2}, & \textit{if } A = 0 \\ \max\left\{\frac{1}{||B||_2}, \frac{|\lambda_i|}{||A||_2}\right\}, & \textit{otherwise} \end{cases}$$

for $1 \leq i \leq n$. Then

$$\bigcup_{i=1}^{n}(\lambda_i - \epsilon r_i, \lambda_i + \epsilon r_i) \subseteq \Lambda_\epsilon^{\mathrm{sym}}(A, B) \subseteq \bigcup_{i=1}^{n}(\lambda_i - r_\epsilon, \lambda_i + r_\epsilon).$$

Proof. The lower inclusion follows directly from the proof of the corresponding portion of Theorem 1.2.9. The upper inclusion, meanwhile, is a consequence of Theorem 5.1.4. □

With this as a backdrop, we now ask: what Hermitian perturbations should we use to obtain shattering for $\Lambda_\epsilon^{\mathrm{sym}}(A, B)$? The most natural extension of the general case is to consider perturbations sampled from the Gaussian Unitary Ensemble (GUE).

Definition 5.2.3. The $n \times n$ *Gaussian Unitary Ensemble* GUE(n) consists of matrices of the form

$$Z = \frac{G + G^H}{\sqrt{2n}}$$

for G an $n \times n$ complex Gaussian random matrix with i.i.d. entries sampled from $\mathcal{N}_{\mathbb{C}}(0, 1)$.

Each $Z \in$ GUE(n) can be interpreted as a symmetrized Ginibre random matrix. Accordingly, analogs of Lemma 2.1.1 and Lemma 1.3.7 – which recall were important building blocks of the proof of shattering – generalize easily, as we demonstrate below. Here, a singular value bound counterpart to Lemma 1.3.7 is obtained as a corollary of a recent result of Aizenman, Peled, Schenker, Shamis, and Sodin [3].

Lemma 5.2.4. *If $Z \in \mathrm{GUE}(n)$ then*

$$\mathbb{P}\left[||Z||_2 \geq 4 + \sqrt{2}\right] \leq 2e^{-n}.$$

Proof. Write $Z = \frac{G+G^H}{\sqrt{2}}$ for G an $n \times n$ Ginibre random matrix. Since $||Z||_2 \leq \sqrt{2}||G||_2$, we have

$$\mathbb{P}\left[||Z||_2 \geq 4 + \sqrt{2}\right] \leq \mathbb{P}\left[||G||_2 \geq 2\sqrt{2} + 1\right]. \tag{5.9}$$

Applying Lemma 2.1.1 completes the proof. □

Lemma 5.2.5 (Aizenman et al. 2017)**.** *Let $M \in \mathbb{C}^{n\times n}$ be Hermitian and let $Z \in \mathrm{GUE}(n)$. For any $t \geq 1$ and an absolute constant $C < \infty$,*

$$\mathbb{P}\left[||(M+Z)^{-1}||_2 \geq tn\right] \leq \frac{C}{t}.$$

Corollary 5.2.6. *Let $M \in \mathbb{C}^{n\times n}$ be Hermitian and let $Z \in \mathrm{GUE}(n)$. For any $\gamma > 0$, $t \leq \frac{\gamma}{n}$, and an absolute constant $C < \infty$,*

$$\mathbb{P}\left[\sigma_n(M+\gamma Z) \leq t\right] \leq \frac{Cnt}{\gamma}.$$

Proof. We have

$$\mathbb{P}\left[\sigma_n(M+\gamma Z) \leq t\right] = \mathbb{P}\left[\left|\left|\left(\frac{1}{\gamma}M + Z\right)^{-1}\right|\right|_2 \geq \frac{\gamma}{t}\right] \leq \frac{Cnt}{\gamma}, \tag{5.10}$$

where the last inequality follows from Lemma 5.2.5. □

Recalling Example 5.1.6, we could alternatively consider perturbing with random diagonal matrices, thereby preserving banded structure if applicable.

Definition 5.2.7. The *essential supremum* of a probability density ρ on $\mathbb{R}$ is

$$||\rho||_\infty = \inf\left\{a : \text{the set } S_a = \{x : \rho(x) > a\} \text{ has measure zero}\right\}.$$

We say that ρ is *bounded* if $||\rho||_\infty < \infty$.

Definition 5.2.8. Let ρ be a bounded probability density on $\mathbb{R}$. The corresponding *diagonal random matrix* D^ρ has nonzero entries sampled according to ρ.

Importantly, both perturbation options satisfy the following key result, which – as we will see – is the backbone of pseudospectral shattering in this specialized setting.

Theorem 5.2.9 (Key Result). *Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix and let $I \subset \mathbb{R}$ be any interval. Suppose the random matrix V satisfies one of the following:*

1. *$V \in \mathrm{GUE}(n)$.*
2. *$V = D^\rho$ for a bounded probability density ρ on $\mathbb{R}$.*

In either case, there exists a constant $C < \infty$ (uniform in A and n) such that

$$\mathbb{P}\left[A + V \text{ has at least two eigenvalues in } I\right] \leq C|I|^2 n^2.$$

Proof. The diagonal case was proved by Minami [100] and subsequently generalized to GUE(n) by Aizenman, Peled, Schenker, Shamis, and Sodin [3]. When $V = D^\rho$, the constant C depends on $||\rho||_\infty$. □

Theorem 5.2.9 pairs naturally with the following technical lemma, which is the final building block we need to prove shattering.

Lemma 5.2.10. *Let (A, B) be a definite pencil. If (A, B) has at least j eigenvalues in the interval $(z_0 - r, z_0 + r)$ for $z_0 \in \mathbb{R}$ and $r > 0$ then*

$$\sigma_{n-j+1}(A - z_0 B) \leq \frac{r||(A, B)||_2^2}{\gamma(A, B)}.$$

Proof. Let X be a nonsingular eigenvector matrix of (A, B) satisfying (5.2). Standard singular value inequalities imply

$$\begin{aligned}\sigma_{n-j+1}(A - z_0 B) &= \sigma_{n-j+1}(X^{-H}(\Lambda_A - z_0\Lambda_B)X^{-1}) \\ &\leq ||X^{-H}||_2 \sigma_{n-j+1}(\Lambda_A - z_0\Lambda_B)||X^{-1}||_2 \qquad (5.11) \\ &= ||X^{-1}||_2^2 \sigma_{n-j+1}(\Lambda_A - z_0\Lambda_B).\end{aligned}$$

Since (A, B) has j eigenvalues in $(z_0 - r, z_0 + r)$, at least j diagonal entries $\alpha_i - z_0\beta_i$ of $\Lambda_A - z_0\Lambda_B$ satisfy

$$|\alpha_i - z_0\beta_i| = |\beta_i| \left| \frac{\alpha_i}{\beta_i} - z_0 \right| \leq |\beta_i| r \leq r, \tag{5.12}$$

meaning $\sigma_{n-j+1}(\Lambda_A - z_0\Lambda_B) \leq r$. Applying this to (5.11) alongside Lemma 5.1.2 yields the final bound. □

Putting everything together, Theorem 5.2.11 presents pseudospectral shattering for (A, B) under GUE perturbations. A direct replication of its proof also implies shattering for diagonal perturbations, though the final form will depend on the distribution ρ. In that case, a counterpart to Corollary 5.2.6 follows[6] from work of Wegner [142].

Theorem 5.2.11. *Let (A, B) be an $n \times n$ definite pencil with $||A||_2, ||B||_2 \leq 1$ and let $(\widetilde{A}, \widetilde{B}) = (A + \gamma Z_1, B + \gamma Z_2)$ for independent $Z_1, Z_2 \in \mathrm{GUE}(n)$ and $\gamma < \frac{\gamma(A,B)}{12\sqrt{2}}$. Taking $\alpha \geq \log_n(\gamma^{-1}) + 2$, choose $z_0 \in (-4 - \omega, -4)$ uniformly at random and construct the grid of points*

$$g = \{z_0 + j\omega : 0 \leq j \leq \lceil 8/\omega \rceil + 1\} \quad \text{for } \omega = \frac{\gamma^4}{n^{4\alpha+3}}.$$

Then $\Lambda_\epsilon^{\mathrm{sym}}(\widetilde{A}, n^\alpha \widetilde{B})$ is shattered with respect to g for $\epsilon = \frac{\gamma^5}{10n^{4\alpha+5}} \leq \frac{\gamma^9}{10n^{13}}$ with probability at least $1 - O(\frac{1}{n})$.

Proof. We condition on the events $||Z_1||_2, ||Z_2||_2 < 6$ and $\sigma_n(\widetilde{B}) > n^{-\alpha}$, which imply the following:

1. **The perturbed and scaled pencil $(\widetilde{A}, n^\alpha \widetilde{B})$ is definite with $\gamma(\widetilde{A}, n^\alpha \widetilde{B}) \geq \frac{\gamma(A,B)}{2}$.** If $||Z_1||_2, ||Z_2||_2 < 6$ then $||A - \widetilde{A}||_2, ||B - \widetilde{B}||_2 < 6\gamma$, meaning

$$\frac{\sqrt{||\widetilde{A} - A||_2^2 + ||\widetilde{B} - B||_2^2}}{\gamma(A, B)} < \frac{6\sqrt{2}\gamma}{\gamma(A, B)} < \frac{1}{2}. \tag{5.13}$$

Recalling (5.3), this implies that $(\widetilde{A}, \widetilde{B})$ is definite with $\gamma(\widetilde{A}, \widetilde{B}) \geq \frac{\gamma(A,B)}{2}$. We finish by noting $\gamma(\widetilde{A}, n^\alpha \widetilde{B}) \geq \gamma(\widetilde{A}, \widetilde{B})$.

[6] Combine [3, Equation 1.2] with Markov's inequality, noting that $\sigma_n(M + D^\rho) < t$ implies that $M + D^\rho$ has at least one eigenvalue in $(-t, t)$.

2. The spectrum of $(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in $(-4, 4)$.

Since $\widetilde{B}$ is almost surely invertible, we have

$$\rho(n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}) \leq ||n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}||_2 \leq \frac{||\widetilde{A}||_2}{n^\alpha \sigma_n(\widetilde{B})}. \tag{5.14}$$

Now $\sigma_n(\widetilde{B}) > n^{-\alpha}$ and $||\widetilde{A}||_2 \leq ||A||_2 + \gamma||Z_1||_2 < 4$ since $||Z_1||_2 < 6$, $||A||_2 \leq 1$, and $\gamma < \frac{1}{2}$ (the latter following from the fact that $\gamma(A, B) \leq \sqrt{2}$ if both A and B have spectral norm at most one), so (5.14) becomes $\rho(n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}) < 4$. Recalling that $(\widetilde{A}, n^\alpha \widetilde{B})$ and $n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$ have the same eigenvalues and moreover that the eigenvalues of $(\widetilde{A}, n^\alpha \widetilde{B})$ are real since the pencil is definite, we conclude $\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \subseteq (-4, 4)$.

3. For $\delta > 0$, the event $\mathrm{gap}(\widetilde{A}, n^\alpha \widetilde{B}) \leq \delta$ occurs with probability at most $C\frac{n^{\alpha+2}\delta}{\gamma^4}$.

Let $\mathcal{N} \subseteq \mathbb{R}$ be a minimal $\frac{\delta}{2}$-net covering $(-4, 4)$, where $|\mathcal{N}| \leq \lceil \frac{16}{\delta} \rceil$. Since $\Lambda(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in $(-4, 4)$ it is easy to see

$$\begin{aligned} \mathbb{P}\left[\mathrm{gap}(\widetilde{A}, n^\alpha \widetilde{B}) \leq \delta\right] &= \mathbb{P}\left[|\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \cap (y - \delta, y + \delta)| \geq 2 \text{ for some } y \in \mathcal{N}\right] \\ &\leq |\mathcal{N}| \max_{y \in \mathcal{N}} \mathbb{P}\left[|\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \cap (y - \delta, y + \delta)| \geq 2\right]. \end{aligned} \tag{5.15}$$

Now for any $y \in \mathcal{N}$, $|\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \cap (y - \delta, y + \delta)| \geq 2$ implies

$$\sigma_{n-1}(\widetilde{A} - yn^\alpha \widetilde{B}) \leq \frac{\delta||(\widetilde{A}, n^\alpha \widetilde{B})||_2^2}{\gamma(\widetilde{A}, n^\alpha \widetilde{B})} \tag{5.16}$$

by Lemma 5.2.10. Applying $\gamma(\widetilde{A}, n^\alpha \widetilde{B}) \geq \frac{\gamma(A,B)}{2} \geq 8\gamma$ and $||(\widetilde{A}, n^\alpha \widetilde{B})||_2 \leq ||\widetilde{A}||_2 + n^\alpha||\widetilde{B}||_2 \leq 8n^\alpha$, this becomes

$$\frac{8n^{2\alpha}\delta}{\gamma} \geq \sigma_{n-1}(\widetilde{A} - yn^\alpha \widetilde{B}) = \gamma\sigma_{n-1}\left(\frac{1}{\gamma}[A - yn^\alpha \widetilde{B}] + Z_1\right). \tag{5.17}$$

Thus, $|\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \cap (y - \delta, y + \delta)| \geq 2$ implies $\sigma_{n-1}(M + Z_1) \leq \frac{8n^{2\alpha}\delta}{\gamma^2}$ for $M = \frac{1}{\gamma}(A - yn^\alpha \widetilde{B})$. But $M + Z_1$ is Hermitian and the singular values of a Hermitian matrix are the absolute values of its eigenvalues, so this is equivalent to $M + Z_1$ having at least two eigenvalues in the interval $[-\frac{8n^{2\alpha}\delta}{\gamma^2}, \frac{8n^{2\alpha}\delta}{\gamma^2}]$. By Theorem 5.2.9, this occurs with probability at most

$C\frac{n^{4\alpha+2}\delta^2}{\gamma^4}$ for some absolute constant C. Plugging this into the union bound (5.15) and allowing C to absorb constants, we obtain

$$\mathbb{P}\left[\mathrm{gap}(\widetilde{A}, n^\alpha \widetilde{B}) \leq \delta\right] \leq \left\lceil \frac{16}{\delta} \right\rceil \left(C\frac{n^{4\alpha+2}\delta^2}{\gamma^4} \right) = C\frac{n^{4\alpha+2}\delta}{\gamma^4}. \tag{5.18}$$

So far, letting E_{cond} be the event that $||Z_1||_2, ||Z_2||_2 < 6$ and $\sigma_n(\widetilde{B}) > n^{-\alpha}$ – i.e., what we conditioned on at the start of the proof – we have shown

$$\mathbb{P}\left[\mathrm{gap}(\widetilde{A}, n^\alpha \widetilde{B}) > \delta,\ \Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \subseteq (-4, 4) \mid E_{\text{cond}}\right] \geq 1 - C\frac{n^{4\alpha+2}\delta}{\gamma^4}. \tag{5.19}$$

Choosing $\delta = \frac{\gamma^4}{n^{4\alpha+3}}$, this implies $\mathrm{gap}(\widetilde{A}, n^\alpha \widetilde{B}) > \omega$ and $\Lambda(\widetilde{A}, n^\alpha \widetilde{B}) \subseteq (-4, 4)$ with probability at least $1 - \frac{C}{n}$. In this case, by construction, each eigenvalue of $(\widetilde{A}, n^\alpha \widetilde{B})$ belongs to a unique interval (g_i, g_{i+1}) for $g_i \in g$. Moreover, the eigenvalues are well separated from the grid points with high probability: applying the same geometric argument made in the proof of (2.3.1), we have

$$\mathbb{P}\left[\min_{\lambda \in \Lambda(\widetilde{A}, n^\alpha \widetilde{B})} \mathrm{dist}_g(\lambda) \leq \frac{\omega}{2n^2}\right] \leq \frac{1}{n}. \tag{5.20}$$

Thus, a simple union bound implies that each eigenvalue of $(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in a unique grid interval, and is at least $\frac{\omega}{2n^2}$-away from the nearest grid point, with probability at least $1 - \frac{C+1}{n}$. When this occurs, shattering is guaranteed as long as $\Lambda_\epsilon^{\text{sym}}(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in a union of intervals of radius $\frac{\omega}{2n^2}$ centered at the eigenvalues of $(\widetilde{A}, n^\alpha \widetilde{B})$. Appealing to our version of Bauer-Fike for definite pencils (Theorem 5.2.2), we know that if $\epsilon < \min\left\{\gamma(\widetilde{A}, n^\alpha \widetilde{B}), \sigma_n(n^\alpha \widetilde{B})\right\}$, which can be enforced by taking $\epsilon < \frac{\gamma(A,B)}{2}$, $\Lambda_\epsilon^{\text{sym}}(\widetilde{A}, n^\alpha \widetilde{B})$ is contained in a union of intervals of radius

$$r_\epsilon = \frac{\epsilon(1 + R_\epsilon^2)}{\gamma(\widetilde{A}, n^\alpha \widetilde{B})} \leq \frac{2\epsilon(1 + R_\epsilon^2)}{\gamma(A, B)} \leq \frac{2\epsilon}{\gamma(A, B)}\left[1 + \left(\frac{\epsilon + 4}{1 - \epsilon}\right)^2\right], \tag{5.21}$$

where the last inequality follows from the upper bound on R_ϵ provided by Lemma 1.2.8. Further assuming $\epsilon < \frac{1}{2}$ so that $\frac{\epsilon+4}{1-\epsilon} < 9$ and recalling $12\sqrt{2}\gamma < \gamma(A, B)$, we conclude

$$r_\epsilon \leq \frac{164\epsilon}{12\sqrt{2}\gamma} \leq \frac{10\epsilon}{\gamma}. \tag{5.22}$$

Thus, we achieve shattering by taking $\frac{10\epsilon}{\gamma} \leq \frac{\omega}{n^2} = \frac{\gamma^4}{n^{4\alpha+5}}$ or equivalently $\epsilon \leq \frac{\gamma^5}{10n^{4\alpha+5}}$, which we note does not violate any of our assumptions on ϵ.

In total, we have shown that for $\epsilon = \frac{\gamma^5}{10n^{4\alpha+5}}$

$$\mathbb{P}\left[\Lambda_\epsilon^{\text{sym}}(\widetilde{A}, n^\alpha \widetilde{B}) \text{ is shattered w.r.t } g \mid E_{\text{cond}}\right] \geq 1 - \frac{C+1}{n}. \tag{5.23}$$

Since Lemma 5.2.4 and Corollary 5.2.6 imply

$$\mathbb{P}\left[||Z_1||_2, ||Z_2||_2 < 6,\ \sigma_n(\widetilde{B}) > n^{-\alpha}\right] \geq 1 - C'\frac{n^{1-\alpha}}{\gamma} - 4e^{-n} \tag{5.24}$$

for some absolute constant C', which can be further simplified to $1 - \frac{C'}{n} - 4e^{-n}$ since $\alpha \geq \log_n(\gamma^{-1}) + 2$, we conclude by Bayes' theorem

$$\mathbb{P}\left[\Lambda_\epsilon^{\text{sym}}(\widetilde{A}, n^\alpha \widetilde{B}) \text{ is shattered w.r.t. } g\right] \geq 1 - \frac{C''}{n} \tag{5.25}$$

for some final absolute constant C''. □

We close this section by stating perturbation results for $\Lambda_\epsilon^{\text{sym}}(A,B)$. The latter two mirror Lemma 2.3.3 and Lemma 2.3.4 from Chapter 2.

Lemma 5.2.12. *Let (A,B) be an $n \times n$ pencil. If $A', B' \in \mathbb{C}^{n\times n}$ satisfy $||A - A'||_2, ||B - B'||_2 \leq \eta < \frac{\epsilon}{\sqrt{2}}$ for $A - A'$ and $B - B'$ Hermitian, then*

$$\Lambda^{\text{sym}}_{\epsilon - \sqrt{2}\eta}(A', B') \subseteq \Lambda^{\text{sym}}_\epsilon(A,B).$$

Proof. Suppose $z \in \Lambda^{\text{sym}}_{\epsilon-\sqrt{2}\eta}(A', B')$. In this case, there exist Hermitian $E, F \in \mathbb{C}^{n\times n}$ with $||E||_2^2 + ||F||_2^2 \leq (\epsilon - \sqrt{2}\eta)^2$ such that $z \in \Lambda(A' + E, B' + F)$. Hence, we have $z \in \Lambda(A + (A' - A + E), B + (B' - B + F))$ with

$$\begin{aligned}||A' - A + E||_2^2 + ||B' - B + F||_2^2 &\leq (\eta + ||E||_2)^2 + (\eta + ||F||_2)^2 \\ &= 2\eta^2 + 2\eta(||E||_2 + ||F||_2) + ||E||_2^2 + ||F||_2^2 \\ &\leq 2\eta^2 + 2\sqrt{2}\eta(\epsilon - \sqrt{2}\eta) + (\epsilon - \sqrt{2}\eta)^2 \\ &= \epsilon^2.\end{aligned} \tag{5.26}$$

Here, the second inequality follows from the fact that $||E||_2 + ||F||_2$ takes maximum value $\sqrt{2}(\epsilon - \sqrt{2}\eta)$ subject to the constraint $||E||_2^2 + ||F||_2^2 \leq (\epsilon - \sqrt{2}\eta)^2$. Since $A' - A + E$ and $B' - B + F$ are Hermitian, we conclude $z \in \Lambda_\epsilon^{\text{sym}}(A, B)$. □

Lemma 5.2.13. *Let (A, B) be an $n \times n$ definite pencil and suppose $\Lambda_\epsilon^{\text{sym}}(A, B)$ is shattered with respect to a set of points $\{g_i\} \subset \mathbb{R}$ for some $0 < \epsilon < \gamma(A, B)$. If $A', B' \in \mathbb{C}^{n\times n}$ satisfy $||A - A'||_2, ||B - B'||_2 \leq \eta < \frac{\epsilon}{\sqrt{2}}$ for $A - A'$ and $B - B'$ Hermitian, then each eigenvalue of (A', B') shares a grid interval (g_i, g_{i+1}) with exactly one eigenvalue of (A, B) and $\Lambda^{\text{sym}}_{\epsilon - \sqrt{2}\eta}(A', B')$ is also shattered with respect to g.*

Proof. This follows from a straightforward recreation of the proof of Lemma 2.3.3, noting here that $\gamma(A', B') \geq 1 - \frac{\sqrt{2}\eta}{\gamma(A,B)}$, so (A', B') is definite, and moreover $\Lambda^{\text{sym}}_{\epsilon-\sqrt{2}\eta}(A', B') \subseteq \Lambda_\epsilon^{\text{sym}}(A, B)$ by Lemma 5.2.12. □

Lemma 5.2.14. *Let (A, B) be an $n \times n$ definite pencil and suppose $\Lambda_\epsilon^{\text{sym}}(A, B)$ is shattered with respect to a grid of points $\{g_i\} \subset (-r, r)$ for some $0 < \epsilon < \gamma(A, B)$. Let $A', B' \in \mathbb{C}^{n\times n}$ satisfy $||A - A'||_2, ||B - B'||_2 \leq \eta < \frac{\epsilon}{\sqrt{2}}$ for $A - A'$ and $B - B'$ Hermitian. If (λ, v) is an eigenpair of (A, B) and*

$$\eta \; < \; \frac{\epsilon\gamma(A', B')}{\sqrt{2}||B||_2} \, \frac{||B'||_2 + ||B||_2}{||B'||_2(1 + r^2) + \gamma(A', B')}$$

then there exists an eigenpair (λ', v') of (A', B') such that λ and λ' share a grid interval (g_i, g_{i+1}) and

$$\frac{||v - v'||_2}{||v||_2} \; \leq \; \frac{\sqrt{2}\eta}{\gamma(A', B')} \, \frac{||B||_2||B'||_2(1 + r^2)}{\epsilon||B'||_2 + (\epsilon - \sqrt{2}\eta)||B||_2} \; < \; 1.$$

Proof. By Lemma 5.2.13, we know that (A', B') is definite, $\Lambda^{\text{sym}}_{\epsilon-\sqrt{2}\eta}(A', B')$ is shattered with respect to g, and each eigenvalue of (A, B) shares a unique grid interval with an eigenvalue of (A', B'). Let λ' be the eigenvalue of (A', B') corresponding to λ. By construction, any other eigenvalue $\mu \in \Lambda(A', B')$ belongs to a different grid interval; since $\Lambda_\epsilon^{\text{sym}}(A, B)$ and

$\Lambda^{\text{sym}}_{\epsilon-\sqrt{2}\eta}(A', B')$ are shattered with respect to g and contain intervals of radius $\frac{\epsilon}{||B||_2}$ and $\frac{\epsilon-\sqrt{2}\eta}{||B'||_2}$ around the eigenvalues of (A, B) and (A', B'), respectively, this guarantees

$$|\lambda - \mu| \geq \frac{\epsilon}{||B||_2} + \frac{\epsilon - \sqrt{2}\eta}{||B'||_2} = \frac{\epsilon||B'||_2 + (\epsilon - \sqrt{2}\eta)||B||_2}{||B||_2||B'||_2}. \tag{5.27}$$

Consequently, we have

$$\chi(\lambda, \mu) = \frac{|\lambda - \mu|}{\sqrt{1 + |\lambda|^2}\sqrt{1 + |\mu|^2}} \geq \frac{\epsilon||B'||_2 + (\epsilon - \sqrt{2}\eta)||B||_2}{||B||_2||B'||_2(1 + r^2)}, \tag{5.28}$$

where we note that $|\lambda|, |\mu| \leq r$. Since the criteria on η implies

$$\frac{\sqrt{||A - A'||_2^2 + ||B - B'||_2^2}}{\min_{\lambda' \neq \mu \in \Lambda(A', B')} \chi(\lambda, \mu)} \leq \frac{\sqrt{2}\eta||B||_2||B'||_2(1 + r^2)}{\epsilon||B'||_2 + (\epsilon - \sqrt{2}\eta)||B||_2} < \gamma(A', B'), \tag{5.29}$$

the result follows from Theorem 5.1.5. □

5.3 Structure-Preserving Divide-and-Conquer

Given structured pseudospectral shattering, we consider next structured divide-and-conquer. Here, "structure-preserving" refers to definiteness; we seek a version of divide-and-conquer (specifically a version of **EIG**) that splits one definite pencil into two smaller ones that are also definite. While we saw above that pseudospectral shattering could accommodate banded structure, the same cannot be said for divide-and-conquer; accordingly, we set that aside for the remainder of the chapter.

Algorithmically, we can guarantee that each subproblem in divide-and-conquer is definite by replacing the right and left matrices U_R and U_L with a single matrix $U \in \mathbb{C}^{n \times k}$. If U contains an orthonormal basis for a right deflating subspace of (A, B) – i.e., is U_R from the general setting – the $k \times k$ pencil $(U^H AU, U^H BU)$ is definite. Indeed, we have

$$\begin{aligned} \gamma(U^H AU, U^H BU) &= \min_{||x||_2 = 1} |x^H(U^H AU + iU^H BU)x| \\ &= \min_{||x||_2 = 1} |(Ux)^H(A + iB)Ux| \\ &= \min_{\substack{||y||_2 = 1 \\ y \in \text{range}(U)}} |y^H(A + iB)y| \\ &\geq \gamma(A, B). \end{aligned} \tag{5.30}$$

Moreover, the eigenpairs of (U^HAU, U^HBU) are exactly (λ, v) for (λ, Uv) a corresponding eigenpair of (A, B). Note that (5.30) holds for any matrix with orthonormal columns, and in particular does not depend on U containing a basis for a deflating subspace.

While modifying divide-and-conquer in this way appears straightforward, there is one additional roadblock to consider: in this case, we cannot guarantee a clean pseudospectral bound like $\Lambda_\epsilon(U_L^HAU_R, U_L^HBU_R) \subseteq \Lambda_\epsilon(A, B)$ for $\Lambda_\epsilon^{\text{sym}}(A, B)$. The proof of that result (see Lemma 1.3.5) relied critically on the fact that the columns of U_L spanned a left deflating subspace of (A, B). Since again the right and left deflating subspaces of a definite pencil are not the same in general, the argument made there will not carry over. Instead, we have the following.

Lemma 5.3.1. *Let (A, B) be an $n \times n$ definite pencil and suppose $U \in \mathbb{C}^{n\times k}$ contains an orthonormal basis for a right deflating subspace of (A, B). For any $\epsilon > 0$ and*

$$\epsilon' = \epsilon \left(\frac{\gamma(A, B)}{||(A, B)||_2} \right)^2$$

we have $\Lambda_{\epsilon'}^{\text{sym}}(U^HAU, U^HBU) \subseteq \Lambda_\epsilon^{\text{sym}}(A, B)$.

Proof. Suppose $z \in \Lambda_{\epsilon'}^{\text{sym}}(U^HAU, U^HBU)$. In this case, there exist Hermitian matrices $E, F \in \mathbb{C}^{k\times k}$ with $\sqrt{||E||_2^2 + ||F||_2^2} \le \epsilon'$ such that $z \in \Lambda(U^HAU + E, U^HBU + F)$. If $v \in \mathbb{C}^k$ is a corresponding right eigenvector, then by definition

$$(U^HAU + E)v = z(U^HBU + F)v. \tag{5.31}$$

Consider now X – the right eigenvector matrix of (A, B) satisfying (5.2). Without loss of generality, we may assume that the columns of U span the right deflating subspace corresponding to the first k columns of X. Writing $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ for $X_1 \in \mathbb{C}^{n\times k}$ and $X_2 \in \mathbb{C}^{n\times n-k}$, this implies $X_1 = UR_1$ for invertible $R_1 \in \mathbb{C}^{k\times k}$. Storing an orthonormal basis for the range of X_2 in $W \in \mathbb{C}^{n\times n-k}$, we obtain a block factorization

$$X = \begin{pmatrix} UR_1 & WR_2 \end{pmatrix} = \begin{pmatrix} U & W \end{pmatrix} \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix} \tag{5.32}$$

for another invertible $R_2 \in \mathbb{C}^{n-k \times n-k}$.

With this in mind, let $Q = \begin{pmatrix} U & W \end{pmatrix}$. Since the orthogonal complement of the right deflating subspace corresponding to one of U and W is the left deflating subspace associated to the other, we have

$$Q^H A Q = \begin{pmatrix} U^H A U & 0 \\ 0 & W^H A W \end{pmatrix} \quad \text{and} \quad Q^H B Q = \begin{pmatrix} U^H B U & 0 \\ 0 & W^H B W \end{pmatrix}. \tag{5.33}$$

Hence, it is easy to see

$$\left(Q^H A Q + \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} - z \left[Q^H B Q + \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix} \right] \right) \begin{pmatrix} v \\ 0 \end{pmatrix} = 0 \tag{5.34}$$

or equivalently, noting that Q is invertible but not necessarily unitary,

$$Q^H \left(A + Q^{-H} \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} - z \left[B + Q^{-H} \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \right] \right) Q \begin{pmatrix} v \\ 0 \end{pmatrix} = 0. \tag{5.35}$$

In other words, z is an eigenvalue of $\left(A + Q^{-H} \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}, B + Q^{-H} \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \right)$ corresponding to eigenvector $Q \begin{pmatrix} v \\ 0 \end{pmatrix}$.

We complete the proof by bounding the norms of the matrices $Q^{-H} \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}$ and $Q^{-H} \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}$. To do this, note that $Q^{-1} = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix} X^{-1}$ and therefore

$$||Q^{-1}||_2 \leq ||X^{-1}||_2 \max\{||R_1||_2, ||R_2||_2\} \leq \kappa_2(X). \tag{5.36}$$

The latter inequality follows from the observation $||R_1||_2 = ||X_1||_2$ and $||R_2||_2 = ||X_2||_2$, which implies that both $||R_1||_2$ and $||R_2||_2$ are at most $||X||_2$. Thus, we have

$$\begin{aligned} \left|\left| Q^{-H} \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \right|\right|_2 &\leq \kappa_2(X)^2 ||E||_2 \\ \left|\left| Q^{-H} \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \right|\right|_2 &\leq \kappa_2(X)^2 ||F||_2. \end{aligned} \tag{5.37}$$

Recalling Lemma 5.1.2 and our definition of the symmetric pseudospectrum, we conclude $z \in \Lambda_\epsilon^{\text{sym}}(A, B)$ provided $\epsilon' = \frac{\epsilon}{\kappa_2(X)^2} \leq \epsilon \left(\frac{\gamma(A,B)}{||(A,B)||_2} \right)^2$. □

As we will see, Lemma 5.3.1 suggests that a structured version of divide-and-conquer may not be more efficient than the original if $\frac{\gamma(A,B)}{||(A,B)||_2}$ is small (say, inverse-polynomial in n). For the purposes of constructing an analog of **EIG**, we will assume access to a lower bound on this quantity – and therefore a bound on ϵ' in Lemma 5.3.1 – as a black box.

5.3.1 Dynamically Weighted Halley Iteration

If each subproblem in divide-and-conquer is definite, we are guaranteed that the spectrum remains constrained to the real axis as we recur (in exact arithmetic). With this in mind, we can revisit the Indicator Approximation Problem. In particular, we consider the dynamically weighted Halley iteration mentioned in Chapter 3, which approximates the sign function via $f_k \circ f_{k-1} \circ \cdots \circ f_0$ for

$$f_i(x) = x\frac{a_i x^2 + b_i}{c_i x^2 + d_i}. \tag{5.38}$$

Recall that we can apply this iteration to (A, B) via the inverse-free arithmetic of Chapter 3 as follows:

$$(B_{i+1}\backslash A_{i+1}) = f_i(B_i\backslash A_i); \quad (B_0\backslash A_0) = (B\backslash A). \tag{5.39}$$

At each step, the eigenvalues of (A_i, B_i) are mapped according to f_i.

Suppose that at the i-th step of this process we have $\Lambda(A_i, B_i) \subset [-1, -l_i] \cup [l_i, 1]$ for some $l_i > 0$.[7] Since any iteration hoping to approximate the sign function must drive eigenvalues to ±1, our goal will be to choose a_i, b_i, c_i, d_i so that

$$f_i : [-1, -l_i] \cup [l_i, 1] \;\mapsto\; [-1, -l_{i+1}] \cup [l_{i+1}, 1] \tag{5.40}$$

for $l_i \leq l_{i+1} \leq 1$, with l_{i+1} as close to one as possible. While this is a nontrivial optimization problem in general, a few key observations make things more manageable. First, we can assume without loss of generality that $d_i = 1$. To ensure that f_i fixes ±1, we can next choose $c_i = a_i + b_i - 1$. With these in place, we have $f_i(x) = x\frac{a_i x^2 + b_i}{(a_i + b_i - 1)x^2 + 1}$ with only a_i and

[7] We assume here that zero is not an eigenvalue of (A_i, B_i), as in that case the sign function is not defined.

b_i left to optimize. Noting that f_i will be odd if $a_i, b_i, c_i > 0$, which guarantees that $[-1, -l_i]$ is mapped to $[-1, -l_{i+1}]$ provided $[l_i, 1] \mapsto [l_{i+1}, 1]$, and setting $l_{i+1} = \min_{l_i \leq x \leq 1} f_i(x)$, we obtain a final optimization problem

$$\underset{a_i, b_i}{\text{maximize}}\ l_{i+1} \quad \text{subject to} \quad a_i, b_i > 0 \ \text{ and } \ a_i + b_i > 1. \tag{5.41}$$

Note that forcing f_i to fix ±1 guarantees $l_{i+1} \leq 1$ for all i.

A solution to (5.41) was rigorously derived in work of Nakatsukasa, Bai and Gygi [105, Appendix A], wherein

$$\begin{aligned} b_i &= \sqrt{1+\gamma_i} + \frac{1}{2}\sqrt{8 - 4\gamma_i + \frac{8(2-l_i^2)}{l_i^2\sqrt{1+\gamma_i}}} \quad \text{for } \ \gamma_i = \sqrt[3]{\frac{4(1-l_i^2)}{l_i^4}}, \\ a_i &= \frac{1}{4}(b_i - 1)^2, \\ l_{i+1} &= f_i(l_i) = l_i \frac{a_i l_i^2 + b_i}{(a_i + b_i - 1)l_i^2 + 1}. \end{aligned} \tag{5.42}$$

For this choice of a_i and b_i we have $(a_i, b_i) \to (1, 3)$ as $l_i \to 1$, meaning the corresponding weighted Halley iteration gradually approaches the standard version.

Applying (5.42) to Algorithm 3 produces an inverse-free dynamically weighted Halley iteration (**IF-DWH**), which we present here as Algorithm 9. Note that this routine requires not only that $\Lambda(A, B)$ is contained in a symmetric union of intervals in $[-1, 1]$ but that a lower bound l_0 on the minimum eigenvalue (in magnitude) is known.

Nakatsukasa, Bai, and Gygi originally stated this version of the Halley iteration as part of an algorithm for computing the polar decomposition of a matrix, which was subsequently deployed by Nakatsukasa and Higham in a divide-and-conquer algorithm for the symmetric eigenvalue problem [107]. While the optimized weights presented above were derived in [105] via a direct and exhaustive search, a connection to the work of Zolotarev was later made by Nakatsukasa and Freund [106], who demonstrated that the rational function corresponding to (5.42) can be interpreted as an optimal approximation[8]

[8] Optimal meaning the best (in the infinity norm) rational function approximation $p(x)/q(x)$ for $p(x)$ and $q(x)$ real polynomials of degree three and two, respectively.

Algorithm 9. Inverse-Free Dynamically Weighted Halley Iteration (**IF-DWH**)
Input: $A, B \in \mathbb{C}^{n\times n}$, k a number of iterations, $l_0 > 0$.
Requires: Eigenvalues λ of (A, B) are real with $l_0 < |\lambda| \leq 1$.

```
$A_0 = A$
$B_0 = B$
for $i = 0 : k-1$ do
    $\gamma_i = (4(1-l_i^2)/l_i^4)^{1/3}$
    $b_i = \sqrt{1+\gamma_i} + \frac{1}{2}\sqrt{8 - 4\gamma_i + 8(2-l_i^2)/(l_i^2\sqrt{1+\gamma_i})}$
    $a_i = \frac{1}{4}(b_i - 1)^2$
    $c_i = a_i + b_i - 1$
    $\begin{pmatrix} -B_i \\ A_i \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} R_i \\ 0 \end{pmatrix}$        ▷ Apply Halley iteration
    $C_i = a_i Q_{12}^H A_i + b_i Q_{22}^H B_i$
    $D_i = c_i Q_{12}^H A_i + Q_{22}^H B_i$
    $\begin{pmatrix} -D_i \\ A_i \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \begin{pmatrix} \widehat{R}_i \\ 0 \end{pmatrix}$
    $A_{i+1} = U_{12}^H C_i$
    $B_{i+1} = U_{22}^H B_i$
    $l_{i+1} = l_i(a_i l_i^2 + b_i)/(c_i l_i^2 + 1)$        ▷ Compute next value of $l$
end for
return $(A_k, B_k)$, optionally $l_k$
```

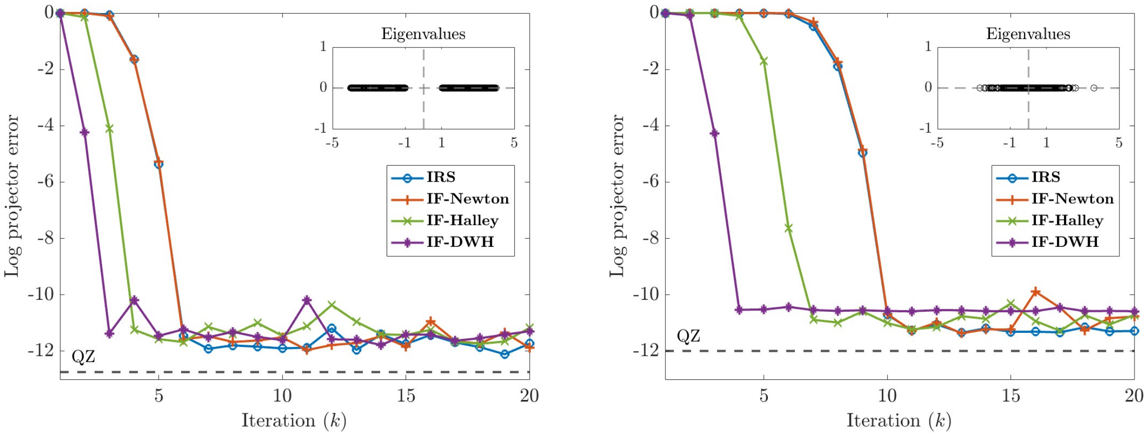
to the sign function on $[-1, -l_i] \cup [l_i, 1]$.

A key question remains: how fast does the weighted Halley iteration converge? We consider first a motivating example.

Example 5.3.2. Construct the 500×500 pencil (A, B) as follows: let B be a random complex Gaussian matrix and set $A = BVDV^H$ for

$$D = \begin{pmatrix} D_+ & 0 \\ 0 & D_- \end{pmatrix} \qquad V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}, \tag{5.43}$$

where V is Haar unitary (with $V_1, V_2 \in \mathbb{C}^{500\times 250}$) and $D_+, D_- \in \mathbb{C}^{250\times 250}$ are diagonal, with nonzero entries sampled from $\mathbb{R}_{>0}$ and $\mathbb{R}_{<0}$, respectively. Note that (A, B) is not necessarily definite. Suppose we are interested in computing the projector P onto the right deflating subspace of (A, B) corresponding to the right half plane, which by construction is given by $P = V_1 V_1^H$. To obtain this projector, we can use any of the iterative methods considered in this thesis – i.e., **IRS**, **IF-Newton**, **IF-Halley**, and now **IF-DWH**.



(a) Well-separated

(b) Poorly-separated

Figure 5.1. Approximation error for different iterative methods of computing spectral projectors. In both figures we approximate a projector of a 500×500 pencil (A, B) constructed according to (5.43). We consider two cases, where the eigenvalues of (A, B) are either well-separated or poorly-separated from the imaginary axis. In each case, we plot the eigenvalues of (A, B) and mark the error produced by QZ – which approximates the projector by computing first a full eigendecomposition and then a QR factorization of corresponding eigenvectors – as a benchmark.

At a given iteration, each of these methods produces a pencil (A_k, B_k) with eigenvalues close to zero or one. An approximate projector can therefore be obtained as $\widetilde{P} = UU^H$ for U a matrix containing the first 250 columns of the U-factor produced by **GRURV** when applied to $\frac{1}{2}B_k^{-1}(A_k + B_k)$ or – in the case of **IRS** – $(A_k + B_k)^{-1}A_k$. Error in this approximate projector can then be measured as $\log_{10}(||P - \widetilde{P}||_2)$. Note that in this approach, **IRS** must apply an initial Möbius transformation mapping the imaginary axis to the unit circle.

Figure 5.1 plots projector error for each method in two cases: one in which the eigenvalues of (A, B) are well-separated from the imaginary axis and one in which they are not. As we might expect, all of the methods require more iterations to converge when eigenvalues are close to the imaginary axis. Regardless, **IF-DWH** is consistently the fastest method. Notably, it requires fewer than half the iterations of **IRS** or **IF-Newton** in both cases, meaning it will be more efficient than either despite requiring (up to) twice as many $2n \times n$ QR factorizations and $n \times n$ matrix multiplications (see Table 3.2). Comparing to **IF-Halley**, which only cuts the number of iterations by roughly $\log_2(3)$, it is clear that dynamic weighting is necessary to outperform these second-order methods.

To run **IF-DWH** here we scale A by $\frac{1}{||B^{-1}A||_2}$, which both drives eigenvalues inside $[-1, 1]$ and implies $l_0 \geq \kappa_2(B^{-1}A)$. This is the only potential drawback to **IF-DWH**, as it may result in eigenvalues significantly closer to the imaginary axis if $||B^{-1}A||_2$ is large. As mentioned above (A, B) is almost certainly not definite in this example; this was done deliberately, aimed at demonstrating that **IF-DWH** is relevant for *any* pencil with real spectrum. Echoing Chapter 3, it also suggests that pursuing weighting schemes for other specialized spectra may be worthwhile in general.

With this example as a backdrop, we consider now theoretical convergence results for **IF-DWH**. Given the complex nature of the weighting scheme, obtaining tight bounds on the number of iterations required to achieve a certain accuracy is difficult. Instead, we recommend keeping track of the parameter l_k, which implies the following straightforward error bound. Note here that the assumption that $\Lambda(A, B)$ is bounded guarantees that B (and later B_k) is invertible.

Lemma 5.3.3. *Suppose (A, B) is a definite pencil with eigenvalues in $(-1, -l_0) \cup (l_0, 1)$ and let $[A_k, B_k, l_k] =$ **IF-DWH**(A, B, k, l_0). Then*

$$||B_k^{-1}A_k - \text{sign}(B^{-1}A)||_2 \leq \frac{||(A, B)||_2}{\gamma(A, B)}(1 - l_k).$$

Proof. Let X be the invertible eigenvector matrix of (A, B) satisfying (5.2). Since (A_k, B_k) has the same right eigenvectors as (A, B), X diagonalizes $B_k^{-1}A_k$. Writing $B_k^{-1}A_k = X\Lambda_k X^{-1}$ for Λ_k diagonal and noting $B^{-1}A = X\Lambda_B^{-1}X^H X^{-H}\Lambda_A X^{-1} = X\Lambda_B^{-1}\Lambda_A X^{-1}$, we have

$$\begin{aligned} ||B_k^{-1}A_k - \text{sign}(B^{-1}A)||_2 &= ||X\Lambda_k X^{-1} - X\text{sign}(\Lambda_B^{-1}\Lambda_A)X^{-1}||_2 \\ &\leq \kappa_2(X)||\Lambda_k - \text{sign}(\Lambda_B^{-1}\Lambda_A)||_2 \qquad (5.44) \\ &\leq \kappa_2(X)(1 - l_k), \end{aligned}$$

Table 5.1. Evolution of l_k in **IF-DWH** for six initial bounds $\Lambda(A, B) \subseteq (-1, -l_0) \cup (l_0, 1)$. The entry -Inf indicates that l_k is indistinguishable from one in double precision.

	Initial lower bound (l_0)					
	10^{-1}	10^{-3}	10^{-5}	10^{-7}	10^{-9}	10^{-11}
$\log_{10}(1 - l_1)$	-0.872	-0.124	-0.024	-0.005	-0.001	-0.0002
$\log_{10}(1 - l_2)$	-4.329	-1.431	-0.646	-0.326	-0.176	-0.099
$\log_{10}(1 - l_3)$	-14.841	-6.074	-3.583	-2.404	-1.718	-1.275
$\log_{10}(1 - l_4)$	-15.654	-15.654	-12.555	-9.014	-6.949	-5.595
$\log_{10}(1 - l_5)$	-Inf	-Inf	-Inf	-15.955	-15.955	-15.654

where the last inequality follows from $\Lambda(A_k, B_k) \subseteq (-1, -l_k) \cup (l_k, 1)$. We complete the proof by bounding $\kappa_2(X)$ via Lemma 5.1.2. $\square$

To build intuition, Table 5.1 records l_k over five iterations of **IF-DWH** for a handful of starting values l_0. We see here the power of the weighting scheme, which consistently drives l_k to one in only a handful of steps.

While this is promising, we would still like to have at least a rough upper bound on the number of iterations required by **IF-DWH**. Fortunately, this can be done by using **IF-Halley** as a baseline. That is, since the standard Halley iteration coefficients are included in the search space of (5.41), we can expect **IF-DWH** to converge at least as fast as **IF-Halley**. With this in mind, we state a similar error bound for the latter. In this case, we drop the requirement that (A, B) has eigenvalues in $[-1, 1]$ since it is not necessary to run the standard Halley iteration.

Lemma 5.3.4. *Suppose (A, B) is a definite pencil with eigenvalues in $(-R, -r) \cup (r, R)$ for some $0 < r < R$ with $r < \frac{1}{R}$. If $[A_k, B_k] =$ **IF-Halley**(A, B, k) then*

$$||B_k^{-1}A_k - \text{sign}(B^{-1}A)||_2 \leq \frac{||(A, B)||_2}{\gamma(A, B)} \left(\frac{2\alpha^{3^k}}{1 - \alpha^{3^k}} \right)$$

for $\alpha = \frac{1-r}{1+r}$.

Proof. This result can be obtained by repeating the proof Lemma 5.3.3. In this case, we have $\Lambda(A,B) \subseteq C_\alpha$ for $\alpha = \frac{1-r}{1+r}$ and therefore, recalling Lemma 3.3.6, $\Lambda(A_k,B_k) \subset C_{\alpha^{3^k}}$. The bound then follows by noting that the maximum distance between $+1$ and any point in $C_\alpha^+ \cap \mathbb{R}$ (equivalently -1 and $C_\alpha^- \cap \mathbb{R}$) is $\frac{2\alpha}{1-\alpha}$. □

Remark 5.3.5. The decision to use **IF-DWH** here represents yet another way of approaching the Indicator Approximation Problem, which may be solved by choosing a method that performs optimally on specific subsets of S and $\mathbb{C} - S$ (in this case intervals on the real axis).

5.3.2 Diagonalization in Exact Arithmetic

We are now ready to state our specialized version of divide-and-conquer. The resulting **EIG-DWH** (presented below as Algorithm 10) can be interpreted as a variant of **EIG** that both enforces definiteness and leverages the efficiency gains of **IF-DWH**. Like the original, it assumes access to a guarantee of pseudospectral shattering for a corresponding set of grid points $g = \{g_i\} \subset \mathbb{R}$. Unlike **EIG**, however, it is somewhat divorced from the specific pseudospectral shattering result presented in the previous section. In particular, we do not bake in the norm assumptions or grid specifications made in Theorem 5.2.11. Instead, we take as inputs lower and upper bounds $\gamma(A,B) > \gamma$ and $||(A,B)||_2 \leq c$ and assume $g \subset (-r,r)$ for some $r > 0$. Our motivation for doing this is rooted primarily in the fact that general bounds available for $\gamma(\widetilde{A}, n^\alpha \widetilde{B})$ are unsatisfactory, potentially forcing ϵ to become prohibitively small as we recur via Lemma 5.3.1, which we discuss in more detail below.

Before that, we consider an analog of Theorem 4.1.4 – i.e., a guarantee that **EIG-DWH** performs as expected for the listed parameters. Note here that, unlike **EIG**, **EIG-DWH** does not require a bound on $\sigma_n(B)$ to produce an eigenvector guarantee (though the assumption that $\Lambda_\epsilon^{\mathrm{sym}}(A,B)$ is bounded implies that B must be invertible).

Algorithm 10. Definite Divide-and-Conquer Eigensolver (**EIG-DWH**)
Input: $n \in \mathbb{N}_+$, $A, B \in \mathbb{C}^{m \times m}$, $\epsilon, \gamma, c, r > 0$, $g = \{g_i\} \subset \mathbb{R}$ a grid of points with equal spacing ω, $\beta \in (0,1)$ a desired eigenvector accuracy, and $\theta \in (0,1)$ a failure probability.
Requires: $m \leq n$, (A, B) definite with $\gamma(A,B) \geq \gamma$ and $||(A,B)||_2 \leq c$, $g \subset (-r, r)$, and $\Lambda_\epsilon^{\text{sym}}(A,B)$ shattered with respect to g.
Output: X an invertible matrix and (Λ_A, Λ_B) a diagonal pencil. The eigenvalues of (Λ_A, Λ_B) each share an interval of g with a unique eigenvalue of (A, B) and each column of X is an approximate right unit eigenvector of (A, B).

```
1:  **if** $m = 1$ **then**
2:      $X = 1$; $\Lambda_A = A$; $\Lambda_B = B$
3:  **else**
4:      $\zeta = \lfloor \log_2 \lceil 2r/\omega \rceil + 1 \rfloor$
5:      $\delta = \min\left\{ \frac{4\theta}{4\theta + 3\zeta n^2(n-1)}, \sqrt{\frac{\theta}{3(n-1)}} \frac{\epsilon^2\gamma^4}{800nc^6}, \frac{1}{16nc^2}\sqrt{\frac{\theta}{3(n-1)}} \left( \frac{\sqrt{2}\beta\epsilon\gamma^3}{c^2[c(1+r^2)+\beta\gamma]} \right)^2 \right\}.$
6:      Choose a grid point $g_i \in g$
7:      $(\mathcal{A}, \mathcal{B}) = (A - g_i B, 2rB)$; $l = \frac{\epsilon}{2rc}$
8:      **while** $l < 1 - \frac{2\delta\gamma}{c}$ **do**
9:          $[\mathcal{A}, \mathcal{B}, l] =$ **IF-DWH**$(\mathcal{A}, \mathcal{B}, 1, l)$
10:     **end while**
11:     $[U, R_1, R_2, V] =$ **GRURV**$(2, 2\mathcal{B}, \mathcal{A} + \mathcal{B}, -1, 1)$
12:     $r = \# \left\{ i : \left| \frac{R_2(i,i)}{R_1(i,i)} \right| \geq 2\sqrt{\frac{\theta}{3\zeta(n-1)}} \frac{1-\delta}{n} \right\}$
13:     **if** $r < \lfloor \frac{m}{2} \rfloor$ or $r > \lceil \frac{m}{2} \rceil$ **then**
14:         Return to line 6, executing a binary search over the grid points if necessary.
15:     **else**
16:         $U =$ **GRURV**$(2, 2\mathcal{B}, \mathcal{A} + \mathcal{B}, -1, 1)$
17:         $U_r = U( : , 1 : r)$
18:         $U =$ **GRURV**$(2, 2\mathcal{B}, \mathcal{A} - \mathcal{B}, -1, 1)$
19:         $U_{m-r} = U( : , 1 : m - r)$
20:         $$(A_{11}, B_{11}) = (U_r^H A U_r,\ U_r^H B U_r)$$
            $$(A_{22}, B_{22}) = (U_{m-r}^H A U_{m-r},\ U_{m-r}^H B U_{m-r})$$
21:         $g_R = \{z \in g : z > g_i\}$;   $g_L = \{z \in g : z < g_i\}$
22:         $[\widehat{X}, \widehat{\Lambda}_A, \widehat{\Lambda}_B] =$ **EIG-DWH**$(n, A_{11}, B_{11}, \frac{4\epsilon\gamma^2}{5c^2}, \gamma, c, r, g_R, \frac{1}{3}\beta, \theta)$
23:         $[\widetilde{X}, \widetilde{\Lambda}_A, \widetilde{\Lambda}_B] =$ **EIG-DWH**$(n, A_{22}, B_{22}, \frac{4\epsilon\gamma^2}{5c^2}, \gamma, c, r, g_L, \frac{1}{3}\beta, \theta)$
24:         $$X = \begin{pmatrix} U_r\widehat{X} & 0 \\ 0 & U_{m-r}\widetilde{X} \end{pmatrix},\ \Lambda_A = \begin{pmatrix} \widehat{\Lambda}_A & 0 \\ 0 & \widetilde{\Lambda}_A \end{pmatrix},\ \Lambda_B = \begin{pmatrix} \widehat{\Lambda}_B & 0 \\ 0 & \widetilde{\Lambda}_B \end{pmatrix}$$
25:     **end if**
26: **end if**
27: **return** $X, \Lambda_A, \Lambda_B$
```

Proposition 5.3.6. *Let (A, B) and $g = \{g_i\} \subset \mathbb{R}$ be a definite pencil and set of grid points satisfying the requirements of* **EIG-DWH**. *For any choice of $\theta, \beta \in (0, 1)$, exact-arithmetic* **EIG-DWH** *applied to (A, B) and g satisfies the following with probability at least $1 - \theta$.*

1. *The recursive procedure converges and each eigenvalue of the diagonal pencil (Λ_A, Λ_B) shares a grid interval (g_i, g_{i+1}) with a unique eigenvalue of (A, B).*
2. *Each column x_i of X satisfies $||x_i - v_i||_2 \leq \beta$ for a right unit eigenvector v_i of (A, B).*

Proposition 5.3.6 can be obtained by modifying the proof of Theorem 4.1.4 accordingly. Rather than repeating this proof in detail, we provide a sketch, noting below the adjustments that need to be made:

1. Once again, success for **EIG-DWH** is predicated on the validity of its recursive calls. In this case, there is one additional requirement to check – i.e., that (A_{11}, B_{11}) and (A_{22}, B_{22}) are definite with $\gamma(A_{11}, B_{11}), \gamma(A_{22}, B_{22}) \geq \gamma$. This follows from (5.30).

2. Since $\Lambda_\epsilon^{\text{sym}}(A, B)$ contains the interval of radius $\frac{\epsilon}{||B||_2} \geq \frac{\epsilon}{c}$ around each eigenvalue (see Theorem 5.2.2), the shattering guarantee $\Lambda_\epsilon^{\text{sym}}(A, B) \cap g = \emptyset$ implies that the eigenvalues of $(A - g_i B, B)$ are contained in $(-2r, -\frac{\epsilon}{c}) \cup (\frac{\epsilon}{c}, 2r)$ for any grid point $g_i \in g$. Each pencil $(\mathcal{A}, \mathcal{B})$ input to **IF-DWH** therefore satisfies $\Lambda(\mathcal{A}, \mathcal{B}) \subseteq (-1, -\frac{\epsilon}{2rc}) \cup (\frac{\epsilon}{2rc}, 1)$. Hence the initial choice of l in line 8.

3. In line 9, **IF-DWH** is employed to compute the projector
$$P_{>g_i} = \frac{1}{2}(\text{sign}(B^{-1}A - g_i I) + I). \tag{5.45}$$
To guarantee $||\frac{1}{2}\mathcal{B}^{-1}(\mathcal{A} + \mathcal{B}) - P_{>g_i}||_2 \leq \delta$ upon exit in line 10, we need $||\mathcal{B}^{-1}\mathcal{A} - \text{sign}(B^{-1}A - g_i I)||_2 \leq 2\delta$. Noting
$$\text{sign}\left(\frac{1}{2r}B^{-1}(A - g_i B)\right) = \text{sign}(B^{-1}(A - g_i B)) = \text{sign}(B^{-1}A - g_i I), \tag{5.46}$$
Lemma 5.3.3 implies that we should run **IF-DWH** until $l \geq 1 - \frac{2\delta\gamma}{c}$. Note that the number of iterations this requires can be loosely upper bounded via Lemma 5.3.4.

4. Because the spectrum is constrained to the real axis, an optimal split that divides it into disjoint sets of size $\lfloor \frac{m}{2} \rfloor$ and $\lceil \frac{m}{2} \rceil$ always exists. Ensuring optimality at each split guarantees that **EIG-DWH** is called exactly $n-1$ times on problems of size $m > 1$.

5. For any $\nu_1 \in (0,1)$ suppose $\delta < \frac{4\nu_1}{4\nu_1^2 + n^2}$ and
$$r = \#\left\{ i : \frac{R_2(i,i)}{R_1(i,i)} \geq \frac{2\nu_1(1-\delta)}{n} \right\}. \tag{5.47}$$
in line 12. In this case, requiring $r = \text{rank}(P_{>g_i})$ for every grid point checked, a given step of **EIG-DWH** finds an optimal split with probability at least $1 - \zeta\nu_1^2$ (see step four in the proof of Theorem 4.1.4).

6. Once a split is found, lines 16-19 compute orthonormal bases for the corresponding right deflating subspaces, replacing the calls to **DEFLATE** in **EIG**. Accordingly, a straightforward replication of Theorem 4.1.3 implies that for $\nu_2 \in (0,1)$ the matrices U_r and U_{m-r} are computed (independently) to within spectral norm error $2\sqrt{\frac{n\delta}{\nu_2}}$ with probability at least $1 - \nu_2^2$.

7. For an appropriate choice of δ, avoiding failure in items 4 and 5 above will guarantee success for one step of divide-and-conquer. Hence, we set $\zeta\nu_1^2 = \nu_2^2 = \frac{\theta}{3(n-1)}$, thereby guaranteeing that a simple union bound will imply a total failure probability for **EIG-DWH** of at most θ.

8. It remains to find the aforementioned choice of δ. Suppose $U \in \mathbb{C}^{m \times r}$ is the true matrix approximated by U_r – i.e., satisfying $||U_r - U||_2 \leq 2\sqrt{\frac{n\delta}{\nu_2}}$. By Lemma 5.3.1, $\Lambda_{\frac{\epsilon\gamma^2}{c^2}}(U^H AU, U^H BU)$ is shattered with respect to g. Hence, recalling Lemma 5.2.13, we maintain shattering for the subsequent calls to **EIG-DWH** by requiring
$$||A_{11} - U^H AU||_2, ||B_{11} - U^H BU||_2 \leq \frac{\epsilon\gamma^2}{5\sqrt{2}c^2}. \tag{5.48}$$
Noting
$$||A_{11} - U^H AU||_2 \leq 2||U_r - U||_2||A||_2 \leq 4c\sqrt{\frac{n\delta}{\nu_2}}, \tag{5.49}$$

it is sufficient to take $4c\sqrt{\frac{n\delta}{\nu_2}} \leq \frac{\epsilon\gamma^2}{5\sqrt{2}c^2}$ or equivalently $\delta \leq \sqrt{\frac{\theta}{3(n-1)}}\frac{\epsilon^2\gamma^4}{800nc^6}$, applying our choice of ν_2.

9. The final requirement on δ follows from Lemma 5.2.14, where we enforce that $||A_{11} - U^HAU||_2, ||B_{11} - U^HBU||_2$ is small enough to guarantee eigenvector error is at most β. Note that, unlike in **EIG**, this does not require a bound on $\sigma_n(B)$.

As in the general case, **EIG-DWH** implies a straightforward diagonalization algorithm. Given input matrices $A, B \in \mathbb{C}^{n\times n}$ with $||A||_2, ||B||_2 \leq 1$ and $\gamma(A, B)$ known, a variant of **RPD** proceeds as follows:

1. Construct $(\widetilde{A}, n^\alpha\widetilde{B})$ and $g = \{g_i\} \subset \mathbb{R}$ in accordance with Theorem 5.2.11.

2. Call $[X, \Lambda_A, \Lambda_B] =$ **EIG-DWH**$(n, \widetilde{A}, n^\alpha\widetilde{B}, \frac{\gamma^5}{10n^{4\alpha+5}}, \frac{1}{2}\gamma(A, B), 8n^\alpha, 4 + \omega, g, \beta, \theta)$ for an appropriate choice of $\beta, \theta \in (0, 1)$.

3. Undo the n^α scaling to construct a diagonalization.

Note that in this approach ϵ shrinks by a factor of $\frac{\gamma(A,B)^2}{320n^\alpha}$ at each step of **EIG-DWH**. Assuming $\gamma(A, B) = O(1)$ and recalling that the recursive depth of divide-and-conquer is $O(\log(n))$, this eventually implies $\epsilon = O(\frac{1}{\text{poly}(n)^{\log(n)}})$. This has the primary consequence of driving up the number of iterations required to compute the projectors $P_{>g_i}$, or at least driving up the upper bound available from Lemma 5.3.4. In the worst case, this may incur an additional log factor in complexity compared to the general case.

There are two loose points in the preceding analysis that could be tightened: Lemma 5.3.1 and the bound $\gamma(\widetilde{A}, n^\alpha\widetilde{B}) \geq \gamma(\widetilde{A}, \widetilde{B})$. In many cases, both are extremely relaxed; the latter, for example, is tight only when $\gamma(A, B)$ achieves its minimum at a unit vector x satisfying $x^HBx = 0$. Absent improvements, theoretical bounds for **EIG-DWH**, and any corresponding diagonalization algorithm, are not any better than the general results presented in Chapter 4. Nevertheless, this approach is likely to be much faster in practice, where theoretical requirements (like the n^{α} scaling or the decision to run divide-and-conquer to 1×1 subproblems) are likely unnecessary.

Chapter 6

Precision Bounds in Floating-Point Arithmetic

The preceding chapters presented pseudospectral divide-and-conquer in exact arithmetic. With this in mind, we close this thesis by considering an alternative floating-point setting, deriving precision bounds for the main building blocks of divide-and-conquer as stated in Chapter 4. Similar bounds for the specialization considered in Chapter 5 are left to future work. Throughout, we aim to both codify the stability of divide-and-conquer observed experimentally in Section 4.3 and further demonstrate the efficacy of working inverse-free.

Recall our model of floating-point arithmetic:

$$fl(x \circ y) = (x \circ y)(1 + \Delta), \ \ |\Delta| \leq \mathbf{u}. \tag{6.1}$$

Here, $\circ$ is any operation from the set $\{+, -, \times, \div\}$ and $\mathbf{u}$ is the corresponding machine precision. In this case, $\log_2(1/\mathbf{u})$ bits of precision are required to achieve (6.1). Our goal is to produce lower bounds on $\mathbf{u}$ – and therefore upper bounds on the number of bits

Table 6.1. Precision requirements for key steps of pseudospectral divide-and-conquer as performed by **RPD** and the corresponding algorithm of Banks et al. In each case, an $n \times n$ matrix/pencil is diagonalized with backward error ε. Note that both algorithms eventually set $\delta = \text{poly}(\varepsilon, n^{-1})$.

Bits of precision required to...	Banks et al. [16]	**RPD** (Algorithm 8)
Guarantee shattering	$O(\log(\frac{n}{\varepsilon}))$	$O(\log(\frac{n}{\varepsilon}))$
Compute input(s) of (**G**)**RURV** to spectral norm accuracy δ	$O(\log(n)\log^3(\frac{n}{\varepsilon})\log(\frac{n}{\delta\varepsilon}))$	$O(\log(\frac{n}{\delta\varepsilon}))$

of precision – required by a given building block of divide-and-conquer, which will be a function of the problem size n and the desired backward diagonalization accuracy ε.

We can a draw a comparison here to similar precision bounds in the single-matrix case of Banks et al. [16]. These are summarized in Table 6.1, where the corresponding results for **RPD** are derived in this chapter. The precision required by Banks et al. to compute the input of **RURV** – which in their case is an approximate projector obtained via the Newton iteration of the sign function – proves to be a bottleneck for their diagonalization algorithm overall, which they show requires $O(\log(n)\log^4(\frac{n}{\varepsilon}))$ bits of precision to diagonalize an $n \times n$ matrix with backward error ε. The corresponding step of **RPD** is **IRS**, which we demonstrate below requires much lower precision.

Of course, computing the inputs of **RURV** and **GRURV** is not equivalent, as in divide-and-conquer they correspond to approximating a spectral projector explicitly versus implicitly. While we have no reason to anticipate that accounting for this difference will incur an additional log factor in the number of bits, the gap between the bounds listed in Table 6.1 suggests that **RPD** is likely to be more stable in finite-precision arithmetic, even if such an increase turns out to be necessary. If it is not, then the precision required to compute spectral projectors is asymptotically the same as to guarantee shattering. Recalling that a diagonalization of the pencil (A, I) corresponds to a diagonalization of A, this ultimately implies that **RPD** may require lower precision to solve the standard eigenvalue problem without increasing asymptotic complexity.

Guide to Chapter Six: Section 6.1 presents the black-box, finite-precision arithmetic algorithms we assume access to, in particular (fast) matrix multiplication, QR, and Gaussian sampling. Sections 6.2 to 6.4 then derive bounds for pseudospectral shattering, **IRS**, and **GRURV** under these assumptions.

A note on notation: In this chapter, we use bold and capital letters to distinguish finite-arithmetic outputs. That is, we differentiate $\mathbf{MM}(A, B)$, the floating-point product of A and B, from its exact counterpart AB.

6.1 Black-Box Assumptions

We begin by stating the floating-point, black-box algorithms we make use of in the subsequent bounds. First up is an algorithm for Gaussian sampling, which will be used to generate the perturbation matrices G_1 and G_2 as well as the random matrices in **RURV**/**GRURV**. In Assumption 6.1.1 below, c_{N} is a constant.

Assumption 6.1.1 (Gaussian Sampling)**.** There exists a c_{N}-stable Gaussian sampler $N(\sigma)$ that takes $\sigma \in \mathbb{R}_{\geq 0}$ and outputs $N(\sigma)$ satisfying $|N(\sigma) - \mathcal{N}| \leq c_{\mathrm{N}}\sigma\mathbf{u}$ for some $\mathcal{N} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$.

Next are floating-point algorithms for matrix multiplication and full QR. Like our floating-point model (6.1), these are standard [74, Section 3.5 and Chapter 19]. Here, $\mu_{\mathrm{MM}}(n)$ and $\mu_{\mathrm{QR}}(m, n)$ are (small degree) polynomials in n and m, n, respectively.

Assumption 6.1.2 (Matrix Multiplication)**.** There exists a $\mu_{\mathrm{MM}}(n)$-stable $n \times n$ multiplication algorithm $\mathbf{MM}(\cdot, \cdot)$ satisfying

$$||\mathbf{MM}(A, B) - AB||_2 \leq \mu_{\mathrm{MM}}(n)\mathbf{u}||A||_2||B||_2$$

in $T_{\mathrm{MM}}(n)$ arithmetic operations.

Assumption 6.1.3 (QR Factorization)**.** There exists a $\mu_{\mathrm{QR}}(m, n)$-stable full QR algorithm $\mathbf{QR}(\cdot)$ satisfying

1. $[Q, R] = \mathbf{QR}(A)$ for $A, R \in \mathbb{C}^{m\times n}$ and $Q \in \mathbb{C}^{m\times m}$.
2. R is exactly upper triangular
3. There exist $A' \in \mathbb{C}^{m\times n}$ and unitary $Q' \in \mathbb{C}^{m\times m}$ such that $A' = Q'R$ with

$$||Q' - Q||_2 \leq \mu_{\text{QR}}(m,n)\mathbf{u} \;\text{ and }\; ||A' - A||_2 \leq \mu_{\text{QR}}(m,n)\mathbf{u}||A||_2,$$

in $T_{\text{QR}}(m,n)$ arithmetic operations.

Assumption 6.1.3 is written generally to accommodate the two different applications of QR in pseudospectral divide-and-conquer – i.e., the $2n \times n$ full factorizations computed by **IRS** and the square $n \times n$ version used by **RURV** and **GRURV**. For the latter, note that it also extends to floating-point QL/RQ algorithms $\mathbf{QL}(\cdot)$ and $\mathbf{RQ}(\cdot)$ (with the same parameters). In an effort to simplify bounds, we state results in terms of $\mu_{\text{QR}}(n) = \mu_{\text{QR}}(2n,n)$ and $T_{\text{QR}}(n) = T_{\text{QR}}(2n,n)$.[1] Throughout, we can always guarantee a truly triangular result from QR/QL in finite precision by forcing entries below/above the diagonal to be zero.

While we won't be too particular about the polynomials $\mu_{\text{MM}}(n)$ and $\mu_{\text{QR}}(n)$, we note that they are compatible with the fast linear algebra framework discussed in Section 1.4. That is, QR can be implemented stably (in a mixed sense) using fast matrix multiplication [38], which itself can be formulated to satisfy the forward error bound given by Assumption 6.1.2 [39]. Hence, the bounds presented in this chapter apply to floating-point implementations of fast pseudospectral divide-and-conquer, including a version built on the fastest known matrix multiplication algorithm of Williams et al. [146]. We may additionally assume $T_{\text{QR}}(n) = O(T_{\text{MM}}(n))$.

Finally, we state the logarithmically stable (fast) inversion algorithm assumed by Banks et al. [16, Definition 2.7], which they import from [38]. Since we do not use inversion in **RPD**, we include Assumption 6.1.4 here for comparison purposes. Once again, $\mu_{\text{INV}}(n)$ is a polynomial in n while c_{INV} is a constant.

[1] While this implies that results for **GRURV** can be tightened, the difference is ultimately insignificant.

Assumption 6.1.4 (Matrix Inversion)**.** There exists a $(\mu_{\text{INV}}(n), c_{\text{INV}})$-stable $n \times n$ inversion algorithm $\mathbf{INV}(\cdot)$ satisfying

$$||\mathbf{INV}(A) - A^{-1}||_2 \leq \mu_{\text{INV}}(n)\mathbf{u}\kappa_2(A)^{c_{\text{INV}}\log(n)}||A^{-1}||_2$$

in $T_{\text{INV}}(n)$ arithmetic operations.

Before moving to our analysis, we give a pair of technical lemmas, which are consequences of the black-box assumptions defined above (and more generally our model (6.1) of floating-point computations).

Lemma 6.1.5. *Let $A, B \in \mathbb{C}^{n\times n}$. If C is the floating-point sum of A and B then*

$$||C - (A+B)||_2 \leq \sqrt{n}\mathbf{u}||A+B||_2.$$

Proof. By (6.1) each entry C_{ij} of C satisfies $C_{ij} = (A+B)_{ij}(1+\Delta_{ij})$ for some $|\Delta_{ij}| \leq \mathbf{u}$. Consequently,

$$[C - (A+B)]_{ij} = (A+B)_{ij}\Delta_{ij} \tag{6.2}$$

and therefore

$$||C-(A+B)||_F^2 \leq \sum_{ij}|(A+B)_{ij}\Delta_{ij}|^2 \leq \mathbf{u}^2\sum_{ij}|(A+B)_{ij}|^2 = \mathbf{u}^2||A+B||_F^2. \tag{6.3}$$

We complete the proof by noting $||C-(A+B)||_2 \leq ||C-(A+B)||_F$ and $||A+B||_F \leq \sqrt{n}||A+B||_2$. □

Lemma 6.1.6. *Let $[Q, R] = \mathbf{QR}(A)$ for*

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}, \quad \textit{and} \quad R = \begin{pmatrix} R' \\ 0 \end{pmatrix},$$

where all blocks are $n \times n$. Define the following matrices:

1. *$\widetilde{R}$ – the floating-point sum of $\mathbf{MM}(Q_{11}^H, A_1)$ and $\mathbf{MM}(Q_{21}^H, A_2)$.*

2. *E – the floating-point sum of $\mathbf{MM}(Q_{12}^H, A_1)$ and $\mathbf{MM}(Q_{22}^H, A_2)$.*

If $\mu_{\mathrm{QR}}(n)\mathbf{u}, \mu_{\mathrm{MM}}(n)\mathbf{u} \leq 1$*, then*

$$||\widetilde{R} - R||_2, ||E||_2 \leq 4\left(\mu_{\mathrm{QR}}(n) + \mu_{\mathrm{MM}}(n) + 2\sqrt{n}\right)\mathbf{u}||A||_2.$$

Proof. By Assumption 6.1.3, there exist matrices $\widehat{A} \in \mathbb{C}^{2n\times n}$ and $\widehat{Q} \in \mathbb{C}^{2n\times 2n}$ such that $\widehat{Q}$ is truly unitary, $\widehat{A} = \widehat{Q}R$, $||Q - \widehat{Q}||_2 \leq \mu_{\mathrm{QR}}(n)\mathbf{u}$, and $||A - \widehat{A}||_2 \leq \mu_{\mathrm{QR}}(n)\mathbf{u}||A||_2$. Let

$$\widehat{A} = \begin{pmatrix} \widehat{A}_1 \\ \widehat{A}_2 \end{pmatrix} \text{ and } \widehat{Q} = \begin{pmatrix} \widehat{Q}_{11} & \widehat{Q}_{12} \\ \widehat{Q}_{21} & \widehat{Q}_{22} \end{pmatrix}, \tag{6.4}$$

where again all blocks are $n \times n$. Consider first $\mathbf{MM}(Q_{11}^H, A_1)$. Applying Assumption 6.1.2 we observe

$$\begin{aligned} ||\mathbf{MM}(Q_{11}^H, A_1) - \widehat{Q}_{11}^H\widehat{A}_1||_2 &\leq ||\mathbf{MM}(Q_{11}^H, A_1) - Q_{11}^H A_1||_2 + ||Q_{11}^H A_1 - \widehat{Q}_{11}^H\widehat{A}_1||_2 \\ &\leq \mu_{\mathrm{MM}}(n)\mathbf{u}||Q_{11}||_2||A_1||_2 + ||Q_{11}^H A_1 - \widehat{Q}_{11}^H A_1||_2 + ||\widehat{Q}_{11}^H A_1 - \widehat{Q}_{11}^H\widehat{A}_1||_2 \\ &\leq \mu_{\mathrm{MM}}(n)\mathbf{u}||Q||_2||A||_2 + ||Q - \widehat{Q}||_2||A||_2 + ||\widehat{Q}||_2||A - \widehat{A}||_2. \end{aligned} \tag{6.5}$$

Since $||\widehat{Q}||_2 = 1$ and therefore $||Q||_2 \leq ||\widehat{Q}||_2 + \mu_{\mathrm{QR}}(n)\mathbf{u} = 1 + \mu_{\mathrm{QR}}(n)\mathbf{u}$, (6.5) implies

$$\begin{aligned} ||\mathbf{MM}(Q_{11}^H, A_1) - \widehat{Q}_{11}^H\widehat{A}_1||_2 &\leq [2\mu_{\mathrm{QR}}(n) + \mu_{\mathrm{MM}}(n)(1 + \mu_{\mathrm{QR}}(n)\mathbf{u})]\,\mathbf{u}||A||_2 \\ &\leq 2(\mu_{\mathrm{QR}}(n) + \mu_{\mathrm{MM}}(n))\mathbf{u}||A||_2. \end{aligned} \tag{6.6}$$

Repeating this argument, swapping blocks accordingly, we obtain the same result for $||\mathbf{MM}(Q_{21}^H, A_2) - \widehat{Q}_{21}^H\widehat{A}_2||_2$, $||\mathbf{MM}(Q_{12}^H, A_1) - \widehat{Q}_{12}^H\widehat{A}_1||_2$, and $||\mathbf{MM}(Q_{22}^H, A_2) - \widehat{Q}_{22}^H\widehat{A}_2||_2$. To now bound $||\widetilde{R} - R'||_2$, note that

$$R' = \widehat{Q}_{11}^H\widehat{A}_1 + \widehat{Q}_{21}^H\widehat{A}_2 \tag{6.7}$$

since $\widehat{A} = \widehat{Q}R$. Consequently,

$$\begin{aligned} ||\widetilde{R} - R'||_2 \leq ||\widetilde{R}&-(\mathbf{MM}(Q_{11}^H, A_1) + \mathbf{MM}(Q_{21}^H, A_2))||_2 \\ &+ ||\mathbf{MM}(Q_{11}^H, A_1) - \widehat{Q}_{11}^H\widehat{A}_1||_2 \\ &+ ||\mathbf{MM}(Q_{21}^H, A_2) - \widehat{Q}_{21}^H\widehat{A}_2||_2. \end{aligned} \tag{6.8}$$

By Lemma 6.1.5, we can bound the first term by $\sqrt{n}\mathbf{u}||\mathbf{MM}(Q_{11}^H, A_1) + \mathbf{MM}(Q_{21}^H, A_2)||_2$, where

$$\begin{aligned}
||\mathbf{MM}(Q_{11}^H, A_1) + \mathbf{MM}(Q_{21}^H, A_2)||_2 &\le ||\mathbf{MM}(Q_{11}^H, A_1)||_2 + ||\mathbf{MM}(Q_{21}^H, A_2)||_2 \\
&\le ||Q_{11}^H A_1||_2 + \mu_{\mathrm{MM}}(n)\mathbf{u}||Q_{11}||_2||A_1||_2 \\
&\qquad + ||Q_{21}^H A_2||_2 + \mu_{\mathrm{MM}}(n)\mathbf{u}||Q_{21}||_2||A_2||_2 \qquad (6.9) \\
&\le 2(1+\mu_{\mathrm{MM}}(n)\mathbf{u})(1+\mu_{\mathrm{QR}}(n)\mathbf{u})||A||_2 \\
&\le 8||A||_2.
\end{aligned}$$

Applying this to (6.8) alongside (6.6) yields

$$||\widetilde{R} - R'||_2 \le 4\left(\mu_{\mathrm{QR}}(n) + \mu_{\mathrm{MM}}(n) + 2\sqrt{n}\right)\mathbf{u}||A||_2. \tag{6.10}$$

We obtain the same bound for $||E||_2$ by repeating this argument with $\mathbf{MM}(Q_{12}^H, A_1)$ and $\mathbf{MM}(Q_{22}^H, A_2)$ and noting that $\widehat{Q}_{12}^H\widehat{A}_1 + \widehat{Q}_{22}^H\widehat{A}_2 = 0$. □

In the subsequent sections, as in the previous two lemmas, we tacitly assume that input matrices can be represented exactly on our floating-point machine. This is done to simplify the analysis, discarding what amounts to negligible additive errors due to floating-point representation.

6.2 Finite-Precision Shattering

In this section, we consider how floating-point computations impact pseudospectral shattering. We begin with $\Lambda_\epsilon(\widetilde{A}, n^\alpha\widetilde{B})$. A straightforward extension of [16, Theorem 3.13] implies the following finite-precision counterpart to Theorem 2.3.1.

Theorem 6.2.1. *Let $A, B \in \mathbb{C}^{n\times n}$ with $||A||_2, ||B||_2 \le 1$ and let $0 < \gamma < \frac{1}{2}$. Further, let $\omega = \frac{\gamma^4}{4}n^{\frac{-8\alpha+13}{3}}$ and construct the grid $g = \mathrm{grid}(z, \omega, \lceil 8/\omega\rceil, \lceil 8/\omega\rceil)$ for $\alpha > 0$ and z chosen uniformly at random from the square with bottom left corner $-4-4i$ and side length ω. On a floating-point machine with precision $\mathbf{u}$, suppose $G_1, G_2 \in \mathbb{C}^{n\times n}$ satisfy $G_k(i,j) = \mathrm{N}(\frac{1}{\sqrt{n}})$*

for $1 \leq i, j \leq n$ *and* $k = 1, 2$*. If* $\widetilde{A} = A + \gamma G_1$ *and* $\widetilde{B} = B + \gamma G_2$ *(again in finite precision) then* $\Lambda_\epsilon(\widetilde{A}, n^\alpha \widetilde{B})$ *is shattered with respect to* g *for*

$$\epsilon = \frac{1}{2} \cdot \frac{\gamma^5}{64n^{\frac{11\alpha+25}{3}} + \gamma^5}$$

with probability at least $\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right]$ *provided*

$$\mathbf{u} \leq \frac{1}{2(3 + c_{\mathrm{N}})n^{\alpha+\frac{1}{2}}} \cdot \frac{\gamma^5}{64n^{\frac{11\alpha+25}{3}} + \gamma^5}.$$

Proof. Combining Assumption 6.1.1 and Lemma 6.1.5, $\widetilde{A}$ and $\widetilde{B}$ are at most $(3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}$ away (in the spectral norm) from their exact-arithmetic counterparts. Accommodating the n^α scaling on $\widetilde{B}$ and recalling that shattering is achieved in exact arithmetic for $\epsilon = \frac{\gamma^5}{64n^{\frac{11\alpha+25}{3}} + \gamma^5}$ with probability at least $\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right]$, Lemma 2.3.3 implies that it is sufficient to take

$$(3 + c_{\mathrm{N}})n^{\alpha+\frac{1}{2}}\mathbf{u} \leq \frac{1}{2} \cdot \frac{\gamma^5}{64n^{\frac{11\alpha+25}{3}} + \gamma^5}, \tag{6.11}$$

which is equivalent to the listed requirement on $\mathbf{u}$. □

To put this result in context, note that in producing a diagonalization of (A, B) with (spectral norm) accuracy ε, **RPD** sets $\gamma = O(\varepsilon)$. Hence, Theorem 6.2.1 implies that $O(\log(\frac{n}{\varepsilon}))$ bits of precision are required to obtain pseudospectral shattering as part of a floating-point diagonalization algorithm. Importantly, this is the same asymptotic precision derived by Banks et al. for single-matrix shattering (see Table 6.1).

Recall that in Chapter 2 we also proved shattering for $\Lambda_\epsilon(n^{-\alpha}\widetilde{B}^{-1}\widetilde{A})$. With this in mind, we pursue a similar precision bound that will guarantee shattering for the product matrix $n^{-\alpha}\widetilde{B}^{-1}\widetilde{A}$. Given Figure 2.1, we expect that higher precision will be necessary here, as we must account for error incurred by inverting $\widetilde{B}$ and multiplying by $\widetilde{A}$ in addition to the error already baked into $\widetilde{A}$ and $\widetilde{B}$. This is in part captured by the following intermediate result.

Lemma 6.2.2. *Suppose* $A_1, A_2, B_1, B_2 \in \mathbb{C}^{n\times n}$ *with* $||A_1 - A_2||_2, ||B_1 - B_2||_2 \leq \delta$ *for* $\delta < \sigma_n(B_1)$. *Then*

$$||B_1^{-1}A_1 - B_2^{-1}A_2||_2 \leq \delta ||B_1^{-1}||_2 \left(1 + \frac{||A_1||_2 + \delta}{\sigma_n(B_1) - \delta}\right).$$

Proof. We have

$$\begin{aligned} ||B_1^{-1}A_1 - B_2^{-1}A_2||_2 &= ||B_1^{-1}A_1 - B_1^{-1}A_2 + B_1^{-1}A_2 - B_2^{-1}A_2||_2 \\ &\leq ||B_1^{-1}||_2||A_1 - A_2||_2 + ||B_1^{-1} - B_2^{-1}||_2||A_2||_2 \\ &\leq \delta||B_1^{-1}||_2 + ||B_1^{-1}||_2||B_2 - B_1||_2||B_2^{-1}||_2||A_2||_2 \qquad (6.12) \\ &\leq \delta||B_1^{-1}||_2 \left(1 + \frac{||A_2||_2}{\sigma_n(B_2)}\right) \\ &\leq \delta||B_1^{-1}||_2 \left(1 + \frac{||A_1||_2 + \delta}{\sigma_n(B_1) - \delta}\right), \end{aligned}$$

where the last inequality follows from Lemma 1.7.2. □

Combining Lemma 6.2.2 with Assumption 6.1.2 and Assumption 6.1.4 yields the following floating-point counterpart to Proposition 2.3.2. Note that we use fast inversion here under the assumption that shattering of $\Lambda_\epsilon(n^{-\alpha}\widetilde{B}^{-1}\widetilde{A})$ would be similarly used as part of a fast diagonalization algorithm for (A, B).

Theorem 6.2.3. *Let* $A, B \in \mathbb{C}^{n\times n}$ *with* $||A||_2, ||B||_2 \leq 1$ *and let* $0 < \gamma < \frac{1}{2}$. *Further let* $\omega = \frac{\gamma^4}{4}n^{-\frac{8\alpha+13}{3}}$ *and construct the grid* $g = \mathrm{grid}(z, \omega, \lceil 8/\omega \rceil, \lceil 8/\omega \rceil)$ *for* $\alpha > 0$ *and* z *chosen uniformly at random from the square with bottom left corner* $-4-4i$ *and side length* ω. *On a floating-point machine with precision* $\mathbf{u}$, *suppose* $G_1, G_2 \in \mathbb{C}^{n\times n}$ *satisfy* $G_k(i,j) = \mathrm{N}(\frac{1}{\sqrt{n}})$ *for* $1 \leq i, j \leq n$ *and* $k = 1, 2$. *If* $\widetilde{A} = A + \gamma G_1$, $\widetilde{B} = B + \gamma G_2$ *(again in finite precision) and further*

$$M = \mathbf{MM}(\mathbf{INV}(\widetilde{B}), \widetilde{A}),$$

then $\Lambda_\epsilon(n^{-\alpha}M)$ *is shattered with respect to* g *for*

$$\epsilon = \frac{\gamma^5}{32}n^{-\frac{11\alpha+25}{3}}$$

with probability at least $\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right]$ *provided*

$$\mathbf{u} \le \frac{1}{21\mu_{\mathrm{INV}}(n)} \cdot \frac{1}{(6n^\alpha + 1)^{c_{\mathrm{INV}}\log(n)}} \cdot \frac{\gamma^5}{64} n^{-\frac{11\alpha+25}{3}}.$$

Proof. Set $(A_2, B_2) = (\widetilde{A}, \widetilde{B})$ and let (A_1, B_1) be the corresponding exact-arithmetic pencil (perturbed with true Ginibre matrices). Further let $X = n^{-\alpha}B_1^{-1}A_1$ and $X' = n^{-\alpha}B_2^{-1}A_2$. Here, X is the exact-arithmetic product covered by Proposition 2.3.2 while X' is the exact product corresponding to (A_2, B_2), equivalently a floating-point version of X that assumes exact inversion and matrix multiplication. Throughout, we assume access to the events that guarantee shattering in Proposition 2.3.2, in particular $\sigma_n(B_1) \ge n^{-\alpha}$, and $||A_1||_2, ||B_1||_2 \le 3$, which occur with probability at least $\left[1 - \frac{82}{n} - \frac{531441}{16n^2}\right]\left[1 - \frac{n^{2-2\alpha}}{\gamma^2} - 4e^{-n}\right]$.

As in the proof of Theorem 6.2.1, we start by noting $||A_1 - A_2||_2, ||B_1 - B_2||_2 \le (3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}$. Since $\mathbf{u} \le \frac{1}{2(3+c_{\mathrm{N}})n^{\alpha+1/2}}$ and therefore $(3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u} \le \frac{1}{2n^\alpha} < \sigma_n(B_1)$, Lemma 6.2.2 implies

$$||X - X'||_2 \le 2(3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}(1 + 3n^\alpha). \tag{6.13}$$

With this in mind, we next seek a bound on $||n^{-\alpha}M - X'||_2$. To do this, let $C = \mathbf{INV}(B_2)$. Applying Assumption 6.1.4, we have

$$||C - B_2^{-1}||_2 \le \mu_{\mathrm{INV}}(n)\mathbf{u}\kappa_2(B_2)^{c_{\mathrm{INV}}\log(n)}||B_2^{-1}||_2. \tag{6.14}$$

By Lemma 1.7.2, $||B_2||_2 \le 3 + (3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}$ and $\sigma_n(B_2) \ge n^{-\alpha} - (3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}$, so this can be simplified to

$$\begin{aligned} ||C - B_2^{-1}||_2 &\le \mu_{\mathrm{INV}}(n)\mathbf{u}\left(\frac{(3 + (3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u})n^\alpha}{1 - [(3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}]n^\alpha}\right)^{c_{\mathrm{INV}}\log(n)} \frac{n^\alpha}{1 - [(3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u}]n^\alpha} \\ &\le \mu_{\mathrm{INV}}(n)\mathbf{u}(6n^\alpha + 1)^{c_{\mathrm{INV}}\log(n)}(2n^\alpha), \end{aligned} \tag{6.15}$$

where we again use the fact that $(3 + c_{\mathrm{N}})\sqrt{n}\mathbf{u} \le \frac{1}{2n^\alpha}$. Now $M = \mathbf{MM}(C, A_2)$, so by Assumption 6.1.2

$$||M - CA_2||_2 \le \mu_{\mathrm{MM}}(n)\mathbf{u}||C||_2||A_2||_2. \tag{6.16}$$

Bounding $||C||_2$ via (6.15) as

$$\begin{aligned} ||C||_2 &\leq ||B_2^{-1}||_2 + \mu_{\text{INV}}(n)\mathbf{u}(6n^\alpha+1)^{c_{\text{INV}}\log(n)}(2n^\alpha) \\ &\leq 2n^\alpha\left[1+\mu_{\text{INV}}(n)\mathbf{u}(6n^\alpha+1)^{c_{\text{INV}}\log(n)}\right] \end{aligned} \tag{6.17}$$

and further noting $||A_2||_2 \leq ||A||_2 + (3+c_{\text{N}})\sqrt{n}\mathbf{u} \leq 3 + (3+c_{\text{N}})\sqrt{n}\mathbf{u}$, we conclude

$$\begin{aligned} ||M - CA_2||_2 &\leq \mu_{\text{MM}}(n)\mathbf{u}(6n^\alpha+1)\left[1+\mu_{\text{INV}}(n)\mathbf{u}(6n^\alpha+1)^{c_{\text{INV}}\log(n)}\right] \\ &= \mu_{\text{MM}}(n)\mathbf{u}(6n^\alpha+1) + \mu_{\text{MM}}(n)\mu_{\text{INV}}(n)\mathbf{u}^2(6n^\alpha+1)^{c_{\text{INV}}\log(n)+1}. \end{aligned} \tag{6.18}$$

Putting everything together, we have

$$\begin{aligned} ||n^{-\alpha}M - X'||_2 &= ||n^{-\alpha}M - n^{-\alpha}CA_2 + n^{-\alpha}CA_2 - n^{-\alpha}B_2^{-1}A_2||_2 \\ &\leq n^{-\alpha}||M-CA_2||_2 + n^{-\alpha}||C - B_2^{-1}||_2||A_2||_2 \\ &\leq 7\left[\mu_{\text{MM}}(n)\mathbf{u} + \mu_{\text{MM}}(n)\mu_{\text{INV}}(n)\mathbf{u}^2(6n^\alpha+1)^{c_{\text{INV}}\log(n)}\right. \\ &\qquad \left. + \mu_{\text{INV}}(n)\mathbf{u}(6n^\alpha+1)^{c_{\text{INV}}\log(n)}\right] \end{aligned} \tag{6.19}$$

after applying $n^{-\alpha}(6n^\alpha+1) \leq 7$ and $||A_2||_2 \leq 3 + (3+c_{\text{N}})\sqrt{n}\mathbf{u} \leq \frac{7}{2}$ to simplify constants. The last term in this expression clearly dominates; assuming each piece of (6.19) is bounded by the last one, we obtain

$$||n^{-\alpha}M - X'||_2 \leq 21\mu_{\text{INV}}(n)\mathbf{u}(6n^\alpha+1)^{c_{\text{INV}}\log(n)}. \tag{6.20}$$

Combining this with (6.13) yields our final error bound:

$$||n^{-\alpha}M - X||_2 \leq 21\mu_{\text{INV}}(n)\mathbf{u}(6n^\alpha+1)^{c_{\text{INV}}\log(n)} + 2(3+c_{\text{N}})\sqrt{n}\mathbf{u}(1+3n^\alpha). \tag{6.21}$$

Since Proposition 2.3.2 implies that $\Lambda_\epsilon(X)$ is shattered for $\epsilon = \frac{\gamma^5}{16}n^{-\frac{11\alpha+25}{3}}$, we ensure shattering for $\Lambda_\epsilon(n^{-\alpha}M)$ with $\epsilon = \frac{\gamma^5}{32}n^{-\frac{11\alpha+25}{3}}$ by requiring that each piece of (6.21) is bounded by $\frac{\gamma^5}{64}n^{-\frac{11\alpha+25}{3}}$, which is guaranteed as long as

$$\mathbf{u} \leq \frac{1}{21\mu_{\text{INV}}(n)} \cdot \frac{1}{(6n^\alpha+1)^{c_{\text{INV}}\log(n)}} \cdot \frac{\gamma^5}{64}n^{-\frac{11\alpha+25}{3}}. \tag{6.22}$$

Note that this requirement on $\mathbf{u}$ satisfies the assumptions used throughout the proof. □

Comparing Theorem 6.2.3 with Theorem 6.2.1 makes clear the practical cost of forming $\widetilde{B}^{-1}\widetilde{A}$. In particular, finite-precision shattering for the pencil requires $O(\log(\frac{n}{\varepsilon}))$ bits of precision versus $O(\log(\frac{n}{\varepsilon}) + \log^2(n))$ for the product matrix, where the polylogarithmic increase in the latter is rooted in the logarithmic stability of **INV**. This captures rigorously the phenomenon displayed in Figure 2.1 and underscores again the necessity of avoiding matrix inversion.

6.3 General IRS Bounds

We turn next to **IRS**. Motivated by Remark 3.2.4, our goal is to obtain general precision bounds for the routine, capable of capturing its performance in a variety of settings, including (but not limited to) divide-and-conquer. If $[\widetilde{A}_p, \widetilde{B}_p] = \mathbf{IRS}(A, B, p)$ on our floating-point machine, we seek the precision $\mathbf{u}$ that guarantees

$$||\widetilde{A}_p - A_p||_2, ||\widetilde{B}_p - B_p||_2 \leq \delta \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 \tag{6.23}$$

for arbitrary $\delta > 0$ and A_p, B_p a corresponding set of exact-arithmetic outputs, which satisfy $A_p^{-1}B_p = (A^{-1}B)^{2^p}$ exactly. Intuitively, such a precision $\mathbf{u}$ will depend on the accuracy parameter δ, the size of the pencil (A, B), p the number steps of squaring, and an appropriately chosen condition number. Noting that exact-arithmetic repeated squaring can at most decrease the norms of the input matrices, we can think of (6.23) as a weak forward error bound.

Analyzing **IRS** in this way marks a departure from the literature. While error in the routine has been bounded rigorously in finite-precision arithmetic before, most notably by Malyshev [93, 94] and Bai, Demmel, and Gu [7], results typically center the spectral projector application and make use of standard $O(n^3)$ matrix multiplication/QR routines. We aim to address both of these shortcomings here, presenting general bounds that are explicitly compatible with fast matrix multiplication (via the black-box assumptions from Section 6.1).

6.3.1 Condition Number and Technical Lemmas

We start by defining a condition number κ_{IRS}. We justify our choice in the following subsection, where we show that κ_{IRS} naturally bounds error in repeated squaring.

Definition 6.3.1. Given $A, B \in \mathbb{C}^{n\times n}$ and $p \geq 1$, define the block matrix

$$D^p_{(A,B)} = \begin{pmatrix} B & & & \\ -A & B & & \\ & -A & \ddots & \\ & & \ddots & B \\ & & & -A \end{pmatrix} \in \mathbb{C}^{2^p n \times (2^p-1)n}.$$

The *condition number* of **IRS** corresponding to the inputs A, B, and p is

$$\kappa_{\text{IRS}}(A,B,p) = \sigma_{\min}(D^p_{(A,B)})^{-1} \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 .$$

Note that κ_{IRS} satisfies a number of properties we should expect from a suitable condition number; it is invariant to both swapping A and B and scaling the pencil (A,B), and it also satisfies $\kappa_{\text{IRS}}(A,B,p) \geq 1$ for any A, B and p.[2] Recalling the discussion from Section 3.2, we might ask how κ_{IRS} compares to the quantities $\omega_{(A,B)}$ and $d_{(A,B)}$ of Malyshev [95] and Bai, Demmel, and Gu [7]. The following lemma provides an answer.

Lemma 6.3.2. *Let (A,B) be an $n \times n$ regular pencil. Then for any $p \geq 1$ we have*

$$\sigma_{\min}(D^p_{(A,B)}) \geq d_{(A,B)} \geq \frac{\sqrt{\sigma_n(AA^H + BB^H)}}{14\omega_{(A,B)}}$$

Proof. Let $m = 2^p$ and define the $mn \times mn$ block matrix

$$M_p(A,B) = \begin{pmatrix} -A & & & -B \\ B & -A & & \\ & \ddots & \ddots & \\ & & B & -A \end{pmatrix}. \tag{6.24}$$

To first show $\sigma_{\min}(D^p_{(A,B)}) \geq \sigma_{\min}(M_p(A,B))$, let $x = [x_1\ x_2\ \cdots\ x_{m-1}]^T \in \mathbb{C}^{(m-1)n}$ be the unit vector satisfying $\sigma_{\min}(D^p_{(A,B)}) = ||D^p_{(A,B)}x||_2$, where $x_i \in \mathbb{C}^n$ for each i. Padding x

[2] This follows form the observation $\sigma_{\min}(D^p_{(A,B)}) \leq \sigma_{\min}\binom{A}{B}$.

with zeros to obtain another unit vector

$$y = [x_{m-1} \;\; x_{m-2} \;\; \cdots \;\; x_1 \;\; 0] \in \mathbb{C}^{mn} \tag{6.25}$$

it is easy to see $||M_p(A,B)y||_2 = ||D^p_{(A,B)}x||_2$ and therefore

$$\sigma_{\min}(M_p(A,B)) \leq ||M_p(A,B)y||_2 = ||D^p_{(A,B)}x||_2 = \sigma_{\min}(D^p_{(A,B)}). \tag{6.26}$$

The first inequality now follows from an observation of Bai, Demmel, and Gu, who show that $M_p(A,B)$ is unitarily equivalent to the block matrix $\operatorname{diag}(-A+e^{i\theta_1}B, \ldots, -A+e^{i\theta_m}B)$ for $e^{i\theta_1}, \ldots, e^{i\theta_m}$ the m^{th} roots of -1. Hence, we have

$$\sigma_{\min}(M_p(A,B)) = \min_{1 \leq j \leq m} \sigma_n(-A + e^{i\theta_j}B) \geq \min_{\theta} \sigma_n(-A + e^{i\theta}B) = d_{(A,B)}. \tag{6.27}$$

The remaining inequality can be derived from [95, Theorem 3]. Letting $LL^H = AA^H + BB^H$ be a Cholesky factorization (which exists since (A,B) is regular) and setting $A_0 = L^{-1}A$ and $B_0 = L^{-1}B$, we have

$$\frac{1}{14\omega_{(A,B)}} < \frac{1}{\max_\phi ||(B_0 - e^{i\phi}A_0)^{-1}||_2} \leq ||L^{-1}||_2 \min_\phi (B - e^{i\phi}A) = \frac{d_{(A,B)}}{\sigma_n(L)}. \tag{6.28}$$

We complete the proof by rearranging and recalling $\sigma_i(L)^2 = \sigma_i(AA^H + BB^H)$ for all i. □

Unlike $d^{-1}_{(A,B)}$ and $\omega_{(A,B)}$, $\kappa_{\text{IRS}}(A,B,p)$ is not necessarily infinite if (A,B) has an eigenvalue on the unit circle. In fact, $\kappa_{\text{IRS}}(A,B,p)$ is always finite if A or B is nonsingular. This underscores the general utility of κ_{IRS}, which can be used to bound error in repeated squaring even when (A,B) is ill-posed for the spectral projector application. Instead, as we will see, κ_{IRS} is infinite when the block QR factorization computed by **IRS** cannot be controlled by perturbation bounds, which are necessary to obtain a result like (6.23).

There is another important property unique to κ_{IRS}: it includes an explicit dependence on p, the number of steps of squaring. While $\kappa_{\text{IRS}}(A,B,p)$ increases with p, Lemma 6.3.2 implies the p-independent upper bound

$$\kappa_{\text{IRS}}(A,B,p) \leq d^{-1}_{(A,B)} \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2. \tag{6.29}$$

Thinking of p as an input to the procedure not only provides a sharper condition number but also allows us to quantify the stability of **IRS** in terms of the number of steps taken (and in particular its dependence on n). Again, this is aimed at producing general bounds. While it is common to consider the setting $p \to \infty$ when using **IRS** to compute spectral projectors, other applications may come with a fixed value of p, in which case the upper bound (6.29) could be loose.

To complete this section, we state a pair of technical lemmas due to Malyshev [93, Lemmas 4.1 and 4.2], which we'll need later on.

Lemma 6.3.3 (Malyshev 1992)**.** *Suppose $R \in \mathbb{C}^{m\times m}$ is nonsingular and $E \in \mathbb{C}^{n\times m}$ for $m \geq n$. There exists a matrix $S \in \mathbb{C}^{(m+n)\times(m+n)}$ such that*

1. $(I+S)\binom{R}{E} = \binom{R'}{0}$.
2. $(I+S)^H(I+S) = I$
3. $||S||_2 \leq ||ER^{-1}||_2 \leq ||E||_2||R^{-1}||_2$.

Proof. We take the opportunity to correct a small error in Malyshev's proof. Define

$$\widetilde{S} = \begin{pmatrix} 0 & (ER^{-1})^H \\ -ER^{-1} & 0 \end{pmatrix}. \tag{6.30}$$

Then the matrix

$$S = \begin{pmatrix} [I+(ER^{-1})^H ER^{-1}]^{-1/2} & 0 \\ 0 & [I+ER^{-1}(ER^{-1})^H]^{-1/2} \end{pmatrix}(I+\widetilde{S}) - I \tag{6.31}$$

satisfies the listed requirements. □

Lemma 6.3.4 (Malyshev 1992)**.** *Let $A \in \mathbb{C}^{m\times n}$ be full rank and suppose*

$$A = Q_1 \begin{pmatrix} K_1 & L_1 \\ 0 & M_1 \end{pmatrix} = Q_2 \begin{pmatrix} K_2 & L_2 \\ 0 & M_2 \end{pmatrix}$$

for $Q_1, Q_2 \in \mathbb{C}^{m\times m}$ unitary, $K_1, K_2 \in \mathbb{C}^{k\times k}$ nonsingular, and $M_1, M_2 \in \mathbb{C}^{(m-k)\times(n-k)}$ full rank. Then there exist unitary matrices $P \in \mathbb{C}^{k\times k}$ and $Q \in \mathbb{C}^{(n-k)\times(n-k)}$ such that $K_2 = PK_1$, $L_2 = PL_1$, and $M_2 = QM_1$.

Remark 6.3.5. Why is it important for R to be nonsingular in Lemma 6.3.3? While Malyshev's proof uses R^{-1} explicitly to construct S, we can always find a unitary matrix that zeros out E, even when R is singular. Indeed, we could obtain $I + S$ by simply computing a full QR factorization of $\binom{R}{E}$. The key here is the norm bound on S. We want to guarantee that if E is sufficiently close to zero, then $I + S$ is essentially the identity. In this way, Lemma 6.3.3 can be interpreted as a perturbation result similar to our full QR bounds from Chapter 3. Eventually we will apply this lemma to a matrix R whose smallest singular value can be bounded from below by $\sigma_{\min}(D^p_{(A,B)})$, and this is in fact our main motivation for the definition of κ_{IRS}.

6.3.2 Error Analysis

We are now ready to bound error in repeated squaring. To simplify the analysis, we will not track the individual polynomials $\mu_{\mathrm{MM}}(n)$ and $\mu_{\mathrm{QR}}(n)$, instead working with a "general polynomial"

$$\mu(n) = \max\left\{\mu_{\mathrm{MM}}(n), \mu_{\mathrm{QR}}(n), \sqrt{n}\right\} \tag{6.32}$$

and the associated quantity $\tau = \mu(n)\mathbf{u}$. Throughout, we can think of τ as small, corresponding to a choice $\mathbf{u} < \mu(n)^{-1}$.

First up is a lemma that bounds norm growth in **IRS**. Because finite-precision **IRS** repeatedly multiplies the inputs by pieces of nearly unitary matrices, we expect that norms should grow by (at most) small constants. Here, $\widetilde{A}_j$ and $\widetilde{B}_j$ are the outputs of j steps of finite-precision **IRS**, beginning with the input matrices $\widetilde{A}_0 = A$ and $\widetilde{B}_0 = B$. In this notation – and in terms of the black-box algorithms defined in Section 6.1 – each iteration of floating-point **IRS** consists of the following.

1. $[\widetilde{Q}, R] = \mathbf{QR}\left(\begin{bmatrix} \widetilde{B}_j \\ -\widetilde{A}_j \end{bmatrix}\right)$ with $\widetilde{Q} = \begin{pmatrix} \widetilde{Q}_{11} & \widetilde{Q}_{12} \\ \widetilde{Q}_{21} & \widetilde{Q}_{22} \end{pmatrix}$
2. $\widetilde{A}_{j+1} = \mathbf{MM}(\widetilde{Q}_{12}^H, \widetilde{A}_j)$

3. $\widetilde{B}_{j+1} = \mathbf{MM}(\widetilde{Q}_{22}^H, \widetilde{B}_j)$

Lemma 6.3.6. *At any step j, $||\binom{\widetilde{A}_{j+1}}{\widetilde{B}_{j+1}}||_2 \leq (1+2\tau)^2 ||\binom{\widetilde{A}_j}{\widetilde{B}_j}||_2$.*

Proof. As noted above, $\widetilde{A}_{j+1} = \mathbf{MM}(\widetilde{Q}_{12}^H, \widetilde{A}_j)$ and $\widetilde{B}_{j+1} = \mathbf{MM}(\widetilde{Q}_{22}^H, \widetilde{B}_j)$ for $\widetilde{Q}_{12}$ and $\widetilde{Q}_{22}$ blocks of a nearly unitary $\widetilde{Q}$ obtained by computing a finite-precision, full QR factorization of $\binom{\widetilde{B}_j}{-\widetilde{A}_j}$. With this in mind, write

$$\left|\left|\begin{pmatrix}\widetilde{A}_{j+1}\\ \widetilde{B}_{j+1}\end{pmatrix}\right|\right|_2 \leq \left|\left|\begin{pmatrix}\widetilde{A}_{j+1} - \widetilde{Q}_{12}^H\widetilde{A}_j\\ \widetilde{B}_{j+1} - \widetilde{Q}_{22}^H\widetilde{B}_j\end{pmatrix}\right|\right|_2 + \left|\left|\begin{pmatrix}\widetilde{Q}_{12}^H\widetilde{A}_j\\ \widetilde{Q}_{22}^H\widetilde{B}_j\end{pmatrix}\right|\right|_2. \tag{6.33}$$

By Assumption 6.1.2, $||\widetilde{A}_{j+1} - \widetilde{Q}_{12}^H\widetilde{A}_j||_2 \leq \tau||\widetilde{Q}_{12}||_2||\widetilde{A}_j||_2$ and $||\widetilde{B}_{j+1} - \widetilde{Q}_{22}^H\widetilde{B}_j||_2 \leq \tau||\widetilde{Q}_{22}||_2||\widetilde{B}_j||_2$, so

$$\left|\left|\begin{pmatrix}\widetilde{A}_{j+1} - \widetilde{Q}_{12}^H\widetilde{A}_j\\ \widetilde{B}_{j+1} - \widetilde{Q}_{22}^H\widetilde{B}_j\end{pmatrix}\right|\right|_2 \leq \sqrt{2}\tau(1+\tau)\left|\left|\begin{pmatrix}\widetilde{A}_j\\ \widetilde{B}_j\end{pmatrix}\right|\right|_2 \tag{6.34}$$

since $\widetilde{Q}_{12}$ and $\widetilde{Q}_{22}$ satisfy $||\widetilde{Q}_{12}||_2, ||\widetilde{Q}_{22}||_2 \leq ||\widetilde{Q}||_2 \leq 1+\tau$ and $||\widetilde{A}_j||_2, ||\widetilde{B}_j||_2 \leq ||\binom{\widetilde{A}_j}{\widetilde{B}_j}||_2$. Similarly,

$$\left|\left|\begin{pmatrix}\widetilde{Q}_{12}^H\widetilde{A}_j\\ \widetilde{Q}_{22}^H\widetilde{B}_j\end{pmatrix}\right|\right|_2 \leq \left|\left|\begin{pmatrix}\widetilde{Q}_{12}^H & 0\\ 0 & \widetilde{Q}_{22}^H\end{pmatrix}\right|\right|_2 \left|\left|\begin{pmatrix}\widetilde{A}_j\\ \widetilde{B}_j\end{pmatrix}\right|\right|_2 \leq (1+\tau)\left|\left|\begin{pmatrix}\widetilde{A}_j\\ \widetilde{B}_j\end{pmatrix}\right|\right|_2. \tag{6.35}$$

We obtain the final inequality by combining (6.34) and (6.35) and using the loose[3] upper bound $(\sqrt{2}\tau+1)(1+\tau) \leq (1+2\tau)^2$. □

The main result of this section can now be stated as follows.

Theorem 6.3.7. *Given $A, B \in \mathbb{C}^{n\times n}$ and $p \geq 1$, let $[\widetilde{A}_p, \widetilde{B}_p] = \mathbf{IRS}(A, B, p)$ on a floating-point machine with precision $\mathbf{u}$. For $\delta \in (0,1)$ and $\mu(n)$ as in (6.32) suppose*

$$\mathbf{u} \leq \frac{\delta}{324\mu(n)\kappa_{\mathrm{IRS}}(A,B,p)\max\{p^2+4p-5, 1\}}.$$

Then there exist matrices $\mathring{A}_p, \mathring{B}_p \in \mathbb{C}^{n\times n}$ such that $\mathring{A}_p^{-1}\mathring{B}_p = (A^{-1}B)^{2^p}$ and

$$||\widetilde{A}_p - \mathring{A}_p||_2, ||\widetilde{B}_p - \mathring{B}_p||_2 \leq \delta\left|\left|\begin{pmatrix}A\\ B\end{pmatrix}\right|\right|_2.$$

[3] We use this bound for convenience to simplify constants. As we will see, it does not significantly impact the final result.

Proof. Deriving this bound is fairly lengthy, so we once again break the proof into pieces. The high-level strategy (to keep in mind throughout) can be summarized as follows. Consider the $2^p n \times (2^p+1)n$ block matrix

$$M = \begin{pmatrix} -A & B & & & \\ & -A & B & & \\ & & \ddots & \ddots & \\ & & & -A & B \end{pmatrix}. \tag{6.36}$$

As we demonstrate below, the floating-point matrices used to obtain $\widetilde{A}_p$ and $\widetilde{B}_p$ via **IRS** can be built into an approximate block QR factorization of M (containing $\widetilde{A}_p$ and $\widetilde{B}_p$). Our goal will be to derive a nearby, *exact* block QR factorization of M, which will contain exact outputs $\mathring{A}_p$ and $\mathring{B}_p$ with bounds on $||\widetilde{A}_p - \mathring{A}_p||_2$ and $||\widetilde{B}_p - \mathring{B}_p||_2$ available. Since it will be relevant later on, note that the middle $2^p n \times (2^p-1)n$ block of M is the matrix $D^p_{(A,B)}$ from Definition 6.3.1.

This strategy for analyzing **IRS** is due to Malyshev [93]. In essence, we demonstrate here that his approach can accommodate fast matrix multiplication; in particular, it does not require doing QR via standard Householder reflectors, as is assumed in [93].

Step One: What happens if we apply the output of IRS to M in blocks?
Consider the first iteration of **IRS**, which computes

$$\left[\widetilde{Q}_1, \begin{pmatrix} R_1 \\ 0 \end{pmatrix}\right] = \mathbf{QR}\left(\begin{bmatrix} B \\ -A \end{bmatrix}\right) \tag{6.37}$$

for nearly unitary $\widetilde{Q}_1 \in \mathbb{C}^{2n\times 2n}$ and upper triangular $R_1 \in \mathbb{C}^{n\times n}$. Let $\widetilde{P}_1$ be the matrix

$$\widetilde{P}_1 = \begin{pmatrix} \widetilde{Q}_1 & & \\ & \ddots & \\ & & \widetilde{Q}_1 \end{pmatrix} \in \mathbb{C}^{2^p n \times 2^p n} \tag{6.38}$$

containing 2^{p-1} copies of $\widetilde{Q}_1$ on its diagonal. Further, let $\widetilde{M}_1$ be a floating-point approximation of $\widetilde{P}_1^H M$ obtained by applying **MM** (and finite-precision matrix addition) in $n \times n$

blocks. It is easy to see that $\widetilde{M_1}$ has block structure

$$\widetilde{M_1} = \begin{pmatrix} * & \widetilde{R}_1 & * & & & & & \\ -\widetilde{A}_1 & \widetilde{E}_1 & \widetilde{B}_1 & & & & & \\ & & * & \widetilde{R}_1 & * & & & \\ & & -\widetilde{A}_1 & \widetilde{E}_1 & \widetilde{B}_1 & & & \\ & & & & \ddots & \ddots & & \\ & & & & & * & \widetilde{R}_1 & * \\ & & & & & -\widetilde{A}_1 & \widetilde{E}_1 & \widetilde{B}_1 \end{pmatrix}, \tag{6.39}$$

where $*$ blocks are arbitrary. We use finite-arithmetic block matrix multiplication here – as opposed to a separate black-box algorithm for large, non-square matrices – to guarantee that $\widetilde{A}_1$ and $\widetilde{B}_1$ appear in (6.39). Note that the zero blocks of $\widetilde{M_1}$ are computed exactly by **MM**. Moreover, $\widetilde{R}_1$ and $\widetilde{E}_1$ are covered by Lemma 6.1.6 – i.e.,

$$||\widetilde{R}_1 - R_1||_2, ||\widetilde{E}_1||_2 \leq 16\tau \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2. \tag{6.40}$$

Step Two: Repeat this argument for the next iteration.

The second step of **IRS** computes

$$\left[\widetilde{U}, \begin{pmatrix} R_2 \\ 0 \end{pmatrix}\right] = \mathbf{QR}\left(\begin{bmatrix} \widetilde{B}_1 \\ -\widetilde{A}_1 \end{bmatrix}\right), \tag{6.41}$$

for $\widetilde{U} \in \mathbb{C}^{2n \times 2n}$ and $R_2 \in \mathbb{C}^{n \times n}$. Breaking $\widetilde{U}$ into $n \times n$ blocks $\widetilde{U} = \begin{pmatrix} \widetilde{U}_{11} & \widetilde{U}_{12} \\ \widetilde{U}_{21} & \widetilde{U}_{22} \end{pmatrix}$ and constructing

$$\widetilde{Q}_2 = \begin{pmatrix} I_n & & & \\ & \widetilde{U}_{11} & & \widetilde{U}_{12} \\ & & I_n & \\ & \widetilde{U}_{21} & & \widetilde{U}_{22} \end{pmatrix} \in \mathbb{C}^{4n \times 4n}, \tag{6.42}$$

let $\widetilde{P}_2$ be the matrix containing 2^{p-2} copies of $\widetilde{Q}_2$ on its diagonal – that is,

$$\widetilde{P}_2 = \begin{pmatrix} \widetilde{Q}_2 & & \\ & \ddots & \\ & & \widetilde{Q}_2 \end{pmatrix} \in \mathbb{C}^{2^p n \times 2^p n}. \tag{6.43}$$

Again applying **MM** in $n \times n$ blocks to left-multiply $\widetilde{M_1}$ by $\widetilde{P}_2$, we obtain $\widetilde{M_2}$ – a finite-precision version of $\widetilde{P}_2^H\widetilde{M_1}$ consisting of 2^{p-2} blocks of the form

$$\begin{pmatrix} * & \widetilde{R}_1 & * & & \\ * & * & \widetilde{R}_2 & * & * \\ & & * & \widetilde{R}_1 & * \\ -\widetilde{A}_2 & \mathbf{MM}(\widetilde{U}_{12}^H, \widetilde{E}_1) & \widetilde{E}_2 & \mathbf{MM}(\widetilde{U}_{22}^H, \widetilde{E}_1) & \widetilde{B}_2 \end{pmatrix}. \tag{6.44}$$

Step Three: Generalize to an arbitrary step of IRS.

The process outlined above yields a sequence of $2^pn \times (2^p+1)n$ matrices $\widetilde{M_1}, \ldots, \widetilde{M_p}$. Each $\widetilde{M_i}$ is an approximation of the exact product $\widehat{M_i} = \widetilde{P}_i^H\widetilde{P}_{i-1}^H \cdots \widetilde{P}_1^H M$ for a corresponding set of nearly unitary matrices $\widetilde{P}_1, \ldots, \widetilde{P}_p$, each of which is constructed from the blocks of a $2n \times 2n$ nearly unitary matrix as in (6.42). Moreover, $\widetilde{M_i}$ consists of 2^{p-i} blocks with structure

$$\begin{pmatrix} * & * & * \\ -\widetilde{A}_i & \widetilde{\Delta}_i & \widetilde{B}_i \end{pmatrix} \tag{6.45}$$

for $\widetilde{\Delta}_i$ a small $n \times (2^i-1)n$ matrix. Indeed, the center $n \times n$ block $\widetilde{E}_i$ of $\widetilde{\Delta}_i$ satisfies

$$||\widetilde{E}_i||_2 \le 16\tau \left|\left| \begin{pmatrix} \widetilde{B}_{i-1} \\ -\widetilde{A}_{i-1} \end{pmatrix} \right|\right|_2 \le 16\tau(1+2\tau)^{2i-2} \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|. \tag{6.46}$$

Step Four: Construct a corresponding set of exact-arithmetic block matrices.

Suppose that $\widetilde{P}_i$ is constructed from the blocks of the nearly unitary matrix $\widetilde{Q} \in \mathbb{C}^{2n\times 2n}$. Since we use **QR** to obtain $\widetilde{Q}$, as described above, we know by Assumption 6.1.3 that there exists a truly unitary matrix $Q \in \mathbb{C}^{2n\times 2n}$ such that $||\widetilde{Q} - Q||_2 \le \mu_{\mathrm{QR}}(n)\mathbf{u} \le \tau$. With this in mind, let P_i be the truly unitary matrix that has the same block structure as $\widetilde{P}_i$ but swaps the blocks of $\widetilde{Q}$ for the corresponding blocks of Q and define the $2^pn \times (2^p+1)n$ matrices $M_i = P_i^H \cdots P_1^H M$.

We now have two sets of exact-arithmetic matrices to work with: M_i and $\widehat{M_i}$. The former can be thought of as an exact-arithmetic counterpart of $\widetilde{M_i}$ while $\widehat{M_i}$ is an intermediate matrix, obtained via exact multiplication with the nearly unitary $\widetilde{P}_i$. Since

$||P_i - \widetilde{P}_i||_2 \leq \tau$ by construction, we can easily bound $||\widehat{M}_i - M_i||_2$ recursively:

$$\begin{aligned} ||\widehat{M}_i - M_i||_2 &= ||\widehat{M}_i - P_i^H \widehat{M}_{i-1} + P_i^H \widehat{M}_{i-1} - M_i||_2 \\ &\leq ||\widehat{M}_i - P_i^H \widehat{M}_{i-1}||_2 + ||P_i^H \widehat{M}_{i-1} - M_i||_2 \\ &\leq ||\widetilde{P}_i - P_i||_2 ||\widehat{M}_{i-1}||_2 + ||P_i||_2 ||\widehat{M}_{i-1} - M_{i-1}||_2 \\ &\leq \tau(1+\tau)^{i-1}||M||_2 + ||\widehat{M}_{i-1} - M_{i-1}||_2 \end{aligned} \tag{6.47}$$

The base case here is $||\widehat{M}_1 - M_1||_2 \leq ||\widetilde{P}_1 - P_1||_2 ||M||_2 \leq \tau||M||_2$, so by induction we obtain

$$||\widehat{M}_i - M_i||_2 \leq \left(\sum_{j=0}^{i-1}(1+\tau)^j\right)\tau||M||_2 = \left[(1+\tau)^i - 1\right]||M||_2. \tag{6.48}$$

Note that $\widehat{M}_i$ and M_i have the same block structure as $\widetilde{M}_i$, a consequence of the fact that each P_i has the same block structure as $\widetilde{P}_i$. Following (6.45), label the blocks of $\widehat{M}_i$ and M_i as

$$\begin{pmatrix} * & * & * \\ -\widehat{A}_i & \widehat{\Delta}_i & \widehat{B}_i \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} * & * & * \\ -A_i & \Delta_i & B_i \end{pmatrix}, \tag{6.49}$$

respectively, where again $*$ blocks are arbitrary and $\widehat{\Delta}_i, \Delta_i \in \mathbb{C}^{n\times(2^i-1)n}$.

Step Five: Bound $||\widetilde{A}_i - A_i||_2$ and $||\widetilde{B}_i - B_i||_2$.

The matrices A_i and B_i in (6.49) are not necessarily the result of applying i steps of exact-arithmetic repeated squaring to A and B. This follows from the fact that the unitary matrices P_i used to obtain A_i and B_i from M do not correspond to true QR factorizations of exact outputs. Rather, Assumption 6.1.3 implies that each P_i is constructed from the Q-factor of a matrix nearby $\left(\begin{smallmatrix} \widetilde{B}_j \\ -\widetilde{A}_j \end{smallmatrix}\right)$. Nevertheless, we are still interested in bounding $||\widetilde{A}_i - A_i||_2$ and $||\widetilde{B}_i - B_i||_2$ as an intermediate step. The heuristic to keep in mind is the following: while A_i and B_i are not the exact outputs we're in search of, they should be close to a pair that will satisfy the guarantees of the theorem.

Consider first $||\widetilde{A}_i - A_i||_2$. Since

$$||\widetilde{A}_i - A_i||_2 \leq ||\widetilde{A}_i - \widehat{A}_i||_2 + ||\widehat{A}_i - A_i||_2 \leq ||\widetilde{A}_i - \widehat{A}_i||_2 + ||\widehat{M}_i - M_i||_2, \tag{6.50}$$

and given (6.48), we can bound $||\widetilde{A}_i - A_i||_2$ via (6.50) by bounding $||\widetilde{A}_i - \widehat{A}_i||_2$, which records the error due to finite-precision, block matrix multiplication. With this in mind, suppose $\widetilde{A}_i = \mathbf{MM}(\widetilde{Q}_{12}^H, \widetilde{A}_{i-1})$ for $\widetilde{Q}_{12}$ an $n \times n$ block of a nearly unitary $2n \times 2n$ matrix. In this case, $\widehat{A}_i = \widetilde{Q}_{12}^H \widehat{A}_{i-1}$ and we have

$$\begin{aligned} ||\widetilde{A}_i - \widehat{A}_i||_2 &= ||\mathbf{MM}(\widetilde{Q}_{12}^H, \widetilde{A}_{i-1}) - \widetilde{Q}_{12}^H \widehat{A}_{i-1}||_2 \\ &= ||\mathbf{MM}(\widetilde{Q}_{12}^H, \widetilde{A}_{i-1}) - \widetilde{Q}_{12}^H \widetilde{A}_{i-1} + \widetilde{Q}_{12}^H \widetilde{A}_{i-1} - \widetilde{Q}_{12}^H \widehat{A}_{i-1}||_2 \\ &\leq ||\mathbf{MM}(\widetilde{Q}_{12}^H, \widetilde{A}_{i-1}) - \widetilde{Q}_{12}^H \widetilde{A}_{i-1}||_2 + ||\widetilde{Q}_{12}^H \widetilde{A}_{i-1} - \widetilde{Q}_{12}^H \widehat{A}_{i-1}||_2. \end{aligned} \tag{6.51}$$

Applying our black-box assumptions and Lemma 6.3.6, we have

$$\begin{aligned} ||\widetilde{A}_i - \widehat{A}_i||_2 &\leq \tau ||\widetilde{Q}_{12}||_2 ||\widetilde{A}_{i-1}||_2 + ||\widetilde{Q}_{12}||_2 ||\widetilde{A}_{i-1} - \widehat{A}_{i-1}||_2 \\ &\leq \tau(1+\tau) \left|\left| \begin{pmatrix} \widetilde{A}_{i-1} \\ \widetilde{B}_{i-1} \end{pmatrix} \right|\right|_2 + (1+\tau) ||\widetilde{A}_{i-1} - \widehat{A}_{i-1}||_2 \\ &\leq \tau(1+\tau)(1+2\tau)^{2i-2} \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 + (1+\tau)||\widetilde{A}_{i-1} - \widehat{A}_{i-1}||_2. \end{aligned} \tag{6.52}$$

Once again we obtain a recursive bound. In this notation $\widetilde{A}_0 = \widehat{A}_0 = A$, so the base case here is simply the error in one finite-precision $n \times n$ matrix multiplication – i.e.,

$$||\widetilde{A}_1 - \widehat{A}_1||_2 \leq \tau(1+\tau)||A||_2 \leq \tau(1+\tau) \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 . \tag{6.53}$$

Thus, we conclude inductively

$$\begin{aligned} ||\widetilde{A}_i - \widehat{A}_i||_2 &\leq \tau \left(\sum_{j=1}^{i} (1+\tau)^{i-j+1} (1+2\tau)^{2j-2} \right) \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 \\ &= \tau(1+2\tau)^{2i} \left(\sum_{j=1}^{i} \left[\frac{1+\tau}{(1+2\tau)^2} \right]^{i-j+1} \right) \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 \\ &= \tau(1+2\tau)^{2i} \cdot \frac{1+\tau}{(1+2\tau)^2} \cdot \frac{1 - \left(\frac{1+\tau}{(1+2\tau)^2}\right)^i}{1 - \frac{1+\tau}{(1+2\tau)^2}} \cdot \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 \\ &= \tau(1+\tau) \frac{(1+2\tau)^{2i} - (1+\tau)^i}{(1+2\tau)^2 - (1+\tau)} \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right| \\ &= \frac{1+\tau}{3+4\tau} \left[(1+2\tau)^{2i} - (1+\tau)^i \right] \left|\left| \begin{pmatrix} A \\ B \end{pmatrix} \right|\right|_2 . \end{aligned} \tag{6.54}$$

Combining this with (6.48) and $\frac{1+\tau}{3+4\tau} < 1$, and noting $||M||_2 \leq ||A||_2 + ||B||_2 \leq 2||\left(\begin{smallmatrix}A\\B\end{smallmatrix}\right)||_2$, we have

$$\begin{aligned}||\widetilde{A}_i - A_i||_2 &\leq \left[(1+2\tau)^{2i} - (1+\tau)^i\right] \left|\left|\begin{pmatrix}A\\B\end{pmatrix}\right|\right|_2 + 2\left[(1+\tau)^i - 1\right] \left|\left|\begin{pmatrix}A\\B\end{pmatrix}\right|\right|_2 \\ &= \left[(1+2\tau)^{2i} + (1+\tau)^i - 2\right] \left|\left|\begin{pmatrix}A\\B\end{pmatrix}\right|\right|_2 .\end{aligned} \tag{6.55}$$

Repeating this argument implies the same bound for $||\widetilde{B}_i - B_i||_2$.

Step Six: Show that $||\Delta_i||_2$ is small.

If M_i *was* obtained from M via exact-arithmetic repeated squaring, we would have $\Delta_i = 0$. Hence, the norm of Δ_i is an indication of how far A_i and B_i are from exact outputs. With this in mind, we next derive a bound on $||\Delta_i||_2$.

We start by bounding $||\widehat{\Delta}_i||_2$, beginning with its middle $n \times n$ block $\widehat{E}_i$, which corresponds to $\widetilde{E}_i$ in $\widetilde{M}_i$. If we again assume that $\widetilde{P}_i$ is built from the $2n \times 2n$ nearly unitary matrix $\widetilde{Q} = \begin{pmatrix}\widetilde{Q}_{11} & \widetilde{Q}_{12}\\ \widetilde{Q}_{21} & \widetilde{Q}_{22}\end{pmatrix}$, we know that $\widetilde{E}_i$ is the finite-arithmetic sum of $\mathbf{MM}(\widetilde{Q}^H_{12}, \widetilde{B}_{i-1})$ and $\mathbf{MM}(\widetilde{Q}^H_{22}, -\widetilde{A}_{i-1})$ while $\widehat{E}_i = \widetilde{Q}^H_{12}\widehat{B}_{i-1} - \widetilde{Q}^H_{22}\widehat{A}_{i-1}$. Hence, we have

$$\begin{aligned}||\widetilde{E}_i - \widehat{E}_i||_2 &\leq ||\widetilde{E}_i - (\mathbf{MM}(\widetilde{Q}^H_{12}, \widetilde{B}_{i-1}) + \mathbf{MM}(\widetilde{Q}^H_{22}, -\widetilde{A}_{i-1}))||_2 \\ &\qquad + ||\mathbf{MM}(\widetilde{Q}^H_{12}, \widetilde{B}_{i-1}) - \widetilde{Q}^H_{12}\widehat{B}_{i-1}||_2 + ||\mathbf{MM}(\widetilde{Q}^H_{22}, \widetilde{A}_{i-1}) - \widetilde{Q}^H_{22}\widehat{A}_{i-1}||_2.\end{aligned} \tag{6.56}$$

By Lemma 6.1.5, the first term in this expression can be bounded by

$$\begin{aligned}&\tau||\mathbf{MM}(\widetilde{Q}^H_{12}, \widetilde{B}_{i-1}) + \mathbf{MM}(\widetilde{Q}^H_{22}, -\widetilde{A}_{i-1})||_2 \\ &\qquad \leq \tau\left[||\mathbf{MM}(\widetilde{Q}^H_{12}, \widetilde{B}_{i-1})||_2 + ||\mathbf{MM}(\widetilde{Q}^H_{22}, -\widetilde{A}_{i-1})||_2\right] \\ &\qquad \leq \tau\left[||\widetilde{Q}^H_{12}\widetilde{B}_{i-1}||_2 + \tau||\widetilde{Q}_{12}||_2||\widetilde{B}_{i-1}||_2 + ||\widetilde{Q}^H_{22}\widetilde{A}_{i-1}||_2 + \tau||\widetilde{Q}_{22}||_2||\widetilde{A}_{i-1}||_2\right] \\ &\qquad \leq \tau(1+\tau)^2\left[||\widetilde{B}_{i-1}||_2 + ||\widetilde{A}_{i-1}||_2\right] \\ &\qquad \leq 2\tau(1+\tau)^2 \left|\left|\begin{pmatrix}\widetilde{A}_{i-1}\\ \widetilde{B}_{i-1}\end{pmatrix}\right|\right|_2 \\ &\qquad \leq 2\tau(1+\tau)^2(1+2\tau)^{2i-2} \left|\left|\begin{pmatrix}A\\B\end{pmatrix}\right|\right|_2 ,\end{aligned} \tag{6.57}$$

where the last inequality follows from Lemma 6.3.6. Using (6.54), the remaining terms of (6.56) satisfy the following:

$$\begin{aligned}
&||\mathbf{MM}(\widetilde{Q}_{22}^H, \widehat{A}_{i-1}) - \widetilde{Q}_{22}^H \widehat{A}_{i-1}||_2 \\
&\qquad \le ||\mathbf{MM}(\widetilde{Q}_{22}^H, \widetilde{A}_{i-1}) - \widetilde{Q}_{22}^H \widetilde{A}_{i-1}||_2 + ||\widetilde{Q}_{22}^H \widetilde{A}_{i-1} - \widetilde{Q}_{22}^H \widehat{A}_{i-1}||_2 \\
&\qquad \le \tau ||\widetilde{Q}_{22}||_2 ||\widetilde{A}_{i-1}||_2 + ||\widetilde{Q}_{22}||_2 ||\widetilde{A}_{i-1} - \widehat{A}_{i-1}||_2 \\
&\qquad \le \tau(1+\tau) ||\widetilde{A}_{i-1}||_2 + (1+\tau) ||\widetilde{A}_{i-1} - \widehat{A}_{i-1}||_2 \\
&\qquad \le \left[(1+\tau)^2(1+2\tau)^{2i-2} - (1+\tau)^i\right] \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 .
\end{aligned} \tag{6.58}$$

Putting everything together, we obtain

$$||\widetilde{E}_i - \widehat{E}_i||_2 \le 2\left[(1+\tau)^3(1+2\tau)^{2i-2} - (1+\tau)^i\right] \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 . \tag{6.59}$$

To extend this bound to all of $\widehat{\Delta}_i$, note that

$$\widehat{\Delta}_i = \begin{pmatrix} \widetilde{Q}_{12}^H \widehat{\Delta}_{i-1} & \widehat{E}_i & \widetilde{Q}_{22}^H \widehat{\Delta}_{i-1} \end{pmatrix} \tag{6.60}$$

for the same $\widetilde{Q}_{12}$ and $\widetilde{Q}_{22}$ used above. Hence, applying both (6.46) and (6.59), we have

$$\begin{aligned}
||\widehat{\Delta}_i||_2 &\le \left|\left|\begin{bmatrix} \widetilde{Q}_{12}^H \widehat{\Delta}_{i-1} & \widetilde{Q}_{22}^H \widehat{\Delta}_{i-1} \end{bmatrix}\right|\right|_2 + ||\widehat{E}_i||_2 \\
&\le \left|\left|\begin{bmatrix} \widetilde{Q}_{12}^H & \widetilde{Q}_{22}^H \end{bmatrix}\right|\right|_2 ||\widehat{\Delta}_{i-1}||_2 + ||\widetilde{E}_i||_2 + ||\widetilde{E}_i - \widehat{E}_i||_2 \\
&\le (1+\tau)||\widehat{\Delta}_{i-1}||_2 + 2\left[(8\tau + (1+\tau)^3)(1+2\tau)^{2i-2} - (1+\tau)^i\right] \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 \\
&\le (1+\tau)||\widehat{\Delta}_{i-1}||_2 + 2\left[(1+15\tau)(1+2\tau)^{2i-2} - (1+\tau)^i\right] \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 ,
\end{aligned} \tag{6.61}$$

where we obtain the final inequality via $8\tau + (1+\tau)^3 \le 1 + 15\tau$. Observing $||\widehat{\Delta}_1||_2 = ||\widehat{E}_1||_2 \le 28\tau ||\binom{A}{B}||_2$, (6.61) implies inductively

$$||\widehat{\Delta}_i||_2 \le 2(1+\tau)^i \left[\frac{14\tau}{1+\tau} + (1+15\tau)\sum_{j=1}^{i-1} \frac{(1+2\tau)^{2j}}{(1+\tau)^{j+1}} - (i-1)\right] \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 . \tag{6.62}$$

We therefore conclude,

$$||\widehat{\Delta}_i||_2 \le 2\left(14\tau(1+\tau)^{i-1} + (i-1)\left[(1+15\tau)(1+2\tau)^{2i} - (1+\tau)^i\right]\right) \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 , \tag{6.63}$$

which we obtain by bounding the sum in (6.62) as

$$(1+\tau)^i \sum_{j=1}^{i-1} \frac{(1+2\tau)^{2j}}{(1+\tau)^{j+1}} = \frac{(1+2\tau)^{2i}}{1+\tau} \sum_{j=1}^{i-1} \left[\frac{1+\tau}{(1+2\tau)^2}\right]^{i-j} \leq (i-1)(1+2\tau)^{2i}, \tag{6.64}$$

noting $\frac{1+\tau}{(1+2\tau)^2} < 1$. Combining (6.63) with (6.48) we have a final bound

$$\begin{aligned} ||\Delta_i||_2 &\leq ||\widehat{\Delta}_i||_2 + ||\widehat{\Delta}_i - \Delta_i||_2 \\ &\leq ||\widehat{\Delta}_i||_2 + ||\widehat{M_i} - M_i||_2 \\ &\leq 2\left(14\tau(1+\tau)^{i-1} + (i-1)\left[(1+15\tau)(1+2\tau)^{2i} - (1+\tau)^i\right]\right. \\ &\qquad\qquad \left. + (1+\tau)^i - 1\right) \left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2 . \end{aligned} \tag{6.65}$$

Step Seven: Obtain $\mathring{A}_p, \mathring{B}_p$ by transforming M_p to block upper triangular. When $i = p$, the matrix M_i consists of only one block of the form (6.49). Hence, we have shown so far

$$\widehat{P}_p^H \widehat{P}_{p-1}^H \cdots \widehat{P}_1^H M = \begin{pmatrix} * & * & * \\ -A_p & \Delta_p & B_p \end{pmatrix}, \tag{6.66}$$

where each $\widehat{P}_i$ is unitary, $\Delta_p \in \mathbb{C}^{n\times(2^p-1)n}$ is small, and A_p and B_p are close to our finite-precision outputs $\widetilde{A}_p$ and $\widetilde{B}_p$. Letting $\Pi \in \mathbb{C}^{(2^p+1)n\times(2^p+1)n}$ be the permutation matrix that swaps the blocks of (6.66) containing $-A_p$ and Δ_p, we have constructed an exact, almost-block-QR factorization

$$\widehat{P}_p^H \widehat{P}_{p-1} \cdots \widehat{P}_1^H M\Pi = \begin{pmatrix} * & * & * \\ \Delta_p & -A_p & B_p \end{pmatrix}. \tag{6.67}$$

Equivalently, recalling that the middle $2^p n \times (2^p-1)n$ block of M is $D^p_{(A,B)}$, we have found an exact factorization $\widehat{P}_p^H \cdots \widehat{P}_1^H D^p_{(A,B)} = \binom{*}{\Delta_p}$.

Label the $*$ block of this matrix as $F \in \mathbb{C}^{(2^p-1)n\times(2^p-1)n}$. By Lemma 6.3.3, there exists $S \in \mathbb{C}^{2^p n\times 2^p n}$ such that $I+S$ is unitary, $(I+S)\binom{F}{\Delta_p} = \binom{F'}{0}$, and

$$||S||_2 \leq ||\Delta_p||_2 ||F^{-1}||_2 \leq \frac{||\Delta_p||_2}{\sigma_{\min}(D^p_{(A,B)}) - ||\Delta_p||_2}, \tag{6.68}$$

assuming $\sigma_{\min}(D^p_{(A,B)}) > ||\Delta_p||_2$. Supposing this is the case, let

$$(I+S)\widehat{P}_p\widehat{P}_{p-1}\cdots\widehat{P}_1^H M\Pi = \begin{pmatrix} * & * & * \\ 0 & -\mathring{A}_p & \mathring{B}_p \end{pmatrix} \tag{6.69}$$

and note

$$||\mathring{A}_p - A_p||_2, ||\mathring{B}_p - B_p||_2 \leq ||S||_2||M||_2 \leq \frac{2||\Delta_p||_2}{\sigma_{\min}(D^p_{(A,B)}) - ||\Delta_p||_2}\left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2. \tag{6.70}$$

Combining (6.70) with (6.55), we obtain a final bound

$$\begin{aligned} ||\widetilde{A}_p - \mathring{A}_p||_2 &\leq ||\widetilde{A}_p - A_p||_2 + ||A_p - \mathring{A}_p||_2 \\ &\leq \left[(1+2\tau)^{2p} + (1+\tau)^p - 2 + \frac{2||\Delta_p||_2}{\sigma_{\min}(D^p_{(A,B)}) - ||\Delta_p||_2}\right]\left|\left|\begin{pmatrix} A \\ B \end{pmatrix}\right|\right|_2, \end{aligned} \tag{6.71}$$

which also applies to $||\widetilde{B}_p - \mathring{B}_p||_2$.

Step Eight: Bound τ by enforcing $||\widetilde{A}_p - \mathring{A}_p||_2, ||\widetilde{B}_p - \mathring{B}_p||_2 \leq \delta||\binom{A}{B}||_2$.

Given (6.71), we obtain the desired bound on $||\widetilde{A}_p - \mathring{A}_p||_2$ and $||\widetilde{B} - \mathring{B}_p||_2$ provided each of $(1+2\tau)^{2p} - 1$, $(1+\tau)^p - 1$, and $\frac{2||\Delta_p||_2}{\sigma_{\min}(D^p_{(A,B)}) - ||\Delta_p||_2}$ is at most $\frac{\delta}{3}$. We focus on the latter, since it is the largest. Here, we note that taking $||\Delta_p||_2 \leq \frac{\delta}{9}\sigma_{\min}(D^p_{(A,B)})$ guarantees not only that the bound (6.68) holds but also

$$\frac{2||\Delta_p||_2}{\sigma_{\min}(D^p_{(A,B)}) - ||\Delta_p||_2} \leq \frac{2\delta}{9-\delta} < \frac{\delta}{3}, \tag{6.72}$$

as desired. Appealing to (6.65) and Definition 6.3.1, we obtain $||\Delta_p||_2 \leq \frac{\delta}{9}\sigma_{\min}(D^p_{(A,B)})$ by requiring that each of $14\tau(1+\tau)^{p-1}$, $(p-1)\left[(1+15\tau)(1+2\tau)^{2p} - (1+\tau)^p\right]$, and $(1+\tau)^p - 1$ is bounded by $\frac{\delta}{54\kappa_{\text{IRS}}(A,B,p)}$. Once again, we focus on the largest of these terms, which in this case is $X = (p-1)\left[(1+15\tau)(1+2\tau)^{2p} - (1+\tau)^p\right]$, assuming $p > 1$.

We begin by rewriting X as follows:

$$\begin{aligned} X &= (p-1)(1+2\tau)^{2p}\left[1 + 15\tau - \left(\frac{1+\tau}{(1+2\tau)^2}\right)^p\right] \\ &= (p-1)(1+2\tau)^{2p}\left[1 + 15\tau - \left[1 - \tau\left(\frac{3+4\tau}{(1+2\tau)^2}\right)\right]^p\right]. \end{aligned} \tag{6.73}$$

Since $\tau\left(\frac{3+4\tau}{(1+2\tau)^2}\right) \leq 1$, we can bound X from above via Bernoulli's inequality

$$\begin{aligned} X &\leq (p-1)(1+2\tau)^{2p}\left[1+15\tau-\left[1-p\tau\left(\frac{3+4\tau}{(1+2\tau)^2}\right)\right]\right] \\ &= (p-1)(1+2\tau)^{2p}\tau\left[15+p\left(\frac{3+4\tau}{(1+2\tau)^2}\right)\right] \\ &\leq 3\tau(p-1)(p+5)(1+2\tau)^{2p}, \end{aligned} \tag{6.74}$$

where the last inequality follows by loosely bounding $\frac{3+4\tau}{(1+2\tau)^2} \leq 3$. Finally assuming $(1+2\tau)^{2p} \leq 2$, we obtain a final bound

$$X \leq 6\tau(p-1)(p+5) = 6\tau(p^2+4p-5), \tag{6.75}$$

which implies a criterion on τ:

$$\tau \leq \frac{\delta}{324\kappa_{\mathrm{IRS}}(A,B,p)(p^2+4p-5)}. \tag{6.76}$$

Note that if $p=1$, and therefore $X=0$, we require instead $15\tau \leq \frac{\delta}{18\kappa_{\mathrm{IRS}}(A,B,p)}$ above, which is clearly satisfied by (6.76). It is similarly not hard to show that this requirement on τ guarantees the remaining bounds and therefore yields $||\widetilde{A}_p - \mathring{A}_p||_2, ||\widetilde{B}_p - \mathring{B}_p||_2 \leq \delta||\binom{A}{B}||_2$.

It remains to show that $\mathring{A}_p$ and $\mathring{B}_p$ can be obtained via exact-arithmetic repeated squaring. This follows from Lemma 6.3.4; exact-arithmetic repeated squaring implies an alternative block-QR factorization of $M\Pi$, which is equivalent to (6.69) up to a rotation/reflection. Since such a rotation/reflection can be baked into the final QR factorization computed by exact-arithmetic repeated squaring (which is agnostic to the specific QR factorizations used), $\mathring{A}_p$ and $\mathring{B}_p$ are indeed exact outputs of repeated squaring satisfying $\mathring{A}_p^{-1}\mathring{B}_p = (A^{-1}B)^{2^p}$. □

Theorem 6.3.7 implies that the number of bits of precision required for **IRS** to compute $\widetilde{A}_p$ and $\widetilde{B}_p$ to within $\delta||\binom{A}{B}||_2$ of a corresponding set of exact outputs is at most

$$\log_2(1/\mathbf{u}) = O\left(\log_2(1/\delta) + \log_2(\mu(n)) + \log_2(\kappa_{\mathrm{IRS}}(A,B,p)) + \log_2(p)\right). \tag{6.77}$$

Its proof provides further intuition for κ_{IRS}; when this condition number is infinite, an approximate block QR factorization of $D^p_{(A,B)}$ cannot be transformed into a true factorization via Lemma 6.3.3, blocking the pathway to $\mathring{A}_p$ and $\mathring{B}_p$.

Taking a step back, we ask: how does this apply to **RPD**? To diagonalize an $n \times n$ pencil to spectral norm accuracy ε, **RPD** applies **IRS** to transformed pencils $(\mathcal{A}, \mathcal{B})$, which satisfy $||\mathcal{A}||_2, ||\mathcal{B}||_2 = O(n^\alpha)$ and $d_{(\mathcal{A},\mathcal{B})} = \Omega(\text{poly}(\varepsilon, n^{-1}))$. Hence, (6.29) implies that κ_{IRS} is at most $O(\text{poly}(n, \varepsilon^{-1}))$ at any point in divide-and-conquer. Recalling from the proof of Proposition 4.2.2 that **RPD** also takes $p = O(\log(\frac{n}{\varepsilon}))$, (6.77) implies that **RPD** requires $O(\log(\frac{n}{\delta\varepsilon}) + \log(\log(\frac{n}{\varepsilon})))$ bits of precision to compute the inputs of **GRURV** to within spectral norm error $\delta > 0$. Referring back to Table 6.1, this is a significant improvement over the corresponding step of Banks et al. [16, Theorem 4.9].

Remark 6.3.8. An alternative to Theorem 6.3.7 can be obtained via repeated application of the QR perturbation bounds from Chapter 3, though the resulting precision requirement is much worse than derived above.

6.4 Two-Matrix GRURV

We close this chapter with a **GRURV** bound. Since **RPD** only applies **GRURV** to products of the form $A_1^{-1}A_2$ and $A_1A_2^{-1}$, it is sufficient[4] to consider the following cases: $\mathbf{GRURV}(2, A_1, A_2, -1, 1)$ and $\mathbf{GRURV}(2, A_1, A_2, 1, -1)$. In terms of our black-box algorithms, the first of these proceeds as follows:

1. $[\widetilde{U}_2, R_2, \widetilde{V}] = \mathbf{RURV}(A_2)$
2. $\widetilde{X} = \mathbf{MM}(A_1^H, \widetilde{U}_2)$
3. $[\widetilde{U}, R_1^H] = \mathbf{QL}(\widetilde{X})$

As in the previous section, we do not use the superscript $\sim$ on R_1 and R_2 since we can guarantee that these are exactly upper triangular.

[4] Bounds for arbitrary products follow naturally from those presented here.

There are three sources of error here: **RURV** in step one, matrix multiplication in step two, and QL in step three. Each of these is fairly straightforward to bound; the latter two are controlled by Assumption 6.1.2 and Assumption 6.1.3, respectively, while finite-precision **RURV** is covered by a lemma of Banks et al. [16, Lemma C.17] (which is itself built on Assumption 6.1.1). Combining these results yields a mixed stability bound for floating-point **GRURV**, which we present below as Theorem 6.4.1. Informally, this implies that (with high probability) the R_1 and R_2 factors computed in floating-point arithmetic belong to an exact rank-revealing factorization of a nearby[5] product $\widehat{A}_1^{-1}\widehat{A}_2$, provided $\mathbf{u}$ is sufficiently small.

For simplicity, we once again state this bound in terms of $\tau = \mu(n)\mathbf{u}$ for $\mu(n)$ as in (6.32). We also assume

$$4c_{\mathrm{N}}\tau||A_2||_2 \leq \frac{1}{4} \leq ||A_2||_2 \text{ and } 1 \leq \min\{\mu_{\mathrm{MM}}(n), \mu_{\mathrm{QR}}(n), c_{\mathrm{N}}\} \tag{6.78}$$

to gain access to the **RURV** bound of Banks et al. [16, Lemma C.17].

Theorem 6.4.1. *Given $A_1, A_2 \in \mathbb{C}^{n\times n}$, let*

$$[\widetilde{U}, R_1, R_2, \widetilde{V}] = \mathbf{GRURV}(2, A_1, A_2, -1, 1)$$

on a floating-point machine with precision $\mathbf{u}$. If $\tau = \mu(n)\mathbf{u}$ for $\mu(n)$ the polynomial (6.32) *and further* (6.78) *holds, then there exist matrices $\widehat{A}_1$, $\widehat{A}_2 \in \mathbb{C}^{n\times n}$ and unitary U, $V \in \mathbb{C}^{n\times n}$, such that $\widehat{A}_1^{-1}\widehat{A}_2 = UR_1^{-1}R_2V$ and*

1. *V is Haar distributed.*
2. *$||\widetilde{U} - U||_2 \leq \tau$.*
3. *$||\widehat{A}_1 - A_1||_2 \leq \tau(\tau^2 + 3\tau^2 + 3)||A_1||_2$.*
4. *For every $\theta \in (0,1)$ and $t > 2\sqrt{2} + 1$, the event that both*

[5] Here, "nearby" is measured by $||\widehat{A}_1 - A_1||_2$ and $||\widehat{A}_2 - A_2||_2$ not $||\widehat{A}_1^{-1}\widehat{A}_2 - A_1^{-1}A_2||_2$.

- $||\widetilde{V} - V||_2 \leq 2\tau \frac{n^{3/2}}{\theta} (4tc_{\mathrm{N}} + 5)$
- $||\widehat{A}_2 - A_2||_2 \leq \tau \left(\frac{n^{3/2}}{\theta} c_{\mathrm{N}} [9t + 10] + 2 \right) ||A_2||_2$

occurs with probability at least $1 - 2e\theta^2 - 2e^{-t^2 n}$.

Proof. We make use of the three-step outline for **GRURV** given above, beginning with $[\widetilde{U}_2, R_2, \widetilde{V}] = \mathbf{RURV}(A_2)$. By [16, Lemma C.17], we know there exist U_2, V and $\widehat{A}_2$ such that U_2 is unitary, V is Haar distributed, and $\widehat{A}_2 = U_2 R_2 V$, where $||\widetilde{U}_2 - U_2||_2$, $||\widetilde{V} - V||_2$, and $||\widetilde{A}_2 - A_2||_2$ can all be bounded. In particular, $||\widetilde{U}_2 - U_2||_2 \leq \tau$. Consequently, Assumption 6.1.2 guarantees that if $\widetilde{X} = \mathbf{MM}(A_1^H, \widetilde{U}_2)$ then

$$||\widetilde{X} - A_1^H \widetilde{U}_2||_2 \leq \tau ||\widetilde{U}_2||_2 ||A_1||_2 \leq \tau(1+\tau)||A_1||_2. \tag{6.79}$$

The final step of **GRURV** computes $[\widetilde{U}, R_1^H] = \mathbf{QL}(\widetilde{X})$. Here, Assumption 6.1.3 implies that there exist matrices U and $\widehat{X}$ such that U is unitary and $\widehat{X} = U R_1^H$, with $||\widetilde{U} - U||_2 \leq \tau$ and $||\widetilde{X} - \widehat{X}||_2 \leq \tau ||\widetilde{X}||_2$. With all of this in mind, let $\widehat{A}_1 = U_2 \widehat{X}^H$. By construction, we have

$$\widehat{A}_1^{-1} \widehat{A}_2 = (U_2 \widehat{X}^H)^{-1} U_2 R_2 V = (R_1 U^H)^{-1} R_2 V = U R_1^{-1} R_2 V, \tag{6.80}$$

which means $U R_1^{-1} R_2 V$ is an exact (generalized) rank-revealing factorization of $\widehat{A}_1^{-1} \widehat{A}_2$.

We can now collect the listed properties/bounds. Items (1) and (4) follow directly from [16, Lemma C.17], while item (2) was derived from Assumption 6.1.3 above. To complete the proof, we observe

$$\begin{aligned}
||\widehat{A}_1 - A_1||_2 &= ||A_1 - U_2 \widehat{X}^H||_2 \\
&= ||A_1^H U_2 - \widehat{X}||_2 \\
&= ||A_1^H U_2 - A_1^H \widetilde{U}_2 + A_1^H \widetilde{U}_2 - \widetilde{X} + \widetilde{X} - \widehat{X}||_2 \\
&\leq ||A_1^H (U_2 - \widetilde{U}_2)||_2 + ||A_1^H \widetilde{U}_2 - \widetilde{X}||_2 + ||\widetilde{X} - \widehat{X}||_2 \\
&\leq [\tau + \tau(1+\tau)] ||A_2||_2 + \tau ||\widetilde{X}||_2,
\end{aligned} \tag{6.81}$$

where $||\widetilde{X}||_2 \leq (1+\tau)^2 ||A_1||_2$ by (6.79). □

Since Theorem 6.4.1 is a bit cumbersome, we present a simplified (i.e., loose) version as Corollary 6.4.2, which can be obtained by taking $\theta = \frac{1}{\sqrt{n}}$ and $t = 4$.

Corollary 6.4.2. *Given $A_1, A_2 \in \mathbb{C}^{n \times n}$, let*

$$[\widetilde{U}, R_1, R_2, \widetilde{V}] = \textbf{GRURV}(2, A_1, A_2, -1, 1)$$

on a floating-point machine with precision $\mathbf{u}$*. Let* $\mu(n)$ *as in* (6.32) *and suppose* (6.78) *holds for* $\tau = \mu(n)\mathbf{u}$*. If for* $\delta \in (0,1)$ *we have*

$$\mathbf{u} = O\left(\frac{\delta}{n^2 \mu(n) c_{\mathrm{N}}}\right)$$

then with probability at least $1 - O(\frac{1}{n})$ *there exist* $\widehat{A}_1, \widehat{A}_2, U, V \in \mathbb{C}^{n \times n}$ *such that* U *is unitary,* V *is Haar distributed,* $\widehat{A}_1^{-1}\widehat{A}_2 = U R_1^{-1} R_2 V$ *and the following bounds hold:*

1. $||\widetilde{U} - U||_2, ||\widetilde{V} - V||_2 \leq O(\delta)$
2. $||\widehat{A}_1 - A_1||_2 \leq O(\delta)||A_1||_2$
3. $||\widehat{A}_2 - A_2||_2 \leq O(\delta)||A_2||_2$.

While these results apply explicitly to $\textbf{GRURV}(2, A_1, A_2, -1, 1)$, they also cover $\textbf{GRURV}(2, A_1, A_2, 1, -1)$. The latter only swaps **RURV** and **QL** for **RULV** and **QR**, which satisfy the same guarantees.

Content Acknowledgement: Portions of this chapter are repurposed from the following submitted works:

- J. Demmel, I. Dumitriu, and R. Schneider. Generalized Pseudospectral Shattering and Inverse-Free Matrix Pencil Diagonalization. arXiv:2306.03700, 2023.
- R. Schneider. When is fast, implicit squaring of $A^{-1}B$ stable? arXiv:2310.00193, 2023.

The dissertation author was the primary investigator and author of both.

Appendix A

Toward a High-Performance Implementation

The formulations of pseudospectral divide-and-conquer presented in this thesis, like the single-matrix version of Banks et al. [16], are primarily theoretical. In an effort to rigorously prove optimal complexity, **EIG**, **RPD**, and even **EIG-DWH** set parameters to extremal values, which are necessary to cover edge cases of the theory but unlikely to be required in practice. With this in mind, we dedicate this appendix to discussing practical modifications to pseudospectral divide-and-conquer, which are aimed at making the algorithm more amenable to high-performance implementation.

We start by restating the observations from Section 4.3: both the n^α scaling and the choice to run divide-and-conquer to subproblems of size 1×1 can realistically be dropped in a practical implementation. In some sense, these are linked; if only a few splits are made, we no longer need our grid g to cover all of $\Lambda_\epsilon(A, B)$ or even $\Lambda(A, B)$, and the scaling was primarily aimed at obtaining bounds for both.

We can next revisit our definition of a successful split. Both **EIG** and **EIG-DWH** require at each step that the chosen split is optimal – i.e., it is at least as close to 50/50 as can be guaranteed to exist. Since the complexity of computing the next subproblems after a split is found is asymptotically the same as checking one grid line/point, divide-and-conquer spends most of its time searching for a split. As a result, requiring optimality is likely more expensive than is worthwhile. Of course, any implementation should still impose some requirements on potential splits to avoid sectioning off only a handful of eigenvalues at each step.

As alluded to in Section 1.3, computing U_L and U_R completely independently is also more expensive than necessary. For our purposes, doing so simplified the analysis of **EIG**, allowing us to control the probability that one of these matrices was computed incorrectly with a straightforward union bound. If the decision is made to compute U_L and U_R independently regardless, we recommended including a check on the approximate rank of each (as computed by **GRURV**) in **DEFLATE**. This can be done essentially for free and provides an easy indication that either **GRURV** failed or **IRS** was not run to high enough accuracy.

Finally, we note a few additional ways randomness may be incorporated in an implementation, which have not yet been explored rigorously. First, a random Möbius transformation can be applied to the input pencil as an alternative to scaling, which can easily be undone and will almost surely wipe out infinite eigenvalues if B is initially singular. Second, the phenomenon displayed in Table 4.1 can be exploited. That is, running the algorithm multiple times with different initial perturbations and averaging eigenvalues may produce better approximations (and identify true eigenvalues for singular pencils, as in the last example of Section 4.3).

Bibliography

[1] B. Adlerborn, B. Kågström, and D. Kressner. A parallel QZ algorithm for distributed memory HPC systems. *SIAM Journal on Scientific Computing*, 36(5):C480–C503, 2014.

[2] L. V. Ahlfors. *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable.* International Series on Pure and Applied Math. McGraw-Hill, Third edition, 1979.

[3] 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.

[4] A. H. Al-Mohy and N. J. Higham. A New Scaling and Squaring Algorithm for the Matrix Exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3):970–989, 2010.

[5] J. Alman and V. V. Williams. A refined laser method and faster matrix multiplication. In *Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '21, page 522–539, USA, 2021. Society for Industrial and Applied Mathematics.

[6] M. Arioli, B. Codenotti, and C. Fassino. The Padé method for computing the matrix exponential. *Linear Algebra and its Applications*, 240, 1996.

[7] Z. Bai, J. Demmel, and M. Gu. An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems. *Numerische Mathematik*, 76:279–308, 1997.

[8] Z. D. Bai. Circular law. *The Annals of Probability*, 25(1):494–529, 1997.

[9] G. Ballard. Avoiding Communication in Dense Linear Algebra. PhD Thesis, 2013.

[10] G. Ballard, E. Carson, J. Demmel, M. Hoemmen, N. Knight, and O. Schwartz. Communication lower bounds and optimal algorithms for numerical linear algebra. *Acta Numerica*, 23:1–155, 2014.

[11] G. Ballard, J. Demmel, and I. Dumitriu. Minimizing Communication for Eigenproblems and the Singular Value Decomposition. Technical Report UCB/EECS-2011-14, EECS Department, University of California, Berkeley, Feb 2011.

[12] G. Ballard, J. Demmel, I. Dumitriu, and A. Rusciano. A generalized randomized rank-revealing factorization. arXiv:1909.06524, 2019.

[13] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing Communication in Numerical Linear Algebra. *SIAM Journal on Matrix Analysis and Applications*, 32(3):866–901, 2011.

[14] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Sequential Communication Bounds for Fast Linear Algebra. Technical Report UCB/EECS-2012-36, EECS Department, University of California, Berkeley, Mar 2012.

[15] J. Banks, J. Garza-Vargas, A. Kulkarni, and N. Srivastava. Overlaps, Eigenvalue Gaps, and Pseudospectrum under real Ginibre and Absolutely Continuous Perturbations. arXiv:2005.08930, 2020.

[16] J. Banks, J. Garza-Vargas, A. Kulkarni, and N. Srivastava. Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time. *Foundations of Computational Mathematics*, pages 1–89, 2022.

[17] J. Banks, A. Kulkarni, S. Mukherjee, and N. Srivastava. Gaussian Regularization of the Pseudospectrum and Davies' Conjecture. *Communications on Pure and Applied Mathematics*, 74:2114–2131, 10 2021.

[18] F. L. Bauer and C. T. Fike. Norms and Exclusion Theorems. *Numerische Mathematik*, 2(1):137–141, 1960.

[19] A. Beavers and E. Denman. A new similarity transformation method for eigenvalues and eigenvectors. *Mathematical Biosciences*, 21(1):143–169, 1974.

[20] P. Benner and R. Byers. Evaluating products of matrix pencils and collapsing matrix products. *Numerical Linear Algebra with Applications*, 8(6-7):357–380, 2001.

[21] P. Benner and R. Byers. An arithmetic for matrix pencils: Theory and new algorithms. *Numerische Mathematik*, 103:539–573, 2006.

[22] R. Bhatia. Pinching, trimming, truncating, and averaging of matrices. *The American Mathematical Monthly*, 107(7):602–608, 2000.

[23] R. Bhatia. *Fourier Series*. Classroom Resource Materials. Mathematical Association of America, 2005.

[24] R. Bhatia and K. Mukherjea. Variation of the Unitary Part of a Matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(3):1007–1014, 1994.

[25] R. Bhattacharjee, G. Dexter, C. Musco, A. Ray, S. Sachdeva, and D. P. Woodruff. Universal Matrix Sparsifiers and Fast Deterministic Algorithms for Linear Algebra. arXiv:2305.05826, 2024.

[26] D. Bini and G. Lotti. Stability of fast algorithms for matrix multiplication. *Numerische Mathematik*, 36:63–72, 1980.

[27] R. P. Brent. Algorithms for matrix multiplication. Technical Report TR-CS-70-157, Department of Computer Science, Stanford, Mar 1970.

[28] A. Y. Bulgakov and S. Godunov. Circular dichotomy of the spectrum of a matrix. *Siberian Mathematical Journal*, 29:734–744, 1988.

[29] T. F. Chan. Rank revealing QR factorizations. *Linear Algebra and its Applications*, 88-89:67–82, 1987.

[30] E. K.-w. Chu. Exclusion theorems and the perturbation analysis of the generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 24(5):1114–1125, 1987.

[31] E. K.-w. Chu. Perturbation of eigenvalues for matrix polynomials via the Bauer-Fike theorems. *SIAM Journal on Matrix Analysis and Applications*, 25(2):551–573, 2003.

[32] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9(3):251–280, 1990.

[33] C. R. Crawford. The Numerical Solution of the Generalized Eigenvalue Problem. PhD Thesis, 1970.

[34] C. R. Crawford. A Stable Generalized Eigenvalue Problem. *SIAM Journal on Numerical Analysis*, 13(6):854–860, 1976.

[35] C. R. Crawford and Y. S. Moon. Finding a positive definite linear combination of two Hermitian matrices. *Linear Algebra and its Applications*, 51:37–48, 1983.

[36] E. B. Davies. Approximate Diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1051–1064, 2008.

[37] C. Davis and W. M. Kahan. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[38] J. Demmel, I. Dumitriu, and O. Holtz. Fast linear algebra is stable. *Numerische Mathematik*, 108:59–91, 2007.

[39] J. Demmel, I. Dumitriu, O. Holtz, and R. Kleinberg. Fast matrix multiplication is stable. *Numerische Mathematik*, 106, 2006.

[40] J. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil A–λB — robust software with error bounds and applications. Part I: theory and algorithms. *ACM Trans. Math. Softw.*, 19(2):160–174, 1993.

[41] J. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil A–λB — robust software with error bounds and applications. Part II: software and applications. *ACM Trans. Math. Softw.*, 19(2):175–201, 1993.

[42] J. W. Demmel and N. J. Higham. Stability of block algorithms with fast level-3 BLAS. *ACM Transactions on Mathematical Software*, 18(3):274–291, 1992.

[43] J. W. Demmel and B. Kågström. Computing stable eigendecompositions of matrix pencils. *Linear Algebra and its Applications*, 88-89:139–186, 1987.

[44] F. M. Dopico and V. Noferini. Root polynomials and their role in the theory of matrix polynomials. *Linear Algebra and its Applications*, 584:37–78, 2020.

[45] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, pages 219–249, 2011.

[46] J. A. Duersch and M. Gu. Randomized QR with Column Pivoting. *SIAM Journal on Scientific Computing*, 39(4):C263–C291, 2017.

[47] I. Dumitriu. Eigenvalue Statistics for Beta-Ensembles. PhD Thesis, 2003.

[48] A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

[49] A. Edelman and N. R. Rao. Random matrix theory. *Acta Numerica*, 14:233–297, 2005.

[50] L. Elsner and P. Lancaster. The spectral variation of pencils of matrices. *Journal of Computational Mathematics*, 3(3):262–274, 1985.

[51] L. Elsner and J.-g. Sun. Perturbation theorems for the generalized eigenvalue problem. *Linear Algebra and its Applications*, 48:341–357, 1982.

[52] E. N. Epperly, J. A. Tropp, and R. J. Webber. XTrace: Making the Most of Every Sample in Stochastic Trace Estimation. *SIAM Journal on Matrix Analysis and Applications*, 45(1):1–23, 2024.

[53] T. Ericsson and A. Ruhe. The spectral transformation Lánczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Mathematics of Computation*, 35:1251–1268, 1980.

[54] B. Farrell and R. Vershynin. Smoothed analysis of symmetric random matrices with continuous distributions. *Proceedings of the American Mathematical Society*, 144, 12 2012.

[55] B. Ford and G. Hall. The generalized eigenvalue problem in quantum chemistry. *Computer Physics Communications*, 8(5):337–348, 1974.

[56] J. G. F. Francis. The QR Transformation A Unitary Analogue to the LR Transformation—Part 1. *The Computer Journal*, 4(3):265–271, 01 1961.

[57] J. G. F. Francis. The QR Transformation—Part 2. *The Computer Journal*, 4(4):332–345, 01 1962.

[58] V. Fraysse, M. Gueury, F. Nicoud, and V. Toumazou. Spectral portraits for matrix pencils. 1996.

[59] G. Galli and M. Parrinello. Large scale electronic structure calculations. *Phys. Rev. Lett.*, 69:3547–3550, 1992.

[60] J. D. Garduber and A. J. Laub. A generalization of the matrix-sign-function solution for algebraic Riccati equations. *International Journal of Control*, 44(3):823–832, 1986.

[61] A. Gopal and P.-G. Martinsson. The PowerURV algorithm for computing rank-revealing full factorizations. arXiv:1812.06007, 2018.

[62] R. M. Gower and P. Richtárik. Randomized Iterative Methods for Linear Systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[63] A. Greenbaum. Personal communication.

[64] A. Greenbaum, R.-C. Li, and M. L. Overton. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM Review*, 62(2):463–482, 2020.

[65] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, Third edition, 2018.

[66] 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.

[67] M. R. Guarracino, C. Cifarelli, O. Seref, and P. M. Pardalos. A classification method based on generalized eigenvalue problems. *Optimization Methods and Software*, 22(1):73–81, 2007.

[68] C.-H. Guo, N. J. Higham, and F. Tisseur. An Improved Arc Algorithm for Detecting Definite Hermitian Pairs. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1131–1151, 2010.

[69] J. Haddock, D. Needell, E. Rebrova, and W. Swartworth. Quantile-Based Iterative Methods for Corrupted Systems of Linear Equations. *SIAM Journal on Matrix Analysis and Applications*, 43(2):605–637, 2022.

[70] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, volume 14. 1996.

[71] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[72] D. Heller. A Survey of Parallel Algorithms in Numerical Linear Algebra. *SIAM Review*, 20(4):740–777, 1978.

[73] N. J. Higham. Exploiting fast matrix multiplication within the level 3 BLAS. *ACM Trans. Math. Softw.*, 16(4):352–368, Dec 1990.

[74] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Second edition, 2002.

[75] N. J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Review*, 51(4):747–764, 2009.

[76] J.-W. Hong and H. T. Kung. I/O complexity: The red-blue pebble game. In *Proceedings of the Thirteenth Annual ACM Symposium on Theory of Computing*, STOC '81, page 326–333, New York, NY, USA, 1981. Association for Computing Machinery.

[77] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Second edition, 2012.

[78] J. L. Howland. The sign matrix and the separation of matrix eigenvalues. *Linear Algebra and its Applications*, 49:221–232, 1983.

[79] Y. Hua and T. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824, 1990.

[80] M. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990.

[81] D. Irony, S. Toledo, and A. Tiskin. Communication lower bounds for distributed-memory matrix multiplication. *Journal of Parallel and Distributed Computing*, 64(9):1017–1026, 2004.

[82] L. Kaufman. Some Thoughts on the QZ Algorithm for Solving the Generalized Eigenvalue Problem. *ACM Trans. Math. Softw.*, 3(1):65–75, 1977.

[83] C. Kenney and A. J. Laub. Rational Iterative Methods for the Matrix Sign Function. *SIAM Journal on Matrix Analysis and Applications*, 12(2):273–291, 1991.

[84] B. Kågström and D. Kressner. Multishift variants of the QZ algorithm with aggressive early deflation. *SIAM Journal on Matrix Analysis and Applications*, 29(1):199–227, 2007.

[85] L. Kronecker. *Algebraische Reduction der Schaaren bilinearer Formen.* Sitzungsberichte der Preussischen Akademie der Wissenschaften. 1890.

[86] V. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *USSR Computational Mathematics and Mathematical Physics*, 1(3):637–657, 1962.

[87] R. Kyng, Y. T. Lee, R. Peng, S. Sachdeva, and D. A. Spielman. Sparsified Cholesky and multigrid solvers for connection laplacians. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 842–850. Association for Computing Machinery, 2016.

[88] H. Li and Y. Wei. Improved rigorous perturbation bounds for the LU and QR factorizations. *Numerical Linear Algebra with Applications*, 22(6):1115–1130, 2015.

[89] R.-C. Li. On Perturbations of Matrix Pencils with Real spectra. *Mathematics of Computation*, 62(205):231–265, 1994.

[90] M. Lotz and V. Noferini. Wilkinson's bus: Weak condition numbers, with an application to singular polynomial eigenproblems. *Foundations of Computational Mathematics*, 20, 2020.

[91] A. N. Malyshev. Computing invariant subspaces of a regular linear pencil of matrices. *Siberian Mathematical Journal*, 30:559–567, 1989.

[92] A. N. Malyshev. Guaranteed accuracy in spectral problems of linear algebra. *Trudy Instituta Matematiki Sibirskogo Otdeleniya AN SSSR*, 17:19–104, 1990.

[93] A. N. Malyshev. Guaranteed accuracy in spectral problems of linear algebra. I. *Siberian Advances in Mathematics*, 2(1):144–197, 1992.

[94] A. N. Malyshev. Guaranteed accuracy in spectral problems of linear algebra. II. *Siberian Advances in Mathematics*, 2(2):153–204, 1992.

[95] A. N. Malyshev. Parallel algorithm for solving some spectral problems of linear algebra. *Linear Algebra and its Applications*, 188-189:489–520, 1993.

[96] V. A. Mandelshtam and H. S. Taylor. Harmonic inversion of time signals and its applications. *The Journal of Chemical Physics*, 107(17):6756–6769, 11 1997.

[97] O. Mangasarian and E. Wild. Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):69–74, 2006.

[98] P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[99] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff. Hutch++: Optimal Stochastic Trace Estimation. In *2021 Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155.

[100] N. Minami. Local fluctuation of the spectrum of a multidimensional Anderson tight binding model. *Communications in Mathematical Physics*, 177(3):709–725, 1996.

[101] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[102] C. B. Moler and G. W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM Journal on Numerical Analysis*, 10(2):241–256, 1973.

[103] R. Murray, J. Demmel, M. W. Mahoney, N. B. Erichson, M. Melnichenko, O. A. Malik, L. Grigori, P. Luszczek, M. Derezinski, M. E. Lopes, T. Liang, H. Luo, and J. Dongarra. Randomized Numerical Linear Algebra: A Perspective on the Field with an Eye to Software. Technical Report UCB/EECS-2023-19, EECS Department, University of California, Berkeley, Feb 2023.

[104] Y. Nakatsukasa. Perturbation behavior of a multiple eigenvalue in generalized Hermitian eigenvalue problems. *BIT Numerical Mathematics*, 50:109–121, 2010.

[105] Y. Nakatsukasa, Z. Bai, and F. Gygi. Optimizing Halley's Iteration for Computing the Matrix Polar Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2700–2720, 2010.

[106] 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.

[107] Y. Nakatsukasa and N. J. Higham. Stable and Efficient Spectral Divide and Conquer Algorithms for the Symmetric Eigenvalue Decomposition and the SVD. *SIAM Journal on Scientific Computing*, 35(3):A1325–A1349, 2013.

[108] P. H. Petkov. Componentwise Perturbation Analysis of the QR Decomposition of a Matrix. *Mathematics*, 10(24), 2022.

[109] 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.

[110] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.

[111] R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989.

[112] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):21–es, 2007.

[113] M. Rudelson and R. Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[114] T. Sakurai and H. Sugiura. A projection method for generalized eigenvalue problems using numerical integration. *Journal of Computational and Applied Mathematics*, 159(1):119–128, 2003.

[115] A. H. Sameh and J. A. Wisniewski. A trace minimization algorithm for the generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 19(6):1243–1259, 1982.

[116] 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.

[117] A. Schönhage. Partial and Total Matrix Multiplication. *SIAM Journal on Computing*, 10(3):434–455, 1981.

[118] X. Shi and Y. Wei. A sharp version of Bauer–Fike's theorem. *Journal of Computational and Applied Mathematics*, 236(13):3218–3227, 2012.

[119] A. Sobczyk, M. Mladenović, and M. Luisier. Invariant subspaces and PCA in nearly matrix multiplication time. arXiv:2311.10459, 2024.

[120] D. A. Spielman and S.-H. Teng. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52:76–84, 2009.

[121] G. W. Stewart. On the Sensitivity of the Eigenvalue Problem $Ax = \lambda Bx$. *SIAM Journal on Numerical Analysis*, 9(4):669–686, 1972.

[122] G. W. Stewart. Gershgorin theory for the generalized eigenvalue problem $Ax = \lambda Bx$. *Mathematics of Computation*, 29(130):600–606, 1975.

[123] G. W. Stewart. Perturbation Bounds for the QR Factorization of a Matrix. *SIAM Journal on Numerical Analysis*, 14(3):509–518, 1977.

[124] G. W. Stewart. Pertubation bounds for the definite generalized eigenvalue problem. *Linear Algebra and its Applications*, 23:69–85, 1979.

[125] G. W. Stewart. Updating a rank-revealing ULV decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14(2):494–499, 1993.

[126] G. W. Stewart and J.-g. Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Elsevier Science, 1990.

[127] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13:354–356, 1969.

[128] T. Strohmer and R. Vershynin. A Randomized Kaczmarz Algorithm with Exponential Convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009.

[129] J.-g. Sun. A note on Stewart's theorem for definite matrix pairs. *Linear Algebra and its Applications*, 48:331–339, 1982.

[130] J.-g. Sun. Perturbation bounds for the Cholesky and QR factorizations. *BIT Numerical Mathematics*, 31(2):341–352, 1991.

[131] J.-g. Sun. On perturbation bounds for the QR factorization. *Linear Algebra and its Applications*, 215:95–111, 1995.

[132] T. Tao and V. Vu. Random matrices: the distribution of the smallest singular values. *Geometric and Functional Analysis*, 20, 2009.

[133] L. N. Trefethen. Pseudospectra of matrices. In *Numerical Analysis*, pages 234–266, 1991.

[134] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2020.

[135] J. A. Tropp and R. J. Webber. Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications, 2023.

[136] F. Uhlig. On computing the generalized Crawford number of a matrix. *Linear Algebra and its Applications*, 438(4):1923–1935, 2013.

[137] P. Van Dooren. Reducing subspaces: Definitions, properties and algorithms. In *Matrix Pencils*, pages 58–73. Springer Berlin Heidelberg, 1983.

[138] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[139] R. C. Ward. The combination shift QZ algorithm. *SIAM Journal on Numerical Analysis*, 12(6):835–853, 1975.

[140] D. S. Watkins. Performance of the QZ algorithm in the presence of infinite eigenvalues. *SIAM Journal on Matrix Analysis and Applications*, 22(2):364–375, 2000.

[141] D. S. Watkins. The QR Algorithm Revisited. *SIAM Review*, 50(1):133–145, 2008.

[142] F. Wegner. Bounds on the density of states in disordered systems. *Zeitschrift für Physik B Condensed Matter*, 44:9–15, 1981.

[143] K. Weierstrass. Zur Theorie der bilinearen und quadratischen Formen. *Monatsh. Akad. Wiss. Berline*, pages 310–38, 1867.

[144] J. Wilkinson. Kronecker's canonical form and the QZ algorithm. *Linear Algebra and its Applications*, 28:285–303, 1979.

[145] V. V. Williams. Multiplying Matrices Faster than Coppersmith-Winograd. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '12, page 887–898. Association for Computing Machinery, 2012.

[146] V. V. Williams, Y. Xu, Z. Xu, and R. Zhou. New bounds for matrix multiplication: from alpha to omega. arXiv:2307.07970, 2023.

[147] J. Wishart. The Generalized Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika*, 20A(1/2):32–52, 1928.

[148] Y. I. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Zap. Imp. Akad. Nauk.*, 30:1–59, 1877.