Algorithms and Perturbation Theory for Matrix Eigenvalue Problems and the Singular Value Decomposition

By

YUJI NAKATSUKASA

B.S. (University of Tokyo) 2005
M.S. (University of Tokyo) 2007

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Roland Freund (Chair)

François Gygi

Naoki Saito

Committee in Charge

2011

Contents

Yuji Nakatsukasa
September 2011
Applied Mathematics

Algorithms and Perturbation Theory for Matrix Eigenvalue Problems and the Singular Value Decomposition

Abstract

This dissertation is about algorithmic and theoretical developments for eigenvalue problems in numerical linear algebra.

The first part of this dissertation proposes algorithms for two important matrix decompositions, the symmetric eigenvalue decomposition and the singular value decomposition. Recent advances and changes in computational architectures have made it necessary for basic linear algebra algorithms to be well-adapted for parallel computing. A few decades ago, an algorithm was considered faster if it required fewer arithmetic operations. This is not the case anymore, and now it is vital to minimize both arithmetic and communication when designing algorithms that are well-suited for high performance scientific computing. Unfortunately, for the above two matrix decompositions, no known algorithm minimizes communication without needing significantly more arithmetic. The development of such algorithms is the main theme of the first half of the dissertation. Our algorithms have great potential as the future approach to computing these matrix decompositions.

The second part of this dissertation explores eigenvalue perturbation theory. Besides being of theoretical interest, perturbation theory is a useful tool that plays important roles in many applications. For example, it is frequently employed in the stability analysis of a numerical algorithm, for examining whether a given problem is well-conditioned under perturbation, or for bounding errors of a computed solution. However, there are a number of phenomena that still cannot be explained by existing theory. We make contributions by deriving refined eigenvalue perturbation bounds for Hermitian block tridiagonal matrices and generalized Hermitian eigenproblems, giving explanations for the perturbation behavior of a multiple generalized eigenvalue, presenting refined eigenvector perturbation bounds, and developing new Gerschgorin-type eigenvalue inclusion sets.

Acknowledgments

Roland Freund advised me for this dissertation work and I am grateful to his support, encouragement and sharing mathematical insights. His view towards numerical linear algebra greatly influenced my perspective for research directions.

A number of individuals have given me valuable suggestions and advices during my graduate studies. With Kensuke Aishima I had exciting communications on the dqds algorithm. I thank Zhaojun Bai for many discussions and introducing me to the polar decomposition. François Gygi introduced me to the computational aspects of molecular dynamics. Nick Higham always inspired me with his insights and knowledge, and gave me countless suggestions to a number of my manuscripts. Ren-Cang Li provided his expertise on eigenvalue perturbation theory and discussions with him has always been exciting and illuminating. Beresford Parlett shared his wisdom on tridiagonal and bidiagonal matrices. Efrem Rensi made our office an excellent place to work in and I enjoyed many discussions with him inside and outside the office. Naoki Saito introduced me to Laplacian eigenvalues, and provided help beyond mathematics, always being attentive despite his busy schedule. Françoise Tisseur generously spared her time to go over my manuscripts and presentations and gave me invaluable suggestions. Ichitaro Yamazaki offered great help with coding, especially for the work on dqds. Shao-Liang Zhang introduced me to the field of numerical linear algebra, and has been understanding and welcoming whenever I visited Japan.

Beyond my studies, I would like to extend my special acknowledgment to Celia Davis, who provided me with inestimable support and advices during difficult times.

And last but not least, I would like to thank my family for always being incredibly patient, understanding and supportive. Without their support this work would not have been completed.

This work was partially supported by NSF grant OCI-0749217 and DOE grant DE-FC02-06ER25794.

CHAPTER 1

Introduction

This dissertation is about design and perturbation analysis of numerical algorithms for matrix eigenvalue problems and the singular value decomposition. The most computationally challenging part of many problems in scientific computing is often to solve large-scale matrix equations, the two most frequently arising of which are linear systems of equations and eigenvalue problems. The rapid development of computer technology makes possible simulations and computations of ever larger scale. It is vital that the algorithms are well-suited for the computing architecture to make full use of the computational power, and that we understand the underlying theory to ensure the computed quantities are meaningful and reliable. These two issues are the fundamental motivation for this dissertation work.

For a numerical algorithm to maximize the computational resources, not only does it need to do as few arithmetic operations as possible, it must be suitable for parallel or pipelined processing. In particular, on the emerging multicore and heterogeneous computing systems, communication costs have exceeded arithmetic costs by orders of magnitude, and the gap is growing exponentially over time. Hence it is vital to minimize both arithmetic and communication when designing algorithms that are well-suited for high performance scientific computing.

The asymptotic communication lower bound is analyzed in [**9**], and algorithms that attain these asymptotic lower bounds are said to minimize communication. There has been much progress in the development of such algorithms in numerical linear algebra, and communication-minimizing implementations are now known for many basic matrix operations such as matrix multiplications, Cholesky factorization and QR decomposition [**9**]. However, the reduction in communication cost sometimes comes at the expense of significantly more arithmetic. This includes the existing communication-minimizing algorithms for computing the symmetric eigendecomposition and the singular value decomposition (SVD). These algorithms also suffer from potential numerical instability.

In the first part of this dissertation we propose algorithms for these two decompositions that minimize communication while having arithmetic cost within a factor 3 of that for the most efficient existing algorithms. The essential cost for each of these algorithms is in performing QR decompositions, of which we require no more than 6 for the symmetric eigenproblem, and 12 for the SVD in IEEE double precision arithmetic. We establish backward stability of these algorithms under mild assumptions. Our algorithms perform comparably to conventional algorithms on our preliminary numerical experiments. Their performance is expected to improve significantly on highly parallel computing architectures where communication dominates arithmetic. Therefore the algorithms we propose here have

great potential as future algorithms for the two important matrix decompositions. This is overall the primary contribution of the dissertation.

We also discuss developments in the dqds algorithm for computing singular values of bidiagonal matrices. This topic has been under investigation since the discovery of the LR and QR algorithm in the 1950s and the 60s, and is considered classical numerical linear algebra which is more or less a matured field. We propose a new deflation strategy to speed up the process, and implement a Fortran code that often runs significantly faster than the widely-used LAPACK routine. Moreover, we argue that the new algorithm can be naturally implemented in a pipelined (parallel) fashion. Therefore by our approach we gain two-fold speedups compared with the traditional dqds implementation.

The second part of the dissertation explores perturbation bounds in numerical linear algebra. Perturbation theory is an important tool in numerical analysis, which finds use in stability analysis, error estimates and algorithm developments. For instance, the stability proof of the algorithms we develop in Part 1 depends largely on perturbation bounds for the matrix polar decomposition. Although perturbation theory is a well-studied subject, there are still a number of phenomena that cannot be explained by existing results. In this dissertation we attempt to fill in some of these gaps by making contributions to several topics in eigenvalue perturbation theory. Understanding such phenomena can give us new insights into unexplained behavior of numerical algorithms, providing directions for improvements.

The subjects that we cover include new perturbation bounds for Hermitian block tridiagonal eigenvalue problems and generalized Hermitian definite eigenvalue problems, analysis on perturbation behavior of a multiple generalized eigenvalue, refined perturbation bounds for eigenvectors and eigenspaces of Hermitian matrices, and Gerschgorin-type theory for standard and generalized eigenvalue problems. Among these, perhaps the contribution of most practical significance is the new bound for Hermitian block tridiagonal matrices, because Hermitian (block) tridiagonal matrices arise frequently in applications and during computation, and our theory may lead to an algorithm that computes some eigenvalues accurately from appropriately chosen submatrices that are much smaller than the original matrix.

The dissertation is organized as follows. Chapter 2 is an overview of the dissertation, in which we give the mathematical backgrounds and highlight our contributions. In Chapter 3 we describe our recently-proposed algorithm for computing the polar decomposition. The algorithm minimizes communication and is backward stable, and serves as the fundamental basis for the two following chapters. Then in Chapter 4 describe our efficient, communication-minimizing algorithm for the symmetric eigendecomposition. We prove our that algorithm is backward stable under suitable assumptions. Chapter 5 develops our SVD algorithm, which is also efficient, communication-minimizing and backward stable. In Chapter 6 we take a different, more classical approach to computing the SVD, and consider techniques for speeding up the computation of the singular values of a bidiagonal matrix. This concludes Part 1, the algorithmic developments.

Chapter 7 starts the study on perturbation theory, in which we derive new perturbation bounds for eigenvalues of Hermitian matrices with block tridiagonal structure. Our bounds can be arbitrarily tighter than any known bound. In Chapter 8 we extend well-known eigenvalue perturbation results for standard Hermitian eigenproblems to generalized

Hermitian eigenproblems. In Chapter 9 we investigate the perturbation behavior of a multiple generalized eigenvalue, whose peculiar behavior was long observed but remained an open problem. In Chapter 10 we discuss the perturbation of eigenvectors, and describe an improvement on the famous Davis-Kahan theory. We also present refined bounds for computed approximate eigenvectors obtained via the Rayleigh-Ritz process. Finally, Chapter 11 discusses Gerschgorin-type theorems, in which we derive eigenvalue inclusion sets that are inexpensive to compute and applicable to generalized eigenproblems. Chapter 12 concludes the dissertation by discussing directions for future research.

CHAPTER 2

Overview and summary of contributions

This chapter gives an overview of this dissertation. We collect mathematical backgrounds that will be used in the developments in the sequel, motivate our investigation and summarize our contribution of each subsequent chapter.

2.1. Notations

There is some notation that we use throughout the dissertation, and we collect them below.

We adopt the Householder convention, in which lowercase letters denote vectors and uppercase letters denote matrices. To specify matrix coordinates we use MATLAB notation, in which $V(i, j:k)$ denotes the jth to kth elements of the ith row of V, and $V(:, \text{end})$ is the last column of V. $v(k)$ denotes the kth element of a vector v. A^T denotes the transpose and A^* is the Hermitian conjugate of A.

$\sigma_i(X)$ denotes the ith singular value of X in descending order of magnitude, unless otherwise specified at the beginning of each chapter. $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ denote the smallest and largest singular values of X respectively.

$\lambda_i(A)$ denotes the ith eigenvalue of a square matrix A, where the method of ordering is specified each chapter as necessary.

I_k is the identity matrix of order k, and we omit subscripts when the size is clear from the context.

Regarding vector and matrix norms, $\|\cdot\|_p$ denotes the matrix or vector p-norm ($p = 1, 2, \infty$) [**56**, Ch. 2]. For any m-by-n matrix A, it is defined by

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \tag{2.1}$$

where the vector norm $\|x\|_p$ for any vector $x = [x_1, x_2, \ldots, x_n]^*$ is defined by

$$\|x\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}}. \tag{2.2}$$

An important and frequently used case is when $p = 2$, which yields the spectral (also frequently called the 2–) norm $\|A\|_2 = \sigma_{\max}(A)$. $\|\cdot\|_F$ denotes the matrix Frobenius norm, defined by $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^2}$, where A_{ij} denotes the (i,j)th element of A. $\|\cdot\|$ denotes an arbitrary unitarily invariant norm satisfying $\|VAU\| = \|A\|$ for any unitary matrices U and V. Such norms include the spectral norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_F$. $\kappa_2(A)$ denotes the 2–norm condition number $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \sigma_{\max}(A)/\sigma_{\min}(A)$.

2.2. The symmetric eigendecomposition and the singular value decomposition

An n-by-n real symmetric (or complex Hermitian) matrix A has a full set of eigenvectors that are orthogonal to each other, and has the eigendecomposition

$$A = V\Lambda V^*, \tag{2.3}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix, whose diagonal elements are called the *eigenvalues*. $V = [v_1, \ldots, v_n]$ is a unitary matrix $V^*V = VV^* = I$, and we can see that $Av_i = \lambda_i v_i$, so the ith column of V is the *eigenvector* corresponding to the eigenvalue λ_i. We can write (2.3) as the outer-product expansion

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^*. \tag{2.4}$$

We now consider general rectangular matrices A. Any rectangular matrix $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ has the singular value decomposition (SVD) [**56**]

$$A = U\Sigma V^*, \tag{2.5}$$

where $U \in \mathbb{C}^{m\times n} = [u_1, \ldots, u_n]$ and $V \in \mathbb{C}^{n\times n} = [v_1, \ldots, v_n]$ have orthonormal columns and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is real, where $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ are called the *singular values* of A. (2.5) has the outer-product experssion

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^*. \tag{2.6}$$

The vectors u_i and v_i are called the left and right *singular vectors* corresponding to σ_i respectively. The SVD is a versatile and important decomposition, both practically and theoretically. For example, the notion of the rank of a matrix is best described using the singular values, and the low-rank approximation problem of A can be solved via the SVD [**56**]. Other applications in which the SVD plays a central role include those discussed in [**21, 86, 156**].

2.2.1. Standard algorithms. The standard algorithms for computing the symmetric eigendecomposition and the SVD are based on first reducing the matrix to condensed form (tridiagonal form for the eigenproblem and bidiagonal for the SVD) via Householder transformations.

For the symmetric eigendecomposition, using $2(n-2)$ Householder transformations we introduce zeros in each column except for the tridiagonal entries, as shown below for the case $n = 5$.

$$A = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix} \overset{H_L, H_R}{\rightarrow} \begin{bmatrix} * & * & & & \\ * & * & * & * & * \\ & * & * & * & * \\ & * & * & * & * \\ & * & * & * & * \end{bmatrix} \overset{H_L, H_R}{\rightarrow} \begin{bmatrix} * & * & & & \\ * & * & * & & \\ & * & * & * & * \\ & & * & * & * \\ & & * & * & * \end{bmatrix} \overset{H_L, H_R}{\rightarrow} \begin{bmatrix} * & * & & & \\ * & * & * & & \\ & * & * & * & \\ & & * & * & * \\ & & & * & * \end{bmatrix}.$$

Here H_L and H_R above the ith arrow indicate the application of left and right multiplication by a Householder reflector Q_i^T and Q_i respectively. A and the resulting tridiagonal matrix T

are related by $A = QTQ^T$ where $Q = \prod_{i=1}^{n-2} Q_i$. After we obtain T, we compute the symmetric tridiagonal eigendecomposition $T = V\Lambda V^T$, for which a number of reliable algorithms exist, including the symmetric tridiagonal QR [**127**], divide-and-conquer [**63**] and bisection algorithm [**56**]. Hence we get the eigendecomposition $A = (QV)\Lambda(QV)^T$, whose eigenvalues are the diagonals of Λ and the corresponding eigenvectors are the columns of QV.

Similarly, for computing the SVD we first reduce the matrix to bidiagonal form as follows.

$$A = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \overset{H_L}{\rightarrow} \begin{bmatrix} * & * & * & * \\ & * & * & * \\ & * & * & * \\ & * & * & * \end{bmatrix} \overset{H_R}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & * \\ & * & * & * \\ & * & * & * \end{bmatrix}$$

$$\overset{H_L}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & * \\ & & * & * \\ & & * & * \end{bmatrix} \overset{H_R}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & \\ & & * & * \\ & & * & * \end{bmatrix} \overset{H_L}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & \\ & & * & * \\ & & & * \end{bmatrix} \equiv B.$$

Here H_L and H_R above the ith arrow indicate the application of left and right multiplication by Householder reflectors U_i^T and V_i respectively. A and the resulting bidiagonal matrix B are related by $A = UTV^T$ where $U = \prod_{i=1}^{n-2} U_i$ and $V = \prod_{i=1}^{n-2} V_i$. We then compute the SVD of the bidiagonal matrix $B = U_B \Sigma_B V_B^T$, which can be done in many ways, for example the QR algorithm, the MRRR algorithm (based on dqds and inverse iteration; its recent developments are described in Willem's PhD thesis [**166**]) and divide-and-conquer [**62**]. Hence we get the SVD $A = (UU_B)\Sigma(V_B V)^T$, whose singular values are the diagonals of Σ and the left and right singular vectors are the columns of UU_B and $V_B V$ respectively.

2.3. The polar decomposition

Any rectangular matrix $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$ has a polar decomposition [**75**]

$$A = U_p H, \tag{2.7}$$

where U_p has orthonormal columns $U_p^* U_p = I$ and H is Hermitian positive definite. $H = (A^*A)^{1/2}$ is always unique and U_p is also unique if A has full column rank. In most of this dissertation we deal with full column-rank matrices.

In particular, the polar decomposition has a close connection to the SVD, for if $A = U\Sigma V^*$ is an SVD then $A = (UV^*)(V\Sigma V^*) = U_p H$, where $U_p = UV^*$. In this dissertation we always use the subscript p in U_p to denote the unitary polar factor, to avoid confusion with the orthogonal factor U in the SVD.

The unitary polar factor U_p has the distinguished property that it is the nearest unitary matrix to A, that is, $\|A - U_p\| = \min\{\|A - Q\| : Q^*Q = I\}$ for any unitarily invariant norm [**75**, Ch. 8], and this makes U_p an important quantity in many applications. U_p also plays a role in the orthogonal Procrustes problem $\min_{Q^*Q=I} \|A - BQ\|_F$, for which the unitary polar factor of B^*A is the solution [**75, 59**].

The Hermitian polar factor H plays a role in the nearest positive semidefinite matrix to A, in that $X = ((A + A^*)/2 + H)/2$ is the unique minimizer of $\|A - X\|_F$ over all Hermitian

positive semidefinite matrices [**75**, p.199]. Other applications of the polar decomposition include factor analysis and satellite tracking [**73, 75**].

The most well-known method for computing the unitary polar factor of a nonsingular matrix A is the Newton iteration

$$X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-*}\right), \quad X_0 = A. \tag{2.8}$$

In practice a scaling technique is used for efficiency if A is ill-conditioned, as described in Chapter 3. The main cost of (2.8) is clearly in forming the explicit inverse X_k^{-*}. This is a potential cause of both numerical instability and high communicaiton cost.

Chapter 3 focuses on the computation of the polar decomposition of square nonsingular matrices, in which we propose a new algorithm that is based on QR decompositions and matrix multiplications and hence inverse-free unlike (2.8), therefore minimizes communication. We also prove backward stability of our new algorithm.

Throughout the first part of this dissertation the polar decomposition plays a major role, being the basis for the algorithms for the symmetric eigendecomposition (discussed in Chapter 4) and the SVD (Chapter 5).

2.4. Backward stability of an algorithm

Suppose that one wants $Y = f(X)$ and obtains the computed approximation $\widehat{Y}$. The forward error is defined by the difference between Y and $\widehat{Y}$. In many cases, however, numerically evaluating the forward error of a computed solution is difficult, and performing forward error analysis of an algorithm is complicated. An alternative approach, called *backward* error analysis, is to seek the smallest ΔX such that $\widehat{Y} = f(X + \Delta X)$, which asks the smallest perturbation in the input X such that the computed $\widehat{Y}$ is the exact solution. This process of backward error analysis has had much success in deepening our understanding of algorithm behaviors. An algorithm is said to be *backward stable* if it provides a computed solution that has small backward error $\|\Delta X\|_F \simeq \epsilon \|X\|_F$, where ϵ here indicates a scalar of order machine precision. We refer to Higham's book [**74**] for much more on stability analysis.

Here we discuss what backward stability means in the context of computing the symmetric eigendecomposition, the SVD and the polar decomposition. For the symmetric eigendecomposition $A = V\Lambda V^* \in \mathbb{C}^{n\times n}$, the computed solution $\widehat{V}\widehat{\Lambda}\widehat{V}^*$ needs to be equal to a small perturbation of A, that is,

$$A + \Delta A = \widehat{V}\widehat{\Lambda}\widehat{V}^*, \tag{2.9}$$

where $\|\Delta A\|_F \leq \epsilon \|A\|_F$. Regarding the eigenvector matrix V, the best we can hope for is that it is numerically orthogonal, that is,

$$\frac{\|V^*V - I\|_F}{\sqrt{n}} \simeq \epsilon, \tag{2.10}$$

where $\sqrt{n}$ here is a normalization factor that accounts for the matrix dimension n. The computed result $\widehat{V}\widehat{\Lambda}\widehat{V}^*$ is a backward stable eigendecomposition of A if (2.9) and (2.10) are both satisfied.

Similarly, a computed SVD $\widehat{U}\widehat{\Sigma}\widehat{V}^*$ of $A \in \mathbb{C}^{m\times n}$ is said to be backward stable if the following hold:

$$A + \Delta A = \widehat{U}\widehat{\Sigma}\widehat{V}^*, \ \frac{\|V^*V - I\|_F}{\sqrt{m}} \simeq \epsilon, \ \frac{\|U^*U - I\|_F}{\sqrt{n}} \simeq \epsilon. \tag{2.11}$$

The computed polar factors $\widehat{U}_p, \widehat{H}$ of $A = U_pH$ is a backward stable solution if they satisfy [**75**, p. 209]

$$A + \Delta A = \widehat{U}_p\widehat{H}, \quad H + \Delta H = \widehat{H} \quad \text{and} \quad \widehat{U}_p^*\widehat{U}_p = I + \epsilon, \tag{2.12}$$

where $\|\Delta A\|_F \leq \epsilon\|A\|_F$ and $\|\Delta H\|_F \leq \epsilon\|H\|_F$.

Contributions in Chapters 3, 4 and 5. A problem with the standard algorithms for the symmetric eigendecomposition and the SVD as summarized above is that they do not minimize communication, as analyzed in [**8**]. In particular, as the matrix size grows the dominant cost tends to lie in the first phase of the algorithm in which the matrix is reduced to condensed forms. In [**8**] a technique is described to reduce communication, which minimizes the number of words communicated on a shared-memory machine at the expense of doing a little more arithmetic. However, the algorithm does not minimize the number of messages, and an extension to parallel, message-passing architectures remains an open problem.

In Chapters 3, 4 and 5 we propose algorithms that minimize both communication (in the asymptotic sense, both number of words and messages) and arithmetic (up to a constant factor). In addition, we prove that all the proposed algorithms are backward stable under mild assumptions. We propose our polar decomposition algorithm in Chapter 3. Much of this chapter is based on [**120**].

We propose an algorithm for the symmetric eigendecomposition in Chapter 4 and for the SVD in Chapter 5. The fundamental building block for these algorithms is the computation of the polar decomposition described in Chapter 3.

2.5. The dqds algorithm

In this dissertation we contribute to the computation of the SVD in two different ways. The first is the material of Chapter 5, the development of an efficient and communication-minimizing algorithm, as we just summarized. The second direction, which is the focus of Chapter 6, is to follow and enhance the standard path of reducing the matrix to bidiagonal form and then computing its SVD. For the second step a standard approach is to first invoke the dqds (differential quotient difference with shifts) algorithm proposed by Fernando and Parlett [**45**], the state-of-the-art algorithm for computing the singular values of a bidiagonal matrix B.

Below is a brief description of dqds. For a bidiagonal matrix B, let $B^{(0)} := B$. dqds computes a sequence of matrices $B^{(m)}$ for $m = 1, 2, \ldots$, expressed as

$$B^{(m)} = \begin{bmatrix} \sqrt{q_1^{(m)}} & \sqrt{e_1^{(m)}} & & \\ & \sqrt{q_2^{(m)}} & \ddots & \\ & & \ddots & \sqrt{e_{n-1}^{(m)}} \\ & & & \sqrt{q_n^{(m)}} \end{bmatrix}. \tag{2.13}$$

Below is a pseudocode of the dqds algorithm.

Algorithm 1 The dqds algorithm

Inputs: $q_i^{(0)} = (B(i,i))^2$ $(i = 1, 2, \ldots, n)$; $e_i^{(0)} = (B(i,i+1))^2$ $(i = 1, 2, \ldots, n-1)$

for $m = 0, 1, \ldots$ **do**
  choose shift $s^{(m)} (\geq 0)$
  $d_1^{(m+1)} = q_1^{(m)} - s^{(m)}$
  for $i = 1, \ldots, n-1$ **do**
    $q_i^{(m+1)} = d_i^{(m+1)} + e_i^{(m)}$
    $e_i^{(m+1)} = e_i^{(m)} q_{i+1}^{(m)} / q_i^{(m+1)}$
    $d_{i+1}^{(m+1)} = d_i^{(m+1)} q_{i+1}^{(m)} / q_i^{(m+1)} - s^{(m)}$
  end for
  $q_n^{(m+1)} = d_n^{(m+1)}$
end for

dqds is mathematically equivalent to the Cholesky LR algorithm applied to $B^T B$ with shifts, expressed as

$$(B^{(m+1)})^T B^{(m+1)} = B^{(m)} (B^{(m)})^T - s^{(m)} I, \tag{2.14}$$

where $B^{(m)}$ is the bidiagonal matrix of the form (2.13) obtained after m dqds iterations. It is a classical result ([**165**, p. 546], see also [**45**] and [**1**] for discussions specific to dqds) that as long as $\sqrt{s^{(m)}}$ is chosen to be smaller than $B^{(m)}$'s smallest singular value $\sigma_{\min}(B^{(m)})$ so that the Cholesky decomposition (2.14) exists, the iterate $B^{(m)}$ converges to a diagonal matrix of (shifted) singular values, that is, $q_i^{(m+1)} \to \sqrt{\sigma_i{}^2 - S}$ as $m \to \infty$ for all i, where $S = \sum_{m=0}^{\infty} s^{(m)}$ is the sum of the previously applied shifts. Moreover, the asymptotic convergence rate of the off-diagonal elements is described by

$$\lim_{m \to \infty} \frac{e_i^{(m+1)}}{e_i^{(m)}} = \frac{\sigma_{i+1}{}^2 - S}{\sigma_i{}^2 - S} < 1 \quad \text{for} \quad i = 1, \ldots, n-1. \tag{2.15}$$

Therefore, the convergence of $e_i^{(m)}$ for $1 \leq i \leq n-2$ is linear, while the bottom off-diagonal $e_{n-1}^{(m)}$ converges superlinearly if $\sigma_n^2 - \sum_{m=0}^{\infty} s^{(m)} = 0$. In view of this, practical deflation

strategies, such as that adopted in DLASQ, check if any of the off-diagonal elements, particularly the bottom one, is small enough to be deflated [**2, 132**]. For detailed descriptions and analyses of dqds, see [**1, 45, 132**].

Choosing an effective shift $s^{(m)}$ is not always a trivial matter. Moreover, computing a shift before each dqds iteration limits the execution of dqds to sequential implementations. Specifically, because of the requirement that (2.14) be well-defined, a good shift $s^{(m)}$ can be estimated only after the whole matrix $B^{(m)}$ is available. This means that if one wishes to run simultaneous dqds iterations in parallel (more precisely in a pipelined fashion), then one has to resort to a crude shift or a zero shift, but this significantly retards the convergence of the bottom off-diagonal element $e_{n-1}^{(m)}$, which means the overall dqds is slow.

In Chapter 6 we propose a solution to this problem by using a new deflation strategy based on the technique of aggressive early deflation, which we describe next.

2.6. Aggressive early deflation

Aggressive early deflation introduced by Braman, Byers and Mathias in 2002 [**16**] aims to deflate eigenvalues long before conventional deflating strategies do so by looking for converged eigenvalues in a $k \times k$ deflation window. This is in contrast to conventional deflating strategies, which typically look only at the bottom subdiagonal element (or two consecutive subdiagonals as described in [**46, 47**]). In this section, we briefly review aggressive early deflation, and discuss its specialization to the symmetric tridiagonal case.

2.6.1. Nonsymmetric case. Let H be an irreducible $n \times n$ Hessenberg matrix obtained after several Hessenberg QR iterations. H can be partitioned as

$$H = \begin{array}{c} \\ n-k-1 \\ 1 \\ k \end{array} \begin{array}{c} \begin{array}{ccc} n-k-1 & 1 & k \end{array} \\ \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ 0 & H_{32} & H_{33} \end{bmatrix} \end{array}, \tag{2.16}$$

where $k \geq 1$ is the *window size* for aggressive early deflation. To perform aggressive early deflation one computes a Schur decomposition $H_{33} = VTV^*$ and considers the unitary transformation

$$\begin{bmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & V \end{bmatrix}^* \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ 0 & H_{32} & H_{33} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & V \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} & H_{13}V \\ H_{21} & H_{22} & H_{23}V \\ 0 & t & T \end{bmatrix}, \tag{2.17}$$

where t is a $k \times 1$ vector, referred to as the *spike vector*. In practice, many of the trailing elements of t are negligibly small so that they can be set to zero. If k_ℓ elements of t are set to zero, then the corresponding k_ℓ eigenvalues are effectively deflated. In practice, more deflatable eigenvalues may be found by recomputing the Schur decomposition with a different eigenvalue ordering.

The leading $(n - k_\ell) \times (n - k_\ell)$ submatrix of (2.17) is then reduced to a Hessenberg form, then the process is repeated of applying multiple QR iterations and executing aggressive early deflation.

This process of aggressive early deflation often dramatically improves the performance of the QR algorithm. Kressner [**94**] shows that the process can be regarded as extracting converged Ritz vectors by the Krylov-Schur algorithm described by Stewart [**144**].

Braman et al. [**16**] shows that $|t_\ell|$, the ℓth element of t, has the expression

$$|t_\ell| = \frac{\left|\prod_{i=n-k}^{n-1} h_{i+1,i}\right|}{\left|\prod_{i\neq\ell}(\mu_i - \mu_\ell)\right| |x_{k,\ell}|}, \tag{2.18}$$

where μ_i $(1 \le i \le k)$ is the ith diagonal of T and $x_{k,\ell}$ is the last element of the eigenvector x corresponding to μ_ℓ. (2.18) partially explains why $|t_\ell|$ can be negligibly small even when none of the subdiagonal elements $h_{i+1,i}$ is.

2.6.2. Symmetric case. Since aggressive early deflation is so effective for the Hessenberg QR algorithm, a similar improvement can be expected in the symmetric tridiagonal case. Here we consider aggressive early deflation applied to the symmetric tridiagonal QR algorithm. Let A be a symmetric tridiagonal matrix, defined by the diagonal elements a_i and off-diagonals b_i, that is,

$$A = \mathrm{tridiag}\left\{ \begin{array}{ccccccccccc} & b_1 & & b_2 & & . & & b_{n-2} & & b_{n-1} & \\ a_1 & & a_2 & & . & & . & & a_{n-1} & & a_n \\ & b_1 & & b_2 & & . & & b_{n-2} & & b_{n-1} & \end{array} \right\}. \tag{2.19}$$

The off-diagonals b_i are assumed to be positive without loss of generality.

Let $A_2 = VDV^T$ be an eigendecomposition of A's lower-right $k \times k$ submatrix A_2, where the diagonals of D are in decreasing order of magnitude. Then, we have

$$\begin{bmatrix} I & \\ & V \end{bmatrix}^T A \begin{bmatrix} I & \\ & V \end{bmatrix} = \begin{bmatrix} A_1 & u_{n-k}t^T \\ tu_{n-k}^T & D \end{bmatrix}, \tag{2.20}$$

where A_1 is the upper-left $(n-k) \times (n-k)$ submatrix of A, $u_{n-k} = [0, 0, \ldots, 1]^T \in \mathbb{R}^{(n-k)\times 1}$ and the spike vector $t = [t_1, \ldots, t_k]^T$ is given by $t = b_{n-k}V(1,:)^T$. If k_ℓ elements of t are smaller than a tolerance τ, for example $\tau = \epsilon\|A\|_2$, then Weyl's theorem [**127**] ensures that the k_ℓ corresponding diagonal elements of D approximate the eigenvalues of A with errors bounded by τ. Hence, we deflate these elements as converged eigenvalues and obtain the symmetric matrix of size $n - k_\ell$ of the form

$$\begin{bmatrix} A_1 & u_{n-k}\tilde{t}^T \\ \tilde{t}u_{n-k}^T & D \end{bmatrix},$$

where $\tilde{t} = [t_1, \ldots, t_{k-k_\ell}]^T$ and $\tilde{D} = \mathrm{diag}(d_1, \ldots, d_{k-k_\ell})$. Now, the bottom-right $(k - k_\ell + 1) \times (k - k_\ell + 1)$ arrowhead matrix needs to be tridiagonalized before we proceed to the next QR iteration. This tridiagonalization can be done in $O(k^2)$ flops by the algorithms in [**123, 169**].

Contrary to the nonsymmetric case, in the symmetric case there is no need to consider another eigendecomposition of A_2 with a different eigenvalue ordering, because it does not change the number of deflatable eigenvalues. The QR algorithm is known to be backward stable, although it can be forward unstable as shown by Parlett and Le [**131**]. In the symmetric case the backward stability of the QR algorithm implies the computed eigenvalues are correct to $\epsilon\|A\|_2$, so they are accurate in the absolute sense. For bidiagonal matrices the

dqds algorithm computes singular values with high relative accuracy, so in our algorithm development in Chapter 6 we ensure relative accuracy is maintained when aggressive early deflation is incorporated into dqds.

We give a separate treatment of aggressive early deflation for the symmetric tridiagonal QR algorithm in Chapter 7, in which we show that many elements of t are negligible if A is in the asymptotic, nearly-diagonal form.

Contributions in Chapter 6. In Chapter 6 we discuss new deflation strategies for dqds based on aggressive early deflation. We take full advantage of the bidiagonal structure to efficiently carry out the process. Moreover, we address the parallelizability issue of dqds raised at the end of Section 2.5. Specifically, we show that when equipped with aggressive early deflation, dqds can mainly use zero shifts without sacrificing the overall speed. Therefore a parallel, pipelined implementation of dqds becomes possible, which can further speed up the execution time significantly. In addition, the new deflation strategy in itself speeds up the dqds iteration, which we demonstrate by showing that our sequential Fortran codes run faster than the latest LAPACK code. This chapter is based on [**119**].

Eigenvalue perturbation theory

2.7. Hermitian eigenproblems

2.7.1. Max-min and Cauchy interlacing theorems. For a Hermitian matrix A, the ith largest eigenvalue $\lambda_i(A)$ has the Courant-Fischer max-min and min-max characterizations given in the following theorem.

THEOREM 2.1 (Max-min and min-max characterizations).

$$\begin{aligned}\lambda_i(A) &= \max_{\dim(S)=i} \min_{x\in S, \|x\|_2=1} x^T A x \\ &= \min_{\dim(S)=n-i+1} \max_{x\in S, \|x\|_2=1} x^T A x.\end{aligned}$$

For a proof, see for example [**56**, Ch. 8] and [**80**, p. 179].

Using the max-min characterization we can prove Cauchy's interlacing theorem [**80**, p. 186], whose statement is as follows.

Let A_1 be a matrix obtained by deleting the kth rows and columns from A. Then A_1 is Hermitian, and denote its ith largest eigenvalue by $\lambda_i(A_1)$. Then we have

THEOREM 2.2 (Cauchy's interlacing theorem).

$$\lambda_1(A) \geq \lambda_1(A_1) \geq \lambda_2(A) \geq \cdots \geq \lambda_{n-1}(A_1) \geq \lambda_n(A). \tag{2.21}$$

See [**80**, p.186] for a proof. The result can be naturally extended to the case where more than one columns and rows are deleted, see [**127**, Sec. 10.1].

2.7.2. Weyl's theorem. From the max-min characterization also follows Weyl's theorem, which bounds the difference between the ith eigenvalue of two Hermitian matrices.

THEOREM 2.3 (Weyl's theorem). *Let the eigenvalues of the Hermitian matrices A and $A+E$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n$ respectively. Then $|\lambda_i - \widetilde{\lambda}_i| \leq \|E\|_2$ for $i = 1, \ldots, n$.*

Weyl's theorem is a simple and beautiful result that gives a uniform bound for all the perturbed eigenvalues. It is sharp in the sense that for any Hermitian A, there exists E such that $|\lambda_i - \widetilde{\lambda}_i| = \|E\|_2$ is attained for all i.

When the matrix has certain structures, however, the Weyl bound can be a severe overestimate. We next discuss bounds that can be much tighter under certain circumstances.

2.7.3. Quadratic eigenvalue perturbation bounds. There exists a quadratic bound for block-partitioned matrices undergoing off-diagonal perturbation [**109, 127**] that relates the eigenvalues of

$$A = \begin{bmatrix} A_1 & E^* \\ E & A_2 \end{bmatrix} \quad \text{and} \quad \widetilde{A} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \tag{2.22}$$

by

$$|\lambda_i(A) - \lambda_i(\widetilde{A})| \leq \|E\|_2^2/\delta_i \tag{2.23}$$

$$\leq \|E\|_2^2/\delta. \tag{2.24}$$

Here, $\lambda_i(X)$ denotes the ith smallest eigenvalue of a Hermitian matrix X. δ_i measures the "gap" between the spectra of A_1 and A_2, defined by $\delta_i \equiv \min_j |\lambda_i(\widetilde{A}) - \lambda_j(A_2)|$ if $\lambda_i(\widetilde{A}) \in \lambda(A_1)$ and $\delta_i \equiv \min_j |\lambda_i(\widetilde{A}) - \lambda_j(A_1)|$ if $\lambda_i(\widetilde{A}) \in \lambda(A_2)$, and $\delta = \min_i \delta_i = \min_{i,j} |\lambda_i(A_1) - \lambda_j(A_2)|$. Here $\lambda(A_i)$ denotes the spectra (set of eigenvalues) of A_i. (2.24) is of use if some information on $\widetilde{A}$ is available, so that a lower bound of δ or δ_i is known.

In a relatively recent paper [**97**] Li and Li give an elegant improvement to the bound, proving that

$$|\lambda_i(A) - \lambda_i(\widetilde{A})| \leq \frac{2\|E\|_2^2}{\delta_i + \sqrt{\delta_i^2 + 4\|E\|_2^2}} \tag{2.25}$$

$$\leq \frac{2\|E\|_2^2}{\delta + \sqrt{\delta^2 + 4\|E\|_2^2}}. \tag{2.26}$$

These bounds can be shown to be always sharper than both the Weyl bound and the quadratic bounds (2.23), (2.24).

When the perturbation matrix E is small, the above quadratic bounds are often much tighter than linear bounds, such as the bound obtained by Weyl's theorem. Such bounds are of interest for example in the context of the Rayleigh-Ritz process, see Section 2.10.2.

The above quadratic bounds suggest that for highly structured Hermitian matrices undergoing structured perturbation, the eigenvalue perturbation can be shown to be much smaller than the generic Weyl bound. In Chapter 7 we push this observation one step further and consider blockwise perturbation of a Hermitian block tridiagonal matrix. We derive bounds that are much tighter than any of the above bounds under certain conditions, roughly that the spectra of the diagonal blocks are well-separated. This chapter is based on [**116**].

2.8. Generalized eigenvalue problems

Eigenvalue problems of the form $Ax = \lambda Bx$ for general square matrices A, B are called generalized eigenvalue problems (GEP). A pair (λ, x) with $x \neq 0$ such that $Ax = \lambda Bx$ is called an eigenpair of the generalized eigenproblem, and we denote the problem by the pair (A, B), called a matrix pair. When B is nonsingular the eigenvalues are those of the matrix $B^{-1}A$. If $Bx = 0$ and $Ax \neq 0$ we say (A, B) has an infinite eigenvalue with eigenvector x. If there exists x such that $Ax = Bx = 0$ then we say the pair (A, B) is singular. In this dissertation we consider only regular pairs (not singular) unless otherwise stated.

An important special case is when A and B are Hermitian and B is positive definite, in which case the pair (A, B) is called a Hermitian definite pair. This case has many connections with the standard Hermitian eigenvalue problem. For example, a Hermitian definite pair is known to have real eigenvalues [**56**, Ch. 8], and there exists a nonsingular W such that $W^*AW = \Lambda$ is a diagonal matrix of eigenvalues and $W^*BW = I$ [**56**, Sec.8.7]. The ith column of W^{-1} is the eigenvector of the Hermitian definite pair corresponding to the ith eigenvalue, which is ith diagonal of Λ.

In this dissertation when we treat Hermitian definite pairs, we deal only with the case where B is positive definite. This type of problem appears in practice most often, and is frequently simply called a generalized Hermitian eigenvalue problem [**6**, Ch.5]. Note that in

the literature a matrix pair is often called Hermitian definite if $\alpha A + \beta B$ is positive definite for some scalars α and β [**146**, pp.281]. When $\alpha A+\beta B$ is positive definite, we can reduce the problem to the positive definite case $A - \theta(\alpha A + \beta B)$, noting that this pair has eigenvalues $\theta_i = \lambda_i/(\beta + \alpha\lambda_i)$.

To illustrate the strong connection between Hermitian eigenproblems and generalized Hermitian eigenproblems, we now review that the max-min and Cauchy interlacing theorems for Hermitian eigenproblems can be extended to Hermitian definite pairs.

2.8.1. Max-min and Cauchy interlacing theorems for generalized Hermitian eigenvalue problems. Here we denote the eigenvalues of an n-by-n Hermitian definite pair (A, B) by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

The min-max principle [**14, 127, 146**] is often stated for the standard Hermitian eigenvalue case $B = I$, but it can be easily extended to generalized Hermitian definite pairs. Namely,

$$\lambda_j = \min_{\mathbb{S}_{n-j+1}} \max_{x\in\mathbb{S}_{n-j+1}} \frac{x^*Ax}{x^*Bx}, \quad \lambda_j = \max_{\mathbb{S}_j} \min_{x\in\mathbb{S}_j} \frac{x^*Ax}{x^*Bx}, \tag{2.27}$$

where $\mathbb{S}_j$ denotes a j-dimensional subspace of $\mathbb{C}^n$. To see this, denoting by $B^{1/2}$ the unique Hermitian positive definite square root of B [**75**, Ch. 6], let Q_j be a N-by-j matrix whose columns form a B-orthonormal basis of $\mathbb{S}_j$. Then there exists $Q_j^\perp$ such that $[Q_j\ Q_j^\perp]$ is square B-orthonormal, so that

$$[Q_j\ Q_j^\perp]^H A[Q_j\ Q_j^\perp] = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad [Q_j\ Q_j^\perp]^H B[Q_j\ Q_j^\perp] = I_N,$$

where A_{11} is j-by-j. Note that the eigenvalues of the pair (A, B) are equal to those of the Hermitian matrix $[Q_jQ_j^\perp]^H A[Q_jQ_j^\perp]$. Then by the standard max-min principle we have

$$\lambda_j \geq \lambda_{\min}(A_{11}) = \min_{x\in\mathrm{span}\{Q_j\}} \frac{x^*Ax}{x^*Bx} = \min_{x\in\mathbb{S}_j} \frac{x^*Ax}{x^*Bx}.$$

Since this inequality holds for any $\mathbb{S}_j$, and equality is attained when $\mathbb{S}_j$ spans the B-orthonormal eigenvectors corresponding to the j largest eigenvectors of $A-\lambda B$, the max-min principle is proved.

In the same way, we can get the min-max principle in (2.27). In particular, (2.27) gives

$$\lambda_1 = \max_x \frac{x^*Ax}{x^*Bx}, \quad \lambda_n = \min_x \frac{x^*Ax}{x^*Bx}.$$

The Cauchy interlacing property [**127**] can also be extended to Hermitian definite pairs. Let A_1 and B_1 be obtained by deleting the kth rows and columns from both A and B, respectively. Then A_1 and B_1 are still Hermitian and B_1 is still positive definite. Denote the eigenvalues of (A_1, B_1) by

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n-1}.$$

Then by (2.27) and the same argument as that in the proof for the standard case [**80**, p.186], one can prove that

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \cdots \geq \lambda_j \geq \mu_j \geq \lambda_{j+1} \geq \cdots \geq \mu_{n-1} \geq \lambda_n. \tag{2.28}$$

Following the this path of extending results for standard Hermitian eigenvalue problems to Hermitian definite pairs, in Chapter 8 we derive eigenvalue perturbation bounds for Hermitian definite pairs. In particular, we derive Weyl-type linear bounds for the generic case and quadratic bounds for the block-diagonal case. This chapter is based on [**104, 113**].

2.9. Eigenvalue first-order expansion

2.9.1. Simple eigenvalue. Let A and E be n-by-n Hermitian matrices. Denote by $\lambda_i(t)$ the ith eigenvalue of $A + tE$, and define the vector-valued function $x(t)$ such that $(A + tE)x(t) = \lambda_i(t)x(t)$ where $\|x(t)\|_2 = 1$ for some $t \in [0, 1]$. If $\lambda_i(t)$ is simple, then

$$\frac{\partial \lambda_i(t)}{\partial t} = x(t)^* E x(t). \tag{2.29}$$

Analogously, for the non-Hermitian case, let $\lambda(t)$ be a simple eigenvalue of a non-Hermitian matrix A with left and right normalized eigenvectors $y(t)^*$ and $x(t)$. Then

$$\frac{\partial \lambda(t)}{\partial t} = \frac{y(t)^* E x(t)}{y(t)^* x(t)}, \tag{2.30}$$

so $1/|y(t)^*x(t)|$ can be considered the condition number of $\lambda(t)$. One way to derive (2.30) is to use Gerschgorin's theorem together with a diagonal similarity transformation, see [**146**, p. 183]. Another approach is to differentiate $(A + tE)x(t) = \lambda(t)x(t)$ with respect to t and left-multiply $y(t)^*$ [**56**, p. 323]

For a generalized eigenvalue problem $Ax = \lambda Bx$, let (x, λ) be a simple eigenpair such that there exist nonsingular matrices $X = (x, X_2)$ and $Y = (y, Y_2)$ that satisfy

$$Y^*AX = \begin{bmatrix} \lambda & 0 \\ 0 & J_A \end{bmatrix}, \quad Y^*BX = \begin{bmatrix} 1 & 0 \\ 0 & J_B \end{bmatrix}, \tag{2.31}$$

where the pair (J_A, J_B) does not have an eigenvalue equal to λ. Then the perturbed pair $(A + tE, B + tE)$ has an eigenpair $(\lambda(t), x(t))$ such that $(A + tE)x(t) = \lambda(t)(B + tF)x(t)$ with $\lambda(0) = \lambda$ and

$$\left.\frac{\partial \lambda(t)}{\partial t}\right|_{t=0} = \frac{y^*(E - \lambda F)x}{y^*Bx}. \tag{2.32}$$

2.9.2. Multiple eigenvalue. The above results have a natural extension to a multiple eigenvalue. For an n-by-n matrix pair (A, B), suppose that λ_0 is a nondefective finite multiple eigenvalue of multiplicity r (we discuss the infinite and defective cases later in Section 9.6.3), so that there exist nonsingular matrices $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ with $X_1, Y_1 \in \mathbb{C}^{n\times r}$ that satisfy

$$Y^*AX = \begin{bmatrix} \lambda_0 I_r & 0 \\ 0 & J_A \end{bmatrix}, \quad Y^*BX = \begin{bmatrix} I_r & 0 \\ 0 & J_B \end{bmatrix}. \tag{2.33}$$

Here the spectrum of the pair (J_A, J_B) does not contain λ_0. Then, the pair $(A+\epsilon E, B+\epsilon F)$ has eigenvalues $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_r$ admitting the first order expansion [**95, 112**]

$$\widetilde{\lambda}_i = \lambda_0 + \lambda_i(Y_1^*(E - \lambda_0 F)X_1)\epsilon + o(\epsilon), \quad i = 1, 2, \ldots, r. \tag{2.34}$$

In Chapter 9 we deal with perturbation of a multiple eigenvalue of a Hermitian definite pair, which exhibits an interesting behavior. In particular, it is observed by Stewart and Sun [**146**, p. 300] that a generalized multiple eigenvalue generally split into r simple eigenvalues when perturbation is applied, each of which having different sensitivities (unlike in the standard Hermitian case). This has remained an open problem. We explain this phenomenon in two ways. We first derive r different perturbation bounds for a multiple eigenvalue of multiplicity r when a Hermitian definite pair (A, B) is perturbed to $(A+E, B+F)$. We then analyze the condition numbers, defined based on (2.34), and show that they take different values when $B \neq I$. This chapter is based on [**114, 117**].

2.10. Perturbation of eigenvectors

The CS decomposition [**124, 141**] is a fundamental basis for eigenvector perturbation theory.

THEOREM 2.4. *For any unitary matrix Q and its 2-by-2 partition $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ where $Q_{11} \in \mathbb{C}^{k\times\ell}$, there exist $U_1 \in \mathbb{C}^{k\times k}$, $U_2 \in \mathbb{C}^{(n-k)\times(n-k)}$, $V_1 \in \mathbb{C}^{\ell\times\ell}$ and $V_2 \in \mathbb{C}^{(n-\ell)\times(n-\ell)}$ such that*

$$\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}^* \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} = \left[\begin{array}{ccc|ccc} I & & & 0 & & \\ & C & & & -S & \\ & & 0 & & & I \\ \hline 0 & & & I & & \\ & S & & & C & \\ & & I & & & 0 \end{array}\right]. \tag{2.35}$$

Here $C = \operatorname{diag}(\cos\theta_1, \ldots, \cos\theta_p)$ and $S = \operatorname{diag}(\sin\theta_1, \ldots, \sin\theta_p)$ for $0 \le \theta_1 \le \cdots \le \theta_p$, in which $p = \min\{k, n-k, \ell, n-\ell\}$.

Now note that any product of two unitary matrices W and V is also unitary: $(W^*V)^*(W^*V) = I$. When the CS decomposition is applied to $Q := W^*V = \begin{bmatrix} W_1^* \\ W_2^* \end{bmatrix} [V_1\ V_2]$ where $W_1 \in \mathbb{C}^{n\times k}$ and $V_1 \in \mathbb{C}^{n\times\ell}$, the nonnegative quantities $0 \le \theta_1 \le \cdots \le \theta_p$ are called the canonical angles between W_1 and V_1. We see from (2.35) that the nonzero angles match those between W_2 and V_2.

An important consequence of the CS decomposition is that the norms of the diagonal matrix of canonical angles are equal to those of the (inner) products of the subspaces. For example, we have

$$\|W_2^* V_1\| = \|\sin\angle(V_1, W_1)\|. \tag{2.36}$$

Here $\sin\angle(V_1, W_1)$ denotes a diagonal matrix whose singular values are the tangents of the canonical angles between the matrices V_1 and W_1 with orthonormal columns.

2.10.1. Davis-Kahan $\sin\theta$ and $\tan\theta$ theorems. The $\tan\theta$ and $\sin\theta$ theorems are two of the four main results in the classical and celebrated paper by Davis and Kahan [**29**]. They are a useful tool for examining the quality of an approximate eigenspace.

The statement of the $\tan\theta$ theorem is as follows. Let A be an n-by-n Hermitian matrix, and let $X = [X_1\ X_2]$ where $X_1 \in \mathbb{C}^{n\times k}$ be an exact unitary eigenvector matrix of A so that $X^*AX = \mathrm{diag}(\Lambda_1, \Lambda_2)$ is diagonal. Also let $Q_1 \in \mathbb{C}^{n\times k}$ have orthogonal columns $Q_1^*Q_1 = I_k$, and define the residual matrix

$$R = AQ_1 - Q_1A_1, \quad \text{where} \quad A_1 = Q_1^*AQ_1. \tag{2.37}$$

The eigenvalues of A_1 are called the Ritz values with respect to Q_1. Suppose that the Ritz values $\lambda(A_1)$ lie entirely above (or below) $\lambda(\Lambda_2)$, the exact eigenvalues corresponding to X_2. Specifically, suppose that there exists $\delta > 0$ such that $\lambda(A_1)$ lies entirely in $[\beta, \alpha]$ while $\lambda(\Lambda_2)$ lies entirely in $[\alpha+\delta, \infty)$, or in $(-\infty, \beta-\delta]$. Then, the $\tan\theta$ theorem gives an upper bound for the tangents of the canonical angles between Q_1 and X_1,

$$\|\tan\angle(Q_1, X_1)\| \le \frac{\|R\|}{\delta}, \tag{2.38}$$

where $\|\cdot\|$ denotes any unitarily invariant norm. $\tan\angle(Q_1, X_1)$ is a diagonal matrix whose singular values are the tangents of the k canonical angles.

The $\sin\theta$ theorem, on the other hand, asserts the same bound, but in terms of the sine instead of tangent:

$$\|\sin\angle(Q_1, X_1)\| \le \frac{\|R\|}{\delta}. \tag{2.39}$$

We discuss some aspects on the derivation. The proofs of the Davis-Kahan theory are centered around the following fact [**146**, p. 251].

LEMMA 2.1. *Let A and B be square matrices such that $1/\|A^{-1}\| - \|B\| = \delta > 0$. Let C satisfies*

$$AX - XB = C.$$

Then we must have $\|X\| \le \frac{\|C\|}{\delta}$.

Proving this is simply done by $\|C\| \ge \|AX\| - \|BX\| \ge (\sigma_{\min}(A) - \|B\|)\|X\| = \delta\|X\|$.

We now present a proof of the $\sin\theta$ theorem because it is short and the technique is the basis for much of the discussion in Chapter 10. In the equation $R = AQ_1 - Q_1M$, right-multiply X_2^* and use the fact $AX_2 = X_2\Lambda_2$ to get $\Lambda_2X_2^*Q_1 - X_2^*Q_1M = X_2^*R$. Now use Lemma 2.1 with $A := \Lambda_2 - sI$ and $B := M - sI$ for an appropriate scalar s, along with the fact $\|X_2^*Q_1\| = \|\sin\angle(Q_1, X_1)\|$ by (2.36), to get the $\sin\theta$ theorem (2.39).

An important practical use of the $\tan\theta$ and $\sin\theta$ theorems is to assess the quality of an approximation to the partial eigenpairs (Λ_1, X_1) of a large Hermitian matrix A obtained by the Rayleigh-Ritz process, see next subsection.

Let us now compare the $\tan\theta$ (2.38) and $\sin\theta$ (2.39) theorems. The (2.38) is clearly sharper than (2.39), because $\tan\theta \ge \sin\theta$ for any $0 \le \theta < \frac{\pi}{2}$. In particular, for the spectral norm, when $\|R\|_2 > \delta$ (2.39) is useless but (2.38) still provides nontrivial information. However, the $\sin\theta$ theorem holds more generally than the $\tan\theta$ theorem in two respects. First, the bound (2.39) holds with A_1 replaced with any k-by-k Hermitian matrix M (the choice affects δ) and R replaced with $AQ_1 - Q_1M$. The $\tan\theta$ theorem takes $M = Q_1^*AQ_1$, which is a special but important choice because it arises naturally in practice as described

above, and it is optimal in the sense that it minimizes $\|R\|$ for any unitarily invariant norm [**145**, p.252].

Second, and more importantly for the discussion in Chapter 10, the hypothesis on the situation of the spectra of A_1 and Λ_2 is less restrictive in the $\sin\theta$ theorem, allowing the Ritz values $\lambda(A_1)$ to lie on both sides of the exact eigenvalues $\lambda(\Lambda_2)$ corresponding to X_2, or vise versa. Specifically, in addition to the situation described above, the bound (2.39) holds also in either of the two cases:

(a) $\lambda(\Lambda_2)$ lies in $[a, b]$ and $\lambda(A_1)$ lies in the *union* of $(-\infty, a-\delta]$ and $[b+\delta, \infty)$.
(b) $\lambda(A_1)$ lies in $[a, b]$ and $\lambda(\Lambda_2)$ lies in the *union* of $(-\infty, a-\delta]$ and $[b+\delta, \infty)$.

We note that in the literature these two cases have not been treated separately. In particular, as discussed above, the original $\tan\theta$ theorem imposes the Ritz values $\lambda(A_1)$ to lie entirely above (or below) the eigenvalues $\lambda(\Lambda_2)$, allowing neither of the two cases.

We note that for the $\sin\theta$ thoerem, the requirement on the spectrums can be further relaxed so that δ is defined by the nearest distance between the nearest eigenvalues of A_1 and Λ_2, at the expense of a slightly larger constant $\pi/2$. This is discussed in [**15**], [**14**, p. 212].

The first part of Chapter 10 discusses eigenvector perturbation analyses based on the Davis-Kahan theorems. There we show that the $\tan\theta$ theorem holds under more relaxed conditions. Specifically, we show that the condition on the spectra in the $\tan\theta$ theorem can be relaxed, in that the bound (2.38) still holds true in case (a) above, but not in case (b). This work is based on [**118**].

2.10.2. The Rayleigh-Ritz process and theorems of Saad and Knyazev. For a Hermitian matrix A that is large (and sparse) enough to make it infeasible to compute its complete eigendecomposition, we must be content with computing a portion of its eigenpairs. The standard way of doing this is to form a k-dimensional subspace where $k \ll n$, represented by a n-by-k matrix Q_1 with orthonormal columns, which approximately contains the desired eigenvectors, and then extract approximate eigenpairs (called the *Ritz pairs*) from it by means of the *Rayleigh-Ritz process*. The Rayleigh-Ritz process computes the eigendecomposition of a k-by-k Hermitian matrix $Q_1^* A Q_1 = Y \widehat{\Lambda} Y^*$, from which the Ritz values are taken as the diagonals of $\widehat{\Lambda}$ and the Ritz vectors as the columns of $\widehat{X} = Q_1 Y$. The Ritz pairs $(\widehat{\lambda}, \widehat{x})$ thus obtained satisfy

- $\widehat{x} \in \operatorname{span}(Q_1)$,
- $A\widehat{x} - \widehat{\lambda}\widehat{x} \perp Q_1$.

A natural question is to ask how accurate the Ritz pairs are as an approximation to the exact eigenpairs. The accuracy of Ritz values can be bounded by observing that for a unitary matrix $[Q_1\ Q_2]$,

$$[Q_1\ Q_2]^* A [Q_1\ Q_2] = \begin{bmatrix} Q_1^* A Q_1 & Q_1^* A Q_2 \\ Q_2^* A Q_1 & Q_2^* A Q_2 \end{bmatrix}, \tag{2.40}$$

and that for any unitarily invariant norm

$$\|Q_2^* A Q_1\| = \|Q_2^* A Q_1 Y\| = \|Q_2^* (A\widehat{X} - \widehat{X}\Lambda)\| = \|Q_2^* R\| = \|R\|, \tag{2.41}$$

where $R = A\widehat{X} - \widehat{X}\Lambda$ is the residual. Here we used the facts $Q_2^*Q_1 = 0$ and $Q_1^*R = 0$, which follows from the second property of Ritz pairs mentioned above. From these facts, combined with Weyl's theorem, we conclude that the eigenvalues of $Q_1^*AQ_1$ match some of those of A to $\|R\|_2$. For individual Ritz values, the corresponding residual norm $\|A\widehat{x} - \widehat{\lambda}\widehat{x}\|_2$ is a bound for the distance to the closest exact eigenvalue. Moreover, by using one of the quadratic perturbation bounds summarized above in Section 2.7.3 one can often get more accurate bounds for the difference between the eigenvalues of $Q_1^*AQ_1$ and A, see [**97, 109, 127**].

Now we turn to the accuracy of the Ritz vectors. Letting $A_1 = Q_1^*AQ_1$ in (2.40) we can invoke the Davis-Kahan $\tan\theta$ (2.38) or $\sin\theta$ (2.39) theorem to measure the nearness between the whole subspace Q_1 and a set of eigenvectors X_1, provided that some gap information on the spectra of A_1 and Λ_2 is available.

Now we consider comparing individual Ritz vector (or a set of some Ritz vectors) and eigenvector(s). As in the Davis-Kahan theory, we bound the angle between approximate and exact eigenvectors $\widehat{x}$ and x. Saad [**134**] proves the following theorem.

THEOREM 2.5. *Let A be a Hermitian matrix and let (λ, x) be any of its eigenpairs. Let $(\widehat{\lambda}, \widehat{x})$ be a Ritz pair obtained from the subspace span(Q_1), such that $\widehat{\lambda}$ is the closest Ritz value to λ. Suppose $\delta > 0$, where δ is the distance between λ and the set of Ritz values other than $\widehat{\lambda}$. Then*

$$\sin\angle(x, \widehat{x}) \leq \sin\angle(x, Q_1)\sqrt{1 + \frac{\|R\|_2^2}{\delta^2}},$$

where $R = A\widehat{X} - \widehat{X}\Lambda$, so that (2.41) holds.

We note that $\sin\angle(x, \widehat{x}) \leq \sin\angle(x, Q_1)$ because $x \in \text{span}(Q_1)$, so the above theorem claims that the Ritz vector is optimal up to a constant $\sqrt{1 + \frac{\|R\|_2^2}{\delta^2}}$. This does not mean, however, that performing the Rayleigh-Ritz process is always a reliable way to extract approximate eigenpairs from a subspace; care is needed especially when computing interior eigenpairs, such as using the Harmonic Rayleigh-Ritz process. For more on this issue see, for example, [**48, 111, 137, 145**].

We note that in [**143, 145**] Stewart presents a generalization of Saad's theorem to the non-Hermitian case.

In [**90**, Thm. 4.3] Knyazev derives a generalization of Saad's result to the subspace case, in which $x, \widehat{x}$ are replaced with sets of eigenvectors and Ritz vectors. Knyazev gives the bound for linear operators, but here we specialize to matrices.

THEOREM 2.6. *Let A be a Hermitian matrix. Let $(\widehat{\Lambda}_1, \widehat{X}_1)$ be a set of k_1 ($< k$) Ritz pairs obtained from the k-dimensional subspace span(Q_1). Also let (Λ_1, X_1) be a set of k_1 exact eigenpairs, whose eigenvalues lie in the interval $[\lambda - d, \lambda + d]$ for some λ and $d > 0$. Suppose that $\delta = \min|diag(\widehat{\Lambda}_2) - \lambda| - d > 0$, where $\widehat{\Lambda}_2$ are the Ritz values corresponding to $\widehat{X}_2$, the orthogonal complement of $\widehat{X}_1$. Then*

$$\sin\angle(X_1, \widehat{X}_1) \leq \sin\angle(X_1, Q_1)\sqrt{1 + \frac{\|R\|_2^2}{\delta^2}}.$$

In the second half of Chapter 10 we derive new bounds for the quality of computed approximate eigenvectors obtained via the Rayleigh-Ritz process that can be much tighter than the bounds by Saad and Knyazev in a situation that arises naturally in large-scale eigenvalue problems. We also consider extensions to the singular value decomposition, deriving analogous results in the context of projection methods (called Petrov-Galerkin) for computing some of the singular triplets. Chapter 10 also introduces what we call the $\cos\theta$ theorem, which asserts the distance instead of nearness between two subspaces. We discuss how we might use the the $\cos\theta$ theorem to efficiently execute an inexact Rayleigh-Ritz process.

2.11. Gerschgorin's theorem

Gerschgorin's theorem [**53**] defines in the complex plane a set including all the eigenvalues for any given square matrix. It is a very well known result and a useful tool that finds use in estimating eigenvalues and perturbation analysis.

THEOREM 2.7 (Gerschgorin's theorem). *All the eigenvalues of an $n \times n$ complex matrix A are contained in $\Gamma(A) \equiv \bigcup_{i=1}^{n} \Gamma_i(A)$, where $\Gamma_i(A)$ is A's ith Gerschgorin disk defined by*

$$\Gamma_i(A) \equiv \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

$\Gamma(A)$ is called the Gerschgorin set of A. Note that it is a union of n disks, each of which can be computed by $O(n)$ arithmetic operations. $\Gamma(A)$ is a useful tool for inexpensively bounding eigenvalues. Owing to its simple expression, it is applied in a variety of applications. We already mentioned some uses of it above, and will discuss more in the following chapters.

The main contribution of Chapter 11 is that we present a Gerschgorin-type set applicable to generalized eigenvalue problems. Such extensions have been considered in the literature but their usage has been limited, due to the use of non-standard metric or high computational complexity. We derive a set that is simple and as inexpensive to compute as $\Gamma(A)$ for standard eigenproblems. This is based on [**115**].

We also present a new eigenvalue inclusion region for standard eigenproblems that is sharper than both $\Gamma(A)$ and the set called Brauer's ovals of Cassini. The computational cost of the new set is generally expensive but becomes lower for highly sparse matrices. Furthermore, unlike many extensions of $\Gamma(A)$, our set is a union of n sets, just like the original $\Gamma(A)$.

Part 1

Algorithms for matrix decompositions

CHAPTER 3

Communication minimizing algorithm for the polar decomposition

In this chapter, we consider the computation of the unitary polar factor U_p of the polar decomposition $A = U_pH$ (recall Section 2.3) of $A \in \mathbb{C}^{m\times n}$ $(m \geq n)$, where $U_p \in \mathbb{C}^{m\times n}$ is a unitary matrix $U_p^*U_p = I$ and $H \in \mathbb{C}^{n\times n}$ is a unique Hermitian positive semidefinite matrix.

We first review the scaled Newton iteration, currently the most widely used method for computing the polar decomposition. We then discuss the Halley iteration, and introduce a dynamically weighted Halley (DWH) iteration. We prove that the new method is globally and asymptotically cubically convergent. For matrices with condition number no greater than 10^{16}, the DWH method needs at most six iterations for convergence with the tolerance 10^{-16}. The Halley iteration can be implemented using QR decompositions and matrix multiplications, without explicit matrix inversions unlike the scaled Newton iteration. We prove that our QR-based algorithm QDWH is backward stabile provided that column pivoting and row sorting (or pivoting) are used for computing the QR decomposition.

Introduction. We discuss algorithms for computing the polar decomposition. We will primarily discuss computing the unitary polar factor U_p, because once once the computed unitary polar factor $\widehat{U}_p = \lim_{k\to\infty} X_k$ is obtained, we compute $\widehat{H}$ [**75**, Sec. 8.8] by

$$\widehat{H} = \frac{1}{2}(\widehat{U}_p^*A + (\widehat{U}_p^*A)^*). \tag{3.1}$$

The unitary polar factor U_p in the polar decomposition $A = U_pH$ can be sensitive to perturbation depending on the smallest two singular values of A [**58, 75**]. By contrast, H is known to be always well-conditioned, in that the Hermitian polar factors of A and $A + \Delta A$ differ by no more than $\sqrt{2}\|\Delta A\|_F$ in Frobenius norm. As we shall see, this property plays an important role in the backward stability proof of our new algorithms we develop.

The most popular method for computing the polar factor of a square nonsingular matrix is the scaled Newton (SN) method [**75**, p. 202]. Recently, Byers and Xu [**20**] presented a suboptimal scaling strategy for the Newton method. They showed that the convergence to within a tolerance of 10^{-16} can be reached in at most nine iterations for matrices with condition number no greater than 10^{16}. Furthermore, they claim that Newton's method with suboptimal scaling is backward stable, provided that the matrix inverses are computed in a mixed forward-backward stable way. We note that there is a recent note [**88**] to indicate some incompleteness of rounding error analysis presented in [**20**].

Successful as Newton's method is, it requires explicit matrix inversion at each iteration. Besides the potential numerical stability issue in finite precision arithmetic, explicit matrix inversion is also expensive in communication costs. On the emerging multicore and heterogeneous computing systems, communication costs have exceeded arithmetic costs by orders of magnitude, and the gap is growing exponentially over time [**9, 61, 138**]. The goal of this chapter is to investigate numerical methods for computing the polar decomposition to minimize the communication costs by using communication friendly matrix operations such as the QR decomposition (without pivoting) [**34**].

In fact, inverse free methods for computing the polar decomposition have been studied in [**26, 19**]. A QR decomposition-based implementation of a variant of the SN method is investigated. Unfortunately, the numerical instability of such an inverse free method has been independently discovered by both studies.

We first propose a dynamically weighted Halley (DWH) method for computing the polar decomposition. We prove that the DWH method converges globally with asymptotically cubic rate. We show that in exact arithmetic, for matrices with condition number $\kappa_2(A) \leq 10^{16}$, no more than six iterations are needed for convergence with the tolerance 10^{-16}. We then discuss an implementation of the DWH method based on the QR decomposition. Extensive numerical tests indicate that the QR-based DWH (QDWH) method is backward stable. The arithmetic cost of the QDWH method is about two to three times that of the SN method, depending on the specific implementation one uses. However, the communication cost of the QDWH method is significantly lower than that of the SN method. The QDWH method is an attractive alternative method for the emerging multicore and heterogeneous computing architectures.

We primarily deal with the polar decomposition of square and nonsingular matrices. The QDWH method is readily applicable to rectangular and singular matrices, whereas the SN method needs to initially use a rank-revealing QR factorization to enable its applicability to more general matrices [**75**, p. 196].

The rest of this chapter is organized as follows. In Section 3.1, we review Newton's method and its variants. In Section 3.2, we study Halley's iteration and derive a dynamical weighting scheme. A global convergence proof of the new method is given. We also show that the cubic convergence makes acceptable a looser convergence criterion than that for the SN iteration. Section 3.3 discusses implementation issues, in which we show how the DWH method can be computed based on the matrix QR decompositions. Numerical examples are shown in Section 3.5. In Section 3.4 we prove that our QR-based algotirhm is backward stable, provided that pivoting is used in computing the QR decompositions. In Section 3.6, we give a detailed solution for the max-min problem that arises in the derivation of the DWH method. In Section we describe our original motivation for studying the polar decomposition, an orthogonal Procrustes problem arising from the subspace alignment process in first-principles molecular dynamics simulations of electronic structure calculations [**4, 44, 66, 67**]. We include this section because this application of the polar decomposition seems not to be stated explicitly in the literature.

Throughout this and the next two chapters, ϵ denotes a matrix or scalar whose norm is a small multiple of the machine epsilon. Whether a specific ϵ represents a matrix or a scalar

should be clear from the context. Following [**127**, p. 111] we allow ϵ to take diffent values in different appearances, so in particular we shall freely use identities such as $\epsilon + \epsilon = \epsilon$. Unless otherwise stated, $\widehat{A}$ denotes a computed version of A.

3.1. Newton-type methods

3.1.1. Scaled Newton iteration. The most well-known method for computing the unitary polar factor of a nonsingular matrix A is the Newton iteration

$$X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-*}\right), \quad X_0 = A. \tag{3.2}$$

It can be shown that the iterates X_k converge quadratically to the polar factor U_p of A and that all singular values $\sigma_i(X_k) \to 1$ as $k \to \infty$ [**75**, p. 202]. However, the initial phase of convergence is slow when A has a singular value that has large relative distance from 1, that is, when a singular value σ exists such that $\max(|1-\sigma|/\sigma, |1-\sigma|/1) \gg 1$. In order to speed up the initial phase, we can apply the SN iteration

$$X_{k+1} = \frac{1}{2}\left(\zeta_k X_k + (\zeta_k X_k)^{-*}\right), \quad X_0 = A, \tag{3.3}$$

where ζ_k is a scaling factor. The frequently used $(1,\infty)$–norm scaling and Frobenius norm scaling are known to work well in practice [**73, 75**]. The rigorous convergence theory is established for the so-called optimal scaling $\zeta_k^{\text{opt}} = (\sigma_{\min}(X_k)\sigma_{\max}(X_k))^{-1/2}$ [**73**]. However, it is not a practical scaling since it is too expensive to compute $\sigma_{\min}(X_k)$ and $\sigma_{\max}(X_k)$ at every iteration. Recently, Byers and Xu [**20**] proposed the following suboptimal scaling:

$$\zeta_0 = 1/\sqrt{\alpha\beta}, \quad \zeta_1 = \sqrt{2\sqrt{\alpha\beta}/(\alpha+\beta)}\,, \quad \zeta_k = 1/\sqrt{\rho(\zeta_{k-1})} \quad \text{for} \quad k = 2,3,\ldots, \tag{3.4}$$

where $\alpha = \|A\|_2$, $\beta = \|A^{-1}\|_2^{-1}$ and $\rho(\zeta) = (\zeta + \zeta^{-1})/2$. It is called a suboptimal scaling since at the kth iteration, it minimizes the width of the interval containing all the singular values of the kth iterate X_k.

THEOREM 3.1 (see [**20**]). *The iterates X_k generated by the SN iteration* (3.3) *with the suboptimal scaling* (3.4) *converge quadratically to the unitary polar factor U_p of A. Moreover, convergence to within a tolerance 10^{-16} is reached within nine iterations if $\kappa_2(A) \le 10^{16}$.*

The following is a pseudocode for the scaled Newton iteration with BX scaling.

Scaled Newton method:

1: $X_0 = A$ and $X_{-1} + I$.
2: $\zeta_0 = 1/\sqrt{\alpha\beta}$ and $k = 0$
3: **repeat**
4: $\quad X_{k+1} = (\zeta_k X_k + (\zeta_k X_k)^{-*})/2$
5: $\quad$ **if** $k = 0$ **then**
6: $\quad\quad \zeta_1 = \sqrt{2\sqrt{\alpha\beta}/(\alpha+\beta)}$
7: $\quad$ **else**
8: $\quad\quad \zeta_{k+1} = \sqrt{2/(\zeta_k + 1/\zeta_k)}$
9: $\quad$ **end if**
10: $\quad k = k + 1$

11: **until** convergence
12: $U = X_k$

In practice, it is sufficient to use some rough estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β. For example, one may take $\widehat{\alpha} = \|A\|_F$ and $\widehat{\beta} = 1/\|A^{-1}\|_F$. It is shown that, for any estimates $\widehat{\alpha}$ and $\widehat{\beta}$ such that $0 < \widehat{\beta} \leq \|A^{-1}\|_2^{-1} \leq \|A\|_2 \leq \widehat{\alpha}$ and $\widehat{\alpha}/\widehat{\beta} < 10^{16}$, the iteration converges within nine iterations [**20**]. It is also known experimentally that the SN iteration with Higham's $(1, \infty)-$scaling [**73**] needs about the same number of iterations.

It is claimed in [**89, 20**] that the SN iteration is backward stable provided that the inverse X_k^{-1} is computed in a mixed forward-backward stable way. For example, one can use a bidiagonal reduction-based matrix inverse algorithm as presented in [**20**]. In that case, the arithmetic cost of each iteration increases to $6n^3$ instead of $2n^3$ when the inverse is computed using the LU factorization with partial pivoting. We note that inversion based on QR factorization without column pivoting does not guarantee backward stability of the SN iteration (see [**89**]).

3.1.2. Newton iteration variant. The Newton variant (NV) iteration is

$$Y_{k+1} = 2Y_k \left(I + Y_k^* Y_k\right)^{-1}, \quad Y_0 = A. \tag{3.5}$$

It can be shown that $Y_k = X_k^{-*}$ for $k \geq 1$, where X_k is generated by the Newton iteration (3.2) [**78**], [**75**, p. 202]. Note that iteration (3.5) is applicable to singular and rectangular matrices. To speed up the convergence, we can use a scaled version of iteration (3.5). Substituting $\eta_k Y_k$ into Y_k in (3.5) yields the scaled Newton variant (SNV) iteration

$$Y_{k+1} = 2\eta_k Y_k \left(I + \eta_k^2 Y_k^* Y_k\right)^{-1}, \quad Y_0 = A, \tag{3.6}$$

where η_k is the scaling factor. A proper choice of η_k is the one such that $Y_0 = X_0$ and $Y_k = X_k^{-*}$ for $k \geq 1$, where X_k is generated by the SN iteration (3.3). It implies that $\eta_0 = \zeta_0$ and $\eta_k = 1/\zeta_k$.

Since Y_k^{-1} is not computed in the SNV iteration (3.6), the $(1, \infty)$–norm scaling or Frobenius norm scaling is not applicable. How to efficiently scale the SNV iteration (3.6) is listed as Problem 8.26 in [**75**, p. 219]. One solution to the problem is to use the suboptimal scaling (3.4), which yields the following iteration for the scaling of the SNV iteration (3.6):

$$\eta_0 = 1/\sqrt{\alpha\beta}, \quad \eta_1 = \sqrt{\frac{\alpha + \beta}{2\sqrt{\alpha\beta}}}, \quad \eta_k = \sqrt{\rho(\eta_{k-1})} \quad \text{for} \quad k = 2, 3, \ldots. \tag{3.7}$$

From the connection with the Newton iteration, it follows from Theorem 3.1 that $Y_k \to U_p^{-*} = U_p$ as $k \to \infty$.

The SN iteration with the suboptimal scaling (3.4) and the SNV iteration with the scaling (3.7) are mathematically equivalent, provided that the same scalars α and β are used. However, the practical implementation of the SN iteration involves explicit matrix inverses. This is usually done by means of the LU factorization with partial pivoting. Pivoting makes necessarily a large amount of data communication and slows down the total computation time [**9, 138**]. As pointed out in [**75**, p. 219], the SNV iteration (3.6) can be implemented using a QR decomposition (without column pivoting). Computing a QR decomposition can be done in a communication friendly way with great performance on modern multicore

and heterogeneous systems [**68**]. Therefore, the QR-based SNV method is an attractive alternative. Unfortunately, as shown in Section 3.5, the SNV iteration (3.6) is not stable for ill-conditioned matrices, even with the QR decomposition-based implementation. The instability had also been independently reported in early studies [**26, 19**]. In the next section, we will exploit an alternative iteration to develop an inverse free method.

3.2. Halley's method and dynamically weighted Halley

Halley's iteration for computing the polar factor of a nonsingular matrix A is

$$X_{k+1} = X_k(3I + X_k^*X_k)(I + 3X_k^*X_k)^{-1}, \quad X_0 = A. \tag{3.8}$$

It is a member of the Padé family of iterations [**87**]. Such iterations frequently arise in linear algebra applications, including iterations for matrix sign functions [**75**, Ch.5] and moment matching for model order reduction [**49, 50**]. It is proven that X_k converges globally and that the convergence rate is cubic [**51, 52**]. However, the initial steps of convergence can still be slow when A has a singular value that has large relative distance from 1. For example, consider the 2×2 matrix

$$A = X_0 = \begin{bmatrix} 1 & \\ & x_0 \end{bmatrix}, \quad x_0 = 10^{-10}. \tag{3.9}$$

The polar factor of A is the 2×2 identity matrix. The kth iterate X_k is given by

$$X_k = \begin{bmatrix} 1 & \\ & x_k \end{bmatrix}, \quad x_k = \frac{x_{k-1}(3 + x_{k-1}^2)}{1 + 3x_{k-1}^2}.$$

After one Halley's iteration, $x_1 \approx 3 \times 10^{-10}$. It takes 24 iterations for the iterate X_k to converge to the polar factor within IEEE double precision machine precision, namely, $\|X_{24} - I\|_2 \leq \epsilon_{\mathrm{M}} = 2.2 \times 10^{-16}$.

To accelerate the convergence of Halley's iteration (3.8), let us consider the following DWH iteration:

$$X_{k+1} = X_k(a_kI + b_kX_k^*X_k)(I + c_kX_k^*X_k)^{-1}, \quad X_0 = A/\alpha, \tag{3.10}$$

where $\alpha = \|A\|_2$ and scalars a_k, b_k, and c_k are nonnegative weighting parameters. We choose these weighting parameters suboptimally[1] in the sense that it maximizes ℓ_{k+1} such that the interval $[\ell_{k+1}, 1]$ contains all the singular values of X_{k+1}. Specifically, let $X_k = U\Sigma_kV^*$ be the SVD of X_k and ℓ_k be such that[2]

$$[\sigma_{\min}(X_k), \sigma_{\max}(X_k)] \subseteq [\ell_k, 1] \subset (0, 1] \tag{3.11}$$

[1]The term "suboptimal" follows that of the suboptimal scaling (3.4) for the SN iteration, which minimizes b_{k+1} such that $[1, b_{k+1}]$ contains all the singular values of X_{k+1}.

[2]In (3.11) one can assume a more general interval $[\widehat{\ell}_k, L]$ for any $L > 0$, but setting $L \equiv 1$ causes no loss of generality since the simple scaling $a_{k-1} \leftarrow a_{k-1}/L$, $b_{k-1} \leftarrow b_{k-1}/L$ maps the interval $[\widehat{\ell}_k, L]$ containing $[\sigma_{\min}(X_k), \sigma_{\max}(X_k)]$ to $[\widehat{\ell}_k/L, 1] \equiv [\ell_k, 1]$.

with initial $\sigma_{\min}(X_0) = \beta/\alpha \equiv \ell_0$ and $\beta = 1/\|A^{-1}\|_2$. Then one step of the DWH iteration (3.10) yields

$$\begin{aligned} X_{k+1} &= X_k(a_k I + b_k X_k^* X_k)(I + c_k X_k^* X_k)^{-1} \\ &= U\Sigma_k V^*(a_k I + b_k V\Sigma_k^2 V^*)(I + c_k V\Sigma_k^2 V^*)^{-1} \\ &= U\Sigma_k(a_k I + b_k\Sigma_k^2)(I + c_k\Sigma_k^2)^{-1}V^* \equiv U\Sigma_{k+1}V^*. \end{aligned}$$

Hence the singular values $\sigma_i(X_{k+1})$ of X_{k+1} are given by

$$\sigma_i(X_{k+1}) = g_k(\sigma_i(X_k)), \tag{3.12}$$

where g_k is a rational function defined as

$$g_k(x) = x\frac{a_k + b_k x^2}{1 + c_k x^2}.$$

By (3.11) and (3.12), we have

$$[\sigma_{\min}(X_{k+1}), \sigma_{\max}(X_{k+1})] \subseteq \left[\min_{\ell_k \le x \le 1} g_k(x), \max_{\ell_k \le x \le 1} g_k(x)\right]. \tag{3.13}$$

Since the closeness of the iterate X_{k+1} to the polar factor can be measured by the maximum distance between singular values $\sigma_i(X_{k+1})$ and 1, a suboptimal choice of the triplet (a_k, b_k, c_k) should make the function g_k be bounded

$$0 < g_k(x) \le 1 \quad \text{for} \quad \ell_k \le x \le 1 \tag{3.14}$$

and attain the max-min

$$\max_{a_k, b_k, c_k} \left\{\min_{\ell_k \le x \le 1} g_k(x)\right\}. \tag{3.15}$$

Once these parameters a_k, b_k, and c_k are found to satisfy (3.14) and (3.15), all singular values of X_{k+1} satisfy

$$[\sigma_{\min}(X_{k+1}), \sigma_{\max}(X_{k+1})] \subseteq [\ell_{k+1}, 1] \subset (0, 1], \tag{3.16}$$

where $\ell_{k+1} = \min_{\ell_k \le x \le 1} g_k(x)$.

Let us now consider how to solve the optimization problem (3.14) and (3.15). To satisfy $g_k(x) > 0$, we can impose

$$a_k, b_k, c_k > 0 \tag{3.17}$$

and

$$g_k(1) = 1. \tag{3.18}$$

These conditions ensure that the function $g_k(x)$ is positive and continuously differentiable for $x > 0$ and has a fixed point at 1. Note that (3.18) implies $c_k = a_k + b_k - 1$. By the assumptions (3.17) and (3.18), the optimization problem (3.14) and (3.15) can be stated as follows.

The bounded max-min problem: find $a_k, b_k > 0$ such that $c_k = a_k + b_k - 1 > 0$,

$$0 < g_k(x) \le 1 \quad \text{for} \quad \ell_k \le x \le 1, \tag{3.19}$$

and

$$\max_{a_k, b_k > 0} \left\{ \min_{\ell_k \le x \le 1} g_k(x) \right\} \tag{3.20}$$

is attained.

In Section 3.6, we show that the solution of the optimization problem (3.19) and (3.20) is given by

$$a_k = h(\ell_k), \quad b_k = (a_k - 1)^2/4, \tag{3.21}$$

where

$$h(\ell) = \sqrt{1+d} + \frac{1}{2}\sqrt{8 - 4d + \frac{8(2-\ell^2)}{\ell^2\sqrt{1+d}}}, \quad d = \sqrt[3]{\frac{4(1-\ell^2)}{\ell^4}}. \tag{3.22}$$

Similar to the SN iteration (3.3) with the suboptimal scaling (3.4), we see that the weighting parameters a_k, b_k and $c_k = a_k + b_k - 1$ of the DWH iteration (3.10) can be generated by simple scalar iterations in which the initial value ℓ_0 depends on the extreme singular values of the original matrix A.

In summary, we derive the DWH iteration (3.10) for computing the polar factor of A, where the weighting parameters a_k and b_k are generated by the scalar iterations (3.21), c_k is defined by $c_k = a_k + b_k - 1$, and

$$\ell_0 = \frac{\beta}{\alpha}, \quad \ell_k = \frac{\ell_{k-1}(a_{k-1} + b_{k-1}\ell_{k-1}^2)}{1 + c_{k-1}\ell_{k-1}^2} \quad \text{for} \quad k = 1, 2, \ldots, \tag{3.23}$$

where $\alpha = \|A\|_2$ and $\beta = 1/\|A^{-1}\|_2$.

Before we prove the global convergence of the DWH iteration (3.10), let us recall the 2×2 matrix A defined as (3.9). The kth DWH iterate X_k is given by

$$X_k = \begin{bmatrix} 1 & \\ & x_k \end{bmatrix}, \quad x_k = \frac{x_{k-1}(a_k + b_k x_{k-1}^2)}{1 + c_k x_{k-1}^2}.$$

Since $\alpha = 1$ and $\ell_0 = 10^{-10}$, by (3.21), we have $a_0 \simeq 1.17 \times 10^7$, $b_0 \simeq 3.42 \times 10^{13}$, and $c_0 \simeq 3.42 \times 10^{13}$. After one iteration, x_0 is mapped to $x_1 \simeq 1.17 \times 10^{-3}$, which is much closer to the target value 1 than the first iterate computed by Halley's iteration (3.8). In fact, it takes only five DWH iterations to approximate the polar factor within the machine precision $\|X_5 - I\|_2 \le \epsilon_{\mathrm{M}}$. It is a factor of five times faster than the Halley iteration. To explain the fast convergence of the DWH iteration, the plots of Figure 3.2.1 show the typical mapping functions g_k from the singular values of X_k to that of X_{k+1} by the Halley iteration (3.8) and the DWH iteration (3.10). We can see that the singular values of X_k are mapped much closer to 1 by the DWH iteration than the Halley iteration.

It is worth noting that the DWH plot in Figure 3.2.1 exhibits an equioscillation behavior on $[\ell, 0]$. This behavior is always observed for any $\ell \in (0, 1)$, and we can explain this from

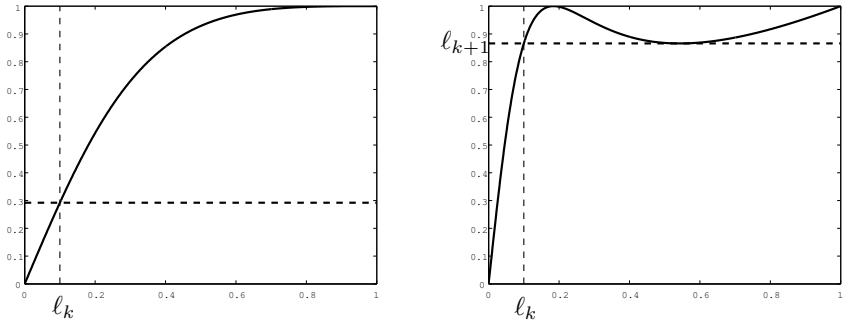

FIGURE 3.2.1. The mapping functions $g(\sigma) = \sigma(3+\sigma^2)/(1+3\sigma^2)$ of the Halley iteration (left) and $g_k(\sigma) = \sigma(a_k + b_k\sigma^2)/(1+c_k\sigma^2)$ of the DWH iteration (right).

the point of view of approximation theory [**154**]. Specifically, since the goal is to map the singular values to 1, a natural approach is to let $g_k(x)$ be as close as possible to the function that maps all values on $x \in [\ell, 1]$ to the constant value 1. Then it is known that the best rational approximation to a given continuous function must equioscillate [**154, 162**]. In this sense DWH can be regarded as a powerful rational approximation for the constant function on $[\ell, 1]$ under the constraint $f(0) = 0$.

THEOREM 3.2. *For a nonsingular A, the iterates X_k generated by the DWH iteration* (3.10) *converge to the polar factor U_p of A. The asymptotic convergence factor is cubic.*

PROOF. We first prove the convergence of the iterates X_k. This is equivalent to showing that the singular values $\sigma_i(X_k) \to 1$ as $k \to \infty$ for all $1 \le i \le n$. By (3.16), we have $[\sigma_{\min}(X_k), \sigma_{\max}(X_k)] \subseteq [\ell_k, 1]$. Hence it suffices to prove $\ell_k \to 1$ as $k \to \infty$.

Using (3.21), (3.23), and $c_k = a_k + b_k - 1$, we derive

$$\frac{1}{\ell_{k+1}} - 1 = F(a_k, \ell_k)\left(\frac{1}{\ell_k} - 1\right),$$

where

$$F(a, \ell) = \frac{((a-1)\ell - 2)^2}{4a + (a-1)^2\ell^2}.$$

Note that $F(a_k, \ell_k) \ge 0$ since $a_k > 0$. All we need to show is that there is a positive constant $\delta < 1$ such that $F(a_k, \ell_k) \le \delta$ for all k. In fact, we have $3 \le a \le \frac{2+\ell}{\ell}$ (see (3.64) in Section 3.6), and on this interval $F(a, \ell)$ is a decreasing function of a:

$$\frac{\partial F}{\partial a} = \frac{4(1+\ell)(\ell^2(a-1)^2 - 4)}{(\ell^2 + a^2\ell^2 + 2a(2-\ell^2))^2} \le 0 \quad \text{on} \quad 3 \le a \le \frac{2+\ell}{\ell}.$$

Therefore, we have

$$F(a_k, \ell_k) \le F(3, \ell_k) = \frac{(3-1)^2\ell_k^2 - 4(3-1)\ell_k + 4}{(3-1)^2\ell_k^2 + 4 \cdot 3} = \frac{(1-\ell_k)^2}{\ell_k^2 + 3} \le \frac{1}{3} = \delta.$$

This completes the proof of the global convergence of the DWH iteration.

Now we consider the asymptotic rate of convergence. By the above argument,

$$|1-\ell_{k+1}| = \left| F(a_k,\ell_k)\left(\frac{1}{\ell_k}-1\right)\ell_{k+1}\right| \le \left|\frac{\ell_{k+1}(1-\ell_k)^2}{\ell_k^2+3}\left(\frac{1}{\ell_k}-1\right)\right| = \frac{\ell_{k+1}\,|1-\ell_k|^3}{\ell_k(\ell_k^2+3)}.$$

By the fact that $\ell_k \to 1$, we conclude that the DWH is asymptotically cubically convergent. □

REMARK 1. It is shown in Section 3.6 that a_k satisfies

$$3 \le a_k \le \frac{2+\ell_k}{\ell_k} \qquad \text{for} \quad k \ge 0,$$

where $0 < \ell_k \le 1$. Note that as $\ell_k \to 1$, $(a_k, b_k, c_k) \to (3, 1, 3)$. These are the weighting parameters of the Halley iteration (3.8).

REMARK 2. For simplicity of exposition, we used the exact extreme singular values of the original matrix A in the analysis, namely, $\alpha = \sigma_{\max}(A)$, $\beta = \sigma_{\min}(A)$, and $\ell_0 = \beta/\alpha = 1/\kappa_2(A)$. In fact, estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β are sufficient as long as the inclusion property $[\sigma_{\min}(A/\widehat{\alpha}), \sigma_{\max}(A/\widehat{\alpha})] \subseteq [\widehat{\ell}_0, 1]$ holds, where $\widehat{\ell}_0 = \widehat{\beta}/\widehat{\alpha}$.

REMARK 3. In [**52**], Gander has observed the slow convergence with respect to small singular values in Halley's iteration. As a remedy and generalization to rectangle and rank-deficient matrices, he proposed the following weighting parameters:

$$a_k = \frac{2\tau-3}{\tau-2}, \quad b_k = \frac{1}{\tau-2}, \quad c_k = \frac{\tau}{\tau-2}, \tag{3.24}$$

where τ is a prescribed parameter. When $\tau = 3$, it is the Halley iteration. It is proved that, for any $\tau > 2$, the resulting method converges globally and quadratically [**96**]. In practice, Gander [**52**] suggests taking $\tau = 2 + \epsilon_{\mathrm{M}}/\delta$ for $\delta > 10\epsilon_{\mathrm{M}}$ and $\tau = 2.1$ for $\epsilon_{\mathrm{M}} < \delta \le 10\epsilon_{\mathrm{M}}$, where ϵ_{M} is the machine epsilon and δ is the convergence tolerance. This stems from the observation that taking τ close to 2 results in both speed-up and instability. We here set the tolerance δ small enough, in which case $\tau = 2.1$. Note that Gander's iteration switches from iteration (3.10) with static weighting parameter (3.24) to the standard Halley iteration (3.8) after a certain number of iterations. To find the appropriate switching strategy, it is noticed that about $s = -\log(\ell_0)$ steps are needed for the smallest singular value to increase to the size of 1, where $\ell_0 = \beta/\alpha = \sigma_{\min}(X_0)$. Therefore, the switching is done after s iterations using $\tau = 2.1$. Unfortunately, the convergence of Gander's iteration can still be slow. For the 2×2 matrix in (3.9), Gander's iteration needs 14 iterations to converge. In Section 3.5, we see that as many as 20 iterations are needed for some cases.

To derive a stopping criterion for the DWH iteration (3.10), we note that, once convergence sets in, $\ell_k \simeq 1$ so that $(a_k, b_k, c_k) \simeq (3, 1, 3)$. Therefore, we will just need to consider a proper stopping criterion for Halley's iteration (3.8). We note that in the SN iteration with Higham's $(1,\infty)$–norm scaling, switching to the unscaled Newton iteration is recommended [**75, 89**]. As for the DWH iteration, this switching is not necessary because we have $(a_k, b_k, c_k) \to (3, 1, 3)$. This is also true for the SN iteration with suboptimal scaling, which guarantees the scaling factor $\zeta_k \to 1$.

We first have the following lemma.

LEMMA 3.1. *For Halley's iteration* (3.8), *if* $\|X_{k-1} - U\|_2 = \|I - \Sigma_{k-1}\|_2 = \epsilon \ll 1$, *then up to the first order in* ϵ,

$$\|X_{k-1} - U\| = (1 + O(\epsilon^2))\|X_k - X_{k-1}\|.$$

PROOF. Writing $X_{k-1} = U\Sigma_{k-1}V$ we have

$$\begin{aligned} X_k - X_{k-1} &= X_{k-1}(3I + X_{k-1}^* X_{k-1})(I + 3X_{k-1}^* X_{k-1})^{-1} - X_{k-1} \\ &= 2X_{k-1}(I - X_{k-1}^* X_{k-1})(I + 3X_{k-1}^* X_{k-1})^{-1} \\ &= 2U(I - \Sigma_{k-1}^2)\Sigma_{k-1}(I + 3\Sigma_{k-1}^2)^{-1}V^*. \end{aligned} \tag{3.25}$$

Taking an unitarily invariant norm and using the inequality $\|AB\| \le \|A\| \cdot \|B\|_2$, we get

$$\begin{aligned} \|X_k - X_{k-1}\| &\le 2\|U(I - \Sigma_{k-1})V^*\| \cdot \|\Sigma_{k-1}(I + \Sigma_{k-1})(I + 3\Sigma_{k-1}^2)^{-1}\|_2 \\ &= 2\|X_{k-1} - U\| \cdot \|\Sigma_{k-1}(I + \Sigma_{k-1})(I + 3\Sigma_{k-1}^2)^{-1}\|_2 \\ &\equiv 2\|X_{k-1} - U\| \cdot \|f(\Sigma_{k-1})\|_2, \end{aligned} \tag{3.26}$$

where $f(x) = x(1+x)/(1+3x^2)$ is a continuouse and differentiable function. It is easy to see that $f(x)$ is increasing on $(0,1)$ and decreasing on $(1,\infty)$. It attains the maximum $1/2$ at $x = 1$. Hence we can write $f(1-\epsilon) = 1/2 - O(\epsilon^2)$ for $\epsilon \ll 1$. Consequently, we have $\|f(\Sigma_{k-1})\|_2 = \max_i |f(\sigma_i)| = 1/2 - O(\epsilon^2)$. By (3.26), it follows that $\|X_k - X_{k-1}\| \le (1 - O(\epsilon^2))\|X_{k-1} - U\|$.

We can prove $\|X_k - X_{k-1}\| \ge (1 - O(\epsilon^2))\|X_{k-1} - U\|$ similarly by noticing from (3.25) that $X_{k-1} - U = \frac{1}{2}(X_k - X_{k-1})(V(I+\Sigma_{k-1})\Sigma_{k-1}(I+3\Sigma_{k-1}^2)^{-1}V^*)^{-1}$ and using $1/f(1+\epsilon) = 2 + O(\epsilon^2)$. □

Now, by the identity $U(\Sigma_{k-1} - I)V^* = X_{k-1} - U$, we have

$$\begin{aligned} \|X_k - U\| &= \|U(\Sigma_{k-1}(3I + \Sigma_{k-1}^2)(I + 3\Sigma_{k-1}^2)^{-1} - I)V^*\| \\ &= \|U(\Sigma_{k-1} - I)^3(I + 3\Sigma_{k-1}^2)^{-1}V^*\| \\ &\le \|X_{k-1} - U\|^3 \cdot \|(I + 3X_{k-1}^* X_{k-1})^{-1}\|_2, \end{aligned}$$

where we used the inequality $\|AB\| \le \|A\| \cdot \|B\|_2$ again. When close to convergence, $\|X_{k-1} - U\|_2 \ll 1$. Hence, by Lemma 3.1, we have

$$\|X_k - U\| \lesssim \|X_k - X_{k-1}\|^3 \cdot \|(I + 3X_{k-1}^* X_{k-1})^{-1}\|_2.$$

This suggests that we accept X_k when

$$\|X_k - X_{k-1}\|_F \le \left(\frac{\epsilon_{\mathrm{M}}}{\|(I + 3X_{k-1}^* X_{k-1})^{-1}\|_2} \right)^{1/3}. \tag{3.27}$$

Close to convergence $X_{k-1}^* X_{k-1} \simeq I$, the test (3.27) is effectively

$$\|X_k - X_{k-1}\|_F \le (4\epsilon_{\mathrm{M}})^{1/3}. \tag{3.28}$$

We recall that for quadratically convergent methods such as the SN method (3.3) and its variant (3.6), the following stopping criterion is suggested [**75**, p. 208]:

$$\|X_k - X_{k-1}\|_F \le (2\epsilon_{\mathrm{M}})^{1/2}. \tag{3.29}$$

In [**20**], it is noted that theoretically the SN iteration with the suboptimal scaling converges in at most nine steps for any matrix of condition number less than 10^{16}. It is based on the bound $\|X_k - U\|_2 \leq b_k - 1$, where b_k can be obtained by a simple scalar iteration. Consequently, the first k satisfying $|1 - b_k| < 10^{-16}$ provides an upper bound on the number of iteration counts.

We can derive a similar result for the DWH iteration (3.10). By the interval (3.16) that bounds the singular values of the iterate X_k, we have

$$\|X_k - U\|_2 = |1 - \sigma_{\min}(X_k)| \leq |1 - \ell_k|.$$

Hence, by finding the first k such that $|1 - \ell_k| < 10^{-16}$, we obtain the number of iterations needed for the DWH iteration to convergence. Specifically, by using the scalar recursion (3.23) with $\ell_0 = 1/\kappa_2(A)$, we have the following upper bounds for the number of DWH iterations:

$\kappa_2(A)$	10^1	10^2	10^5	10^8	10^{10}	10^{12}	10^{16}
SN, SNV	5	6	7	8	8	9	9
DWH	3	4	5	5	5	5	6

The result suggests that the DWH iteration converges within at most six steps for any matrix with condition number $\kappa_2(A) \leq 10^{16}$. The number of DWH iterations is about one-third fewer than the number of SN iterations (3.3) with the suboptimal scaling (3.4).

3.3. QR-based implementations

In this section, we discuss an implementation of the DWH iteration (3.10) using the QR decomposition. The QR-based implementation is more desirable than those involving explicit inverses for enhancing parallelizability. Numerical examples in Section 3.5 suggest that it also improves the numerical stability.

First we have the following basic result, given in [**75**, p. 219] and based on the results in [**171**].

THEOREM 3.3. *Let* $\begin{bmatrix} \eta X \\ I \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R$ *be a QR decomposition of* $\begin{bmatrix} \eta X \\ I \end{bmatrix}$, *where* $X, Q_1 \in \mathbb{C}^{m\times n}$ *and* $Q_2, R \in \mathbb{C}^{n\times n}$. *Then*

$$Q_1 Q_2^* = \eta X (I + \eta^2 X^* X)^{-1}. \tag{3.30}$$

PROOF. By the polar decomposition

$$\begin{bmatrix} \eta X \\ I \end{bmatrix} = \widetilde{U}\widetilde{H}, \tag{3.31}$$

we have $\widetilde{H}^2 = I + \eta^2 X^* X$ and $\widetilde{H} = (I + \eta^2 X^* X)^{1/2}$. Note that $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$ and $\widetilde{U}$ span the column space of $\begin{bmatrix} \eta X \\ I \end{bmatrix}$ and that they both have orthonormal columns. Hence it follows that

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \widetilde{U} W \tag{3.32}$$

for some orthogonal matrix W. By (3.31) and (3.32), we have

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} \eta X \\ I \end{bmatrix} (I + \eta^2 X^* X)^{-1/2} W.$$

The identity (3.30) can now be verified by a straightforward calculation. □

By Theorem 3.3, we immediately derive that the SNV iteration (3.6) is mathematically equivalent to the following iteration, referred to as a QR-based scaled Newton variant (QSNV):

$$\begin{cases} \begin{bmatrix} \eta_k X_k \\ I \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R, \\ X_{k+1} = 2 Q_1 Q_2^*, \end{cases} \tag{3.33}$$

with the initial $X_0 = A$, where the scaling factor η_k is defined as (3.7). The following is a pseudocode of the QSNV iteration:

QSNV algorithm:

$X_0 = A$
$\eta_0 = 1/\sqrt{\alpha\beta}$ and $k = 0$
repeat
  compute QR decomposition $\begin{bmatrix} \eta_k X_k \\ I \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R$
  $X_{k+1} = 2 Q_1 Q_2^*$
  if $k = 0$ **then**
    $\eta_1 = \sqrt{(\alpha + \beta)/(2\sqrt{\alpha\beta})}$
  else
    $\eta_{k+1} = \sqrt{(\eta_k + 1/\eta_k)/2}$
  end if
  $k = k + 1$
until convergence
$U_p = X_k$

Now we consider the DWH iteration (3.10). Iteration (3.10) can be equivalently written as

$$X_{k+1} = \frac{b_k}{c_k} X_k + \left(a_k - \frac{b_k}{c_k} \right) X_k (I + c_k X_k^* X_k)^{-1}, \quad X_0 = A/\alpha, \tag{3.34}$$

where the weighting triplet (a_k, b_k, c_k) is defined as (3.21). By Theorem 3.3, iteration (3.34) can be written using the QR decomposition as follows, referred to as the QR-based dynamically weighted Halley (QDWH) iteration:

$$\text{(3.35)} \qquad \begin{cases} \begin{bmatrix} \sqrt{c_k}X_k \\ I \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R, \\ X_{k+1} = \dfrac{b_k}{c_k}X_k + \dfrac{1}{\sqrt{c_k}}\left(a_k - \dfrac{b_k}{c_k}\right) Q_1 Q_2^*. \end{cases}$$

The following is a pseudocode of the QDWH iteration:

QDWH algorithm:

1: $X_0 = A/\alpha$, $\ell_0 = \beta/\alpha$
2: $k = 0$
3: **repeat**
4: $\quad a_k = h(\ell_k),\ b_k = (a_k - 1)^2/4,\ c_k = a_k + b_k - 1$
5: $\quad$ compute QR decomposition $\begin{bmatrix} \sqrt{c_k}X_k \\ I \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R$
6: $\quad X_{k+1} = (b_k/c_k)X_k + (1/\sqrt{c_k})\,(a_k - b_k/c_k)\,Q_1 Q_2^*$
7: $\quad \ell_{k+1} = \ell_k(a_k + b_k\ell_k^2)/(1 + c_k\ell_k^2)$
8: $\quad k = k + 1$
9: **until** convergence
10: $U_p = X_k$

For the practical implementations of the QSNV and QDWH methods, we only need estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β satisfying $0 < \widehat{\beta} \le \sigma_{\min}(A) \le \sigma_{\max}(A) \le \widehat{\alpha}$. We can simply use $\widehat{\alpha} = \|A\|_F$. An estimate of $\beta = \sigma_{\min}(A)$ is a nontrivial task [**25, 69, 71**]. For the SN method, the estimate $\widehat{\beta} = 1/\|A^{-1}\|_F$ is suggested in [**20**]. However, it is not practical for the QSNV and QDWH methods since A^{-1} is not calculated explicitly. By the inequality $\|A\|_1/\sqrt{n} \le \|A\|_2 \le \sqrt{n}\|A\|_1$, we have $\beta = \sigma_{\min}(A) = \|A^{-1}\|_2^{-1} \ge (\sqrt{n}\|A^{-1}\|_1)^{-1}$. Therefore, we may use the lower bound of β as an estimate, i.e.,

$$\text{(3.36)} \qquad \widehat{\beta} = (\gamma\sqrt{n})^{-1},$$

where γ is the LAPACK 1-norm estimate of A^{-1} [**74**, Chap. 15]. In Section 3.5, we will examine the effect of the estimate $\widehat{\beta}$ on the convergence of the QDWH iteration. The numerical examples suggest that it is harmless to use a rough estimate $\widehat{\beta}$ as far as $\widehat{\ell_0} = \widehat{\beta}/\widehat{\alpha}$ is a lower bound of $\sigma_{\min}(X_0)$. We note that QDWH and Gander's algorithm use the normalized initial matrix $X_0 = A/\alpha$, whereas SN and QSNV use $X_0 = A$. However, the scalars α and β need to be provided in all these methods.

To end this section, let us consider the arithmetic cost of the QDWH method. Note that the QSNV and QDWH iterations share the same computational kernel, namely,

(a) compute $\begin{bmatrix} \eta X \\ I \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R$, and

(b) form $\widehat{X} = Q_1 Q_2^*$.

A straightforward implementation is to first compute the QR decomposition of the $2n \times n$ matrix by a dense QR decomposition using the LAPACK routine `DGEQRF` [**3**]. The cost is $\frac{10}{3}n^3$ flops. Then we form Q_1 and Q_2 explicitly by using `DORGQR`. Its cost is $\frac{10}{3}n^3$ flops. Finally, we compute the product $Q_1Q_2^*$ by the matrix-matrix multiplication routine `DGEMM` in BLAS with the cost $2n^3$ flops. In summary, the arithmetic cost of each QDHW iteration is $\frac{26}{3}n^3$ flops. Since it generally takes at most six iterations to converge, the total cost of the QDWH method is at most $52n^3$ flops.

In contrast, the cost of each SN iteration is $2n^3$ flops if the matrix inversion is computed by LU factorization-based routines `DGETRF` and `DGETRI` in LAPACK. Together with the fact that it generally needs at most nine steps to converge, the total cost of the SN method is at most $18n^3$ flops. Therefore, the cost of the QDWH method is about three times more than that of the SN method. If the matrix inversion in the SN iteration is calculated using a bidiagonal reduction-based algorithm for backward stability [**20**], then it increases the cost to $6n^3$ flops per iteration. This makes the total cost up to $54n^3$ flops. In this case, the costs of the SN and QDWH methods are about the same.

We note that it is possible to reduce the cost of the QDWH method by exploiting the diagonal block in the QR decomposition step. We can first compute the QR decomposition of ηX and then carefully reduce the augmented matrix into a triangular form by using Givens rotations. The cost per QDWH iteration is reduced to $(16/3)n^3$ flops. Thus the cost of six iterations of QDWH iterations is thereby bounded by $32n^3$ flops.

3.4. Backward stability proof

Recall in Theorems 4.1 and 5.1 that the crucial assumption for QDWH-eig and QDWH-SVD to be backward stable is that QDWH computes the polar decomposition in a backward stable manner. Here we develop backward stability analyses of algorithms for computing the polar decomposition. We first present a sufficient condition for the computed polar decomposition to be backward stable. We then show that QDWH satisfies the condition provided that row/column pivoting is used in computing the QR decompositions.

3.4.1. Sufficient condition for an algorithm to be backward stable. Throughout this section we define $f(A)$ by $f(A) = Uf(\Sigma)V^*$, where $A = U\Sigma V^*$ is an SVD with $\Sigma = \mathrm{diag}(\sigma_i)$ and $f(\Sigma) = \mathrm{diag}(f(\sigma_i))$. Note that $f(A)$ does not depend on the particular choice of A's SVD.

The unitary polar factors of $f(X)$ and X are identical for any rectangular X. We begin with a lemma that gives a sufficient condition for the unitary polar factor of a computed $f(X)$ to provide a backward stable polar decomposition of X.

LEMMA 3.2. *Let X be an m-by-n ($m \geq n$) matrix, and let $\widehat{Y}$ be a computed approximation to $Y = f(X)$. Suppose that $\widehat{Y}$ is computed in a mixed backward-forward manner, so that*

$$\widehat{Y} = f(\widetilde{X}) + \epsilon\|\widehat{Y}\|_2, \quad \textit{where} \quad \widetilde{X} = X + \epsilon\|X\|_2. \tag{3.37}$$

Let $m = \sigma_{\min}(\widetilde{X}), M = \sigma_{\max}(\widetilde{X})$. Suppose further that $m > 0$, and that the function $f(x)$ satisfies

$$\frac{f(x)}{\max_{m\le x\le M} f(x)} \ge \frac{x}{dM} \quad \textit{for all} \quad x \in [m, M], \tag{3.38}$$

where $d \ge 1$ is a modest constant[3]*. Then the unitary polar factor U_p of $\widehat{Y} = U_p H_{\widehat{Y}}$ provides a backward stable polar decomposition for X, that is,*

$$X - U_p\widehat{H} = \epsilon\|X\|_2, \quad \textit{where} \quad \widehat{H} = \frac{1}{2}((U_p^*X) + (U_p^*X)^*). \tag{3.39}$$

In addition, $\widehat{H}$ approximates the exact Hermitian polar factor H of X to relative accuracy ϵ:

$$\widehat{H} = H + \epsilon\|H\|_2. \tag{3.40}$$

PROOF. We first consider the square case $m = n$. Let $\widetilde{X} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^*$ be an SVD where $\widetilde{\Sigma} = \mathrm{diag}(\widetilde{\sigma}_i)$, so that $f(\widetilde{X}) = \widetilde{U}f(\widetilde{\Sigma})\widetilde{V}^*$. Also let $\widetilde{H} = \widetilde{V}\widetilde{\Sigma}\widetilde{V}^*$ be the Hermitian polar factor of $\widetilde{X}$.

It is known [**75**, p. 200],[**14**, p. 215] that the Hermitian polar factor is well-conditioned: the Hermitian polar factor of $A + \Delta A$ differs from A by at most $\sqrt{2}\|\Delta A\|_F$ in the Frobenius norm. Hence by $f(\widetilde{X}) = \widehat{Y} + \epsilon\|\widehat{Y}\|_2$ we must have $H_{\widehat{Y}} = f(\widetilde{H}) + \epsilon\|\widehat{Y}\|_2$[4], where $f(\widetilde{H}) = \widetilde{V}f(\widetilde{\Sigma})\widetilde{V}^*$ is the exact Hermitian polar factor of $f(\widetilde{X})$. Hence we have

$$\widehat{Y} = U_p H_{\widehat{Y}} = U_p f(\widetilde{H}) + \epsilon\|\widehat{Y}\|_2 = U_p\widetilde{V}f(\widetilde{\Sigma})\widetilde{V}^* + \epsilon\|\widehat{Y}\|_2.$$

Together with $\widehat{Y} = \widetilde{U}f(\widetilde{\Sigma})\widetilde{V}^* + \epsilon\|\widehat{Y}\|_2$ and $\|\widehat{Y}\|_2 = \|f(\widetilde{\Sigma})\|_2 + \epsilon\|\widehat{Y}\|_2 = (1+\epsilon)\|f(\widetilde{\Sigma})\|_2$ we get

$$\widetilde{U}f(\widetilde{\Sigma})\widetilde{V}^* = U_p\widetilde{V}f(\widetilde{\Sigma})\widetilde{V}^* + \epsilon\|f(\widetilde{\Sigma})\|_2. \tag{3.41}$$

Now we right-multiply (3.41) by $\widetilde{V}f(\widetilde{\Sigma})^{-1}\widetilde{\Sigma}\widetilde{V}^* = \widetilde{V}\mathrm{diag}(\widetilde{\sigma}_1/f(\widetilde{\sigma}_1), \ldots, \widetilde{\sigma}_n/f(\widetilde{\sigma}_n))\widetilde{V}^*$. In doing so we note that

$$\begin{aligned}
\|\widetilde{V}f(\widetilde{\Sigma})^{-1}\widetilde{\Sigma}\widetilde{V}^*\|_2 &= \max_i \frac{\widetilde{\sigma}_i}{f(\widetilde{\sigma}_i)} \\
&\le \max_{m\le x\le M} \frac{x}{f(x)} \qquad (\text{because } \widetilde{\sigma}_i \in [m, M]) \\
&\le \frac{dM}{\max_{m\le x\le M} f(x)} \qquad (\text{by (3.38)}) \\
&\le \frac{dM}{\max_i f(\widetilde{\sigma}_i)} = \frac{d\max_i \widetilde{\sigma}_i}{\max_i f(\widetilde{\sigma}_i)} = \frac{d\|\widetilde{X}\|_2}{\|f(\widetilde{\Sigma})\|_2}.
\end{aligned}$$

[3]The condition (3.38) for the case $x = M$ forces $d \ge 1$.

[4]This $\epsilon\|\widehat{Y}\|_2$ can be magnified by a factor $\simeq \log n$ (not $\sqrt{2n}$ as in a general case) when converting the Frobenius norm to ths spectral norm [**14**, p. 321]. We regard low-degree polynomials of n as modest constants, so we used $\epsilon \log n = \epsilon$.

Since in our convention $d\epsilon = \epsilon$, we therefore get

$$\widetilde{U}\widetilde{\Sigma}\widetilde{V}^* = U_p\widetilde{V}\widetilde{\Sigma}\widetilde{V}^* + \epsilon\|\widetilde{X}\|_2. \tag{3.42}$$

Left-multiplying (3.42) by U_p^* and using $\|\widetilde{X}\|_2 = (1+\epsilon)\|X\|_2$, $\widetilde{X} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^*$ and $\widetilde{H} = \widetilde{V}\widetilde{\Sigma}\widetilde{V}^*$ we get

$$U_p^*\widetilde{X} = \widetilde{H} + \epsilon\|X\|_2. \tag{3.43}$$

Therefore we conclude that the backward error of the polar decomposition of X is

$$\begin{aligned} X - U_p\widehat{H} &= X - U_p \cdot \frac{1}{2}(U_p^*X + (U_p^*X)^*) \\ &= \frac{1}{2}(X - U_pXU_p) \\ &= \frac{1}{2}U_p(U_p^*X - XU_p) \\ &= \frac{1}{2}U_p\left((\widetilde{H} + \epsilon\|X\|_2) - (\widetilde{H} + \epsilon\|X\|_2)^*\right) \\ &= \epsilon\|X\|_2, \end{aligned}$$

where we used (3.43) to get the fourth equality and $\widetilde{H} = \widetilde{H}^*$ for the last equality. This proves (3.39).

To get (3.40), first note that since $\widetilde{H}$ is the (well-conditioned) Hermitian polar factor of $\widetilde{X} = X + \epsilon\|X\|_2$, we must have $H = \widetilde{H} + \epsilon\|X\|_2$. Combining this with (3.43) and $\widehat{H} = \frac{1}{2}((U_p^*X) + (U_p^*X)^*)$ we conclude that

$$\begin{aligned} \widehat{H} - H &= \frac{1}{2}((U_p^*X) + (U_p^*X)^*) - H \\ &= \frac{1}{2}\left((\widetilde{H} + \epsilon\|X\|_2) + (\widetilde{H} + \epsilon\|X\|_2)^*\right) - (\widetilde{H} + \epsilon\|X\|_2) \\ &= \epsilon\|X\|_2 = \epsilon\|H\|_2. \end{aligned}$$

Finally, consider the rectangular case $m > n$. In this case we can make the same argument as above by appending $m-n$ columns of zeros to the right of $X, \widetilde{X}$ and $\widehat{Y}$ (e.g., replace X with $[X\ 0_{m,m-n}]$). Note that in this case we have $m = \sigma_n(\widetilde{X})$ and the exact Hermitian polar factors $H, \widetilde{H}, f(H), f(\widetilde{H})$ take the form $\begin{bmatrix} H & 0_{n,m-n} \\ 0_{m-n,n} & 0_{m-n,m-n} \end{bmatrix}$, and $\widehat{H}$ takes the form $\begin{bmatrix} \widehat{H}_{n,n} & E^* \\ E & 0_{m-n,m-n} \end{bmatrix}$. Here E is an $(m-n)$-by-n matrix with $\|E\|_2 = \epsilon\|H\|_2$, which is an innocuous artifact of the analysis that does not appear in actual computation that involves no m-by-m matrices. □

The requirement (3.38) on $f(x)$ warrants further remark. The condition ensures that an initially large singular value $\sigma_i(X) \simeq \|X\|_2$ must stay large in the relative sense through the mapping f, that is, $f(\sigma_i(X)) \simeq \|f(X)\|_2$. If, on the contrary, $\widetilde{d} = \frac{\sigma_i(X)}{\|X\|_2} \cdot \frac{\|f(X)\|_2}{f(\sigma_i(X))} \gg 1$ for some i then then the backward error bound (or the $\epsilon\|\widetilde{X}\|_2$ term in (3.42)) increases by a factor $\widetilde{d}$. A qualitative explanation is that if $f(\sigma_i(X)) \ll \max_j f(\sigma_j(X))$ then the accuracy of the singular vector corresponding to $\sigma_i(X)$ is lost by a factor $\frac{\|f(X)\|_2}{f(\sigma_i(X))}$, because such singular

vectors are obscured by the presence of the singular vectors corresponding to the singular values equal (and close) to $\|f(X)\|_2$.

A simple example illustrates the effect of such "unstable mappings" of singular values. Let $\mu = 10^{-10}$ and let $A = U\mathrm{diag}(1, \sqrt{\mu}, \mu)V^T$ where U, V are randomly generated unitary matrices. We have the polar factors $U_p = UV^T$ and $H = \frac{1}{2}(U_p^T A + (U_p^T A)^T)$. Now let $f(x)$ be a function that maps $f(1) = \mu, f(\sqrt{\mu}) = \mu$ and $f(\mu) = 1$, which is an unstable mapping because (3.38) holds only for $d \geq 1/\mu = 10^{10}$. In MATLAB we compute the SVD $f(A) + E = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ where $\|E\|_2 = \epsilon \simeq 1.1 \times 10^{-16}$, then form $\widetilde{U}_p = \widetilde{U}\widetilde{V}^T$ and $\widetilde{H} = \frac{1}{2}(\widetilde{U}_p^T A + (\widetilde{U}_p^T A)^T)$. Then we had $\|A - \widetilde{U}_p\widetilde{H}\|_2/\|A\|_2 \simeq \|H - \widetilde{H}\|_2/\|H\|_2 \simeq 10^{-6}$, which shows a large backward error due to the unstable mapping. Contrarily, for a stable mapping $f(x)$ such that $f(1) = 1, f(\sqrt{\mu}) = 10^{-1}$ and $f(\mu) = 0.01$ so that (3.38) holds with $d = 1$, running the same process for 10^5 randomly generated U, V and E we always had small backward errors $\|A - \widetilde{U}_p\widetilde{H}\|_2/\|A\|_2 \simeq \|H - \widetilde{H}\|_2/\|H\|_2 \leq 2.1 \cdot 10^{-15}$.

As we shall see below, such an unstable mapping of singular values does not happen in the QDWH or the scaled Newton iterations (which are backward stable), but does happen in the unstable algorithm QSNV mentioned in [**120**]. To illustrate the idea, we show in Figures 3.4.1, 3.4.2 and 3.4.3 plots of $\frac{f(x)}{\max_{m \leq x \leq M} f(x)} = \frac{f(x)}{\|f(x)\|_\infty}$ and $\frac{x}{M}$ for the three methods: the QDWH iteration (left), the scaled Newton iteration (middle), and the (unstable) QSNV iteration (right). The plots show the case when $\kappa_2(X) = 20$ and optimal scaling/weighting parameters are used. Observe that $\frac{f(x)}{\|f(x)\|_\infty}$ lies above $\frac{x}{M}$ in the left two plots (indicating (3.38) holds with $d = 1$), but not in the right plot. In fact, below we see that for QSNV we have $\frac{f(x)}{\|f(x)\|_\infty} / \frac{x}{M} \simeq 2/\sqrt{\kappa_2(X)}$ for $x \simeq M$, so (3.38) holds only with $d \geq \frac{1}{2}\sqrt{\kappa_2(X)}$.

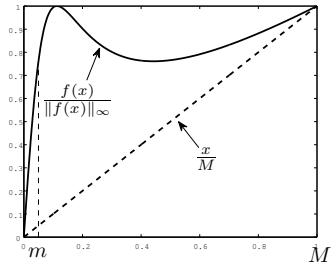

FIGURE 3.4.1. QDWH iteration. This is a "stable" mapping because (3.38) holds with $d = 1$.

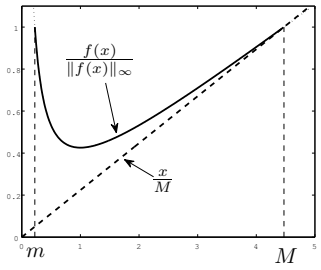

FIGURE 3.4.2. Scaled Newton iteration. This is a "stable" mapping because (3.38) holds with $d = 1$.

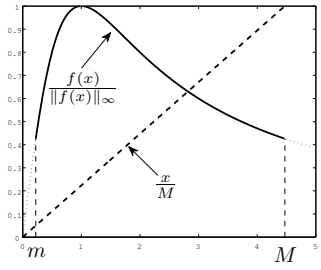

FIGURE 3.4.3. QSNV iteration. This is an "unstable" mapping because (3.38) holds only for $d \gtrsim 2/\sqrt{\kappa_2(X)}$.

We also note that although (3.40) shows the computed Hermitian polar factor $\widehat{H}$ approximates the exact H to working accuracy, $\widehat{H}$ is not guaranteed to be positive semidefinite. However, since eigenvalues of Hermitian matrices are well-conditioned (3.40) implies that $\widehat{H}$ can be indefinite only if A (and hence H) is ill-conditioned, and the negative eigenvalues (if any) are of size $\epsilon\|A\|_2$ in absolute value.

Now we state the main result of this section.

THEOREM 3.4. *Let A be a rectangular matrix. Consider an iteration X_k for $k = 0, 1, \ldots$ for computing its unitary polar factor, expressed (in exact arithmetic) as $X_{k+1} = f_k(X_k) = U f_k(\Sigma) V^*$ and $X_0 = A/\alpha$ for a positive constant α. Suppose that for $k = 0, \ldots,$ the computed $\widehat{X}_k$ and the function f_k satisfy the two conditions*

- *$\widehat{X}_{k+1}$ is computed from $\widehat{X}_k$ in a mixed backward-forward manner, so that*

$$\widehat{X}_{k+1} = f_k(\widetilde{X}_k) + \epsilon \|\widehat{X}_{k+1}\|_2 \quad \text{where} \quad \widetilde{X} = \widehat{X}_k + \epsilon \|\widehat{X}_k\|_2. \tag{3.44}$$

  Here for notational convenience we let $\widehat{X}_0 = X_0$.
- *The function $f_k(x)$ satisfies*

$$\frac{f_k(x)}{\max_{m_k \le x \le M_k} f_k(x)} \ge \frac{x}{d M_k} \quad \text{on} \quad [m_k, M_k], \tag{3.45}$$

  where $m_k = \sigma_{\min}(\widetilde{X}_k), M_k = \sigma_{\max}(\widetilde{X}_k)$ and d is a modest constant.

Finally, suppose that $\min_{k\to\infty} X_k = \widehat{U}_p$ is unitary to working accuracy, hence $\widehat{U}_p^* \widehat{U}_p = I + \epsilon$. Then $\widehat{U}_p$ is a backward stable unitary polar factor of A, that is,

$$A - \widehat{U}_p \widehat{H} = \epsilon \|A\|_2, \tag{3.46}$$

where $\widehat{H} = \frac{1}{2}(\widehat{U}_p^* A + (\widehat{U}_p^* A)^*)$. In addition, $\widehat{H}$ is a forward stable Hermitian polar factor of A:

$$\widehat{H} - H = \epsilon \|H\|_2.$$

PROOF. Let $\widehat{X}_j = U_{p,j} H_j$ be the exact polar decomposition for $j = 0, 1, \ldots$. We first show that for any given k, the unitary polar factor $U_{p,k}$ of $\widehat{X}_k$ gives a backward stable polar decomposition for any $\widehat{X}_j$ for $j = 0, \ldots, k$, that is,

$$\widehat{X}_j - U_{p,k} \widehat{H}_j = \epsilon \|\widehat{X}_j\|_2 \quad \text{and} \quad \|H_j - \widehat{H}_j\|_2 = \epsilon \|H_j\|_2 \quad \text{for} \quad j = 0, 1, \ldots, k, \tag{3.47}$$

where $\widehat{H}_j = \frac{1}{2}((U_{p,k}^* \widehat{X}_j) + (U_{p,k}^* \widehat{X}_j)^*)$.

Showing (3.47) for the case $j = k$ is trivial. The case $j = k - 1$ follows immediately from using Lemma 3.2 (substituting $\widehat{X}_{k-1}$ into X and $\widehat{X}_k$ into $\widehat{Y}$) and the assumptions (3.44),(3.45).

Next we consider $j = k - 2$. Since (3.47) holds for $j = k - 1$ we have

$$\widehat{X}_{k-1} = U_{p,k} \widehat{H}_{k-1} + \epsilon \|\widehat{X}_{k-1}\|_2 = U_{p,k} H_{k-1} + \epsilon \|\widehat{X}_{k-1}\|_2.$$

By the assumption (3.44) for the case $k := k - 2$ we also have

$$\widehat{X}_{k-1} = f_{k-2}(\widehat{X}_{k-2} + \epsilon \|\widehat{X}_{k-2}\|_2) + \epsilon \|\widehat{X}_{k-1}\|_2,$$

so it follows that

$$U_{p,k} H_{k-1} = f_{k-2}(\widehat{X}_{k-2} + \epsilon \|\widehat{X}_{k-2}\|_2) + \epsilon \|\widehat{X}_{k-1}\|_2.$$

Since $U_{p,k} H_{k-1}$ is a polar decomposition, this means that $U_{p,k}$ is the unitary polar factor of a perturbed version of $f_{k-2}(\widehat{X}_{k-2})$, satisfying the mixed backward-forward error model as in (3.37). Hence we can invoke Lemma 3.2 by letting $\widehat{Y} := U_{p,k} H_{k-1}$ and $X := \widehat{X}_{k-2}$, which shows (3.47) is satisfied for $j = k - 2$.

By repeating the same argument we can prove (3.47) for $j = k-3, k-4, \ldots, 0$. In particular, letting $k \to \infty$ and $j = 0$ in (3.47), together with $\widehat{X}_0 = X_0 = A/\alpha$, yields

$$A - U_p\widehat{H} = \epsilon\|A\|_2, \quad \text{and} \quad \|H - \widehat{H}\|_2 = \epsilon\|H\|_2, \tag{3.48}$$

where U_p is the unitary polar factor of $\widehat{U}_p = \lim_{k\to\infty} \widehat{X}_k$ and $\widehat{H} = \frac{1}{2}(U_p^*A + (U_p^*A)^*)$.

The only gap between (3.48) and (3.46) is that $\widehat{U}_p$ is not exactly unitary in (3.46). To prove (3.46), note that $\widehat{U}_p^*\widehat{U}_p = I + \epsilon$ implies $\widehat{U}_p = U_p + \epsilon$, because the unitary polar factor U_p is the nearest unitary matrix to $\widehat{U}_p$. Hence we get

$$\begin{aligned}
&A - \widehat{U}_p \cdot \frac{1}{2}(\widehat{U}_p^*A + (\widehat{U}_p^*A)^*) \\
\leq &\left(A - U_p \cdot \frac{1}{2}(U_p^*A + (U_p^*A)^*)\right) + \frac{1}{2}\left(U_p \cdot (U_p^*A + (U_p^*A)^*) - \widehat{U}_p \cdot (\widehat{U}_p^*A + (\widehat{U}_p^*A)^*)\right) \\
= &\epsilon\|A\|_2 + \frac{1}{2}\left(U_p \cdot (U_p^*A + (U_p^*A)^*) - (U_p+\epsilon) \cdot ((U_p+\epsilon)^*A + ((U_p+\epsilon)^*A)^*)\right) \\
= &\epsilon\|A\|_2,
\end{aligned}$$

which is (3.46).

Finally, we use the second equation in (3.47) letting $j = 0, k \to \infty$, which, together with $\widehat{X}_0 = A/\alpha$ and $\widehat{U}_p = U_p + \epsilon$, yields

$$\begin{aligned}
\widehat{H} - H &= \frac{1}{2}(\widehat{U}_p^*A + (\widehat{U}_p^*A)^*) - H \\
&\leq \frac{1}{2}\left(\widehat{U}_p^*A + (\widehat{U}_p^*A)^*) - (U_p^*A + (U_p^*A)^*\right) + \frac{1}{2}(U_p^*A + (U_p^*A)^*) - H \\
&= \epsilon\|A\|_2 + (\widehat{H} - H) \\
&= \epsilon\|A\|_2 + \epsilon\|H\|_2 = \epsilon\|H\|_2.
\end{aligned}$$

□

3.4.2. QDWH with row/column pivoting is backward stable. Using Theorem 3.4 we prove that the QDWH algorithm is backward stable, provided that the QR decompositions are computed with row and column pivoting.

For what follows it helps to note that (in exact arithmetic) the QDWH iteration preserves the singular vectors while mapping all the singular values to 1 at a (sub)optimal rate, that is, if $X_0 = A/\alpha = U\Sigma_0 V^*$ is the SVD, then $X_k = U\Sigma_k V^*$ where $\Sigma_k = f_{k-1}(\Sigma_{k-1}) = f_{k-1}(f_{k-2}(\ldots f_0(\Sigma_0))$ where $f_k(x)$ maps each singular value by the rational function $f_k(x) = x\frac{a_k + b_k x^2}{1 + c_k x^2}$. We have $\Sigma_k \to I$ as $k \to \infty$.

In view of Theorem 3.4, it suffices to prove the two conditions (3.44) and (3.45) are satisfied throughout the QDWH iterations. We treat these separately.

3.4.2.1. *Proving mixed forward-backward stability of QDWH iteration.* The goal of this subsection is to prove the first condition (3.44) is satisfied in QDWH. Here we express a general QDWH iteration as $Y = f(X)$ where $f(x) = x\frac{a+bx^2}{1+cx^2}$, and we shall show that the computed value $\widehat{Y}$ satisfies the mixed forward-backward stable error model (3.37). Assume

without loss of generality that $\|\widehat{X}\|_2 = 1$ (which is necessarily true if $\alpha = \|A\|_2$ in (3.35)). Since $f(1) = 1$ we have $\|Y\|_2 = 1$, and hence our goal is to show

$$\widehat{Y} = f(\widehat{X}) + \epsilon \quad \text{where} \quad \widehat{X} = X + \epsilon. \tag{3.49}$$

We use the following result concerning the row-wise stability of a Householder QR factorization with column/row pivoting.

THEOREM 3.5. [**74**, Sec.19.4] *Let $\widehat{Q}, \widehat{R}$ be the computed QR factors of A obtained via the Householder QR algorithm with row and column pivoting. Then*

$$(A + \Delta A)\Pi = \widehat{Q}\widehat{R},$$

where Π is a permutation matrix and $\|\Delta A(i,:)\|_2 \le d\epsilon\|A(i,:)\|_2$ for all i for a modest constant $d \ge 1$[5]*. .*

We use this theorem to prove (3.49). If the QR factorization of $\begin{bmatrix} \sqrt{c}X \\ I \end{bmatrix}$ in (3.35) is computed with column and row pivoting, then by Theorem 3.5 the computed $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{R}$ satisfy

$$\left(\begin{bmatrix} \sqrt{c}X \\ I \end{bmatrix} + \Delta X \right) \Pi = \begin{bmatrix} \sqrt{c}(X + \epsilon) \\ I + \epsilon_1 \end{bmatrix} \Pi = \begin{bmatrix} \widehat{Q}_1 \\ \widehat{Q}_2 \end{bmatrix} \widehat{R}, \tag{3.50}$$

where we used $\|X\|_2 = 1$. Here ϵ_1 is of order ϵ, but unlike ϵ takes a fixed value in all appearances. Since $\Pi \cdot \Pi^* = I$, (3.50) is equivalent to

$$\begin{bmatrix} \sqrt{c}(X + \epsilon) \\ I + \epsilon_1 \end{bmatrix} = \begin{bmatrix} \widehat{Q}_1 \\ \widehat{Q}_2 \end{bmatrix} \widehat{R}\Pi^*.$$

It follows that $\begin{bmatrix} \widehat{Q}_1 \\ \widehat{Q}_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \end{bmatrix}$ ($\epsilon_{11}, \epsilon_{21}$ account for the loss of orthogonality caused by rounding error in forming $\widehat{Q}_1$ and $\widehat{Q}_2$) is the exact orthogonal factor of $\begin{bmatrix} \sqrt{c}(X + \epsilon) \\ I + \epsilon_1 \end{bmatrix} := \begin{bmatrix} \sqrt{c}\widehat{X} \\ I + \epsilon_1 \end{bmatrix}$. Therefore, the computed product $fl(\widehat{Q}_1\widehat{Q}_2^*)$ satisfies ([**19**], proof given in Section 3.4.4) Using Proposition 3.1 in Section 3.6 we have

$$(\widehat{Q}_1 + \epsilon_{11})(\widehat{Q}_2 + \epsilon_{21})^* = \sqrt{c}\widehat{X}((I + \epsilon_1)^*(I + \epsilon_1) + c\widehat{X}^*\widehat{X})^{-1}(I + \epsilon_1)^*,$$

and so

$$\widehat{Q}_1\widehat{Q}_2^* = \sqrt{c}\widehat{X}((I + \epsilon_1)^*(I + \epsilon_1) + c\widehat{X}^*\widehat{X})^{-1}(I + \epsilon_1)^* - \widehat{Q}_1\epsilon_{21}^* - \epsilon_{11}\widehat{Q}_2^*,$$

Since $-\widehat{Q}_1\epsilon_{21}^* - \epsilon_{11}\widehat{Q}_2^* = \epsilon$, accounting for the multiplication rounding error we get

$$fl(\widehat{Q}_1\widehat{Q}_2^*) = \sqrt{c}\widehat{X}((I + \epsilon_1)^*(I + \epsilon_1) + c\widehat{X}^*\widehat{X})^{-1}(I + \epsilon_1)^* + \epsilon. \tag{3.51}$$

Therefore we get

$$\widehat{Y} = \frac{b}{c}X + \sqrt{c}\widehat{X}((I + \epsilon_1)^*(I + \epsilon_1) + c\widehat{X}^*\widehat{X})^{-1}(I + \epsilon_1)^* + \epsilon,$$

[5] d here can be bounded by $\sqrt{m}(1 + \sqrt{2})^{n-1}$, which grows exponentially with n, so d becomes very large for moderately large m. However experiments show d is usually much smaller in practice.

where the last term ϵ here includes the rounding error caused by performing the addition (we used the fact that the norms of both terms in the addition are bounded by 1). Note that

$$\begin{aligned}((I+\epsilon_1)^*(I+\epsilon_1)+c\widehat{X}^*\widehat{X})^{-1} &= \Big((I+(\epsilon_1+\epsilon_1^*+\epsilon_1^*\epsilon_1)(I+c\widehat{X}^*\widehat{X})^{-1})(I+c\widehat{X}^*\widehat{X})\Big)^{-1} \\ &= (I+c\widehat{X}^*\widehat{X})^{-1}(I+(\epsilon_1+\epsilon_1^*+\epsilon_1^*\epsilon_1)(I+c\widehat{X}^*\widehat{X})^{-1})^{-1}.\end{aligned}$$

Note that $\|(I+c\widehat{X}^*\widehat{X})^{-1}\|_2 \le 1$ because the singular values of $I+c\widehat{X}^*\widehat{X}$ are all larger than 1. Hence $\|(\epsilon_1+\epsilon_1^*+\epsilon_1^*\epsilon_1)(I+c\widehat{X}^*\widehat{X})^{-1}\|_2 = \epsilon$, so the above equation can be written $(I+c\widehat{X}^*\widehat{X})^{-1}+\epsilon$. Therefore we get

$$\widehat{Y} = \frac{b}{c}X + \left(a-\frac{b}{c}\right)\widehat{X}(I+c\widehat{X}^*\widehat{X})^{-1}(I+\epsilon_1)+\epsilon.$$

Since $\| \left(a-\frac{b}{c}\right)\widehat{X}(I+c\widehat{X}^*\widehat{X})^{-1}\|_2 \le 1+\epsilon$ (which follows from $f(x)\le 1$ and $f'(1)\le 1$), $c \ge b$ (which follows from $c = a+b-1$ and $a \ge 3$) and $X = \widehat{X}+\epsilon$, we conclude that

$$\widehat{Y} = \frac{b}{c}\widehat{X} + \left(a-\frac{b}{c}\right)\widehat{X}(I+c\widehat{X}^*\widehat{X})^{-1}+\epsilon.$$

Since we have $\frac{b}{c}\widehat{X} + \left(a-\frac{b}{c}\right)\widehat{X}(I+c\widehat{X}^*\widehat{X})^{-1} = \widehat{X}(aI+b\widehat{X}^*\widehat{X})(I+c\widehat{X}^*\widehat{X})^{-1}$, we have proved that

$$\widehat{Y} = \widehat{X}(aI+b\widehat{X}^*\widehat{X})(I+c\widehat{X}^*\widehat{X})^{-1}+\epsilon,$$

which is (3.49). □

3.4.2.2. *Proving* (3.45). First note that Section 3.6 shows $f(x)$ generally takes one local maximum and local minimum on $[0,1]$, and is monotonically increasing on $[0,\infty)$, see Figure 3.4.4.

We separately consider the two cases $\|A\|_2/\alpha \ge 1$ and $\|A\|_2/\alpha < 1$. When $\|A\|_2/\alpha \ge 1$, we have $\|X_k\|_2 = f_{k-1}(f_{k-2}(\dots f_0(\|A\|_2/\alpha)) \ge 1$, because $f'(x) \ge 0$ on $[0,\infty]$. Hence (3.45) is satisfied if

$$\frac{f(x)}{x} \ge \frac{f(M)}{dM} \quad \text{on} \quad [0,M],$$

where $M \ge 1$. This holds with $d = 1$, because basic algebra shows the function $f(x)/x = \frac{a+bx^2}{1+cx^2}$ is decreasing on $[1,\infty)$. Hence QDWH is backward stable in this case.

Next, when $\|A\|_2/\alpha < 1$, we have $\|X_k\|_2 \ge \|X_0\|_2 = \|A\|_2/\alpha$ for all $k \ge 0$. This is because $f(x) \ge x$ on $[0,1]$. To see this, note that

$$f(x)-x = \frac{ax+bx^3}{1+cx^2} - x = \frac{(a-1)x+(b-c)x^3}{1+cx^2}.$$

Using $b = (a-1)^2/4$ and $c = a+b-1$, we have

$$(a-1)x+(b-c)x^3 = (a-1)x+(1-a)x^3 = (a-1)x(1-x^2) \ge 0.$$

Using $f(x) \le 1$ on $[0,1]$ we see that (3.45) is satisfied if

$$\frac{f(x)}{x} \ge \frac{1}{dM} \quad \text{on} \quad [0,M],$$

where $M \geq \|A\|_2/\alpha$. Since $f(x) \geq x$ on $[0,1]$ this holds for $d = 1/M$. $1/M \leq \alpha/\|A\|_2$ is modest unless α is a severe overestimate of $\|A\|_2$. Since we can inexpensively get an upper bound $\|A\|_F \leq \sqrt{n}\|A\|_2$, it follows that QDWH is always backward stable.

Note that the above analysis does not involve ℓ_0, suggesting the estimate of $\sigma_{\min}(X_0)$ plays no role on the backward stability of QDWH. Of course, it plays a fundamental role on the convergence speed.

□

For clarity, Figure 3.4.4 plots the functions $y = f(x)$ and $y = x$ for the case $\ell = 0.1$, which shows a typical behavior of $f(x)$.

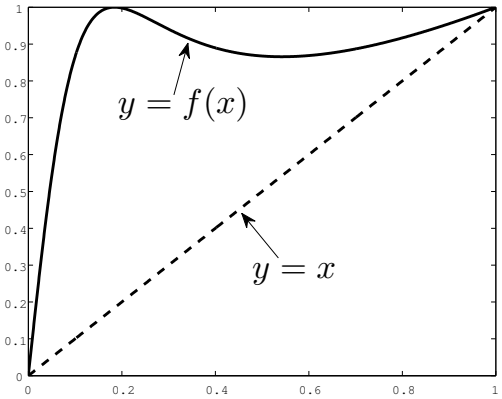

FIGURE 3.4.4. Plots when $\ell = 0.1$. Verify that $x \leq f(x) \leq 1$ on $[0,1]$.

The results of the two subsection combined with Theorem 3.4 prove that QDWH with column and row pivoting is backward stable.

Remark: In practice, even without pivoting QDWH exhibits excellent backward stability. This is the case even when the computed QR decomposition has terrible row-wise backward stability, as seen in the following example. Let

$$A = \begin{bmatrix} 1 & 1 \\ \sqrt{2} & \sqrt{3} \end{bmatrix} \operatorname{diag}(10^{-15}, 1).$$

We have $\kappa_2(A) \simeq 1.2 \times 10^{16}$. Letting $\alpha = 2$, the computed QR decomposition in the first QDWH iteration $\begin{bmatrix} \sqrt{c_0}X_0 \\ I \end{bmatrix} \simeq \begin{bmatrix} \widehat{Q}_1 \\ \widehat{Q}_2 \end{bmatrix} \widehat{R}$ satisfies

$$\|\widehat{Q}_1\widehat{R} - \sqrt{c_0}X_0\|_F / \|\sqrt{c_0}X_0\|_F = 2 \times 10^{-16},$$

but the row-wise backward error of the second block is

$$\|\widehat{Q}_2\widehat{R} - I\|_F = 1.3 \times 10^{-5}.$$

Hence the row-wise backward stability of the first QR factorization is terrible. However, the computed matrices $\widehat{U}_p, \widehat{H}$ after 6 QDWH iterations satisfy

$$\|A - \widehat{U}_p\widehat{H}\|_F/\|A\|_F = 4.9 \times 10^{-16},$$

so QDWH performed in a backward stable manner.

Whether or not we can prove backward stability of QDWH without pivoting is an open problem.

3.4.3. Backward stability of other polar decomposition algorithms. The analysis of Theorem 3.4 is general enough to let us investigate backward stability of other polar decomposition algorithms. Here we use the theorem to show that the scaled Newton iteration is backward stable provided that matrix inverses are computed in a mixed backward-forward stable manner. The backward stability has been studied in [**20, 89**], and through very complicated analyses [**89**] proves its backward stability when Higham's $(1,\infty)$–scaling [**73**] is used. [**20**] derives the backward stability through much simpler arguments, but [**88**] indicates some incompleteness of the analysis. We believe our proof here is much simpler than that of both [**89**] and [**20**].

We also give an explanation of why QSNV (3.33), which is mathematically equivalent to the scaled Newton iteration, fails to be backward stable.

3.4.3.1. *The scaled Newton iteration is backward stable.* As is well known, the scaled Newton iteration is an effective method for computing the unitary polar factor of a square nonsingular matrix [**75**]. The iteration is expressed as

$$X_{k+1} = \frac{1}{2}\left(\zeta_k X_k + (\zeta_k X_k)^{-*}\right), \quad X_0 = A, \tag{3.52}$$

where $\zeta_k > 0$ is a scaling factor, practical and effective choices of which are Higham's $(1,\infty)$–norm scaling [**73**] and the suboptimal scaling in [**20**]. Experimentally, it has numerical backward stability comparable to that of QDWH.

We shall show that we can establish backward stability of the scaled Newton iteration by using Theorem 3.4, because if inverses are computed in a mixed backward-forward stable manner then showing the first in Theorem 3.4 is simple, and the second condition follows by the fact $\frac{1}{2}(x + x^{-1}) \geq x$, when (nearly) optimal scaling is used. Below we explain these in more detail.

Proving (3.44). If the computed inverse $\widehat{Z} = fl(X^{-1})$ is mixed backward-forward stable then we can write $\widehat{Z} = (X + \epsilon\|X\|_2)^{-1} + \epsilon\|\widehat{Z}\|_2 \equiv \widetilde{X}^{-1} + \epsilon\|\widehat{Z}\|_2$. Therefore the computed version $\widehat{Y}$ of $Y = \frac{1}{2}\left(\zeta X + (\zeta X)^{-*}\right)$ satisfies

$$\begin{aligned}
\widehat{Y} &= \frac{1}{2}\left(\zeta X + \widehat{Z}/\zeta\right) + \epsilon \max\{\|\zeta X\|_2, \|\widehat{Z}/\zeta\|_2\} \\
&= \frac{1}{2}\left(\zeta X + (\widetilde{X}^{-*} + \epsilon\|\widehat{Z}\|_2)/\zeta\right) + \epsilon \max\{\|\zeta X\|_2, \|\widehat{Z}/\zeta\|_2\} \\
&= \frac{1}{2}\left(\zeta(\widetilde{X} + \epsilon\|X\|_2) + (\widetilde{X}^{-*} + \epsilon\|\widehat{Z}\|_2)/\zeta\right) + \epsilon \max\{\|\zeta X\|_2, \|\widehat{Z}/\zeta\|_2\} \\
&= \frac{1}{2}\left(\zeta\widetilde{X} + (\zeta\widetilde{X})^{-*}\right) + \epsilon \max\{\|\zeta X\|_2, \|\widehat{Z}/\zeta\|_2\}.
\end{aligned}$$

This implies $\|\widehat{Y}\|_2 \geq \frac{1}{2}\max\{\|\zeta\widetilde{X}\|_2, \|\widetilde{X}^{-1}/\zeta\|_2\} \simeq \frac{1}{2}\max\{\|\zeta X\|_2, \|\widehat{Z}/\zeta\|_2\}$, so we conclude that $\widehat{Y} = \frac{1}{2}\left(\zeta\widehat{X} + (\zeta\widehat{X})^{-*}\right) + \epsilon\|\widehat{Y}\|_2$. Hence the iteration (3.3) is computed in a mixed backward-forward stable manner.

Proving (3.45). In the above notation, the second condition (3.45) is $\frac{g(\zeta x)}{\max_{m\leq x\leq M} g(\zeta x)} \geq \frac{x}{cM}$ on $[m, M]$, where $g(x) = \frac{1}{2}(x + x^{-1})$, $m = \sigma_{\min}(\widehat{X})$ and $M = \sigma_{\max}(\widehat{X})$. Note that $\max_{m\leq x\leq M} g(x) = \max(g(\zeta m), g(\zeta M))$, because the function $g(\zeta x)$ can take its maximum on a closed positive interval only at the endpoints. Hence we need to show

$$\frac{g(\zeta x)}{\max(g(\zeta m), g(\zeta M))} \geq \frac{x}{cM} \quad \text{on} \quad [m, M], \tag{3.53}$$

for a modest constant c. We consider the cases $g(\zeta m) \leq g(\zeta M)$ (which happens when $\zeta \geq \zeta_{opt}$, where ζ_{opt} is the optimal scaling parameter which satisfies $\|\zeta_{opt}X\|_2 = \|(\zeta_{opt}X)^{-1}\|_2$) and $g(\zeta m) < g(\zeta M)$ separately.

In the first case, the condition (3.53) becomes $\frac{g(\zeta x)}{g(\zeta M)} \geq \frac{x}{cM}$ on $[m, M]$. This is equivalent to

$$\frac{g(\zeta x)}{\zeta x} \geq \frac{1}{d} \cdot \frac{g(\zeta M)}{\zeta M} \quad \text{on} \quad [m, M]. \tag{3.54}$$

(3.54) holds with $d = 1$, because $\frac{g(\zeta x)}{\zeta x} = \frac{1}{2}(1 + (\zeta x)^{-2})$ is a decreasing function on $(0, M]$ and equality holds when $x = M$.

In the second case, similar arguments show the inequality is equivalent to the condition $\frac{g(\zeta x)}{\zeta x} \geq \frac{g(\zeta m)}{c\zeta M}$ on $[m, M]$. At $x = M$ the function $\frac{g(\zeta x)}{\zeta x}$ takes its minimium

$$\frac{g(\zeta M)}{\zeta M} = \left(\frac{g(\zeta M)}{g(\zeta m)}\right) \cdot \frac{g(\zeta m)}{\zeta M},$$

so $\frac{g(\zeta x)}{\zeta x} \geq \frac{g(\zeta m)}{d\zeta M}$ holds with $d = \frac{g(m)}{g(M)}$. Hence backward stability can be lost only if $g(M) \ll g(m)$, which happens when $\frac{1}{m} \gg M$, which means $\zeta \ll \zeta_{opt}$. Since ζ is generally obtained by estimating $\zeta_{opt} = 1/\sqrt{\sigma_{\min}(X)\sigma_{\max}(X)}$, it follows that backward stability is maintained unless $\sigma_{\min}(X)$ and/or $\sigma_{\max}(X)$ are severely overestimated. In practice using $\|X\|_F$ ensures $\sigma_{\max}(X)$ is not overestimated by more than a factor $\sqrt{n}$, so we conclude that the scaled Newton iteration is backward stable provided that the estimate of $\sigma_{\min}(X)$ is not a terrible overestimate.

3.4.3.2. *The QSNV iteration is not backward stable.* We observed in Chapter 3 that the QSNV algorithm, although mathematically equivalent to scaled Newton iteration, is not backward stable. We can explain this instability by showing that d in (3.45) must be extremely large when $\kappa_2(A) \gg 1$. In the QSNV iteration we have $f(x) = 2\eta x(1+\eta^2x^2)^{-1}$. For simplicity suppose that the optimal scaling factor $\eta = 1/\sqrt{\sigma_{\min}(X)\sigma_{\max}(X)}$ is chosen. (3.45) at $x = \sigma_{\max}(X)$ becomes $\|f(X)\|_2/f(\sigma_{\max}(X_k)) \leq \frac{1}{d}$. Since $\|X_{k+1}\|_2 \leq \max_{0\leq x\leq\infty} f(x) = 1$ and $f(\sigma_{\max}(X_k)) = f(\kappa_2(X)) \lesssim 2/\sqrt{\kappa_2(X)}$, it follows that we need $d \geq \sqrt{\kappa_2(X)}/2$. Therefore even if the optimal scaling factor is chosen, QSNV violates the second condition by a large factor for ill-conditioned matrices. We note that the loss of backward stability by a

factor $\sqrt{\kappa_2(A)}$ accurately reflects the backward stability observed in numerical experiments in Section 3.5.

3.4.4. Proof of (3.51)**.** We first prove the following result, which is basically a copy of Proposition 1 in [**19**]:

PROPOSITION 3.1. *Suppose* $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \in \mathbb{C}^{m\times n}$ *has full column rank, where* $B_1 \in \mathbb{C}^{(m-n)\times n}$ *and* $B_2 \in \mathbb{C}^{n\times n}$*. Also suppose*

$$B = V \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \bar{R}$$

is a decomposition such that $V^*V = I_m$*,* $V_1 \in \mathbb{C}^{(m-n)\times n}$*,* $V_2 \in \mathbb{C}^{n\times n}$ *and* $R \in \mathbb{C}^{n\times n}$*. Then,*

$$V_1 = B_1(B^*B)^{-1/2}W, \quad V_2 = B_2(B^*B)^{-1/2}W, \quad \bar{R} = W^*(B^*B)^{1/2}. \tag{3.55}$$

for some unitary matrix W*.*

PROOF. $(B^*B)^{1/2}$ is nonsingular because B has full column rank, so B has the decomposition

$$B = QR = \begin{bmatrix} B_1(B^*B)^{-1/2} \\ B_2(B^*B)^{-1/2} \end{bmatrix} (B^*B)^{1/2},$$

where $Q \in \mathbb{C}^{m\times n}$ has orthonormal columns. By setting $B = QR = V \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix}$ we have

$$V^*Q = \begin{bmatrix} \bar{R}R^{-1} \\ 0 \end{bmatrix}.$$

Since V is unitary, we have

$$(V^*Q)^*(V^*Q) = I = (\bar{R}R^{-1})^*\bar{R}R^{-1},$$

so $\bar{R}R^{-1}$ is also unitary. Hence by letting $\bar{R}R^{-1} = W^*$, we get

$$\begin{aligned} B &= \begin{bmatrix} B_1(B^*B)^{-1/2} \\ B_2(B^*B)^{-1/2} \end{bmatrix} R \\ &= \begin{bmatrix} B_1(B^*B)^{-1/2}W \\ B_2(B^*B)^{-1/2}W \end{bmatrix} W^*R = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \bar{R}, \end{aligned}$$

which is (3.55). □

Now, to prove (3.51), recall that $\begin{bmatrix} \widehat{Q}_1 \\ \widehat{Q}_2 \end{bmatrix} + \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix}$ is the exact orthogonal factor of the QR factorization of $\begin{bmatrix} \sqrt{c_k}Y_k \\ I + Z \end{bmatrix}$, where $\|\Delta_1\|_2, \|\Delta_2\|_2 \le d\epsilon$. Using Proposition 3.1 we have

$$(\widehat{Q}_1 + \Delta_1)(\widehat{Q}_2 + \Delta_2)^* = \sqrt{c_k}Y_k((I+Z)^*(I+Z) + c_kY_k^*Y_k)^{-1}(I+Z)^*.$$

Hence

$$\widehat{Q}_1\widehat{Q}_2^* = \sqrt{c_k}Y_k((I+Z)^*(I+Z) + c_kY_k^*Y_k)^{-1}(I+Z)^* - \widehat{Q}_1\Delta_2^* - \Delta_1\widehat{Q}_2^*,$$

from which (3.51) follows by letting $\Delta' = -\widehat{Q}_1\Delta_2^* - \Delta_1\widehat{Q}_2^*$ and observing $\|\Delta'\|_2 \le \|\Delta_2\| + \|\Delta_1\|_2 \le d\epsilon$, and accounting for the multiplication rounding error Z_0.

3.5. Numerical examples

This section shows several numerical experiments to demonstrate the numerical behaviors of the QDWH method. All numerical experiments were performed in MATLAB 7.4.0 and run on a PC with Intel® Core™ 2 Duo processor. The machine epsilon is $\epsilon_{\mathrm{M}} \simeq 2.2 \times 10^{-16}$. The stopping criterion (3.28) is used for the cubically convergent methods, namely, Halley, Gander, DWH, and QDWH iterations. For the quadratically convergent Newton-type methods, namely, SN, NV, SNV, and QSNV iterations, the stopping criterion (3.29) is applied. Since A^{-1} is computed explicitly in the SN iteration, the estimates of extreme singular values are $\widehat{\alpha} = \|A\|_F$ and $\widehat{\beta} = 1/\|A^{-1}\|_F$. Otherwise, we use the estimates $\widehat{\alpha} = \|A\|_F$ and $\widehat{\beta}$ as in (3.36).

The accuracy of the computed polar decomposition is tested by the residual norm $\mathrm{res} = \|A - \widehat{U}\widehat{H}\|_F/\|A\|_F$, where $\widehat{U}$ is the computed polar factor of A and $\widehat{H}$ is the computed Hermitian factor given by $\widehat{H} = \frac{1}{2}(\widehat{U}^*A + (\widehat{U}^*A)^*)$. Recalling (2.12), we measure the backward error by the residual norm $\mathrm{res} = \|A - \widehat{U}\widehat{H}\|_F/\|A\|_F$ and say the solution is backward stable if it is smaller than $c\epsilon_{\mathrm{M}}$ for a moderate constant c. We note that the analysis in Section 3.4 shows that $H - \widehat{H}$, the error in H, is also small if the residual is small and $\widehat{U}_p$ is numerically orthogonal.

Example 1. This example shows the effectiveness of the dynamical weighting in terms of the number of iterations. Let A be 20×20 diagonal matrices such that the diagonal elements form a geometric series with $a_{11} = 1/\kappa$ and $a_{nn} = 1$. The condition numbers of the matrices A are $\kappa = 10, 10^2, 10^5, \ldots, 10^{20}$. The reason for picking a diagonal matrix is to minimize the effects of rounding errors. The following data show the iteration counts and residual norms of three variants of Halley's method.

κ		10	10^2	10^5	10^{10}	10^{15}	10^{20}
iter	Halley (3.8)	5	7	14	24	35	45
	Gander (3.24)	6	7	9	14	18	24
	DWH (3.10)	4	4	5	5	6	6
res	Halley (3.8)	4.7e-16	5.4e-16	2.4e-16	1.1e-16	1.0e-16	1.1e-16
	Gander (3.24)	7.6e-16	7.5e-16	8.0e-16	7.4e-16	8.0e-16	6.4e-16
	DWH (3.10)	4.9e-16	3.8e-16	3.1e-16	5.7e-16	6.6e-16	5.4e-16

From the above table, we see that Gander's iteration is faster than Halley's iteration but still increases substantially with the increase of the condition numbers. The DWH iteration converges the fastest, all within six steps as predicted in Section 3.2.

Example 2. The purpose of this example is to show that three variants of the Newton iteration are numerically unstable. Consider the simple 3×3 matrix $A = U\Sigma V^T$, where

$\Sigma = \mathrm{diag}\{10^8, 1, 10^{-8}\}$,

$$U = \begin{bmatrix} \sin\theta & 0 & \cos\theta \\ 0 & 1 & 0 \\ -\cos\theta & 0 & \sin\theta \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} \sin\theta & \cos\theta & 0 \\ -\cos\theta & \sin\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \theta = \pi/3.$$

The following table shows that three variants of Newton's iteration, namely, the NV iteration (3.5), SNV iteration (3.6), and QSNV iteration (3.33), are numerically unstable. The QR-based implementation in the QSNV iteration improves the backward stability, but it is still not numerically backward stable to machine precision.

	SN	NV	SNV	QSNV	DWH	QDWH
iter	9	31	9	9	6	6
res	1.1e-16	4.9e-3	5.1e-3	1.1e-9	3.1e-11	3.3e-16

The instability of the SNV method, including QSNV, has been observed in previous studies [**26, 19**]. This numerical observation led us to give up QSNV and turn to the study of a Halley-type iteration. We note that, from the last column of the previous table, the QDWH method performed in a backward stable manner to machine precision.

Example 3. The purpose of this example is to compare the SN and QDWH methods on numerical stability and convergence rate. The bidiagonal reduction-based matrix inversion method is used in the SN method to guarantee the numerical backward stability. We construct three groups of 20×20 test matrices using the MATLAB function `gallery('randsvd', 20, kappa)`, where the condition number `kappa` is set to be 10^2, 10^8, and 10^{15}, respectively. The following table shows the minimum and maximum numbers of iterations and residual norms from 100 test runs.

$\kappa_2(A)$		10^2		10^8		10^{15}	
		min	max	min	max	min	max
iter	QDWH	4	5	5	5	6	6
	SN	6	6	8	8	9	9
res	QDWH	4.2e-16	7.8e-16	4.7e-16	8.1e-16	2.8e-16	7.1e-16
	SN	4.3e-16	6.5e-16	5.8e-16	9.5e-16	3.4e-16	1.2e-15

We observe that both SN and QDWH methods exhibit excellent numerical stability. The QDWH method needs about two-thirds as many iterations as the SN method does, as discussed in Section 3.2. We have also tested many other types of matrices such as extremely ill-conditioned Hilbert matrices. In all our experiments, the QDWH method converged within six iterations and performed in a backward stable manner.

Example 4. In this example, we examine the sufficiency of the QDWH stopping criterion (3.28), which is looser than the one used for SN and QSNV. We generated 100 test matrices as in Example 3, where the condition number `kappa` is set to be 10^8. The following table shows the values $\|X_k - X_{k-1}\|_F$, the corresponding residual norms $\|A - \widehat{U}\widehat{H}\|_F/\|A\|_F$, and the distance from orthogonality $\|X_k^* X_k - I\|_F$ at the iterations $k = 4, 5, 6$.

k	4		5		6	
	min	max	min	max	min	max
$\|X_k - X_{k-1}\|_F$	4.2e-2	6.1e-2	1.7e-7	5.1e-7	1.5e-15	2.4e-15
res	6.6e-8	2.2e-7	4.7e-16	8.1e-16	4.8e-16	7.8e-16
$\|X_k^* X_k - I\|_F$	3.6e-7	1.0e-6	1.9e-15	3.0e-15	2.0e-15	3.2e-15

As we can see, when the QDWH stops at $k = 5$ after satisfying the stopping criterion (3.28), the residual norms and the distance from orthogonality are at the order of 10^{-15} or smaller. Therefore, the stopping criterion (3.28) is a reliable and realistic stopping criterion.

Example 5. In this example, we investigate the impact of estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of $\alpha = \sigma_{\max}(A)$ and $\beta = \sigma_{\min}(A)$ on the convergence of the QDWH method. Since $\|A\|_F/\sqrt{n} \le \|A\|_2 \le \|A\|_F$, $\widehat{\alpha} = \|A\|_F$ is a safe and reliable choice (see Remark 2). Let us focus on the effect of the estimate $\widehat{\beta}$. Let $A \in \mathbb{R}^{20\times 20}$ be generated by using `randsvd` as in Example 3 with $\kappa_2(A) = 10^8$. The following table shows the number of QDWH iterations and residual errors for different estimates $\widehat{\beta}$.

$\widehat{\beta}/\beta$	10^{-9}	10^{-6}	10^{-3}	1	10^3	10^6	10^9
iter	6	6	6	5	12	18	24
res	5.8e-16	6.2e-16	7.3e-16	5.8e-16	6.1e-16	8.2e-16	9.3e-16

These results suggest that a severely overestimated $\widehat{\beta}$ slows down the convergence substantially but that an underestimated $\widehat{\beta}$ is essentially harmless on the convergence rate and numerical stability. We further performed many tests for other types of matrices and drew the same conclusion. Hence, in practice, it is important to make sure that $\widehat{\beta} \le \sigma_{\min}(A)$ if possible. This observation has led us to use the estimate in (3.36). Why such crude estimates of $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ work so well is a topic of future study.

3.6. Solving the max-min problem

The reader can skip this section without loss of continuity.

In this section, we consider an analytic solution of the optimization problem (3.19) and (3.20). In [**121**], Nie describes a scheme to reformulate the problem as a standard semidefinite programming (SDP) problem so that we can solve it by using an SDP software such as SeDuMi [**148**].

Let us restate the optimization problem (3.19) and (3.20) as follows:

Let

$$g(x; a, b) = \frac{x(a + bx^2)}{1 + (a + b - 1)x^2},$$

where $(a, b) \in \mathcal{D} = \{(a, b) \mid a > 0, b > 0 \text{ and } a + b > 1\}$. Let ℓ be a prescribed constant and $0 < \ell \le 1$. Find $(a_*, b_*) \in \mathcal{D}$ such that

$$0 < g(x; a_*, b_*) \le 1 \quad \text{for} \quad \ell \le x \le 1, \tag{3.56}$$

and (a_*, b_*) attains the max-min

$$\max_{(a,b)\in\mathcal{D}} \left\{ \min_{\ell\le x\le 1} g(x;a,b) \right\}. \tag{3.57}$$

We first consider the case $0 < \ell < 1$ and treat the case $\ell = 1$ at the end.

3.6.1. Partition of $\mathcal{D}$. First we note that $g(x;a,b)$ is a continuously differentiable odd function of x. The first and second partial derivatives of $g(x;a,b)$ with respect to x are

$$\partial_x g(x;a,b) = \frac{b(a+b-1)x^4 - (a(a+b-1)-3b)x^2 + a}{(1+(a+b-1)x^2)^2} \tag{3.58}$$

and

$$\partial_{xx} g(x;a,b) = \frac{2(a-1)(a+b)x((a+b-1)x^2-3)}{(1+(a+b-1)x^2)^3}. \tag{3.59}$$

The derivative of $g(x;a,b)$ with respect to a is given by

$$\partial_a g(x;a,b) = \frac{x(1-x^2)(1+bx^2)}{(1+(a+b-1)x^2)^2}. \tag{3.60}$$

It is easy to see that $g(x;a,b)$ is a strictly increasing function of a on $0 < x < 1$.

By some basic algebra manipulation, we derive the following lemma.

LEMMA 3.3. *Consider the domain $(a,b) \in \mathcal{D}$. If $a > \Gamma \equiv \frac{1}{2}(1-b+\sqrt{1+34b+b^2})$, then $g(x;a,b)$ has two real positive critical points $0 < x_m(a,b) < x_M(a,b)$. If $a = \Gamma$, then $g(x;a,b)$ has one critical point $0 < x_m(a,b)$. If $a < \Gamma$, then $g(x;a,b)$ has no real critical points. Furthermore, $x_m(a,b) > 1$ if and only if $1 < a < 3$ and $a < b+2$, and $x_M(a,b) > 1$ if and only if $1 < a < 3$ or $a > b+2$.*

In view of Lemma 3.3, we partition $\mathcal{D}$ into the following four domains:

- $\mathcal{D}_{\rm I} = \{(a,b) \mid a > b+2\}$.
- $\mathcal{D}_{\rm II} = \{(a,b) \mid \Gamma \le a \le b+2,\ b \ge 1\}$.
- $\mathcal{D}_{\rm III} = \{(a,b) \mid 1-b < a < \Gamma\}$.
- $\mathcal{D}_{\rm IV} = \{(a,b) \mid \Gamma \le a \le b+2,\ b < 1\}$.

These four domains are illustrated by Figure 3.6.1.

3.6.2. Exclusion of $\mathcal{D}_{\rm I}$, $\mathcal{D}_{\rm III}$, and $\mathcal{D}_{\rm IV}$. We show that domains $\mathcal{D}_{\rm I}$, $\mathcal{D}_{\rm III}$, and $\mathcal{D}_{\rm IV}$ can be immediately excluded for further considerations since, when (a,b) are in these domains, either the condition (3.56) is violated or there is no maximum value satisfying (3.57).

When $(a,b) \in \mathcal{D}_{\rm I}$, $g(x;a,b)$ has the critical points $x_m(a,b) < 1$ and $x_M(a,b) > 1$. By (3.58), we have $\partial_x g(1;a,b) < 0$. Noting that $g(1;a,b) = 1$, there must be an x such that $\ell < x \le 1$ and $g(x;a,b) > 1$. This violates the constraint (3.56). Hence, domain $\mathcal{D}_{\rm I}$ is excluded from further consideration.

When $(a,b) \in \mathcal{D}_{\rm III}$, $g(x;a,b)$ has no critical point. By (3.58), we have $\partial_x g(x;a,b) > 0$ for $x \in [0,1]$, so $g(x;a,b)$ is strictly increasing on $[0,1]$. In addition, $g(0;a,b) = 0$, and $g(1;a,b) = 1$. The condition (3.56) is satisfied. However, it follows from (3.60) that $h(a,b) = \min_{\ell\le x\le 1} g(x;a,b)$ is a strictly increasing function of a. Since $\mathcal{D}_{\rm III}$ is right-end open

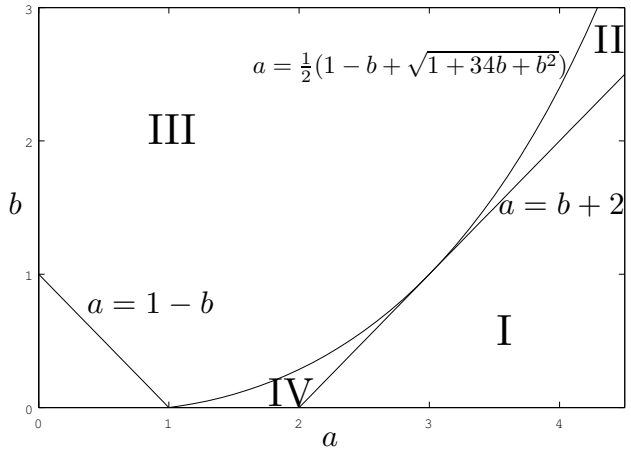

FIGURE 3.6.1. Partition of domain $\mathcal{D}$.

with respect to a, i.e., the boundary curve $a = \Gamma$ is not included, $h(a, b)$ will not have a maximum on $\mathcal{D}_{\mathrm{III}}$.[6] Hence the domain $\mathcal{D}_{\mathrm{III}}$ can be removed from consideration.

Finally, when $(a, b) \in \mathcal{D}_{\mathrm{IV}}$, the critical points $x_m(a, b), x_M(a, b) > 1$. Similar to the discussion of domain $\mathcal{D}_{\mathrm{III}}$, we can show that $\partial_x g(x; a, b) > 0$ on $x \in [0, 1]$ and that $g(0; a, b) = 0$ and $g(1; a, b) = 1$. Hence, the condition (3.56) is satisfied. By (3.60), $h(a, b) = \min_{l \le x \le 1} g(x; a, b)$ is a strictly increasing function of a. Since $\mathcal{D}_{\mathrm{IV}}$ includes the boundary line $a = b + 2$, $h(a, b)$ has the maximum (with respect to a) only on the boundary line $a = b + 2$. On the boundary line, $g(x; b + 2, b)$ is an increasing function of b since $\partial_b g(x; b + 2, b) > 0$. Hence, $H(b) = \min_{\ell \le x \le 1} g(x; b + 2, b)$ is a strictly increasing function of b. However, $\mathcal{D}_{\mathrm{IV}}$ does not include the point $(a, b) = (3, 1)$; therefore, $H(b)$ has no maximum. Consequently, domain $\mathcal{D}_{\mathrm{IV}}$ can be removed from consideration.

3.6.3. Searching on domain $\mathcal{D}_{\mathrm{II}}$. Let us focus on domain $\mathcal{D}_{\mathrm{II}}$. When $(a, b) \in \mathcal{D}_{\mathrm{II}}$, the critical points satisfy $x_m(a, b), x_M(a, b) \le 1$. (We define $x_M(a, b) = x_m(a, b)$ when $a = \Gamma$.) By (3.59), we have $\partial_{xx} g(x; a, b) \le 0$ at $x = x_m(a, b)$ and $\partial_{xx} g(x; a, b) \ge 0$ at $x = x_M(a, b)$, where both equalities hold only when $a = \Gamma$. Therefore, we have the following lemma.

LEMMA 3.4. *When $(a, b) \in \mathcal{D}_{\mathrm{II}}$ and $a > \Gamma$, $g(x; a, b)$ has the local maximum at $x_m(a, b)$ and the local minimum at $x_M(a, b)$. When $(a, b) \in \mathcal{D}_{\mathrm{II}}$ and $a = \Gamma$, $g(x; a, b)$ is monotonically increasing on $[0, 1]$.*

3.6.3.1. *Further partition of $\mathcal{D}_{\mathrm{II}}$.* To find the subregion $\mathcal{D}_{\mathrm{II}}^0$ of $\mathcal{D}_{\mathrm{II}}$ in which (3.56) is satisfied, let us further divide domain $\mathcal{D}_{\mathrm{II}}$ into two subdomains:

- $\mathcal{D}_{\mathrm{II}}^a = \{(a, b) \mid (a, b) \in \mathcal{D}_{\mathrm{II}} \text{ and } x_m(a, b) < \ell\}$.
- $\mathcal{D}_{\mathrm{II}}^b = \{(a, b) \mid (a, b) \in \mathcal{D}_{\mathrm{II}} \text{ and } x_m(a, b) \ge \ell\}$.

[6]Here we are using a basic result from calculus that says a strictly increasing function $f(x)$ has no maximum value on a right-open interval.

When $(a,b) \in \mathcal{D}_{\rm II}^a$, by Lemma 3.4, we know that $g(x;a,b)$ does not have a local maximum on $[\ell,1]$. Since a differentiable function on a closed interval takes its maximum at either the endpoints or the local maximum, we have $\max_{\ell \le x \le 1} g(x;a,b) = \max\{g(\ell;a,b), g(1;a,b)\}$. Noting that $g(1;a,b) = 1$, we have the following lemma.

LEMMA 3.5. *For $(a,b) \in \mathcal{D}_{\rm II}^a$, $g(\ell;a,b) \le 1$ is the necessary and sufficient condition to meet* (3.56).

We now show that the condition $g(\ell;a,b) \le 1$ is violated for (a,b) in a subset of $\mathcal{D}_{\rm II}^a$. Consider the case $g(\ell;a,b) = 1$. It implies that $a = b\ell + 1 + 1/\ell \ \equiv a_1(b)$. Let us further partition $\mathcal{D}_{\rm II}^a$ into two subdomains:

- $\mathcal{D}_{\rm II}^{a,1} = \{(a,b) \mid (a,b) \in \mathcal{D}_{\rm II}^a \text{ and } a \le a_1(b)\}$.
- $\mathcal{D}_{\rm II}^{a,2} = \{(a,b) \mid (a,b) \in \mathcal{D}_{\rm II}^a \text{ and } a > a_1(b)\}$.

When $(a,b) \in \mathcal{D}_{\rm II}^{a,1}$, by (3.60), $g(\ell;a,b)$ is a strictly increasing function of a. Since $g(\ell;a_1(b),b) = 1$, it follows that, for any $\Delta a \ge 0$, we have $g(\ell;a_1(b) - \Delta a, b) \le 1$. Using Lemma 3.5 and noting that any point in $\mathcal{D}_{\rm II}^{a,1}$ can be written as $(a_1(b) - \Delta a, b)$ for some $\Delta a \ge 0$, it follows that, for $(a,b) \in \mathcal{D}_{\rm II}^{a,1}$, the condition (3.56) is met.

When $(a,b) \in \mathcal{D}_{\rm II}^{a,2}$, we have $g(\ell;a,b) > 1$, and so (3.56) is violated since $g(\ell;a_1(b) + \Delta a, b) > 1$ for any $\Delta a > 0$. Therefore, $\mathcal{D}_{\rm II}^{a,2}$ is excluded from further consideration.

Next consider $(a,b) \in \mathcal{D}_{\rm II}^b$. By Lemma 3.4, $g(x;a,b)$ is increasing on $[\ell, x_m(a,b)]$, decreasing on $[x_m(a,b), x_M(a,b)]$, and increasing on $[x_M(a,b),1]$. Therefore, it follows that $\max_{\ell \le x \le 1} g(x(a,b);a,b) = \max\{g(x_m(a,b);a,b), g(1;a,b)\}$. Noting that $g(1;a,b) = 1$, we have the following result.

LEMMA 3.6. *For $(a,b) \in \mathcal{D}_{\rm II}^b$, $g(x_m(a,b);a,b) \le 1$ is the necessary and sufficient condition to meet* (3.56).

We show that the condition $g(x_m(a,b);a,b) \le 1$ is violated for (a,b) in a subset of $\mathcal{D}_{\rm II}^b$. Consider the case $g(x_m(a,b);a,b) = 1$, which implies $a = 2\sqrt{b} + 1 \ \equiv a_2(b)$, which we get by solving $g(x_m(a,b);a,b) = 1$ and $\partial_x g(x_m(a,b);a,b) = 0$ for a. Let us first partition $\mathcal{D}_{\rm II}^b$ into two subdomains:

- $\mathcal{D}_{\rm II}^{b,1} = \{(a,b) \mid (a,b) \in \mathcal{D}_{\rm II}^b \text{ and } a \le a_2(b)\}$.
- $\mathcal{D}_{\rm II}^{b,2} = \{(a,b) \mid (a,b) \in \mathcal{D}_{\rm II}^b \text{ and } a > a_2(b)\}$.

By the same argument as the one we used to exclude domain $\mathcal{D}_{\rm II}^{a,2}$, we can show that (3.56) is satisfied when $(a,b) \in \mathcal{D}_{\rm II}^{b,1}$ and is violated when $(a,b) \in \mathcal{D}_{\rm II}^{b,2}$. Therefore, $\mathcal{D}_{\rm II}^{b,2}$ is excluded.

3.6.3.2. *Characterization of* $\boldsymbol{\mathcal{D}_{\rm II}^0}$. By the above arguments we conclude that, only when $(a,b) \in \mathcal{D}_{\rm II}^0 = \mathcal{D}_{\rm II}^{a,1} \bigcup \mathcal{D}_{\rm II}^{b,1}$, the condition (3.56) is satisfied. We next identify the boundary of $\mathcal{D}_{\rm II}^0$. We first note that the line $a = b + 2$ cannot be the boundary of $\mathcal{D}_{\rm II}^0$ since on the line, $\partial_x g(1;b+2,b) = 0$ and $\partial_{xx} g(1;b+2,b) > 0$, there exists x such that $\ell < x \le 1$ and $g(x;a,b) > 1$, which violates the the condition (3.56). Consequently, the boundary of $\mathcal{D}_{\rm II}^0$ consists of the following:

- $a = \Gamma$ and
- $a = a_1(b)$ and $x_m(a,b) < \ell$ or

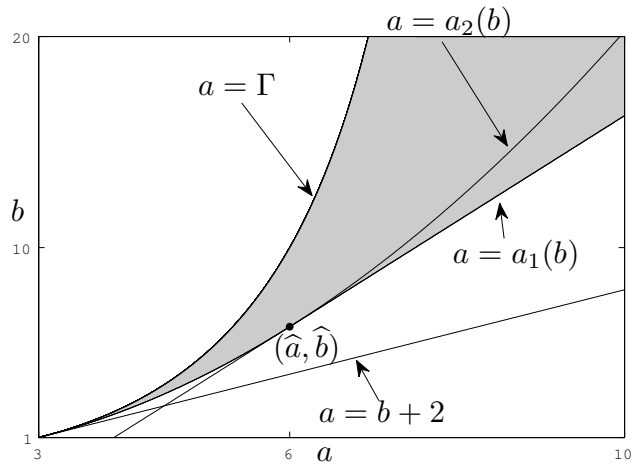

FIGURE 3.6.2. Shaded region is $\mathcal{D}_{\rm II}^0$ for $\ell = 0.4$.

- $a = a_2(b)$ and $x_m(a, b) \geq \ell$.

Basic algebra shows that $a_1(b) > \Gamma$ and $a_2(b) > \Gamma$ on $b \geq 1$, so $a = \Gamma$ is the left-side boundary of $\mathcal{D}_{\rm II}^0$. To determine the right-side boundary, we note that $a_1(b)$ is the tangent line of the curve $a_2(b)$ at $(\widehat{a}, \widehat{b}) \equiv (\frac{2+\ell}{\ell}, \frac{1}{\ell^2})$. This also means that $x_m(\widehat{a}, \widehat{b}) = \ell$. Furthermore, through basic algebra manipulation, we can verify that

(i) $\frac{d}{db} x_m(a_1(b), b) < 0$ on $b > 1/\ell^2$,
(ii) $\frac{d}{db} x_m(a_2(b), b) < 0$ on $b \geq 1$.

From the above facts, we conclude that the right-side boundary with respect to a of $\mathcal{D}_{\rm II}^0$ is $a = a_2(b)$ for $1 \leq b \leq \widehat{b}$ and $a = a_1(b)$ for $\widehat{b} > b$. Using (3.60), we see that any point on the left of this boundary satisfies (3.56), so we conclude that

$$\mathcal{D}_{\rm II}^0 = \{(a, b) \mid (\Gamma \leq a \leq a_2(b) \text{ and } 1 \leq b \leq \widehat{b}) \text{ and } (\Gamma \leq a \leq a_1(b) \text{ and } b > \widehat{b})\}. \tag{3.61}$$

The shaded region in Figure 3.6.2 illustrates the region $\mathcal{D}_{\rm II}^0$ for the case $\ell = 0.4$.

3.6.3.3. *Optimal solution on boundary of* $\boldsymbol{\mathcal{D}_{\rm II}^0}$. Now we need to consider only the region $\mathcal{D}_{\rm II}^0$, defined in (3.61). Recall that $h(a, b) = \min_{\ell \leq x \leq 1} g(x; a, b)$ is a strictly increasing function of a. Hence, for a fixed b, $h(a, b)$ can be maximized at only the right-side boundary. Therefore, the optimal solution will occur on the right-side boundary of $\mathcal{D}_{\rm II}^0$, i.e., on the curve $a = a_2(b)$ for $b \in [1, \widehat{b}]$ or the line $a = a_1(b)$ for $b \in (\widehat{b}, \infty)$.

In fact, the line $a = a_1(b)$ for $b \in (\widehat{b}, \infty)$ is removed from consideration. This is because $\partial_b g(x; a_1(b), b) < 0$ on $\ell < x < 1$ and $\min_{\ell \leq x \leq 1} g\,(x; a_1(b), b)$ is a strictly decreasing function of b and does not reach its maximum on the left-open interval $b \in (\widehat{b}, \infty)$.

Now let us consider the curve $a = a_2(b)$ for $b \in [1, \widehat{b}]$. Rewriting $a = a_2(b) = 2\sqrt{b} + 1$ as a function of a, we have

$$b_2(a) = (a - 1)^2/4, \quad 3 \leq a \leq \widehat{a}. \tag{3.62}$$

By Lemma 3.4 and the fact that $x_m(a,b) \geq \ell$ on the curve $a = a_2(b)$, we know that $g(x; a, b_2(a))$ is increasing on $x \in [\ell, x_m(a,b)]$, decreasing on $x \in [x_m(a,b), x_M(a,b)]$, and increasing again on $x \in [x_M(a,b), 1]$. It follows that

$$\min_{\ell \leq x \leq 1} g(x; a, b_2(a)) = \min\{s_1(a), s_2(a)\}, \tag{3.63}$$

where

$$s_1(a) \equiv g(\ell; a, b_2(a)) = \frac{\ell(4a + (a-1)^2\ell^2)}{4 + (a+3)(a-1)\ell^2},$$
$$s_2(a) \equiv g(x_M(a,b); a, b_2(a)) = \frac{4a^{3/2}}{(a+3)\sqrt{(a+3)(a-1)}}.$$

The following lemma is readily verified.

LEMMA 3.7. *$s_1(a)$ is increasing and $s_2(a)$ is decreasing on $a \in [3, \widehat{a}]$. Furthermore, $s_1(3) \leq s_2(3)$, and $s_1(\widehat{a}) \geq s_2(\widehat{a})$.*

Lemma 3.7 implies that there exists $a_* \in [3, \widehat{a}]$ such that

$$s_1(a_*) = s_2(a_*). \tag{3.64}$$

Solving (3.64) for a_* yields $a_* = h(\ell)$, where $h(\ell)$ is as defined in (3.22). Note that Lemma 3.7 also implies that $\min_{\ell \leq x \leq 1} g(x; a, b_2(a))$ is increasing on $a \in [3, a_*]$ and decreasing on $a \in [a_*, \widehat{a}]$ with respect to a. Therefore, $\min_{\ell \leq x \leq 1} g(x; a, b_2(a))$ is maximized at $a = a_*$.

By (3.62), the optimal value of b is given by $b_* = \frac{1}{4}(a_* - 1)^2$. (a_*, b_*) attains the max-min in (3.57), and the value is given by

$$g(\ell; a_*, b_*) = \max_{a,b \in \mathcal{D}} \{ \min_{\ell \leq x \leq 1} g(x; a, b) \} = \frac{\ell(a_* + b_*\ell^2)}{1 + (a_* + b_* - 1)\ell^2}.$$

The max-min value $g(\ell; a_*, b_*)$ is used to update ℓ in (3.23). Finally, we note that if $\ell = 1$, the solution gives $a_* = 3$ and $b_* = 1$. In this case, the DWH iteration (3.10) and the Halley iteration (3.8) coincide.

3.6.4. Proofs of $\frac{d}{db} x_m(a_1(b), b) < 0$ on $b \geq 1$ and $\frac{d}{db} x_m(a_2(b), b) < 0$ on $b \geq 1$. First we prove $\frac{d}{db} x_m(a_1(b), b) < 0$ on $b \geq 1$, which is easier. Solving $\partial_x g(x; a_2(b), b) = 0$ for x yields $x = \frac{1}{\sqrt{b}}, \sqrt{\frac{1+2\sqrt{b}}{2\sqrt{b}+b}}$. Since $x_m(a_2(b), b) \leq x_M(a_2(b), b)$ it follows that $x_m(a_2(b), b) = \frac{1}{\sqrt{b}}$, hence $\frac{d}{db} x_m(a_2(b), b) < 0$.

We next prove $\frac{d}{db} x_m(a_2(b), b) < 0$ on $b \geq 1$. Solving $\partial_x g(x; a_2(b), b) = 0$ for x yields

$$x_m(a_1(b), b) = \frac{4\beta}{\sqrt{2}b\ell(1 + b\ell(1+\ell))t(b,\ell)},$$

where $\alpha = 1 + \ell + b\ell + b\ell^3 + b^2\ell^3 + b^2\ell^4$, $\beta = b\ell^2\left(1 + \ell + b\ell^2\right)(1 + b\ell(1+\ell))$ and $t(b,\ell) = \alpha - \sqrt{\alpha^2 - 4\beta}$. Note that

$$\alpha^2 - 4\beta = (1+\ell)(1+b\ell)\left(1 + b\ell^2\right)\left(1 + \ell + b\ell - 6b\ell^2 + b\ell^3 + b^2\ell^3 + b^2\ell^4\right) > 0,$$

because

$$\begin{aligned}
&1+\ell+b\ell-6b\ell^2+b\ell^3+b^2\ell^3+b^2\ell^4\\
=&\ell^3(1+\ell)b^2+\ell(1-6\ell+\ell^2)b+\ell+1\\
=&\ell^3(1+\ell)\left(b+\frac{(1-6\ell+\ell^2)}{2\ell^2(1+\ell)}\right)^2-\frac{(1-6\ell+\ell^2)^2}{4\ell(1+\ell)}+\ell+1\\
>&\ell^3(1+\ell)\left(\frac{1}{\ell^2}+\frac{(1-6\ell+\ell^2)}{2\ell^2(1+\ell)}\right)^2-\frac{(1-6\ell+\ell^2)^2}{4\ell(1+\ell)}+\ell+1\\
=&2(\ell-2+1/\ell)=2(\sqrt{\ell}-1/\sqrt{\ell})^2>0,
\end{aligned}$$

where we used $b>\frac{1}{\ell^2},\frac{(1-6\ell+\ell^2)}{2\ell^2(1+\ell)}\geq\frac{1}{\ell^2}$ to get the first inequality and $0<\ell<1$ to get the last. Therefore,

$$\begin{aligned}
&\frac{d}{db}x_m(a_1(b),b)\\
&=\frac{-2\sqrt{2}\ell\left(1+\ell+b\ell^2\right)(t_1+t_2+t_3)}{\left(\alpha-\sqrt{\alpha^2-4\beta}\right)^2},
\end{aligned}$$

where

$$\begin{aligned}
t_1&=\ell+\ell^3+2b\ell^3+2b\ell^4,\\
t_2&=\ell^2(1+\ell+b\ell+b\ell^3+b^2\ell^3+b^2\ell^4+\sqrt{\alpha^2-4\beta}),\\
t_3&=\frac{\ell(1+\ell)}{\sqrt{\alpha^2-4\beta}}\left(1+(-2+b)\ell+(1-3b)\ell^2+3(-1+b)b\ell^3+\left(b-6b^2\right)\ell^4+b^2(3+2b)\ell^5+2b^3\ell^6\right).
\end{aligned}$$

Since t_1 and t_2 are clearly positive, it follows that to prove $\frac{d}{db}x_m(a_1(b),b)<0$ it suffices to show $t_3>0$, that is,

$$1+(-2+b)\ell+(1-3b)\ell^2+3(-1+b)b\ell^3+\left(b-6b^2\right)\ell^4+b^2(3+2b)\ell^5+2b^3\ell^6>0.$$

We can show this by using $0<\ell<1$ and $b>\frac{1}{\ell^2}$ as follows.

$$\begin{aligned}
&1+(-2+b)\ell+(1-3b)\ell^2+3(-1+b)b\ell^3+\left(b-6b^2\right)\ell^4+b^2(3+2b)\ell^5+2b^3\ell^6\\
>&(1-2\ell+\ell^2)+b\ell-3b\ell^2-3b\ell^3+5b^2\ell^3+b\ell^4-4b^2\ell^4+3b^2\ell^5\\
>&(1-\ell)^2+3b(b\ell^2-1)\ell^3+4(1-\ell)b^2\ell^3+b\ell(1+\ell^3-3\ell+b\ell^2)\\
>&(1-\ell)^2+4(1-\ell)b^2\ell^3+b\ell(1-\ell)^2(2+\ell)>0. \quad \square
\end{aligned}$$

3.7. Application in molecular dynamics simulations

The power of recent supercomputers has enabled Molecular dynamcis (MD) simulations of order nanoseconds with $O(10^3)$ atoms. At each time step of the MD simulation, one needs to compute an approximate electronic ground state by solving the Kohn-Sham equation with relatively high accuracy [**42**]. Solving the Kohn-Sham equation is often done by the Self-Consistent Field (SCF) iterations (e.g., [**108**]).

Here we discuss a strategy for obtaining a trial wavefunction via an extrapolation scheme for accelerating the SCF convergence in the MD simulation, first introduced in [**4**]. Such

obtained trial wavefunctions often speed up the SCF convergence, thereby significantly enhancing the computational speed. Here, the process called subspace alignment must be performed before the extrapolation can be applied. The conventional method for both processes involves computing a matrix inverse, which is a potential source of numerical instability, and also computationally demanding because pivoting, which is often required when computing matrix inverses, causes high communication cost, which has exceeded arithmetic cost by orders of magnitude [**61, 9, 138**].

In MD simulation, at each time step t we solve the Kohn-Sham equation

$$\mathcal{H}(\rho(\Phi(t)), t)\Phi(t) = \Phi(t)\Lambda(t), \tag{3.65}$$

where $\mathcal{H}(\rho, t) \in \mathbb{C}^{N\times N}$ is the discretized Hamiltonian using plane-wave discretization, $\Phi(t)$ is N-by-m ($N \gg m$, N is the number of basis functions and m is the number of electrons)and represents the wavefunction at time step t, and $\rho(\Phi) = \mathrm{diag}(\sum_i f_i\phi_i\phi_i^*)$ is the charge density, where f_i is a certain weight function.

The KS equation (3.65) is a nonlinear eigenvalue problem in which the source of nonlinearity is in the eigenvector dependency of the Hamiltonian, and we are interested in computing the m smallest eigenpairs. One common method for solving (3.65) is by means of the Self-Consistent Field (SCF) iterations, in which we first fix the Hamiltonian $\mathcal{H}(\rho_{in}, t)$, compute the linear eigenvalue problem

$$\mathcal{H}(\rho, t)\Phi = \Phi\Lambda(t),$$

and then compare ρ_{in} and $\rho_{out} = \mathrm{diag}(\sum_{i=0}^{m} f_i\phi_i\phi_i^*)$, where $\Phi = [\phi_1\ \phi_2\ \ldots \phi_m]$.

Suppose we have computed $\Phi(t)$ and $\Phi(t-1)$, the electronic ground state wavefunctions at the current and previous time steps, respectively. The simplest choice of trial wavefunction $\widehat{\Phi}$ for the next time step $\Phi(t+1)$ is to use $\Phi(t)$. To get a better trial wavefunction, in [**4**] a linear extrapolation technique is proposed, which computes $\widehat{\Phi}$ by

$$\widehat{\Phi} = 2\Phi(t) - \Phi(t-1). \tag{3.66}$$

However, a direct application of (3.66) does not work because $\Phi(t)$ is frequently a discontinuous function of t, for the following three possible reasons. First, if there exist multiple eigenvalues in the Kohn-Sham equation, arbitrary rotations may be formed in the wavefunctions corresponding to those states. Second, one might be computing only the subspace spanned by the eigenvectors rather than the each eigenvector in the Kohn-Sham energy minimization, since the Kohn-Sham energy is unitarily invariant, i.e., $E(\Phi) = E(\Phi Q)$ for any unitary matrix Q. Finally, $\Phi(t)$ becomes discontinuous also when Kohn-Sham eigenstates cross the Fermi level [**4**].

To overcome this difficulty, one needs to apply a unitary transformation (such as "rotation") U to bring back $\Phi(t)$ to the same "manifold" as $\Phi(t-1)$. We do this by solving the orthogonal Procrustes problem

$$\min_{U^HU=I} \|\Phi(t-1) - \Phi(t)U\|_F. \tag{3.67}$$

The solution U to this problem is known to be the unitary polar factor of the matrix $\Phi(t)^T\Phi(t-1)$.

After solving (3.67), we perform the linear extrapolation

$$\widehat{\Phi} = 2\Phi(t)U - \Phi(t-1). \tag{3.68}$$

In this way, the the subspaces span$\{\Phi(t-1)\}$ and span$\{\Phi(t)\}$ are "aligned". The process is called *subspace-alignment* [**4**].

CHAPTER 4

Efficient, communication minimizing algorithm for the symmetric eigenvalue problem

In this chapter, we propose a new algorithm QDWH-eig for computing the symmetric eigenvalue decomposition. QDWH-eig minimizes communication in the asymptotic sense while simultaneously having arithmetic operation costs within a factor 3 of that for the most efficient existing algorithms. The essential cost for each of the two algorithms is in performing QR decompositions, of which we require no more than 6 for matrices of the original size. We establish backward stability of the algorithms under mild assumptions. Compared with similar known communication-minimizing algorithms based on spectral divide-and-conquer, QDWH-eig is highly preferable in terms of both speed and stability. In our preliminary numerical experiments using a sequential machine with 4 cores, QDWH-eig required the same order of runtime as the best available standard algorithms. Their performance is expected to improve substantially on highly parallel computing architectures where communication dominates arithmetic.

Introduction. Our focus in this chapter is to devise an algorithm for the symmetric eigendecomposition that minimize both communication (asymptotically, that is, in the big-Oh sense) and arithmetic (up to a small constant factor compared with the best available algorithm, < 3 in our algorithm).

Some recent progress has been made towards devising communication-minimizing algorithms for the symmetric eigendecomposition and the singular value decomposition. Ballard, Demmel and Dumitriu [**8**] propose a family of spectral divide-and-conquer algorithms, which we call BDD, that are applicable to eigenproblems and the SVD. BDD requires only QR decompositions and matrix multiplications, so it minimizes communication, both bandwidth and latency costs, on two-level sequential machines and parallel machines (up to polylog factors). It is argued in [**8**] that BDD converges in at most about $-\log_2 \epsilon = 53$ iterations, so the overall algorithm minimizes communication in the asymptotic, big-Ω sense. However, BDD generally needs significantly more arithmetic than conventional algorithms, and it loses backward stability when an eigenvalue exists close to a splitting point. See Sections 4.1.4, 4.1.5 and 5.1.4 for details.

As discussed in Chapter 2, conventional algorithms for the symmetric eigenproblem and the SVD initially reduce the matrix to condensed form (tridiagonal or bidiagonal), after which specialized and efficient algorithms are used to complete the decomposition. Such algorithms minimize arithmetic costs, at least up to a small constant. However, unfortunately

there is no known way of performing the initial reduction with minimal communication. An implementation that minimizes the bandwidth cost is given in [**9**], but it does not minimize the latency cost.

In this chapter we derive a new algorithm QDWH-eig for computing the symmetric eigendecomposition that asymptotically minimize communication while at the same time minimizing arithmetic cost up to small (< 3 for the eigenproblem and < 2 for the SVD) constant factors. Ours is the first algorithm to attain both minimization properties. In addition, we prove that the algoritm is backward stable. The assumption we make to prove the backward stability of QDWH-eig is that the polar decompositions computed by QDWH are backward stable.

The tool underlying QDWH-eig is the QDWH algorithm for the polar decomposition, proposed in the previous chapter. The key fact is that the positive and negative invariant subspaces of a Hermitian matrix can be efficiently computed via the unitary polar factor. This observation leads to our spectral divide-and-conquer algorithm QDWH-eig. For a Hermitian matrix $A \in \mathbb{C}^{n\times n}$, the dominant cost of QDWH-eig is in performing six or fewer QR decompositions of $2n$-by-n matrices. QDWH-eig generally converges in much fewer iterations than BDD and other spectral divide-and-conquer algorithms proposed in the literature.

We note that by requiring smaller number of iterations QDWH-eig also reduces communication cost, so it is cheaper than BDD also in communication, although in the big-Ω argument this effect is absorbed as a constant.

We perform numerical experiments with QDWH-eig on a sequential machine using a small number of processors, and employing the conventional LAPACK-implemented QR decomposition algorithm that does not minimize communication. Even under such conditions, the performance of QDWH-eig is comparable to that of the standard algorithms in terms of both speed and backward stability. On massively parallel computing architectures we expect the communication-optimality of QDWH-eig will improve the performance significantly.

This chapter is organized as follows. In Section 4.1 we develop the algorithm QDWH-eig. We then establish its backward stability in Section 4.1.4. In Section 4.1.5 we compare QDWH-eig with other algorithms for symmetric eigenproblems. Section 4.2 addresses practical implementation issues and techniques to further enhance the efficiency. Numerical experiments are shown in Section 4.3.

We develop algorithms for complex matrices $A \in \mathbb{C}^{m\times n}$, but note that if A is real then all the operations can be carried out by using only real arithmetic.

4.1. Algorithm QDWH-eig

Throughout the section A denotes a symmetric (or Hermitian) matrix. For simplicity we always call the eigendecomposition $A = V\Lambda V^*$ the symmetric eigendecomposition, whether A is real symmetric or complex Hermitian. This section develops QDWH-eig, our QR-based symmetric eigendecomposition algorithm. QDWH-eig is based on a spectral divide-and-conquer idea, which is to compute invariant subspaces corresponding to eigenvalues lying in certain intervals.

4.1.1. Computing invariant subspace via the polar decomposition. The goal here is to compute an invariant subspace of A corresponding to the positive (or negative)

eigenvalues. Suppose that A has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0 > \lambda_{k+1} \geq \cdots \geq \lambda_n$ (for simplicity we assume A is nonsingular; the singular case is discussed in Section 4.2.3). The first step is to realize the connection between the polar decomposition and the eigendecomposition of symmetric matrices. Let $A = U_p H$ be the polar decomposition and let $A = [V_1\ V_2] \begin{bmatrix} \Lambda_+ & \\ & \Lambda_- \end{bmatrix} [V_1\ V_2]^*$ be an eigendecomposition where $\mathrm{diag}(\Lambda_+) = \{\lambda_1, \ldots, \lambda_k\}$ are the positive eigenvalues. Then, U_p and V are related by $U_p = [V_1\ V_2] \begin{bmatrix} I_k & \\ & -I_{n-k} \end{bmatrix} [V_1\ V_2]^*$, because

$$U_p H = \left([V_1\ V_2] \begin{bmatrix} I_k & \\ & -I_{n-k} \end{bmatrix} [V_1\ V_2]^* \right) \cdot \left([V_1\ V_2] \begin{bmatrix} \Lambda_+ & \\ & |\Lambda_-| \end{bmatrix} [V_1\ V_2]^* \right). \tag{4.1}$$

An alternative way to understand (4.1) is to note that the polar decomposition $A = U_p H$ and matrix sign decomposition [**75**, Ch. 5] $A = (A(A^2)^{-1/2}) \cdot (A^2)^{1/2}$ are equivalent when A is symmetric. We prefer to regard (4.1) as a polar decomposition because, as we shall see, it lets us derive an SVD algorithm in a unified fashion.

Suppose we have computed U_p in (4.1) using the QDWH algorithm. Note that

$$U_p + I = [V_1\ V_2] \begin{bmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{bmatrix} [V_1\ V_2]^* + I = [V_1\ V_2] \begin{bmatrix} 2I_k & 0 \\ 0 & 0 \end{bmatrix} [V_1\ V_2]^*,$$

so the symmetric matrix $C = \frac{1}{2}(U_p + I) = V_1 V_1^*$ is an orthogonal projector onto $\mathrm{span}(V_1)$, which is the invariant subspace that we want. Hence we can compute $\mathrm{span}(V_1)$ by computing an orthogonal basis for the column space of C. One way of doing this is to perform QR with pivoting, as is suggested in [**170, 171**]. However, pivoting is expensive in communication cost.

We advocate using subspace iteration [**127**, Ch. 14] with $r = \mathrm{round}(\|C\|_F^2)$ vectors (since the eigenvalues of C are either 0 or 1, in exact arithmetic r is the precise rank of C). Subspace iteration converges with the convergence factor $|\lambda_{r+1}|/|\lambda_k|$ for the kth eigenvalue [**145**], so $\lambda_r = 1$ and $\lambda_{r+1} = 0$ means a *single* iteration of subspace iteration yields the desired subspace $\mathrm{span}(V_1)$. In practice sometimes more than one iteration is needed for subspace iteration with $\widehat{C}$ to converge, and we terminate subspace iteration and accept the computed matrices $\widehat{V}_1$ and its orthogonal complement $\widehat{V}_2$ (which we get as the orthogonal complement of $\widehat{V}_1$ via accumulating the Householder reflectors) once the conditions

$$\widehat{C}\widehat{V}_1 = \widehat{V}_1 + \epsilon \quad \text{and} \quad \widehat{C}\widehat{V}_2 = \epsilon \tag{4.2}$$

are satisfied. Recall that ϵ denotes a matrix (or scalar) of order machine epsilon, whose values differ in different appearances. We provide more details of a practical implementation of subspace iteration in Section 4.2.2.

We then have a matrix $\widehat{V} = [\widehat{V}_1\ \widehat{V}_2]$ such that $\widehat{V}^* A \widehat{V} = \left[\begin{smallmatrix} A_1 & E^* \\ E & A_2 \end{smallmatrix}\right]$. E is the backward error of the spectral divide-and-conquer, which is acceptable if $\|E\|_2/\|A\|_2 = \epsilon$.

4.1.2. Algorithm. The entire eigendecomposition can be computed by repeatedly applying the spectral divide-and-conquer algorithm on the submatrices $V_1^* A V_1$ and $V_2^* A V_2$. Algorithm 2 gives a pseudocode for QDWH-eig.

Algorithm 2 QDWH-eig: computes an eigendecomposition of a symmetric (Hermitian) matrix A

1: Choose σ, estimate of the *median* of eig(A)
2: Compute polar factor U_p of $A - \sigma I = U_p H$ by the QDWH algorithm
3: Compute $V_1 \in \mathbb{C}^{n\times k}$ such that $\frac{1}{2}(U_p + I) = V_1 V_1^*$ via subspace iteration, then form a unitary matrix $V = [V_1\ V_2]$
4: Compute $A_1 = V_1^* A V_1$ and $A_2 = V_2^* A V_2$
5: Repeat steps 1–4 with $A := A_1, A_2$ until A is diagonalized

Algorithm 2 introduces a shift σ, designed to make $A - \sigma I$ have similar numbers of positive and negative eigenvalues so that the sizes of A_1 and A_2 are about the same. σ is the splitting point that separate the subspaces $\widehat{V}_1$ and $\widehat{V}_2$. To keep the exposition simple, we defer the discussion of practical implementation issues such as the choice of σ to Section 4.2.

We note that, as mentioned in [**8**], spectral divide-and-conquer type algorithms such as QDWH-eig for symmetric eigendecompositions deal effectively with multiple (or clusters of) eigenvalues. This is because if a diagonal block A_j has all its eigenvalues lying in $[\lambda_0-\epsilon, \lambda_0+\epsilon]$, then A_j must be nearly diagonal, $A_j = \lambda_0 I + \epsilon$. Upon detecting such a submatrix A_j, in QDWH-eig we stop performing spectral divide-and-conquer on A_j and return the value of the (nearly) multiple eigenvalue λ_0, its multiplicity, along with its corresponding invariant subspace V_j which is orthogonal to working accuracy.

4.1.3. Communication/arithmetic cost. As we saw above, QDWH-eig uses the same computational kernels as the eigendecomposition algorithm BDD in [**8**], namely QR decompositions and matrix multiplications. Hence, just like BDD, QDWH-eig asymptotically minimizes communication, both bandwidth and latency costs.

As for arithmetic cost, assuming that a good splitting point σ is always taken (this assumption is nontrivial but outside the scope of this dissertation), one spectral divide-and-conquer results in two submatrices of size $\simeq n/2$. Since the arithmetic cost scales cubically with the matrix size, the overall arithmetic cost is approximately $\sum_{i=0}^{\infty}(2^{-i})^3\beta = \frac{8}{7}\beta$ flops along the critical path (noting that further divide-and-conquer of A_1, A_2 can be done in parallel), where β is the number of flops needed for one run of spectral divide-and-conquer for a n-by-n matrix. In Section 4.2.5 we examine the flop counts in detail and we show that $\frac{8}{7}\beta = 26n^3$.

4.1.4. Backward stability proof. Here we establish the backward stability of QDWH-eig. For notational simplicity we let $A \leftarrow A - \sigma I$, so that A has both negative and positive eigenvalues.

We prove that QDWH-eig is backward stable, provided that the polar decomposition $A = U_p H$ computed at step 2 of QDWH-eig is backward stable. Specifically, our assumption is that the computed factors $\widehat{U}_p, \widehat{H}$ satisfy

$$A = \widehat{U}_p \widehat{H} + \epsilon\|A\|_2, \quad \widehat{U}_p^* \widehat{U}_p - I = \epsilon. \tag{4.3}$$

THEOREM 4.1. *Suppose that the polar decompositions computed by QDWH within QDWH-eig are backward stable so that* (4.3) *holds. Then QDWH-eig computes the symmetric eigendecomposition in a backward stable manner.*

PROOF. It suffices to prove that one recursion of steps 1-4 of Algorithm 2 computes an invariant subspace of A in a backward stable manner, that is, $\|E\|_2 = \epsilon\|A\|_2$ where E contains the off-diagonal blocks of $\widehat{V}^*A\widehat{V}$. We note that $\widehat{V}^*\widehat{V} = I + \epsilon$, because $\widehat{V} = [\widehat{V}_1\ \widehat{V}_2]$ computed by the Householder QR decomposition is always unitary to working accuracy [**56**]. Together with the subspace iteration stopping criterion (4.2) and $\widehat{C} = \frac{1}{2}(\widehat{U}_p + I)$ we have

$$\frac{1}{2}(\widehat{U}_p + I)[\widehat{V}_1\ \widehat{V}_2] = [\widehat{V}_1\ 0] + \epsilon,$$

so right-multiplying $2\widehat{V}^*$ we get

$$\begin{aligned}\widehat{U}_p &= 2[\widehat{V}_1\ 0][\widehat{V}_1\ \widehat{V}_2]^* - I + \epsilon \\ &= [\widehat{V}_1\ -\widehat{V}_2][\widehat{V}_1\ \widehat{V}_2]^* + \epsilon \\ &= \widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^* + \epsilon. \qquad (4.4)\end{aligned}$$

Using (4.3) and (4.4) we get

$$\begin{aligned}A &= \left(\widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^* + \epsilon\right)\widehat{H} + \epsilon\|A\|_2 \\ &= \widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^*\widehat{H} + \epsilon\|A\|_2, \qquad (4.5)\end{aligned}$$

where we used $\|\widehat{H}\|_2 \simeq \|A\|_2$, which follows from (4.3). Hence, using $\widehat{V}^*\widehat{V} - I = \epsilon$ we obtain

$$\widehat{V}^*A\widehat{V} = \begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^*\widehat{H}\widehat{V} + \epsilon\|A\|_2.$$

Therefore, to prove $\|E\|_2 = \epsilon\|A\|_2$ it suffices to prove the off-diagonal blocks X_{21}, X_{12} of $\widehat{V}^*\widehat{H}\widehat{V} = \left[\begin{smallmatrix} X_{11} & X_{12} \\ X_{21} & X_{22}\end{smallmatrix}\right]$ can be expressed as $\epsilon\|A\|_2$. We obtain, using (4.5),

$$\begin{aligned}0 &= A - A^* \\ &= \left(\widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^*\widehat{H} + \epsilon\|A\|_2\right) - \left(\widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^*\widehat{H} + \epsilon\|A\|_2\right)^* \\ &= \left(\widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^*\widehat{H} - \widehat{H}^*\widehat{V}\begin{bmatrix} I_k & \\ & -I_{n-k}\end{bmatrix}\widehat{V}^*\right) + \epsilon\|A\|_2. \qquad (4.6)\end{aligned}$$

Here we used the fact $\widehat{H}^* = \widehat{H}$, which follows from (3.1). Hence, using $\widehat{V}^*\widehat{V} - I = \epsilon$ and multiplying (4.6) by $\begin{bmatrix} I_k & \\ & -I_{n-k} \end{bmatrix} \widehat{V}^*$ on the left and $\widehat{V}$ on the right we get

$$\begin{aligned} \epsilon\|A\|_2 &= \left(\widehat{V}^*\widehat{H}\widehat{V} - \begin{bmatrix} I_k & \\ & -I_{n-k} \end{bmatrix} \widehat{V}^*\widehat{H}\widehat{V} \begin{bmatrix} I_k & \\ & -I_{n-k} \end{bmatrix} \right) \\ &= \begin{bmatrix} 0 & 2X_{12} \\ 2X_{21} & 0 \end{bmatrix}. \end{aligned}$$

Therefore it follows that $\|X_{21}\|_2/\|A\|_2 = \epsilon$ and $\|X_{12}\|_2/\|A\|_2 = \epsilon$ as required. □

The conclusion is that provided that QDWH (for the polar decomposition) is backward stable, QDWH-eig (for the eigendecomposition) also enjoys backward stability. QDWH performed backward stably in all experiments in [**120**], but no proof is given there. We shall prove that QDWH is backward stable (thus so is QDWH-eig), provided that row and column pivoting are used for computing the QR decomposition in (3.35). To keep the focus on the algorithmic developments, we defer this analysis to Section 3.4.

To our knowledge Theorem 4.1 is the first backward stability proof for a spectral divide-and-conquer algorithm for symmetric eigenproblems that makes no assumption on the splitting point σ. This is in contrast to the backward stability analysis in [**7**], which applies to the BDD algorithm and proves that the backward error is bounded by a factor proportional to ϵ/d, where d is the smallest distance between an eigenvalue of $A - \sigma I$ and the splitting points, which are ± 1 in BDD. Hence the backward error can be large when σ is close to a splitting point (we observed the instability in our experiments, see Section 4.3.1.1). As discussed in [**7**], this is precisely when BDD's convergence is slow. Hence, in "difficult" cases in which an eigenvalue of $A - \sigma I$ lies close to a splitting point, BDD suffers from potential instability and slow convergence.

QDWH-eig has neither problem, even in such difficult cases, which corresponds to the situation where QDWH needs to compute the polar decomposition of a nearly singular matrix A. This is because QDWH converges within 6 iterations for any $\kappa_2(A) < 10^{16}$, and the backward stability of QDWH is independent of $\kappa_2(A)$, as shown in Section 3.4.

Other spectral divide-and-conquer algorithms for the symmetric eigenproblem are proposed in [**170**] and [**171**], but neither includes backward stability analysis in the presence of rounding errors, and their convergence is slow in difficult cases.

4.1.5. Comparison with other known methods. Table 4.1.1 compares four algorithms for symmetric eigenproblems: QDWH-eig, BDD [**8**], ZZ (algorithms in [**170, 171**]) and the standard algorithm which performs tridiagonalization followed by the symmetric tridiagonal QR algorithm [**56**]. The first three methods are spectral divide-and-conquer algorithms, which can be implemented in a communication-minimizing way (for ZZ we need to replace performing QR with pivoting by subspace iteration or the randomized algorithm suggested in [**8**]). The standard method is the best known algorithm in terms of minimizing arithmetic. Table 4.1.1 shows whether the algorithm minimizes communication, the arithmetic cost in flops, existence and conditions of a backward stability proof, the theoretical maximum iteration count and the maximum matrix dimension involved for execution. For ZZ the flop count and maximum iteration are those of algorithm QUAD in [**171**]; other

algorithms in [**170, 171**] behave similarly. We obtained Max. iteration by the number of iterations needed in the worst practical case, when an eigenvalue exists within ϵ of the splitting points, which are $-\log_2 u = 53$ for BDD and ZZ. Some of the algorithms in [**170, 171**] involve $2n$-by-$2n$ matrices.

TABLE 4.1.1. Comparison of algorithms for the symmetric eigendecomposition.

	QDWH-eig	BDD[**8**]	ZZ[**170, 171**]	standard
Minimize communication?	$\surd$	$\surd$	$\surd$	$\times$
Arithmetic	$26n^3$	$\simeq 1000n^3$	$\simeq 250n^3$	$9n^3$
Backward stability proof	$\surd$	conditional	none	$\surd$
Max. iteration	6	53	53	
Matrix dimension involved	$2n$-by-n	$2n$-by-$2n$	$2n$-by-n	n-by-n

We note that the arithmetic costs of the spectral divide-and-conquer algorithms depend largely on how separated the splitting point σ is from A's spectrum. For BDD and ZZ the table shows the "worst" case, in which A has an eigenvalue within distance ϵ of the splitting points, so practically the largest possible number of iterations is needed for convergence. In practice σ is usually more well-separated from A's spectrum, and the arithmetic costs for the first three algorithms would be lower accordingly. Table 4.1.1 nonetheless illustrates a general picture of the arithmetic costs, which is reflected in our numerical examples in Section 4.3.1.1.

Compared with the communication-minimizing algorithms BDD and ZZ, QDWH-eig is much cheaper in arithmetic cost (by a factor 10 or larger), primarily because the maximum number of iterations needed is much smaller. This difference comes from the nature of the iteration parameters: QDWH uses dynamical parameters as in (3.10), which dramatically speed up convergence in the initial stage, while maintaining the asymptotic cubic convergence. BDD and ZZ, on the other hand, use iterations with static parameters. Specifically, BDD implicitly performs repeated squaring of eigenvalues, mapping the eigenvalues inside $(-1, 1)$ to 0 and those outside to ∞. The convergence is asymptotically quadratic but the initial convergence is slow when eigenvalues close to ± 1 exist. A similar argument holds for ZZ and all the algorithms in [**170, 171**].

As noted in the introduction, because QDWH-eig needs much fewer iterations it is cheaper than BDD and ZZ also in communication.

To summarize, the cost of QDWH-eig is smaller than that of BDD and ZZ by a large constant factor, in both arithmetic and communication. However, we do not claim our algorithms are better than BDD in all aspects, because BDD is applicable in more general settings, namely the generalized and non-Hermitian eigenproblem.

Compared with the standard tridiagonalization-based algorithm, QDWH-eig has the advantage that it minimizes communication. The arithmetic cost of QDWH-eig is higher, but only by a factor smaller than 3. On computing architectures where communication cost dominates arithmetic, we expect QDWH-eig to become the preferred algorithm.

4.2. Practical considerations

This section collects detailed implementation issues of QDWH-eig and QDWH-SVD.

4.2.1. Choosing splitting point σ. The splitting point σ is ideally chosen to be the median of eig(A), so that the bisection results in two matrices A_1 and A_2 of order $\simeq n/2$. To estimate the median in our experiments we suggest using the median of diag(A). This choice makes both matrices A_1 and A_2 have at least dimension one. [**8**] suggests choosing σ to be around the center of an interval in which the eigenvalues of A exist (obtained e.g. via Gerschgorin's theorem). If σ is the center of the Gerschgorin bounds we can also prove $\dim(A_1), \dim(A_2) \geq 1$. Our choice simply is based on the facts that taking diagonal elements of A is cheaper (although the Gerschgorin costs are nominally $O(n^2)$), and when A is close to diagonal form, likely to be better because it ensures $\dim(A_1) \simeq \dim(A_2)$. However for general A a better strategy may very well exist, and more study is needed. Another possible strategy of complexity $O(n^2)$ is to compute the eigenvalues of the tridiagonal part of A, then take their median (worth trying).

4.2.2. Implementing subspace iteration. Algorithm 3 gives a pseudocode for the subspace iteration applied to C.

Algorithm 3 Subspace iteration: compute invariant subspaces V_1, V_2 of C

1: Choose initial matrix X
2: Compute QR decomposition $X = [V_1\ V_2]R$
3: Stop if (4.2), otherwise $X := CX$ and go to step 2

A practical and efficient choice of initial matrix X for the subspace iteration is the set of $r+\widetilde{r}$ ($\widehat{r} \ll r$ is a small constant used as a safe-guard buffer space; in our experiments we used $\widetilde{r} = 3$) columns of $C = \frac{1}{2}(U_p - I)$ of largest column norms. This is based on the observation that we are computing the column space of C. We use the randomized algorithm [**8**] as a safeguard strategy to remedy the unfortunate case where the initial matrix was orthogonal to a desired vector. Our approach of taking large columns of C generally works well and saves one matrix multiplication. As noted in Section 4.1.1, in exact arithmetic subspace iteration converges in just one step. In this case the process completes by computing the full QR decomposition $X = [\widehat{V}_1\ \widehat{V}_2]R$, in which we note that the first multiplication by C need not be computed because $CX = X$. We then verify that the condition (4.2) holds. In our experiments we terminated subspace iteration when both $\|C\widehat{V}_1 - \widehat{V}_1\|_F$ and $\|C\widehat{V}_2\|_F$ are smaller than $10\sqrt{n}u$.

In finite precision arithmetic the eigenvalues are not exactly 0 or 1 but close to them, so subspace iteration may not converge to satisfy (4.2) in just one iteration. In all our experiments, two subspace iterations was enough to either successfully yielded an invariant subspace to working accuracy, occasionally achieving higher accuracy than with one iteration, or reaching stagnation, in which (4.2) is not satisfied but further subspace iteration does not help (note that stagnation in subspace iteration does not necessarily imply failure of QDWH-eig). In either case, after two steps of subspace iteration we obtain $\widehat{V}_1$ and $\widehat{V}_2$ by

first performing the "economy" QR decomposition $X = QR$, then computing a full QR decomposition $CQ = [\widehat{V}_1\ \widehat{V}_2]\begin{bmatrix} R \\ 0 \end{bmatrix}$.

Theorem 4.1 shows that if subspace iteration successfully computes $\widehat{V}_1, \widehat{V}_2$ satisfying the condition (4.2) then $\|E\|_F/\|A\|_F$ is negligible, so we need not verify this. When stagnation happens in subspace iteration this may not be the case, so we need to check whether $\|E\|_F/\|A\|_F \leq \epsilon$ holds or not. To do this, we follow the strategy suggested in [**32**], and after computing $[V_1\ V_2]^* A [V_1\ V_2]$ we choose r such that the Frobenius norm of the off-diagonal block of size $(n-r)$-by-r is minimized.

$\|E\|_F/\|A\|_F$ may still fail to be negligibly small in two unlikely events:

(i). The initial matrix was nearly orthogonal to a subspace of $\frac{1}{2}(U_p - I)$.

(ii). $\frac{1}{2}(U_p + I)$ had eigenvalues far from both 0 and 1.

Case (i) can be remedied by rerunning subspace iteration using a different initial matrix, an effective candidate of which is to employ the randomized algorithm, which takes the randomized matrix $\frac{1}{2}(U_p - I)W$ where W is a random Haar distributed orthogonal matrix. [**8**] shows that the QR decomposition of $\frac{1}{2}(U_p - I)W$ is rank-revealing, hence the initial matrix X contains the desired subspace and subspace iteration succeeds with high probability.

Case (ii) indicates that the $\widehat{U}_p$ computed by QDWH failed to be unitary to working accuracy. This can happen when the splitting point σ was extremely close to an eigenvalue of A, making $A - \sigma I$ nearly singular, and the smallest singular value of X_0 was severely overestimated. As discussed in [**8**], $A - \sigma I$ is nearly singular with an extremely low probability $O(\epsilon)$. If (ii) does happen nonetheless (which is signified in practice when the remedy (i) does not yield small $\|E\|_F/\|A\|_F$), then we choose a different σ and rerun steps 1-4 of QDWH-eig. We never had to resort to this remedy in our experiments.

4.2.3. When σ is equal to an eigenvalue. Even when σ is equal to an eigenvalue of A in QDWH-eig, the QDWH iteration for computing the polar decomposition of $A - \sigma I$ does not break down (unlike the scaled Newton iteration [**75**, Ch. 8], which requires the matrix to be nonsingular) but it computes the partial isometry U_p in the canonical polar decomposition of A [**75**, p. 194]. In terms of the QDWH-eig execution this causes little problem, because in this case the matrix $\frac{1}{2}(U_p - I)$ has eigenvalues $1, 0$ or 0.5. Subspace iteration has no difficulty finding an invariant subspace V_1 corresponding to eigenvalues 1 and 0.5. V_1 is then an invariant subspace corresponding to the nonnegative (including 0) eigenvalues of $A - \sigma I$.

In practice, such a situation rarely arises, because rounding errors usually cause the zero singular values to be perturbed to a small positive value, and QDWH eventually maps them to 1, in which case the singularity of $A - \sigma I$ is treated unexceptionally by QDWH.

4.2.4. Faster QDWH iterations. The QDWH iterate (3.35) is mathematically equivalent to (3.10), which can be computed via

$$\text{(4.7)} \qquad \begin{cases} Z &= I + c_k X_k^* X_k, \quad W = \mathrm{chol}(Z), \\ X_{k+1} &= \frac{b_k}{c_k} X_k + \frac{1}{\sqrt{c_k}} \left(a_k - \frac{b_k}{c_k} \right) (X_k W^{-1}) W^{-*}. \end{cases}$$

Here chol(Z) denotes the Cholesky factor of Z. The arithmetic cost of this is forming the Hermitian positive definite matrix Z (n^3 flops), computing its Cholesky factorization ($\frac{1}{3}n^3$ flops) and two triangular substitutions ($2n^3$ flops). Therefore this implementation requires $10n^3/3$ flops, which is cheaper than computing the QDWH iterate (3.35) via explicitly forming the QR decomposition, which needs $16n^3/3$ flops. Furthermore, the implementation (4.7) involves only n-by-n matrices, not $2n$-by-n. Finally, the Cholesky decomposition and triangular substitution both have a known arithmetic and communication-minimizing implementation [**9, 10**]. Therefore in general (4.7) is expected to be considerably faster than (3.35).

However, if $\kappa_2(Z)$ is large then the Cholesky factorization and triangular substitution have error bounds proportional to $\epsilon\kappa_2(Z)$ [**74**]. This affects the backward stability of the QDWH iteration (which we confirmed in our experiment). Note that the implementation (3.35) is also subject to errors, involving a QR decomposition of the matrix $\left[\begin{smallmatrix}\sqrt{c_k}X_k\\ I\end{smallmatrix}\right]$, but this matrix has a much smaller condition number $\simeq \sqrt{\kappa_2(Z)}$ when X_k is ill-conditioned, and we shall see that for (3.35) a large condition number does not affect the stability of the computed polar decomposition.

Although (3.35) is subject to large errors when $\kappa_2(Z) \gg 1$, since $\kappa_2(Z) \leq 1 + c_k\|X_k\|^2$ and $\|X_k\|_2 \leq 1$ (provided that $\alpha \geq \|A\|_2$), it follows that when $c_k \ll 1/\epsilon$, say $c_k \lesssim 100$, we can safely compute X_{k+1} via (4.7) with forward error roughly bounded by $c_k\epsilon$ (we note that the QR-based implementation (3.35) involves a matrix of condition number $\leq \sqrt{c_k}$, which is the square root of $\kappa_2(Z)$ when $c_k \gg 1$). Fortunately we know a priori that c_k converges to 3, and the convergence is fast enough so that $c_k \leq 100$ for $k \geq 2$ for any practical choice $\ell_0 > 10^{-16}$. In our experiments we switch from (3.35) to (4.7) once c_k becomes smaller than 100. In particular, if $\ell_0 > 10^5$ then we have $c_k \leq 100$ for $k \geq 1$, so we need just one iteration of (3.35).

4.2.5. Detailed flop counts. By counting the number of flops for applying Householder reflectors and noting that Q_2 is an upper-triangular matrix, we see that one QDWH iteration (3.35) for a general square A requires $(5 + \frac{1}{3})n^3$ flops. When A is symmetric X_k is also symmetric for all $k \geq 0$, so we can save $\frac{1}{2}n^3$ flops by using the fact that $Q_1Q_2^*$ is also symmetric. The same applies to the Cholesky-based algorithm above.

Now we evaluate the arithmetic cost for QDWH-eig. Recall from Section 4.1.3 that the total flop count is $\frac{8}{7}\beta$ along the critical path. We now evaluate β, where we drop the terms smaller than $O(n^3)$. For computing the polar decomposition $A - \sigma I = U_pH$, we note that for practical dense matrices of size sufficiently smaller than 10^5, we get $\ell_0 > 10^5$ (we choose ℓ_0 by estimating $\sigma_{\min}(X_0)$ using a condition number estimator) with high probability (if this is not the case we can try a different σ), so computing U_p needs $(5+\frac{1}{3}-\frac{1}{2})n^3+4\cdot(3+\frac{1}{3}-\frac{1}{2})n^3$ flops. Then subspace iteration follows, which in most cases needs just one iteration. This needs $\left(\frac{4}{3}+\frac{4}{3}\right)n^3\cdot\frac{7}{8}$ flops for forming the full decomposition $X = [\widehat{V}_1\ \widehat{V}_2]R$, and an additional $2n^3$ flops for verifying (4.2). We then form $A_1 = \widehat{V}_1^*A\widehat{V}_1$ and $A_2 = \widehat{V}_2^*A\widehat{V}_2$, which by taking advantage of the symmetry can be done in $\left(1+\frac{1}{4}\right)n^3$ flops each. We also need to perform updates $\widehat{V}_1 := \widehat{V}_1\widehat{V}_{21}$ and $\widehat{V}_2 := \widehat{V}_2\widehat{V}_{22}$, where V_{21} and V_{22} are splitting subspaces of A_1 and A_2. Each of these needs n^3 flops. Since the last two computations can be done completely

in parallel, the total flop count along the critical path is

$$\beta = \left((5+\frac{1}{3}-\frac{1}{2})+4\cdot(3+\frac{1}{3}-\frac{1}{2})+\left(\frac{4}{3}+\frac{4}{3}\right)\cdot\frac{7}{8}+2+\left(1+\frac{1}{4}\right)+1\right)n^3$$

$$= \left(22+\frac{3}{4}\right)n^3, \tag{4.8}$$

so the total arithmetic cost of QDWH-eig is $\frac{8}{7}\cdot\left(22+\frac{3}{4}\right)n^3 = 26n^3$. If we ignore the parallelizability, we have $\beta = 25n^3$ and the total flop count becomes $\frac{5}{4}\beta = (31+\frac{1}{4})n^3$.

4.3. Numerical experiments

Here we present numerical experiments to demonstrate the speed and stability of QDWH-eig. and to compare it with known algorithms in the literature. All the experiments were carried out on a desktop machine with a quad core, Intel Core i7 2.67GHz Processor and 12GB of main memory, using double precision arithmetic with rounding unit $u = 2^{-53}$. As a general remark, when running the QDWH iterations (3.35), we estimated $\alpha \simeq \|A\|_2$ by MATLAB's function `normest`(A), and estimated $\sigma_{\min}(X_0)$ using the condition number estimator `1/condest`(X_0). Wrong estimates can cause QDWH to require one or two additional iterations, but not instability or misconvergence, so rough estimates that are accurate to a factor (say) 5 are completely acceptable.

4.3.1. Symmetric eigendecomposition.

4.3.1.1. *Spectral divide-and-conquer algorithms.* This section compares spectral divide-and-conquer-type algorithms for computing the symmetric eigendecomposition. The algorithms we compare are QDWH-eig, BDD [**8**] and ZZ (the algorithm QUAD in [**171**]; the behavior of other algorithms in [**171, 170**] was all similar). Recall that these methods can be implemented in a communication-minimizing manner. We also test QDWH-eig with QR decompositions computed with row/column pivoting, which is shown as "QDWH-eig(p)".

We compare how the methods behave for problems of varying difficulties. Our experiments were carried out as follows. We set $n = 100$ and generated n-by-n symmetric matrices as $A = V\Lambda V^T$, where V is a random orthogonal matrix, generated via the QR decomposition of a random matrix generated by MATLAB function `randn(n)`. Λ is a diagonal matrix whose diagonals form a geometric series $1, r, r^2, \ldots$, with ratio $r = -\kappa^{-1/(n-1)}$, where $\kappa = \kappa_2(A)$ is the prescribed condition number, which we let be $10^2, 10^5$ and 10^{15}. A's eigenvalue closest to 0 is κ^{-1}.

To compute an invariant subspace V_1 corresponding to the positive eigenvalues of A, we apply QDWH-eig on A, and BDD and ZZ on $A - I$, because the latter two compute an invariant subspace corresponding to eigenvalues inside (and outside) an interval, which here we set $(-1, 1)$. Spectral divide-and-conquer algorithms generally face difficulty when the matrix has an eigenvalue close to the splitting points (0 for QDWH and ± 1 for BDD and ZZ), so in our setting $\kappa_2(A)$ is a precise measure of the problem difficulty.

We generated 100 matrices for each value of $\kappa = 10^2, 10^8, 10^{15}$, and report the maximum and minimum values of the iteration counts, shown as "iter" in the below table, and the backward error, assessed as in [**7**] by the residual $\|A\widehat{V}_1 - \widehat{V}_1(\widehat{V}_1^* A\widehat{V}_1)\|_F/\|A\|_F$, shown as

"res", where $\widehat{V}_1$ is the computed n-by-k eigenvector matrix. We obtained $k = 50 = \frac{n}{2}$ in all cases, and verified that all the eigenvalues of $\widehat{V}_1^* A \widehat{V}_1$ are numerically nonnegative, larger than $-u$, indicating the computed $\widehat{V}_1$ indeed approximates the positive eigenspace. The reason we compute only an invariant subspace and not the entire eigendecomposition (which is shown below) is to highlight the behavior of the algorithms when the problem becomes ill-conditioned.

Here and in the experiments below, although not shown in the table the distance from orthogonality $\|\widehat{V}_1^* \widehat{V}_1 - I_k\|_F/\sqrt{n}$ was of order ϵ for all the methods.

TABLE 4.3.1. Iteration count and residual of spectral divide-and-conquer algorithms.

$\kappa_2(A)$		10^2		10^8		10^{15}	
		min	max	min	max	min	max
iter	QDWH-eig(p)	4	5	5	5	6	6
	QDWH-eig	4	5	5	5	6	6
	ZZ	12	12	32	32	55	56
	BDD	12	13	32	32	54	55
res	QDWH-eig(p)	5.6e-16	6.1e-16	5.8e-16	6.5e-16	6.4e-16	7.3e-16
	QDWH-eig	8.5e-16	9.4e-16	8.5e-16	9.7e-16	8.4e-16	9.8e-16
	ZZ	1.6e-15	1.9e-15	2.6e-15	2.9e-15	2.9e-15	4.1e-15
	BDD	2.1e-15	2.9e-14	2.4e-13	3.3e-12	3.8e-10	4.0e-8

Observations:

- QDWH-eig converges within 6 iterations in every case, whereas ZZ and BDD need many more iterations, especially in the difficult cases where $\kappa_2(A)$ is large.
- QDWH-eig performed in a backward stable way throughout. Pivoting in computing the QR decompositions does not seem to play a role in the backward stability (recall that we prove backward stability of QDWH-eig when pivoting is employed). BDD loses backward stability when $\kappa_2(A)$ is large, which reflects the backward error bound given in [**7**]. ZZ performed in a backward stable manner, but it has no known analysis.

The experiments suggest that QDWH-eig is significantly superior to the other spectral divide-and-conquer algorithms that minimize communication. Pivoting does not appear necessary in practice. Although QR with column pivoting can be done with the same arithmetic cost as without pivoting, the communication cost generally becomes higher, so pivoting is better avoided for efficiency. In light of these observations, below we focus on QDWH-eig without using pivoting.

4.3.1.2. *Experiments using NAG MATLAB toolbox.* The NAG MATLAB toolbox [**152**] provides access from within MATLAB to NAG library routines, including LAPACK routines (contained in Chapter F). This enables different symmetric eigensolvers to be used: namely, after reducing the matrix to tridiagonal form $T = Q^* A Q$, we can compute the eigendecomposition of T via (i) the QR algorithm, (ii) divide-and conquer [**63**], or (iii) the MR^3 algorithm

[**38**]. The MATLAB function `eig` automatically runs the QR algorithm, but as we shall see, the performance of the three approaches can differ significantly.

The tables below show the result of MATLAB experiments with QDWH-eig, compared with the above standard solvers, which minimize arithmetic but not communication. "QR" stands for the algorithm that employs tridiagonalization followed by tridiagonal QR, "D-C" uses the divide-and-conquer algorithm [**63**] (not to be confused with the spectral divide-and-conquer algorithms that include QDWH-eig, BDD and ZZ) and "MR^3" uses the solver based on multiple relatively robust representations [**38**] after tridiagonalization. We used a machine with an Intel Core i7 2.67GHz Processor (4 cores, 8 threads), and 12GB RAM.

Here we generated Hermitian matrices A by generating a random matrix B by the MATLAB function $B =$`randn(n)`, then letting $A = \frac{1}{2}(B + B^*)$. We tested for 10 different cases of A, and Table 4.3.2 shows the average runtime. For reference, the time needed for the tridiagonalization phase is shown as "tridiag". $\times$ means the memory ran out before the method finished the execution.

TABLE 4.3.2. Runtime(s) for eigendecomposition algorithms.

n	1000	2000	3000	4000	5000
QDWH-eig	3.4	18.8	57.0	125	236
ZZ	13.4	91.1	314	685	1280
BDD	32.6	221	725	1570	$\times$
QR	2.9	27.9	100.2	238	459
D-C	0.55	3.5	11.5	26.7	50.9
MR^3	0.90	4.8	14.5	33.4	61.5
tridiag	0.45	1.79	5.9	13.8	25.6

We see that D-C is the fastest algorithm and in fact, the dominant part of its runtime is consumed in the tridiagonalization step. The runtime of MR^3 is comparable to D-C but is generally slightly longer. While QDWH-eig is notably slower than D-C and MR^3, it is faster than QR for $n \geq 2000$.

The results may suggest that D-C and MR^3 are faster than QDWH-eig. However we recall that our machine has only 4 cores and we used MATLAB's built-in function `qr` for computing the QR decompositions, which does not minimize communication. Recent progress in implementing communication-optimal QR decomposition [**33**] suggests that communication-avoiding QR often run significantly faster than standard QR. When combined with such implementations, we expect that QDWH will be the preferred algorithm on a more parallel computing architecture in which communication dominates arithmetic.

Table 4.3.3 shows the largest backward error $\|\widehat{V}\widehat{\Lambda}\widehat{V}^T - A\|_F/\|A\|_F$ of 10 runs. The experimental backward stability is acceptable for all the methods but BDD. It is perhaps worth noting that the backward errors of QDWH-eig were generally noticeably smaller than the other methods (by more than a factor 3), while those of MR^3 were about a magnitude larger.

Table 4.3.4 shows the distance from orthogonality of the computed $\widehat{V}$.

TABLE 4.3.3. Backward error $\|A - \widehat{V}\widehat{\Lambda}\widehat{V}^T\|_F/\|A\|_F$.

n	1000	2000	3000	4000	5000
QDWH-eig	1.7e-15	1.8e-15	1.8e-15	1.9e-15	2.0e-15
ZZ	5.9e-15	8.2e-15	9.7e-15	1.1e-14	1.4e-14
BDD	3.4e-13	4.8e-12	2.3e-12	8.4e-11	$\times$
QR	9.1e-15	1.3e-14	1.5e-14	1.8e-14	2.0e-14
D-C	3.7e-15	4.8e-15	5.5e-15	6.2e-15	6.8e-15
MR^3	9.8e-14	4.9e-14	9.4e-14	1.2e-13	1.4e-13

TABLE 4.3.4. Orthogonality of $\widehat{V}$: $\|\widehat{V}^T\widehat{V} - I\|_F/\sqrt{n}$.

n	1000	2000	3000	4000	5000
QDWH-eig	6.9e-16	7.7e-16	8.7e-16	9.6e-16	1.0e-15
ZZ	2.4e-15	3.4e-15	4.0e-15	4.4e-15	5.3e-15
BDD	2.7e-15	3.4e-15	3.9e-15	4.2e-15	4.9e-15
QR	7.5e-15	1.0e-14	1.3e-14	1.4e-14	1.6e-14
D-C	3.1e-15	3.9e-15	4.6e-15	5.1e-15	5.7e-15
MR^3	1.3e-13	1.4e-13	1.6e-13	2.0e-13	2.4e-13

The spectral divide-and-conquer algorithms generally produce $\widehat{V}$ that are orthogonal to working accuracy.

4.3.2. Summary of numerical experiments. The results of our experiments can be summarized as follows.

- QDWH-eig has excellent numerical backward stability, even without pivoting for computing the QR decompositions.
- QDWH-eig is generally much faster than known algorithms that minimize communication, and significantly faster than MATLAB's default algorithms for large matrices.
- On our shared-memory machine, divide-and-conquer following reduction (to tridiagonal or bidiagonal form) is faster than QDWH-eig. Most of the runtime of divide-and-conquer is consumed in the reduction stage. This reduction stage becomes a communication-bottleneck in parallel computing, as there is no known way to do it while minimizing communication, in particular the latency cost. Hence we can expect that on massively parallel computing architectures the communication cost dominates arithmetic cost, making our QDWH-based algorithms preferable. This will be investigated in future work.

CHAPTER 5

Efficient, communication minimizing algorithm for the SVD

In this chapter we propose an SVD algorithm QDWH-SVD that minimizes communication. As the name suggests, QDWH-SVD is also based on the QDWH algorithm we developed in Chapter 3, just like QDWH-eig. QDWH-SVD requires the computation of no more than 12 QR decompositions, and the number is smaller for well-conditioned matrices. The arithmetic cost of QDWH-SVD is smaller than twice that of a standard SVD algorithm. Its backward stability follows immediately from that of QDWH and QDWH-eig. QDWH-SVD performs comparably to conventional algorithms on our preliminary numerical experiments, and its performance is expected to improve significantly on highly parallel computing architectures where communication dominates arithmetic.

Introduction. We now turn to algorithms for the singular value decomposition. The algorithm development is simple because we are well-prepared with the necessary tools presented in the previous two chapters. Our SVD algorithm QDWH-SVD is simply a combination of the QDWH algorithm, which computes the polar decomposition $A = U_pH$, and QDWH-eig, which computes the eigendecomposition $H = V\Sigma V^*$. Then the SVD of A is $A = (U_pV)\Sigma V^* = U\Sigma V^*$. Backward stability of QDWH-SVD follows immediately from that of QDWH and QDWH-eig. The essential cost for computing the SVD of $A \in \mathbb{C}^{m\times n}$ is in performing QR decompositions of no more than six $(m + n)$-by-n matrices and six $2n$-by-n matrices. We argue that for square matrices, computing one SVD via QDWH-SVD is cheaper than computing two symmetric eigendecompositions via QDWH-eig. Numerical experiments show the competitiveness of QDWH-SVD with conventional SVD algorithms, even without a QR decomposition algorithm that minimizes communication.

5.1. Algorithm QDWH-SVD

In this section we describe our SVD algorithm QDWH-SVD.

5.1.1. Algorithm. Higham and Papadimitriou [**76, 77**] propose a framework of computing the SVD via the polar decomposition and the eigendecomposition: given the polar decomposition $A = U_pH$ and the symmetric eigendecomposition $H = V\Sigma V^*$, the SVD is obtained by $A = (U_pV)\Sigma V^*$. They suggest using a Padé-type method for the polar decomposition and any standard method for the symmetric eigendecomposition.

Our SVD algorithm QDWH-SVD follows this path but replaces both steps with QDWH-based algorithms: it computes the polar decomposition by the QDWH algorithm (3.35), and

the symmetric eigendecomposition via QDWH-eig. Algorithm 4 is a pseudocode of our SVD algorithm QDWH-SVD.

Algorithm 4 QDWH-SVD: compute the SVD of a general matrix A

1: Compute the polar decomposition $A = U_pH$ via QDWH.
2: Compute the symmetric eigendecomposition $H = V\Sigma V^*$ via QDWH-eig.
3: Form $U = U_pV$. $A = U\Sigma V^*$ is the SVD.

5.1.1.1. *Rectangular case.* QDWH-SVD works for rectangular matrices A because, as mentioned in [**120**], the QDWH algorithm can compute the polar decomposition of rectangular matrices with full column rank. Hence the case $m > n$ poses no problem as long as the matrix has full column rank. However, for practical efficiency, when $m \gg n$ it is preferred to perform an initial QR decomposition of $A = QR$ to reduce the computation to the SVD of a square n-by-n matrix $R = U_R\Sigma V^*$. Then the SVD is obtained as $A = (QU_R)\Sigma V^*$. This approach maintains backward stability of the overall SVD because the QR decomposition $A = QR$ computed via Householder QR is backward stable. Numerical experiments in Section 5.3.0.2 confirm that this way of computing the SVD is faster when $m \gg n$. We do not need to consider the case $m < n$ because if $A = U\Sigma V^*$ then the SVD of A^* is just $V\Sigma U^*$.

5.1.1.2. *Singular and rank-deficient case.* Some care is needed when A is rank-deficient, or equivalently when it has zero singular values, because QDWH computes[1] $A = U_pH$ where U_p is not unitary but a partial isometry [**75**, p. 194]. However, H is still the unique Hermitian polar factor of A, and its has the same number of zero singular values as A. Suppose the computed eigendecomposition is $H = [V_1\ V_2]\mathrm{diag}(\Sigma_1, 0_{r\times r})[V_1\ V_2]^*$, or in practice suppose H has r computed eigenvalues of order ϵ. Then we obtain the "economy" SVD $A = (U_pV_1)\Sigma_1V_1^*$. This also provides A's null space V_2.

5.1.2. Backward stability. Higham and Papadimitriou [**76**] show that the computed SVD is backward stable provided that both the polar decomposition and the eigendecomposition are computed in a backward stable manner. Therefore, the backward stability of QDWH-SVD immediately follows from that of QDWH and QDWH-eig.

THEOREM 5.1. *Suppose in QDWH-SVD that all the polar decompositions computed by QDWH (in steps 1 and 2) are backward stable. Then QDWH-SVD computes the SVD in a backward stable manner, that is,* $\|A - \widehat{U}\widehat{\Sigma}\widehat{V}^*\|_F/\|A\|_F = \epsilon$.

The proof is done simply by combining the result in [**76**] and Theorem 4.1. As was the case for QDWH-eig, combined with the analysis in Section 3.4 we conclude that QDWH-SVD is backward stable if row and column pivoting is used when computing the QR decompositions.

5.1.3. Communication/arithmetic cost. QDWH-SVD minimizes communication cost in the asymptotic sense, because just like QDWH-eig, it uses only QR decompositions and matrix multiplications.

[1]In this case we need to choose $0 < \ell_0 \le \sigma_{n-r}(X_0)$ to ensure QDWH converges within 6 iterations, where $\sigma_{n-r}(X_0)$ is the smallest positive singular value.

The arithmetic cost is essentially the sum of those of QDWH and QDWH-eig, and as we will see in Section 5.2.2, in the square case ranges from $34n^3$ to $52n^3$ flops depending on $\kappa_2(A)$. We see that the cost of computing the SVD of a general square matrix is less than twice that of computing an eigendecomposition of a symmetric matrix of the same size by QDWH-eig.

5.1.4. Comparison with other known methods. As described in [**8**], BDD for the symmetric eigenproblem can be applied to compute the SVD by computing the eigendecomposition of $\left[\begin{smallmatrix} 0 & A^* \\ A & 0 \end{smallmatrix}\right]$ (we simply call this algorithm BDD). This algorithm still minimizes communication in the big-Ω sense.

Table 5.1.1 compares three SVD algorithms: QDWH-SVD, BDD and the standard algorithm which performs bidiagonalization followed by bidiagonal QR. Compared with the

TABLE 5.1.1. Comparison of algorithms for the SVD.

	QDWH-SVD	BDD for SVD [**8**]	Standard
Minimize communication?	$\surd$	$\surd$	$\times$
Arithmetic	$34n^3 \sim 52n^3$	$\simeq 8000n^3$	$26n^3$
Backward stability proof	$\surd$	conditional	$\surd$
Max. iteration	6	53	
Matrix dimension involved	$(m+n)$-by-n	$2(m+n)$-by-$2(m+n)$	n-by-n

communication-minimizing algorithm BDD, our method is clearly much cheaper in arithmetic cost, by about a factor 200. This factor can be roughly understood as follows: a factor > 8 from the required iteration number, another factor $\simeq 8$ because [**8**] forms a $2n$-by-$2n$ matrix (which soon becomes dense as iterations proceed), and another factor $\simeq 2$ from the fact that QDWH-eig computes the column space of the augmented matrix whereas [**8**] needs its orthogonal complement, which is more expensive. The small iteration count makes the communication cost of QDWH-SVD smaller than that of BDD by a large constant factor, for the same reason QDWH-eig is cheaper than BDD in communication for symmetric eigenproblems.

Compared with the standard bidiagonalization-based SVD algorithm, QDWH-SVD has the obvious advantage that it minimizes communication. The arithmetic cost of QDWH-SVD is higher, but only by a factor no larger than 2. Just like QDWH-eig for symmetric eigenproblems, QDWH-SVD is preferred when communication cost dominates arithmetic cost.

5.2. Practical considerations

5.2.1. When H has negative computed eigenvalues. The eigenvalues of H computed by QDWH-SVD are not guaranteed to be nonnegative, although in exact arithmetic they are the singular values of A. However, since QDWH-SVD is backward stable and singular values are always well-conditioned, the negative eigenvalues of H can appear only when A's singular values of order $\epsilon\|A\|_2$ are perturbed by rounding errors. Hence these perturbed values are necessarily negligibly small, so there is no harm in regarding them as 0, or taking their absolute values as suggested in [**76**].

5.2.2. Detailed flop counts. Compared with QDWH-eig, QDWH-SVD has the additional arithmetic cost is in computing U_p and H. The cost for computing U_p largely depends on the condition number of A, which determines how many iterations of (4.7) and (3.35) are needed. We summarize them in the below table, which is obtained by monitoring c_k and the convergence $\ell_k \to 1$ with $\ell_0 = 1/\kappa_2(A)$ until $|\ell_k - 1| \leq 2 \times 10^{-16}$. Since computing H by

TABLE 5.2.1. Arithmetic cost and iteration breakdown of QDWH.

$\kappa_2(A)$	1.1	1.5	10	10^3	10^5	10^{10}	10^{16}
flops/n^3	$6+\frac{2}{3}$	10	$13+\frac{1}{3}$	$15+\frac{1}{3}$	$18+\frac{2}{3}$	$20+\frac{2}{3}$	24
# of (3.35)	0	0	0	1	1	2	2
# of (4.7)	2	3	4	3	4	3	4

(3.1) requires $2n^3$ flops, together with $\frac{8}{7}\beta = 26n^3$ we conclude that the total flop count for QDWH-SVD ranges from $(34+\frac{2}{3})n^3$ to $52n^3$.

5.3. Numerical experiments

We now compare different algorithms for computing the SVD. We generate rectangular matrices by forming $A = U\Sigma V^*$, where U, V are random orthogonal matrices and Σ is a diagonal matrix of singular values, which form an arithmetic sequence.

5.3.0.1. *Square nonsingular case.* Using the NAG MATLAB toolbox we can compute the SVD by first reducing the matrix to bidiagonal form, then computing a bidiagonal SVD via either (i) divide-and conquer [**62**], which we call DCSVD, or (ii) the QR algorithm, which we call QRSVD. We note that MATLAB's function `svd` calls QRSVD, which is sometimes considered more stable, because the computed singular values of the bidiagonal matrix are shown to have high relative accuracy [**35**].

Below we compare the speed of the four algorithms QDWH-SVD, BDD, QRSVD and DCSVD[2].

Different matrix sizes. We set $\kappa_2(A) = 1.5$ and varied the matrix size n. Tables 5.3.1 and 5.3.2 show the average runtime and largest backward error $\|A - \widehat{U}\widehat{\Sigma}\widehat{V}^T\|_F/\|A\|_F$ of 10 runs. The time needed for the bidiagonalization phase is shown as "bidiag". divide-and-conquer

TABLE 5.3.1. Runtime(s) for SVD algorithms, varying matrix size.

n	1000	2000	3000	4000	5000
QDWH-SVD	5.1	26.8	80.1	175	334
BDD	197	1416	×	×	×
QRSVD	16.7	149	540	1313	2727
DCSVD	1.78	12.2	36.4	78.4	138
bidiag	0.67	6.1	19.3	46.2	84.1

is the fastest, most of whose runtime is in the bidiagonalization step for large matrices.

[2]MR^3 for the bidiagonal SVD was not available with the NAG toolbox as of writing, but its relative performance should resemble that for the symmetric eigenproblem.

QDWH-SVD is notably faster than QRSVD (which MATLAB uses by default) but still slower than DCSVD. The same comment as in the previous subsection regarding parallel computing applies here. The backward errors of QDWH-SVD, QRSVD and DCSVD were all acceptably small, and that of QDWH-SVD was smaller than the rest by a factor about 3.

The distance from orthogonality of the computed $\widehat{U}, \widehat{V}$ were also acceptably small, again QDWH-SVD yielding the smallest.

TABLE 5.3.2. Backward error $\|A - \widehat{U}\widehat{\Sigma}\widehat{V}^T\|_F/\|A\|_F$.

n	1000	2000	3000	4000	5000
QDWH-SVD	1.9e-15	2.1e-15	2.2e-15	2.4e-15	2.4e-15
BDD	1.9e-12	3.7e-11	×	×	×
QRSVD	1.1e-14	1.5e-14	1.9e-14	2.2e-14	2.5e-14
DCSVD	4.4e-15	5.7e-15	6.8e-15	7.6e-15	8.6e-15

TABLE 5.3.3. Orthogonality of computed $\widehat{U}, \widehat{V}$: $\max\{\|\widehat{U}^T\widehat{U} - I\|_F/\sqrt{n}, \|\widehat{V}^T\widehat{V} - I\|_F/\sqrt{n}\}$.

n	1000	2000	3000	4000	5000
QDWH-SVD	8.9e-16	9.4e-16	9.8e-15	1.0e-15	1.0e-15
BDD	8.7e-14	1.2e-13	×	×	×
QRSVD	8.7e-15	1.1e-14	1.4e-14	1.6e-14	1.9e-14
DCSVD	3.6e-15	4.7e-15	5.4e-15	6.0e-15	6.7e-15

Different condition numbers. Here we fix the matrix size $n = 5000$ and varied the condition number $\kappa_2(A)$. Tables 5.3.4, 5.3.5 and 5.3.6 show the results.

TABLE 5.3.4. Runtime(s) for SVD algorithms, varying condition number.

$\kappa_2(A)$	1.1	1.5	10	10^5	10^{10}	10^{15}
QDWH-SVD	335	334	357	389	418	419
QRSVD	3012	2727	1861	1639	1261	885
DCSVD	138	138	126	125	113	108

TABLE 5.3.5. Backward error $\|A - \widehat{U}\widehat{\Sigma}\widehat{V}^T\|_F/\|A\|_F$.

$\kappa_2(A)$	1.1	1.5	10	10^5	10^{10}	10^{15}
QDWH-SVD	2.4e-15	2.4e-15	2.3e-15	2.4e-15	2.3e-15	2.3e-15
QRSVD	2.4e-14	2.5e-14	2.3e-14	2.2e-14	2.1e-14	2.1e-14
DCSVD	8.4e-15	8.6e-15	8.0e-15	7.7e-15	7.8e-15	7.9e-15

In all cases, the runtime of QDWH-SVD was shorter than twice that of QDWH-eig (472 seconds) for matrices of the same size. In this sense we verify that computing the SVD is no

TABLE 5.3.6. Orthogonality of computed $\widehat{U}, \widehat{V}$: $\max\{\|\widehat{U}^T\widehat{U} - I\|_F/\sqrt{n}, \|\widehat{V}^T\widehat{V} - I\|_F/\sqrt{n}\}$.

$\kappa_2(A)$	1.1	1.5	10	10^5	10^{10}	10^{15}
QDWH-SVD	9.2e-16	1.0e-15	9.3e-16	1.1e-15	1.0e-15	1.0e-15
QRSVD	1.8e-14	1.9e-14	1.6e-14	1.4e-14	1.4e-14	1.5e-14
DCSVD	6.6e-15	6.7e-15	7.0e-15	6.8e-15	6.9e-15	7.0e-15

more expensive than computing the symmetric eigendecomposition twice. QDWH-SVD is somewhat faster for smaller $\kappa_2(A)$, because computing the polar decomposition by QDWH in step 1 is easier for smaller $\kappa_2(A)$.

By contrast, QRSVD is significantly faster for ill-conditioned matrices. A possible explanation is that computing orthogonal eigenvectors via inverse iteration becomes more difficult when eigenvalues are clustered. DCSVD performs consistently and is always the fastest.

For the backward error and distance from orthogonality we observed the same behavior as above, QDWH-SVD giving the smallest errors.

5.3.0.2. *Rectangular matrices.* As mentioned in Section 5.1.1.1, there are two ways to compute the SVD of rectangular matrices via QDWH-SVD: (i) simply run QDWH-SVD, and (ii) first compute a QR decomposition and then apply QDWH-SVD to R (we call the approach QR-QDWH-SVD). Here we compare the two approaches.

We fix $n = 500$ and $\kappa_2(A) = 10^5$, and define m-by-n matrices A where m varies from n to $200n$. Table 5.3.7 shows the runtime. Comparing QDWH-SVD and DCSVD we observe

TABLE 5.3.7. Runtime(s) for SVD algorithms, rectangular matrices.

ratio m/n	1	2	5	10	50	100	200
QDWH-SVD	1.3	1.4	1.9	2.7	9.1	17.5	33.8
QR-QDWH-SVD	1.3	1.4	1.5	1.8	3.7	6.4	11.7
QRSVD	1.4	1.74	3.4	6.2	30.0	56.2	111
DCSVD	0.33	0.34	0.78	1.6	9.6	18.2	34.9
QR-DCSVD	0.33	0.41	0.61	0.78	2.8	5.4	10.8

that divide-and-conquer is again faster unless $m \gg n$. However, the runtime of QDWH-SVD scales better as m grows and for $m \geq 50n$ the performance of the two was similar.

We also verify that the second approach of first computing $A = QR$ is more efficient, especially when $m \gg n$. Note that this is also true for the bidiagonalization methods, as discussed in in [**23**]. We illustrated this above by QR-DCSVD, which first computes $A = QR$ and then compute the SVD of R via bidiagonalization-divide-and-conquer. The backward error in all the cases were of order ϵ.

5.3.1. Summary of numerical experiments. The results of our experiments can be summarized as follows.

- Just like QDWH-eig, QDWH-SVD has excellent numerical backward stability, even without pivoting.

- For the same matrix size, QDWH-SVD generally takes less than twice as much time as QDWH-eig.
- Just like for the symmetric eigenproblem, divide-and-conquer following reduction (to tridiagonal or bidiagonal form) is faster than our algorithms, both for the eigendecomposition and the SVD. We expect that on massively parallel computing architectures the relative performance of QDWH-SVD will increase.

CHAPTER 6

dqds with aggressive early deflation for computing singular values of bidiagonal matrices

This chapter deals with standard algorithms for the SVD, unlike the previous two chapters which developed completely new algorithms for the symmetric eigendecomposition and the SVD. In particular, we focus on the dqds algorithm (reviewed in Section 2.5), which computes all the singular values of a bidiagonal matrix to high relative accuracy. In this chapter we incorporate into dqds the technique of aggressive early deflation, which has been applied successfully to the Hessenberg QR algorithm. Extensive numerical experiments show that aggressive early deflation often reduces the dqds runtime significantly. In addition, our theoretical analysis suggests that with aggressive early deflation, the performance of dqds is largely independent of the shift strategy. We confirm through experiments that the zero-shift version is often as fast as the shifted version. All of our proposed algorithms compute all the singular values to high relative accuracy.

Introduction. The differential quotient difference with shifts (dqds), as summarized in Section 2.5, computes all the singular values of a bidiagonal matrix to high relative accuracy [**45**]. Its efficient implementation has been developed and is now available as an LAPACK subroutine DLASQ [**132**]. Because of its guaranteed relative accuracy and efficiency, dqds has now replaced the QR algorithm [**35**], which had been the default algorithm to compute the singular values of a bidiagonal matrix. The standard way of computing the singular values of a general matrix is to first apply suitable orthogonal transformations to reduce the matrix to bidiagonal form, then use dqds [**31**]. dqds is also a major computational kernel in the MRRR algorithm for computing the eigenvalues of a symmetric tridiagonal matrix [**37, 38, 39**].

The aggressive early deflation strategy, introduced in [**16**] and summarized in Section 2.6, is known to greatly improve the performance of the Hessenberg QR algorithm for computing the eigenvalues of a general square matrix by deflating converged eigenvalues much earlier than a conventional deflation strategy does. Our primary contribution here is the proposal of two deflation strategies for dqds based on aggressive early deflation. The first strategy is a direct specialization of aggressive early deflation to dqds. The second strategy, which takes full advantage of the bidiagonal structure of the matrix, is computationally more efficient. Both of our proposed strategies guarantee high relative accuracy of all the computed singular values. The results of extensive numerical experiments demonstrate that performing aggressive early deflation significantly reduces the solution time of dqds in many cases. We

observed speedups of up to 50, and in all our experiments the second strategy was at least as fast as DLASQ for any matrix larger than 3000.

With a conventional deflation strategy, the zero-shift version dqd is too slow to be practical. Hence, DLASQ adopts a sophisticated shift strategy to improve the convergence of the bottom off-diagonal element [**132**]. We demonstrate both theoretically and experimentally that when aggressive early deflation is incorporated, dqd is often as fast as dqds. Besides making it unnecessary to compute a shift, this observation provides a potential to parallelize our algorithm by running multiple dqd iterations in a pipelined fashion, which has been successfully done in the QR algorithm [**5, 110, 157**].

The structure of the chapter is as follows. Sections 6.1 and 6.2 describe our two aggressive early deflation strategies for dqds. Then in Section 6.3 we present convergence analysis of dqds with aggressive early deflation, which explains its speedup and motivates the use of dqd. Numerical results are presented in Section 6.4 to illustrate the efficiency and accuracy of our algorithms.

In this chapter, we use ϵ denote the machine precision, and ϵ_x with any subscript represents a number such that $|\epsilon_x| \leq \epsilon$.

Recall that the dqds algorithm computes the singular values of the bidiagonal matrix B of the following form:

$$B = \mathrm{bidiag}\left(\begin{array}{ccccccc} \sqrt{q_1} & \sqrt{e_1} & \cdot & \cdot & \sqrt{e_{n-2}} & \sqrt{e_{n-1}} & \\ & \sqrt{q_1} & \cdot & \cdot & \cdot & \sqrt{q_{n-1}} & \sqrt{q_n} \end{array} \right), \tag{6.1}$$

6.1. Aggressive early deflation for dqds - version 1: Aggdef(1)

In this section, we describe our first aggressive early deflation strategy for dqds, which is referred to as Aggdef(1), and is more or less a direct application of aggressive early deflation to the bidiagonal case.

6.1.1. Algorithm. Let $B_2 = U\Sigma V^T$ be the SVD of the lower-right $k \times k$ submatrix B_2 of a bidiagonal matrix B as in (6.1), where the singular values appear in decreasing order of magnitude. Then, we compute the orthogonal transformation

$$\begin{bmatrix} I & & \\ & 1 & \\ & & U^T \end{bmatrix} B \begin{bmatrix} I & & \\ & 1 & \\ & & V \end{bmatrix} = \begin{bmatrix} & & \\ & B_1 & \begin{matrix} t^T \\ \Sigma \end{matrix} \end{bmatrix}, \tag{6.2}$$

where the right-hand side matrix has the nonzero pattern

$$\begin{bmatrix} \ddots & \ddots & & & & & \\ & * & * & & & & \\ & & * & * & * & * & \\ & & & * & & & \\ & & & & * & & \\ & & & & & * & \end{bmatrix}. \tag{6.3}$$

We now look for elements of the spike vector $t^T = \sqrt{e_{n-k}}V(1,:)$ that are small enough to be neglected, and deflate the corresponding diagonals of Σ to obtain the reduced $(n-k_\ell)\times(n-k_\ell)$

matrix, where k_ℓ is the number of negligible elements in t. This matrix needs to be re-bidiagonalized in order to return to the dqds iterations. Algorithm 5 shows the pseudocode of this aggressive deflation strategy, which we call Aggdef(1).

Algorithm 5 Aggressive early deflation - version 1: Aggdef(1)

Inputs: Bidiagonal B, window size k, sum of previous shifts S

1: compute the singular values of B_2, the lower-right $k \times k$ submatrix of B
2: compute the spike vector t in (6.2)
3: find negligible elements in t and deflate converged singular values
4: bidiagonalize matrix of form (6.3)

Below we discuss the details of each step of Aggdef(1).

Computing the singular values of B_2. On line 1 of Aggdef(1), we use standard dqds (without aggressive early deflation) to compute the singular values of B_2. This generally requires $O(k^2)$ flops.

Computing the spike vector. To compute the spike vector t on line 2, the first elements of the right singular vectors V of B_2 need to be computed. This can be done by computing the full SVD of B_2, which requires at least $O(k^2)$ flops. We can reduce the cost by noting that only the first element of each singular vector is needed to compute t. This corresponds to the Gauss quadrature, whose computational algorithms are discussed in [**57, 55**]. However this approach generally still requires $O(k^2)$ flops.

When to neglect elements of the spike vector. Basic singular value perturbation theory [**142**, p.69] tells us that the perturbation on the computed singular values caused by neglecting the ℓth element t_ℓ of t is bounded by $|t_\ell|$. Since the unconverged singular values are greater than $\sqrt{S}$ where S is the sum of previous shifts, we can safely neglect elements of t that are smaller than $\sqrt{S}\epsilon$ (ϵ is the machine precision) without causing loss of relative accuracy of any singular value.

Rebidiagonalization process. Since the upper-left part of the matrix is already bidiagonal, we only need to bidiagonalize the bottom-right $(k_0+1) \times (k_0+1)$ part of the matrix of the form (6.3), where $k_0 = k - k_\ell$.

We use a 4×4 ($k_0 = 3$) example to illustrate our bidiagonalization process, which is based on a sequence of Givens rotations:

$$\begin{bmatrix} * & * & * & * \\ & * & & \\ & & * & \\ & & & * \end{bmatrix} \overset{G_R(3,4)}{\rightarrow} \begin{bmatrix} * & * & * & 0 \\ & * & & \\ & & * & + \\ & & + & * \end{bmatrix} \overset{G_L(3,4)}{\rightarrow} \begin{bmatrix} * & * & * & \\ & * & & \\ & & * & * \\ & & 0 & * \end{bmatrix} \overset{G_R(2,3)}{\rightarrow} \begin{bmatrix} * & * & 0 & \\ & * & + & \\ & + & * & * \\ & & & * \end{bmatrix}$$

$$\overset{G_L(2,3)}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & + \\ & 0 & * & * \\ & & & * \end{bmatrix} \overset{G_R(3,4)}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & 0 \\ & & * & * \\ & & + & * \end{bmatrix} \overset{G_L(3,4)}{\rightarrow} \begin{bmatrix} * & * & & \\ & * & * & \\ & & * & * \\ & & 0 & * \end{bmatrix}.$$

Here, $G_L(i,j)$ (or $G_R(i,j)$) above the arrow indicates the application of a Givens rotation from the left (or right) to the ith and jth rows (or columns). "0" indicates an element that was zeroed out by the rotation, and "+" is a nonzero that was newly created. By counting

the number of rotations, we can show that the total flops required for this process is at most $18k_0^2$, which is generally $O(k^2)$. We note that this process can be regarded as a bidiagonal version of the tridiagonalization algorithm of an arrowhead matrix discussed in [**123, 169**].

Maintaining high relative accuracy. The computation of the spike vector t and the bidiagonalization process described above can cause errors of order $\epsilon\|B_2\|_2$ in finite precision arithmetic. This may result in loss of relative accuracy for small singular values. To avoid this, we dynamically adjust the deflation window size (shrink from input size k) so that B_2 does not contain elements that are larger than $c\sqrt{S}$, where S is the sum of previous shifts and c is a modest constant. In our experiments we let $c = 1.0$.

6.2. Aggressive early deflation for dqds - version 2: Aggdef(2)

Numerical experiments in Section 6.4 illustrate that Aggdef(1) described above significantly reduces the number of dqds iterations in many cases. However, computing the spike vector t and rebidiagonalizing the matrix generally require at least $O(k^2)$ flops, which can be expensive. Furthermore, Aggdef(1) requires the computation of the square roots of q_i and e_i, and it needs to consider a safe window size to guarantee the high relative accuracy. In this section, we discuss an alternative deflation strategy, Aggdef(2), which addresses these issues by seeking one deflatable singular value at a time.

6.2.1. Process to deflate one singular value. To introduce Aggdef(2) we first describe Aggdef(2)-1, a simplified process to deflate one smallest singular value. As before, B_2 is the lower-right $k \times k$ submatrix of B.

Algorithm 6 Aggdef(2)-1, process for deflating one singular value

Inputs: Bidiagonal B, window size k, sum of previous shifts S

1: compute $s = (\sigma_{\min}(B_2))^2$
2: compute $\widehat{B}_2$ such that $\widehat{B}_2^T\widehat{B}_2 = B_2^TB_2 - sI$ by dstqds. Set $\widehat{B}_2(\text{end},\text{end}) \leftarrow 0$ if it is negligible (see (6.10)), otherwise exit
3: compute $\breve{B}_2 = \widehat{B}_2\prod_{i=1}^{i_0} G_R(k-i,k)$ for $i_0 = 1,\ldots,k-2$ until w as in (6.4) becomes negligible (see (6.12)), then $w \leftarrow 0$. Exit if w never becomes negligible
4: compute $\widetilde{B}_2$ such that $\widetilde{B}_2^T\widetilde{B}_2 = \breve{B}_2^T\breve{B}_2 + sI$ by dstqds and update B by replacing B_2 with $\widetilde{B}_2$

On the first line of Aggdef(2)-1, only the smallest singular value of B_2 is computed using dqds, which requires $O(k)$ flops.

On lines 2 and 4, we use the dstqds algorithm [**38, 39**], which was originally developed to obtain the LDL^T decomposition of a shifted tridiagonal matrix in a mixed forward-backward stable manner in the relative sense. We slightly modify this algorithm to reflect the bidiagonal structure. This allows us to compute the k-by-k bidiagonal $\widehat{B}_2$ with $\widehat{q}_{n-k+i} = (\widehat{B}_2(i,i))^2$ and $\widehat{e}_{n-k+i} = (\widehat{B}_2(i,i+1))^2$ from B_2 such that $\widehat{B}_2^T\widehat{B}_2 = B_2^TB_2 - sI$ in about $5k$ flops, without losing relative accuracy of the computed singular values. Algorithm 7 shows the pseudocode of our dstqds algorithm.

Algorithm 7 differential stationary qds (dstqds)

Inputs: $s, q_i = (B(i,i))^2\ (i = n-k+1, \ldots, n), e_i = (B(i,i+1))^2\ (i = n-k+1, \ldots, n-1)$

1: $d = -s$
2: $\widehat{q}_{n-k+1} = q_{n-k+1} + d$
3: **for** $i := n-k+1, \cdots, n-1$ **do**
4: $\quad \widehat{e}_i = q_i e_i / \widehat{q}_i$
5: $\quad d = d e_i / \widehat{q}_i - s$
6: $\quad \widehat{q}_{i+1} = q_{i+1} + d$
7: **end for**

The bottom diagonal element of $\widehat{B}_2$ is 0 in exact arithmetic. However in practice its computed value is nonzero, and we safely set it to 0 when it is small enough, as detailed below in (6.10). In exact arithmetic, the bidiagonal elements of $\widehat{B}_2$ are all positive except for the bottom zero element. However in finite precision arithmetic, negative elements could appear. When negative elements exist besides at the bottom diagonal, this indicates a breakdown of the Cholesky factorization. When this happens, we exit Aggdef(2)-1 and return to the dqds iterations.

On line 3 of Aggdef(2)-1, to determine if $\sqrt{s+S}$ can be deflated as a converged singular value, we apply a sequence of i_0 ($\leq k-2$) Givens transformations (note that they are not strictly Givens rotations: we apply matrices of the form $\begin{bmatrix} c & s \\ s & -c \end{bmatrix}$ to specified columns, where $c^2 + s^2 = 1$) to $\widehat{B}_2$ on the right to compute $\breve{B}_2 = \widehat{B}_2 \prod_{i=1}^{i_0} G_R(k-i, k)$, where $G_R(k-i,k)$ is the Givens transformation acting on the $(k-i)$th and kth columns. (6.4) describes this process for the case $k = 5$ and $i_0 = 3$:

$$(6.4) \qquad \begin{bmatrix} * & * & & & \\ & * & * & & \\ & & * & * & \\ & & & * & * \\ & & & & \end{bmatrix} \to \begin{bmatrix} * & * & & & \\ & * & * & & \\ & & * & * & w \\ & & & * & 0 \\ & & & & \end{bmatrix} \to \begin{bmatrix} * & * & & & \\ & * & * & & w \\ & & * & * & 0 \\ & & & * & 0 \\ & & & & \end{bmatrix} \to \begin{bmatrix} * & * & & & w \\ & * & * & & 0 \\ & & * & * & 0 \\ & & & * & 0 \\ & & & & \end{bmatrix}.$$

Here, "0" represents the element that was zeroed out and "w" is the nonzero that was newly created by the transformation. The Givens transformations are applied so that all but the bottom diagonals of the matrices in (6.4) are positive. We stop applying the transformations once w becomes negligibly small so that (6.12) is satisfied.

Let us denote $x = \sqrt{w}$ and express the effects of the ith Givens transformation in (6.4) as follows:

$$(6.5) \qquad \begin{bmatrix} * & * & & & \\ & * & \sqrt{\widehat{e}_j} & & \\ & & \sqrt{\widehat{q}_{j+1}} & * & \sqrt{x} \\ & & & * & 0 \\ & & & & \end{bmatrix} \begin{matrix} \\ G_R(k-i,k) \\ \to \\ \\ \\ \end{matrix} \begin{bmatrix} * & * & & & \\ & * & \sqrt{\breve{e}_j} & & \sqrt{\breve{x}} \\ & & \sqrt{\breve{q}_{j+1}} & * & 0 \\ & & & * & 0 \\ & & & & \end{bmatrix},$$

where $j = n - i - 1$. The triplet $(\breve{q}_{j+1}, \breve{e}_j, \breve{x})$ can be computed from $(\widehat{q}_{j+1}, \widehat{e}_j, x)$ by

$$\breve{q}_{j+1} = \widehat{q}_{j+1} + x, \quad \breve{e}_j = \frac{\widehat{q}_{j+1}\widehat{e}_j}{\widehat{q}_{j+1} + x}, \quad \breve{x} = \frac{x\widehat{e}_j}{\widehat{q}_{j+1} + x}. \tag{6.6}$$

Hence, Aggdef(2)-1 can be executed without computing square roots. Note that (6.6) provides a decreasing factor of the element x, i.e., $\breve{x} < x\frac{\widehat{e}_j}{\widehat{q}_{j+1}}$. Now, since $\widehat{B}_2$ converges to a diagonal matrix, the diagonal element $\widehat{q}_{j+1}$ is typically larger than the off-diagonal element $\widehat{e}_j$. This suggests that the size of $\breve{x}$ tends to decrease as it is chased up, i.e., $0 < \breve{x} \ll x$ if $\widehat{q}_{j+1} \gg \widehat{e}_j$. In practice, we observed that $\breve{x}$ often becomes negligible long before it reaches the top, that is, $i_0 \ll k - 2$.

6.2.2. Theoretical justifications. Here we express the key steps of Aggdef(2)-1 in matrix notations when it deflates a singular value. We denote by B and $\widetilde{B}$ the input and output n-by-n bidiagonals of Aggdef(2)-1 respectively.

Line 2 of Aggdef(2)-1 computes $\widehat{B}_2$ such that $\widehat{B}_2^T\widehat{B}_2 = B_2^T B_2 - sI$. Then, line 3 post-multiplies $\widehat{B}_2$ by the unitary matrix $\prod_{i=1}^{i_0} G_R(k-i, k) \equiv \begin{bmatrix} 1 & \\ & Q \end{bmatrix}$, where Q is a $(k-1)\times(k-1)$ unitary matrix. Once w in (6.4) becomes negligible it is set to 0, and we thus obtain the bidiagonal matrix $\breve{B}_2$ such that

$$\begin{aligned} \breve{B}_2^T\breve{B}_2 &= \left(\widehat{B}_2\begin{bmatrix} 1 & \\ & Q \end{bmatrix} - \begin{bmatrix} & w \end{bmatrix}\right)^T \left(\widehat{B}_2\begin{bmatrix} 1 & \\ & Q \end{bmatrix} - \begin{bmatrix} & w \end{bmatrix}\right) \\ &\equiv \begin{bmatrix} 1 & \\ & Q^T \end{bmatrix} B_2^T B_2 \begin{bmatrix} 1 & \\ & Q \end{bmatrix} - sI + E. \end{aligned} \tag{6.7}$$

We will carefully examine the "error matrix" E later in Section 6.2.3.

Finally, line 4 computes $\widetilde{B}_2$ such that

$$\begin{aligned} \widetilde{B}_2^T\widetilde{B}_2 &= \breve{B}_2^T\breve{B}_2 + sI \\ &= \begin{bmatrix} 1 & \\ & Q^T \end{bmatrix} B_2^T B_2 \begin{bmatrix} 1 & \\ & Q \end{bmatrix} + E. \end{aligned}$$

Since denoting $u_1 = [1, 0, 0, \cdots, 0]^T \in \mathbb{R}^{k\times 1}$ and $u_{n-k} = [0, 0, 0, \cdots, 1]^T \in \mathbb{R}^{n-k\times 1}$ we have

$$B^T B = \begin{bmatrix} B_1^T B_1 & \sqrt{q_{n-k}e_{n-k}}u_{n-k}u_1^T \\ \sqrt{q_{n-k}e_{n-k}}u_1 u_{n-k}^T & B_2^T B_2 + e_{n-k}u_1 u_1^T \end{bmatrix},$$

noting that $u_1^T \begin{bmatrix} 1 & \\ & Q \end{bmatrix} = u_1^T$ we obtain

$$\begin{aligned}
\widetilde{B}^T\widetilde{B} &= \begin{bmatrix} B_1^TB_1 & \sqrt{q_{n-k}e_{n-k}}u_{n-k}u_1^T \\ \sqrt{q_{n-k}e_{n-k}}u_1u_{n-k}^T & \widetilde{B}_2^T\widetilde{B}_2 + e_{n-k}u_1u_1^T \end{bmatrix} \\
&= \begin{bmatrix} I_{n-k} & & \\ & 1 & \\ & & Q^T \end{bmatrix} \begin{bmatrix} B_1^TB_1 & \sqrt{q_{n-k}e_{n-k}}u_{n-k}u_1^T \\ \sqrt{q_{n-k}e_{n-k}}u_1u_{n-k}^T & B_2^TB_2 + e_{n-k}u_1u_1^T \end{bmatrix} \begin{bmatrix} I_{n-k} & & \\ & 1 & \\ & & Q \end{bmatrix} \\
&\quad + \begin{bmatrix} & & \\ & & \\ & & E \end{bmatrix} \qquad (6.8) \\
&= \begin{bmatrix} I_{n-k+1} & \\ & Q^T \end{bmatrix} B^TB \begin{bmatrix} I_{n-k+1} & \\ & Q \end{bmatrix} + \begin{bmatrix} & \\ & E \end{bmatrix}. \qquad (6.9)
\end{aligned}$$

Later in Section 6.2.3 we show that the condition (6.12) implies $\|E\|_2$ is small enough to ensure $|\lambda_i(\widetilde{B}^T\widetilde{B} + SI) - \lambda_i(B^TB + SI)| \le 2cS\epsilon$ for a modest constant c, from which we conclude that high relative accuracy of the computed singular values is maintained.

The above arguments tell us that the entire process of Aggdef(2)-1 (which is to peel off the submatrix B_2, "shift" it by sI, multiply a unitary matrix, shift it back, then copy it back to the original B_2) is a valid process only because the unitary matrix $\prod_{i=1}^{i_0} G_R(k-i,k)$ we right-multiply to $\widehat{B}_2$ preserves the first column: multiplying a general unitary matrix destroys the crucial equality $u_1^T \begin{bmatrix} 1 & \\ & Q \end{bmatrix} = u_1^T$, and is not allowed here.

6.2.3. When to neglect elements. In this section, we derive conditions that ensure neglecting the bottom diagonal element $\sqrt{\widehat{q}_n}$ of $\widehat{B}_2$ and the error matrix E in (6.7) does not cause loss of relative accuracy of the computed singular values.

We first examine when it is safe to neglect a nonzero computed $\widehat{q}_n$.

First suppose that $\widehat{q}_n$ is positive. Since setting $\widehat{q}_n$ to zero only changes the bottom diagonal of $\widehat{B}_2^T\widehat{B}_2 + (s+S)I$ by $\widehat{q}_n$, Weyl's theorem ensures that high relative accuracy of the unconverged singular values is maintained if $\widehat{q}_n < cS\epsilon$ for a modest constant c.

Next consider the case $\widehat{q}_n < 0$. dstqds of Algorithm 7 computes $\widehat{q}_n$ as $\widehat{q}_n = q_n + d$, where d does not depend on q_n. Hence, setting $\widehat{q}_n$ to 0 is equivalent to replacing q_n of the original matrix $B_2^TB_2$ with $q_n - \widehat{q}_n$. Weyl's theorem applied to $B_2B_2^T + SI$ guarantees that high relative accuracy of the singular values is preserved if $|\widehat{q}_n| < cS\epsilon$.

In summary, we can safely neglect $\widehat{q}_n$ if

$$|\widehat{q}_n| \le cS\epsilon. \qquad (6.10)$$

We next examine when to neglect $w = \sqrt{x}$ (or equivalently E) when applying the Givens transformations. After setting $\widehat{q}_n$ to zero and applying i_0 Givens transformations to $\widehat{B}_2$, we

have $\widehat{B}_2^T\widehat{B}_2 + sI = B_2^T B_2$, where $\widehat{B}_2$ is of the form

$$\widehat{B}_2 = \begin{bmatrix} * & * & & & \\ & \sqrt{\widehat{q}_j} & \sqrt{\breve{e}_j} & & \sqrt{x} \\ & & * & * & \\ & & & * & \\ & & & & 0 \end{bmatrix}, \tag{6.11}$$

where $j = n - i_0 - 1$ is the row index of x. Then, recalling $x = w^2$, we see that E as in (6.7), (6.9) is

$$E = \begin{bmatrix} & & & -\sqrt{x\widehat{q}_j} \\ & & & -\sqrt{x\breve{e}_j} \\ & & & \\ -\sqrt{x\widehat{q}_j} & -\sqrt{x\breve{e}_j} & & x \end{bmatrix}.$$

Hence, Weyl's theorem ensures that the perturbation to the eigenvalues of $\widehat{B}_2^T\widehat{B}_2+(S+s)I$ caused by setting x to zero is bounded by $\|E\|_2 \le \sqrt{x(\widehat{q}_j + \breve{e}_j)} + x$. Therefore, it is safe to neglect x when $\sqrt{x(\widehat{q}_j + \breve{e}_j)} \le cS\epsilon$ and $x \le cS\epsilon$, or equivalently

$$x(\widehat{q}_j + \breve{e}_j) \le (cS\epsilon)^2 \quad \text{and} \quad x \le cS\epsilon. \tag{6.12}$$

In our numerical experiments, we set $c = 1.0$ in both (6.10) and (6.12).

We note that as the dqds iterations proceed, the sum of the previous shifts S typically becomes larger than $\widehat{q}_n, \widehat{q}_j$ and $\breve{e}_j$, so that the three inequalities all become more likely to hold. As a result, more singular values are expected to be deflated.

In the discussion here and in Section 6.1 we use only the Weyl bound. If some information on the gap between singular values is available, a sharper, quadratic perturbation bound can be used [**97, 109**]. We do not use such bounds here because estimating the gap is a nontrivial task, involving the whole matrix B instead of just B_2 or $\widehat{B}_2$, and experiments suggest that the improvement we get is marginal.

Let us give more details. In practice we are unwilling to spend $O(n)$ flops for estimating the gap, so instead we estimate the gap using only $\widehat{B}_2$. One choice is to estimate a lower bound for the smallest singular value $\sigma_{\min}$ of the top-left $(k-1)$-by-$(k-1)$ submatrix of B_2, and apply the bound in [**97**] to obtain the bound

$$|\sigma_i(\widehat{B}_2) - \sigma_i(\widehat{B}_{2,0})| \le \frac{2x}{\sigma_{\min} + \sqrt{\sigma_{\min}^2 + 4x}}, \tag{6.13}$$

where $\widehat{B}_{2,0}$ is the matrix obtained by setting x to 0 in (6.11). We emphasize that (6.13) is not a bound in terms of the entire matrix B, which is what we need to guarantee the desired accuracy. In practice estimating $\sigma_{\min}$ can be also costly, so we attempt to estimate it simply by $\sqrt{\widehat{q}_{n-1}}$. Combining this with (6.13), we tried neglecting the x values when

$$\frac{2x}{\sqrt{\widehat{q}_{n-1}} + \sqrt{\widehat{q}_{n-1} + 4x}} \le \sqrt{cS\epsilon}. \tag{6.14}$$

We observed through experiments that using this criterion sometimes results in loss of relative accuracy. Moreover, there was no performance gain on average, no particular case giving

more than 5% speedup. To guarantee relative accuracy while using a quadratic perturbation bound we need a more complicated and restrictive criterion than (6.14), which is unlikely to provide a faster implementation. Therefore we decide to use the simple and safe criterion (6.12) using Weyl's theorem.

6.2.4. High relative accuracy of Aggdef(2)-1 in floating point arithmetic. Here we show that Aggdef(2)-1 preserves high relative accuracy of singular values. We use the standard model of floating point arithmetic

$$fl(x \circ y) = (x \circ y)(1+\delta) = (x \circ y)/(1+\eta),$$

where $\circ \in \{+, -, \times, \div\}$ and $\delta,\ \eta$ satisfy

$$(1+\epsilon)^{-1}(x \circ y) \leq fl(x \circ y) \leq (1+\epsilon)(x \circ y).$$

For the error analysis below, we need to define $\widehat{B}_2$ clearly. In this subsection, we let $\widehat{B}_2$ be the first bidiagonal matrix in (6.4). In other words, $\widehat{B}_2$ is obtained by computing the dstqds transform from B_2 in floating point arithmetic, then setting the bottom element $\widehat{q}_n$ to 0, supposing that (6.10) is satisfied.

First we show that high relative accuracy of singular values of the lower right submatrices B_2 is preserved. We do this by using direct mixed stability analysis with respect to B_2, $\widehat{B}_2$, $\breve{B}_2$, $\widetilde{B}_2$, using an argument similar to that in [**45**, sec. 7.2].

Let us first analyze the transformation from B_2 to $\widehat{B}_2$. We introduce two ideal matrices $\dot{B}_2$, $\ddot{\widehat{B}}_2$ satisfying $\ddot{\widehat{B}}_2^T \ddot{\widehat{B}}_2 = \dot{B}_2^T \dot{B}_2 - sI$ for all but the bottom element $\ddot{\widehat{q}}_n$ of $\ddot{\widehat{B}}_2$, which is set to 0. We seek such $\dot{B}_2$ and $\ddot{\widehat{B}}_2$ so that $\dot{B}_2$ is a small relative perturbation of B_2 and $\ddot{\widehat{B}}_2$ is a small relative perturbation of $\widehat{B}_2$. In this subsection, we use a *dot* to denote backward type ideal matrices, and a *double dot* to denote forward type ideal matrices. The ith main and off-diagonals of $\dot{B}_2$ are $\dot{q}_i$ and $\dot{e}_i$, and those of $\ddot{\widehat{B}}_2$ are $\ddot{\widehat{q}}_i$ and $\ddot{\widehat{e}}_i$.

All the results in this subsection state errors in terms of the relative error, and we use the statement "$\dot{q}_i$ differs from q_i by $\alpha\epsilon$" to mean $(1+\epsilon)^{-\alpha}\dot{q}_i \leq q_i \leq (1+\epsilon)^{\alpha}\dot{q}_i$ ($\simeq (1+\alpha\epsilon)\dot{q}_i$). Below we specify the values of d as in Algorithm 7 and x as in (6.6) by denoting them with subscripts d_i and x_i.

LEMMA 6.1. *Concerning the mixed stability analysis in the transformation from B_2 to $\widehat{B}_2$, $\dot{q}_i$ differs from q_i by ϵ and $\dot{e}_i$ differs from e_i by 3ϵ, and $\widehat{q}_i$ differs from $\ddot{\widehat{q}}_i$ by 2ϵ and $\widehat{e}_i$ differs from $\ddot{\widehat{e}}_i$ by 2ϵ.*

PROOF. From the dstqds transform, we have

$$\begin{aligned}
\widehat{e}_i &= (q_i e_i/\widehat{q}_i)(1+\epsilon_{i,*1})(1+\epsilon_{i,/}),\\
d_{i+1} &= ((d_i e_i/\widehat{q}_i)(1+\epsilon_{i+1,*2})(1+\epsilon_{i,/}) - s)(1+\epsilon_{i+1,-}),\\
\widehat{q}_{i+1} &= (q_{i+1} + d_{i+1})(1+\epsilon_{i+1,+}).
\end{aligned}$$

From these equalities for d_{i+1} and $\widehat{q}_{i+1}$, we have

$$\frac{d_{i+1}}{1+\epsilon_{i+1,-}} = \frac{d_i e_i (1+\epsilon_{i+1,*2})(1+\epsilon_{i,/})}{(q_i + d_i)(1+\epsilon_{i,+})} - s.$$

This tells us how to define $\dot{B}_2$. We let them be

$$\dot{d}_{i+1} = d_{i+1}/(1+\epsilon_{i+1,-}), \tag{6.15}$$

$$\dot{q}_{i+1} = q_{i+1}/(1+\epsilon_{i+1,-}), \tag{6.16}$$

$$\dot{e}_i = e_i(1+\epsilon_{i+1,*2})(1+\epsilon_{i+1,/})/(1+\epsilon_{i,+}). \tag{6.17}$$

Then we see that

$$\dot{d}_{i+1} = \dot{d}_i\dot{e}_i/(\dot{q}_i + \dot{d}_i) - s,$$

so the recurrence for $\dot{d}_{i+1}$ of the dstqds transformation is satisfied. We then define the elements of the ideal $\ddot{\widehat{B}}_2$ as

$$\ddot{\widehat{q}}_{i+1} = \widehat{q}_{i+1}/(1+\epsilon_{i+1,+})(1+\epsilon_{i+1,-}), \tag{6.18}$$

$$\ddot{\widehat{e}}_i = \widehat{e}_i(1+\epsilon_{i+1,*2})/(1+\epsilon_{i,*1}). \tag{6.19}$$

Then the dstqds transformation from $\dot{B}_2$ to $\ddot{\widehat{B}}_2$, expressed in matrix form as $\ddot{\widehat{B}}_2^T\ddot{\widehat{B}}_2 = \dot{B}_2^T\dot{B}_2 + sI$, is satisfied. □

We next prove two lemmas regarding the connections between $\widehat{B}_2$, $\breve{B}_2$ and $\widetilde{B}_2$, and their corresponding ideal matrices denoted with dots. Similarly to $\widehat{B}_2$, the bidiagonal matrix $\breve{B}_2$ is here defined as the $(k-1)\times(k-1)$ deflated matrix obtained after applying the Givens transformations and setting x to 0.

LEMMA 6.2. *Concerning the mixed stability analysis in the transformation from $\widehat{B}_2$ to $\breve{B}_2$, we have $\dot{\widehat{q}}_i = \widehat{q}_i$, and $\dot{\widehat{e}}_i$ differs from $\widehat{e}_i$ by 3ϵ, $\breve{q}_i$ differs from $\ddot{\breve{q}}_i$ by ϵ and $\breve{e}_i$ differs from $\ddot{\breve{e}}_i$ by 2ϵ.*

PROOF. Recalling (6.6) we have

$$\begin{aligned}
\breve{q}_{i+1} &= (\widehat{q}_{i+1} + x_{i+1})(1+\epsilon_{i+1,+}),\\
\breve{e}_i &= \frac{\widehat{q}_{i+1}\widehat{e}_i(1+\epsilon_{i+1,*1})(1+\epsilon_{i+1,/})}{(\widehat{q}_{i+1}+x_{i+1})(1+\epsilon_{i+1,+})},\\
x_i &= \frac{x_{i+1}\widehat{e}_i(1+\epsilon_{i+1,*2})(1+\epsilon_{i+1,/})}{(\widehat{q}_{i+1}+x_{i+1})(1+\epsilon_{i+1,+})}.
\end{aligned}$$

Hence, we define the variables for the ideal matrix $\dot{\widehat{B}}_2$ as

$$\begin{aligned}
\dot{x}_i &= x_i,\\
\dot{\widehat{q}}_{i+1} &= \widehat{q}_{i+1},\\
\dot{\widehat{e}}_i &= \widehat{e}_i(1+\epsilon_{i+1,*2})(1+\epsilon_{i+1,/})/(1+\epsilon_{i+1,+}).
\end{aligned}$$

Then it follows that

$$\dot{x}_i = \frac{\dot{x}_{i+1}\dot{\widehat{e}}_i}{\dot{\widehat{q}}_{i+1} + \dot{x}_{i+1}},$$

so the recurrence for $\dot{x}_i$ is satisfied.

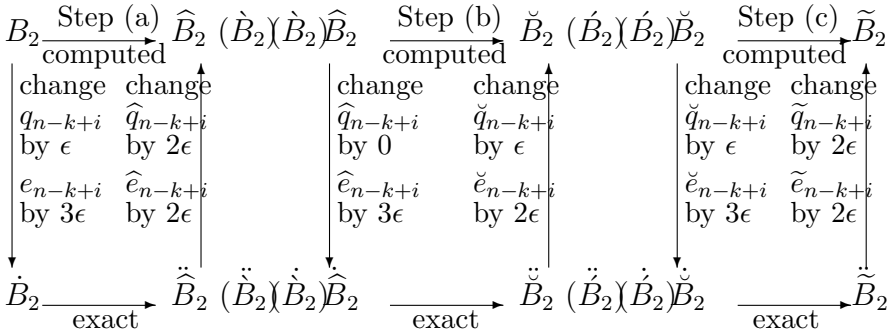

FIGURE 6.2.1. Effect of roundoff

Similarly, we define the variables for the ideal matrix $\ddot{\breve{B}}_2$ as

$$\ddot{\breve{q}}_{i+1} = \breve{q}_{i+1}/(1+\epsilon_{i+1,+}),$$
$$\ddot{\breve{e}}_i = \breve{e}_i(1+\epsilon_{i+1,*2})/(1+\epsilon_{i+1,*1}).$$

Then the transformation from $\dot{\widehat{B}}_2$ to $\ddot{\breve{B}}_2$ is realized in exact arithmetic. □

LEMMA 6.3. *Concerning the mixed stability analysis in the transformation from $\breve{B}_2$ to $\widetilde{B}_2$, $\dot{\breve{q}}_i$ differs from $\breve{q}_i$ by ϵ, $\dot{\breve{e}}_i$ differs from $\breve{e}_i$ by 3ϵ, $\widetilde{q}_i$ differs from $\ddot{\widetilde{q}}_i$ by 2ϵ and $\widetilde{e}_i$ differs from $\ddot{\widetilde{e}}_i$ by 2ϵ.*

PROOF. The proof is the same as in Lemma 6.1. □

The above results are summarized in Figure 6.2.1. We will discuss the matrices $\grave{B}_2$, $\acute{B}_2$ shortly.

Combining Lemma 6.1, (6.10) and a result by Demmel and Kahan [**35**, Corollary 2] that shows that the relative perturbation of bidiagonal elements produces only small relative perturbation in the singular values, we see that relative accuracy of the deflated singular value is preserved. We next show that Aggdef(2)-1 preserves high relative accuracy of all singular values of the whole bidiagonal matrix B.

The key idea of the proof below is to define bidiagonal matrices $\grave{B}_2$ and $\acute{B}_2$ satisfying $\grave{B}_2^T\grave{B}_2 = \widehat{B}_2^T\widehat{B}_2 + sI$ and $\acute{B}_2^T\acute{B}_2 = \breve{B}_2^T\breve{B}_2 + sI$ in exact arithmetic, so that we can discuss solely in terms of matrices that are not shifted by $-sI$. We first consider the bidiagonal matrices $\grave{B}_2$ and $\ddot{\grave{B}}_2$ satisfying $\grave{B}_2^T\grave{B}_2 = \widehat{B}_2^T\widehat{B}_2 + sI$ and $\ddot{\grave{B}}_2^T\ddot{\grave{B}}_2 = \ddot{\widehat{B}}_2^T\ddot{\widehat{B}}_2 + sI$. We have the following lemma.

LEMMA 6.4. *Concerning the relative errors between $\grave{B}_2$ and $\ddot{\grave{B}}_2$, $\ddot{\grave{q}}_{n-k+i}$ differs from $\ddot{\widehat{q}}_{n-k+i}$ by $4i\epsilon$ and $\ddot{\grave{e}}_{n-k+i}$ differs from $\ddot{\widehat{e}}_{n-k+i}$ by $4(i+1)\epsilon$ for $i = 1, \ldots, k$.*

PROOF. The dstqds transformation from $\widehat{B}_2$ to $\grave{B}_2$ gives

$$\grave{e}_i = \widehat{q}_i\widehat{e}_i/\grave{q}_i, \quad \grave{d}_{i+1} = \grave{d}_i\widehat{e}_i/\grave{q}_i + s, \quad \grave{q}_{i+1} = \widehat{q}_{i+1} + \grave{d}_{i+1}. \tag{6.20}$$

Hence

$$\grave{d}_{i+1} = \frac{\grave{d}_i \widehat{e}_i}{\widehat{q}_i + \grave{d}_i} + s = \frac{\widehat{e}_i}{\widehat{q}_i/\grave{d}_i + 1} + s. \tag{6.21}$$

Regarding the variables of $\ddot{B}$, by Lemma 6.1 the relative perturbation of $\widehat{q}_i$, $\widehat{e}_i$ from $\ddot{\widehat{q}}_i$, $\ddot{\widehat{e}}_i$ are both 2ϵ, that is,

$$(1+\epsilon)^{-2}\widehat{q}_i \le \ddot{\widehat{q}}_i \le (1+\epsilon)^2 \widehat{q}_i \tag{6.22}$$

$$(1+\epsilon)^{-2}\widehat{e}_i \le \ddot{\widehat{e}}_i \le (1+\epsilon)^2 \widehat{e}_i. \tag{6.23}$$

Moreover, similarly to (6.21) we have

$$\ddot{d}_{i+1} = \frac{\widehat{e}_i}{\ddot{\widehat{q}}_i/\ddot{d}_i + 1} + s.$$

Note that in the computation of $\grave{d}_i$, $\ddot{d}_i$, subtraction is not involved and $\grave{d}_{n-k+1} = \ddot{d}_{n-k+1} = s$. We claim that the relative perturbation of $\grave{d}_{n-k+i}$ of $\widehat{B}_2$ from $\ddot{d}_{n-k+i}$ of $\ddot{\widehat{B}}_2$ is less than $4i\epsilon$:

$$(1+\epsilon)^{-4i}\grave{d}_{n-k+i} \le \ddot{d}_{n-k+i} \le (1+\epsilon)^{4i}\grave{d}_{n-k+i} \tag{6.24}$$

for $i = 1, \ldots, k$. We can prove (6.24) by backward induction on i. For $i = k$ it is obvious. Next, if (6.24) holds for $i = j-1$, then for $i = j$ we have

$$\begin{aligned}
\ddot{d}_{n-k+j} &= \frac{\ddot{\widehat{e}}_{n-k+j-1}}{\ddot{\widehat{q}}_{n-k+j-1}/\ddot{d}_{n-k+j-1} + 1} + s \\
&\le \frac{\widehat{e}_{n-k+j-1}(1+\epsilon)^2}{\widehat{q}_{n-k+j-1}(1+\epsilon)^{-2}/\grave{d}_{n-k+j-1}(1+\epsilon)^{4(j-1)} + 1} + s \\
&\le \frac{\grave{d}_{n-k+j-1}\widehat{e}_{n-k+j-1}(1+\epsilon)^{4j}}{\widehat{q}_{n-k+j-1}/\grave{d}_{n-k+j-1} + (1+\epsilon)^{4j-2}} + s \\
&\le \frac{\grave{d}_{n-k+j-1}\widehat{e}_{n-k+j-1}(1+\epsilon)^{4j}}{\widehat{q}_{n-k+j-1}/\grave{d}_{n-k+j-1} + 1} + (1+\epsilon)^{4j}s \\
&= (1+\epsilon)^{4j}\grave{d}_{n-k+j}.
\end{aligned}$$

The first inequality in (6.24) can be shown similarly. Using (6.20), (6.23) and (6.24) we get

$$\begin{aligned}
(1+\epsilon)^{-4i}\grave{q}_{n-k+i} &= (1+\epsilon)^{-4i}(\widehat{q}_{n-k+i} + \grave{d}_{n-k+i}) \\
&\le \ddot{\widehat{q}}_{n-k+i} + \ddot{d}_{n-k+i} (= \ddot{q}_{n-k+i}) \\
&\le (1+\epsilon)^{4i}(\widehat{q}_{n-k+i} + \grave{d}_{n-k+i}) \\
&= (1+\epsilon)^{4i}\grave{q}_{n-k+i}.
\end{aligned}$$

Therefore,

$$(1+\epsilon)^{-4i}\grave{q}_{n-k+i} \le \ddot{q}_{n-k+i} \le (1+\epsilon)^{4i}\grave{q}_{n-k+i} \tag{6.25}$$

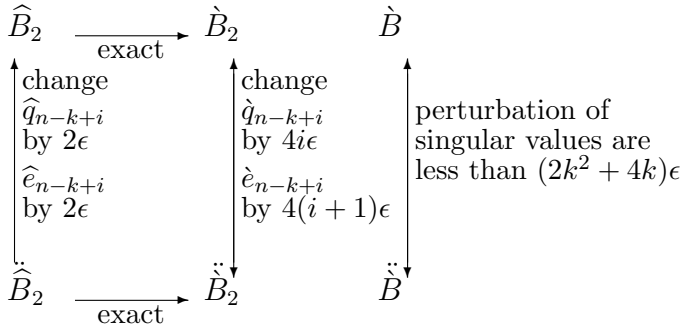

FIGURE 6.2.2. Effect of roundoff for singular values of $\ddot{\grave{B}}$ and $\grave{B}$

for $i = 1, \ldots, k$. Therefore the relative error between $\grave{q}_{n-k+i}$ and $\ddot{\grave{q}}_{n-k+i}$ is $4i\epsilon$. Similarly, we estimate the relative error between $\grave{e}_{n-k+i}$ and $\ddot{\grave{e}}_{n-k+i}$. We see that

$$\begin{aligned}
(1+\epsilon)^{-4(i+1)}\grave{e}_{n-k+i} &= (1+\epsilon)^{-4(i+1)}\widehat{q}_{n-k+i}\widehat{e}_{n-k+i}/\grave{q}_{n-k+i} \\
&\le \ddot{\widehat{q}}_{n-k+i}\ddot{\widehat{e}}_{n-k+i}/\ddot{\grave{q}}_{n-k+i} (= \ddot{\grave{e}}_{n-k+i}) \\
&\le (1+\epsilon)^{4(i+1)}\widehat{q}_{n-k+i}\widehat{e}_{n-k+i}/\grave{q}_{n-k+i} \\
&= (1+\epsilon)^{4(i+1)}\grave{e}_{n-k+i},
\end{aligned}$$

where we used (6.20), (6.23), (6.24) and (6.25). Therefore the relative error between $\grave{e}_{n-k+i}$ and $\ddot{\grave{e}}_{n-k+i}$ is $4(i+1)\epsilon$ for $i = 1, \ldots, k-1$. □

Now, from the $n \times n$ bidiagonal matrix B we define a new $n \times n$ bidiagonal matrix $\dot{B}$ obtained by replacing the lower right $k \times k$ part by $\dot{B}_2$. By Lemma 6.1 and Demmel-Kahan's result [**35**, Corollary 2], we have

$$\prod_{i=1}^{k}\sqrt{(1+\epsilon)^{-3}}\prod_{i=1}^{k-1}\sqrt{(1+\epsilon)^{-1}}\sigma_i \le \dot{\sigma}_i \le \prod_{i=1}^{k}\sqrt{(1+\epsilon)^{3}}\prod_{i=1}^{k-1}\sqrt{(1+\epsilon)}\sigma_i,$$

where $\dot{\sigma}_i$ denotes the ith singular value of $\dot{B}$. Here the square roots come from the facts $B_{i,i} = \sqrt{q_i}$ and $B_{i,i+1} = \sqrt{e_i}$. Using $\prod_{i=1}^{k}\sqrt{(1+\epsilon)^3}\prod_{i=1}^{k-1}\sqrt{(1+\epsilon)} \le \prod_{i=1}^{k}(1+\epsilon)^2 \le \exp(2k\epsilon)$ and an analogous inequality for the lower bound we get

$$\exp(2k\epsilon)^{-1}\sigma_i \le \dot{\sigma}_i \le \exp(2k\epsilon)\sigma_i. \quad \text{for} \quad i = 1, \ldots, n, \tag{6.26}$$

Similarly, we introduce $\grave{B}$, $\ddot{\grave{B}}$ whose lower right submatrix is replaced by $\grave{B}_2$, $\ddot{\grave{B}}_2$. By Lemma 6.4 and Demmel-Kahan's result, we have

$$\exp((2k^2+4k)\epsilon)^{-1}\ddot{\grave{\sigma}}_i \le \grave{\sigma}_i \le \exp((2k^2+4k)\epsilon)\ddot{\grave{\sigma}}_i \quad \text{for} \quad i = 1, \ldots, n, \tag{6.27}$$

where $\ddot{\grave{\sigma}}_i$ and $\grave{\sigma}_i$ are the singular values of $\ddot{\grave{B}}$ and $\grave{B}$. This analysis is summarized in Figure 6.2.2.

Recall that assuming (6.10) is satisfied, we set $\widehat{q}_n$ and $\ddot{\widehat{q}}_n$ to 0. We next bound the effect of the operation $\ddot{\widehat{q}}_n \leftarrow 0$ on the singular values of $\dot{B}_2$ and $\ddot{\grave{B}}_2$. Noting that the bounds in Lemma 6.1 hold for the bottom elements $\widehat{q}_n$ and $\ddot{\widehat{q}}_n$ even before setting them to 0, by the

argument leading to (6.10) we obtain (recall that $\sqrt{\sigma_i^2+S}$ are the singular values to be computed)

$$(1-c(1+\epsilon)^2\epsilon)(\dot{\sigma}_i^2+S) \le \ddot{\sigma}_i^2+S \le (1+c(1+\epsilon)^2\epsilon)(\dot{\sigma}_i^2+S) \quad \text{for} \quad i=1,\ldots,n. \tag{6.28}$$

For simplicity we rewrite (6.28) as

$$\exp(c'\epsilon)^{-1}\sqrt{\dot{\sigma}_i^2+S} \le \sqrt{\ddot{\sigma}_i^2+S} \le \exp(c'\epsilon)\sqrt{\dot{\sigma}_i^2+S} \quad \text{for} \quad i=1,\ldots,n, \tag{6.29}$$

where $c'(\approx c/2)$ is a suitable constant such that the original inequality (6.28) is satisfied.

Since $S \ge 0$, we see that (6.26) implies $\exp(2k\epsilon)^{-1}\sqrt{\sigma_i^2+S} \le \sqrt{\dot{\sigma}_i^2+S} \le \exp(2k\epsilon)\sqrt{\sigma_i^2+S}$, and an analogous inequality holds for (6.27). Combining the three bounds we obtain a bound for the relative error in Step (a) in Figure 6.2.1

$$\exp((2k^2+6k+c')\epsilon)^{-1}\sqrt{\sigma_i^2+S} \le \sqrt{\grave{\sigma}_i^2+S} \le \exp((2k^2+6k+c')\epsilon)\sqrt{\sigma_i^2+S} \tag{6.30}$$

for $i=1,\ldots,n$.

We next discuss Step (b) in Figure 6.2.1. Similarly to the above discussion, we define a bidiagonal matrix $\dot{\grave{B}}_2$ satisfying $\dot{\grave{B}}_2^T\dot{\grave{B}}_2 = \dot{\widehat{B}}_2^T\dot{\widehat{B}}_2 + sI$ in exact arithmetic.

LEMMA 6.5. *$\dot{\grave{q}}_{n-k+i}$ differs from $\grave{q}_{n-k+i}$ by $3i\epsilon$ and $\dot{\grave{e}}_{n-k+i}$ differs from $\grave{e}_{n-k+i}$ by $3(i+1)\epsilon$ for $i=1,\ldots,k$.*

PROOF. Similarly to (6.22) and (6.23), by Lemma 6.2 we have

$$\dot{\widehat{q}}_i = \widehat{q}_i \tag{6.31}$$

$$(1+\epsilon)^{-3}\widehat{e}_i \le \dot{\widehat{e}}_i \le (1+\epsilon)^3\widehat{e}_i \tag{6.32}$$

Therefore, similarly to (6.24), we have

$$(1+\epsilon)^{-3i}\grave{d}_{n-k+i} \le \dot{\grave{d}}_{n-k+i} \le (1+\epsilon)^{3i}\grave{d}_{n-k+i}, \tag{6.33}$$

so the proof is completed as in Lemma 6.4. □

Therefore, we have

$$\exp((3k^2/2+3k)\epsilon)^{-1}\grave{\sigma}_i \le \dot{\grave{\sigma}}_i \le \exp((3k^2/2+3k)\epsilon)\grave{\sigma}_i \quad \text{for} \quad i=1,\ldots,n. \tag{6.34}$$

We next define and compare the bidiagonal matrices $\acute{B}_2$ and $\ddot{\acute{B}}_2$ satisfying $\acute{B}_2^T\acute{B}_2 = \breve{B}_2^T\breve{B}_2 + sI$ and $\ddot{\acute{B}}_2^T\ddot{\acute{B}}_2 = \ddot{\breve{B}}_2^T\ddot{\breve{B}}_2 + sI$ in exact arithmetic.

LEMMA 6.6. *$\acute{q}_{n-k+i}$ differs from $\ddot{\acute{q}}_{n-k+i}$ by $3i\epsilon$ and $\acute{e}_{n-k+i}$ differs from $\ddot{\acute{e}}_{n-k+i}$ by $3(i+1)\epsilon$ for $i=1,\ldots,k$.*

PROOF. By Lemma 6.2 we have

$$(1+\epsilon)^{-1}\breve{q}_i \le \ddot{\breve{q}}_i \le (1+\epsilon)\breve{q}_i$$
$$(1+\epsilon)^{-2}\breve{e}_i \le \ddot{\breve{e}}_i \le (1+\epsilon)^2\breve{e}_i.$$

Therefore, similarly to (6.24), we have

$$(1+\epsilon)^{-3i}\acute{d}_{n-k+i} \le \ddot{\acute{d}}_{n-k+i} \le (1+\epsilon)^{3i}\acute{d}_{n-k+i}. \tag{6.35}$$

The same argument as in Lemma 6.4 completes the proof. □

Therefore, we have

$$\exp((3k^2/2+3k)\epsilon)^{-1}\ddot{\acute{\sigma}}_i \le \acute{\sigma}_i \le \exp((3k^2/2+3k)\epsilon)\ddot{\acute{\sigma}}_i \quad \text{for} \quad i=1,\dots,n. \tag{6.36}$$

Recall that the kth column of $\breve{B}_2$ is set to the zero vector when (6.12) is satisfied, and hence by Lemma 6.2 we see that

$$(1-2c(1+\epsilon)\epsilon)(\dot{\grave{\sigma}}_i^2+S) \le (\ddot{\acute{\sigma}}_i^2+S) \le (1+2c(1+\epsilon)\epsilon)(\dot{\grave{\sigma}}_i^2+S) \quad \text{for} \quad i=1,\dots,n.$$

For simplicity, we rewrite the inequalities using a suitable constant $c''(\approx c)$ as

$$\exp(c''\epsilon)^{-1}\sqrt{\dot{\grave{\sigma}}_i^2+S} \le \sqrt{\ddot{\acute{\sigma}}_i^2+S} \le \exp(c''\epsilon)\sqrt{\dot{\grave{\sigma}}_i^2+S} \quad \text{for} \quad i=1,\dots,n. \tag{6.37}$$

Combining (6.34), (6.36) and (6.37) we get

$$\exp((3k^2+6k+c'')\epsilon)^{-1}\sqrt{\grave{\sigma}_i^2+S} \le \sqrt{\acute{\sigma}_i^2+S} \le \exp((3k^2+6k+c'')\epsilon)\sqrt{\grave{\sigma}_i^2+S} \tag{6.38}$$

for $i=1,\dots,n$.

Finally, we bound the relative error caused in Step (c). Let $\dot{\acute{B}}_2$ be a bidiagonal matrix satisfying $\dot{\acute{B}}_2^T\dot{\acute{B}}_2 = \dot{\breve{B}}_2^T\dot{\breve{B}}_2 + sI$ in exact arithmetic. We have the following lemma comparing $\acute{B}_2$ and $\dot{\acute{B}}_2$.

Lemma 6.7. *$\acute{q}_{n-k+i}$ differs from $\dot{\acute{q}}_{n-k+i}$ by $4i\epsilon$ and $\acute{e}_{n-k+i}$ differs from $\dot{\acute{e}}_{n-k+i}$ by $4(i+1)\epsilon$ for $i=1,\dots,k$.*

Proof. By Lemma 6.3 we have

$$(1+\epsilon)^{-1}\breve{q}_i \le \dot{\breve{q}}_i \le (1+\epsilon)\breve{q}_i$$
$$(1+\epsilon)^{-3}\breve{e}_i \le \dot{\breve{e}}_i \le (1+\epsilon)^3\breve{e}_i.$$

Therefore, similarly to (6.24), we have

$$(1+\epsilon)^{-4i}\acute{d}_{n-k+i} \le \dot{\acute{d}}_{n-k+i} \le (1+\epsilon)^{4i}\acute{d}_{n-k+i}. \tag{6.39}$$

The same argument as in Lemma 6.4 completes the proof. □

From this lemma, we get

$$\exp((2k^2+4k)\epsilon)^{-1}\acute{\sigma}_i \le \dot{\acute{\sigma}}_i \le \exp((2k^2+4k)\epsilon)\acute{\sigma}_i \quad \text{for} \quad i=1,\dots,n,$$

with the aid of Demmel-Kahan's result. Moreover, using Lemma 6.3 we get

$$\exp(2k\epsilon)^{-1}\dot{\acute{\sigma}}_i \le \widetilde{\sigma}_i \le \exp(2k\epsilon)\dot{\acute{\sigma}}_i \quad \text{for} \quad i=1,\dots,n.$$

Therefore, we obtain

$$\exp((2k^2+6k)\epsilon)^{-1}\acute{\sigma}_i \le \widetilde{\sigma}_i \le \exp(2k^2+6k)\epsilon)\acute{\sigma}_i, \tag{6.40}$$

for $i=1,\dots,n$.

Now we present the main result of this subsection.

THEOREM 6.1. *Aggdef(2)-1 preserves high relative accuracy. The singular values* $\sigma_1 > \cdots > \sigma_n$ *of* B *and* $\widehat{\sigma}_1 > \cdots > \widehat{\sigma}_n$ *of* $\widehat{B}$ *and the sum of previous shifts* S *satisfy*

$$\exp((7k^2+18k+C)\epsilon)^{-1}\sqrt{\sigma_i^2+S} \le \sqrt{\widehat{\sigma}_i^2+S} \le \exp((7k^2+18k+C)\epsilon)\sqrt{\sigma_i^2+S} \tag{6.41}$$

for $i = 1, \ldots, n$, *where* $C = c' + c''$.

PROOF. Combine (6.30), (6.38) and (6.40). □

We note that in practice we always let the window size k be $k \le \sqrt{n}$ (see Section 6.4.2), so the bound (6.41) gives an relative error bound of order $O(n\epsilon)$, which has the same order as the bound for one dqds iteration derived in [**45**]. Also note that in our experiments $C \simeq 1.5$, because we let $c = 1.0$. We conclude that executing Aggdef(2)-1 does not affect the relative accuracy of dqds.

As discussed below, in Aggdef(2) we execute Aggdef(2)-1 repeatedly to deflate $\ell(>1)$ singular values. In this case we have

$$\exp((7k^2+18k+C)\ell\epsilon)^{-1}\sqrt{\sigma_i^2+S} \le \sqrt{\widehat{\sigma}_i^2+S} \le \exp((7k^2+18k+C)\ell\epsilon)\sqrt{\sigma_i^2+S}$$

for $i = 1, \ldots, n$, where ℓ is the number of deflated singular values by Aggdef(2).

6.2.5. Overall algorithm Aggdef(2). As mentioned above, Aggdef(2)-1 deflates only one singular value. To deflate $\ell(>1)$ singular values we execute Aggdef(2), which is mathematically equivalent to ℓ runs of Aggdef(2)-1, but is cheaper saving ℓ calls of dstqds. Algorithm 8 is its pseudocode.

Algorithm 8 Aggressive early deflation - version 2: Aggdef(2)

Inputs: Bidiagonal B, window size k, sum of previous shifts S

1: $C = B_2$, $\ell = 0$
2: compute $s_{\ell+1} = (\sigma_{\min}(C))^2$
3: compute $\widehat{B}_2$ such that $\widehat{B}_2^T\widehat{B}_2 = C^TC - s_{\ell+1}I$ by dstqds. Set $\widehat{B}_2(\text{end},\text{end}) \leftarrow 0$ if (6.10) holds, otherwise go to line 6
4: compute $\breve{B}_2 = \widehat{B}_2\prod_{i=1}^{i_0} G_R(k-i,k)$ for $i_0 = 1, \ldots, k-2$ until (6.12) holds. Go to line 6 if (6.12) never holds
5: $C := \breve{B}_2(1:\text{end}-1, 1:\text{end}-1)$, $\ell \leftarrow \ell+1$, $k \leftarrow k-1$, go to line 2
6: compute $\widetilde{B}_2$ such that $\widetilde{B}_2^T\widetilde{B}_2 = C^TC + \sum_{i=1}^{\ell} s_iI$ by dstqds and update B by replacing B_2 with $\mathrm{diag}(\tilde{B}_2, \mathrm{diag}(\sqrt{\sum_{j=1}^{\ell} s_j}, \ldots, \sqrt{\sum_{j=1}^{2} s_j}, \sqrt{s_1}))$.

6.2.6. Relation between Aggdef(2) and other methods. In this subsection, we examine the relation between Aggdef(2) and previously-proposed methods including Aggdef(1).

6.2.6.1. *Comparison with Aggdef(1).* First, it should be stressed that Aggdef(2) is computationally more efficient than Aggdef(1). Specifically, in contrast to Aggdef(1), which always needs $O(k^2)$ flops, Aggdef(2) requires $O(k\ell)$ flops when it deflates ℓ singular values. Hence, even when only a small number of singular values are deflated (when $\ell \ll k$), Aggdef(2) wastes very little work. In addition, as we saw above, unlike Aggdef(1), Aggdef(2) preserves high relative accuracy of the computed singular values, regardless of the window size k.

However, it is important to note that Aggdef(1) and Aggdef(2) are not mathematically equivalent, although closely related. To see the relation between the two, let us define the k-by-k unitary matrices $Q_i = \prod_{j=1}^{i} G_R(k-j,k)$ for $i = 1, \ldots, k-2$. After i Givens transformations are applied on line 4 of Aggdef(2), $Q_i(k-i:k,k)$ (the $i+1$ bottom part of the last column) is parallel to the corresponding part of $v = [v_1, \ldots, v_k]^T$, the null vector of $\widehat{B}_2$. This can be seen by recalling that $\widehat{B}_2$ is upper-bidiagonal and the bottom $i+1$ elements of $\widehat{B}_2 Q_i(:,k)$ are all zeros. Note that v is also the right-singular vector corresponding to $\sigma_{\min}(B_2)$.

Hence in particular, after $k-2$ (the largest possible number) Givens transformations have been applied we have $Q_{k-2}(2:k,k) = v(2:k)/\sqrt{1-v_1^2}$. It follows that w in (6.4) is $w = \sqrt{\widehat{e}_{n-k+1}}v_2/\sqrt{1-v_1^2} = -\sqrt{\widehat{q}_{n-k+1}}v_1/\sqrt{1-v_1^2}$, where we used the fact $\sqrt{\widehat{q}_{n-k+1}}v_1 + \sqrt{\widehat{e}_{n-k+1}}v_2 = 0$. Hence recalling (6.12) and $x = w^2$, we conclude that Aggdef(2) deflates $\sqrt{S+s}$ as a converged singular value if $\frac{|v_1|}{\sqrt{1-|v_1|^2}} < \min\{S\epsilon/\sqrt{\widehat{q}_{n-k+1}(\widehat{q}_{n-k+1}+\widehat{e}_{n-k+1})}, \sqrt{S\epsilon/\widehat{q}_{n-k+1}}\}$. On the other hand, as we discussed in Section 6.1.1, Aggdef(1) deflates $\sqrt{S+s}$ if $|v_1| < \sqrt{S}\epsilon/\sqrt{e_{n-k+1}}$. We conclude that Aggdef(1) and Aggdef(2) are similar in the sense that both deflate the smallest singular value of B_2 when the first element of its right singular vector v_1 is small enough, albeit with different tolerances.

The fundamental difference between Aggdef(1) and Aggdef(2) is that Aggdef(1) deflates all the deflatable singular values at once, while Aggdef(2) attempts to deflate singular values one by one from the smallest ones. As a result, Aggdef(2) deflates only the smallest singular values of B_2, while Aggdef(1) can detect the converged singular values that are not among the smallest. Consequently, sometimes fewer singular values are deflated by Aggdef(2). However this is not a serious defect of Aggdef(2), since as we show in Section 6.3.1, smaller singular values are more likely to be deflated via aggressive early deflation. The numerical results in Section 6.4 also show that the total numbers of singular values deflated by the two strategies are typically about the same.

6.2.6.2. *Relation with Sorensen's deflation strategy.* A deflation strategy closely related to Aggdef(2) is that proposed by Sorensen ([**139**], [**6**, Sec.4.5.7]) for restarting the Arnoldi or Lanczos algorithm. Just like Aggdef(2), the strategy attempts to deflate one converged eigenvalue at a time. Its idea can be readily applied for deflating a converged eigenvalue in a k-by-k bottom-right submatrix A_2 of a n-by-n symmetric tridiagonal matrix A as in (2.19). An outline of the process is as follows: Let (λ, v) be an eigenpair of A_2 with $v(k) \neq 0$. Sorensen defines the special k-by-k orthogonal matrix $Q_S = L + vu_k^T$, where $u_k = [0, \ldots, 1]$ and L is lower triangular with nonnegative diagonals except $L(k,k) = 0$ (see [**139, 6**] for

a precise formulation of L). L also has the property that for $i = 1, \ldots, k-1$, the below-diagonal part of the ith column $L(i+1:k, i)$ is parallel to $v(i+1:k)$. In [**139**] it is shown that $\mathrm{diag}(I_{n-k}, Q_S^T) A \mathrm{diag}(I_{n-k}, Q_S) = \begin{bmatrix} A_1 & t^T \\ t & \widehat{A}_2 \end{bmatrix}$ for $\widehat{A}_2 = \mathrm{diag}(T, \lambda)$ and $t = [\widehat{b}_{n-k}, 0, \ldots, 0, b_{n-k} v(1)]^T$, where T is a $(k-1)$-by-$(k-1)$ tridiagonal matrix. Therefore λ can be deflated if $b_{n-k} v(1)$ is negligibly small.

Now we discuss the close connection between Sorensen's deflation strategy and Aggdef(2). Recall the definition $Q_i = \prod_{j=1}^{i} G_R(n-j, n)$. We shall see that Q_{k-2} and Q_S are closely related. To do so, we first claim that a unitary matrix with the properties of Q_S is uniquely determined by the vector v. To see this, note that [**60**] shows that for any vector v there exists a unique unitary upper Hessenberg matrix expressed as a product of $n-1$ Givens transformations, whose last column is v. We also note that such unitary Hessenberg matrices is discussed in [**130**]. Now, by permuting the columns of Q_S by right-multiplying the permutation matrix P such that $P(i, i+1) = 1$ for $i = 1, \ldots, k-1$ and $P(k, 1) = 1$, and taking the transpose we get a unitary upper Hessenberg matrix $P^T Q_S^T$ whose last column is v. Therefore we conclude that such Q_S is unique. Recalling that for $i = 1, \ldots, k-2$ the last column of Q_i is parallel to $[0, \ldots, 0, v(k-i:k)^T]^T$, and noting that the irreducibility of B_2 ensures $v(k) \neq 0$ and that the diagonals of Q_i are positive (because the diagonals of (6.4) are positive), we conclude that $Q_S = Q_{k-2} \cdot G_R(n-k+1, n)$. Here, the Givens transformation $G_R(n-k+1, n)$ zeros the top-right w if applied to the last matrix in (6.4)[1]. Conversely, Q_i can be obtained in the same way as Q_S, by replacing v with a unit vector parallel to $[0, 0, \ldots, 0, v_{k-i}, \ldots, v_k]^T$. Therefore recalling (6.9), we see that performing Aggdef(2) can be regarded as successively and implicitly forming $\mathrm{diag}(I_{n-k}, Q_i^T) B^T B \mathrm{diag}(I_{n-k}, Q_i)$ where Q_i is a truncated version of Q_S.

Table 6.2.1 summarizes the relation between aggressive early deflation (AED) as in [**16**], Sorensen's strategy, Aggdef(1) and Aggdef(2).

TABLE 6.2.1. Summary of deflation strategies.

	Hessenberg	bidiagonal
deflate all at once	AED [**16**]	Aggdef(1)
deflate one at a time	Sorensen [**139**]	[**92**], Aggdef(2)

While being mathematically nearly equivalent to Sorensen's strategy, Aggdef(2) achieves significant improvements in both efficiency and stability. To emphasize this point, we compare Aggdef(2) with another, more straightforward extension of Sorensen's deflation strategy for the bidiagonal case, described in [**92**]. The authors in [**92**] use both the left and right singular vectors of a target singular value to form two unitary matrices Q_S and P_S (determined by letting y be the singular vectors that determines the unitary matrix) such that $\mathrm{diag}(I_{n-k}, P_S^T) \cdot B \cdot \mathrm{diag}(I_{n-k}, Q_S)$ is bidiagonal except for the nonzero $(n-k, n)$th element, and its bottom diagonal is "isolated". Computing this orthogonal transformation requires at

[1]We do not allow applying the transformation $G_R(n-k+1, n)$ in Aggdef(2) for the reason mentioned in Section 6.2.2.

least $O(k^2)$ flops. It can also cause loss of relative accuracy of the computed singular values. In contrast, Aggdef(2) completely bypasses P_S (whose effect is implicitly "taken care of" by the two dstqds transformations) and applies the truncated version of Q_S without forming it. The resulting rewards are immense: the cost is reduced to $O(k)$ and high relative accuracy is guaranteed.

6.3. Convergence analysis

In this section we develop convergence analyses of dqds with aggressive early deflation. Specifically, we derive convergence factors of the x elements in (6.5), which explain why aggressive early deflation improves the performance of dqds. Our analyses also show that aggressive early deflation makes a sophisticated shift strategy less vital for convergence speed, making the zero-shift variant dqd a competitive alternative to dqds.

Our analysis focuses on Aggdef(2), because it is more efficient and always stable, and outperforms Aggdef(1) in all our experiments. In addition, as we discussed in Section 6.2.6.1, the two strategies are mathematically closely related. We start by estimating the value of the x elements as in (6.5) in Aggdef(2). Then in Section 6.3.2 we study the impact on x of one dqds iteration.

6.3.1. Bound for x elements. As reviewed in Section 2.5, in the dqds algorithm the diagonals $\sqrt{q_i}$ of B converge to the singular values in descending order of magnitude, and the ith off-diagonal element $\sqrt{e_i}$ converge to zero with the convergence factor $\frac{{\sigma_{i+1}}^2-S}{{\sigma_i}^2-S}$ [**1**]. In view of this, here we assume that q_i are roughly ordered in descending order of magnitude, and that the off-diagonals e_i are small so that $e_i \ll q_i$.

Under these assumptions we claim that the dstqds step changes the matrix B_2 only slightly, except for the bottom element $\sqrt{q_n}$ which is mapped to 0. To see this, note that since $s < q_n$, we have $\widehat{q}_{n-k+1} = q_{n-k+1} - s \simeq q_{n-k+1}$, which also implies $\widehat{e}_{n-k+1} \simeq e_{n-k+1}$. Now since by assumption we have $e_i \ll q_i \simeq \widehat{q}_i$, so $d \simeq -s$ throughout the dstqds transformation. Therefore the claim follows.

Now consider the size of the x element in Aggdef(2)-1 when it is chased up to the top, $(n-k+1, n)$ element. For definiteness here we denote by x_i the value of x after i transformations are applied. Initially we have $x_0 = \widehat{e}_{n-1} \simeq e_{n-1}$. Then the Givens transformations are applied, and as seen in (6.6), the ith transformation reduces x by a factor $\frac{\widehat{e}_{n-i-1}}{\widehat{q}_{n-i}+x_i}$. Therefore after the application of $k-2$ transformations x becomes

$$x = \widehat{e}_{n-1} \prod_{i=1}^{k-2} \frac{\widehat{e}_{n-i-1}}{\widehat{q}_{n-i} + x_i} \leq \widehat{e}_{n-1} \prod_{i=1}^{k-2} \frac{\widehat{e}_{n-i-1}}{\widehat{q}_{n-i}} \simeq e_{n-1} \prod_{i=1}^{k-2} \frac{e_{n-i-1}}{q_{n-i}}. \tag{6.42}$$

This is easily seen to be small when $q_{n-i} \gg e_{n-i-1}$ for $i = 1, \ldots, k-2$, which necessarily holds in the asymptotic stage of dqds convergence. Note that asymptotically we have $x_i \ll \widehat{q}_{n-i}$, so that $\widehat{q}_{n-i} + x_i \simeq \widehat{q}_{n-i}$ for $i = 1, \ldots, k-2$, and so all the inequalities and approximations in (6.42) become an equality.

Now we consider deflating the $\ell(\geq 2)$th smallest singular value of B_2 via the ℓth run of Aggdef(2)-1 in Aggdef(2), assuming that the smallest $\ell - 1$ singular values have been deflated. Here we denote by x_ℓ the x element after the Givens transformations. Under

the same asymptotic assumptions as above we see that after the maximum $(k-\ell-1)$ transformations the x_ℓ element is

$$x_\ell = \widehat{e}_{n-\ell} \prod_{i=1}^{k-\ell-1} \frac{\widehat{e}_{n-i-\ell}}{\widehat{q}_{n-i-\ell+1} + x_i} \leq \widehat{e}_{n-\ell} \prod_{i=1}^{k-\ell-1} \frac{\widehat{e}_{n-i-\ell}}{\widehat{q}_{n-i-\ell+1}} \simeq e_{n-\ell} \prod_{i=1}^{k-\ell-1} \frac{e_{n-i-\ell}}{q_{n-i-\ell+1}}. \tag{6.43}$$

Several remarks regarding (6.43) are in order.

- While the analyses in [**16, 94**] are applicable to the more general Hessenberg matrix, our result exhibits several useful simplifications by specializing in the bidiagonal case. Our bound (6.43) on x_ℓ involves only the elements of B_2. On the other hand, the bound (2.18) in [**16**], which bound the spike vector elements in Aggdef(1) (after interpreting the problem in terms of computing the eigenvalues of B^TB), is difficult to use in practice because it requires information about the eigenvector.
- By (6.43) we see that x_ℓ is typically larger for large ℓ. This is because for large ℓ, fewer Givens transformations are applied, and $e_{n-\ell}$ tends to be smaller for small ℓ, because as can be seen by (2.15), e_{n-i} typically converges via the dqds iterations with a smaller convergence factor for small i. This suggests that Aggdef(2) detects most of the deflatable singular values of B_2, because it looks for deflatable singular values from the smallest ones. Together with the discussion in Section 6.2.6.1, we argue that the number of singular values deflated by Aggdef(1) and Aggdef(2) are expected to be similar. Our numerical experiments confirm that this is true in most cases.
- (6.43) indicates that x_ℓ can be regarded as converged when $e_{n-\ell}\prod_{i=1}^{k-\ell-1}\frac{e_{n-i-\ell}}{q_{n-i-\ell+1}}$ is small. This is essentially proportional to the product of the off-diagonal elements $e_{n-i-\ell}$ for $i = 0, 1, \ldots, k-\ell-1$, because once convergence reaches the asymptotic stage the denominators $q_{n-i-\ell+1}$ converge to the constants $\sigma_{n-i-\ell+1}$. Hence, (6.43) shows that x_ℓ can be deflated when the product $\prod_{i=0}^{k-\ell-1} e_{n-i-\ell}$ is negligibly small, which can be true even if none of $e_{n-i-\ell}$ is.

6.3.2. Impact of one dqd iteration. Here we study the convergence factor of x_ℓ when one dqd (without shift, we discuss the effect of shifts shortly) iteration is run. As we reviewed in the introduction, in the asymptotic stage we have $q_i \simeq \sigma_i^2$, and $e_i \to 0$ with the convergence factor $\sigma_{i+1}^2/\sigma_i^2$ [**1**]. This is an appropriate measure for the convergence factor of the dqd(s) algorithm with a conventional deflation strategy. On the other hand, when Aggdef(2) is employed, the convergence factor of x_ℓ is a more natural way to measure convergence. In this section, we discuss the impact of running one dqd iteration on x_ℓ.

In this subsection we denote by $\widetilde{B}$ the bidiagonal matrix obtained by running one dqd step on B, and let $\widetilde{q}_i, \widetilde{e}_i$ be the bidiagonal elements of $\widetilde{B}$. Similarly denote by $\widetilde{x}$ the value of x when Aggdef(2) is applied to $\widetilde{B}$. Then, in the asymptotic stage we have $\widetilde{q}_i \simeq q_i \simeq \sigma_i^2$ and $\widetilde{e}_i \simeq \frac{\sigma_{i+1}^2}{\sigma_i^2} e_i$. Then by the estimate (6.43) we have

$$\widetilde{x}_\ell \simeq \widetilde{e}_{n-\ell} \prod_{i=1}^{k-\ell-1} \frac{\widetilde{e}_{n-i-\ell}}{\widetilde{q}_{n-i-\ell+1}}.$$

It follows that

$$\begin{aligned}\frac{\widetilde{x}_\ell}{x_\ell} &\simeq \frac{\prod_{i=0}^{k-\ell-1}\widetilde{e}_{n-i-\ell}}{\prod_{i=0}^{k-\ell-1} e_{n-i-\ell}} \cdot \frac{\prod_{i=1}^{k-\ell-1}\widetilde{q}_{n-i-\ell+1}}{\prod_{i=1}^{k-\ell-1}\widetilde{q}_{n-i-\ell+1}} \\ &\simeq \frac{\prod_{i=0}^{k-\ell-1} e_{n-i-\ell}(\sigma^2_{n-i-\ell+1}/\sigma^2_{n-i-\ell})}{\prod_{i=0}^{k-\ell-1} e_{n-i-\ell}} \cdot \frac{\prod_{i=1}^{k-\ell-1}\widetilde{q}_{n-i-\ell+1}}{\prod_{i=1}^{k-\ell-1}\widetilde{q}_{n-i-\ell+1}} \\ &= \prod_{i=0}^{k-\ell-1} \frac{\sigma^2_{n-i-\ell+1}}{\sigma^2_{n-i-\ell}} = \frac{\sigma^2_{n-\ell+1}}{\sigma^2_{n-k+1}}. \end{aligned} \tag{6.44}$$

(6.44) suggests that asymptotically one dqd iteration reduces the magnitude of x_ℓ by an approximate factor $\dfrac{\sigma^2_{n-\ell+1}}{\sigma^2_{n-k+1}}$. This is generally much smaller than $\dfrac{\sigma^2_{n-\ell+1}}{\sigma^2_{n-\ell}}$, the asymptotic convergence factor of $e_{n-\ell}$ (which needs to be small to deflate ℓ smallest singular values by a conventional deflation strategy). We note that the estimate (6.44) is still sharp in the asymptotic sense.

Simple example. The following example illustrates the sharpness of the convergence estimate (6.44). Consider the bidiagonal matrix B of size $n = 1000$, defined by $q_i = n+1-i, e_i = 0.1$ for $i = 1, \ldots, n-1$, that is,

$$B = \mathrm{bidiag}\begin{pmatrix} \sqrt{1000} & \sqrt{0.1} & & & & \sqrt{0.1} & & \sqrt{0.1} & \\ & \ddots & \ddots & & & & \sqrt{2} & & \sqrt{1} \end{pmatrix}. \tag{6.45}$$

To see the sharpness of the estimate (6.44), we set the window size to $k = 30$, and ran a dqd iteration (without shifts, an experiment with shifts will appear shortly) on B, then computed $\dfrac{\widehat{x}_\ell}{x_\ell} \Big/ \dfrac{\sigma^2_{n-\ell+1}}{\sigma^2_{n-k+1}}$. We observed that[2]

$$1 + 6.79 \times 10^{-3} \le \frac{\widehat{x}_\ell}{x_\ell} \Big/ \frac{\sigma^2_{n-\ell+1}}{\sigma^2_{n-k+1}} \le 1 + 5.996 \times 10^{-2} \tag{6.46}$$

for $\ell = 1, \ldots, k-2$. This suggests that (6.44) is a sharp estimate of the convergence factor $\widehat{x}_\ell/x_\ell$. This behavior is not specific to the particular choice of B, and similar results were obtained generally with any graded diagonally-dominant bidiagonal matrix.

6.3.3. Effect of shifts and motivations for dqd. (6.44) suggests that the introduction of shifts may have little effect on the number of deflatable eigenvalues in dqds with aggressive early deflation. Specifically, when a shift s ($< \sigma^2_n$) is applied to dqds, the convergence factor estimate (6.44) becomes $\dfrac{\sigma^2_{n-\ell+1} - s}{\sigma^2_{n-k+1} - s}$. When a conventional deflation strategy is used, we can regard $k = 2$ and $\ell = 1$, in which case this factor is greatly reduced with an appropriate shift $s \simeq \sigma^2_n$. In fact, the zero-shift strategy dqd results in prohibitively slow

[2]Assuming $S = 1$, about ten singular values were deflated by Aggdef(2), so only $\ell \le 10$ will appear in actual computation. To fully investigate the convergence behavior of x here we kept running Aggdef(2) regardless of whether (6.12) was satisfied.

convergence of the bottom off-diagonal element, so a sophisticated shift strategy is imperative. However, when aggressive early deflation is adopted so that $k > 2$, we see that the estimate (6.44) is close to that of dqd, that is,

$$\frac{\sigma_{n-\ell+1}^2 - s}{\sigma_{n-k+1}^2 - s} \simeq \frac{\sigma_{n-\ell+1}^2}{\sigma_{n-k+1}^2} \tag{6.47}$$

for $\ell = 2, 3, \ldots, k-2$, because during a typical run of dqds we have $\sigma_{n-\ell+1} \gg \sigma_n > s$ for such ℓ. Hence, when dqds is equipped with aggressive early deflation, shifts may not be essential for the performance. This observation motivates the usage of dqd instead of dqds.

Using dqd instead of dqds has many advantages: dqd has a smaller chance of an overflow or underflow [**132**] and smaller computational cost per iteration, not to mention the obvious fact that computing the shifts is unnecessary[3]. Furthermore, because the shifts are prescribed to be zero, dqd can be parallelized by running multiple dqd in a pipelined manner, just as multiple steps of the QR algorithm can be executed in parallel when multiple shifts are chosen in advance [**5, 157, 110**]. We note that it has been difficult to parallelize dqds in such a way, since an effective shift usually cannot be determined until the previous dqds iteration is completed.

Simple example. To justify the above observation, we again use our example matrix (6.45), and run five dqds iterations with the Johnson shift [**85**] to obtain $\widetilde{B}$, and compute the values $\widetilde{x}_\ell$ by running Aggdef(2). Similarly we obtain $\widehat{B}$ by running five dqd iterations and compute $\widehat{x}_\ell$. Figure 6.3.1 shows plots of $x_\ell, \widehat{x}_\ell$ and $\widetilde{x}_\ell$ for $\ell = 1, \ldots, 15$.

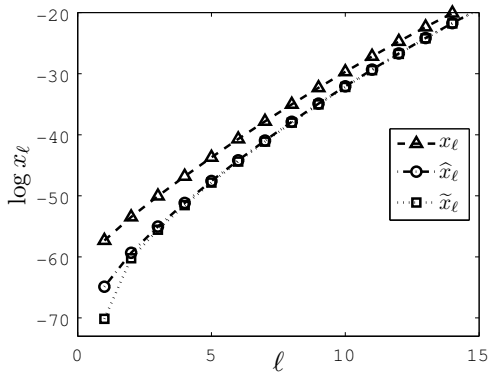

FIGURE 6.3.1. ℓ-log x_ℓ plots for matrix B in (6.45). $\widehat{x}_\ell$ and $\widetilde{x}_\ell$ are obtained from matrices after running 5 dqd and dqds iterations respectively.

We make two remarks on Figure 6.3.1. First, running Aggdef(2) on the original B already lets us deflate about nine singular values (since we need $x_\ell \simeq 10^{-30}$ to satisfy (6.12)). This is because B has a graded and diagonally dominant structure that typically arises in the asymptotic stage of dqds convergence, which is favorable for Aggdef(2). Running dqd (or

[3]Choosing the shift is a delicate task because a shift larger than the smallest singular value results in breakdown of the Cholesky factorization that is implicitly computed by dqds. In DLASQ, whenever a shift violates the positivity, the dqds iteration is rerun with a smaller shift.

dqds) iterations generally reduces the magnitude of x_ℓ, and for $\widehat{B}$ and $\widetilde{B}$ we can deflate one more singular value by Aggdef(2).

Second, the values of $\widehat{x}_\ell$ and $\widetilde{x}_\ell$ are remarkably similar for all ℓ but $\ell = 1$. This reflects our above estimates (6.44) and (6.47), which suggest shifts can help the convergence only of the smallest singular value. Furthermore, as can be seen in Figure 6.3.1, the smallest singular value tends to be already converged in the context of Aggdef(2), so enhancing its convergence is not necessary. Therefore we conclude that shifts may have little or no effect on the overall deflation efficiency of Aggdef(2), suggesting that a zero shift is sufficient and preferred for parallelizability.

6.4. Numerical experiments

This section shows results of numerical experiments to compare the performance of different versions of dqds.

6.4.1. Pseudocodes. Algorithm 9 shows the pseudocode of dqds with aggressive early deflation.

Algorithm 9 dqds with aggressive early deflation

Inputs: Bidiagonal matrix $B \in \mathbb{R}^{n\times n}$, deflation frequency p

1: **while** size of B is larger than $\sqrt{n}$ **do**
2: run p iterations of dqds
3: perform aggressive early deflation
4: **end while**
5: run dqds until all singular values are computed

On the third line, either Aggdef(1) or Aggdef(2) may be invoked. After the matrix size is reduced to smaller than $\sqrt{n}$ we simply use the standard dqds algorithm because the remaining part needs only $O(n)$ flops, the same as one dqds iteration for the original matrix. Along with aggressive early deflation we invoke the conventional deflation strategy after each dqds iteration, just like in the Hessenberg QR case [**16**].

As motivated in the previous section, we shall also examine the performance of the zero-shift version, dqd with aggressive early deflation. Algorithm 10 is its pseudocode.

Algorithm 10 dqd with aggressive early deflation

Inputs: Bidiagonal matrix $B \in \mathbb{R}^{n\times n}$, deflation frequency p

1: **while** size of B is larger than $\sqrt{n}$ **do**
2: run one iteration of dqds, followed by $p-1$ iterations of dqd
3: perform aggressive early deflation
4: **end while**
5: run dqds until all singular values are computed

Note that on line 2 of Algorithm 10, one dqds iteration is run prior to the dqd iterations. This can be done even if we run the p iterations (1 dqds and $p-1$ dqd) parallely, because

this requires only the first shift. The single dqds iteration is important for efficiency because typically a large shift s can be taken after a significant number of singular values are deflated by aggressive early deflation.

6.4.2. Choice of parameters. Two parameters need to be specified when executing Aggdef in Algorithm 9 or 10: the frequency p with which we invoke Aggdef, and the window size k.

We first discuss our choice of the frequency p. It is preferable to set p small enough to take full advantage of aggressive early deflation. This is the case especially for Aggdef(2), because each execution requires only $O(n\ell)$ flops, and experiments indicate that the execution of Aggdef(2) typically takes less than 4% of the overall runtime. In our experiments we let $p = 16$. This choice is based on the fact that when we run dqd iterations in parallel, we need $p \geq n_p$ where n_p is the number of processors run in a pipelined fashion. Experiments suggest that setting p too large (say $p > 300$) may noticeably, but not severely, deteriorate the performance on a sequential implementation.

In practice, when a significant number of singular values are deflated by Aggdef, the performance can often be further improved by performing another aggressive early deflation before starting the next dqds iteration. In our experiments we performed another aggressive early deflation when three or more singular values were deflated by Aggdef. A similar strategy is suggested in [**16**] for the Hessenberg QR algorithm.

We now discuss our choice of the window size k. The idea is to choose k flexibly, using the information of the bottom-right part of B. From the estimate of x_ℓ in (6.42) we see that increasing k reduces the size of x as long as $e_{n-k+1} < q_{n-k+2}$ holds. Hence we compare e_{n-i+1} and q_{n-i+2} for $i = 1, 2, \ldots$, and set k to be the largest i such that $e_{n-i+1} < q_{n-i+2}$. When this results in $k \leq 10$, we skip Aggdef and go on to the next dqds iteration to avoid wasting effort. The choice sometimes makes k too large (e.g., $k = n$ when B is diagonally dominant), so we set a safeguard upper bound $k \leq \sqrt{n}$. In addition, from (6.12) and (6.42) we see that a singular value can be deflated once $\prod_{i=1}^{k-2} \frac{e_{n-i-1}}{q_{n-i}}$ is negligible, so we compute the products $\prod_{i=11}^{k-2} \frac{e_{n-i-1}}{q_{n-i}}$ (we start taking the product from $i = 11$ because we want to deflate more than one singular value; in view of (6.43), with $i = 11$ we can expect $\simeq 10$ deflations to occur) and decide to stop increasing k once the product becomes smaller than ϵ^2.

Through experiments we observed that the above choice of p and k is effective, achieving speedups of on average about 25% for matrices $n \geq 10000$, compared with any static choice such as $p = k = \sqrt{n}$.

6.4.3. Experiment details. We compare the performance of the following four algorithms[4].

(1) DLASQ: dqds subroutine of LAPACK version 3.2.2.
(2) dqds+agg1: Algorithm 9, call Aggdef(1) on line 3.
(3) dqds+agg2: Algorithm 9, call Aggdef(2) on line 3.
(4) dqd+agg2: Algorithm 10, call Aggdef(2) on line 3.

[4]dqd with Aggdef(1) can be implemented, but we do not present its results because dqd+agg2 was faster in all our experiments.

We implemented our algorithms in Fortran by incorporating our deflation strategies into the LAPACK routines dlasqx.f (x ranges from 1 to 6). Hence, our codes perform the same shift and splitting strategies implemented in DLASQ[5]. When running dqds+agg1, we used the LAPACK subroutine DBDSQR to compute the singular values of B_2 and the spike vector t in (6.2). All experiments were conducted on a single core of a desktop machine with a quad core, Intel Core i7 2.67GHz Processor and 12GB of main memory. For compilation we used the f95 compiler and the optimization flag -O3, and linked the codes to BLAS and LAPACK.

TABLE 6.4.1. Test bidiagonal matrices

	n	Description of the bidiagonal matrix B	Source
1	30000	$\sqrt{q_i} = n + 1 - i,\ \sqrt{e_i} = 1$	
2	30000	$\sqrt{q_{i-1}} = \beta\sqrt{q_i},\ \sqrt{e_i} = \sqrt{q_i},\ \beta = 1.01$	[**45**]
3	30000	Toeplitz: $\sqrt{q_i} = 1,\ \sqrt{e_i} = 2$	[**45**]
4	30000	$\sqrt{q_{2i-1}} = n + 1 - i,\ \sqrt{q_{2i}} = i,\ \sqrt{e_i} = (n-i)/5$	[**129**]
5	30000	$\sqrt{q_{i+1}} = \beta\sqrt{q_i}\ (i \geq n/2),\ \sqrt{q_{n/2}} = 1,$ $\sqrt{q_{i-1}} = \beta\sqrt{q_i}\ (i \leq n/2),\ \sqrt{e_i} = 1,\ \beta = 1.01$	
6	30000	Cholesky factor of tridiagonal $(1, 2, 1)$ matrix	[**107, 132**]
7	30000	Cholesky factor of Laguerre matrix	[**107**]
8	30000	Cholesky factor of Hermite recurrence matrix	[**107**]
9	30000	Cholesky factor of Wilkinson matrix	[**107**]
10	30000	Cholesky factor of Clement matrix	[**107**]
11	13786	matrix from electronic structure calculations	[**133**]
12	16023	matrix from electronic structure calculations	[**133**]

Table 6.4.1 gives a brief description of our test matrices. Matrix 1 is "nearly diagonal", for which aggressive early deflation is particularly effective. Matrix 2 is a "nicely graded" matrix [**45**], for which DLASQ needs relatively few ($\ll 4n$) iterations. Matrix 3 is a Toeplitz matrix [**45**], which has uniform diagonals and off-diagonals. Matrix 4 has highly oscillatory diagonal entires. Matrix 5 is a perversely graded matrix, designed to be difficult for DLASQ. Matrices 6-10 are the Cholesky factors of test tridiagonal matrices taken from [**107**]. For matrices 8-10, we applied an appropriate shift to make the tridiagonal matrix positive definite before computing the Cholesky factor. Matrices 11 and 12 have the property that some of the singular values are tightly clustered.

6.4.4. Results. Results are shown in Figures 6.4.1-6.4.4, which compare the total runtime, number of dqds iterations, percentage of singular values deflated via aggressive early deflation, and the percentage of the time spent performing aggressive early deflation relative to the overall runtime. We executed ten runs and took the average. The numbers in parentheses show the performance of DLASQ for each matrix: the runtime in seconds in Figure 6.4.1 and the iteration counts divided by the matrix size n in Figure 6.4.2. Although not shown in the figures, in all our experiments we confirmed the singular values are computed

[5]It is possible that when equipped with aggressive early deflation, a different shift strategy for dqds is more efficient than that used in DLASQ. This is a possible topic of future study.

to high relative accuracy. Specifically, the maximum element-wise relative difference of the singular values computed by our algorithms from those computed by DLASQ was smaller than both 1.5×10^{-13} and $n\epsilon$ for each problem.

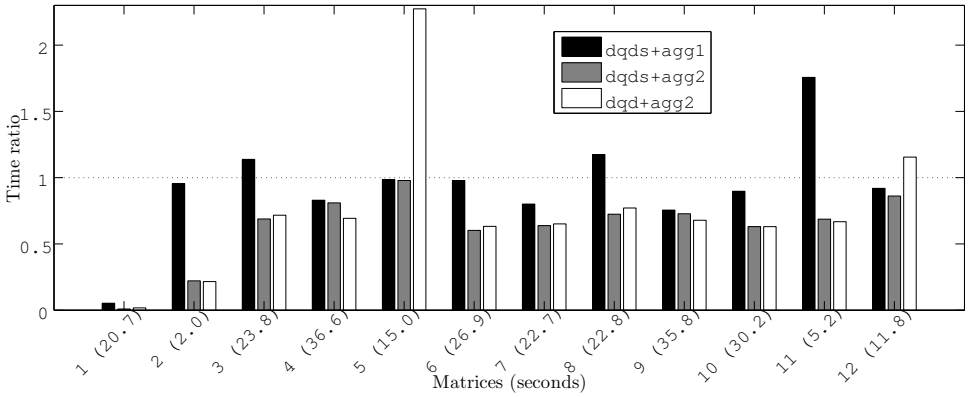

FIGURE 6.4.1. Ratio of time/DLASQ time.

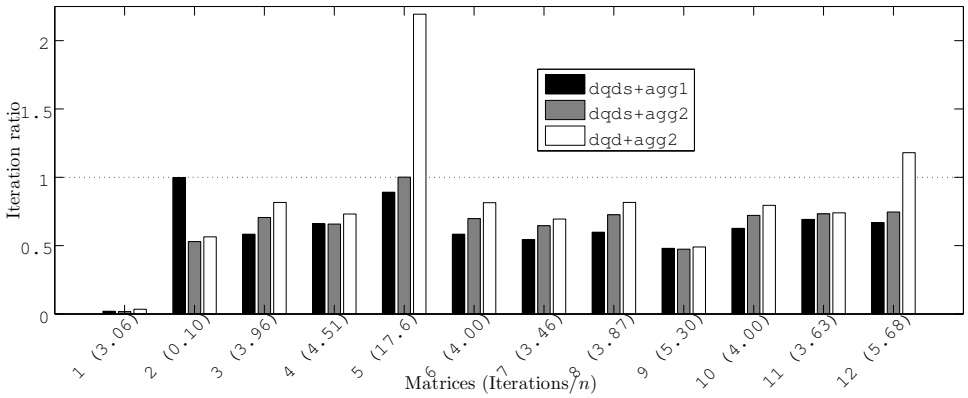

FIGURE 6.4.2. Ratio of iteration/DLASQ iteration.

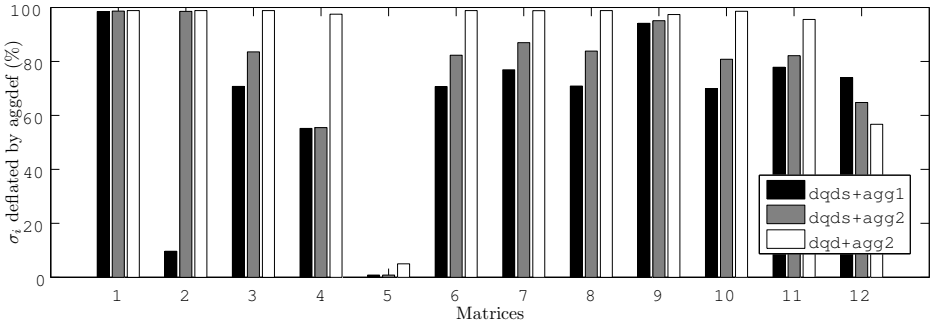

FIGURE 6.4.3. Percentage of singular values deflated by aggressive early deflation.

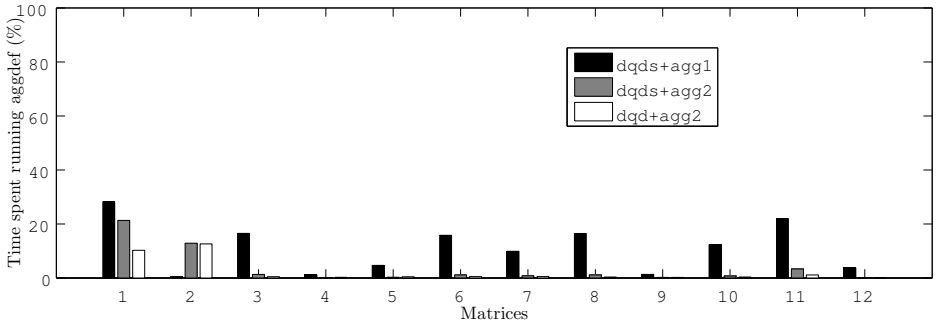

FIGURE 6.4.4. Percentage of time spent executing aggressive early deflation.

The results show that aggressive early deflation, particularly Aggdef(2), can significantly reduce both the runtime and iteration count of DLASQ. We obtained speedups of up to 50 with dqds+agg2 and dqd+agg2.

dqds+agg2 was notably faster than DLASQ in most cases, and never slower. There was no performance gain for the "difficult" Matrix 5, for which many dqds iterations are needed before the iteration reaches the asymptotic stage where the matrix is graded and diagonally dominant, after which Aggdef(2) becomes effective. dqds+agg2 was also at least as fast as dqds+agg1 in all our experiments. This is because as discussed in Section 6.2.1, Aggdef(1) requires at least $O(k^2)$ flops, while Aggdef(2) needs only $O(k\ell)$ flops when it deflates $\ell \leq k$ singular values.

We see from Figures 6.4.2 and 6.4.3 that dqds+agg1 and dqds+agg2 usually require about the same number of iterations and deflate similar numbers of singular values by Aggdef. The exception in Matrix 2 is due to the fact that the safe window size enforced in Aggdef(1) (described in Section 6.1.1) is often much smaller than k (determined as in Section 6.4.2), making Aggdef(1) less efficient. For dqd+agg2, usually most of the singular values are deflated by Aggdef(2). This is because with zero-shifts the bottom off-diagonal converges much slower, making the conventional deflation strategy ineffective.

Finally, in many cases dqd+agg2 was the fastest algorithm requiring comparable numbers of iterations to dqds+agg2, except for problems that are difficult (iteration $\geq 5n$) for DLASQ. As we mentioned earlier, this is in major contrast to dqd with a conventional deflation strategy, which is impractical due to the slow convergence of each off-diagonal element. Furthermore, as can be seen in Figure 6.4.4, with Aggdef(2) the time spent executing aggressive early deflation is typically less than 4%[6]. This observation makes the parallel implementation of dqd+agg2 particularly promising since it is already the fastest of the tested algorithms in many cases, and its parallel implementation is expected to speed up the dqd runs, which are essentially taking up more than 95% of the time.

We also tested with more than 500 other bidiagonal matrices, including the 405 bidiagonal matrices from the tester in the development of DLASQ [**106**], 9 test matrices from [**132**],

[6]Exceptions are in "easy" cases, such as matrices 1 and 2, where dqd+agg2 requires much fewer iterations than $4n$. In such cases dqd+agg2 spends relatively more time executing Aggdef(2) recursively. This is by no means a pathological case, because dqd+agg2 is already very fast with a sequential implementation.

and 6 matrices that arise in electronic structure calculations [**133**]. We show in Figure 6.4.5 a scatter plot of the runtime ratio over DLASQ against the matrix size for dqds+agg2 and dqd+agg2. To keep the plot simple we do not show dqds+agg1, which was never faster than dqds+agg2. Also, to ensure that the computed time is measured reliably, we show the results only for 285 matrices for which DLASQ needed more than 0.01 second.

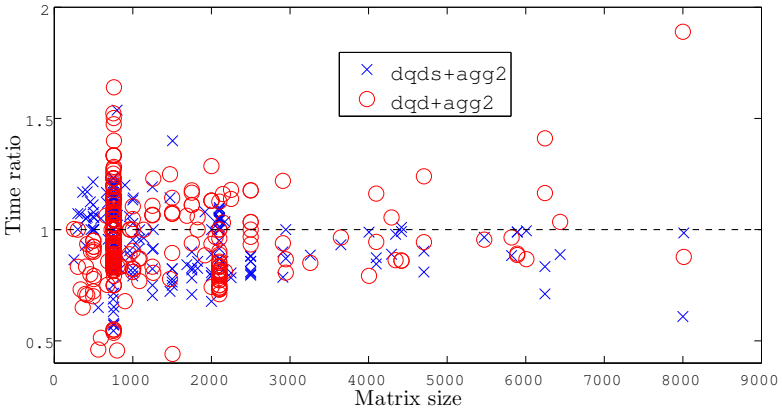

FIGURE 6.4.5. Ratio of time/DLASQ time for test matrices.

We note that many of these matrices represent "difficult" cases (DLASQ needs more than $5n$ iterations), as they were generated for checking the algorithm robustness. In such cases, many dqd(s) iterations are needed for the matrix to reach the asymptotic graded structure, during which using Aggdef(2) may not be of much help. Nonetheless, dqds+agg2 was always at least as fast as DLASQ for all matrices larger than 3000. Moreover, dqds+agg2 was never slower than DLASQ by more than 0.016 second, so we argue that in practice it is never slower. The speed of dqd+agg2 varied more depending on the matrices, taking up to 0.15 second more than or 1.9 times as much time as DLASQ.

Summary and future work. We proposed two algorithms dqds+agg1 and dqds+agg2 to incorporate aggressive early deflation into dqds for computing the singular values of bidiagonal matrices to high relative accuracy. We presented numerical results to demonstrate that aggressive early deflation can significantly speed up dqds. In particular, dqds+agg2 is at least as fast as the LAPACK implementation of dqds, and is often much faster. The zero-shifting strategy exhibits even more promising results with the potential to be parallelized. We plan to report the implementation and performance of a parallel version of dqd+agg2 in a future work.

Part 2

Eigenvalue perturbation theory

CHAPTER 7

Eigenvalue perturbation bounds for Hermitian block tridiagonal matrices

The first chapter on eigenvalue perturbation theory concerns changes in eigenvalues when a structured Hermitian matrix undergoes a perturbation. In particular, we consider block tridiagonal matrices and derive new eigenvalue perturbation bounds, which can be arbitrarily tighter than the generic Weyl bound. The main message of this chapter is that an eigenvalue is insensitive to blockwise perturbation, if it is well-separated from the spectrum of the diagonal blocks nearby the perturbed blocks. Our bound is particularly effective when the matrix is block-diagonally dominant and graded. Our approach is to obtain eigenvalue bounds via bounding eigenvector components, which is based on the observation that an eigenvalue is insensitive to componentwise perturbation if the corresponding eigenvector components are small. We use the same idea to explain two well-known phenomena, one concerning aggressive early deflation used in the symmetric tridiagonal QR algorithm and the other concerning the extremal eigenvalues of Wilkinson matrices.

Introduction. As we reviewed in Chapter 2, Eigenvalue perturbation theory for Hermitian matrices is a well-studied subject with many known results, see for example [**146**, Ch.4] [**56**, Ch.8], [**31**, Ch.4]. Among them, Weyl's theorem is perhaps the simplest and most well-known, which states that the eigenvalues of the Hermitian matrices A and $A + E$ differ at most by $\|E\|_2$. In fact, when the perturbation E is allowed to be an arbitrary Hermitian matrix, Weyl's theorem gives the smallest possible bound that is attainable.

Hermitian matrices that arise in practice frequently have special sparse structures, important examples of which being banded and block tridiagonal structures. For such structured matrices, perturbation of eigenvalues is often much smaller than any known bound guarantees. The goal of this chapter is to treat block tridiagonal Hermitian matrices and derive eigenvalue perturbation bounds that can be much sharper than known general bounds, such as Weyl's theorem.

The key observation of this chapter, to be made in Section 7.1, is that an eigenvalue is insensitive to componentwise perturbations if the corresponding eigenvector components are small. Our approach is to obtain bounds for eigenvector components, from which we obtain eigenvalue bounds. In this framework we first give new eigenvalue perturbation bounds for the simplest, 2-by-2 block case. In particular, we identify a situation in which the perturbation bound of an eigenvalue scales cubically with the norm of the perturbation.

We then discuss the general block tridiagonal case, in which we show that an eigenvalue is insensitive to blockwise perturbation, if it is well-separated from the spectrum of the diagonal blocks nearby the perturbed blocks.

Finally, to demonstrate the effectiveness of our approach, we show that our framework successfully explains the following two well-known phenomena: (i) Aggressive early deflation applied to the symmetric tridiagonal QR algorithm deflates many eigenvalues even when no off-diagonal element is negligibly small. (ii) Wilkinson matrices have many pairs of nearly equal eigenvalues.

A number of studies exist in the literature regarding bounds for eigenvectors, especially in the tridiagonal case, for which explicit formulas exist for the eigenvector components using the determinants of submatrices [**127**, Sec. 7.9]. In [**27**] Cuppen gives an explanation for the exponential decay in eigenvector components of tridiagonal matrices, which often lets the divide-and-conquer algorithm run much faster than its estimated cost suggests. The derivation of our eigenvector bounds are much in the same vein as Cuppen's argument. A difference here is that we use the bounds to show the insensitivity of eigenvalues. In [**126**] Parlett investigates the localization behavior of eigenvectors (or an invariant subspace) corresponding to a cluster of m eigenvalues, and notes that accurate eigenvalues and nearly orthogonal eigenvectors can be computed from appropriately chosen m submatrices, which allow overlaps within one another. One implication of this is that setting certain subdiagonals to zero has negligible influence on some of the eigenvalues. A similar claim is made by our Theorem 7.2, which holds for block tridiagonal matrices and whether or not the eigenvalue belongs to a cluster. Finally, in a recent paper [**128**] Parlett considers symmetric banded matrices and links the disjointness of an eigenvalue from Gerschgorin disks with bounds of off-diagonal parts of L and U in the LDU decomposition, from which exponential decay in eigenvector components can be deduced. A similar message is conveyed by our Lemma 7.3, but unlike in [**128**] we give direct bounds for the eigenvector components, and we use them to obtain eigenvalue bounds. Moreover, Parlett and Vömel [**125**] study detecting such eigenvector decay behavior to devise a process to efficiently compute some of the eigenvalues of a symmetric tridiagonal matrix. Our results here may be used to foster such developments. Finally, [**82**] considers general Hermitian matrices and shows for any eigenpair (λ_i, x) of A that the interval $[\lambda_i - \|Ex\|_2, \lambda_i + \|Ex\|]$ contains an eigenvalue of $A+E$, and gives a condition under which the interval must contain ith eigenvalue of $A + E$.

The rest of this chapter is organized as follows. In Section 7.1 we outline our basic idea of deriving eigenvalue perturbation bounds via bounding eigenvector components. Section 7.2 treats the 2-by-2 block case and presents a new bound. Section 7.3 discusses the block tridiagonal case. In Section 7.4 we investigate the above two case studies.

Notations: $\lambda_i(X)$ denotes the ith smallest eigenvalue of a Hermitian matrix X. For simplicity we use λ_i, $\widehat{\lambda}_i$ and $\lambda_i(t)$ to denote the ith smallest eigenvalue of $A, A + E$ and $A + tE$ for $t \in [0, 1]$ respectively. $\lambda(A)$ denotes A's spectrum, the set of eigenvalues. We only use the matrix spectral norm $\|\cdot\|_2$.

7.1. Basic approach

We first recall the partial derivative of a simple eigenvalue [**146**].

LEMMA 7.1. *Let A and E be n-by-n Hermitian matrices. Denote by $\lambda_i(t)$ the ith eigenvalue of $A+tE$, and define the vector-valued function $x(t)$ such that $(A+tE)x(t) = \lambda_i(t)x(t)$ where $\|x(t)\|_2 = 1$ for some $t \in [0,1]$. If $\lambda_i(t)$ is simple, then*

$$\frac{\partial \lambda_i(t)}{\partial t} = x(t)^* E x(t). \tag{7.1}$$

Our main observation here is that if $x(t)$ has small components in the positions corresponding to the dominant elements of E, then $\frac{\partial \lambda_i(t)}{\partial t}$ is small. For example, suppose that E is nonzero only in the (j,j)th element. Then we have $\left|\frac{\partial \lambda_i(t)}{\partial t}\right| \le \|E\|_2 |x_j(t)|^2$, where $x_j(t)$ is jth element of $x(t)$. Hence if we know a bound for $|x_j(t)|$ for all $t \in [0,1]$, then we can integrate (7.1) over $0 \le t \le 1$ to obtain a bound for $|\lambda_i - \widehat{\lambda}_i| = |\lambda_i(0) - \lambda_i(1)|$. In the sequel we shall describe in detail how this observation can be exploited to derive eigenvalue perturbation bounds.

It is important to note that Lemma 7.1 assumes that λ_i is a simple eigenvalue of A. Special treatment is needed to get the derivative of multiple eigenvalues. This is shown in Section 7.5 at the end of this chapter, in which we show that everything we discuss below carries over even in the presence of multiple eigenvalues. In particular, when $\lambda_i(t)$ is multiple (7.1) still holds for a certain choice of eigenvector $x(t)$ of $\lambda_i(t)$. We defer the treatment of multiple eigenvalues until the final section of the chapter, because it only causes complications to the analysis that are not fundamental to the eigenvalue behavior. Hence for simplicity until 7.5 we assume that $\lambda_i(t)$ is simple for all t, so that the normalized eigenvector is unique up to a factor $e^{\mathrm{i}\theta}$.

7.2. 2-by-2 block case

In this section we consider the 2-by-2 case. Specifically, we study the difference between eigenvalues of the Hermitian matrices A and $A+E$, where

$$A = \begin{bmatrix} A_{11} & A_{21}^* \\ A_{21} & A_{22} \end{bmatrix} \quad \text{and} \quad E = \begin{bmatrix} E_{11} & E_{21}^* \\ E_{21} & E_{22} \end{bmatrix}, \tag{7.2}$$

in which A_{jj} and E_{jj} are square and of the same size for $j = \{1,2\}$.

Since $\lambda_i(0) = \lambda_i$ and $\lambda_i(1) = \widehat{\lambda}_i$, from (7.1) it follows that

$$|\lambda_i - \widehat{\lambda}_i| = \left| \int_0^1 x(t)^* E x(t) dt \right| \tag{7.3}$$

$$\le \left| \int_0^1 x_1(t)^* E_{11} x_1(t) dt \right| + 2 \left| \int_0^1 x_2(t)^* E_{21} x_1(t) dt \right| + \left| \int_0^1 x_2(t)^* E_{22} x_2(t) dt \right|, \tag{7.4}$$

where we block-partitioned $x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$ so that $x_1(t)$ and A_{11} have the same number of rows. The key observation here is that the latter two terms in (7.4) are small if $\|x_2(t)\|_2$ is small for all $t \in [0,1]$. We obtain an upper bound for $\|x_2(t)\|_2$ by the next lemma.

LEMMA 7.2. *Suppose that $\lambda_i \notin \lambda(A_{22})$ is the ith smallest eigenvalue of A as defined in (7.2). Let $Ax = \lambda_i x$ such that $\|x\|_2 = 1$. Then, partitioning $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ as above we have*

$$\|x_2\|_2 \leq \frac{\|A_{21}\|_2}{\min|\lambda_i - \lambda(A_{22})|}. \tag{7.5}$$

PROOF. The bottom k rows of $Ax = \lambda_i x$ is

$$A_{21}x_1 + A_{22}x_2 = \lambda_i x_2,$$

so we have

$$x_2 = (\lambda_i I - A_{22})^{-1} A_{21} x_1.$$

Taking norms we get

$$\|x_2\|_2 \leq \|(\lambda_i I - A_{22})^{-1}\|_2 \|A_{21}\|_2 \|x_1\|_2 \leq \frac{\|A_{21}\|_2}{\min|\lambda_i - \lambda(A_{22})|}, \tag{7.6}$$

where we used $\|x_1\|_2 \leq \|x\|_2 = 1$ to get the last inequality. □

We note that Lemma 7.2 is just a special case of the Davis-Kahan generalized $\sin\theta$ theorem [**29**, Thm. 6.1], in which the two subspaces have dimensions 1 and $n-k$. Specifically, 7.2 bounds the sin of the angle between an eigenvector x and the first $n-k$ columns of the identity matrix I. Moreover, by the first inequality of (7.6), we have $\frac{\|x_2\|_2}{\|x_1\|_2} \leq \|(\lambda_i I - A_{22})^{-1}\|_2 \|A_{21}\|_2$. This can be regarded essentially as the Davis-Kahan generalized $\tan\theta$ theorem [**29**, Thm. 6.3] for the 1-dimensional case, which gives $\tan\theta \leq \frac{\|A_{21}\|_2}{gap}$, where in the 1-dimensional case $\tan\theta = \frac{\|x_2\|_2}{\|x_1\|_2}$. *gap* is the distance between λ_i and $\lambda(A_{22})$ such that $\lambda(A_{22}) \in [\lambda_i + gap, \infty)$ or $\lambda(A_{22}) \in (-\infty, \lambda_i - gap]$. Note that Lemma 7.2 requires less assumption on the situation between λ_i and $\lambda(A_{22})$, because (7.5) holds when $\lambda(A_{22})$ (the set of Ritz values) lies both below and above λ_i. This is precisely what we prove in Chapter 10.

Clearly, if $\lambda_i \notin \lambda(A_{11})$ then (7.5) holds with x_2 replaced with x_1 and A_{22} replaced with A_{11}. This applies to this entire section, but for definiteness we only present results assuming $\lambda_i \notin \lambda(A_{22})$.

We note that (7.5) is valid for any λ_i and its normalized eigenvector x, whether or not λ_i is a multiple eigenvalue. It follows that in the multiple case, all the vectors that span the corresponding eigenspace satisfy (7.5).

We now derive refined eigenvalue perturbation bounds by combining Lemmas 7.1 and 7.2. As before, let $(\lambda_i(t), x(t))$ be the ith smallest eigenpair such that $(A + tE)x(t) = \lambda_i(t)x(t)$ with $\|x(t)\|_2 = 1$. When $\min|\lambda_i - \lambda(A_{22})| > 2\|E\|_2$, using (7.5) we get an upper bound for

$\|x_2(t)\|_2$ for all $t \in [0,1]$:

$$\begin{aligned}\|x_2(t)\|_2 &\le \frac{\|A_{21}+tE_{21}\|_2}{\min|\lambda_i(t)-\lambda(A_{22}+tE_{22})|}\\ &\le \frac{\|A_{21}\|_2+t\|E_{21}\|_2}{\min|\lambda_i(0)-\lambda(A_{22})|-2t\|E\|_2} \quad \text{(by Weyl's theorem)}\\ &\le \frac{\|A_{21}\|_2+\|E_{21}\|_2}{\min|\lambda_i-\lambda(A_{22})|-2\|E\|_2}. \qquad (7.7)\end{aligned}$$

We now present a perturbation bound for λ_i.

THEOREM 7.1. *Let λ_i and $\widehat{\lambda}_i$ be the ith eigenvalue of A and $A+E$ as in* (7.2) *respectively, and define* $\tau_i = \dfrac{\|A_{21}\|_2+\|E_{21}\|_2}{\min|\lambda_i-\lambda(A_{22})|-2\|E\|_2}$. *Then for each i, if $\tau_i > 0$ then*

$$\left|\lambda_i-\widehat{\lambda}_i\right| \le \|E_{11}\|_2 + 2\|E_{21}\|_2\tau_i + \|E_{22}\|_2\tau_i^2. \qquad (7.8)$$

PROOF. Substituting (7.7) into (7.4) we get

$$\begin{aligned}\left|\lambda_i-\widehat{\lambda}_i\right| &\le \left|\int_0^1 \|E_{11}\|_2\|x_1(t)\|_2^2 dt\right| + 2\left|\int_0^1 \|E_{21}\|_2\|x_1(t)\|_2\|x_2(t)\|_2 dt\right| + \left|\int_0^1 \|E_{22}\|_2\|x_2(t)\|_2^2 dt\right|\\ &\le \|E_{11}\|_2 + 2\|E_{21}\|_2\tau_i + \|E_{22}\|_2\tau_i^2,\end{aligned}$$

which is (7.8). □

We make three points on Theorem 7.1.

- $\tau_i < 1$ is a necessary condition for (7.8) to be tighter than the Weyl bound $\|E\|_2$. $\min|\lambda_i-\lambda(A_{22})| > 2\|E\|_2+\|A_{21}\|_2+\|E_{21}\|_2$. If λ_i is far from the spectrum of A_{22} so that $\tau_i \ll 1$ and $\|E_{11}\|_2 \ll \|E\|_2$, then (7.8) is much smaller than $\|E\|_2$.
- When A_{21}, E_{11}, E_{22} are all zero (i.e., when a block-diagonal matrix undergoes an off-diagonal perturbation), (7.8) becomes

$$\left|\lambda_i-\widehat{\lambda}_i\right| \le \frac{2\|E_{21}\|_2^2}{\min|\lambda_i-\lambda(A_{22})|-2\|E_{21}\|_2}, \qquad (7.9)$$

  which shows the perturbation must be $O(\|E_{21}\|^2)$ if λ_i is not an eigenvalue of A_{22}.

  We note that much work has been done for such structured perturbation. For example, under the same assumption of off-diagonal perturbation, [**109, 97**] prove the quadratic residual bounds

$$\begin{aligned}|\lambda_i-\widehat{\lambda}_i| &\le \frac{\|E_{21}\|_2^2}{\min|\lambda_i(A)-\lambda(A_{22})|}\\ &\le \frac{2\|E_{21}\|_2^2}{\min|\lambda_i(A)-\lambda(A_{22})|+\sqrt{\min|\lambda_i(A)-\lambda(A_{22})|^2+4\|E_{21}\|_2^2}}. \qquad (7.10)\end{aligned}$$

  Our bound (7.8) (or (7.9)) is not as tight as the bounds in (7.10). However, (7.8) has the advantage that it is applicable for a general perturbation, not necessarily off-diagonal.

- (7.8) also reveals that if $E = \begin{bmatrix} 0 & 0 \\ 0 & E_{22} \end{bmatrix}$ and A_{21} is small, then λ_i is particularly insensitive to the perturbation E_{22}: the bound (7.8) becomes proportional to $\|E_{22}\|_2 \|A_{21}\|_2^2$.

  For example, consider the n-by-n matrices $\begin{bmatrix} A_{11} & v \\ v^* & \varepsilon \end{bmatrix}$ and $\begin{bmatrix} A_{11} & v \\ v^* & 0 \end{bmatrix}$ where A_{11} is nonsingular. These matrices have one eigenvalue that matches up to ε, and $n-1$ eigenvalues that match up to $O(\varepsilon\|v\|_2^2)$. Note that when $\|v\|_2 = O(\varepsilon)$, $\varepsilon\|v\|_2^2$ scales *cubically* with ε.

7.3. Block tridiagonal case

Here we consider the block tridiagonal case and apply the idea we used above to obtain a refined eigenvalue perturbation bound. Let A and E be Hermitian block tridiagonal matrices defined by

$$A = \begin{bmatrix} A_1 & B_1^* & & \\ B_1 & \ddots & \ddots & \\ & \ddots & \ddots & B_{n-1}^* \\ & & B_{n-1} & A_n \end{bmatrix} \quad \text{and} \quad E = \begin{bmatrix} \ddots & \ddots & & & \\ \ddots & 0 & 0 & & \\ & 0 & \Delta A_s & \Delta B_s^* & \\ & & \Delta B_s & 0 & 0 \\ & & & 0 & \ddots \end{bmatrix}, \tag{7.11}$$

where $A_j, \in \mathbb{C}^{n_j \times n_j}$, $B_j \in \mathbb{C}^{n_j \times n_{j+1}}$. The size of ΔA_s and ΔB_s match those of A_s and B_s. Here we consider perturbation in a single block, so E is zero except for the sth blocks ΔA_s and ΔB_s. When more than one block is perturbed we can apply the below argument repeatedly.

We obtain an upper bound for $|\lambda_i - \widehat{\lambda}_i|$ by bounding the magnitude of the eigenvector components corresponding to the sth and $(s+1)$th blocks. As before we let $(\lambda_i(t), x(t))$ be the ith eigenpair such that $(A + tE)x(t) = \lambda_i(t)x(t)$ for $t \in [0,1]$. To prove a useful upper bound for the blocks of the eigenvector $x(t)$ corresponding to $\lambda_i(t)$ for all $t \in [0,1]$, we make the following Assumption 1. Here we say "a belongs to the jth block of A" if

$$a \in [\lambda_{\min}(A_j) - \eta_j, \lambda_{\max}(A_j) + \eta_j] \quad \text{where} \quad \eta_j = \|B_j\|_2 + \|B_{j-1}\|_2 + \|E\|_2, \tag{7.12}$$

in which for convenience we define $B_0 = 0, B_n = 0$. Note that a can belong to more than one block.

Assumption 1. *There exists an integer $\ell > 0$ such that λ_i does not belong to the first $s + \ell$ blocks of A.*

Roughly, the assumption demands that λ_i be far away from the eigenvalues of $A_1, \ldots, A_{s+\ell}$, and that the norms of E and $B_1, \ldots, B_{s+\ell}$ are not too large. A typical case where the assumption holds is when $A_1, A_2, \ldots, A_n$ have a graded structure, so that the eigenvalues of A_i are smaller (or larger) than those of A_j for all (i,j) with $i < j$. For example, consider

the tridiagonal matrix

$$A + E = \text{tridiag} \left\{ \begin{array}{ccccccccc} & 1 & & 1 & . & 1 & & 1 & \\ 1000 & & 999 & & . & & 2 & & 1 \\ & 1 & & 1 & . & 1 & & 1 & \end{array} \right\}, \tag{7.13}$$

where E is zero except for the first off-diagonals, which are 1. This means we have $s = 1$. In this case we set the block size to one, and see that the interval as in (7.12) for the jth block is $[\lambda_{\min}(A_j) - \eta_j, \lambda_{\max}(A_j) + \eta_j] \subseteq [(1001 - j) - 3, (1001 - j) + 3]$ for all j, in which for $2 \leq j \leq 999$ the intervals are equal. Hence for any eigenvalue $\lambda_i \leq 100$, λ_i does not belong to the first 897 blocks, so Assumption 1 is valid with $\ell \geq 896$. More generally, for any $\lambda_i \in [k, k+1]$ for any integer $k < 997$, Assumption 1 holds with $\ell = 997 - k$.

We also note that the below argument hold exactly the same for the case where λ_i does not belong to the last $n - s + 1 + \ell$ blocks, but for definiteness we proceed under Assumption 1.

We now derive an upper bound for the eigenvector components.

LEMMA 7.3. *Let A and E be Hermitian block tridiagonal matrices as in* (7.11)*. Suppose that A's ith eigenvalue λ_i satisfies Assumption 1, so that defining $gap_j = \min |\lambda_i - \lambda(A_j)|$, we have $gap_j > \|E\|_2 + \|B_j\|_2 + \|B_{j-1}\|_2$ for $j = 1, \ldots, s + \ell$. Let $(A + tE)x(t) = \lambda_i(t)x(t)$ and $\|x(t)\|_2 = 1$ for $t \in [0, 1]$, and block-partition $x(t)^* = [x_1(t)^*\ x_2(t)^*\ \ldots\ x_n(t)^*]^*$ such that $x_j(t)$ and A_j have the same numbers of rows. Let*[1]

$$\delta_0 = \frac{\|B_s\|_2 + \|\Delta B_s\|_2}{gap_s - \|E\|_2 - \|\Delta A_S\|_2 - \|B_{s-1}\|_2}, \quad \delta_1 = \frac{\|B_{s+1}\|_2}{gap_{s+1} - \|E\|_2 - \|B_s\|_2 - \|\Delta B_s\|_2}, \tag{7.14}$$

and

$$\delta_j = \frac{\|B_{s+j}\|_2}{gap_{s+j} - \|E\|_2 - \|B_{s+j-1}\|_2} \quad \text{for} \quad j = 2, \ldots, \ell, \tag{7.15}$$

and suppose that the denominators in (7.14) *are both positive. Then, for all $t \in [0, 1]$ we have*

$$\|x_s(t)\|_2 \leq \prod_{j=0}^{\ell} \delta_j, \tag{7.16}$$

$$\|x_{s+1}(t)\|_2 \leq \prod_{j=1}^{\ell} \delta_j. \tag{7.17}$$

PROOF. We first show that $\|x_1(t)\|_2 \leq \|x_2(t)\|_2 \leq \cdots \leq \|x_{s+\ell}(t)\|_2$. The first block of $(A + tE)x(t) = \lambda_i(t)x(t)$ is

$$A_1 x_1(t) + B_1^* x_2(t) = \lambda_i(t) x_1(t),$$

so we have

$$x_1(t) = (\lambda_i(t) I - A_1)^{-1} B_1^* x_2(t).$$

[1] δ_j and gap_s depend also on i, but we omit the subscript for simplicity.

Now since by Weyl's theorem we have $\lambda_i(t) \in [\lambda_i - \|E\|_2, \lambda_i + \|E\|_2]$ for all $t \in [0,1]$, it follows that $\|(\lambda_i(t)I - A_1)^{-1}\|_2 \leq 1/(gap_1 - \|E\|_2)$. Therefore, $\|x_1(t)\|_2/\|x_2(t)\|_2$ can be bounded by

$$\frac{\|x_1(t)\|_2}{\|x_2(t)\|_2} \leq \frac{\|B_1\|_2}{gap_1 - \|E\|_2} \leq 1,$$

where the last inequality follows from Assumption 1.

Next, the second block of $(A + tE)x(t) = \lambda_i(t)x(t)$ is

$$B_1 x_1(t) + A_2 x_2(t) + B_2^* x_3(t) = \lambda_i(t) x_2(t),$$

so we have

$$x_2(t) = (\lambda_i(t)I - A_2)^{-1}(B_1 x_1(t) + B_2^* x_3(t)).$$

Using $\|(\lambda_i(t)I - A_2)^{-1}\|_2 \leq 1/(gap_2 - \|E\|_2)$ we get

$$\begin{aligned}\|x_2(t)\|_2 &\leq \frac{\|B_1\|_2\|x_1(t)\|_2 + \|B_2\|_2\|x_3(t)\|_2}{gap_2 - \|E\|_2} \\ &\leq \frac{\|B_1\|_2\|x_2(t)\|_2 + \|B_2\|_2\|x_3(t)\|_2}{gap_2 - \|E\|_2}, \quad \text{(because } \|x_1(t)\|_2 \leq \|x_2(t)\|_2)\end{aligned}$$

and so

$$\frac{\|x_2(t)\|_2}{\|x_3(t)\|_2} \leq \frac{\|B_2\|_2}{gap_2 - \|E\|_2 - \|B_1\|_2}.$$

By Assumption 1 this is no larger than 1, so $\|x_2(t)\|_2 \leq \|x_3(t)\|_2$.

By the same argument we can prove $\|x_1(t)\|_2 \leq \|x_2(t)\|_2 \leq \cdots \leq \|x_{s+\ell}(t)\|_2$ for all $t \in [0,1]$.

Next consider the sth block of $(A + tE)x(t) = \lambda_i(t)x(t)$, which is

$$B_{s-1}x_{s-1}(t) + (A_s + t\Delta A_s)x_s(t) + (B_s + t\Delta B_s)^H x_{s+1}(t) = \lambda_i(t)x_s(t),$$

so we have

$$x_s(t) = (\lambda_i(t)I - A_s - t\Delta A_s)^{-1}\left(B_{s-1}x_{s-1}(t) + (B_s + t\Delta B_s)^H x_{s+1}(t)\right).$$

Using $\|(\lambda_i(t)I - A_s - t\Delta A_s)^{-1}\|_2 \leq 1/(gap_s - \|E\|_2 - \|\Delta A_S\|_2)$ and $\|x_{s-1}(t)\|_2 \leq \|x_s(t)\|_2$ we get

$$\|x_s(t)\|_2 \leq \frac{\|B_{s-1}\|_2\|x_s(t)\|_2 + \|B_s + t\Delta B_s\|_2\|x_{s+1}(t)\|_2}{gap_s - \|E\|_2 - \|\Delta A_S\|_2}.$$

Hence we get $\dfrac{\|x_s(t)\|_2}{\|x_{s+1}(t)\|_2} \leq \dfrac{\|B_s\|_2 + \|\Delta B_s\|_2}{gap_s - \|E\|_2 - \|\Delta A_S\|_2 - \|B_{s-1}\|_2} = \delta_0$ for all $t \in [0,1]$.

The $(s+1)$th block of $(A + tE)x(t) = \lambda_i(t)x(t)$ is

$$(B_s + t\Delta B_s)x_s(t) + A_{s+1}x_{s+1}(t) + B_{s+1}^H x_{s+2}(t) = \lambda_i(t)x_{s+1}(t),$$

so we get

$$x_{s+1}(t) = (\lambda_i(t)I - A_{s+1})^{-1}\left((B_s + t\Delta B_s)x_s(t) + B_{s+1}^H x_{s+2}(t)\right),$$

and hence $\dfrac{\|x_{s+1}(t)\|_2}{\|x_{s+2}(t)\|_2} \le \dfrac{\|B_{s+1}\|_2}{gap_{s+1} - \|E\|_2 - \|B_s\|_2 - \|\Delta B_s\|_2} = \delta_1$. Similarly we can prove that

$$\frac{\|x_{s+j}(t)\|_2}{\|x_{s+j+1}(t)\|_2} \le \delta_j \quad \text{for} \quad j = 1, \ldots, \ell.$$

Together with $\|x_{s+\ell+1}\|_2 \le \|x\|_2 = 1$ it follows that for all $t \in [0,1]$,

$$\|x_s(t)\|_2 \le \prod_{j=0}^{\ell} \delta_j \|x_{s+\ell+1}(t)\|_2 \le \prod_{j=0}^{\ell} \delta_j,$$

$$\|x_{s+1}(t)\|_2 \le \prod_{j=1}^{\ell} \delta_j \|x_{s+\ell+1}(t)\|_2 \le \prod_{j=1}^{\ell} \delta_j.$$

□

Above we showed how a small eigenvector component implies a small eigenvalue bound. We note that Jiang [**84**] discusses the relation between the convergence of Ritz values obtained in the Lanczos process and eigenvector components of tridiagonal matrices. In patricular, [**84**] argues that a Ritz value must be close to an exact eigenvalue if the corresponding eigenvector of the tridiagonal submatrix has a small bottom element. We argue that Lemma 7.3 can be used to extend this to the block Lanczos method (e.g., [**6**, Ch. 4.6]). Assuming for simplicity that deflation does not occur, after j steps of block Lanczos with block size p we have $AV = VT + [0\ \widehat{V}R]$ where $V \in \mathbb{C}^{n\times jp}$ and $\widehat{V} \in \mathbb{C}^{n\times p}$ have orthonormal columns and $V^H\widehat{V} = 0$. $T \in \mathbb{C}^{jp\times jp}$ is a symmetric banded matrix with bandwidth $2p+1$, and $R \in \mathbb{C}^{p\times p}$. Then we see that letting $U = [V\ \widehat{V}\ V_2]$ be a square unitary matrix, we have

$$U^H A U = \begin{bmatrix} T_{11} & \ddots & & \\ \ddots & \ddots & T_{j,j-1}^H & \\ & T_{j,j-1} & T_{jj} & R^H \\ & & R & A_2 \end{bmatrix}.$$

Note that the top-left $jp \times jp$ submatrix is equal to T. Now, if an eigenvalue λ of T does not belong to the last $s > 0$ blocks of T (which is more likely to happen if λ is an extremal eigenvalue), then by Lemma 7.3 we can show that the bottom block x_p of the eigenvector x corresponding to λ is small. Since the Ritz value λ has residual $\|Ay - \lambda y\|_2 \le \|x_p\|_2\|R\|_2$ where $y = Ux$, this in turn implies that the there must exist an exact eigenvalue of $U^H AU$ lying within distance of $\|R\|_2\|x_p\|_2$ from the Ritz value λ.

We now return to the matrices A and $A + E$ as in (7.11) and present a perturbation bound for λ_i.

THEOREM 7.2. *Let λ_i and $\widehat{\lambda}_i$ be the ith eigenvalue of A and $A+E$ as in* (7.11) *respectively, and let δ_i be as in* (7.14). *Suppose that λ_i satisfies Assumption 1. Then*

$$\left|\lambda_i - \widehat{\lambda}_i\right| \le \|\Delta A_s\|_2 \left(\prod_{j=0}^{\ell} \delta_j\right)^2 + 2\|\Delta B_s\|_2 \delta_0 \left(\prod_{j=1}^{\ell} \delta_j\right)^2. \tag{7.18}$$

PROOF. Using (7.3) we have

$$\begin{aligned}|\lambda_i - \widehat{\lambda}_i| &= \left| \int_0^1 x(t)^* E x(t) dt \right| \\ &\le \left| \int_0^1 x_s(t)^* \Delta A_s x_s(t) dt \right| + 2 \left| \int_0^1 x_{s+1}(t)^* \Delta B_s x_s(t) dt \right| \\ &\le \|\Delta A_s\|_2 \left| \int_0^1 \|x_s(t)\|_2^2 dt \right| + 2\|\Delta B_s\|_2 \left| \int_0^1 \|x_s(t)\|_2 \|x_{s+1}(t)\|_2 dt \right|.\end{aligned}$$

Substituting (7.16) and (7.17) we get

$$\begin{aligned}\left|\lambda_i - \widehat{\lambda}_i\right| &\le \|\Delta A_s\|_2 \left| \left(\prod_{j=0}^{\ell} \delta_j \right)^2 \int_0^1 dt \right| + 2\|\Delta B_s\|_2 \left| \prod_{j=0}^{\ell} \delta_j \prod_{j=1}^{\ell} \delta_j \int_0^1 dt \right| \\ &= \|\Delta A_s\|_2 \left(\prod_{j=0}^{\ell} \delta_j \right)^2 + 2\|\Delta B_s\|_2 \delta_0 \left(\prod_{j=1}^{\ell} \delta_j \right)^2.\end{aligned}$$

□

Below are two remarks on Theorem 7.2.

- Since the bound in (7.18) is proportional to the product of δ_j^2, the bound can be negligibly small if ℓ is large and each δ_j is sufficiently smaller than 1 (say 0.5). Hence Theorem 7.2 shows that λ_i is insensitive to perturbation in far-away blocks, if its separation from the spectrum of the diagonal blocks nearby the perturbed ones is large compared with the off-diagonal blocks. We illustrate this below by an example in Section 7.4.1.
- When the bound (7.14) is smaller than the Weyl bound $\|E\|_2$, we can obtain sharper bounds by using the results recursively, that is, the new bound (7.18) can be used to redefine $\delta_j := \frac{\|B_{s+j}\|_2}{gap_{s+j} - \Delta - \|B_{s+j-1}\|_2}$, where Δ is the right-hand side of (7.18). The new δ_j is smaller than the one in (7.14), and this in turn yields a refined bound (7.18) computed from the new δ_j.

7.4. Two case studies

Here we present two examples to demonstrate the sharpness of our approach. Specifically, we explain

(1) why aggressive early deflation can deflate many eigenvalues as "converged" when applied to the symmetric tridiagonal QR algorithm.
(2) why Wilkinson matrices have many pairs of nearly equal eigenvalues.

In both cases A and E are symmetric tridiagonal, and we denote

$$A + E = \begin{bmatrix} a_1 & b_1 & & \\ b_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & b_{n-1} & a_n \end{bmatrix}, \tag{7.19}$$

where E is zero except for a few off-diagonal elements, as specified below. We assume without loss of generality that $b_j \geq 0$ for all j. Note that when b_j are all nonzero the eigenvalues are known to be always simple [**127**], so the treatment of multiple eigenvalues becomes unnecessary.

In both case studies, we will bound the effect on an eigenvalue λ_i of setting some b_j to 0. We note that [**84**] made the observation that setting b_j to 0 perturbs an eigenvalue extremely insensitively if its eigenvector corresponding to the jth element is negligibly small. However [**84**] does not explain when or why the eigenvector element tends to be negligible. Our approach throughout has been to show that for eigenvalues that are well-separated from the spectrum of the blocks nearby the perturbed blocks, the corresponding eigenvector elements can be bounded without computing them.

7.4.1. Aggressive early deflation applied to symmetric tridiagonal QR. Recall the description in Section 2.6 of aggressive early deflation applied to the symmetric tridiagonal QR algorithm. Here we restate it using the notations of this section. Let $A + E$ as in (7.19) be a matrix obtained in the course of the symmetric tridiagonal QR algorithm. Here we let $A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$, where A_1 is $s \times s$ for an integer parameter s. E has only one off-diagonal b_s. Let $A_2 = VDV^T$ be an eigendecomposition, where the diagonals of D are arranged in decreasing order of magnitude. Then, we have

$$\begin{bmatrix} I & \\ & V \end{bmatrix}^T (A + E) \begin{bmatrix} I & \\ & V \end{bmatrix} = \begin{bmatrix} A_1 & t^T \\ t & D \end{bmatrix}, \tag{7.20}$$

where the vector t is given by $t = b_s V(1,:)^T$ where $V(1,:)$ denotes the first row of V. It often happens in practice that many elements of t are negligibly small, in which case aggressive early deflation regards D's corresponding eigenvalues as converged and deflate them. This is the case even when none of the off-diagonals of A is particularly small.

This must mean that many eigenvalues of the two matrices A and $A + E$, particularly the ones that belong to the bottom-right block, must be nearly equal, or equivalently that the perturbation of the eigenvalues by the sth off-diagonal b_s is negligible. Here we give an explanation to this under an assumption that is typically valid for a tridiagonal matrix appearing in the course of the QR algorithm.

It is well-known that under mild assumptions the tridiagonal QR algorithm converges, in that the diagonals converge to the eigenvalues in descending order of magnitude, and the off-diagonal elements converge to zero [**155**]. In light of this, we can reasonably expect the diagonals a_j to be roughly ordered in descending order of their magnitudes, and that the off-diagonals b_j are small. Hence for a target (small) eigenvalue $\lambda_i \in \lambda(A_2)$, we suppose that Assumption 1 is satisfied, or that there exists an integer $\ell > 0$ such that $|a_j - \lambda_i| > b_{j-1} + b_j + b_s$ for $j = 1, \ldots, s + \ell$.

Under these assumptions, to bound $\left|\lambda_i - \widehat{\lambda}_i\right|$ we can use Theorem 7.2 in which we let all block sizes be 1-by-1. Since $\Delta A_S = 0$, we have $\delta_0 = \frac{2b_s}{gap_s - b_s - b_{s-1}}$, $\delta_1 = \frac{b_{s+1}}{gap_{s+1} - 3b_s}$ $\delta_j =$

$\frac{b_{s+j}}{gap_{s+j}-b_s-b_{s+j-1}}$ for $j \geq 2$, and so using $gap_j = |a_j - \lambda_i|$ we get

$$\left|\lambda_i - \widehat{\lambda}_i\right| \leq 2b_s \frac{2b_s}{|a_s - \lambda_i| - b_s - b_{s-1}} \left(\frac{b_{s+1}}{gap_{s+1} - 3b_s} \prod_{j=2}^{\ell} \frac{b_{s+j}}{|a_{s+j} - \lambda_i| - b_s - b_{s+j-1}} \right)^2 . \tag{7.21}$$

Simple example. To illustrate the result, let $A+E$ be the 1000-by-1000 tridiagonal matrix as in (7.13), where here we let E be zero except for the 900th off-diagonals, which are 1 (i.e., $s = 900$). Note that none of the off-diagonals is negligibly small. We focus on λ_i (the ith smallest eigenvalue of A) for $i = 1, \ldots, 9$, which are smaller than 10. For such λ_i we have $\ell = 87$ (since $\lambda_i \notin [a_j - \eta_j, a_j + \eta_j] = [998 - j, 1004 - j]$ for $j \leq 987$; recall (7.12)), and so (7.21) gives a bound

$$\begin{aligned}
\left|\lambda_i - \widehat{\lambda}_i\right| &\leq 2b_s \frac{2b_s}{|a_s - \lambda_i| - b_s - b_{s-1}} \left(\frac{b_{s+1}}{gap_{s+1} - 3b_s} \prod_{j=2}^{\ell} \frac{b_{s+j}}{|a_{s+j} - \lambda_i| - b_s - b_{s+j-1}} \right)^2 \\
&= 2\frac{2}{|100 - \lambda_i| - 2} \left(\frac{1}{|100 - 1 - \lambda_i| - 3} \prod_{j=2}^{87} \frac{1}{|100 - j - \lambda_i| - 2} \right)^2 \\
&< 2\frac{2}{|100 - 10| - 2} \left(\frac{1}{|100 - 1 - 10| - 3} \prod_{j=1}^{87} \frac{1}{|100 - j - 10| - 2} \right)^2 \\
&< 1.05 \times 10^{-266} \qquad (7.22)
\end{aligned}$$

for $i = 1, \ldots, 9$. This shows that all the eigenvalues of A_2 that are smaller than 10 can be hardly perturbed by setting the off-diagonal b_s to 0.

The same argument as above applied to $i = 1, \ldots, 80$ shows that more than 80 eigenvalues of A_2 match 80 eigenvalues of A to within accuracy 10^{-16}. The general conclusion is that if the diagonal elements of A are roughly graded and the off-diagonals are not too large (compared with the difference between the diagonal elements), then we can show by Theorem 7.2 that the smallest eigenvalues of A are determined accurately by a much smaller lower-right submatrix of A.

We note that [**94**] shows for the non-symmetric Hessenberg QR algorithm that the process of aggressive early deflation can be regarded as extracting converged Ritz vectors by the Krylov-Schur algorithm. Although we treat only the the symmetric tridiagonal case, the advantage of our analysis above is that it gives computable bounds for the accuracy of $\widehat{\lambda}_i$.

7.4.2. Eigenvalues of Wilkinson's matrix. Wilkinson's matrix [**165**] appears both in applications and for theoretical purposes, whose famous $2n+1 = 21$ case is

$$W_{21}^{+} = \begin{bmatrix} 10 & 1 & & & & & & & & & \\ 1 & 9 & \ddots & & & & & & & & \\ & \ddots & \ddots & 1 & & & & & & & \\ & & 1 & 1 & 1 & & & & & & \\ & & & 1 & 0 & 1 & & & & & \\ & & & & 1 & 1 & \ddots & & & & \\ & & & & & \ddots & \ddots & \ddots & & & \\ & & & & & & \ddots & 9 & 1 & & \\ & & & & & & & 1 & 10 & & \end{bmatrix}. \tag{7.23}$$

Such matrices are known to have many pairs of extremely close eigenvalues. For example, the two largest eigenvalues of W_{21}^{+} agree up to about 7×10^{-14}. In general, it is observed that for sufficiently large n the largest eigenvalues of W_{2n+1}^{+} always have an extremely close pair, while smaller eigenvalues have a pair that is not as close. [**165**, p.308] notes that in general the two largest eigenvalues of the matrix W_{2n+1}^{+} agree up to roughly $(n!)^{-2}$, but does not explain this in detail. We shall give an explanation using the framework we described in this chapter. Define $(2n+1)$-by-$(2n+1)$ matrices A and E such that $A + E = W_{2n+1}^{+}$ by

$$A = \begin{bmatrix} n & 1 & & & & & & & & & \\ 1 & n-1 & \ddots & & & & & & & & \\ & \ddots & \ddots & 1 & & & & & & & \\ & & 1 & 1 & 0 & & & & & & \\ & & & 0 & 0 & 0 & & & & & \\ & & & & 0 & 1 & 1 & & & & \\ & & & & & 1 & \ddots & \ddots & & & \\ & & & & & & \ddots & n-1 & 1 & & \\ & & & & & & & 1 & n & & \end{bmatrix}, E = \begin{bmatrix} & & \ddots & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & & \\ \ddots & \ddots & \ddots & & & & & & & & \\ & \ddots & & 1 & & & & & & & \\ & & 1 & 0 & 1 & & & & & & \\ & & & 1 & & \ddots & & & & & \\ & & & & \ddots & \ddots & \ddots & & & & \\ & & & & & \ddots & & & & & \end{bmatrix}. \tag{7.24}$$

Note that A has n pairs of multiple eigenvalues of multiplicity 2, and a simple eigenvalue 0. Hence to prove that large eigenvalues of $A + E = W_{2n+1}^{+}$ (those close to n) must have a very close pair, it suffices to show that large eigenvalues of A are extremely insensitive to the perturbation E.

7.4.2.1. *Bounding* $\left|\widehat{\lambda}_i - \lambda_i\right|$. Unfortunately, the setting (7.24) does not satisfy Assumption 1 which we used to derive Theorem 7.2, because any large eigenvalue of A belongs to some of the top and bottom blocks, which here are 1-by-1 blocks, that is, simply the diagonal elements of A. For example, the two largest eigenvalues of A belong to the top and bottom blocks.

Here we show that by a slight modification of the derivation of Theorem 7.2 we can still get sharp bounds for $|\lambda_i - \widehat{\lambda}_i|$ for large i.

Below we suppose $n > 4$. First we consider the two largest eigenvalues of A, either of which we denote by λ_i (i.e., i can be either $2n$ or $2n+1$). As before, define $x(t) = [x_1(t) \ \dots \ x_{2n+1}(t)]^*$ such that $(A+tE)x(t) = \lambda_i(t)x(t)$.

First, the $(n+1)$st row of $(A+tE)x(t) = \lambda_i(t)x(t)$ is

$$\lambda_i(t)x_{n+1}(t) = t(x_n(t) + x_{n+2}(t)),$$

hence

$$|x_{n+1}(t)| \le \frac{2t\max(|x_n(t)|, |x_{n+2}(t)|)}{|\lambda_i(t)|}, \quad t \in [0,1]. \tag{7.25}$$

We separately consider the two cases $|x_n(t)| \ge |x_{n+2}(t)|$ and $|x_n(t)| < |x_{n+2}(t)|$, and show that in both cases a tight bound can be obtained for $|\lambda_i(A+E) - \lambda_i(A)|$.

Suppose $|x_n(t)| \ge |x_{n+2}(t)|$. Then we also have $|x_n(t)| \ge |x_{n+1}(t)|$ in view of (7.25). From the nth row of $(A+tE)x(t) = \lambda_i(t)x(t)$ we similarly get

$$|x_n(t)| \le \frac{|x_{n-1}(t)| + |x_{n+1}(t)|}{|\lambda_i(t) - 1|}, \quad t \in [0,1]. \tag{7.26}$$

Now since $n < \lambda_i(t) < n+1$ for all $t \in [0,1]$[2] we must have $|x_{n-1}(t)| \ge |x_n(t)| \ge |x_{n+1}(t)|$. Substituting this into (7.26) yields $|x_n(t)| \le \dfrac{|x_{n-1}(t)|}{|\lambda_i(t) - 1| - 1}$. Therefore we have

$$|x_n(t)| \le \frac{|x_{n-1}(t)|}{n-2} \quad \text{for} \quad t \in [0,1].$$

By a similar argument we find for all $t \in [0,1]$ that

$$|x_{n-j}(t)| \le \frac{|x_{n-j-1}(t)|}{n-j-2} \quad \text{for} \quad j = 0, \dots, n-3. \tag{7.27}$$

Hence together with (7.25) we get

$$|x_{n+1}(t)| \le \frac{2t}{n-1}|x_2| \prod_{j=0}^{n-3} \frac{1}{n-j-2} \le \frac{2t}{n-1} \prod_{j=0}^{n-3} \frac{1}{n-j-2}, \tag{7.28}$$

$$|x_n(t)| \le \prod_{j=0}^{n-3} \frac{1}{n-j-2}. \tag{7.29}$$

[2]We can get $n < \lambda(t) < n+1$ by first following the same argument using $n - \|E\|_2 < \lambda(t) < n+1$, which follows from Weyl's and Gerschgorin's theorems.

We now plug these into (7.4) to get

$$\begin{aligned}|\lambda_i(A+E)-\lambda_i(A)| &\le \left|\int_0^1 x(t)^* E x(t) dt\right| \\ &\le \int_0^1 2(|x_n(t)|+|x_{n+2}(t)|)|x_{n+1}(t)|dt \\ &\le \frac{4}{n-1}\left(\prod_{j=0}^{n-3}\frac{1}{n-j-2}\right)^2 \int_0^1 t dt \\ &= \frac{2}{n-1}\left(\prod_{j=0}^{n-3}\frac{1}{n-j-2}\right)^2. \end{aligned} \tag{7.30}$$

The case $|x_n(t)| < |x_{n+2}(t)|$ can also be treated similarly, and we get the same result.

Finally, since (7.30) holds for both $i = 2n$ and $i = 2n+1$, we conclude that

$$|\lambda_{2j}(W_{2n+1}^+) - \lambda_{2j+1}(W_{2n+1}^+)| \le \frac{4}{n-1}\left(\prod_{j=0}^{n-3}\frac{1}{n-j-2}\right)^2. \tag{7.31}$$

We easily appreciate that the bound (7.31) roughly scales as $1/(n-1)((n-2)!)^2$ as $n \to \infty$, which supports the claim in [**165**].

We also note that by a similar argument we can prove for $j \ge 1$ that the $2j-1$th and $2j$th largest eigenvalues of W_{2n+1}^+ match to $O((n-j)^{-1}((n-j-1)!)^{-2})$, which is small for small j, but not as small for larger j. This is an accurate description of what is well known about the eigenvalues of Wilkinson matrices.

In [**168**] Ye investigates tridiagonal matrices with nearly multiple eigenvalues, motivated also by the Wilkinson matrix. We note that we can give another explanation for the nearly multiple eigenvalue by combining ours with Ye's. Specifically, we first consider the block partition $W_{2n+1}^+ = \begin{bmatrix} W_1 & E^* \\ E & W_2 \end{bmatrix}$ where W_1 is $(n+1)$-by-$(n+1)$, and E contains one off-diagonal of W_{2n+1}^+. We can use Theorem 7.2 to show that the largest eigenvalues of W_1 and W_2 are nearly the same; let the distance be δ. Furthermore, we can use Lemma 7.3 to show the corresponding eigenvectors decay exponentially, so that the eigenvector component for W_1 is of order $1/n!$ at the bottom, and that for W_2 is of order $1/n!$ at the top. We can then use Theorem 2.1 of [**168**] to show that W_{2n+1}^+ must have two eigenvalue within distance $d + O(1/n!)$. However, this bound is not as tight as the bound (7.31), being roughly its square root.

7.5. Effect of the presence of multiple eigenvalues

In the text we assumed that all the eigenvalues of $A + tE$ are simple for all $t \in [0,1]$. Here we treat the case where multiple eigenvalues exist, and show that all the results we proved still hold exactly the same.

We note that [**11, 12**] indicate that multiple eigenvalues can be ignored when we take integrals such as (7.3), because $A + tE$ can only have multiple eigenvalues on a set of t of

measure zero, and hence (7.1) can be integrated on t such that $A + tE$ has only simple eigenvalues. However when A, E are both allowed to be arbitrary Hermitian matrices[3] we cannot use this argument, which can be seen by a simple counterexample $A = E = I$, for which $A + tE$ has a multiple eigenvalue for all $0 \leq t \leq 1$. Hence in a general setting we need a different approach.

7.5.1. Multiple eigenvalue first order perturbation expansion. First we review a known result on multiple eigenvalue first order perturbation expansion [**150, 112, 95**]. Suppose that a Hermitian matrix A has a multiple eigenvalue λ_0 of multiplicity r. There exists a unitary matrix $Q = [Q_1, Q_2]$, where Q_1 has r columns, such that

$$Q^*AQ = \begin{bmatrix} \lambda_0 I & 0 \\ 0 & \Lambda \end{bmatrix}, \tag{7.32}$$

where Λ is a diagonal matrix that contains eigenvalues not equal to λ_0. Then, the matrix $A + \varepsilon E$ has eigenvalues $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_r$ admitting the first order expansion

$$\widehat{\lambda}_i = \lambda_0 + \mu_i(Q_1^*EQ_1)\varepsilon + o(\varepsilon), \quad \text{for} \quad i = 1, \ldots, r, \tag{7.33}$$

where $\mu_i(Q_1^*EQ_1)$ denotes the ith eigenvalue of the r-by-r matrix $Q_1^*EQ_1$.

Using (7.33), we obtain the partial derivative corresponding to (7.1) when $A + tE$ has a multiple eigenvalue $\lambda_i(t) = \lambda_{i+1}(t) = \cdots = \lambda_{i+r-1}(t)$ of multiplicity r, with corresponding invariant subspace $Q_1(t)$:

$$\frac{\partial \lambda_{i+j-1}(t)}{\partial t} = \mu_j(Q_1(t)^*EQ_1(t)) \quad \text{for} \quad j = 1, \ldots, r. \tag{7.34}$$

Now, let $Q_1(t)^*EQ_1(t) = U^*DU$ be the eigendecomposition where the diagonals of D are arranged in descending order. Then $D = UQ_1(t)EQ_1(t)U^* = \tilde{Q}_1(t)E\tilde{Q}_1(t)$, where $\tilde{Q}_1(t) = Q_1(t)U^*$, so $\mu_j(Q_1(t)^*EQ_1(t)) = q_j(t)^*Eq_j(t)$, where $q_j(t)$ denotes the jth column of $\tilde{Q}_1(t)$. Therefore we can write

$$\frac{\partial \lambda_{i+j-1}(t)}{\partial t} = q_j(t)^*Eq_j(t) \quad \text{for} \quad j = 1, \ldots, r. \tag{7.35}$$

Now, since any vector of the form $Q_1(t)v$ is an eigenvector corresponding to the eigenvalue $\lambda_i(t)$, so is $q_j(t)$. We conclude that we can always write the first order perturbation expansion of $\lambda_i(t)$ in the form (7.1), in which when $\lambda_i(t)$ is a multiple eigenvalue $x(t)$ represents a particular eigenvector among the many possible choices.

Finally, since all our eigenvector bounds (such as (7.5), (7.16) and (7.17)) hold regardless of whether λ_i is a multiple eigenvalue or not, we conclude that all the bounds in the text hold exactly the same without the assumption that $\lambda_i(t)$ is simple for all $t \in [0, 1]$.

[3] [**12**] makes the assumption that λ is a simple eigenvalue at $t = 0$.

7.5.2. Note on the trailing term. Here we refine the expansion (7.33) by showing that the trailing term is $O(\varepsilon^2)$ instead of $o(\varepsilon)$. To see this, we write $E = \begin{bmatrix} E_{11} & E_{21} \\ E_{21}^* & E_{22} \end{bmatrix}$, and see in (7.32) that

$$Q^*(A+\varepsilon E)Q = \begin{bmatrix} \lambda_0 I + \varepsilon Q_1^* E_{11} Q_1 & \varepsilon Q_1^* E_{21} Q_2 \\ \varepsilon Q_2^* E_{21}^* Q_1 & \Lambda + \varepsilon Q_2^* E_{22} Q_2 \end{bmatrix}.$$

For sufficiently small ε there is a positive *gap* in the spectrums of the matrices $\lambda_0 I + \varepsilon Q_1^* E_{11} Q_1$ and $\Lambda + \varepsilon Q_2^* E_{22} Q_2$. Hence, using the quadratic eigenvalue perturbation bounds in [**97**] we see that the ith eigenvalue of $A+\varepsilon E$ and those of $\begin{bmatrix} \lambda_0 I + \varepsilon Q_1^* E_{11} Q_1 & 0 \\ 0 & \Lambda + \varepsilon Q_2^* E_{22} Q_2 \end{bmatrix}$ differ at most by $\dfrac{2\|\varepsilon E\|_2^2}{gap + \sqrt{gap^2 + 4\|\varepsilon E\|_2^2}}$. This is of size $O(\varepsilon^2)$ because $gap > 0$. Therefore we conclude that (7.33) can be replaced by

$$\widehat{\lambda}_i = \lambda_0 + \mu_i(Q_1^* E Q_1)\varepsilon + O(\varepsilon^2) \quad \text{for} \quad i = 1, 2, \ldots, r. \tag{7.36}$$

CHAPTER 8

Perturbation of generalized eigenvalues

This chapter concerns eigenvalue perturbation bounds for generalized Hermitian definite eigenvalue problems. The goal here is to extend well-known bounds for the standard Hermitian eigenproblems to the generalized case. Specifically, we extend two kinds of bounds, the generic Weyl bound and the quadratic bounds under off-diagonal perturbation.

We first derive Weyl-type eigenvalue perturbation bounds that hold under general perturbation. The results provide a one-to-one correspondence between the original and perturbed eigenvalues, and give a uniform bound. We give both absolute and relative perturbation results, defined in the standard Euclidean metric instead of the chordal metric commonly used in the context of generalized eigenproblems.

We then turn to the case where 2-by-2 block-diagonal Hermitian definite pairs undergo off-diagonal perturbation, and show that eigenvalue perturbation generally scales quadratically with the norms of the perturbation matrices, thereby extending the result by Li and Li [**97**]. We also briefly treat non-Hermitian pairs.

Introduction. This chapter is concerned with eigenvalue perturbation in a generalized Hermitian eigenvalue problem $Ax = \lambda Bx$ where $A, B \in \mathbb{C}^{n\times n}$ are Hermitian and B is positive definite. The theme here is to derive results that are natural extensions of well-known results for Hermitian eigenproblems $Ax = \lambda x$.

As reviewed in Section 2.8, for eigenvalues of a Hermitian definite pair, many properties analogous to the eigenvalues of a Hermitian matrix carry over. In particular, the pair has n real and finite eigenvalues satisfying a min-max property similar to that for Hermitian matrices.

Some perturbation results for generalized eigenvalue problems are known, mostly in the chordal metric [**146**, Ch.6.3], [**24, 100, 102**]. Using the chordal metric is a natural choice for a general matrix pair because it can deal uniformly with infinite eigenvalues. However, this metric is not invariant under scaling, and bounds in this metric may be less intuitive than those defined in the standard Euclidean metric. Most importantly, for a Hermitian definite pair we know a priori that no infinite eigenvalues exist, so in this case the Euclidean metric may be a more natural choice. For these reasons, in this chapter we use the standard Euclidean metric to derive eigenvalue perturbation bounds.

Perhaps the best-known perturbation bound for standard Hermitian eigenproblems is Weyl's theorem [**164, 127, 31**], which gives the uniform bound $\|E\|_2$ for the difference between the ith eigenvalues of A and $A + E$ for any i, as we reviewed in Section 2.7.2.

Despite being merely a special case of the Lidskii-Mirsky-Wielandt theorem [**98**], Weyl's theorem stands out as a simple and useful result that

- Orders and pairs up the original and perturbed eigenvalues, so that we can discuss in terms of the matching distance [**146**, pp.167].
- Gives a bound on the largest distance between a perturbed and exact eigenvalue.

Owing to its simple expression and wide applicability, the theorem has been used in many contexts, e.g., in the basic forward error analysis of standard eigenvalue problems [**6**, Ch.4.8]. In order to distinguish this theorem from the variants discussed below, in this chapter we refer to it as the *absolute Weyl theorem*.

The *relative Weyl theorem*, which is applicable when the perturbation is multiplicative, can provide much tighter bounds for small eigenvalues.

THEOREM 8.1 (Relative Weyl theorem). *Let A be Hermitian and X be nonsingular. Let the eigenvalues of A be $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, let the eigenvalues of X^*AX be $\tilde{\lambda}_1 \le \tilde{\lambda}_2 \le \cdots \le \tilde{\lambda}_n$, and let $\epsilon = \|X^*X - I\|_2$. Then the eigenvalues differ by ϵ in the relative sense, i.e.,*

$$\frac{|\lambda_i - \tilde{\lambda}_i|}{|\lambda_i|} \le \epsilon, \quad i = 1, 2, \ldots, n.$$

This important observation leads to a number of relative perturbation results, along with algorithms that compute eigenvalues/singular values to high relative accuracy including small ones, such as the dqds algorithm (recall Chapter 6), bidiagonal QR and Jacobi's method [**35, 36**].

The first goal of this chapter is to derive Weyl-type theorems for Hermitian definite matrix pairs, both the absolute (Section 8.1) and relative (Section 8.2) versions. Our results employ the Euclidean metric, and have the two Weyl-type properties described above. Compared to known results, our absolute Weyl theorem is simpler than some known bounds (e.g., [**100**]), and our relative Weyl theorem assumes no condition on A. By contrast, the relative perturbation results for Hermitian definite pairs obtained in [**98, 101**] are derived under the assumption that A and B are both positive definite, which limits their applicability; Hermitian definite pairs that arise in practice may not have this property (e.g., [**43**]).

The second part of this chapter treats the quadratic eigenvalue perturbation bounds. For standard Hermitian matrices we summarized the known results in Section 2.7.3, which deal with 2-by-2 block-diagonal matrices under off-diagonal perturbation. A natural extension to a Hermitian definite pair (A, B) is when $A = \mathrm{diag}(A_{11}, A_{22})$ and $B = \mathrm{diag}(B_{11}, B_{22})$, and the two matrices undergo Hermitian perturbations E and F respectively, which are both nonzero only in the $(1, 2)$ and $(2, 1)$ off-diagonal blocks. The second part of this chapter deals with such cases and obtain bounds that scale quadratically with the norms of the perturbation matrices $\|E\|_2$ and $\|F\|_2$.

In this chapter, $\lambda_i(A)$ denotes the ith smallest eigenvalue of a Hermitian matrix A, and $\mathrm{eig}(A, B)$ denotes the set of eigenvalues of the matrix pair (A, B) and $\mathrm{eig}(A) \equiv \mathrm{eig}(A, I)$.

8.1. Absolute Weyl theorem

For a Hermitian definite pair (A, B), we have the following generalization of the absolute Weyl theorem [**6**, Sec. 5.7], when only A is perturbed.

THEOREM 8.2. *Suppose that the Hermitian definite pairs* (A, B) *and* $(A + \Delta A, B)$ *have eigenvalues* $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ *and* $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n$*, respectively. Then for all* $i = 1, 2, \ldots, n$*,*

$$|\lambda_i - \tilde{\lambda}_i| \leq \frac{\|\Delta A\|_2}{\lambda_{\min}(B)}. \tag{8.1}$$

PROOF. Define $Z = B^{-1/2}$ (the matrix square root [**75**, Ch.6] of B). A congruence transformation that multiplies by Z from both sides shows that the pair (A, B) is equivalent to the pair (ZAZ, I). Hence, these pairs and the Hermitian matrix ZAZ have the same eigenvalues. Similarly, the pair $(A + \Delta A, B)$ and the Hermitian matrix $Z(A + \Delta A)Z$ have the same eigenvalues.

Now, to compare the eigenvalues of ZAZ and $Z(A+\Delta A)Z$, we observe that $\|Z\Delta AZ\|_2 \leq \|Z^2\|_2\|\Delta A\|_2 = \|B^{-1}\|_2\|\Delta A\|_2 = \|\Delta A\|_2/\lambda_{\min}(B)$, so we obtain (8.1) by using the absolute Weyl theorem applied to ZAZ and $Z(A + \Delta A)Z$. □

Theorem 8.1 takes into account only perturbations in the matrix A. In practical problems, the matrix B may be obtained from data that may include errors, or may be subject to floating-point representation errors. Therefore, we are also interested in the impact of perturbations in B. We shall derive a bound that takes such perturbations into account.

We will use the following Lemma.

LEMMA 8.1. *Suppose* $(A, B + \Delta B)$ *and* $(A - \mu_i \Delta B, B)$ *are Hermitian positive definite pairs. If* μ_i *is the* i*th eigenvalue of the first pair, then it is also the* i*th eigenvalue of the second pair.*

PROOF. Since the two pairs are Hermitian positive definite, there exist nonsingular Z_1 and Z_2 such that $Z_1^* A Z_1 = \Lambda_1$, $Z_1^*(B + \Delta B)Z_1 = I$, $Z_2^*(A - \mu_i \Delta B)Z_2 = \Lambda_2$ and $Z_2^* B Z_2 = I$, where Λ_1 and Λ_2 are diagonal matrices containing the eigenvalues [**56**, Sec.8.7].

Then, denoting $M = A - \mu_i(B + \Delta B)$, we have $Z_1^* M Z_1 = \Lambda_1 - \mu_i I$ and $Z_2^* M Z_2 = \Lambda_2 - \mu_i I$. Note that by assumption both matrices have a zero eigenvalue, which is the ith eigenvalue of the first matrix. Since by Sylvester's law of inertia [**31**, Sec.5.2], $Z_1^* M Z_1$ and $Z_2^* M Z_2$ have the same inertia, it follows that 0 is the ith eigenvalue of both matrices, so μ_i is the ith eigenvalue of both Λ_1 and Λ_2. □

THEOREM 8.3 (Absolute Weyl theorem for generalized eigenvalue problems). *Suppose* (A, B) *and* $(A + \Delta A, B + \Delta B)$ *are Hermitian positive definite pairs, and let* $\lambda_1 \leq \lambda_1 \leq \cdots \leq \lambda_n$ *be the eigenvalues of* (A, B) *and let* $\widehat{\lambda}_1 \leq \widehat{\lambda}_1 \leq \cdots \leq \widehat{\lambda}_n$ *be the eigenvalues of* $(A + \Delta A, B + \Delta B)$*. Then,*

$$|\widehat{\lambda}_i - \lambda_i| \leq \|(B + \Delta B)^{-1}\|_2 \|\Delta A - \lambda_i \Delta B\|_2 \tag{8.2}$$

and

$$|\widehat{\lambda}_i - \lambda_i| \leq \|B^{-1}\|_2 \|\Delta A - \widehat{\lambda}_i \Delta B\|_2 \tag{8.3}$$

for all $1 \leq i \leq n$*.*

PROOF. We prove (8.3). Suppose

$$(A+\Delta A)\widehat{x}_i = \widehat{\lambda}_i(B+\Delta B)\widehat{x}_i. \tag{8.4}$$

This is equivalent to

$$(A+\Delta A-\widehat{\lambda}_i\Delta B)\widehat{x}_i = \widehat{\lambda}_i B\widehat{x}_i. \tag{8.5}$$

(8.5) means the pair $(A+\Delta A-\widehat{\lambda}_i\Delta B, B)$ has an eigenpair $(\widehat{\lambda}_i, \widehat{x}_i)$. Moreover, using Lemma 8.1 by substituting $A+\Delta A$ into A and $\widehat{\lambda}_i$ into μ_i, we know that $\widehat{\lambda}_i$ is the ith eigenvalue. (8.5) can be transformed into a standard Hermitian eigenvalue problem by premultiplying $B^{-1/2}$ (the matrix square root of B^{-1} [**75**, Ch.5]):

$$B^{-1/2}(A+\Delta A-\widehat{\lambda}_i\Delta B)B^{-1/2}y_i = \widehat{\lambda}_i y_i,$$

where we defined $y_i = B^{1/2}\widehat{x}_i$. Noting that the matrix $B^{-1/2}AB^{-1/2}$ and the pair (A,B) have the same eigenvalues, we conclude by using Weyl's theorem that

$$|\lambda_i - \widehat{\lambda}_i| \le \|B^{-1/2}(\Delta A-\widehat{\lambda}_i\Delta B)B^{-1/2}\|_2 \le \|B^{-1}\|_2\|\Delta A-\widehat{\lambda}_i\Delta B\|_2,$$

which is (8.3). (8.2) can be obtained similarly by starting with $Ax_i = \lambda_i Bx_i$ in (8.4).

□

Note that if $\|\Delta B\|_2 < \lambda_{\min}(B)$ then $(A+\Delta A, B+\Delta B)$ is a Hermitian positive definite pair.

Several points are worth noting regarding Theorem 8.3.

- Theorem 8.3 reduces to Theorem 8.2 when $\Delta B = 0$. Moreover, for the standard Hermitian eigenvalue problem ($B = I$ and $\Delta B = 0$), Theorem 8.3 becomes $|\lambda_i(A) - \lambda_i(A+\Delta A)| \le \|\Delta A\|_2$, the absolute Weyl theorem.
- The result is sharp. This can be seen by the simple example

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Delta A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \text{ and } \Delta B = \begin{pmatrix} -0.8 & 0 \\ 0 & 0 \end{pmatrix}. \tag{8.6}$$

  The eigenvalues of (A,B) are $\{2,1\}$ and those of $(A+\Delta A, B+\Delta B)$ are $\{15,0\}$, so $\max_i|\lambda_i - \tilde{\lambda}_i| = 13$. On the other hand, applying $\|A\|_2 = 2, \|\Delta A\|_2 = 1, \|\Delta B\|_2 = 0.8, \lambda_{\min}(B) = 1$ to (8.3) gives $\max_i|\lambda_i - \tilde{\lambda}_i| \le 1/1 + 0.8(2+1)/(1-0.8) = 13$, matching the actual perturbation.
- It is worth comparing our result with that of Stewart and Sun [**146**, Cor. 3.3]. They give a bound

$$\rho(\lambda_i, \tilde{\lambda}_i) \le \frac{\sqrt{\|\Delta A\|_2 + \|\Delta B\|_2}}{\gamma(A,B)}, \tag{8.7}$$

  where $\gamma(A,B) = \min_{\|x\|_2=1}\sqrt{(x^*Ax)^2 + (x^*Bx)^2}$. Here the metric $\rho(a,b) = |a-b|/\sqrt{(1+a^2)(1+b^2)}$ uses the chordal metric. Noting that the distance between any two numbers a and b is less than 1 in the chordal metric, we see that (8.7) does not provide any information when $\|\Delta A\|_2 + \|\Delta B\|_2 > (\gamma(A,B))^2$. In fact, (8.7) is useless for the matrices in (8.6), because $\|\Delta A\|_2 + \|\Delta B\|_2 = 3$ while $(\gamma(A,B))^2 = 2$. On the other hand, Theorem 8.3 gives a nontrivial bound as long as $\|\Delta B\|_2 < \lambda_{\min}(B)$.

However, when $\|\Delta B\|_2 \geq \lambda_{\min}(B)$ our result is not applicable, whereas it may still be that $\|\Delta A\|_2 + \|\Delta B\|_2 > (\gamma(A,B))^2$, in which case (8.7) is a nontrivial bound. Therefore the two bounds are not comparable in general. An advantage of our result is that it is defined in the Euclidean metric, making its application more direct and intuitive.

- In [**100**] a result similar to Theorem 8.3 is proved, using the chordal metric but directly applicable to the Euclidean metric:

$$|\lambda_i - \tilde{\lambda}_i| \leq \frac{1}{\sqrt{\lambda_{\min}(B)\lambda_{\min}(B+\Delta B)}}\|\Delta A\|_2 + \frac{\|A\|_2/\sqrt{\lambda_{\min}(B)} + \|A+\Delta A\|_2/\sqrt{\lambda_{\min}(B+\Delta B)}}{\lambda_{\min}(B)\lambda_{\min}(B+\Delta B)(\|B\|_2^{-1/2} + \|B+\Delta B\|_2^{-1/2})}\|\Delta B\|_2.$$

Compared to this bound, our result is simpler and requires less information.

8.2. Relative Weyl theorem

We now discuss a generalization of the relative Weyl theorem to Hermitian definite pairs. We show two classes of perturbations that preserve relative accuracy of eigenvalues.

First we observe that a simple analogy from the relative Weyl theorem for standard eigenvalue problems does not work, in the sense that the pairs (X^TAX, B) and (A, B) can have totally different eigenvalues for X such that $\|X^*X - I\|_2$ is small. This is seen by the simple example $A = B = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}$ and $X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$; the first pair has eigenvalues $\{1, 1\}$ while those of the second are $\{100, 0.01\}$. Therefore, the allowed types of multiplicative perturbations have to be more restricted. The following result claims that perturbations of the form $(I + \Delta A)^T A(I + \Delta A)$ are acceptable.

THEOREM 8.4 (Relative Weyl theorem for generalized eigenvalue problems 1). *Let a Hermitian definite pair (A, B) have eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, and let $\sqrt{\kappa_2(B)}\|\Delta A\|_2 = \epsilon$. If $\|\Delta B\|_2 < \lambda_{\min}(B)$, then $((I+\Delta A)^T A(I+\Delta A), B+\Delta B)$ is a Hermitian definite pair whose eigenvalues $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n$ satisfy*

$$|\lambda_i - \tilde{\lambda}_i| \leq \left(\epsilon(2+\epsilon) + (1+\epsilon)^2 \frac{\|\Delta B\|_2}{\lambda_{\min}(B) - \|\Delta B\|_2}\right)|\lambda_i|, \quad i = 1, 2, \ldots, n. \tag{8.8}$$

PROOF. The fact that $((I + \Delta A)^T A(I + \Delta A), B + \Delta B)$ is Hermitian definite is trivial. Define $Z = B^{-1/2}$ and $Z_1 = (I + Z\Delta BZ)^{-1/2}$. Note that

$$\|Z_1 Z_1 - I\|_2 \leq \frac{1}{(1 - \|\Delta B\|_2)/\lambda_{\min}(B)} - 1 = \frac{\|\Delta B\|_2}{\lambda_{\min}(B) - \|\Delta B\|_2}. \tag{8.9}$$

We see that the comparison between the eigenvalues of $A - \lambda B$ and $(I + \Delta A)^T A(I + \Delta A) - \lambda(B + \Delta B)$ is equivalent to a comparison between the eigenvalues of the Hermitian matrices ZAZ and $Z_1 Z(I + \Delta A)^* A(I + \Delta A)ZZ_1$, so our goal is to compare the eigenvalues of these two matrices.

The key idea is to consider the matrix $X = I+Z^{-1}\Delta AZ$, which satisfies $Z(I+\Delta A)^*A(I+\Delta A)Z = X^*ZAZX$. Note that $\|Z^{-1}\Delta AZ\|_2 \le \kappa_2(Z)\|\Delta A\|_2 = \sqrt{\kappa_2(B)}\|\Delta A\|_2 (\equiv \epsilon)$, so

$$\|X^*X - I\|_2 = \|Z^{-1}\Delta AZ + (Z^{-1}\Delta AZ)^* + (Z^{-1}\Delta AZ)^*Z^{-1}\Delta AZ\|_2 \le \epsilon(2+\epsilon).$$

Therefore, by using the relative Weyl theorem for ZAZ and $Z(I+\Delta A)^*A(I+\Delta A)Z$ and recalling that $\lambda_i(ZAZ) = \lambda_i$, we obtain

$$\begin{aligned} |\lambda_i(Z(I+\Delta A)^*A(I+\Delta A)Z) - \lambda_i(ZAZ)| &= |\lambda_i(X^*ZAZX) - \lambda_i(ZAZ)| \\ &\le |\lambda_i(ZAZ)| \cdot \|X^*X - I\|_2 \\ &\le \epsilon(2+\epsilon)|\lambda_i|. \end{aligned} \tag{8.10}$$

Now to compare the eigenvalues between $Z(I+\Delta A)^*A(I+\Delta A)Z$ and $Z_1Z(I+\Delta A)^*A(I+\Delta A)ZZ_1$, we use the relative Weyl theorem again to get

$$\begin{aligned} &|\lambda_i(Z_1Z(I+\Delta A)^*A(I+\Delta A)ZZ_1) - \lambda_i(Z(I+\Delta A)^*A(I+\Delta A)Z)| \\ \le&|\lambda_i(Z(I+\Delta A)^*A(I+\Delta A)Z)| \cdot \|Z_1Z_1 - I\|_2 \\ \le&(1+\epsilon)^2|\lambda_i| \cdot \frac{\|\Delta B\|_2}{\lambda_{\min}(B) - \|\Delta B\|_2}. \quad \text{(by (8.10) and (8.9))} \end{aligned}$$

Combining the above yields (8.8):

$$\begin{aligned} &|\lambda_i(Z_1Z(I+\Delta A)^*A(I+\Delta A)ZZ_1) - \lambda_i| \\ \le&\epsilon(2+\epsilon)|\lambda_i| + (1+\epsilon)^2|\lambda_i| \cdot \frac{\|\Delta B\|_2}{\lambda_{\min}(B) - \|\Delta B\|_2} \\ \le&\left(\epsilon(2+\epsilon) + (1+\epsilon)^2\frac{\|\Delta B\|_2}{\lambda_{\min}(B) - \|\Delta B\|_2}\right)|\lambda_i|. \end{aligned}$$

□

The next result shows that a simpler result can be obtained when both perturbations are multiplicative and the pair can be expressed as $((I+\Delta A)^*A(I+\Delta A), (I+\Delta B)^*B(I+\Delta B))$.

THEOREM 8.5 (Relative Weyl theorem for generalized eigenvalue problems 2). *Let the Hermitian definite pairs (A,B) and $((I+\Delta A)^TA(I+\Delta A), (I+\Delta B)^*B(I+\Delta B))$ have eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ and $\tilde\lambda_1 \le \tilde\lambda_2 \le \cdots \le \tilde\lambda_n$, respectively. Suppose that $\sqrt{\kappa_2(B)}\|\Delta A\|_2 = \epsilon$ and $\sqrt{\kappa_2(B)}\|\Delta B\|_2 = \delta < 1$. Then, $\tilde\lambda_i$ $(1 \le i \le n)$ satisfy*

$$|\lambda_i - \tilde\lambda_i| \le \left(\epsilon(2+\epsilon) + \frac{(1+\epsilon)^2\delta(2-\delta)}{(1-\delta)^2}\right)|\lambda_i|, \quad i = 1,2,\ldots,n. \tag{8.11}$$

PROOF. Define $Z = B^{-1/2}$ and consider $Y = I + Z^{-1}\Delta BZ$, which satisfies $(I+\Delta B)^TB(I+\Delta B) = Z^{-1}Y^*YZ^{-1}$. We observe that the pair $((I+\Delta A)^TA(I+\Delta A), (I+\Delta B)^TB(I+\Delta B)) = ((I+\Delta A)^TA(I+\Delta A), Z^{-1}Y^*YZ^{-1})$ has the same eigenvalues as the matrix $Y^{-H}Z(I+\Delta A)^TA(I+\Delta A)ZY^{-1}$. Hence we shall compare the eigenvalues of the matrices ZAZ and $Y^{-H}Z(I+\Delta A)^TA(I+\Delta A)ZY^{-1}$.

Using the same argument as in the proof of Theorem 8.4, we have (cf. (8.10))

$$|\lambda_i(Z(I+\Delta A)^*A(I+\Delta A)Z) - \lambda_i| = \epsilon(2+\epsilon)|\lambda_i|. \tag{8.12}$$

Next we recall that $Y = I + Z^{-1}\Delta BZ$, and see that $\|Z^{-1}\Delta BZ\|_2 \leq \kappa_2(Z)\|\Delta B\|_2 = \sqrt{\kappa_2(B)}\|\Delta B\|_2 (\equiv \delta)$. It follows that the singular values of Y^{-1} lie in $[1/(1+\delta), 1/(1-\delta)]$, so we have

$$\|Y^{-H}Y^{-1} - I\|_2 \leq 1/(1-\delta)^2 - 1 = \frac{\delta(2-\delta)}{(1-\delta)^2}.$$

Therefore, using the relative Weyl theorem and (8.12) we have

$$\begin{aligned}
&|\lambda_i(Y^{-H}Z(I+\Delta A)^*A(I+\Delta A)ZY^{-1}) - \lambda_i(Z(I+\Delta A)^*A(I+\Delta A)Z)| \\
\leq &|\lambda_i(Z(I+\Delta A)^*A(I+\Delta A)Z)| \cdot \|Y^{-H}Y^{-1} - I\|_2 \\
\leq &(1+\epsilon)^2\frac{\delta(2-\delta)}{(1-\delta)^2}|\lambda_i|.
\end{aligned}$$

Therefore, (8.11) is obtained by

$$\begin{aligned}
&|\lambda_i(Y^{-H}Z(I+\Delta A)^*A(I+\Delta A)ZY^{-1}) - \lambda_i| \\
\leq &\epsilon(2+\epsilon)|\lambda_i| + \frac{(1+\epsilon)^2\delta(2-\delta)}{(1-\delta)^2}|\lambda_i|.
\end{aligned}$$

□

Theorems 8.4 and 8.5 do not directly match the relative Weyl theorem for standard eigenvalue problems by letting $B = I$ and $\Delta B = 0$, because a general unitary transformation on A is not allowed.

Nonetheless, our results are consistent, as the following argument indicates. Consider the pair (X^*AX, I). If $\|X^*X - I\|_2 = \epsilon$ and $\epsilon < 1$ then the singular values of X must lie in $[\sqrt{1-\epsilon}, \sqrt{1+\epsilon}]$. Hence, X can be written as $X = U + \Delta U$, in which U is the unitary polar factor of X (the closest unitary matrix to X [**75**, pp.197]) and $\|\Delta U\|_2 \leq 1 - \sqrt{1-\epsilon}$. Then, the pair (X^*AX, I) is rewritten as $(U^*(I + (\Delta UU^*)^*)A(I + \Delta UU^*)U, I)$, which a unitary transformation shows is equivalent to $((I + (\Delta UU^*)^*)A(I + \Delta UU^*), I)$. Noting that $\|\Delta UU^*\|_2 = \|\Delta U\|_2 \leq 1 - \sqrt{1-\epsilon}$ and using Theorem 8.4 (or 8.5) for the pairs (A, I) and $((I + (\Delta UU^*)^*)A(I + \Delta UU^*), I)$, we see that the pair (X^*AX, I) has eigenvalues that match those of the pair (A, I) to relative accuracy $(1 - \sqrt{1-\epsilon})(2 + 1 - \sqrt{1-\epsilon})$. Notice that $(1 - \sqrt{1-\epsilon})(2 + 1 - \sqrt{1-\epsilon}) \simeq \epsilon$ when $\epsilon \ll 1$, yielding the relative Weyl theorem. Hence, Theorems 8.4 and 8.5 become equivalent to the relative Weyl theorem when $B = I$, $\Delta B = 0$ and $\epsilon \ll 1$.

8.3. Quadratic perturbation bounds for Hermitian definite pairs

The rest of this chapter is concerned with the Hermitian definite generalized eigenvalue problem $Ax = \lambda Bx$ for block diagonal matrices $A = \mathrm{diag}(A_{11}, A_{22})$ and $B = \mathrm{diag}(B_{11}, B_{22})$. As before, both A and B are Hermitian and B is positive definite. We establish bounds on how its eigenvalues vary when A and B are perturbed by Hermitian matrices. These bounds are generally of linear order with respect to the perturbations in the diagonal blocks and of quadratic order with respect to the perturbations in the off-diagonal blocks. The results for the case of no perturbations in the diagonal blocks can be used to bound the changes of eigenvalues of a Hermitian definite generalized eigenvalue problem after its off-diagonal

blocks are dropped, a situation that occurs frequently in eigenvalue computations. The presented results extend those of Li and Li [**97**].

In the rest of this chapter we assume that A and B are 2-by-2 block diagonal

$$A = \begin{matrix} {\scriptstyle m}\{ \\ {\scriptstyle n}\{ \end{matrix} \overbrace{\begin{bmatrix} A_{11} & \\ & A_{22} \end{bmatrix}}^{\begin{matrix} m & n \end{matrix}}, \quad B = \begin{matrix} {\scriptstyle m}\{ \\ {\scriptstyle n}\{ \end{matrix} \overbrace{\begin{bmatrix} B_{11} & \\ & B_{22} \end{bmatrix}}^{\begin{matrix} m & n \end{matrix}}. \tag{8.1}$$

When A and B are perturbed to

$$\widetilde{A} \stackrel{\text{def}}{=} A + E = \begin{bmatrix} A_{11} + E_{11} & E_{12} \\ E_{21} & A_{22} + E_{22} \end{bmatrix}, \; \widetilde{B} \stackrel{\text{def}}{=} B + F = \begin{bmatrix} B_{11} + F_{11} & F_{12} \\ F_{21} & B_{22} + F_{22} \end{bmatrix} \tag{8.2}$$

by two Hermitian matrices E and F, we are interested in bounding how much the eigenvalues of (A, B) change. Two kinds of bounds will be established:

- bounds on the difference between the eigenvalues of (A, B) and those of $(\widetilde{A}, \widetilde{B}$;
- bounds on the difference between the eigenvalues of (A_{11}, B_{11}) and some m eigenvalues of $(\widetilde{A}, \widetilde{B})$.

There are two immediate applicable situations of such bounds. The first situation is when we have a nearly block diagonal pair, that is, $E_{ii} = F_{ii} = 0$ in (8.2) and all E_{ij} and F_{ij} for $i \neq j$ are small in magnitude relative to A_{ii} and B_{ii}. Such a situation naturally arises when one uses a Jacobi-type algorithm [**127**, p.353] that iteratively reduces both A and B to diagonal form. The second situation arises from the solution of a large-scale generalized Hermitian eigenvalue problem, where one may have an approximate eigenspace. Projecting the pair onto the approximate eigenspace via the Rayleigh-Ritz process leads to (8.2) with again $E_{ii} = F_{ii} = 0$ for $i = 1, 2$ and some norm estimates on E_{ij} and F_{ij} for $i \neq j$ but usually unknown A_{22} and B_{22}. In such a case, we would like to estimate how well the eigenvalues of (A_{11}, B_{11}) approximate some of those of $(\widetilde{A}, \widetilde{B})$.

The special case when $B = \widetilde{B} = I_N$, the identity matrix can be dealt with well by some existing theories. For example, if all blocks in E have similar magnitudes, we may simply bound the eigenvalue changes using the norms of E by the Weyl-Lidskii theorem [**14, 127, 146**]; if E_{ij} for $i \neq j$ have much smaller magnitudes relative to E_{ii} for $i = 1, 2$, we may write

$$\widehat{A} = A + \begin{bmatrix} & E_{12} \\ E_{21} & \end{bmatrix}, \quad \widetilde{A} = \widehat{A} + \begin{bmatrix} E_{11} & \\ & E_{22} \end{bmatrix},$$

then the eigenvalue differences for A and $\widetilde{A}$ can be bounded in two steps: bounding the differences for A and $\widehat{A}$ and the differences for $\widehat{A}$ and $\widetilde{A}$. The eigenvalue differences for A and $\widehat{A}$ are potentially of the second order in E_{ij} $(i \neq j)$ and are no worse than of the first order in E_{ij} $(i \neq j)$ [**22, 97, 109, 149**], while the eigenvalue differences for $\widehat{A}$ and $\widetilde{A}$ can be again bounded using the norms of E by the Weyl-Lidskii theorem.

The rest of the chapter is organized as follows. In Section 8.3 we present our main results: error bounds on the differences between the eigenvalues of pairs (8.1) and (8.2). These error bounds are usually quadratic in the norms of E_{21} and F_{21}. The case when $B = I$ and $F_{21} = 0$ has been investigated in [**97**] and in fact our bounds here reduce to the ones there

in this case. In Section 8.4 we briefly consider perturbation of partitioned matrices in the non-Hermitian case.

Set $N = m + n$, and denote the eigenvalues of (A, B) and $(\widetilde{A}, \widetilde{B})$ by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \quad \text{and} \quad \widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_N, \tag{8.3}$$

respectively. Here we define the spectrum gap by

$$\eta_i \stackrel{\text{def}}{=} \begin{cases} \min\limits_{\mu_2 \in \text{eig}(A_{22}, B_{22})} |\lambda_i - \mu_2|, & \text{if } \lambda_i \in \text{eig}(A_{11}, B_{11}), \\ \min\limits_{\mu_1 \in \text{eig}(A_{11}, B_{11})} |\lambda_i - \mu_1|, & \text{if } \lambda_i \in \text{eig}(A_{22}, B_{22}), \end{cases} \tag{8.4a}$$

$$\eta \stackrel{\text{def}}{=} \min_{1 \leq i \leq N} \eta_i = \min_{\mu_1 \in \text{eig}(A_{11}, B_{11}), \mu_2 \in \text{eig}(A_{22}, B_{22})} |\mu_1 - \mu_2|. \tag{8.4b}$$

For the sake of this definition, we treat a multiple eigenvalue as different copies of the same value. If the multiple eigenvalue comes from both $\text{eig}(A_{11}, B_{11})$ and $\text{eig}(A_{22}, B_{22})$, each copy is regarded as an eigenvalue of only one of (A_{ii}, B_{ii}) for $i = 1, 2$ but not both.

8.3.1. Special Case. We will start by considering the special case:

$$E_{ii} = F_{ii} = 0, \quad B_{11} = I_m, \quad B_{22} = I_n. \tag{8.5}$$

For this case,

$$\widetilde{A} = \begin{bmatrix} A_{11} & E_{21}^* \\ E_{21} & A_{22} \end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix} I_m & F_{21}^* \\ F_{21} & I_n \end{bmatrix}. \tag{8.6}$$

We shall bound the differences $\widetilde{\lambda}_j - \lambda_j$ via three different approaches. Throughout this subsection we assume that

$$\|F_{21}\|_2 < 1,$$

so that $\widetilde{B}$ is Hermitian positive definite.

Method I. Noting that $I - F_{21}F_{21}^*$ is Hermitian definite, we let

$$X = \begin{bmatrix} I_m & -F_{21}^* \\ 0 & I_n \end{bmatrix}, \quad W = \begin{bmatrix} I_m & 0 \\ 0 & [I - F_{21}F_{21}^*]^{1/2} \end{bmatrix}. \tag{8.7}$$

Then

$$\widehat{B} \stackrel{\text{def}}{=} X^* \widetilde{B} X = \begin{bmatrix} I_m & 0 \\ 0 & I - F_{21}F_{21}^* \end{bmatrix} = W^2, \tag{8.8a}$$

$$\widehat{A} \stackrel{\text{def}}{=} X^* \widetilde{A} X = \begin{bmatrix} A_{11} & -A_{11}F_{21}^* + E_{21}^* \\ -F_{21}A_{11} + E_{21} & A_{22} - E_{21}F_{21}^* - F_{21}E_{21}^* + F_{21}A_{11}F_{21}^* \end{bmatrix}. \tag{8.8b}$$

We now consider the following four eigenvalue problems:

$$\begin{aligned} &\text{EIG (a)}: \quad (\widetilde{A}, \widetilde{B}) \text{ which has the same eigenvalues as } (W^{-1}\widehat{A}W^{-1}, I_N), \\ &\text{EIG (b)}: \quad (\widehat{A}, I_N), \\ &\text{EIG (c)}: \quad \left(\begin{bmatrix} A_{11} & -A_{11}F_{21}^* + E_{21}^* \\ -F_{21}A_{11} + E_{21} & A_{22} \end{bmatrix}, I_N \right), \\ &\text{EIG (d)}: \quad (A, I_N). \end{aligned}$$

Denote the eigenvalues for EIG ($\mathtt{x}$) by $\lambda_j^{(\mathtt{x})}$ in descending order, i.e.,

$$\lambda_1^{(\mathtt{x})} \geq \lambda_2^{(\mathtt{x})} \geq \cdots \geq \lambda_N^{(\mathtt{x})}. \tag{8.9}$$

Then we have $\lambda_j^{(\mathrm{a})} = \widetilde{\lambda}_j$ and $\lambda_j^{(\mathrm{d})} = \lambda_j$ for all j, recalling (8.3) and (8.5). There are existing perturbation bounds for any two adjacent eigenvalue problems in the above list.

(a-b) There exist t_j $(1 \leq j \leq N)$ satisfying

$$1/[\sigma_{\max}(W)]^2 = [\sigma_{\min}(W^{-1})]^2 \leq t_j \leq [\sigma_{\max}(W^{-1})]^2 = 1/[\sigma_{\min}(W)]^2$$

such that

$$\lambda_j^{(\mathrm{a})} = t_j \lambda_j^{(\mathrm{b})}, \text{ or equivalently } \lambda_j^{(\mathrm{b})} = t_j^{-1} \lambda_j^{(\mathrm{a})}, \text{ for } 1 \leq j \leq N.$$

It can be seen that $\sigma_{\max}(W) = 1$ and $\sigma_{\min}(W) = \sqrt{1 - \|F_{21}\|_2^2}$. Thus

$$0 \leq 1 - t_j^{-1} \leq \|F_{21}\|_2^2.$$

(b-c) By Weyl's theorem, we have $|\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}| \leq \|E_{21}F_{21}^* + F_{21}E_{21}^* - F_{21}A_{11}F_{21}^*\|_2$ for $1 \leq j \leq N$.

(c-d) By (2.25), for $1 \leq j \leq N$

$$|\lambda_j^{(\mathrm{c})} - \lambda_j^{(\mathrm{d})}| \leq \frac{2\|A_{11}F_{21}^* - E_{21}^*\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|A_{11}F_{21}^* - E_{21}^*\|_2^2}} \leq \frac{2\|E_{21} - F_{21}A_{11}\|_2^2}{\eta + \sqrt{\eta^2 + 4\|E_{21} - F_{21}A_{11}\|_2^2}}.$$

Combining these three bounds we get for $1 \leq j \leq N$,

$$\begin{aligned} |\widetilde{\lambda}_j - \lambda_j| =& |\lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{d})}| \\ =& |\lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{b})} + \lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})} + \lambda_j^{(\mathrm{c})} - \lambda_j^{(\mathrm{d})}| \\ \leq& |1 - t_j^{-1}|\,|\lambda_j^{(\mathrm{a})}| + |\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}| + |\lambda_j^{(\mathrm{c})} - \lambda_j^{(\mathrm{d})}| \\ \leq& \|F_{21}\|_2^2\,|\widetilde{\lambda}_j| + \|E_{21}F_{21}^* + F_{21}E_{21}^* - F_{21}A_{11}F_{21}^*\|_2 \\ &+ \frac{2\|E_{21} - F_{21}A_{11}\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|E_{21} - F_{21}A_{11}\|_2^2}}. \end{aligned} \tag{8.10}$$

REMARK 4. If $\eta_j > 0$, the right-hand side of (8.10) is of $O(\max\{\|E_{21}\|_2^2, \|F_{21}\|_2^2\})$ for that j. If $\eta > 0$, it is of $O(\max\{\|E_{21}\|_2^2, \|F_{21}\|_2^2\})$ for all $1 \leq j \leq N$. We can establish more bounds between $\widetilde{\lambda}_j$ and λ_j by using various other existing bounds available to bound the differences among $\lambda_j^{(\mathtt{x})}$'s. Interested readers may find them in [**14, 41, 98, 101, 103, 109, 127, 146**].

Symmetrically permute $\widetilde{A}$ and $\widetilde{B}$ in (8.2) to

$$\begin{bmatrix} A_{22} & E_{21} \\ E_{21}^* & A_{11} \end{bmatrix}, \quad \begin{bmatrix} I_n & F_{21} \\ F_{21}^* & I_m \end{bmatrix}$$

and then apply (8.10) to get

$$|\widetilde{\lambda}_j - \lambda_j| \leq \|F_{21}\|_2^2\,|\widetilde{\lambda}_j| + \|E_{21}^*F_{21} + F_{21}^*E_{21} - F_{21}^*A_{22}F_{21}\|_2$$

$$+\frac{2\|E_{21}-A_{22}F_{21}\|_2^2}{\eta_j+\sqrt{\eta_j^2+4\|E_{21}-A_{22}F_{21}\|_2^2}}. \tag{8.11}$$

The following theorem summarizes what we have obtained so far.

THEOREM 8.6. *Assume* (8.5) *and* $\|F_{21}\|_2<1$ *for the Hermitian definite pairs* (A,B) *and* $(\widetilde{A},\widetilde{B})$ *with* A, B, $\widetilde{A}$, *and* $\widetilde{B}$ *as in* (8.1) *and* (8.2). *Denote their eigenvalues as in* (8.3) *and define gaps* η_i *and* η *as in* (8.4). *Then both* (8.10) *and* (8.11) *hold for all* $1\le j\le N$.

We now investigate how accurate the eigenvalues of A_{11} are as approximations to a subset of the eigenvalues of $(\widetilde{A},\widetilde{B})$. Recall (8.5). We have

$$R\overset{\text{def}}{=}\widetilde{A}\begin{bmatrix}I_m\\0\end{bmatrix}-\widetilde{B}\begin{bmatrix}I_m\\0\end{bmatrix}A_{11}=\begin{bmatrix}0\\E_{21}-F_{21}A_{11}\end{bmatrix}. \tag{8.12}$$

Note that $\widetilde{B}=X^{-*}W^2X^{-1}$ and $WX^{-1}\begin{bmatrix}I_m\\0\end{bmatrix}=XW^{-1}\begin{bmatrix}I_m\\0\end{bmatrix}=\begin{bmatrix}I_m\\0\end{bmatrix}$ by (8.7) and (8.8). Thus $\operatorname{eig}(\widetilde{A},\widetilde{B})=\operatorname{eig}(W^{-1}X^*\widetilde{A}XW^{-1})$, and

$$\begin{aligned}W^{-1}X^*R&\equiv\left[W^{-1}X^*\widetilde{A}XW^{-1}\right]\begin{bmatrix}I_m\\0\end{bmatrix}-\begin{bmatrix}I_m\\0\end{bmatrix}A_{11}\\&=\begin{bmatrix}0\\[I-F_{21}F_{21}^*]^{-1/2}(E_{21}-F_{21}A_{11})\end{bmatrix}.\end{aligned} \tag{8.13}$$

Hence we obtain the following.

THEOREM 8.7. *Assume the conditions of* Theorem 8.6. *There are* m *eigenvalues* $\mu_1\ge\cdots\ge\mu_m$ *of* $(\widetilde{A},\widetilde{B})$ *such that*

$$|\mu_j-\theta_j|\le\frac{\|E_{21}-F_{21}A_{11}\|_2}{\sqrt{1-\|F_{21}\|_2^2}}\quad \textit{for } 1\le j\le m,$$

$$\sqrt{\sum_{j=1}^m|\mu_j-\theta_j|^2}\le\frac{\|E_{21}-F_{21}A_{11}\|_{\mathrm{F}}}{\sqrt{1-\|F_{21}\|_2^2}},$$

where $\theta_1\ge\cdots\ge\theta_m$ *are the* m *eigenvalues of* A_{11}.

Method II. Much of this approach is adapted from [**97**] for the standard eigenvalue problem. We shall begin by seeking motivation from a 2-by-2 example.

EXAMPLE 8.1. Consider the 2×2 Hermitian matrices

$$\widetilde{A}=\begin{bmatrix}\alpha&\epsilon^*\\\epsilon&\beta\end{bmatrix},\quad \widetilde{B}=\begin{bmatrix}1&\delta^*\\\delta&1\end{bmatrix}, \tag{8.14}$$

where $\alpha>\beta$ and $1-|\delta|^2>0$ (i.e., B is positive definite). The eigenvalues of $(\widetilde{A},\widetilde{B})$, denoted by $\lambda_\pm$, satisfy

$$(1-|\delta|^2)\lambda^2-(\alpha+\beta-\epsilon^*\delta-\epsilon\delta^*)\lambda+\alpha\beta-|\epsilon|^2=0.$$

Let

$$\begin{aligned}\Delta &= (\alpha+\beta-\epsilon^*\delta-\epsilon\delta^*)^2-4(1-|\delta|^2)(\alpha\beta-|\epsilon|^2)\\ &= (\alpha-\beta)^2+2[(\alpha\delta-\epsilon)^*(\beta\delta-\epsilon)+(\alpha\delta-\epsilon)(\beta\delta-\epsilon)^*]+(\epsilon^*\delta-\epsilon\delta^*)^2.\end{aligned}$$

The eigenvalues of $(\widetilde{A},\widetilde{B})$ are

$$\lambda_\pm=\frac{(\alpha+\beta-\epsilon^*\delta-\epsilon\delta^*)\pm\sqrt{\Delta}}{2(1-|\delta|^2)}.$$

It is not obvious to see $\lambda_+\geq\alpha$ and $\lambda_-\leq\beta$ from this expression. A better way to see $\lambda_+\geq\alpha$ and $\lambda_-\leq\beta$ is through the min-max principle, see Section 2.8.1. Namely

$$\lambda_+=\max_x\frac{x^*\widetilde{A}x}{x^*\widetilde{B}x}\geq\frac{e_1^*\widetilde{A}e_1}{e_1^*\widetilde{B}e_1}=\alpha,\quad \lambda_-=\min_x\frac{x^*\widetilde{A}x}{x^*\widetilde{B}x}\leq\beta,$$

where e_1 is the first column of the identity matrix. Consider

$$\widetilde{A}-\lambda_+\widetilde{B}=\begin{bmatrix}\alpha-\lambda_+ & (\epsilon-\lambda_+\delta)^*\\ \epsilon-\lambda_+\delta & \beta-\lambda_+\end{bmatrix}.$$

Since $\lambda_+\geq\alpha>\beta$, we can define

$$X=\begin{bmatrix}1 & 0\\ -(\beta-\lambda_+)^{-1}(\epsilon-\lambda_+\delta) & 1\end{bmatrix}.$$

We have

$$X^*(\widetilde{A}-\lambda_+\widetilde{B})X=\begin{bmatrix}\alpha-\lambda_+-(\epsilon-\lambda_+\delta)^*(\beta-\lambda_+)^{-1}(\epsilon-\lambda_+\delta) & 0\\ 0 & \beta-\lambda_+\end{bmatrix}$$

which must be singular. Since $\lambda_+\geq\alpha>\beta$, we have

$$\begin{aligned}&\alpha-\lambda_+-(\epsilon-\lambda_+\delta)^*(\beta-\lambda_+)^{-1}(\epsilon-\lambda_+\delta)=0,\\ &\alpha-\lambda_+=\frac{|\epsilon-\lambda_+\delta|^2}{(\beta-\alpha)+(\alpha-\lambda_+)},\\ &(\lambda_+-\alpha)^2+(\alpha-\beta)(\lambda_+-\alpha)-|\epsilon-\lambda_+\delta|^2=0.\end{aligned}$$

The last equation gives, upon noticing that $\lambda_+-\alpha\geq0$, that

$$\lambda_+-\alpha=\frac{2|\epsilon-\lambda_+\delta|^2}{(\alpha-\beta)+\sqrt{(\alpha-\beta)^2+4|\epsilon-\lambda_+\delta|^2}}. \tag{8.15}$$

We also apply (8.15) to $(-\widetilde{A},\widetilde{B})$ to get

$$\beta-\lambda_-=\frac{2|\epsilon-\lambda_-\delta|^2}{(\alpha-\beta)+\sqrt{(\alpha-\beta)^2+4|\epsilon-\lambda_-\delta|^2}}. \tag{8.16}$$

We next show that an inequality similar to (8.15) and (8.16) holds for the general case as stated in Theorem 8.8 below. □

THEOREM 8.8. *Under the conditions of* Theorem 8.6*, we have for all* $1 \le i \le N$

$$|\widetilde{\lambda}_i - \lambda_i| \le \frac{2\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2}{\eta_i + \sqrt{\eta_i^2 + 4\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2}} \le \frac{2\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2}{\eta + \sqrt{\eta^2 + 4\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2}}. \tag{8.17}$$

PROOF. We shall prove (8.17) by induction. We may assume that A_{11} and A_{22} are diagonal with their diagonal entries arranged in descending order, respectively; otherwise replace $\widetilde{A}$ and $\widetilde{B}$ by

$$(U \oplus V)^* \widetilde{A}(U \oplus V) \quad \text{and} \quad (U \oplus V)^* \widetilde{B}(U \oplus V),$$

respectively, where U and V are unitary such that $U^* A_{11} U$ and $V^* A_{22} V$ are in such diagonal forms.

We may perturb the diagonal of A so that all entries are distinct, and apply the continuity argument for the general case.

If $N = 2$, the result is true by Example 8.1. Assume that $N > 2$, and that the result is true for Hermitian matrices of size $N - 1$.

First, we show that (8.17) holds for $i = 1$. Assume that the $(1,1)$th entry of A_{11} equals λ_1. By the min-max principle, we have

$$\widetilde{\lambda}_1 \ge \frac{e_1^* \widetilde{A} e_1}{e_1^* \widetilde{B} e_1} = \lambda_1 .$$

No proof is necessary if $\widetilde{\lambda}_1 = \lambda_1$. Assume $\widetilde{\lambda}_1 > \lambda_1$ and let

$$X = \begin{bmatrix} I_m & 0 \\ -(A_{22} - \widetilde{\lambda}_1 I_n)^{-1}(E_{21} - \widetilde{\lambda}_1 F_{21}) & I_n \end{bmatrix}.$$

Then

$$X^*(\widetilde{A} - \widetilde{\lambda}_1 \widetilde{B})X = \begin{bmatrix} M(\widetilde{\lambda}_1) & 0 \\ 0 & A_{22} - \widetilde{\lambda}_1 I_n \end{bmatrix},$$

where

$$M(\lambda) = A_{11} - \lambda I_m - (E_{21} - \lambda F_{21})^*(A_{22} - \lambda I_n)^{-1}(E_{21} - \lambda F_{21}).$$

Since $\widetilde{A} - \widetilde{\lambda}_1 \widetilde{B}$ and $X^*(\widetilde{A} - \widetilde{\lambda}_1 \widetilde{B})X$ have the same inertia, we see that $M(\widetilde{\lambda}_1)$ has zero as the largest eigenvalue. Notice that the largest eigenvalue of $A_{11} - \widetilde{\lambda}_1 I$ is $\lambda_1 - \widetilde{\lambda}_1 \le 0$. Thus, for $\delta_1 \overset{\text{def}}{=} |\widetilde{\lambda}_1 - \lambda_1| = \widetilde{\lambda}_1 - \lambda_1$, we have

$$\begin{aligned} \delta_1 = |0 - (\lambda_1 - \widetilde{\lambda}_1)| &\le \|(E_{21} - \widetilde{\lambda}_1 F_{21})^*(A_{22} - \widetilde{\lambda}_1 I_n)^{-1}(E_{21} - \widetilde{\lambda}_1 F_{21})\|_2 \\ &\le \|E_{21} - \widetilde{\lambda}_1 F_{21}\|_2^2 \|(A_{22} - \widetilde{\lambda}_1 I_n)^{-1}\|_2 \\ &\le \frac{\|E_{21} - \widetilde{\lambda}_1 F_{21}\|_2^2}{\delta_1 + \eta_1}, \end{aligned}$$

where we have used $[\|A_{22} - \widetilde{\lambda}_1 I_n)^{-1}\|_2]^{-1} = \widetilde{\lambda}_1 - \max_{\mu \in \mathrm{eig}(A_{22})} \mu = \delta_1 + \eta_1$. Consequently,

$$\delta_1 \le \frac{2\|E_{21} - \widetilde{\lambda}_1 F_{21}\|_2^2}{\eta_1 + \sqrt{\eta_1^2 + 4\|E_{21} - \widetilde{\lambda}_1 F_{21}\|_2^2}}$$

as asserted. Similarly, we can prove the result if the $(1,1)$th entry of A_{22} equals λ_1. In this case, we will apply the inertia arguments to $\widetilde{A} - \widetilde{\lambda}_1 \widetilde{B}$ and $Y(\widetilde{A} - \widetilde{\lambda}_1 \widetilde{B})Y^*$ with

$$Y = \begin{bmatrix} I_m & 0 \\ -(E_{21} - \widetilde{\lambda}_1 F_{21})(A_{11} - \widetilde{\lambda}_1 I_m)^{-1} & I_n \end{bmatrix}.$$

Applying the result of the last paragraph to $((-\widetilde{A}, \widetilde{B})$, we see that (8.17) holds for $i = N$.

Now, suppose $1 < i < N$. The result trivially holds if $\widetilde{\lambda}_i = \lambda_i$. Suppose $\widetilde{\lambda}_i \ne \lambda_i$. We may assume that $\lambda_i > \widetilde{\lambda}_i$. Otherwise, replace $\{(A,B), (\widetilde{A}, \widetilde{B}), i\}$ by $\{(-A, B), (-\widetilde{A}, \widetilde{B}), N - i + 1\}$. Delete the row and column of $\widetilde{A}$ that contain the diagonal entry λ_N and delete the corresponding row and column of $\widetilde{B}$ as well. Let $\widehat{A}$ and $\widehat{B}$ be the resulting matrices. Write the eigenvalues of the pair $(\widehat{A}, \widehat{B})$ as $\nu_1 \ge \cdots \ge \nu_{N-1}$. By the interlacing inequalities (2.28), we have

$$\widetilde{\lambda}_i \ge \nu_i \qquad \text{and hence} \qquad \lambda_i - \widetilde{\lambda}_i \le \lambda_i - \nu_i. \tag{8.18}$$

Note that λ_i is the ith largest diagonal entry of $\widehat{A}$. Let $\widehat{\eta}_i$ be the minimum distance between λ_i and the diagonal entries in the diagonal block $\widehat{A}_{jj}$ in $\widehat{A}$ not containing λ_i, where $j \in \{1,2\}$. Then

$$\widehat{\eta}_i \ge \eta_i$$

because $\widehat{A}_{jj}$ may have one fewer diagonal entries than A_{jj}. Let $\widehat{E}_{21}$ and $\widehat{F}_{21}$ be the off-diagonal block of $\widehat{A}$ and $\widehat{B}$, respectively. Then

$$\|\widehat{E}_{21} - \widetilde{\lambda}_i \widehat{F}_{21}\|_2 \le \|E_{21} - \widetilde{\lambda}_i F_{21}\|_2. \tag{8.19}$$

Finally, we have

$$\begin{aligned}
|\widetilde{\lambda}_i - \lambda_i| &= \lambda_i - \widetilde{\lambda}_i && \text{because } \lambda_i > \widetilde{\lambda}_i \\
&\le \lambda_i - \nu_i && \text{by (8.18)} \\
&\le \frac{2\|\widehat{E}_{21} - \widetilde{\lambda}_i \widehat{F}_{21}\|_2^2}{\widehat{\eta}_i + \sqrt{\widehat{\eta}_i^2 + 4\|\widehat{E}_{21} - \widetilde{\lambda}_i \widehat{F}_{21}\|_2^2}} && \text{by induction assumption} \\
&\le \frac{2\|\widehat{E}_{21} - \widetilde{\lambda}_i \widehat{F}_{21}\|_2^2}{\eta_i + \sqrt{\eta_i^2 + 4\|\widehat{E}_{21} - \widetilde{\lambda}_i \widehat{F}_{21}\|_2^2}} && \text{because } \widehat{\eta}_i \ge \eta_i \\
&= \frac{1}{2}\left(\sqrt{\eta_i^2 + 4\|\widehat{E}_{21} - \widetilde{\lambda}_i \widehat{F}_{21}\|_2^2} - \eta_i\right) \\
&\le \frac{1}{2}\left(\sqrt{\eta_i^2 + 4\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2} - \eta_i\right) && \text{by (8.19)}
\end{aligned}$$

$$= \frac{2\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2}{\eta_i + \sqrt{\eta_i^2 + 4\|E_{21} - \widetilde{\lambda}_i F_{21}\|_2^2}}$$

as expected. □

Method III. We now consider the following three eigenvalue problems:

$$\begin{aligned}
&\text{EIG (a)}: \quad (\widetilde{A}, \widetilde{B}) \text{ which has the same eigenvalues as } (\widetilde{B}^{-1/2}\widetilde{A}\widetilde{B}^{-1/2}, I_N),\\
&\text{EIG (b)}: \quad (\widetilde{A}, I_N),\\
&\text{EIG (c)}: \quad (A, I_N).
\end{aligned}$$

Denote the eigenvalues for EIG ($\mathtt{x}$) by $\lambda_j^{(\mathtt{x})}$ in descending order as in (8.9). Then we have $\lambda_j^{(\mathrm{a})} = \widetilde{\lambda}_j$ and $\lambda_j^{(\mathrm{c})} = \lambda_j$ for all j, recalling (8.3). Now we bound the eigenvalue bounds in two steps.

(a-b) There exist t_j $(1 \le j \le N)$ satisfying

$$1/\sigma_{\max}(\widetilde{B}) = \sigma_{\min}(\widetilde{B}^{-1}) \le t_j \le \sigma_{\max}(\widetilde{B}^{-1}) = 1/\sigma_{\min}(\widetilde{B})$$

such that

$$\lambda_j^{(\mathrm{a})} = t_j \lambda_j^{(\mathrm{b})} \quad (\text{or equivalently } \lambda_j^{(\mathrm{b})} = t_j^{-1}\lambda_j^{(\mathrm{a})}) \quad \text{for } 1 \le j \le N.$$

It can be seen that $1 - \sigma_{\max}(F_{21}) = \sigma_{\min}(\widetilde{B}) \le \sigma_{\max}(\widetilde{B}) = 1 + \sigma_{\max}(F_{21})$. Thus $|1 - t_j^{-1}| \le \|F_{21}\|_2$, which will be used later.

(b-c) For $1 \le j \le N$ we have

$$|\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}| \le \frac{2\|E_{21}\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|E_{21}\|_2^2}} \le \frac{2\|E_{21}\|_2^2}{\eta + \sqrt{\eta^2 + 4\|E_{21}\|_2^2}}.$$

Finally for $1 \le j \le N$,

$$\begin{aligned}
|\widetilde{\lambda}_j - \lambda_j| = |\lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{c})}| &= |\lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{b})} + \lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}|\\
&\le |1 - t_j^{-1}|\,|\lambda_j^{(\mathrm{a})}| + |\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}|\\
&\le \|F_{21}\|_2\,|\widetilde{\lambda}_j| + \frac{2\|E_{21}\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|E_{21}\|_2^2}}. \qquad (8.20)
\end{aligned}$$

Remark 5. The derivation here is the shortest (and simplest) among the three methods that lead to (8.10) and (8.11), (8.17), and (8.20), but not without a sacrifice, namely, it is likely the weakest when $\|F_{21}\|_2$ has a much bigger magnitude than $\|E_{21}\|_2^2$ because $\|F_{21}\|_2$ appears linearly in (8.20) *vs.* quadratically in (8.10), (8.11), and (8.17). Note the similarity among the third term in the right-hand side of (8.10), the second and last terms in (8.17), and the second term in the right-hand side of (8.20).

Theorem 8.9. *Under the conditions of* Theorem 8.6*, we have* (8.20) *for all* $1 \le j \le N$.

We point out in passing that Theorems 8.6, 8.8, and 8.9 all reduce to the main result in [**97**].

8.3.2. General Case. We are now looking into the general case, i.e., without assuming (8.5).

LEMMA 8.2. *Let Δ be a Hermitian matrix. If $\delta \overset{\text{def}}{=} \|\Delta\|_2 < 1$, then $I+\Delta$ is positive definite, and*

$$\|(I+\Delta)^{-1/2} - I\|_2 \le \frac{1}{\sqrt{1-\delta}} - 1 = \frac{\delta}{\sqrt{1-\delta}\,(1+\sqrt{1-\delta})}.$$

PROOF. Any eigenvalue of $I+\Delta$ is no smaller than $1-\delta > 0$, so $I+\Delta$ is positive definite. We have

$$\begin{aligned}\|(I+\Delta)^{-1/2} - I\|_2 &= \max_{\mu\in\mathrm{eig}(\Delta)} |(1+\mu)^{-1/2} - 1| \\ &\le \max\left\{(1-\delta)^{-1/2} - 1, 1 - (1+\delta)^{-1/2}\right\} \\ &= (1-\delta)^{-1/2} - 1,\end{aligned}$$

as was to be shown. □

Recall that $A, \widetilde{A}, B, \widetilde{B}, E$ and F are all Hermitian. Set

$$\Delta_{ij} = B_{ii}^{-1/2} F_{ij} B_{jj}^{-1/2}, \tag{8.21a}$$

$$Y = \mathrm{diag}([I+\Delta_{11}]^{-1/2} B_{11}^{-1/2},\, [I+\Delta_{22}]^{-1/2} B_{22}^{-1/2}), \tag{8.21b}$$

$$\widehat{F}_{ij} = [I+\Delta_{ii}]^{-1/2} \Delta_{ij} [I+\Delta_{jj}]^{-1/2} \quad \text{for } i \ne j, \tag{8.21c}$$

$$\widehat{A}_{ii} = B_{ii}^{-1/2} A_{ii} B_{ii}^{-1/2}, \tag{8.21d}$$

$$\widehat{E}_{ij} = [I+\Delta_{ii}]^{-1/2} B_{ii}^{-1/2} E_{ij} B_{jj}^{-1/2} [I+\Delta_{jj}]^{-1/2} \quad \text{for } i \ne j, \tag{8.21e}$$

$$\begin{aligned}\widehat{E}_{ii} &= [I+\Delta_{ii}]^{-1/2} B_{ii}^{-1/2} (A_{ii}+E_{ii}) B_{ii}^{-1/2} [I+\Delta_{ii}]^{-1/2} - \widehat{A}_{ii} \\ &= \left([I+\Delta_{ii}]^{-1/2} - I\right) \widehat{A}_{ii} \left([I+\Delta_{ii}]^{-1/2} - I\right) \\ &\quad + \widehat{A}_{ii} \left([I+\Delta_{ii}]^{-1/2} - I\right) + \left([I+\Delta_{ii}]^{-1/2} - I\right) \widehat{A}_{ii} \\ &\quad + [I+\Delta_{ii}]^{-1/2} B_{ii}^{-1/2} E_{ii} B_{ii}^{-1/2} [I+\Delta_{ii}]^{-1/2},\end{aligned} \tag{8.21f, 8.21g}$$

and set

$$\delta_{ij} = \|\Delta_{ij}\|_2 \le \sqrt{\|B_{ii}^{-1}\|_2 \|B_{jj}^{-1}\|_2}\, \|F_{ij}\|_2, \quad \gamma_{ij} = (1-\delta_{ij})^{-1/2} - 1. \tag{8.22}$$

In obtaining (8.22), we have used $\|B_{ii}^{-1/2}\|_2 = \sqrt{\|B_{ii}^{-1}\|_2}$. To ensure that $\widetilde{B}$ is positive definite, throughout this subsection we assume that

$$\max\{\delta_{11}, \delta_{22}\} < 1, \quad \delta_{12}^2 < (1-\delta_{11})(1-\delta_{22}). \tag{8.23}$$

We can bound $\widehat{E}_{ij}$ and $\widehat{F}_{ij}$ as follows.

$$\|\widehat{E}_{ij}\|_2 \le \frac{\|B_{ii}^{-1/2} E_{ij} B_{jj}^{-1/2}\|_2}{\sqrt{(1-\delta_{ii})(1-\delta_{jj})}} \tag{8.24a}$$

$$\leq \sqrt{\frac{\|B_{ii}^{-1}\|_2\|B_{jj}^{-1}\|_2}{(1-\delta_{ii})(1-\delta_{jj})}}\|E_{ij}\|_2 \quad \text{for } i \neq j, \tag{8.24b}$$

$$\|\widehat{E}_{ii}\|_2 \leq \gamma_{ii}(2+\gamma_{ii})\|\widehat{A}_{ii}\|_2 + \frac{\|B_{ii}^{-1/2}E_{ii}B_{ii}^{-1/2}\|_2}{1-\delta_{ii}} \tag{8.24c}$$

$$\leq \gamma_{ii}(2+\gamma_{ii})\|\widehat{A}_{ii}\|_2 + \frac{\|B_{ii}^{-1}\|_2}{1-\delta_{ii}}\,\|E_{ii}\|_2, \tag{8.24d}$$

$$\|\widehat{F}_{ij}\|_2 \leq \frac{\delta_{ij}}{\sqrt{(1-\delta_{ii})(1-\delta_{jj})}} \quad \text{for } i \neq j. \tag{8.24e}$$

We have used Lemma 8.2 to get (8.24c) from (8.21g). We then have

$$\widehat{A} \stackrel{\text{def}}{=} Y^*\widetilde{A}Y = \begin{bmatrix} \widehat{A}_{11}+\widehat{E}_{11} & \widehat{E}_{12} \\ \widehat{E}_{21} & \widehat{A}_{22}+\widehat{E}_{22} \end{bmatrix}, \; \widehat{B} \stackrel{\text{def}}{=} Y^*\widetilde{B}Y = \begin{bmatrix} I_m & \widehat{F}_{12} \\ \widehat{F}_{21} & I_n \end{bmatrix}.$$

We now consider the following three eigenvalue problems:

$$\begin{aligned} &\text{EIG (a)}: \quad (\widetilde{A},\widetilde{B}) \text{ which is equivalent to } (\widehat{A},\widehat{B}), \\ &\text{EIG (b)}: \quad \left(\begin{bmatrix} \widehat{A}_{11} & \widehat{E}_{12} \\ \widehat{E}_{21} & \widehat{A}_{22} \end{bmatrix}, \begin{bmatrix} I_m & \widehat{F}_{12} \\ \widehat{F}_{21} & I_n \end{bmatrix}\right), \\ &\text{EIG (c)}: \quad \left(\begin{bmatrix} \widehat{A}_{11} & \\ & \widehat{A}_{22} \end{bmatrix}, I_N\right). \end{aligned}$$

Denote the eigenvalues for EIG ($\mathtt{x}$) by $\lambda_j^{(\mathtt{x})}$ in descending order as in (8.9). Then we have $\lambda_j^{(\mathrm{a})} = \widetilde{\lambda}_j$ and $\lambda_j^{(\mathrm{c})} = \lambda_j$ for all j, recalling (8.3). Note $\|\widehat{F}_{21}\|_2 < 1$ because of (8.23). The eigenvalue differences between any two adjacent eigenvalue problems in the above list can be bounded as follows.

(a-b) $|\lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{b})}| \leq \frac{\max_i \|\widehat{E}_{ii}\|_2}{1-\|\widehat{F}_{21}\|_2}$ for $1 \leq j \leq N$. This is because

$$\widehat{B}^{\,-1/2}\,\widehat{A}\,\widehat{B}^{\,-1/2} = \widehat{B}^{\,-1/2}\begin{bmatrix} \widehat{A}_{11} & \widehat{E}_{12} \\ \widehat{E}_{21} & \widehat{A}_{22} \end{bmatrix}\widehat{B}^{\,-1/2} + \widehat{B}^{\,-1/2}\begin{bmatrix} \widehat{E}_{11} & \\ & \widehat{E}_{22} \end{bmatrix}\widehat{B}^{\,-1/2},$$

and $\|\widehat{B}^{-1}\|_2 = \left[1-\|\widehat{F}_{21}\|_2\right]^{-1}$.

(b-c) Use Theorems 8.6, 8.8, or 8.9 to bound $\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}$ to yield three bounds:

$$\begin{aligned} |\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}| \leq & \|\widehat{F}_{21}\|_2^2\,|\widetilde{\lambda}_j| + \|\widehat{E}_{21}\widehat{F}_{21}^* + \widehat{F}_{21}\widehat{E}_{21}^* - \widehat{F}_{21}\widehat{A}_{11}\widehat{F}_{21}^*\|_2 \\ & + \frac{2\|\widehat{E}_{21} - \widehat{F}_{21}\widehat{A}_{11}\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|\widehat{E}_{21} - \widehat{F}_{21}\widehat{A}_{11}\|_2^2}}, \end{aligned} \tag{8.25}$$

$$|\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}| \leq \frac{2\|\widehat{E}_{21} - \widetilde{\lambda}_j\widehat{F}_{21}\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|\widehat{E}_{21} - \widetilde{\lambda}_j\widehat{F}_{21}\|_2^2}}, \tag{8.26}$$

$$|\lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}| \le \|\widehat{F}_{21}\|_2 \, |\widetilde{\lambda}_j| + \frac{2\|\widehat{E}_{21}\|_2^2}{\eta_j + \sqrt{\eta_j^2 + 4\|\widehat{E}_{21}\|_2^2}}. \tag{8.27}$$

Further bounds in terms of the norms of E, F, and B_{ii} can be obtained by using inequalities in (8.24). Finally we use $\widetilde{\lambda}_j - \lambda_j = \lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{c})} = \lambda_j^{(\mathrm{a})} - \lambda_j^{(\mathrm{b})} + \lambda_j^{(\mathrm{b})} - \lambda_j^{(\mathrm{c})}$ to conclude

THEOREM 8.10. *For the Hermitian definite pairs (A, B) and $(\widetilde{A}, \widetilde{B})$ with A, B, $\widetilde{A}$, and $\widetilde{B}$ as in* (8.1) *and* (8.2)*, assume* (8.23)*, where δ_{ij} are as defined in* (8.21) *and* (8.22)*. Denote their eigenvalues as in* (8.3) *and define gaps η_i and η as in* (8.4)*. Then for all $1 \le j \le N$*

$$|\widetilde{\lambda}_j - \lambda_j| \le \frac{\max_i \|\widehat{E}_{ii}\|_2}{1 - \|\widehat{F}_{21}\|_2} + \epsilon_j,$$

where ϵ_j can be taken to be any one of the right-hand sides of (8.25)*,* (8.26)*, and* (8.27)*.*

Next we estimate the differences between the eigenvalues of (A_{11}, B_{11}) (which is the same as those of $\widehat{A}_{11}$) and some m eigenvalues of $(\widetilde{A}, \widetilde{B})$. This will be done in two steps:

(a) bound the differences between the eigenvalues of $\widehat{A}_{11} + \widehat{E}_{11}$ and m of those of $(\widetilde{A}, \widetilde{B})$;
(b) bound the differences between the eigenvalues of $\widehat{A}_{11} + \widehat{E}_{11}$ and those of $\widehat{A}_{11}$.

The first step can be accomplished by using Theorem 8.7, while the second step can be done by invoking Weyl's theorem. We thereby get

THEOREM 8.11. *Assume the conditions of* Theorem 8.10*. There are m eigenvalues $\mu_1 \ge \cdots \ge \mu_m$ of $(\widetilde{A}, \widetilde{B})$ such that*

$$|\mu_j - \theta_j| \le \|\widehat{E}_{11}\|_2 + \frac{\|\widehat{E}_{21} - \widehat{F}_{21}(\widehat{A}_{11} + \widehat{E}_{11})\|_2}{\sqrt{1 - \|\widehat{F}_{21}\|_2^2}} \quad \textit{for } 1 \le j \le m,$$

$$\sqrt{\sum_{j=1}^{m} |\mu_j - \theta_j|^2} \le \|\widehat{E}_{11}\|_{\mathrm{F}} + \frac{\|\widehat{E}_{21} - \widehat{F}_{21}(\widehat{A}_{11} + \widehat{E}_{11})\|_{\mathrm{F}}}{\sqrt{1 - \|\widehat{F}_{21}\|_2^2}},$$

where $\theta_1 \ge \cdots \ge \theta_m$ are the m eigenvalues of (A_{11}, B_{11}).

REMARK 6. Some comments for comparing Theorems 8.10 and Theorem 8.11 are in order:

(a) Theorem 8.10 bounds the changes in all the eigenvalues of (A, B), while Theorem 8.11 bounds only a portion of them, namely those of (A_{11}, B_{11}).
(b) $\widehat{E}_{22}$ does not appear in the bounds in Theorem 8.11, while both $\widehat{E}_{ii}$ show up in those in Theorem 8.10. This could potentially make the bounds in Theorem 8.11 more favorable if $\widehat{E}_{22}$ has much larger magnitude than $\widehat{E}_{11}$. This point will be well manifested in our analysis later in Section 9.2. Also, note that the quantities $\|\widehat{E}_{ij}\|$ are relative quantities as opposed to absolute quantities $\|E_{ij}\|$, because $\widehat{E}_{ij}$ has been multiplied by $B_{ii}^{-1/2}$ and $B_{jj}^{-1/2}$.

(c) Except when ϵ_j is taken to be the right-hand side of (8.27), bounds in Theorem 8.10 are of quadratic order in $\widehat{E}_{21}$ and $\widehat{F}_{21}$, while those in Theorem 8.11 are of linear order in $\widehat{E}_{21}$ and $\widehat{F}_{21}$. All bounds are of linear order in $\widehat{E}_{ii}$ and $\widehat{F}_{ii}$. Thus when $\widehat{E}_{ii} = F_{ii} = 0$ for $i = 1, 2$, Theorem 8.10 may lead to sharper bounds.

(d) Theorem 8.10 requires some gap information among the eigenvalues of (A_{ii}, B_{ii}) for $i = 1, 2$, while Theorem 8.11 does not.

EXAMPLE 8.2. To illustrate these comments in Remark 6, we consider the following parameterized pair

$$(\widetilde{A}(\alpha), \widetilde{B}(\alpha)) \equiv (A + \alpha E, B + \alpha F),$$

where α is a parameter ranging from 0 to 1, $A = \text{diag}(4, 1)$, and $B = \text{diag}(2, 1)$. Two types of perturbations E and F will be considered: the general dense perturbations

$$E = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}, \quad F = \frac{1}{10} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \tag{8.28}$$

and the off-diagonal perturbations

$$E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad F = \frac{1}{10} \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}. \tag{8.29}$$

Denote by $\widetilde{\lambda}_j(\alpha)$ the jth largest eigenvalue of $(\widetilde{A}(\alpha), \widetilde{B}(\alpha))$. Here we take $j = 1$, so $\lambda_1(0) = \lambda_1 = 2$. Figure 8.3.1 shows log-log scale plots for the actual $|\widetilde{\lambda}_1(\alpha) - \lambda_1|$, its bound by Theorem 8.11, and the three bounds by Theorem 8.10 corresponding to ϵ_j being the right-hand sides of (8.25), (8.26), and (8.27), respectively. These three bounds are shown as "Thm 8.10(i)", "Thm 8.10(ii)", and "Thm 8.10(iii)" in the plots. We assumed the gap $\eta_1 = 1$ is known.

In the left plot of Figure 8.3.1 we plot only one curve for the three bounds by Theorem 8.10 because they are visually indistinguishable. The figure illustrates the first two comments. First, the bound by Theorem 8.11 becomes smaller than $|\widetilde{\lambda}_1(\alpha) - \lambda_1|$ for $\alpha \gtrapprox 0.25$. This is not an error but because the bound by Theorem 8.11 is for the distance between $\lambda_1 = 2$ and *an* eigenvalue of $(\widetilde{A}(\alpha), \widetilde{B}(\alpha))$, which may not necessarily be $\widetilde{\lambda}_1(\alpha)$. In fact, for $\alpha \gtrapprox 0.25$ the eigenvalue of $(\widetilde{A}(\alpha), \widetilde{B}(\alpha))$ closer to 2 is $\widetilde{\lambda}_2(\alpha)$. Second, Theorem 8.11 gives a smaller bound than Theorem 8.10, reflecting the fact that E_{22} is much larger than E_{11}.

The right plot of Figure 8.3.1 illustrates the third comment. Specifically, the first two bounds by Theorem 8.10 are much smaller than the other bounds. They reflect the quadratic scaling of $|\widetilde{\lambda}_1(\alpha) - \lambda_1|$ under off-diagonal perturbations, as can be seen by the slope of the plots. □

We now specialize the results so far in this subsection to the case

$$E_{ii} = F_{ii} = 0 \quad \text{for } i = 1, 2. \tag{8.30}$$

This corresponds to a practical situation in eigenvalue computations: what is the effect of dropping off off-diagonal blocks with relatively small magnitudes? Assume (8.30), then

$$\Delta_{ii} = 0, \ \widehat{E}_{ii} = 0, \ \gamma_{ii} = \delta_{ii} = 0, \ \text{for } i = 1, 2,$$

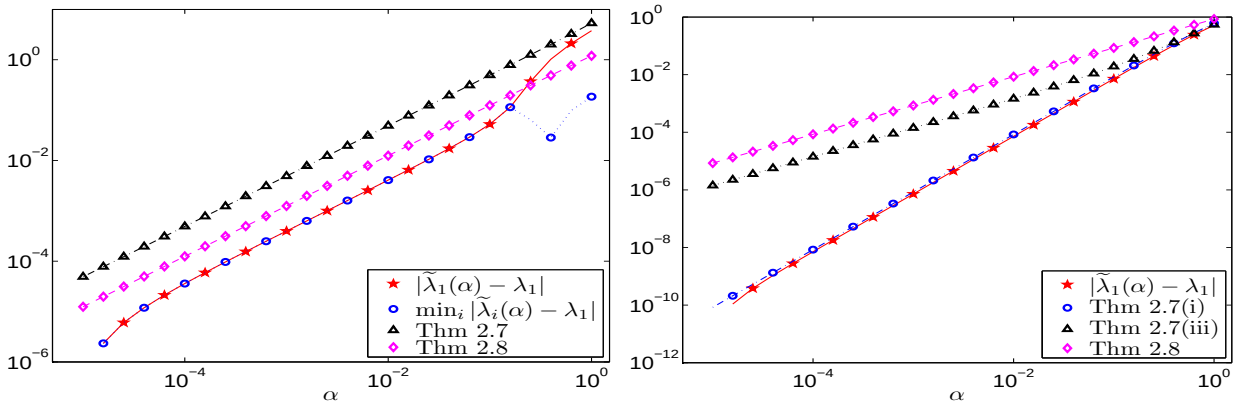

FIGURE 8.3.1. $|\widetilde{\lambda}_1(\alpha) - \lambda_1|$ and its bounds by Theorem 8.10 and 8.11. **Left:** Under perturbation (8.28), the curves for the three bounds by Theorem 8.10 are indistinguishable and the bound by Theorem 8.11 is smaller. It is interesting to notice that the curve for $|\widetilde{\lambda}_1(\alpha) - \lambda_1|$ crosses above the bound by Theorem 8.11 for α about 0.25 or larger. This is because the bound by Theorem 8.11 is actually for $\min_i |\widetilde{\lambda}_i(\alpha) - \lambda_1|$. **Right:** Under perturbation (8.29), the curve for Thm 8.10(ii) is the same as for $|\widetilde{\lambda}_1(\alpha) - \lambda_1|$, and the bound by Theorem 8.11 is larger.

$$\widehat{E}_{21} = B_{22}^{-1/2} E_{21} B_{11}^{-1/2}, \ \widehat{F}_{21} = B_{22}^{-1/2} F_{21} B_{11}^{-1/2}. \tag{8.31}$$

Theorem 8.10 yields

COROLLARY 8.3. *To the conditions of* Theorem 8.10 *add* $E_{ii} = F_{ii} = 0$ *for* $i = 1, 2$. *Let* $\widehat{E}_{21}$ *and* $\widehat{F}_{21}$ *be given as in* (8.31) *and assume* $\|\widehat{F}_{21}\|_2 < 1$. *Then for all* $1 \le j \le N$

$$|\widetilde{\lambda}_j - \lambda_j| \le \epsilon_j, \tag{8.32}$$

where ϵ_j *can be taken to be any one of the right-hand sides of* (8.25), (8.26), *and* (8.27).

At the same time Theorem 8.11 gives

COROLLARY 8.4. *Assume the conditions of* Corollary 8.3. *There are* m *eigenvalues* $\mu_1 \ge \cdots \ge \mu_m$ *of* $\widetilde{A} - \lambda \widetilde{B}$ *such that*

$$|\mu_j - \theta_j| \le \frac{\|\widehat{E}_{21} - \widehat{F}_{21}\widehat{A}_{11}\|_2}{\sqrt{1 - \|\widehat{F}_{21}\|_2^2}} \quad \text{for } 1 \le j \le m,$$

$$\sqrt{\sum_{j=1}^{m} |\mu_j - \theta_j|^2} \le \frac{\|\widehat{E}_{21} - \widehat{F}_{21}\widehat{A}_{11}\|_{\mathrm{F}}}{\sqrt{1 - \|\widehat{F}_{21}\|_2^2}},$$

where $\theta_1 \ge \cdots \ge \theta_m$ *are the* m *eigenvalues of the pair* (A_{11}, B_{11}).

8.4. An extension to non-Hermitian pairs

In this section we will make an attempt to derive a quadratic eigenvalue bound for diagonalizable non-Hermitian pairs subject to off-diagonal perturbations. Specifically, let

$$A = \begin{matrix} {\scriptstyle m}\{ \\ {\scriptstyle n}\{ \end{matrix} \overbrace{\begin{bmatrix} A_{11} & \\ & A_{22} \end{bmatrix}}^{\scriptstyle m \quad \scriptstyle n}, \quad B = \begin{matrix} {\scriptstyle m}\{ \\ {\scriptstyle n}\{ \end{matrix} \overbrace{\begin{bmatrix} B_{11} & \\ & B_{22} \end{bmatrix}}^{\scriptstyle m \quad \scriptstyle n}, \tag{8.33a}$$

$$\widetilde{A} = \begin{bmatrix} A_{11} & E_{12} \\ E_{21} & A_{22} \end{bmatrix}, \qquad \widetilde{B} = \begin{bmatrix} B_{11} & F_{12} \\ F_{21} & B_{22} \end{bmatrix} \tag{8.33b}$$

be non-Hermitian matrices. We assume that B is nonsingular and (A, B) is diagonalizable. So (A, B) has only finite eigenvalues, and there exist nonsingular matrices $X = \operatorname{diag}(X_1, X_2)$ and $Y = \operatorname{diag}(Y_1, Y_2)$ such that $YAX = \Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2)$ and $YBX = I$, where X_1, Y_1 and Λ_1 are m-by-m and Λ is the diagonal matrix of eigenvalues. The last assumption loses little generality, since if (A, B) is regular and diagonalizable but B is singular (hence infinite eigenvalues exist), we can apply the results below to $(A, B - \alpha A)$ for a suitable scalar α such that $B - \alpha A$ is nonsingular (the regularity assumption of (A, B) ensures the existence of such α). The eigenvalues ν of $(A, B - \alpha A)$ and τ of (A, B) are related by $\nu = \tau/(1 - \alpha\tau)$.

We will establish a bound on $|\mu - \widetilde{\mu}|$, where μ is an eigenvalue of (A, B) and $\widetilde{\mu}$ is an eigenvalue of $(\widetilde{A}, \widetilde{B})$.

THEOREM 8.12. *Let $A, B, \widetilde{A}, \widetilde{B}$ be as in* (8.33a) *and* (8.33b). *Suppose that there exist nonsingular matrices $X = diag(X_1, X_2)$ and $Y = diag(Y_1, Y_2)$ such that $YAX = \Lambda$ is diagonal and $YBX = I$. If $\widetilde{\mu}$ is an eigenvalue of $(\widetilde{A}, \widetilde{B})$ such that*

$$\eta_k \stackrel{\text{def}}{=} \min_{\mu \in \operatorname{eig}(A_{kk}, B_{kk})} |\widetilde{\mu} - \mu| > 0$$

for $k = 1$ or 2, then (A, B) has an eigenvalue μ such that

$$|\widetilde{\mu} - \mu| \le \|X\|_2 \|Y\|_2 \|E_{12} - \widetilde{\mu} F_{12}\|_2 \|E_{21} - \widetilde{\mu} F_{21}\|_2 \|(A_{kk} - \widetilde{\mu} B_{kk})^{-1}\|_2 \tag{8.34a}$$

$$\le \frac{\kappa_2(X)\kappa_2(Y)\|E_{12} - \widetilde{\mu} F_{12}\|_2 \|E_{21} - \widetilde{\mu} F_{21}\|_2}{\eta_k}. \tag{8.34b}$$

PROOF. We prove the result only for $k = 2$. The proof for $k = 1$ is entirely analogous.

Suppose that $\widetilde{\mu} \notin \operatorname{eig}(A_{22}, B_{22})$ is an eigenvalue of $(\widetilde{A}, \widetilde{B})$. Thus $\widetilde{A} - \widetilde{\mu}\widetilde{B}$ is singular. Defining the nonsingular matrices

$$W_L = \begin{bmatrix} I & -(E_{12} - \widetilde{\mu} F_{12})(A_{22} - \widetilde{\mu} B_{22})^{-1} \\ 0 & I \end{bmatrix},$$

$$W_R = \begin{bmatrix} I & 0 \\ -(A_{22} - \widetilde{\mu} B_{22})^{-1}(E_{21} - \widetilde{\mu} F_{21}) & I \end{bmatrix},$$

we see that

$$\operatorname{diag}(Y_1, I_n) W_L (\widetilde{A} - \widetilde{\mu}\widetilde{B}) W_R \operatorname{diag}(X_1, I_n)$$

$$= \operatorname{diag}(Y_1, I_n) \begin{bmatrix} A_{11} - \widetilde{\mu} B_{11} - (E_{12} - \widetilde{\mu} F_{12})(A_{22} - \widetilde{\mu} B_{22})^{-1}(E_{21} - \widetilde{\mu} F_{21}) & 0 \\ 0 & A_{22} - \widetilde{\mu} B_{22} \end{bmatrix}$$
$$\times \operatorname{diag}(X_1, I_n)$$
$$= \begin{bmatrix} \Lambda_1 - \widetilde{\mu} I_m - Y_1(E_{12} - \widetilde{\mu} F_{12})(A_{22} - \widetilde{\mu} B_{22})^{-1}(E_{21} - \widetilde{\mu} F_{21}) X_1 & 0 \\ 0 & A_{22} - \widetilde{\mu} B_{22} \end{bmatrix}.$$

Since this matrix is also singular, it follows that $A_{11} - \lambda B_{11}$ must have an eigenvalue μ that satisfies

$$\begin{aligned} |\widetilde{\mu} - \mu| &\le \|Y_1\|_2 \|(E_{12} - \widetilde{\mu} F_{12})(A_{22} - \widetilde{\mu} B_{22})^{-1}(E_{21} - \widetilde{\mu} F_{21})\|_2 \|X_1\|_2 \\ &\le \frac{1}{\eta_k} \|X_1\|_2 \|X_2^{-1}\|_2 \|Y_1\|_2 \|Y_2^{-1}\|_2 \|E_{12} - \widetilde{\mu} F_{12}\|_2 \|E_{21} - \widetilde{\mu} F_{21}\|_2, \end{aligned}$$

where we have used

$$\|(A_{22} - \widetilde{\mu} B_{22})^{-1}\|_2 = \|X_2^{-1}(\Lambda_2 - \widetilde{\mu} I)^{-1} Y_2^{-1}\|_2 \le \|X_2^{-1}\|_2 \|Y_2^{-1}\|_2 / \eta_k.$$

Now use $\|X\|_2 = \max\{\|X_1\|_2, \|X_2\|_2\}$ and $\|X^{-1}\|_2 = \max\{\|X_1^{-1}\|_2, \|X_2^{-1}\|_2\}$ to get (8.34b) for the case $k = 2$. □

When the pairs (A, B) and $(\widetilde{A}, \widetilde{B})$ are Hermitian definite and (8.5) holds, we have $\kappa_2(X) = \kappa_2(Y) = 1$; so (8.34b) reduces to $|\widetilde{\mu} - \mu| \le \|E_{12} - \widetilde{\mu} F_{12}\|_2^2 / \eta_k$. This is similar to our earlier result (8.17), except for the slight difference in the denominator. If we further let $F_{12} = 0$, then the expression (8.34b) becomes exactly that of the quadratic residual bound in [**109**] for Hermitian matrices. However (8.34b) does not give a one-to-one pairing between the eigenvalues of (A, B) and $(\widetilde{A}, \widetilde{B})$.

The assumption that $\eta_k > 0$ is a reasonable one when there is a gap between $\operatorname{eig}(A_{kk}, B_{kk})$ for $k = 1, 2$ because then it is reasonable to expect that $\widetilde{\mu}$ is very near one of them but away from the other.

CHAPTER 9

Perturbation and condition numbers of a multiple generalized eigenvalue

This chapter is concerned with the perturbation behavior of a multiple eigenvalue in generalized eigenvalue problems. In particular, we address a question raised by Stewart and Sun, which claims a multiple generalized eigenvalue has different sensitivities under perturbations. We solve this problem in two ways. First we derive k different eigenvalue perturbation bounds for a multiple generalized eigenvalue when the matrices undergo perturbation of finite norms. Second, we consider the asymptotic perturbation behavior and show that a multiple generalized eigenvalue has multiple condition numbers, even in the Hermitian definite case. We also show how this difference of sensitivities plays a role in the eigenvalue forward error analysis after the Rayleigh-Ritz process, for which we present an approach that provides tight bounds.

Introduction. This chapter is motivated by a question raised in [**146**, p.300], where it was pointed out that multiple eigenvalues of a Hermitian positive definite matrix pair (A, B) (A, B are Hermitian and B is positive definite) tend to behave differently under perturbations. Specifically, they consider pairs (A, B) such as

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 20000 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 10000 \end{pmatrix}, \tag{9.1}$$

which has a multiple eigenvalue $\lambda = 2$ of multiplicity 2. Interestingly, one of the two eigenvalues react more sensitively to perturbations than the other does. For example, define the perturbed pair by $(A + E, B + F)$ where E and F are random Hermitian perturbation matrices with $||E||_2, ||F||_2 \leq 10^{-2}$. We tested for 10^4 such random matrices (details in Section 9.1.4), and observed that one eigenvalue of $(A+E, B+F)$ always lies in the interval $[2 - 1.6 \times 10^{-4}, 2 + 1.6 \times 10^{-4}]$, while the other can be more perturbed and lies in $[2 - 2.0 \times 10^{-2}, 2 + 2.0 \times 10^{-2}]$.

There seems to be a clear sensitivity difference between the two multiple eigenvalues, but a theoretical explanation for this behavior has remained an open problem. For example, a general Weyl-type perturbation result [**113**] applied to (9.1) only gives $|\lambda - 2| \leq 3.03 \times 10^{-2}$, which is a tight bound for the more sensitive eigenvalue, but does not tell anything about the insensitive one.

The purpose of this chapter is to provide reasonings for this behavior. In the first part we consider finite perturbation, that is, we derive perturbation bounds in terms of $\|E\|_2$ and $\|F\|_2$. Theorem 9.1, gives k different perturbation bounds for a multiple eigenvalue of a

Hermitian positive definite pair (A, B) of multiplicity k. Applying the theorem to the pair (9.1) allows us to explain theoretically that the two eigenvalues have different perturbation behaviors.

The fact that multiple eigenvalues react differently to perturbations has several practical consequences. One example we discuss here is the the forward error analysis of computed multiple eigenvalues (Ritz values) obtained by the Rayleigh-Ritz process. We show how the different sensitivities of a multiple eigenvalue plays a role here, and present an approach that yields k different error bounds for a multiple eigenvalue of multiplicity k.

In the second part of this chapter we consider the condition numbers of a multiple generalized eigenvalue, that is, the behavior in the limit $E, F \to 0$. For standard eigenvalue problems, a closed-form expression for the condition numbers of a multiple eigenvalue is known. In particular, they are uniformly 1 in the Hermitian case, and generally take different values in the non-Hermitian case. We consider the generalized eigenvalue problem and identify the condition numbers of a multiple eigenvalue. Our main result is that a multiple eigenvalue generally has multiple condition numbers, even in the Hermitian definite case. The condition numbers are characterized in terms of the singular values of the outer product of the corresponding left and right eigenvectors.

The first order perturbation expansion for a simple eigenvalue is a well-known result [**145, 146**], and that for a multiple eigenvalue is also studied in [**150**] for standard eigenvalue problems, and in [**30**] for generalized eigenvalue problems, including singular matrix pairs. Using such results and following the definition of r condition numbers of a multiple eigenvalue for standard eigenvalue problems introduced in [**151**], we naturally define condition numbers for the generalized case, as shown below in (9.37).

Sun shows in [**151**] that the r condition numbers $\kappa_i(A, \lambda_0)$ for $i = 1, \ldots, r$ of a nondefective multiple eigenvalue of a non-Hermitian matrix are expressed by

$$\kappa_i(A, \lambda_0) = \left(\prod_{j=1}^{i} c_j(A, \lambda_0) \right)^{1/i} \quad \text{for} \quad i = 1, \ldots, r, \tag{9.2}$$

where $c_j(A, \lambda_0)$ are the secants of the canonical angles between the left and right invariant subspaces corresponding to the multiple eigenvalue. When A is non-Hermitian $c_i(A, \lambda_0)$ generally take different values for different i, hence so do $\kappa_i(A, \lambda_0)$ and (9.2) shows that a multiple eigenvalue has multiple condition numbers. Contrarily, in the Hermitian case we have $c_j(A, \lambda_0) \equiv 1$, so $\kappa_i(A, \lambda_0) = 1$ for $i = 1, \ldots, r$. Hence (9.2) also shows the well-known fact that a multiple eigenvalue of a Hermitian matrix always has a uniform condition number 1.

Since a standard non-Hermitian eigenvalue problem can be regarded as a special case of the generalized non-Hermitian eigenvalue problem $Ax = \lambda Bx$, it is clear that a multiple eigenvalue in this case must have multiple condition numbers. On the other hand, in the important case of the generalized Hermitian definite pair (where A and B are Hermitian and B is positive definite), it is not trivial whether or not the condition numbers $\kappa_i(A, B, \lambda_0)$ for $i = 1, \ldots, r$ take different values. In this chapter we identify their expression, which shows that they generally do take r different values. We shall see that there are two sources for this different conditioning, namely the difference between the left and right eigenvectors

(as present in non-Hermitian standard eigenproblems), and the fact that the B-orthonormal eigenvectors have different 2-norms (as present in the case $B \neq I$).

It is important to note that in the Hermitian definite case, a natural choice of metric can be based on the B-based inner product $(x, y)_B = x^*By$ instead of the standard inner product $(x, y) = x^*y$. This leads to the standard eigenvalue problem for the Hermitian matrix $B^{-1/2}AB^{-1/2}$, so in this inner product all the condition numbers are the same (see the second remark in Section 9.5.2). Our discussion in this chapter assumes the use of the standard inner product.

The structure of this chapter is as follows. In Section 9.1 we present k different perturbation bounds for a multiple eigenvalue of a Hermitian positive definite pair. In Section 9.2 we use the results in the previous Chapter to give another explanation. In Section 9.3 we describe an approach that provides refined forward error bounds for a computed multiple eigenvalue after the Rayleigh-Ritz process. Section 9.4 starts the discussion of condition numbers. We present condition numbers of a multiple generalized eigenvalue in both the Hermitian definite and the non-Hermitian cases.

9.1. Perturbation bounds for multiple generalized eigenvalues

Suppose a Hermitian positive definite pair (A, B) has a multiple eigenvalue λ_0 of multiplicity k. In this section we consider a perturbed Hermitian positive definite pair $(A + E, B + F)$. We are interested in the eigenvalues of $(A + E, B + F)$ that are close to λ_0: how the multiple eigenvalue λ_0 is perturbed. Our goal is to give k different perturbation bounds, that is, to derive $0 \leq b_1 \leq b_2 \leq \cdots \leq b_k$ such that there are at least i eigenvalues of $(A + E, B + F)$ that lie in $[\lambda_0 - b_i, \lambda_0 + b_i]$.

9.1.1. Choosing eigenvectors. Since (A, B) is a Hermitian positive definite pair with a multiple eigenvalue λ_0 of multiplicity k, there exists a nonsingular matrix X such that

$$X^*AX = \Lambda = \mathrm{diag}(\lambda_0, \cdots, \lambda_0, \lambda_{k+1}, \cdots, \lambda_n), \quad X^*BX = I_n, \tag{9.3}$$

where $\lambda_{k+i} \neq \lambda_0$ for $i \geq 1$. The columns of X are the right eigenvectors of (A, B), and the first k columns are the eigenvectors corresponding to the multiple eigenvalue λ_0. It is important to note that X is not unique, in that there is freedom of unitary transformations in the first k columns of X. Specifically, for any unitary matrix $Q \in \mathbb{C}^{k\times k}$, X can be replaced by $X \cdot \mathrm{diag}(Q, I_{n-k})$, and (9.3) still holds. Among the possible choices of X, considering the following specific choice X_0 will be essential for our analysis below.

Choice of X_0: Among the possible X that satisfy (9.3), we choose X_0 such that the first k columns of $X_0 = [x_1, x_2, \cdots, x_k, \cdots, x_n]$ are chosen so that they are mutually orthogonal, that is,

$$[x_1, x_2, \ldots, x_k] = U\Sigma I_k \tag{9.4}$$

is the SVD, so that $U = [u_1, u_2, \ldots, u_k] \in \mathbb{C}^{n\times k}, U^*U = I_k$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ with $0 < \sigma_1 \leq \cdots \leq \sigma_k$.

Note that Σ is unique for any choice of unitary matrix Q as shown above. As for obtaining such X_0, given an arbitrary $\hat{X}$ that satisfies (9.3), we can get X_0 by first computing the SVD of the first k columns of $\hat{X}$: $\hat{X}(:, 1 : k) = U\Sigma V^*$, and then letting $X_0(:, 1 : k) = \hat{X}(:, 1 :$

$k)V = U\Sigma$ and $X_0(:, k+1:n) = \hat{X}(:, k+1:n)$ (we use MATLAB notation, denoting by $\hat{X}(:, 1:k)$ the first k columns of $\hat{X}$).

Now, given an integer t ($\leq k$), write

$$X_0^* E X_0 = \begin{bmatrix} E_{11}^{(t)} & E_{12}^{(t)} \\ E_{21}^{(t)} & O \end{bmatrix} + \begin{bmatrix} O & O \\ O & E_{22}^{(t)} \end{bmatrix} \equiv E_1^{(t)} + E_2^{(t)}, \tag{9.5}$$

and

$$X_0^* F X_0 = \begin{bmatrix} F_{11}^{(t)} & F_{12}^{(t)} \\ F_{21}^{(t)} & O \end{bmatrix} + \begin{bmatrix} O & O \\ O & F_{22}^{(t)} \end{bmatrix} \equiv F_1^{(t)} + F_2^{(t)}, \tag{9.6}$$

where $E_{11}^{(t)}$ and $F_{11}^{(t)}$ are t-by-t. Note that the two pairs $(A+E, B+F)$ and $(\Lambda + E_1^{(t)} + E_2^{(t)}, I + F_1^{(t)} + F_2^{(t)})$ are congruent, so they have the same eigenvalues.

Our next task is to bound $\|E_i^{(t)}\|_2$ and $\|F_i^{(t)}\|_2$ ($i = 1, 2$). We shall show that $\|E_1^{(t)}\|_2$ and $\|F_1^{(t)}\|_2$ are small for t such that σ_t is small. This in turn implies that a small interval exists that traps t eigenvalues of $(A+E, B+F)$.

9.1.2. Bounding $\|E_1^{(t)}\|_2$ and $\|F_1^{(t)}\|_2$. To bound $\|E_1^{(t)}\|_2$, we use

$$\|E_1^{(t)}\|_2 = \left\| \begin{bmatrix} E_{11}^{(t)} & E_{12}^{(t)} \\ E_{21}^{(t)} & O \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} 0 & E_{12}^{(t)} \\ E_{21}^{(t)} & 0 \end{bmatrix} \right\|_2 + \left\| \begin{bmatrix} E_{11}^{(t)} & 0 \\ 0 & 0 \end{bmatrix} \right\|_2. \tag{9.7}$$

The first term can be bounded by

$$\left\| \begin{bmatrix} 0 & E_{12}^{(t)} \\ E_{21}^{(t)} & 0 \end{bmatrix} \right\|_2 = \|E_{12}^{(t)}\|_2 \leq \sigma_t \sqrt{\|B^{-1}\|_2} \|E\|_2,$$

because $E_{12}^{(t)} = \Sigma_t U_t^* E X_0(:, t+1:n)$, where $U_t = [u_1, \ldots, u_t]$, $\Sigma_t = \mathrm{diag}(\sigma_1, \ldots, \sigma_t)$, so using $\|\Sigma_t\|_2 = \sigma_t$ and $\|X_0\|_2 = \sqrt{\|B^{-1}\|_2}$ (which follows from $B = X_0^{-H} X_0^{-1}$), we get $\|E_{12}^{(t)}\|_2 \leq \|\Sigma_t\|_2 \|U_t\|_2 \|E\|_2 \|X_0\|_2 = \sigma_t \sqrt{\|B^{-1}\|_2} \|E\|_2$.

The second term of (9.7) can be bounded by

$$\left\| \begin{bmatrix} E_{11}^{(t)} & 0 \\ 0 & 0 \end{bmatrix} \right\|_2 = \left\| E_{11}^{(t)} \right\|_2 \leq \sigma_t^2 \|E\|_2,$$

because $E_{11}^{(t)} = \Sigma_t U_t^* E U_t \Sigma_t$, from which we get $\|E_{11}^{(t)}\|_2 \leq \|\Sigma_t\|_2^2 \cdot \|U_t\|_2^2 \cdot \|E\|_2 = \sigma_t^2 \|E\|_2$.

Substituting these into (9.7) yields

$$\|E_1^{(t)}\| = \left\| \begin{bmatrix} E_{11}^{(t)} & E_{12}^{(t)} \\ E_{21}^{(t)} & O \end{bmatrix} \right\|_2 \leq \sigma_t (\sqrt{\|B^{-1}\|_2} + \sigma_t) \|E\|_2. \tag{9.8}$$

Similarly, we can bound $\|F_1^{(t)}\|_2$ by

$$\|F_1^{(t)}\|_2 = \left\| \begin{bmatrix} F_{11}^{(t)} & F_{12}^{(t)} \\ F_{21}^{(t)} & O \end{bmatrix} \right\|_2 \leq \sigma_t (\sqrt{\|B^{-1}\|_2} + \sigma_t) \|F\|_2. \tag{9.9}$$

As for bounding $\|E_2^{(t)}\|_2$ and $\|F_2^{(t)}\|_2$, it is straightforward to see from (9.5) and (9.6) that $\|E_2^{(t)}\|_2$ and $\|F_2^{(t)}\|_2$ satisfy

$$\|E_2^{(t)}\|_2 \le \|X_0^* E X_0\|_2 \le \|B^{-1}\|_2 \|E\|_2, \quad \|F_2^{(t)}\|_2 \le \|X_0^* F X_0\|_2 \le \|B^{-1}\|_2 \|F\|_2. \tag{9.10}$$

9.1.3. k bounds for multiple generalized eigenvalue. Now we are ready to state our main result of the section.

THEOREM 9.1. *Suppose that (A, B) and $(A+E, B+F)$ are n-by-n Hermitian positive definite pairs, and that (A, B) has a multiple eigenvalue λ_0 of multiplicity k. Let X be a nonsingular matrix that satisfies (9.3). Denote by $0 < \sigma_1 \le \sigma_2 \le \cdots \le \sigma_k$ the singular values of $X(:, 1:k)$, the first k columns of X. Then, for any integer t such that $1 \le t \le k$, $(A+E, B+F)$ has at least t eigenvalues $\widehat{\lambda}_i$ $(i = 1, 2, \ldots, t)$ that satisfy*

$$|\widehat{\lambda}_i - \lambda_0| \le \sigma_t(\sqrt{\|B^{-1}\|_2} + \sigma_t) \frac{\|E\|_2 + |\lambda_0|\|F\|_2}{1 - \|B^{-1}\|_2 \|F\|_2}. \tag{9.11}$$

PROOF. Recall that the pairs $(A+E, B+F)$ and $(\Lambda + E_1^{(t)} + E_2^{(t)}, I + F_1^{(t)} + F_2^{(t)})$ have the same eigenvalues, and note that the pair $(\Lambda + E_2^{(t)}, I + F_2^{(t)})$ has a multiple eigenvalue λ_0 of multiplicity (at least) t. We apply Theorem 8.3 by regarding $(\Lambda + E_1^{(t)} + E_2^{(t)}, I + F_1^{(t)} + F_2^{(t)})$ as a perturbed pair of $(\Lambda + E_2^{(t)}, I + F_2^{(t)})$. Then we see that the pair $(\Lambda + E_1^{(t)} + E_2^{(t)}, I + F_1^{(t)} + F_2^{(t)})$ has at least t eigenvalues $\widehat{\lambda}_i$ $(i = 1, 2, \ldots, t)$ that satisfy

$$\begin{aligned}
|\widehat{\lambda}_i - \lambda_0| \le & \frac{\|E_1^{(t)}\|_2 + |\lambda_0|\|F_1^{(t)}\|_2}{\lambda_{\min}(I + F_1^{(t)} + F_2^{(t)})} \\
\le & \frac{\sigma_t(\sqrt{\|B^{-1}\|_2} + \sigma_t)\|E\|_2 + |\lambda_0|\sigma_t(\sqrt{\|B^{-1}\|_2} + \sigma_t)\|F\|_2}{1 - \|B^{-1}\|_2\|F\|_2} \\
\le & \sigma_t(\sqrt{\|B^{-1}\|_2} + \sigma_t) \frac{\|E\|_2 + |\lambda_0|\|F\|_2}{1 - \|B^{-1}\|_2\|F\|_2},
\end{aligned}$$

where we used (9.8),(9.9) and $\lambda_{\min}(I + F_1^{(t)} + F_2^{(t)}) \ge 1 - \|F_1^{(t)} + F_2^{(t)}\|_2 = 1 - \|X_0^* F X_0\|_2 \ge 1 - \|B^{-1}\|_2\|F\|_2$, which follows from Weyl's theorem. □

We emphasize that inequality (9.11) holds for t (not k) eigenvalues of $(\Lambda + E_1^{(t)} + E_2^{(t)}, I + F_1^{(t)} + F_2^{(t)})$. The upper bound in (9.11) for $t = t_0$ is much smaller than that for $t = k$ if $\sigma_{t_0} \ll \sigma_k$. In such a case, among the k eigenvalues of (A, B) equal to λ_0, there are t_0 eigenvalues that are much less sensitive than the most sensitive one.

9.1.4. Simple example. Let us return to the simple example shown in the introduction and examine the sharpness of our results. For the pair (9.1), we formed perturbed Hermitian positive definite pairs $(A+E, B+F)$ using MATLAB version 7.4 by defining E and F by $\alpha(C^* + C)$, where the entries of the 2-by-2 matrix C are random numbers in $[-1/2, 1/2]$ generated by the MATLAB function `rand` $-$ `0.5` and α is defined by $10^{-2} \cdot$ `rand`$/\|C^* + C\|_2$ to force $\|E\|_2, \|F\|_2 \le 10^{-2}$. Experimenting with 10^4 such pairs, we observed that one

eigenvalue is always trapped in $[2 - 1.6 \times 10^{-4}, 2 + 1.6 \times 10^{-4}]$, but the interval that traps both eigenvalues needs to be as large as $[2 - 2.0 \times 10^{-2}, 2 + 2.0 \times 10^{-2}]$.

We give an explanation for this phenomenon by using Theorem 9.1. Here $|\lambda_0| = 2$, $\|B^{-1}\|_2 = 1$, $\|E\|_2 \le 10^{-2}$, $\|F\|_2 \le 10^{-2}$, $\sigma_1 = 10^{-2}$ and $\sigma_2 = 1$, so letting $t = 1$ in (9.11) we get

$$\begin{aligned} |\lambda_1 - 2| &\le 10^{-2}(1 + 10^{-2}) \cdot \frac{10^{-2} + 2 \times 10^{-2}}{1 - 1 \cdot 10^{-2}} \\ &= 3.0606 \times 10^{-4}, \end{aligned} \tag{9.12}$$

which means at leaset one eigenvalue of $(A+E, B+F)$ exists in $[2-3.1\times10^{-4}, 2+3.1\times10^{-4}]$.

To bound both eigenvalues we let $t = 2$, which yields

$$\begin{aligned} |\lambda_{1,2} - 2| &\le 1 \cdot (1 + 1) \cdot \frac{10^{-2} + 2 \times 10^{-2}}{1 - 1 \cdot 10^{-2}} \\ &= 6.06 \times 10^{-2}, \end{aligned}$$

a bound that is larger than (9.12) by more than a factor of 100.

We observe that these bounds reflect our experiments pretty accurately. Thus we claim Theorem 9.1 is one explanation for the different behaviors of multiple eigenvalues in generalized Hermitian eigenvalue problems.

9.1.5. Comparison with known results. Here we compare Theorem 9.1 with known perturbation results for a multiple eigenvalue in generalized Hermitian eigenvalue problems. To our knowledge, no result has been known that gives different a priori perturbation bounds using only the norms of the perturbation matrices (in the case of the standard eigenvalue problem, different condition numbers of a multiple eigenvalue are derived in [**147, 151**]). Here we give a comparison with known first-order perturbation approximations in [**12**] and [**95**], and see that both of them require more information than just the norms of E and F.

Below is a special case of Corollary 2.3 in [**12**], when (A, B) is a positive definite pair that has a multiple eigenvalue.

THEOREM 9.2. *Let (A, B) be a Hermitian positive definite pair that has a multiple eigenvalue λ_0 of multiplicity k. The perturbed pair $(A+E, B+F)$ has k eigenvalues $\widehat{\lambda}_i$ $(1 \le i \le k)$ such that*

$$exp(-\kappa_i) \le \frac{\widehat{\lambda}_i}{\lambda_0} \le exp(\kappa_i), \tag{9.13}$$

where

$$\kappa_i = \max_{\zeta\in[0,1]} \left| \frac{x_i^*(\zeta)Ex_i(\zeta)}{x_i^*(\zeta)A(\zeta)x_i(\zeta)} - \frac{x_i^*(\zeta)Fx_i(\zeta)}{x_i^*(\zeta)B(\zeta)x_i(\zeta)} \right| \ (\equiv \max_{\zeta\in[0,1]} g_i(\zeta)), \tag{9.14}$$

where the maximum in (9.14) is taken over ζ such that $\lambda_i(\zeta)$ is not a multiple eigenvalue. Here we denoted $A(\zeta) = A + \zeta E$, $B(\zeta) = B + \zeta F$, and let $(\lambda_i(\zeta), x_i(\zeta))$ for $1 \le i \le k$ be the k eigenpairs such that $\lambda_i(0) = \lambda_0$, $\lambda_i(1) = \widehat{\lambda}_i$ and

$$A(\zeta)x_i(\zeta) = \lambda_i(\zeta)B(\zeta)x_i(\zeta), \quad \zeta \in [0, 1].$$

In practice, bounds (9.13) are not available because κ_i cannot be computed. [**12**] suggests obtaining estimates of them by approximating κ_i by taking $g_i(\zeta)$ at a specific ζ, such as $\zeta = 0$, which results in first-order approximations of (9.13). Unfortunately we cannot take $\zeta = 0$ here because $\lambda_i(0)$ is a multiple eigenvalue. Instead, we can for example compute $\widehat{\kappa}_i = g_i(1)$ for $1 \le i \le k$. Substituting such computed $\widehat{\kappa}_i$ into κ_i in (9.13) yields estimates (but not bounds) of the perturbed eigenvalues $\widehat{\lambda}_i$, which are accurate when E and F are small.

To see how accurate these estimates are, let us consider again the simple example (9.1). If $E = 10^{-2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $F = 10^{-2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, using Theorem 9.2 by estimating each κ_i by $g_i(1)$, we get estimates for the eigenvalues of $(A+E, B+F)$: $|\widehat{\lambda}_1 - 2| \lesssim 2.998784 \times 10^{-6}$ and $|\widehat{\lambda}_2 - 2| \lesssim 9.980 \times 10^{-3}$. The true eigenvalues are $\simeq (2 - 2.998789 \times 10^{-6}, 2 + 1.00040 \times 10^{-2})$. The estimates are much sharper than the bounds given by Theorem 9.1 (see previous section). However the estimates are not strict bounds, as seen by the fact that the estimated interval for $\widehat{\lambda}_2$ does not trap the true eigenvalue. More importantly, Theorem 9.2 requires the eigenvectors of a perturbed pair, which is not available if the perturbations are unknown.

Another way to obtain approximations to the perturbed eigenvalues is to use the first-order eigenvalue perturbation expansion result, for example Theorem 4.1 in [**95**]. This involves computing eigenvalues of the matrix $X_k^*(E - \lambda_0 F)X_k$, where X_k is the first k columns of X in (9.3). Note that this also requires more information than just $\|E\|_2$ and $\|F\|_2$.

In summary, the above known results that give approximations to the multiple eigenvalue perturbations are generally sharper than the bound obtained by Theorem 9.1, but require more information of the perturbation matrices E and F. Theorem 9.1 has the advantage that it gives a priori bounds for arbitrary E and F using only their norms, revealing the fact that a multiple eigenvalue has different sensitivities.

9.2. Another explanation

Here we show that another explanation to the behavior can be made using Theorem 8.10 in Chapter 8.

Suppose a Hermitian definite pair (A, B) has a multiple eigenvalue λ_0 of multiplicity m. Then there is an $(m+n)$-by-$(m+n)$ matrix $X = (X_1,\ X_2)$ with $X_1^H X_1 = I_m$ and $X_2^H X_2 = I_n$ such that

$$X^H A X = \begin{array}{c} \\ {\scriptstyle m} \\ {\scriptstyle n} \end{array} \begin{array}{c} \begin{array}{cc} {\scriptstyle m} & {\scriptstyle n} \end{array} \\ \begin{bmatrix} \lambda_0 B_{11} & \\ & A_{22} \end{bmatrix} \end{array}, \quad X^H B X = \begin{array}{c} \\ {\scriptstyle m} \\ {\scriptstyle n} \end{array} \begin{array}{c} \begin{array}{cc} {\scriptstyle m} & {\scriptstyle n} \end{array} \\ \begin{bmatrix} B_{11} & \\ & B_{22} \end{bmatrix} \end{array}. \tag{9.15}$$

This can be seen by letting X_1 and X_2 be the orthogonal factors in the QR decompositions of $\widehat{X}_1$ and $\widehat{X}_2$ respectively, where $(\widehat{X}_1\ \widehat{X}_2)$ is the nonsingular matrix that diagonalizes (A, B) [**127**, p.344].

We may assume that B_{11} is diagonal:

$$B_{11} \equiv \Omega = \mathrm{diag}(\omega_1, \omega_2, \ldots, \omega_m), \quad \omega_1 \ge \omega_2 \ge \cdots \ge \omega_m > 0. \tag{9.16}$$

Otherwise, we can let the eigendecomposition of B_{11} (which is Hermitian and positive definite) be $B_{11} = U\Omega U^H$, where Ω is diagonal, and then perform substitutions $B_{11} \leftarrow \Omega$ and $X_1 \leftarrow X_1 U$.

Suppose (A, B) is perturbed to $(\widetilde{A}, \widetilde{B}) \equiv (A+E, B+F)$, where E and F are Hermitian. Write

$$X^H \widetilde{A} X = \begin{bmatrix} \lambda_0\Omega + E_{11} & E_{12} \\ E_{21} & A_{22} + E_{22} \end{bmatrix}, \quad X^H \widetilde{B} X = \begin{bmatrix} \Omega + F_{11} & F_{12} \\ F_{21} & B_{22} + F_{22} \end{bmatrix}. \tag{9.17}$$

For any given k $(1 \le k \le m)$, we re-partition $X^H AX$, $X^H BX$, $X^H EX$, and $X^H FX$ with k-by-k $(1,1)$ blocks as follows:

$$\begin{bmatrix} A_{11}^{(k)} & 0 \\ 0 & A_{22}^{(k)} \end{bmatrix}, \begin{bmatrix} B_{11}^{(k)} & 0 \\ 0 & B_{22}^{(k)} \end{bmatrix}, \begin{bmatrix} E_{11}^{(k)} & E_{12}^{(k)} \\ E_{21}^{(k)} & E_{22}^{(k)} \end{bmatrix}, \begin{bmatrix} F_{11}^{(k)} & F_{12}^{(k)} \\ F_{21}^{(k)} & F_{22}^{(k)} \end{bmatrix}. \tag{9.18}$$

It can be seen that

$$\begin{aligned} A_{11}^{(k)} &= \lambda_0 \Omega_{(1:k,1:k)}, \quad A_{22}^{(k)} = \mathrm{diag}(\lambda_0 \Omega_{(k+1:m,k+1:m)}, A_{22}), \\ B_{11}^{(k)} &= \quad \Omega_{(1:k,1:k)}, \quad B_{22}^{(k)} = \mathrm{diag}(\quad \Omega_{(k+1:m,k+1:m)}, B_{22}), \end{aligned}$$

Similarly to those in (8.21) define

$$\Delta_{ij}^{(k)} = [B_{ii}^{(k)}]^{-1/2} F_{ij}^{(k)} [B_{jj}^{(k)}]^{-1/2}, \tag{9.19a}$$

$$Y^{(k)} = \mathrm{diag}([I + \Delta_{11}^{(k)}]^{-1/2}[B_{11}^{(k)}]^{-1/2}, [I + \Delta_{22}^{(k)}]^{-1/2}[B_{22}^{(k)}]^{-1/2}), \tag{9.19b}$$

$$\widehat{F}_{ij}^{(k)} = [I + \Delta_{ii}^{(k)}]^{-1/2} \Delta_{ij}^{(k)} [I + \Delta_{jj}^{(k)}]^{-1/2} \quad \text{for } i \ne j, \tag{9.19c}$$

$$\widehat{A}_{ii}^{(k)} = [B_{ii}^{(k)}]^{-1/2} A_{ii}^{(k)} [B_{ii}^{(k)}]^{-1/2} \ (= \lambda_0 I \quad \text{when } i = 1), \tag{9.19d}$$

$$\widehat{E}_{ij}^{(k)} = [I + \Delta_{ii}^{(k)}]^{-1/2} [B_{ii}^{(k)}]^{-1/2} E_{ij}^{(k)} [B_{jj}^{(k)}]^{-1/2} [I + \Delta_{jj}^{(k)}]^{-1/2} \quad \text{for } i \ne j, \tag{9.19e}$$

$$\widehat{E}_{ii}^{(k)} = [I + \Delta_{ii}^{(k)}]^{-1/2} [B_{ii}^{(k)}]^{-1/2} (A_{ii}^{(k)} + E_{ii}^{(k)}) [B_{ii}^{(k)}]^{-1/2} [I + \Delta_{ii}^{(k)}]^{-1/2} - \widehat{A}_{ii}^{(k)}, \tag{9.19f}$$

and

$$\delta_{ij}^{(k)} = \|\Delta_{ij}^{(k)}\|_2 \le \sqrt{\|[B_{ii}^{(k)}]^{-1}\|_2 \|[B_{jj}^{(k)}]^{-1}\|_2}\, \|F_{ij}^{(k)}\|_2, \quad \gamma_{ij}^{(k)} = (1 - \delta_{ij}^{(k)})^{-1/2} - 1. \tag{9.20}$$

We can bound $\widehat{E}_{ij}^{(k)}$ and $\widehat{F}_{ij}^{(k)}$ as follows.

$$\|\widehat{E}_{ij}^{(k)}\|_2 \le \frac{\|[B_{ii}^{(k)}]^{-1/2} E_{ij}^{(k)} [B_{jj}^{(k)}]^{-1/2}\|_2}{\sqrt{(1 - \delta_{ii}^{(k)})(1 - \delta_{jj}^{(k)})}} \tag{9.21a}$$

$$\le \sqrt{\frac{\|[B_{ii}^{(k)}]^{-1}\|_2 \|[B_{jj}^{(k)}]^{-1}\|_2}{(1 - \delta_{ii}^{(k)})(1 - \delta_{jj}^{(k)})}}\, \|E_{ij}^{(k)}\|_2 \quad \text{for } i \ne j, \tag{9.21b}$$

$$\|\widehat{E}_{ii}^{(k)}\|_2 \le \gamma_{ii}^{(k)}(2 + \gamma_{ii}^{(k)}) \|\widehat{A}_{ii}^{(k)}\|_2 + \frac{\|[B_{ii}^{(k)}]^{-1/2} E_{ii}^{(k)} [B_{ii}^{(k)}]^{-1/2}\|_2}{1 - \delta_{ii}^{(k)}} \tag{9.21c}$$

$$\leq \gamma_{ii}^{(k)}(2+\gamma_{ii}^{(k)})\|\widehat{A}_{ii}^{(k)}\|_2 + \frac{\|[B_{ii}^{(k)}]^{-1}\|_2}{1-\delta_{ii}^{(k)}}\,\|E_{ii}^{(k)}\|_2, \tag{9.21d}$$

$$\|\widehat{F}_{ij}^{(k)}\|_2 \leq \frac{\delta_{ij}^{(k)}}{\sqrt{(1-\delta_{ii}^{(k)})(1-\delta_{jj}^{(k)})}} \quad \text{for } i \neq j. \tag{9.21e}$$

The gaps η_j as previously defined, when applied to the current situation with the partitioning as in (9.18), are all zeros, unless when $k = m$ and λ_0 is not an eigenvalue of (A_{22}, B_{22}). This makes Theorem 8.10 less favorable to apply than Theorem 8.11 because of the appearance of $\max_i \|\widehat{E}_{ii}^{(k)}\|_2$. Also we are interested here only in how different copies of λ_0 change due to the perturbation. The following theorem is a consequence of Theorem 8.11.

THEOREM 9.3. *Suppose that the Hermitian definite GEP* (9.15) *is perturbed to* (9.17) *and assume* (9.16). *Let all assignments* (9.18) – (9.20) *hold. Then for any given* k $(1 \leq k \leq m)$, *there are* k *eigenvalues* $\mu_1 \geq \cdots \geq \mu_k$ *of* $(\widetilde{A}, \widetilde{B})$ *such that*

$$|\mu_j - \lambda_0| \leq \|\widehat{E}_{11}^{(k)}\|_2 + \frac{\|\widehat{E}_{21}^{(k)} - \widehat{F}_{21}^{(k)}(\lambda_0 I + \widehat{E}_{11}^{(k)})\|_2}{\sqrt{1-\|\widehat{F}_{21}^{(k)}\|_2^2}} \quad \textit{for } 1 \leq j \leq k, \tag{9.22}$$

$$\sqrt{\sum_{j=1}^{k}|\mu_j - \lambda_0|^2} \leq \|\widehat{E}_{11}^{(k)}\|_{\mathrm{F}} + \frac{\|\widehat{E}_{21}^{(k)} - \widehat{F}_{21}^{(k)}(\lambda_0 I + \widehat{E}_{11}^{(k)})\|_{\mathrm{F}}}{\sqrt{1-\|\widehat{F}_{21}^{(k)}\|_2^2}}. \tag{9.23}$$

What makes this theorem interesting is that the right-hand sides of the inequalities may increase with k, illustrating different sensitivities of different copies of the multiple eigenvalue λ_0.

EXAMPLE 9.1. Consider matrices A and B in (9.1). It can be seen that

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ X^H A X = \begin{bmatrix} 2\times 10^4 & 0 \\ 0 & 2 \end{bmatrix},\ X^H B X = \begin{bmatrix} 10^4 & 0 \\ 0 & 1 \end{bmatrix}.$$

Suppose we perturb A and B by Hermitian matrices E and F with $\max_{i,j}\{|E_{(i,j)}|, |F_{(i,j)}|\} \leq \varepsilon$. Note that $\max_{i,j}\{|(X^H E X)_{(i,j)}|, |(X^H F X)_{(i,j)}|\} \leq \varepsilon$, because X is a permutation matrix. We shall now use Theorem 9.3 to bound how much the two copies of the multiple eigenvalue 2 may be perturbed. The application is done for $k = 1$ and 2. Recall that the right-hand sides of (9.22) and (9.23) depend on k; Let ρ_k denote the right-hand side of (9.22).

$k = 1$**:** $\delta_{11}^{(k)} \leq 10^{-4}\varepsilon$, $\gamma_{11}^{(k)} \leq (\sqrt{1-10^{-4}\varepsilon})^{-1} - 1 \approx \frac{1}{2}\times 10^{-4}\varepsilon$, $\delta_{22}^{(k)} \leq \varepsilon$, and

$$|\widehat{E}_{11}^{(k)}| \leq 2\gamma_{11}^{(k)}(2+\gamma_{11}^{(k)}) + \frac{10^{-4}\varepsilon}{1-10^{-4}\varepsilon} \approx 3\times 10^{-4}\varepsilon,$$

$$|\widehat{E}_{21}^{(k)}|,\ |\widehat{F}_{21}^{(k)}| \leq \frac{10^{-2}\varepsilon}{(1-10^{-4}\varepsilon)(1-\varepsilon)} \approx 10^{-2}\varepsilon.$$

Therefore $\rho_1 \lessapprox 3\times 10^{-4}\varepsilon + 3\times 10^{-2}\varepsilon \approx 3\times 10^{-2}\varepsilon$ after dropping higher order terms in ε. Here and in what follows, this "approximately less than" notation means the inequality holds up to the first order in ε.

$k = 2$**:** Now the blocks in the second row and column are empty. We have

$$\delta_{11}^{(k)} = \|\Delta_{11}^{(k)}\|_2 \le \|\Delta_{11}^{(k)}\|_{\mathrm{F}} \le \sqrt{1 + 2 \cdot 10^{-4} + 10^{-8}}\,\varepsilon \approx (1 + 10^{-4})\varepsilon,$$

$$\gamma_{11}^{(k)} \le \left[1 - (1 + 10^{-4})\varepsilon\right]^{-1/2} - 1 \approx \frac{1}{2}(1 + 10^{-4})\varepsilon,$$

$$\|\widehat{E}_{11}^{(k)}\|_2 \le 2\gamma_{11}^{(k)}(2 + \gamma_{11}^{(k)}) + \frac{(1 + 10^{-4})\varepsilon}{1 - \delta_{11}^{(k)}} \approx 3(1 + 10^{-4})\,\varepsilon.$$

Therefore $\rho_2 \lessapprox 3(1 + 10^{-4})\,\varepsilon$, again after dropping higher order terms in ε.

Putting these two facts together, we conclude that the perturbed pair has one eigenvalue that is away from 2 by approximately no more than $3 \times 10^{-2}\varepsilon$, while its other eigenvalue is away from 2 by approximately no more than $3(1 + 10^{-4})\,\varepsilon$. Further detailed examination reveals that the copy 20000/10000 is much less sensitive to perturbations than the copy 2/1. The bounds are rather sharp. For example in (9.1) if the $(1,1)$th blocks of A and B are perturbed to $2 + \varepsilon$ and $1 - \varepsilon$, respectively, then the more sensitive copy 2/1 is changed to $(2 + \varepsilon)/(1 - \varepsilon) \approx 2 + 3\varepsilon$ whose first order term is $3\,\varepsilon$, barely less than the bound on ρ_2. If A and B are perturbed to

$$A \to \begin{bmatrix} 2 & \varepsilon \\ \varepsilon & 20000 \end{bmatrix}, \quad B \to \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & 10000 \end{bmatrix},$$

where $\varepsilon \ge 0$, then the perturbed pair has eigenvalues, to the first order of ε,

$$2 - 3 \times 10^{-2}\varepsilon,\ 2 + 3 \times 10^{-2}\varepsilon,$$

which suggests that our estimate on ρ_1 is also sharp. □

9.3. Implication in the Rayleigh-Ritz process

In this section we consider the eigenvalue forward error analysis after the Rayleigh-Ritz process. In particular, our focus is on how our observation in Section 9.1 plays a role in this context.

9.3.1. Preliminaries. The Rayleigh-Ritz process is frequently used in an algorithm that computes a subset of eigenvalues and eigenvectors of a large matrix/pair. These algorithms include Lanczos, steepest descent, conjugate gradient, LOBPCG, generalized Davidson and Jacobi-Davidson methods [**6, 91, 127, 137**]. For a N-by-N generalized Hermitian eigenvalue problem $Ax = \lambda Bx$, given $Y \in \mathbb{C}^{N \times m}$ $(N \gg m)$ whose columns ideally contain the desired eigenspace, the Rayleigh-Ritz process computes approximate eigenvalues/eigenvectors by solving an $m \times m$ eigenvalue problem $Y^*AYz_i = \theta_i Y^*BYz_i$ $(i = 1, \ldots, m)$, from which the approximate eigenvalues θ_i (Ritz values) and approximate eigenvectors $w_i = Yz_i$ (Ritz vectors) are obtained such that denoting $W = [w_1\ w_2\ \cdots\ w_m]$, $W^*AW = \Lambda = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_m)$ and $W^*BW = I_m$. We are particularly interested in the case where a multiple Ritz value (of multiplicity k) exists, i.e., when $\theta_0 \equiv \theta_1 = \theta_2 = \cdots = \theta_k$. We also have the residual vectors

$$r_i = Aw_i - \theta_i Bw_i, \quad i = 1, 2, \ldots, m, \tag{9.24}$$

which are nonzero but expected to be small. Conventionally, an eigenvalue forward error analysis that bounds the closeness of the Ritz values to a true eigenvalue is obtained as follows [**6**]. First note that (9.24) is equivalent to

$$B^{-1/2}r_i = (B^{1/2}AB^{-1/2})(B^{1/2}w_i) - \theta_i(B^{1/2}w_i).$$

Suppose the Ritz vectors w_i $(1 \le i \le m)$ are B-orthonormalized, so $\|w_i\|_B = 1$, where $\|v\|_B = \sqrt{v^*Bv}$. Then, it is known [**127**, p.73] that there exists *an* eigenvalue λ of the pair (A, B) such that

$$|\lambda - \theta_i| \le \frac{\|B^{-1/2}r_i\|_2}{\|B^{1/2}w_i\|_2} = \frac{\|r_i\|_{B^{-1}}}{\|w_i\|_B} \le \sqrt{\|B^{-1}\|_2}\|r_i\|_2. \tag{9.25}$$

A quadratic error bound

$$|\lambda - \theta_i| \le \frac{1}{gap} \cdot \left(\frac{\|B^{-1/2}r_i\|_2}{\|B^{1/2}w_i\|_2}\right)^2 \le \frac{\|B^{-1}\|_2\|r_i\|_2^2}{gap} \tag{9.26}$$

can also be used when an estimate of *gap* (the smallest gap between θ_i and any eigenvalue of $A - \lambda B$ but the one closest to θ_i) is available [**6**, Sec.5.7.1].

9.3.2. Questions. Bounds (9.25) and (9.26) have the following caveats, regarding the multiple Ritz value θ_0.

(1) They do not reflect the different perturbation behaviors of a multiple eigenvalue that we observed in Section 9.1.
(2) They only give an interval in which at least one true eigenvalue exists, so there is no known $b_i \ge 0$ such that $[\theta_0 - b_i, \theta_0 + b_i]$ contains at least i true eigenvalues for $2 \le i \le k$.

We use a simple example to illustrate these two issues. Consider a 4-by-4 Hermitian positive definite pair (A, B), defined by

$$A = \mathrm{diag}(10^4, 1, 2, 2) + \begin{bmatrix} 0 & C_1^* \\ C_1 & 0 \end{bmatrix}, \quad B = \mathrm{diag}(10^4, 1, 1, 1) + \begin{bmatrix} 0 & C_2^* \\ C_2 & 0 \end{bmatrix}, \tag{9.27}$$

where $C_1, C_2 \in \mathbb{C}^{2\times 2}$. When C_1 and C_2 are small, the pair (A, B) has two eigenvalues close to 1 and another two close to 2. Furthermore, using Theorem 9.1 we see that among the two eigenvalues close to 1, one has to satisfy $|\lambda - 1| \lesssim 10^{-2}(\|C_1\|_2 + \|C_2\|_2)$, while the other has the bound $|\lambda - 1| \lesssim \|C_1\|_2 + \|C_2\|_2$, suggesting different sensitivities.

Consider for example the case $C_1 = 0.1\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $C_2 = 0.1\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (any choice of sufficiently small random matrices C_1, C_2 yields similar results for the following arguments). Suppose we have an approximate eigenspace spanned by $(1\ 0\ 0\ 0)^T$ and $(0\ 1\ 0\ 0)^T$. The Rayleigh-Ritz process yields the 2-by-2 pair $(\tilde{A}, \tilde{B})$ where $\tilde{A} = \tilde{B} = \mathrm{diag}(10^4\ 1)$, so the Ritz values are both 1. The resulting $B-$orthonormalized Ritz vectors may be for example $w_1 = (10^{-2}\ 0\ 0\ 0)^T$ and $w_2 = (0\ 1\ 0\ 0)^T$, which yield the residual vectors

$$r_1 = Aw_1 - Bw_1 = (0\ 0\ 0\ 10^{-3})^T, \quad r_2 = Aw_2 - Bw_2 = (0\ 0\ 10^{-1}\ 0)^T. \tag{9.28}$$

Hence the forward eigenvalue error bound (9.25) yields

$$|\lambda - 1| \leq \sqrt{\|B^{-1}\|_2}\|r_1\|_2 \leq \sqrt{1/0.9}/1000 < 1.06 \times 10^{-3}, \tag{9.29}$$

which at least one eigenvalue has to satisfy (here we used $\|B^{-1}\|_2 \leq 1/0.9$, which is easily obtained by Weyl's theorem. In practical problems, estimating $\|B^{-1}\|_2$ is a nontrivial task, which is beyond the scope of this chapter).

Note that any set of vectors written as $[w_1\ w_2]Q$, where Q is any unitary matrix, is a pair of Ritz vectors that yields the same Ritz values. Hence one may instead have as Ritz vectors for example $\widetilde{w}_1 = (\frac{1}{100\sqrt{2}}\ \frac{1}{\sqrt{2}}\ 0\ 0)^T$ and $\widetilde{w}_2 = (\frac{1}{100\sqrt{2}}\ -\frac{1}{\sqrt{2}}\ 0\ 0)^T$. In this case, the residual vectors become

$$\widetilde{r}_1 = \frac{1}{\sqrt{2}}(0\ 0\ 10^{-1}\ 10^{-3}), \quad \widetilde{r}_2 = \frac{1}{\sqrt{2}}(0\ 0\ -10^{-1}\ 10^{-3})^T. \tag{9.30}$$

Applying these to (9.25) yields only $|\lambda - 1| \leq 7.5 \times 10^{-2}$, which is a much looser bound than (9.29). Now the obvious question is, how can we ensure to choose the "right" Ritz vectors so we have "good" bound such as (9.28)? This is the first question we raised.

The second question concerns the eigenvalue that is more sensitive to perturbations. It is important to note that the union of 2 bounds using r_1 and r_2 in (9.28) or (9.30) does not necessarily bound two eigenvalues, as is warned in [**127**, Sec.11.5]. How can we obtain a bound that is guaranteed to contain two eigenvalues? And for a general pair, if a multiple Ritz value has multiplicity k, can we get k different bounds to reflect the different sensitivities?

9.3.3. Choosing the "right" Ritz vectors. This subsection answers the above two questions. We consider the issue of choosing the set of "right" Ritz vectors $W = \{w_1, w_2, \ldots, w_k\}$ for a multiple Ritz value θ_0 of multiplicity k. Here, 'right' means the particular choice of Ritz vectors provides the tightest forward error bounds for the Ritz values. In view of the bound (9.25), it is natural to attempt to minimize $\sqrt{\|B^{-1}\|_2}\|r_i\|_2$ for $i = 1, 2, \ldots, k$. We do this by the following approach.

Suppose we have a set of computed Ritz vectors $\widehat{W} \in \mathbb{C}^{N\times k}$ such that $\widehat{W}^* A\widehat{W} = \theta_0 I_k$ and $\widehat{W}^* B\widehat{W} = I_k$. Recall that we can replace $\widehat{W}$ by $\widehat{W}Q$ for any unitary matrix Q. In our approach, we compute $\widehat{R} = A\widehat{W} - \theta_0 B\widehat{W}$ and its SVD: $\widehat{R} = U_R \Sigma_R V_R^*$, where $\Sigma_R = \mathrm{diag}(\sigma_1(\widehat{R}), \ldots, \sigma_k(\widehat{R}))$ with $0 \leq \sigma_1(\widehat{R}) \leq \cdots \leq \sigma_k(\widehat{R})$. Then define W and R by $W = \widehat{W}V_R$ and $R = \widehat{R}V_R = U_R\Sigma_R (= AW - \theta_0 BW)$. This is equivalent to letting $Q = V_R$.

Note that this choice $Q = V_R$ is optimal in the sense that denoting $\bar{R} = \widehat{R}Q$, it minimizes $\|\bar{R}(:, 1:i)\|_2$ over all unitary Q for all integers $i \leq k$. To see this, we use the property of singular values [**142**, p.68] that for any matrix X, its ith smallest singular value $\sigma_i(X)$ is characterized by

$$\sigma_i(X) = \min_{\dim(S)=i} \max_{\substack{w\in S \\ \|w\|_2=1}} \|Xw\|_2. \tag{9.31}$$

Using (9.31) and denoting by S_i the subspace spanned by the first i columns of I_k, we have

$$\|\bar{R}(:, 1:i)\|_2 = \max_{\substack{w\in S_i \\ \|w\|_2=1}} \|\bar{R}w\|_2 = \max_{\substack{w\in S_i \\ \|w\|_2=1}} \|\widehat{R}Qw\|_2 = \max_{\substack{v\in QS_i \\ \|v\|_2=1}} \|\widehat{R}v\|_2 \geq \sigma_i(\widehat{R}), \tag{9.32}$$

because $\dim(S_i) = \dim(QS_i) = i$. Since $\|R(:, 1:i)\|_2 = \sigma_i(\widehat{R})$, we see that the equality in (9.32) is attained for all integers $i \leq k$ when $Q = V_R$, so the claim is proved.

Now, consider $W_2 \in \mathbb{C}^{N\times(N-k)}$ such that $W_1 = [W\ W_2]$ satisfies $W_1^* B W_1 = I_N$ (such W_1 exists, which can be obtained for example by B-orthonormalizing a nonsingular matrix $[W\ \widehat{W_2}]$). Then, we have

$$W_1^* A W_1 = \begin{bmatrix} \theta_0 I & R_2^* \\ R_2 & A_{22} \end{bmatrix}, \tag{9.33}$$

where $R_2 = W_2^* R$. The matrix $W_1^* A W_1$ and the pair (A, B) have the same eigenvalues. Here, since we know that $\|R(:, 1:i)\|_2 = \sigma_i(\widehat{R})$, we have

$$\|R_2(:, 1:i)\|_2 \leq \|R(:, 1:i)\|_2 \|W_2\|_2 = \sigma_i(\widehat{R})\|W_2\|_2 \leq \sigma_i(\widehat{R})\sqrt{\|B^{-1}\|_2},$$

for $i = 1, \ldots, k$. Here we used $\|W_2\|_2 \leq \|W_1\|_2 = \|B^{-1/2}\|_2$. Then, by using Weyl's theorem, we can conclude that for any integer $i \leq k$, there are at least i eigenvalues of the pair (A, B) that satisfy

$$|\lambda - \theta_0| \leq \sigma_i(\widehat{R})\sqrt{\|B^{-1}\|_2}. \tag{9.34}$$

Note that we only need to compute the singular values of $\widehat{R}$ (besides an estimate of $\|B^{-1}\|$) to get (9.34). Note also that for $i = 1$, (9.34) is equivalent to (9.25) obtained by substituting the residual vector $R(:, 1)$, which is the smallest possible error bound for the Ritz value θ_0 one can get using (9.25). Our bound (9.34) gives bounds also for $i \geq 2$.

9.3.4. Simple example. Let us return to the pair (9.27), and again consider the case $C_1 = 0.1\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $C_2 = 0.1\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ to demonstrate the approach described above. Suppose the Ritz vectors $\widehat{W} = [\widehat{w}_1, \widehat{w}_2]$ (not unique: for example, $\hat{W}$ can be $[w_1, w_2]$ or $[\widetilde{w}_1, \widetilde{w}_2]$ in Section 9.3.2) are computed so that $\widehat{W}^* A \widehat{W} = \theta_0 I_2 (= I_2)$ and $\widehat{W}^* B \widehat{W} = I_2$. Our approach computes Σ_R in the SVD $\widehat{R} = (A - B)\widehat{W} = U_R \Sigma_R V_R^*$. In this case, $\Sigma_R = \mathrm{diag}(\sigma_1(\widehat{R}), \sigma_2(\widehat{R})) = \mathrm{diag}(10^{-3}, 10^{-1})$, regardless of the choice of $\widehat{W}$. Then, again using $\|B^{-1}\|_2 \leq 1/0.9$, we use (9.34) to conclude that one eigenvalue satisfies $|\lambda - 1| \leq \sqrt{\|B^{-1}\|_2}\sigma_1(\widehat{R}) < 1.06 \times 10^{-3}$ $(= \delta_1)$, and that two eigenvalues satisfy $|\lambda - 1| \leq \sqrt{\|B^{-1}\|_2}\sigma_2(\widehat{R}) < 1.06 \times 10^{-1}$ $(= \delta_2)$.

The true eigenvalues of (A, B) are $\simeq (1 - 10^{-6}, 1 - 10^{-2}, 2, 2 + 2 \cdot 10^{-2})$. Note that the sensitivity difference between the two eigenvalues close to 1 is a result of the difference between $\sigma_1(\widehat{R})$ and $\sigma_2(\widehat{R})$, which is justified by observing that if we replace the $(1, 1)$ elements of A and B by 1, then $\sigma_1(\widehat{R}) = \sigma_2(\widehat{R}) = 0.1$, and the eigenvalues become $\simeq (1 - 9.7 \cdot 10^{-3}, 1 - 10^{-2}, 2, 2 + 4 \cdot 10^{-2})$, so both eigenvalues exhibit similar sensitivities.

Unfortunately our error estimates δ_1 and δ_2 for the two eigenvalues close to 1 are both overestimates, and in particular we observe that their squares δ_1^2, δ_2^2 seem to be accurate estimates of the true errors. This behavior is rather general, and in fact was true for all randomly constructed C_1 and C_2 that we tried. However we cannot make this observation precise; the known quadratic error bound $\|r_i\|_2^2/gap$ in [**6**] is not helpful here, because $gap \simeq 10^{-2}$ is small (or unavailable in practice when a multiple Ritz value exists) and the bounds will not be improved. We can also apply quadratic residual bounds [**109, 97**] to the matrix

(9.33), in which case we get an error bound $\|R_2\|^2/1 = 10^{-2}$. However this is an error bound for both eigenvalues, and does not describe the less sensitive eigenvalue. A rigorous explanation for the observation is an open problem.

9.4. Condition numbers of a multiple generalized eigenvalue

So far in this chapter we have investigated perturbation bounds of a multiple generalized eigenvalue when the matrices A, B undergo finite perturbation E, F. In the rest of this chapter we focus on the case $E, F \to 0$.

Table 9.4.1 summarizes what follows in the rest of the chapter, which shows the condition numbers κ_i for $i = 1, \ldots, r$ of a nondefective finite multiple eigenvalue in four situations, expressed in terms of $\sigma_1, \ldots, \sigma_r$, the r positive singular values of $X_1 Y_1^H$ (see Section 9.4.1). The contribution of this chapter is that we fill in the second row, that is, we identify the condition numbers of a multiple eigenvalue in a generalized eigenvalue problem, both for the Hermitian definite case and the non-Hermitian case. Here τ is a prescribed positive constant that accounts for perturbation scalings, see Section 9.4.1.

TABLE 9.4.1. Summary of condition numbers κ_i of a multiple eigenvalue for $i = 1, \ldots, r$.

	Hermitian	Non-Hermitian
$Ax = \lambda x$	1	$\left(\prod_{j=1}^{i} \sigma_j\right)^{1/i}$
$Ax = \lambda Bx$	$(1 + \tau\|\lambda_0\|) \min_{1 \le j \le i} \sqrt{\sigma_j \sigma_{i-j+1}}$	$(1 + \tau\|\lambda_0\|) \left(\prod_{j=1}^{i} \sigma_j\right)^{1/i}$

There are a number of related studies in the literature. [**95**] investigates the Hölder condition number, which is essentially κ_1 in our terminology when λ_0 is nondefective. The focus of [**95**] is the effect of the structure of the perturbation on the Hölder condition number, and in Section 9.6.1 we discuss how our results are related to those in [**95**].

An observation that a multiple generalized eigenvalue has different sensitivities under perturbations was first made in [**146**, p.300], which mentions that a multiple eigenvalue of a pair such as $A = \begin{bmatrix} 2000 & 0 \\ 0 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1000 & 0 \\ 0 & 1 \end{bmatrix}$ tends to behave differently under perturbations in A and B. We note that as shown in [**161**], for Hermitian definite pairs, small componentwise relative changes in A and B can introduce only small relative perturbation to any eigenvalue, and it is easy to see the two eigenvalues of the above pair (A, B) have similar perturbation behaviors. However, in terms of "standard" normwise perturbation, that is, when (A, B) is perturbed to $(A + \epsilon E, B + \epsilon F)$ under $\|E\|_2, \|F\|_2 \le 1$ and $\epsilon \to 0$, a multiple eigenvalue can exhibit different behaviors. [**114, 104**] consider the Hermitian definite case and give an explanation for this behavior, presenting r different perturbation bounds for λ_0 under perturbations of finite norm. The approach of this chapter is different in that our focus is on the condition numbers, which are attainable perturbation bounds in the first order sense in the limit $E, F \to 0$. The bounds in [**114, 104**] are valid for non-asymptotic E, F but are less tight (generally not attainable) when $E, F \to 0$.

Our arguments closely follow that of [**151**], in which the condition numbers are called *worst-case* condition numbers, to emphasize the difference from the *typical-case* condition numbers, as presented in [**147**]. In this sense, our results should also be regarded as worst-case condition numbers, in that κ_i are the largest attainable bounds in the first order sense. Experiments show that these bounds are not likely to be attained in practice for randomly generated perturbations, especially for large i (see the example in Section 9.6.1).

9.4.1. Definition. For an n-by-n matrix pair (A, B), suppose that λ_0 is a nondefective finite multiple eigenvalue (we discuss the infinite and defective cases later in Section 9.6.3) of multiplicity r, so that there exist nonsingular matrices $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ with $X_1, Y_1 \in \mathbb{C}^{n\times r}$ that satisfy

$$Y^H A X = \begin{bmatrix} \lambda_0 I_r & 0 \\ 0 & J_A \end{bmatrix}, \quad Y^H B X = \begin{bmatrix} I_r & 0 \\ 0 & J_B \end{bmatrix}. \tag{9.35}$$

Here the spectrum of the pair (J_A, J_B) does not contain λ_0. Then, the pair $(A+\epsilon E, B+\epsilon F)$ has eigenvalues $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_r$ admitting the first order expansion [**112, 95**]

$$\widehat{\lambda}_i = \lambda_0 + \lambda_i(Y_1^H(E - \lambda_0 F)X_1)\epsilon + o(\epsilon), \quad i = 1, 2, \ldots, r, \tag{9.36}$$

where $\lambda_i(Y_1^H(E - \lambda_0 F)X_1)$ are the eigenvalues of $Y_1^H(E - \lambda_0 F)X_1$ for $i = 1, \ldots, r$. In light of (9.36) and following the definition presented in [**151**], we define r condition numbers $\kappa_i(A, B, \lambda_0)$ for $i = 1, \ldots, r$ of the multiple eigenvalue λ_0 as follows.

Definition 9.2. *Let an n-by-n matrix pair (A, B) have the decomposition* (9.35)*, and let $\tau > 0$ be a prescribed constant. We define the condition numbers of λ_0, a multiple eigenvalue of (A, B) of multiplicity r, by*

$$\kappa_i(A, B, \lambda_0) \equiv \sup_{\|E\|_2\le 1, \|F\|_2\le \tau} |\lambda_i(Y_1^H(E - \lambda_0 F)X_1)|, \quad i = 1, \ldots, r, \tag{9.37}$$

where the eigenvalues $\lambda_i(Y_1^H(E - \lambda_0 F)X_1)$ are ordered such that $|\lambda_1(Y_1^H(E - \lambda_0 F)X_1)| \ge |\lambda_2(Y_1^H(E - \lambda_0 F)X_1)| \ge \cdots \ge |\lambda_r(Y_1^H(E - \lambda_0 F)X_1)|$.

In words, $\kappa_i(A, B, \lambda_0)$ measures by how much small changes in A and B can be magnified in the multiple eigenvalue λ_0, in the first order sense. τ is a positive constant that allows for the case where perturbations in A and B occur in different magnitudes, which is a notion adopted for example in [**72**].

9.4.2. Equivalent characterization of $\kappa_i(A, B, \lambda_0)$. Here we show that $\kappa_i(A, B, \lambda_0)$ can be expressed in an equivalent form, just as in the standard ($B = I_n$) case $\kappa_i(A, \lambda_0)$ can be expressed as (9.2) using the secants of the canonical angles $c_j(A, \lambda_0)$ between the left and right invariant subspaces corresponding to λ_0. Note that $c_j(A, \lambda_0) = \sigma_j(X_1 Y_1^H)$ for $j = 1, \ldots, r$. In this chapter we use the quantity $\sigma_j(X_1 Y_1^H)$ instead of the canonical angles to identify the condition numbers, because it allows us to treat generalized eigenvalue problems in a uniform way.

We use the proof of Theorem 2.1 in [**151**], whose crucial identity is

$$\begin{aligned} |\lambda_i(Y_1^H(E - \lambda_0 F)X_1)| &= |\lambda_i((E - \lambda_0 F)X_1 Y_1^H)| = |\lambda_i((E - \lambda_0 F)U\Sigma V^H)| \\ &= |\lambda_i(V^H(E - \lambda_0 F)U\Sigma)| \end{aligned}$$

for $i = 1, \ldots, r$, where $X_1 Y_1^H = U\Sigma V^H$ is the "thin" SVD. Here, to get the first and last equalities we used the fact [**146**, p.27] that for general $X \in \mathbb{C}^{n\times m}$ and $Y \in \mathbb{C}^{m\times n}$, the nonzero eigenvalues of XY and those of YX are the same. Since V and U have orthonormal columns and E, F can take arbitrary matrices with $\|E\|_2 \le 1, \|F\|_2 \le \tau$, it follows that $V^H EU, V^H FU$ can also take arbitrary matrices such that $\|V^H EU\|_2 \le 1, \|V^H FU\|_2 \le \tau$. Hence, redefining $E := V^H EU$ and $F := V^H FU$, we see that the condition numbers $\kappa_i(A, B, \lambda_0)$ have the following equivalent characterization.

LEMMA 9.1. *Under the assumptions in Definition 9.2, suppose that $X_1 Y_1^H = U\Sigma V^H$ is the SVD where $\Sigma = diag(\sigma_i)$ is r-by-r ($\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_r > 0$). Then, $\kappa_i(A, B, \lambda_0)$ in* (9.37) *can be expressed as*

$$\kappa_i(A, B, \lambda_0) \equiv \sup_{\|E\|_2\le 1, \|F\|_2\le\tau} |\lambda_i(\Sigma(E - \lambda_0 F))|, \quad i = 1, \ldots, r. \tag{9.38}$$

Here we have $\sigma_r > 0$ because both X_1 and Y_1 have full column-rank. Note that the size of E and F in (9.38) is r-by-r, which is smaller than n-by-n as in (9.37). In the sequel we use the expression (9.38) to identify $\kappa_i(A, B, \lambda_0)$.

9.5. Hermitian definite pairs

9.5.1. Specifications. When (A, B) is a Hermitian definite pair, all the eigenvalues are always real and nondefective, and there exists a nonsingular matrix X such that [**56**]

$$X^H AX = \begin{bmatrix} \lambda_0 I_r & 0 \\ 0 & \Lambda_1 \end{bmatrix}, \quad X^H BX = I_n, \tag{9.39}$$

where Λ_1 is a diagonal matrix containing the eigenvalues not equal to λ_0. Hence the diagonals of Σ in (9.38) are the r positive singular values of the matrix $X_1 X_1^H$, which are equal to the eigenvalues of the matrix $X_1^H X_1$. Since (A, B) is a Hermitian definite pair it is natural to require that the perturbation matrices preserve the property, so (9.38) becomes the "structured" condition numbers $\kappa_i(A, B, \lambda_0; \mathbb{S})$, expressed by

$$\kappa_i(A, B, \lambda_0; \mathbb{S}) \equiv \sup_{\substack{\|E\|_2\le 1, \|F\|_2\le\tau \\ E=E^H, F=F^H}} |\lambda_i(\Sigma(E - \lambda_0 F))|, \quad i = 1, \ldots, r. \tag{9.40}$$

Denoting $D = \Sigma^{1/2} = \mathrm{diag}(\sqrt{\sigma_1}, \ldots, \sqrt{\sigma_r})$, we see that the eigenvalues of $\Sigma(E - \lambda_0 F)$ are equal to those of the Hermitian matrix $D(E - \lambda_0 F)D$.

We further observe that $E - \lambda_0 F$ can represent an arbitrary Hermitian matrix H with $\|H\|_2 \le 1 + \tau|\lambda_0|$, which can be done by letting $E = H/\|H\|_2$ and $F = -\tau E|\lambda_0|/\lambda_0$. Conversely, it is easily seen that the class of Hermitian matrices H with $\|H\|_2 \le 1 + \tau|\lambda_0|$ includes all the matrices expressed by $E - \lambda_0 F$. Together with the fact that the singular values of a Hermitian matrix are simply the absolute values of the eigenvalues, we have yet another characterization of condition numbers in the Hermitian definite case

$$\kappa_i(A, B, \lambda_0; \mathbb{S}) = (1 + \tau|\lambda_0|) \sup_{\substack{\|H\|_2\le 1 \\ H=H^H}} \sigma_i(DHD), \quad i = 1, \ldots, r. \tag{9.41}$$

9.5.2. Identifying the condition numbers. Now we are ready to identify the condition numbers $\kappa_i(A, B, \lambda_0; \mathbb{S})$ using the expression (9.41).

THEOREM 9.4. *In the Hermitian definite case,* $\kappa_i(A, B, \lambda_0; \mathbb{S})$ *as in* (9.40),(9.41) *is*

$$\kappa_i(A, B, \lambda_0; \mathbb{S}) = (1 + \tau|\lambda_0|) \min_{1 \le j \le i} \sqrt{\sigma_j \sigma_{i-j+1}}, \quad i = 1, \ldots, r. \tag{9.42}$$

Note that $\sqrt{\sigma_j \sigma_{i-j+1}}$ is the geometric mean of σ_j and σ_{i-j+1}, the jth largest and smallest of $(\sigma_1, \sigma_2, \ldots, \sigma_i)$, which is the set of the i largest singular values of $X_1^H X_1$.

Proof In view of (9.41), to prove the theorem it suffices to prove that for any Hermitian H such that $\|H\|_2 = 1$, $\sigma_i(DHD)$ is bounded above by $\min_j \sqrt{\sigma_j \sigma_{k-j+1}}$ for $i = 1, \ldots, r$, and that this bound is attainable.

First, proving attainability is simply done by considering the case where H is zero except for its $i \times i$ leading principal submatrix, which is set to the antidiagonal matrix (which has 1 on the antidiagonals and 0 elsewhere). This choice of H makes the $i \times i$ leading principal submatrix of DHD also an anti-diagonal matrix, whose jth antidiagonal is $\sqrt{\sigma_j \sigma_{i-j+1}}$. Hence we have $\sigma_i(DHD) = \min_j \sqrt{\sigma_j \sigma_{i-j+1}}$.

Our remaining task is to prove that $\min_j \sqrt{\sigma_j \sigma_{i-j+1}}$ is an upper bound of $\sigma_i(DHD)$ for any Hermitian H with $\|H\|_2 \le 1$. Using the max-min characterization of singular values [**142**, p. 68], we have

$$\sigma_i(DHD) = \max_{Q^H Q = I_i} \min_{\|v\|_2 = 1} \|DHDQv\|_2,$$

so it suffices to show that for any $Q \in \mathbb{C}^{r \times i}$ with orthonormal columns, there exists a unit vector v such that $\|DHDQv\|_2 \le \min_j \sqrt{\sigma_j \sigma_{i-j+1}}$.

To prove this, let $j_0 = \operatorname{argmin}_{j \le (i+1)/2} \sqrt{\sigma_j \sigma_{i-j+1}}$. Since for any Q we have $\operatorname{rank}(Q(1 : i - j_0, :)) \le i - j_0$, there are at least j_0 linearly independent vectors in $\mathbb{C}^{i \times 1}$ that are orthogonal to the rows of Q. Therefore there must exist $P \in \mathbb{C}^{i \times j_0}$ with orthonormal columns such that the first $i - j_0$ rows of the r-by-j_0 matrix QP are all zeros. For such P, we have

$$\|DQP\|_2 = \|\operatorname{diag}(0, \ldots, 0, \sqrt{\sigma_{i-j_0+1}}, \ldots, \sqrt{\sigma_r}) QP\|_2 \le \sqrt{\sigma_{i-j_0+1}}.$$

Furthermore, since $\operatorname{rank}(HDQP(1 : j_0 - 1, :)) \le j_0 - 1$, there must exist a unit vector $w \in \mathbb{C}^{j_0 \times 1}$ that is orthogonal to $HDQP(1 : j_0 - 1, :)$, so that the first $j_0 - 1$ rows of $HDQPw$ are all zeros. We easily see that for such w we have $\|DHDQPw\|_2 \le \sqrt{\sigma_{j_0} \sigma_{i-j_0+1}}$. Therefore we have shown that for any $Q \in \mathbb{C}^{k \times i}$ with orthonormal columns there exists a unit vector $v_0 = Pw$ such that

$$\min_{\|v\|_2 = 1} \|DHDQv\|_2 \le \|DHDQv_0\|_2 \le \sqrt{\sigma_{j_0} \sigma_{i-j_0+1}} = \min_j \sqrt{\sigma_i \sigma_{i-j+1}}.$$

□

Three remarks are in order.

- When $B \ne I_n$, σ_i for $i = 1, \ldots, r$ generally take different values, so (9.42) shows that a multiple generalized eigenvalue has multiple condition numbers, which is our main result. Note that the ratio among the condition numbers is bounded by $\kappa_1(A, B, \lambda_0; \mathbb{S}) / \kappa_r(A, B, \lambda_0; \mathbb{S}) \le \sigma_1 / \sigma_r$. Now since $\sigma_{\min}(B^{-1}) \le \sigma_r \le \sigma_1 \le \sigma_{\max}(B^{-1})$, we have $\sigma_1 / \sigma_r \le \sigma_{\max}(B) / \sigma_{\min}(B) = \kappa_2(B)$, the standard 2−norm

condition number of B. It follows that if B is well-conditioned then a multiple eigenvalue of a Hermitian definite pair must have similar condition numbers.

- For standard Hermitian eigenvalue problems ($B = I_n$), we have $d_i \equiv 1$ and $\tau = 0$, so (9.42) reduces to $\kappa_i = 1$ for all i, the well-known result that a multiple eigenvalue of a Hermitian matrix has a uniform condition number 1. When one allows for perturbation in $B = I_n$, the condition numbers are $1 + \tau|\lambda_0|$ regardless of the multiplicity. The second term suggests that larger changes occur in larger eigenvalues. This observation can be summarized as follows: for perturbation in A, all the eigenvalues have the same sensitivity in the absolute sense, while for perturbation in B, all the eigenvalues have the same sensitivity in the relative sense.
- The above arguments show that the difference among condition numbers of a multiple eigenvalue is due to the difference among the r singular values of $X_1 Y_1^H$, the outer product of the left and right eigenvectors corresponding to λ_0. $\sigma_i(X_1 Y_1^H)$ are all 1 in the standard Hermitian case because $X_1 = Y_1$ and it has orthonormal columns. In the standard non-Hermitian case $X_1 \neq Y_1$ and neither is orthogonal, so $\sigma_i(X_1 Y_1^H)$ take r different values. In the generalized Hermitian case we have $X_1 = Y_1$ but X_1 is not orthogonal, so $\sigma_i(X_1 X_1^H)$ again take r different values. Note that one uses the B-based inner product this difference disappears because X_1 is B-orthogonal, recall the remark in the introduction.

9.6. Non-Hermitian pairs

Here we consider the case where (A, B) is a general non-Hermitian pair. In view of (9.38), our task is to bound $|\lambda_i(X\Sigma)|$ for an arbitrary square matrix X such that $\|X\|_2 \leq 1$. This is in fact the exact same problem addressed in [**151**, Thm.3.1]. Hence the analysis there can be directly applied to yield the following result.

THEOREM 9.5. *For a non-Hermitian pair (A, B) that satisfies* (9.35)*, let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ be the positive singular values of the matrix $X_1 Y_1^H$. Then, $\kappa_i(A, B, \lambda_0)$ in* (9.38) *can be expressed as*

$$\kappa_i(A, B, \lambda_0) = (1 + \tau|\lambda_0|) \left(\prod_{j=1}^{i} \sigma_j \right)^{1/i}, \quad i = 1, \ldots, r. \tag{9.43}$$

9.6.1. Structured perturbation. It is instructive to revisit the Hermitian definite case, but now allowing for non-Hermitian perturbations, that is, E, F are general matrices whose norms are bounded by 1. In this case, the condition numbers $\kappa_i(A, B, \lambda_0)$ have the characterization (9.38) (instead of (9.41)), so they have the expression (9.43), the same as that for the non-Hermitian pair.

As might be expected, the condition number under Hermitian perturbation (9.42) is always no larger than that under a non-Hermitian perturbation (9.43):

$$\frac{\kappa_i(A, B, \lambda_0; \mathbb{S})}{1 + \tau|\lambda_0|} = \left(\min_{1 \leq j \leq i} (\sigma_j \sigma_{i-j+1})^i \right)^{1/2i} \leq \left(\prod_{j=1}^{i} (\sigma_j \sigma_{i-j+1}) \right)^{1/2i}$$

$$= \left(\prod_{j=1}^{i} \sigma_j^2\right)^{1/2i} = \left(\prod_{j=1}^{i} \sigma_j\right)^{1/i} = \frac{\kappa_i(A, B, \lambda_0)}{1 + \tau|\lambda_0|}.$$

The above arguments imply that if the singular values of $X_1 Y_1^H$ are the same, then under a general non-Hermitian perturbation the condition numbers $\kappa_i(A, B, \lambda_0)$ are all the same[1], regardless of whether (A, B) is Hermitian definite or non-Hermitian. Therefore, the structure of the perturbation matrices plays an important role in the perturbation sensitivity of a multiple generalized eigenvalue. We note that the standard Hermitian case with $B \equiv I$ is an exception, in which the condition numbers are always all 1 whether or not the perturbation matrices are Hermitian.

This point of view, to focus on the effect of the structure of the perturbation, was investigated extensively in [**95**], in which (Theorem 4.5) it is shown that (among other structures they consider) the Hermitian structure of the perturbation matrices does not have any effect on the Hölder condition number.

At first sight this seems to contradict our results, which show that the Hermitian structure of the perturbation matrices does affect the condition numbers of the multiple eigenvalue λ_0. The explanation is that [**95**] treats only the Hölder condition number, which is equivalent to $\kappa_1(A, B, \lambda_0)$ in the nondefective case. Here we are identifying individual condition numbers of each of the r eigenvalues. In fact, we can see that for $i = 1$, κ_i in (9.42) and (9.43) are the same, both equal to $(1 + \tau|\lambda_0|)\sigma_1$. We can easily see that they are equal also for $i = 2$. The difference between (9.42) and (9.43) starts to take effect only for $i \geq 3$, so λ_0's multiplicity r must be at least 3. In particular, for a simple eigenvalue the Hermitian structure of the perturbation has no effect on the condition number, which is a trivial consequence of the results in [**95**].

9.6.2. Examples. Here we present two simple examples to illustrate the above results and observations.

EXAMPLE 9.3. For the Hermitian definite pair $A = \begin{bmatrix} 2000 & 0 \\ 0 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1000 & 0 \\ 0 & 1 \end{bmatrix}$ presented in [**146**, p.300], we have $\kappa_1(A, B, \lambda_0; \mathbb{S}) = \kappa_1(A, B, \lambda_0) = 3$ and $\kappa_2(A, B, \lambda_0; \mathbb{S}) = \kappa_2(A, B, \lambda_0) = 3/\sqrt{1000}$, which explains why the multiple eigenvalue $\lambda_0 = 2$ has different sensitivities. Note that in this case the structure of the perturbation has no effect on the condition numbers, because the multiplicity of λ_0 is $r < 3$.

EXAMPLE 9.4. We consider a 4-by-4 Hermitian definite pair (A, B) expressed by

$$A = W^H \Lambda W, \quad B = W^H W,$$

where $\Lambda = \mathrm{diag}(1, 1, 1, 2)$ and $W = \mathrm{diag}(1, 2, 100, 1)$, so the eigenvalues of (A, B) are $1, 1, 1, 2$. Since X that diagonalizes A, B (as in (9.39)) is $X = W^{-1} = \mathrm{diag}(1, 0.5, 0.01, 1)$ and X_1 is its first three columns, the singular values of $X_1 X_1^H$ are $\sigma_1 = 1^2, \sigma_2 = 0.5^2, \sigma_3 = 0.01^2$ (where in this example we let $\tau = 1$), hence by (9.42) it follows that $\kappa_1(A, B, 1; \mathbb{S}) = 2, \kappa_2(A, B, 1; \mathbb{S}) = 1$ and $\kappa_3(A, B, 1; \mathbb{S}) = 0.02$. Using MATLAB version 7.10 we generated 10^6 sets of random Hermitian perturbation matrices E and F such that $||E||_2, ||F||_2 \leq 1$, and examined the

[1] In fact the entire first order perturbation expansions become the same.

behavior of the three eigenvalues of the pair $(A+\epsilon E, B+\epsilon F)$ that are closest to $\lambda_0 = 1$, where we let $\epsilon = 10^{-5}$. Specifically, denoting by $\widehat{\lambda}_i$ for $i = 1, 2, 3$ the three eigenvalues of $(A+\epsilon E, B+\epsilon F)$ that are closest to 1 such that $|\widehat{\lambda}_1 - 1| \geq |\widehat{\lambda}_2 - 1| \geq |\widehat{\lambda}_3 - 1|$, we examine how large $|\widehat{\lambda}_i - 1|/\epsilon$ can be for $i = 1, 2, 3$.

We also experimented with non-Hermitian perturbations, in which case we let E, F be arbitrary non-Hermitian matrices with $||E||_2, ||F||_2 \leq 1$. In this case the condition numbers (9.43) are $\kappa_1(A, B, 1) = 2, \kappa_2(A, B, 1) = 2(1 \cdot 0.5^2)^{1/2} = 1$ and $\kappa_3(A, B, 1) = 2(1 \cdot 0.5^2 \cdot 0.01^2)^{1/3} \simeq 0.058$, in which we confirm that the first two are the same as in the above Hermitian case.

Lastly, in order to see how the Hermitian property of the matrices plays a role in the eigenvalue perturbation behaviors, we also tested with a non-Hermitian pair (A, B) that has the same eigenvalues and σ_i (of $X_1 Y_1^H$) as the above Hermitian pair. We formed such a pair (A, B) by defining $A = Y^{-H} \Lambda X^{-1}$ and $B = Y^{-H} X^{-1}$, where $\Lambda = \mathrm{diag}(1, 1, 1, 2)$, Y_1^H (the first 3 rows of Y) is set to $Z\Sigma V^H$ and X_1 (the first 3 columns of X) is set to UZ^{-1}, where U and V are randomly generated matrices with orthonormal columns, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3) = (1^2, 0.5^2, 0.01^2)$ and Z is an arbitrary nonsingular matrix[2]. Elements of the last row of Y and the last column of X were taken as random numbers. Since we have $X_1 Y_1^H = U\Sigma V^H$, we have $\kappa_1(A, B, 1) = 2, \kappa_2(A, B, 1) = 1$ and $\kappa_3(A, B, 1) = 0.058$, the same condition numbers as the above second case with non-Hermitian perturbation, as was intended. The perturbations E and F are taken as arbitrary non-Hermitian matrices.

In summary we tested under three different situations, all of which have the same $\sigma_i(X_1 Y_1^H)$: (i) Both (A, B) and (E, F) are Hermitian (shown as "Her + Her" in Table 9.6.1), (ii) (A, B) is Hermitian but (E, F) is non-Hermitian ("Her + NonHer"), and (iii) Both (A, B) and (E, F) are non-Hermitian ("NonHer + NonHer").

The results are summarized in Table 9.6.1 below, which shows the average and maximum (shown as avg. and max respectively) values of $\Delta\lambda_i/\epsilon = |\widehat{\lambda}_i - 1|/\epsilon$ among the 10^6 runs with randomly generated E and F, along with the condition numbers $\kappa_i(A, B, \lambda_0)$ (which are first order upper bounds for $\Delta\lambda_i/\epsilon$) for $i = 1, 2, 3$.

TABLE 9.6.1. Average and maximum perturbation $\Delta\lambda_i/\epsilon$ of 10^6 runs for $i = 1, 2, 3$.

	Her + Her			Her + NonHer			NonHer + NonHer		
i	avg.	max	κ_i	avg.	max	κ_i	avg.	max	κ_i
1	0.579	1.98	2.0	0.41	1.86	2.0	0.42	1.90	2.0
2	0.141	0.84	1.0	0.136	0.76	1.0	0.137	0.76	1.0
3	0.00018	0.012	0.02	0.00021	0.027	0.058	0.00021	0.027	0.058

We make the following observations.

- We confirm that κ_i is an upper bound of $\max \Delta\lambda_i/\epsilon$ for all i in all three cases (which is necessarily true in the limit $\epsilon \to 0$). Interestingly, for $i = 1$ the bound κ_i is nearly attained while for $i = 2, 3$, $\max \Delta\lambda_i/\epsilon$ is noticeably smaller than κ_i, which

[2]Note that the choice of Z does not affect the condition numbers $\kappa_i(A, B, 1)$.

suggests that for larger i it becomes more and more rare that the largest-possible perturbation is attained.

- Reflecting the fact that κ_i are the same for all the three cases for $i = 1$ and 2, we can see that $\max \Delta\lambda_i/\epsilon$ are similar in all three cases, so two eigenvalues have similar maximum sensitivities regardless of whether A, B, E, F are Hermitian or not. On the contrary, $\max \Delta\lambda_i/\epsilon$ for $i = 3$ show the somewhat different sensitivities of the third eigenvalue depending on the structure of E, F.
- The behavior of the multiple eigenvalue is remarkably similar for the latter two cases, not only in terms of $\max \Delta\lambda_i/\epsilon$ but also $\text{avg.}\Delta\lambda_i/\epsilon$. This reflects the fact that the first order expansions of λ_0 are the same for the two cases, so that the local behavior of an eigenvalue is determined solely by the singular values of $X_1 Y_1^H$, and does not depend on the structure of the matrices A and B.
- Comparing $\text{avg.}\Delta\lambda_i/\epsilon$ with $\max \Delta\lambda_i/\epsilon$, we see that the former is much smaller than the latter for larger i. For $i = 1$ the difference seems less significant.

A precise explanation for the last two observations, which necessarily involves statistical analysis, is an open problem: our discussions deal only with the maximum attainable perturbation $\max \Delta\lambda_i/\epsilon$, not with $\text{avg.}\Delta\lambda_i/\epsilon$.

9.6.3. Defective and infinite cases. So far we have treated only the case where λ_0 is a finite and nondefective multiple eigenvalue. Here we briefly consider the cases where λ_0 is infinite and/or defective.

The case $\lambda_0 = \infty$ can be treated as in [**30, 95**] simply by considering the multiple zero eigenvalue of the pair (B, A), for which the exact same discussion as above is valid.

When λ_0 is defective, Lidskii's perturbation theory [**30, 95**] shows that the leading term in λ_0's perturbation expansion is not linear in ϵ. Specifically, if λ_0 is an eigenvalue of (A, B) of multiplicity $n_1 r$ belonging to a Jordan block of dimension n_1 repeated r times, then there are $n_1 r$ eigenvalues of $(A + \epsilon E, B + \epsilon F)$ admitting the expansion

$$\widehat{\lambda}_{i,\ell} = \lambda_0 + \left(\lambda_i(Y_1^H(E - \lambda_0 F)X_1)\right)^{1/n_1} \epsilon^{1/n_1} + o(\epsilon^{1/n_1}) \tag{9.44}$$

for $i = 1, 2, \ldots, r$ and $\ell = 1, 2, \ldots, n_1$. Here $Y_1^H \in \mathbb{C}^{r\times n}$ and $X_1 \in \mathbb{C}^{n\times r}$ represent the linearly independent left and right eigenvectors of (A, B) corresponding to λ_0, and the value $\left(\lambda_i(Y_1^H(E - \lambda_0 F)X_1)\right)^{1/n_1}$ takes all the n_1 distinct n_1th roots.

We observe in (9.44) that although the leading exponent is different from that in (9.36), the sensitivities of the multiple eigenvalue are still governed by $|\lambda_i(Y_1^H(E - \lambda_0 F)X_1)|$ for $i = 1, \ldots, r$, for which we gave a bound in the above discussions. Hence all our previous results carry over to this case, and the condition numbers of λ_0 with the exponent $1/n_1$, which we define by the theoretical bounds for $\sup_{\|E\|_2\le 1, \|F\|_2\le\tau} |\lambda_i(Y_1^H(E - \lambda_0 F)X_1)|^{1/n_1}$, are

$$\kappa_{i,n_1}(A, B, \lambda_0) = \left((1 + \tau|\lambda_0|)\left(\prod_{j=1}^{i} \sigma_j\right)^{1/i}\right)^{1/n_1}, \quad i = 1, \ldots, r.$$

By (9.44), we must have $|\widehat{\lambda}_{i,\ell} - \lambda_0|/\epsilon^{1/n_1} \le \kappa_{i,n_1}(A, B, \lambda_0)$ for $i = 1, \ldots, r$ in the limit $\epsilon \to 0$ for any E and F. Note that $\kappa_{i,n_1}(A, B, \lambda_0)$ does not depend explicitly on ℓ. We

observe that in the defective case $n_1 \geq 2$, the exponent $1/n_1$ makes the difference among the condition numbers less significant than in the nondefective case. See the example below for an illustration.

9.6.3.1. *Example.* To examine the behavior of a defective multiple eigenvalue, we generate a 7-by-7 pair (A, B) defined by

$$A = Y^{-H} \begin{bmatrix} J & & & \\ & J & & \\ & & J & \\ & & & 2 \end{bmatrix} X^{-1}, \quad \text{and} \quad B = Y^{-H} X^{-1}, \tag{9.45}$$

where $J = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is a 2-by-2 Jordan block. (A, B) has a multiple eigenvalue $\lambda_0 = 1$ of multiplicity six and a simple eigenvalue 2. $Y_1^H \equiv [Y(:,2)\ Y(:,4)\ Y(:,6)]^H = Z\Sigma V^H$ and $X_1 \equiv [X(:,1)\ X(:,3)\ X(:,5)] = UZ^{-1}$ are the left and right eigenvectors corresponding to λ_0, where U and V are random matrices with orthonormal columns and Z is an arbitrary nonsingular matrix. The other rows of Y^H and columns of X do not affect the condition numbers of λ_0, so we let them take random values. We let $\Sigma = \mathrm{diag}(1^2, 0.5^2, 0.01^2)$, so that $\sigma_i(X_1 Y_1^H)$ take the same values as in the non-Hermitian case of the second example in Section 9.6.2.

Recall from (9.44) that perturbation in (A, B) generally makes λ_0 split into $n_1 r$ perturbed eigenvalues $\widehat{\lambda}_{i,\ell}$ for $i = 1, \ldots, r$ and $\ell = 1, \ldots, n_1$. (9.44) also shows that for a fixed i, $|\widehat{\lambda}_{i,\ell} - \lambda_0|$ must be nearly equal for all ℓ up to $o(\epsilon^{1/n_1})$. For the matrix pair (9.45) we have $r = 3$ and $n_1 = 2$, so we separate the six eigenvalues $\widehat{\lambda}_{i,\ell}$ into three groups according to the value of i, so that the two eigenvalues of the ith group have perturbation sensitivity governed by $\left|\lambda_i(Y_1^H(E - \lambda_0 F)X_1)\right|^{1/n_1}$.

With $\tau = 1$, the condition numbers $\kappa_{i,2}(A, B, 1)$ for the ith group for $i = 1, 2, 3$ are $\kappa_{1,2}(A, B, 1) = (2 \cdot 1)^{1/2} = \sqrt{2}, \kappa_{2,2}(A, B, 1) = \left(2 \cdot (1 \cdot 0.5^2)^{1/2}\right)^{1/2} = 1$ and $\kappa_{3,2}(A, B, 1) = \left(2 \cdot (1 \cdot 0.5^2 \cdot 0.01^2)^{1/3}\right)^{1/2} \simeq 0.24$. Comparing these with $\kappa_i(A, B, 1)$ in the example in Section 9.6.2 we see that that although $\sigma_i(X_1 Y_1^H)$ take the same values, the relative difference among the condition numbers is smaller here, due to the exponent 1/2.

Recalling that we must have $|\widehat{\lambda}_{i,\ell} - 1|/\epsilon^{1/2} \leq \kappa_{i,2}(A, B, 1)$ for small ϵ, here we examine how large $|\widehat{\lambda}_{i,\ell} - 1|/\epsilon^{1/2}$ becomes for $i = 1, 2, 3$. To do this, of the six eigenvalues of $(A + \epsilon E, B + \epsilon F)$ close to λ_0, we check the perturbation of the most perturbed, third perturbed, and the fifth perturbed ones.

In the experiment we let E, F be random non-Hermitian matrices with $||E||_2, ||F||_2 \leq 1$, let $\epsilon = 10^{-6}$ and tested with 10^6 pairs. In table 9.6.2 we report the average and maximum values of $|\widehat{\lambda}_{i,\ell} - 1|/\epsilon^{1/2}$ for $i = 1, 2, 3$.

Similarly to the experiments in Section 9.6.2 for nondefective multiple eigenvalues, we see that a defective multiple eigenvalue also exhibits different sensitivities under perturbation. We also tested with Hermitian perturbations $E = E^H$ and $F = F^H$, and obtained nearly the same results as in Table 9.6.1. This suggests that the structure of the perturbation does not play a role here.

TABLE 9.6.2. Defective matrix pair (9.45) with three 2×2 Jordan blocks, average and maximum perturbation $|\widehat{\lambda}_{i,\ell} - 1|/\epsilon^{1/2}$ of 10^6 runs for $i = 1, 2, 3$.

i	avg.	max	$\kappa_{i,2}$
1	0.511	1.21	1.41
2	0.282	0.733	1
3	0.0089	0.138	0.24

In all our experiments we had $|\widehat{\lambda}_{i,1} - \widehat{\lambda}_{i,2}|/\epsilon^{1/2} < 0.04$ for $i = 1, 2, 3$, which matches the theoretical result indicated by (9.44) that for a given i, $|\widehat{\lambda}_{i,\ell} - 1|$ are equal up to $o(\epsilon^{1/n_1})$ for all ℓ.

Finally, a comparison between Table 9.6.2 and the third case of Table 9.6.1 suggests that the relative difference among the multiple eigenvalues is smaller in the defective case, reflecting the last remark before this example.

9.7. Multiple singular value

In this section we identify the condition numbers of a multiple singular value. We start by presenting a first order perturbation expansion of a multiple singular value, which we use to define its condition numbers.

9.7.1. First order perturbation expansion of a multiple singular value. The first order perturbation expansion of a simple singular value is well-known (e.g., [**145**]): suppose a matrix A's simple singular value σ_i has left and right singular vectors u_i and v_i respectively. Then the ith singular value $\widehat{\sigma}_i$ of $A + \epsilon E$ has the first order expansion

$$\widehat{\sigma}_i = \sigma_i + u_i^H E v_i \epsilon + O(\epsilon^2). \tag{9.46}$$

Such a result for a multiple singular value does not appear to be widely known. [**70**] characterizes the first order expansions of a multiple singular value in terms of the eigenvalues of $(A + \epsilon E)^H (A + \epsilon E) - A^H A$, but here we develop a simpler expression, which is completely analogous to that for a simple singular value (9.46).

Suppose $A \in \mathbb{C}^{m \times n}$ has the SVD

$$A = U \Sigma V^H = U \begin{bmatrix} \sigma_0 I_r & \\ & \widehat{\Sigma} \end{bmatrix} V^H,$$

where $\widehat{\Sigma}$ contains the singular values not equal to σ_0 (the multiple singular value of multiplicity r). Then, denoting $U = [U_1 \ U_2]$ and $V = [V_1 \ V_2]$ where U_1 and V_1 have r columns, $A + \epsilon E$ for $E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$ with $\|E\|_2 = 1$ can be expressed as

$$\begin{aligned} A + \epsilon E &= U \begin{bmatrix} \sigma_0 I_r & \\ & \widehat{\Sigma} \end{bmatrix} V^H + \epsilon \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \\ &= U \begin{bmatrix} \sigma_0 I_r + \epsilon U_1^H E V_1 & \epsilon U_1^H E V_2 \\ \epsilon U_2^H E V_1 & \widehat{\Sigma} + \epsilon U_2^H E V_2 \end{bmatrix} V^H. \end{aligned}$$

Hence the singular values of $A+\epsilon E$ are those of $\begin{bmatrix}\sigma_0 I_r + \epsilon U_1^H E V_1 & \epsilon U_1^H E V_2 \\ \epsilon U_2^H E V_1 & \widehat{\Sigma} + \epsilon U_2^H E V_2\end{bmatrix}$. Since $\|U_1^H E V_2\|_2, \|U_2^H E V_1\|_2 \leq 1$, denoting by $\widehat{\sigma}_i$ for $i = 1, \ldots, r$ the singular values of the matrix $\sigma_0 I_r + \epsilon U_1^H E V_1$ and using Theorem 3 in [**97**], we see that there are r singular values η_i ($1 \leq i \leq r$) of $A+\epsilon E$ that satisfy

$$|\eta_i - \widehat{\sigma}_i| \leq \frac{2\epsilon^2}{gap + \sqrt{gap^2 + 4\epsilon^2}},$$

where *gap* denotes the minimum distance between σ_0 and the singular values of $\widehat{\Sigma}$. Since $gap > 0$ by assumption, this bound is $O(\epsilon^2)$, therefore $\widehat{\sigma}_i$ and η_i match up to first order in ϵ. Therefore we conclude that the multiple singular value of A has the first order expansion

$$\widehat{\sigma}_i = \sigma_i(\sigma_0 I_r + \epsilon U_1^H E V_1) + O(\epsilon^2) \quad \text{for} \quad i = 1, \ldots, r. \tag{9.47}$$

Comparing (9.46) with (9.47) we easily see that the latter is a direct generalization of (9.46), much in the same way (9.36) generalizes the expansion of a simple eigenvalue.

9.7.2. Condition numbers of a multiple singular value. Now we identify the perturbation sensitivities of A's multiple singular value σ_0 of multiplicity r. In light of (9.47), a natural way to define the condition numbers of σ_0 analogously to (9.37) is

$$\kappa_i(A, \sigma_0) \equiv \lim_{\epsilon \to 0} \sup_{\|E\|_2 \leq 1} \frac{1}{\epsilon} \left| \sigma_0 - \sigma_i(\sigma_0 I_r + \epsilon U_1^H E V_1) \right| \quad \text{for} \quad i = 1, \ldots, r. \tag{9.48}$$

We show that $\kappa_i(A, \sigma_0)$ for $i = 1, \ldots, r$ are always uniformly 1.

THEOREM 9.6.

$$\kappa_i(A, \sigma_0) = 1 \quad \text{for} \quad i = 1, \ldots, r.$$

Proof By Weyl's theorem [**142**, p. 69] we have $|\sigma_i(\sigma_0 I_r + \epsilon U_1^H E V_1) - \sigma_0| \leq \epsilon$ for any $\epsilon > 0$, hence $\kappa_i(A, \sigma_0) \leq 1$. Therefore it suffices to prove that $\sigma_0 I_r + \epsilon U_1^H E V_1$ can have r singular values equal to $\sigma_0 + \epsilon$, which is true when $U_1^H E V_1 = I_r$. This is the case for example when $E = U_1 V_1^H$. □

CHAPTER 10

Perturbation of eigenvectors

We now turn to the perturbation of eigenvectors or eigenspaces. The celebrated Davis-Kahan $\tan\theta$ theorem bounds the tangent of the angles between an approximate and an exact invariant subspace of a Hermitian matrix. When applicable, it gives a sharper bound than the $\sin\theta$ theorem. However, the $\tan\theta$ theorem requires more restrictive conditions on the spectrums, demanding that the entire approximate eigenvalues (Ritz values) lie above (or below) the set of exact eigenvalues corresponding to the orthogonal complement of the invariant subspace. In this chapter we show that the conditions of the $\tan\theta$ theorem can be relaxed, in that the same bound holds even when the Ritz values lie both below and above the exact eigenvalues, but not vice versa.

We then investigate the Rayleigh-Ritz process and present new bounds for the accuracy of the Ritz vectors, which can be regarded as refined versions of the theorems by Saad and Knyazev. We also derive what we call the $\cos\theta$ and $1/\tan\theta$ theorems, which measure the distance instead of nearness of two subspaces.

Introduction. Recall the description of the Davis-Kahan $\tan\theta$ and $\sin\theta$ theorems in Section 2.10.1. For ease of reference here we restate the $\tan\theta$ theorem.

Let A be an n-by-n Hermitian matrix, and let $X = [X_1\ X_2]$ where $X_1 \in \mathbb{C}^{n\times k}$ be an exact unitary eigenvector matrix of A so that $X^*AX = \operatorname{diag}(\Lambda_1, \Lambda_2)$ is diagonal. Also let $Q_1 \in \mathbb{C}^{n\times k}$ be an orthogonal matrix $Q_1^*Q_1 = I_k$, and define the residual matrix

$$R = AQ_1 - Q_1A_1, \quad \text{where} \quad A_1 = Q_1^*AQ_1. \tag{10.1}$$

The eigenvalues of A_1 are the Ritz values with respect to Q_1. Suppose that the Ritz values $\lambda(A_1)$ lie entirely above (or below) $\lambda(\Lambda_2)$, the exact eigenvalues corresponding to X_2. Specifically, suppose that there exists $\delta > 0$ such that $\lambda(A_1)$ lies entirely in $[\beta, \alpha]$ while $\lambda(\Lambda_2)$ lies entirely in $[\alpha + \delta, \infty)$, or in $(-\infty, \beta - \delta]$. Then, the $\tan\theta$ theorem gives an upper bound for the tangents of the canonical angles between Q_1 and X_1,

$$\|\tan\angle(Q_1, X_1)\| \le \frac{\|R\|}{\delta}, \tag{10.2}$$

where $\|\cdot\|$ denotes any unitarily invariant norm. $\tan\angle(Q_1, X_1)$ is the matrix whose singular values are the tangents of the k canonical angles between the n-by-k orthogonal matrices Q_1 and X_1.

The $\sin\theta$ theorem, on the other hand, asserts the same bound, but in terms of the sine instead of tangent:

$$\|\sin\angle(Q_1, X_1)\| \leq \frac{\|R\|}{\delta}. \tag{10.3}$$

In the context of the Rayleigh-Ritz process, we have the matrix

$$\widetilde{A} = Q^* A Q = \begin{bmatrix} A_1 & \widetilde{R}^* \\ \widetilde{R} & A_2 \end{bmatrix}, \tag{10.4}$$

in which A, Q_1 and A_1 are known. Note that $\|\widetilde{R}\|$ can be computed because $\|\widetilde{R}\| = \|AQ_1 - Q_1 A_1\| = \|R\|$ for any unitarily invariant norm.

Recall the comparison between the $\tan\theta$ theorem (10.2) and the $\sin\theta$ theorem (10.3), as described in Section 2.10.1. In particular, we mentioned that the $\tan\theta$ theorem requires more restricted conditions on the condition of the spectrums of Λ_2 and A_1).

The goal of this chapter is to show that the condition in the $\tan\theta$ theorem can be relaxed by proving that the bound (10.2) still holds true in the first (but not in the second) case in Section 2.10.1. In other words, the conclusion of the $\tan\theta$ theorem is valid even when the Ritz values $\lambda(A_1)$ lie both below and above the exact eigenvalues $\lambda(\Lambda_2)$.

We will also revisit the counterexample described in [**29**] that indicates the restriction on the spectrums is necessary in the $\tan\theta$ theorem. This does not contradict our result because, as we will see, its situation corresponds to the second case above. We also extend the result to the generalized $\tan\theta$ theorem, in which the dimensions of Q_1 and X_1 are allowed to be different.

In the second part this chapter we investigate the Rayleigh-Ritz process, which we reviewed in Section 2.10.2. We derive refined bounds for the angles between exact eigenvectors and approximate eigenvectors (Ritz vectors) computed by the Rayleigh-Ritz process. We first point out the structure and properties of the residual matrix $R = A\widehat{X} - \widehat{X}\widehat{\Lambda}$ that are typically observed in practice when computing extremal eigenpairs of a large Hermitian matrix. We then present bounds for the accuracy of the Ritz vectors, which can be arbitrarily sharper than previously known bounds by Saad (Theorem 2.5) and Knyazev (Theorem 2.6). The bounds are also tighter than those given by directly applying the $\tan\theta$ or $\sin\theta$ theorems. We then derive what might be called the $\cos\theta$ theorem, which measures the distance, not nearness, of subspaces. This hints on an efficient execution of an inexact Rayleigh-Ritz process, whose further investigation is left as a future research topic.

Notations of the chapter: Recall that $\|\cdot\|$ denotes an arbitrary unitarily invariant norm. In this chapter $\lambda(A)$ denotes the spectrum, or the set of eigenvalues of a square matrix A.

10.1. The $\tan\theta$ theorem under relaxed conditions

10.1.1. Preliminaries. We first prove a lemma that we use in the proof of our main result.

LEMMA 10.1. *Let $X \in \mathbb{C}^{m\times n}, Y \in \mathbb{C}^{n\times r}, Z \in \mathbb{C}^{r\times s}$ have the singular value decompositions $X = U_X \Sigma_X V_X^*$, $Y = U_Y \Sigma_Y V_Y^*$ and $Z = U_Z \Sigma_Z V_Z^*$, where the singular values are arranged in*

descending order. Then for any unitarily invariant norm $\|\cdot\|$,

$$\|XYZ\| \le \|Y\|_2 \|\widetilde{\Sigma}_X \widetilde{\Sigma}_Z\|, \tag{10.5}$$

where $\widetilde{\Sigma}_X = diag(\sigma_1(X), \ldots, \sigma_p(X)), \widetilde{\Sigma}_Z = diag(\sigma_1(Z), \ldots, \sigma_p(Z))$ *are diagonal matrices of the* p *largest singular values where* $p = \min\{m, n, r, s\}$. *Moreover, analogous results hold for any combination of* $\{X, Y, Z\}$, *that is,* $\|XYZ\| \le \|X\|_2 \|\widetilde{\Sigma}_Y \widetilde{\Sigma}_Z\|$ *and* $\|XYZ\| \le \|Z\|_2 \|\widetilde{\Sigma}_X \widetilde{\Sigma}_Y\|$.

PROOF. In the majorization property of singular values of a matrix product $\sum_{i=1}^k \sigma_i(AB) \le \sum_{i=1}^k \sigma_i(A)\sigma_i(B)$ for all $k = 1, \ldots, p$ [**81**, p.177], we let $A := X$ and $B := YZ$ to get

$$\begin{aligned}\sum_{i=1}^k \sigma_i(XYZ) &\le \sum_{i=1}^k \sigma_i(X)\sigma_i(YZ) \\ &\le \sum_{i=1}^k \sigma_i(X)\sigma_i(Z)\|Y\|_2 \\ &= \|Y\|_2 \sum_{i=1}^k \sigma_i(\Sigma_X \Sigma_Z) \quad \text{for} \quad k = 1, \ldots, p.\end{aligned}$$

(10.5) now follows from Ky-Fan's theorem [**80**, p.445]. A similar argument proves the inequality for the other two combinations. □

We next recall the CS decomposition (2.35), summarized in Section 2.10.1. Applied to the unitary matrix $W = Q^*X = \begin{bmatrix} Q_1^*X_1 & Q_1^*X_2 \\ Q_2^*X_1 & Q_2^*X_2 \end{bmatrix}$, the CS decomposition states that there exist unitary matrices $U_1 \in \mathbb{C}^{k\times k}$, $U_2 \in \mathbb{C}^{(n-k)\times(n-k)}$, $V_1 \in \mathbb{C}^{k\times k}$ and $V_2 \in \mathbb{C}^{(n-k)\times(n-k)}$ such that $\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}^* W \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}$ can be expressed as $\left[\begin{array}{c|cc} C & 0 & -S \\ \hline 0 & I_{n-2k} & 0 \\ S & 0 & C \end{array}\right]$ when $k < \frac{n}{2}$, $\begin{bmatrix} C & -S \\ S & C \end{bmatrix}$ when $k = \frac{n}{2}$, and $\left[\begin{array}{cc|c} I_{2k-n} & 0 & 0 \\ 0 & C & -S \\ \hline 0 & S & C \end{array}\right]$ when $k > \frac{n}{2}$, where $C = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_p)$ and $S = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_p)$, in which $p = \min\{k, n-k\}$. The nonnegative quantities $\theta_1 \le \cdots \le \theta_p$ are the canonical angles between Q_1 and V_1. Note that they are also the canonical angles between Q_2 and V_2.

10.1.2. Main result. We now prove the $\tan\theta$ theorem under a relaxed condition.

THEOREM 10.1. *Let* $A \in \mathbb{C}^{n\times n}$ *be a Hermitian matrix and let* $X = [X_1\ X_2]$ *be its unitary eigenvector matrix so that* $X^*AX = diag(\Lambda_1, \Lambda_2)$ *is diagonal where* X_1 *and* Λ_1 *have* k *columns. Let* $Q_1 \in \mathbb{C}^{n\times k}$ *be orthogonal, and let* $R = AQ_1 - Q_1A_1$, *where* $A_1 = Q_1^*AQ_1$.

Suppose that $\lambda(\Lambda_2)$ lies in $[a,b]$ and $\lambda(A_1)$ lies in the union of $(-\infty, a-\delta]$ and $[b+\delta,\infty)$. Then

$$\|\tan\angle(Q_1, X_1)\| \le \frac{\|R\|}{\delta}. \tag{10.6}$$

PROOF. Note that $W = Q^*X$ is the unitary eigenvector matrix of $\widetilde{A} = Q^*AQ = \begin{bmatrix} A_1 & \widetilde{R}^* \\ \widetilde{R} & A_2 \end{bmatrix}$ as in (10.4). Partition $W = \begin{bmatrix} Q_1^*X_1 & Q_1^*X_2 \\ Q_2^*X_1 & Q_2^*X_2 \end{bmatrix} = [W_1\ W_2]$, so that the columns of W_2 are the eigenvectors of $\widetilde{A}$ corresponding to $\lambda(\Lambda_2)$. Further partition $W_2 = \begin{bmatrix} Q_1^*X_2 \\ Q_2^*X_2 \end{bmatrix} = \begin{bmatrix} W_2^{(1)} \\ W_2^{(2)} \end{bmatrix}$ so that $W_2^{(1)}$ is k-by-$(n-k)$. The first k rows of $\widetilde{A}W_2 = W_2\Lambda_2$ is

$$A_1W_2^{(1)} + \widetilde{R}^*W_2^{(2)} = W_2^{(1)}\Lambda_2,$$

which is equivalent to

$$A_1W_2^{(1)} - W_2^{(1)}\Lambda_2 = -\widetilde{R}^*W_2^{(2)}. \tag{10.7}$$

For definiteness we discuss the case $k \le \frac{n}{2}$. The case $k > \frac{n}{2}$ can be treated with few modifications. By the CS decomposition we know that there exist unitary matrices $U_1 \in \mathbb{C}^{k\times k}, U_2 \in \mathbb{C}^{(n-k)\times(n-k)}$ and $V \in \mathbb{C}^{(n-k)\times(n-k)}$ such that $W_2^{(1)} = U_1\widetilde{S}V^*$ and $W_2^{(2)} = U_2\widetilde{C}V^*$, where $\widetilde{C} = \operatorname{diag}(I_{n-2k}, C) \in \mathbb{C}^{(n-k)\times(n-k)}$, $\widetilde{S} = [0_{k,n-2k}\ \ -S] \in \mathbb{C}^{k\times(n-k)}$ in which $C = \operatorname{diag}(\cos\theta_1,\dots,\cos\theta_k)$ and $S = \operatorname{diag}(\sin\theta_1,\dots,\sin\theta_k)$. Hence we can express (10.7) as

$$A_1U_1\widetilde{S}V^* - U_1\widetilde{S}V^*\Lambda_2 = -\widetilde{R}^*U_2\widetilde{C}V^*. \tag{10.8}$$

We claim that $\widetilde{C}$ is nonsingular. To see this, suppose on the contrary that there exists i such that $\cos\theta_i = 0$, which makes $\widetilde{C}$ singular. Defining $j = n-2k+i$ this means $W_2^{(2)}Ve_j = 0$ where e_j is the jth column of I_{n-k}, so the jth column of $W_2^{(2)}V$ is all zero.

Now, by $\widetilde{A}W_2 = W_2\Lambda_2$ we have $\widetilde{A}W_2V = W_2V(V^*\Lambda_2V)$. Taking the jth column yields

$$\widetilde{A}W_2Ve_j = W_2V(V^*\Lambda_2V)e_j.$$

Since W_2Ve_j is nonzero only in its first k elements, we get

$$\begin{bmatrix} A_1 \\ \widetilde{R} \end{bmatrix} W_2^{(1)}Ve_j = W_2V(V^*\Lambda_2V)e_j,$$

the first k elements of which is

$$A_1W_2^{(1)}Ve_j = W_2^{(1)}V(V^*\Lambda_2V)e_j.$$

Now define $v = W_2^{(1)}Ve_j$ and let $\gamma = (a+b)/2$. Subtracting γv we get

$$(A_1 - \gamma I)v = W_2^{(1)}V(V^*(\Lambda_2 - \gamma I)V)e_j.$$

Defining $\widehat{A}_1 = A_1 - \gamma I$ and $\widehat{\Lambda}_2 = \Lambda_2 - \gamma I$ and taking the spectral norm we get

$$\|\widehat{A}_1 v\|_2 = \|W_2^{(1)}\widehat{\Lambda}_2 Ve_j\|_2.$$

Note by assumption that defining $c = \frac{1}{2}(b-a)$ the eigenvalues of $\widehat{\Lambda}_2$ lie in $[-c, c]$ and those of $\widehat{A}_1$ lie in the union of $[c+\delta, \infty)$ and $(-\infty, c-\delta]$, so noting that $\|v\|_2 = \|e_j\|_2 = 1$ and $\|W_2^{(1)}\|_2 = \|\widetilde{C}\|_2 \leq 1$, we must have $\sigma_{\min}(\widehat{A}_1) \leq \|W_2^{(1)}\widehat{\Lambda}_2 V e_j\|_2 \leq \|\widehat{\Lambda}_2\|_2$. However, this contradicts the assumptions, which require $\delta + c < \sigma_{\min}(\widehat{A}_1)$ and $\|\widehat{\Lambda}_2\|_2 \leq c$. Therefore we conclude that $\widetilde{C}$ must be invertible.

Hence we can right-multiply $V\widetilde{C}^{-1}$ to (10.8), which yields

$$\begin{aligned} -\widetilde{R}^* U_2 &= A_1 U_1 \widetilde{S} V^* V \widetilde{C}^{-1} - U_1 \widetilde{S} V^* \Lambda_2 V \widetilde{C}^{-1} \\ &= A_1 U_1 \widetilde{S}\widetilde{C}^{-1} - U_1 \widetilde{S}\widetilde{C}^{-1} \cdot (\widetilde{C} V^* \Lambda_2 V \widetilde{C}^{-1}). \end{aligned}$$

As above we introduce a "shift" $\gamma = (a+b)/2$ such that

$$\begin{aligned} -\widetilde{R}^* U_2 &= A_1 U_1 \widetilde{S}\widetilde{C}^{-1} - (\gamma U_1 \widetilde{S}\widetilde{C}^{-1} - \gamma U_1 \widetilde{S}\widetilde{C}^{-1}) - U_1 \widetilde{S}\widetilde{C}^{-1} \cdot (\widetilde{C} V^* \Lambda_2 V \widetilde{C}^{-1}) \\ &= (A_1 - \gamma I) U_1 \widetilde{S}\widetilde{C}^{-1} - U_1 \widetilde{S}\widetilde{C}^{-1} \cdot (\widetilde{C} V^* (\Lambda_2 - \gamma I) V \widetilde{C}^{-1}) \\ &= \widehat{A}_1 U_1 \widetilde{S}\widetilde{C}^{-1} - U_1 \widetilde{S} V^* \widehat{\Lambda}_2 V \widetilde{C}^{-1}. \end{aligned}$$

Taking a unitarily invariant norm and using $\|\widetilde{R}\| = \|R\|$ and the triangular inequality yields

$$\begin{aligned} \|R\| &\geq \|\widehat{A}_1 U_1 \widetilde{S}\widetilde{C}^{-1}\| - \|(U_1\widetilde{S})(V^*\widehat{\Lambda}_2 V)\widetilde{C}^{-1}\| \\ &\geq \sigma_{\min}(\widehat{A}_1)\|\widetilde{S}\widetilde{C}^{-1}\| - \|(U_1\widetilde{S})(V\widehat{\Lambda}_2 V^*)\widetilde{C}^{-1}\|. \end{aligned}$$

We now appeal to Lemma 10.1 substituting $X \leftarrow U_1\widetilde{S}, Y \leftarrow V^*\widehat{\Lambda}_2 V, Z \leftarrow \widetilde{C}^{-1}$. In doing so we note that $\widetilde{\Sigma}_X\widetilde{\Sigma}_Z = \operatorname{diag}(\tan\theta_k, \ldots, \tan\theta_1)$ so $\|\widetilde{\Sigma}_X\widetilde{\Sigma}_Z\| = \|SC^{-1}\| = \|\widetilde{S}\widetilde{C}^{-1}\| = \|\tan\angle(Q_1, X_1)\|$, so we get

$$\begin{aligned} \|R\| &\geq \sigma_{\min}(\widehat{A}_1)\|SC^{-1}\| - \|V^*\widehat{\Lambda}_2 V\|_2\|SC^{-1}\| \\ &= \sigma_{\min}(\widehat{A}_1)\|SC^{-1}\| - \|\widehat{\Lambda}_2\|_2\|SC^{-1}\| \\ &= \|\tan\angle(Q_1, X_1)\| \left(\sigma_{\min}(\widehat{A}_1) - \|\widehat{\Lambda}_2\|_2\right). \end{aligned}$$

Using $\sigma_{\min}(A_1) - \|\Lambda_2\|_2 \geq (c+\delta) - c = \delta$, we conclude that

$$\|\tan\angle(Q_1, X_1)\| \leq \frac{\|R\|}{\sigma_{\min}(A_1) - \|\Lambda_2\|_2} \leq \frac{\|R\|}{\delta}.$$

□

Remarks. Below are two remarks on the $\tan\theta$ theorem with relaxed conditions, Theorem 10.1.

- Practical situations to which the relaxed theorem is applicable but not the original include the following two cases:
  (i) When extremal (both smallest and largest) eigenpairs are sought, for example by the Lanczos algorithm (e.g., [**6, 127**]). In this case Q_1 tends to approximately contain the eigenvectors corresponding to the largest and smallest eigenvalues of A, so we may directly have the situation in Theorem 10.1.

(ii) When internal eigenpairs are sought. In this case the exact (undesired) eigenvalues $\lambda(\Lambda_2)$ lie below and above $\lambda(A_1)$, so Theorem 10.1 is not applicable. However, if the residual $\|R\|$ is sufficiently small then we must have $\lambda(A_1) \simeq \lambda(\Lambda_1)$ and $\lambda(A_2) \simeq \lambda(\Lambda_2)$, in which case the Ritz values $\lambda(A_2)$ lie both below and above the eigenvalues $\lambda(\Lambda_1)$. We can then invoke Theorem 10.1 with the subscripts 1 and 2 swapped, see below for an example.

- For the $\tan 2\theta$ theorem we cannot make a similar relaxation in the conditions on the spectrums. Note that in the $\tan 2\theta$ theorem the gap δ is defined as the separation between the two sets of Ritz values $\lambda(A_1)$ and $\lambda(A_2)$ (instead of $\lambda(\Lambda_2)$), so there is no separate situations in which one spectrum lies both below and above the other, unlike in the $\tan\theta$ theorem. To see that in such cases $\frac{\|R\|}{\widetilde{\delta}}$ (where $\widetilde{\delta}$ is the separation between $\lambda(A_1)$ and $\lambda(A_2)$) is not an upper bound of $\|\frac{1}{2}\tan 2\angle(Q_1, X_1)\|$, we consider the example (10.9) below, in which we have $\frac{\|R\|_2}{\widetilde{\delta}} = \frac{1/\sqrt{2}}{1/\sqrt{2}} = 1$ but $\|\frac{1}{2}\tan 2\angle(Q_1, X_1)\|_2 = \infty$.

The counterexample in [**29**]. [**29**] considers the following example in which the spectrums of A_1 and Λ_2 satisfy the conditions of the $\sin\theta$ theorem but not the original $\tan\theta$ theorem.

$$A = \begin{bmatrix} 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{10.9}$$

A has eigenvalues $0, 1, -1$, and the exact angle between Q_1 and the eigenvector $X_1 = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]^T$ corresponding to the zero eigenvalue satisfies $\tan\angle(Q_1, X_1) = 1$. We can also compute $A_1 = 0$ so $\delta = 1$, and $\|R\|_2 = 1/\sqrt{2}$. In this case $\lambda(\Lambda_2) = \{1, -1\}$ lies on both sides of $A_1 = 0$, which violates the assumption in the original $\tan\theta$ theorem. In fact, $\|R\|_2/\delta = 1/\sqrt{2}$ is not an upper bound of $\|\tan\angle(Q_1, X_1)\|_2 = 1$.

Let us now examine (10.9) in terms of our relaxed $\tan\theta$ theorem, Theorem 10.1. The above setting does not satisfy the assumption in Theorem 10.1 either. In particular, the situation between $\lambda(A_1)$ and $\lambda(\Lambda_2)$ corresponds to the second case (b) in Section 2.10.1, which the relaxed $\tan\theta$ theorem does not cover. However, in light of the fact $\angle(Q_1, X_1) = \angle(Q_2, X_2)$ for all the p canonical angles, we can attempt to bound $\|\tan\angle(Q_1, X_1)\|$ via bounding $\|\tan\angle(Q_2, X_2)\|$. We have $\lambda(A_2) = \pm\frac{1}{\sqrt{2}}$ and $\lambda(\Lambda_1) = 0$, so the assumptions in Theorem 10.1 (in which we swap the subscripts 1 and 2) are satisfied with $\delta = 1/\sqrt{2}$. Therefore we can invoke the $\tan\theta$ theorem, and get the correct and sharp bound $\|\tan\angle(Q_2, X_2)\| \le \|R\|/\delta = 1$. We note that the original $\tan\theta$ theorem still cannot be invoked because the assumptions are violated.

10.2. The generalized $\tan\theta$ theorem with relaxed conditions

[**29**] also proves the *generalized* $\tan\theta$ theorem, in which the dimension of Q_1 is smaller than that of X_1. Here we show that the same relaxation on the condition can be attained for the generalized $\tan\theta$ theorem. We prove the below theorem, in which X_1 now has $\ell(\ge k)$ columns.

THEOREM 10.2. *Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix and let $X = [X_1\ X_2]$ be its unitary eigenvector matrix so that $X^*AX = \mathrm{diag}(\Lambda_1, \Lambda_2)$ is diagonal where X_1 and Λ_1 have $\ell(\geq k)$ columns. Let $Q_1 \in \mathbb{C}^{n\times k}$ be orthogonal, and let $R = AQ_1 - Q_1A_1$, where $A_1 = Q_1^*AQ_1$. Suppose that $\lambda(\Lambda_2)$ lies in $[a, b]$ and $\lambda(A_1)$ lies in the union of $(-\infty, a-\delta]$ and $[b+\delta, \infty)$. Then*

$$\|\tan\angle(Q_1, X_1)\| \leq \frac{\|R\|}{\delta}. \tag{10.10}$$

PROOF. The proof is almost the same as that for Theorem 10.1, so we only highlight the differences.

We discuss the case $k \leq \ell \leq \frac{n}{2}$; other cases are analogous. We partition $W_2 = \begin{bmatrix} Q_1^*X_2 \\ Q_2^*X_2 \end{bmatrix} = \begin{bmatrix} W_2^{(1)} \\ W_2^{(2)} \end{bmatrix}$ where $W_2^{(1)}$ is k-by-$(n-\ell)$. There exist unitary matrices $U_1 \in \mathbb{C}^{k\times k}, U_2 \in \mathbb{C}^{(n-k)\times(n-k)}$ and $V \in \mathbb{C}^{(n-\ell)\times(n-\ell)}$ such that $W_2^{(1)} = U_1\widetilde{S}V^*$ and $W_2^{(2)} = U_2\widetilde{C}V^*$, where $\widetilde{C} = \begin{bmatrix} \mathrm{diag}\,(I_{n-k-\ell}, C) \\ 0_{\ell-k,n-\ell} \end{bmatrix} \in \mathbb{C}^{(n-k)\times(n-\ell)}$ and $\widetilde{S} = [0_{k,n-k-\ell} \quad -S] \in \mathbb{C}^{k\times(n-\ell)}$, in which $C = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_k)$ and $S = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_k)$. We then right-multiply (10.8) by $\mathrm{diag}\,(I_{n-k-\ell}, C^{-1})$, which yields

$$-\widetilde{R}^*U_2 \begin{bmatrix} I_{n-\ell} \\ 0_{\ell-k,n-\ell} \end{bmatrix} = A_1U_1\widetilde{S}\mathrm{diag}\left(I_{n-k-\ell}, C^{-1}\right) - U_1\widetilde{S}V^*\Lambda_2V\mathrm{diag}\left(I_{n-k-\ell}, C^{-1}\right).$$

Noting that the k largest singular values of $\mathrm{diag}\,(I_{n-k-\ell}, C^{-1})$ are $1/\cos\theta_k, \ldots, 1/\cos\theta_1$ and using Lemma 10.1 we get

$$\begin{aligned} \|R\| &\geq \left\| \widetilde{R}^*U_2 \begin{bmatrix} I_{n-\ell} \\ 0_{\ell-k,n-\ell} \end{bmatrix} \right\| \\ &\geq \sigma_{\min}(\widehat{A}_1)\|SC^{-1}\| - \|\widehat{\Lambda}_2\|_2\|SC^{-1}\| \\ &= \|\tan\angle(Q_1, X_1)\| \left(\sigma_{\min}(\widehat{A}_1) - \|\widehat{\Lambda}_2\|_2\right), \end{aligned}$$

which is (10.10). □

10.3. Refined Rayleigh-Ritz approximation bound

The rest of this chapter is concerned with the Rayleigh-Ritz process, which we reviewed in Section 2.10.2.

For simplicity we assume that some of the smallest eigenvalues of A are desired, and so a k-dimensional trial subspace $\mathrm{span}(Q_1)$ approximates the corresponding eigenspace. Consider a Hermitian matrix that is involved in the Rayleigh-Ritz process with the subspace $\mathrm{span}(Q_1)$: For a unitary matrix $[\widehat{X}_1\ \widehat{X}_2]$,

$$[\widehat{X}_1\ \widehat{X}_2]^*A[\widehat{X}_1\ \widehat{X}_2] = \begin{bmatrix} \widehat{\Lambda} & \widetilde{R}^* \\ \widetilde{R} & X_2^*AX_2 \end{bmatrix}, \tag{10.11}$$

where $\widehat{\Lambda} = \mathrm{diag}(\theta_1, \ldots, \theta_k)$ is the matrix of Ritz values arranged in increasing order, and $\widetilde{R} = \widehat{X}_2^*A\widehat{X}_1$. The matrix of Ritz vectors $\widehat{X}_1 = [\widehat{x}_1, \ldots, \widehat{x}_k]$ and Q_1 are related by $\widehat{X} = Q_1Y$

where $Q_1^*AQ_1 = Y\widehat{\Lambda}Y^*$ is the symmetric eigendecomposition of $Q_1^*AQ_1$. After performing the Rayleigh-Ritz process, the norms of each column of $\widetilde{R}$ are known because they are equal to the norm of the residual: $\|\widetilde{R}(:,i)\|_2 = \|A\widehat{x}_i - \theta_i\widehat{x}_i\|_2 = \|R(:,i)\|_2$. An important aspect is that we typically have $\|\widetilde{R}(:,1)\|_2 \lesssim \|\widetilde{R}(:,2)\|_2 \ldots \lesssim \|\widetilde{R}(:,m)\|_2$ with $\|\widetilde{R}(:,1)\|_2 \ll \|\widetilde{R}(:,m)\|_2$, because extremal eigenvalues tend to converge much faster than interior ones [**6, 91**].

In such circumstances, how accurate can we say a Ritz vector (or subspace) is? In particular, for the smallest Ritz pairs (which are typically of foremost interest in applications), can we obtain bounds that are sharper than any existing bound? This was our original to investigate what follows in the remainder of this chapter.

To get an idea of the situation, perhaps it helps to think of a specific example such as $m = 10$ and

$$A = \begin{bmatrix} \mathrm{diag}(0.1, 0.2, \ldots, 1) & \widetilde{R}^* \\ \widetilde{R} & A_2 \end{bmatrix}, \tag{10.12}$$

where $\widetilde{R} = [\widetilde{R}_1\ \widetilde{R}_2\ \ldots\ \widetilde{R}_{10}]$ and $\|\widetilde{R}_1\|_2 = 10^{-5}$ and $\|\widetilde{R}_i\|_2 = 10^{-1}$ for $i = 2, \ldots, 10$.

We note that a number of studies exist concerning the accuracy of Ritz values and vectors. In [**83**] the convergence of Ritz pairs on a general non-Hermitian matrix is studied. The authors of [**136**] investigates the accuracy of Ritz values for the Hermitian case, giving a priori bounds in the sense that only eigenvalues of the matrix and the angle between the subspace and eigenvector of interest are involved. The quality of the Galerkin approximation (a special (orthogonal, for Hermitian A) case of which is the Rayleigh-Ritz) when applied to a general linear operator in a Hilbert space is studied in [**13**]. Here our focus is on the accuracy of the Ritz vectors when applied to a Hermitian matrix A.

10.4. New bounds for the angles between Ritz vectors and exact eigenvectors

Let

$$(\widetilde{A} =)\ Q^*AQ = \begin{bmatrix} A_1 & 0 & \widetilde{R}_1^* \\ 0 & A_2 & \widetilde{R}_2^* \\ \widetilde{R}_1 & \widetilde{R}_2 & A_3 \end{bmatrix}, \tag{10.13}$$

where $A_1 = \mathrm{diag}(\theta_1, \ldots, \theta_\ell)$ and $A_2 = \mathrm{diag}(\theta_{\ell+1}, \ldots, \theta_k)$ are diagonal matrices of Ritz values such that $\theta_1 \le \theta_2 \le \cdots \le \theta_k$. The known quantities are A_1, A_2, the 2-norms of each of the columns of $\widetilde{R}_1, \widetilde{R}_2$. In the context of the previous subsection, we regard the projection subspace Q_1 as k-dimensional, so the eigenvalues of A_1 and A_2 are the Ritz values.

Let $W = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix}$ be the orthognoal eigenvector matrix of $\widetilde{A}$ corresponding to the ℓ smallest eigenvalues $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_\ell)$. Then the second block of $\widetilde{A}W = W\Lambda_1$ gives

$$A_2W_2 + \widetilde{R}_2^*W_3 = W_2\Lambda_1,$$

which is equivalent to

$$A_2W_2 - W_2\Lambda_1 = -\widetilde{R}_2^*W_3.$$

Therefore, assuming $\lambda_\ell < \theta_{\ell+1}$, by Lemma 2.1 we have

$$\|W_2\| \leq \frac{\|\widetilde{R}_2^* W_3\|}{|\lambda_\ell - \theta_{\ell+1}|}.$$

Hence using the fact [**75**, p. 327] that $\|XY\| \leq \|X\|_2 \|Y\|$ for any unitarily invariant norm we have

$$\|W_2\| \leq \frac{\|\widetilde{R}_2\|_2}{|\lambda_\ell - \theta_{\ell+1}|} \|W_3\| \ (\equiv \gamma_{23}\|W_3\|)\,. \tag{10.14}$$

we note in passing that this is equivalent to Knyazev's result in [**90**, Thm. 3.2] for the finite-dimensional case. We believe the derivation here is much simpler.

10.4.1. Refined versions of theorems by Saad and Knyazev. Now the third block of $\widetilde{A}W = W\Lambda_1$ gives

$$\widetilde{R}_1 W_1 + \widetilde{R}_2 W_2 + A_3 W_3 = W_3 \Lambda_1,$$

hence

$$A_3 W_3 - W_3 \Lambda_1 + \widetilde{R}_2 W_2 = -\widetilde{R}_1 W_1.$$

Take any unitarily invariant norm, use Lemma 2.1 and the triangular inequality to get

$$\left(gap(A_3, \Lambda_1) - \|\widetilde{R}_2\|_2 \frac{\|W_2\|}{\|W_3\|} \right) \|W_3\| \leq \|\widetilde{R}_1\| \|W_1\|_2,$$

where $gap(A_3, \Lambda_1)$ is the smallest distance between the sets of eigenvalues of A_3 and Λ_1. Therefore, provided that $gap(A_3, \Lambda_1) - \frac{\|\widetilde{R}_2\|_2^2}{|\lambda_\ell - \theta_{\ell+1}|} > 0$, using (10.14) we get

$$\|W_3\| \leq \frac{\|\widetilde{R}_1\|}{gap(A_3, \Lambda_1) - \frac{\|\widetilde{R}_2\|_2^2}{|\lambda_\ell - \theta_{\ell+1}|}} \|W_1\|_2 \,(\equiv \gamma_{31}\|W_1\|_2)\,, \tag{10.15}$$

which is approximately $\|\widetilde{R}_1\|_2 / gap(A_3, \Lambda_1)$ if $\|\widetilde{R}_2\|_2$ is small enough so that $\frac{\|\widetilde{R}_2\|_2^2}{|\lambda_\ell - \theta_{\ell+1}|} \ll gap(A_3, \Lambda_1)$.

For any unitarily invariant norm, we have

$$\left\| \begin{bmatrix} W_2 \\ W_3 \end{bmatrix} \right\| \leq \sqrt{\|W_2\|^2 + \|W_3\|^2} = \sqrt{1 + \gamma_{23}^2}\|W_3\| = \sqrt{1 + \frac{\|\widetilde{R}_2\|_2^2}{(\lambda_\ell - \theta_{\ell+1})^2}}\|W_3\|. \tag{10.16}$$

This is a *sharper version* of Theorem 4.3 in Knyazev [**90**], which states (using the notations here)

$$\|\sin\angle(\widehat{X}_1, X_1)\|^2 \leq \left[1 + \frac{\left\| \begin{bmatrix} \widetilde{R}_1 \\ \widetilde{R}_2 \end{bmatrix} \right\|^2}{(\lambda_\ell - \theta_{\ell+1})^2} \right] \|\sin\angle(Q_1, X_1)\|^2. \tag{10.17}$$

Noting that $\|\sin\angle(\widehat{X}_1, X_1)\| = \left\| \begin{bmatrix} W_2 \\ W_3 \end{bmatrix} \right\|$ and $\|\sin\angle(Q_1, X_1)\| = \|\sin\angle([\widehat{X}_1\ \widehat{X}_2], X_1)\| = \|W_3\|$, we see that (10.16) is equivalent to

$$\|\sin\angle(\widehat{X}_1, X_1)\|^2 \le \left[1 + \frac{\|\widetilde{R}_2\|_2^2}{(\lambda_\ell - \theta_{\ell+1})^2}\right] \|\sin\angle(Q_1, X_1)\|^2, \tag{10.18}$$

which is clearly a sharper bound than (10.17), although the improvement may be marginal because typically we have $\|\widetilde{R}_2\|_2 \gg \|\widetilde{R}_1\|_2$.

10.4.2. Direct bound for $\|\sin\angle(\widehat{X}_1, X_1)\|$. Another, perhaps more significant, improvement that we can make along this argument is that we obtain upper bounds for $\|\sin\angle(Q_1, X_1)\|$ that are much sharper than existing bounds suggest. This, combined with (10.18), yields direct and tight bounds for $\|\sin\angle(\widehat{X}_1, X_1)\|$.

We first show that we get a new bound on $\|\sin\angle(Q_1, X_1)\| = \|\sin\angle([\widehat{X}_1\ \widehat{X}_2], X_1)\|$ ($\equiv \|W_3\|$), which measures how much of X_1 is contained in the initial trial subspace Q_1. One way of bounding it is by means of the Davis-Kahan generalized $\sin\theta$ theorem [**29**, Thm. 6.1], which yields

$$\|\sin\angle(Q_1, X_1)\| \le \frac{\left\| \begin{bmatrix} \widetilde{R}_1 \\ \widetilde{R}_2 \end{bmatrix} \right\|}{gap(A_3, \Lambda_1)}. \tag{10.19}$$

Instead, here we use (10.15) which is

$$\begin{aligned} \|\sin\angle(Q_1, X_1)\| &\le \frac{\|\widetilde{R}_1\|}{gap(A_3, \Lambda_1) - \frac{\|\widetilde{R}_2\|_2^2}{|\lambda_\ell - \theta_{\ell+1}|}} \|W_1\|_2 \\ &\le \frac{\|\widetilde{R}_1\|}{gap(A_3, \Lambda_1) - \frac{\|\widetilde{R}_2\|_2^2}{|\lambda_\ell - \theta_{\ell+1}|}}. \end{aligned} \tag{10.20}$$

The bound is again approximately $\|\widetilde{R}_1\|_2 / gap(A_3, \Lambda_1)$ provided that $\|\widehat{R}_2\|_2$ is small enough. Since in practice we typically have $\|R_1\| \ll \|R_2\|$, this bound can be much smaller than the Davis-Kahan bound (10.19).

Finally, we plug (10.20) into (10.18) to obtain

$$\|\sin\angle(\widehat{X}_1, X_1)\| \le \sqrt{1 + \frac{\|\widetilde{R}_2\|_2^2}{(\lambda_\ell - \theta_{\ell+1})^2}} \frac{\|\widetilde{R}_1\|}{gap(A_3, \Lambda_1) - \frac{\|\widetilde{R}_2\|_2^2}{|\lambda_\ell - \theta_{\ell+1}|}}.$$

We regard this as a direct bound for $\|\sin\angle(\widehat{X}_1, X_1)\|$ because the bound can be computed from $\|\widetilde{R}_i\|$ and the eigenvalues and Ritz values, involving no angle between subspaces.

One-vector case. Let us briefly consider the one-vector case $\ell = 1$, in which we show that we can derive bounds in terms of the tangent instead of sine, because we have

$$\begin{aligned}\sin\theta = \left\| \begin{bmatrix} W_2 \\ W_3 \end{bmatrix} \right\|_2 &= \sqrt{\|W_2\|_2^2 + \|W_3\|_2^2} \\ &\le \gamma_{31}\sqrt{1+\gamma_{23}^2}\|W_1\| \\ &= \gamma_{31}\sqrt{1+\gamma_{23}^2}w_1 = \gamma_{31}\sqrt{1+\gamma_{23}^2}\cos\theta,\end{aligned}$$

and therefore

$$\tan\theta \le \gamma_{31}\sqrt{1+\gamma_{23}^2}w_1 = \gamma_{31}\sqrt{1+\gamma_{23}^2}.$$

This means $\tan\angle(x_1, \widehat{x}_1) \le \gamma_{31}\sqrt{1+\gamma_{23}^2}$, which is again approximately $\|\widetilde{R}_1\|_2/gap(A_3, \Lambda_1)$.

Recalling the $\tan\theta$ theorem with relaxed conditions discussed earlier in this chapter, we note that this argument compares an exact eigenvector $\Lambda_1 = \lambda_1$ with Ritz values $\lambda(A_3)$ which may lie below and above λ_1, so the situation is the same as in Theorem 10.1. Hence the natural question becomes, when $\ell > 1$, if the Ritz values $\mathrm{eig}(A_3)$ lie both above and below the exact eigenvalues (those close to $\mathrm{eig}(A_1), \mathrm{eig}(A_2)$), do all the results above hold in terms of $\tan\theta$? This is an open question as of writing.

10.4.3. When more than 3 blocks exist : $A_1, \ldots, A_{k+1}$. Let

$$(\widetilde{A} =)\ Q^*AQ = \begin{bmatrix} A_1 & & & \widetilde{R}_1^* \\ & \ddots & & \vdots \\ & & A_k & \widetilde{R}_k^* \\ \widetilde{R}_1 & \cdots & \widetilde{R}_k & A_{k+1} \end{bmatrix}, \tag{10.21}$$

where $A_1 = \mathrm{diag}(\theta_{1,1}, \ldots, \theta_{1,\ell})$ and $A_i = \theta_i$ for $i \ge 2$ are diagonal matrices of Ritz values. The known quantities are $A_1, \ldots, A_k$ and the 2-norms of each column of $\widetilde{R}_1, \ldots, \widetilde{R}_k$.

Let $W = \begin{bmatrix} W_1 \\ \vdots \\ W_{k+1} \end{bmatrix}$ be the orthognoal eigenvector matrix of $\widetilde{A}$ corresponding to the smallest ℓ eigenvalues Λ_1. Then the ith block of $\widetilde{A}W = W\Lambda_1$ gives

$$A_iW_i + \widetilde{R}_i^*W_{k+1} = W_i\Lambda_1,$$

which is equivalent to

$$A_iW_i - W_i\Lambda_1 = -\widetilde{R}_i^*W_{k+1}. \tag{10.22}$$

Therefore by Lemma 2.1 we have

$$\|W_i\| \le \frac{\|\widetilde{R}_i^*W_{k+1}\|}{|\lambda_\ell - \theta_i|} \equiv \frac{\|\widetilde{R}_i^*W_{k+1}\|}{d_i}.$$

Hence

$$\|W_i\| \le \frac{\|\widetilde{R}_i\|_2}{d_i}\|W_{k+1}\|. \tag{10.23}$$

We therefore have the upper bound

$$\|\sin\angle(\widehat{X}_1, X_1)\| = \left\| \begin{bmatrix} W_2 \\ \vdots \\ W_{k+1} \end{bmatrix} \right\| \leq \sqrt{\sum_{i=2}^{k+1} \|W_i\|^2} = \sqrt{1 + \sum_{i=2}^{k} \left(\frac{\|\widetilde{R}_i\|_2}{d_i} \right)^2} \|W_{k+1}\|$$

$$= \sqrt{1 + \sum_{i=2}^{k} \left(\frac{\|\widetilde{R}_i\|_2}{d_i} \right)^2} \|\sin\angle(Q_1, X_1)\|. \tag{10.24}$$

Now the last block of $\widetilde{A}W = W\Lambda_1$ gives

$$\widetilde{R}_1 W_1 + \widetilde{R}_2 W_2 + \ldots + \widetilde{R}_k W_k + A_{k+1} W_{k+1} = W_{k+1}\Lambda_1,$$

hence

$$A_{k+1}W_{k+1} - W_{k+1}\Lambda_1 + \sum_{i=2}^{k} \widetilde{R}_i W_i = -\widetilde{R}_1 W_1.$$

Taking norms, using Lemma 2.1 and the triangular inequality yields

$$gap(A_3, \Lambda_1)\|W_{k+1}\| - \|\sum_{i=2}^{k} \widetilde{R}_i W_i\| \leq \|\widetilde{R}_1\|\|W_1\|_2. \tag{10.25}$$

Now we want to bound the second term $\|\sum_{i=2}^{k} \widetilde{R}_i W_i\|$, in particular we want a bound like $\|\sum_{i=2}^{k} \widetilde{R}_i W_i\| \leq \gamma\|W_{k+1}\|$. One way to do this is to use (10.23) and get

$$\|\sum_{i=2}^{k} \widetilde{R}_i W_i\| \leq \sum_{i=2}^{k} \|\widetilde{R}_i W_i\| \leq \left(\sum_{i=2}^{k} \frac{\|\widetilde{R}_i\|_2^2}{d_i} \right) \|W_{k+1}\|. \tag{10.26}$$

We plug the inequality (10.26) into (10.25) to get a bound on $\|W_{k+1}\|$ as

$$\|\sin\angle(Q_1, X_1)\| = \|W_{k+1}\| \leq \frac{\|\widetilde{R}_1\|}{gap(A_3, \Lambda_1) - \left(\sum_{i=2}^{k} \frac{\|\widetilde{R}_i\|^2}{d_i} \right)} \|W_1\|_2. \tag{10.27}$$

The bounds (10.24) and (10.27) can be much tighter than those in the previous subsection (10.18) and (10.20) respectively, especially when $\|\widetilde{R}_i\|_2$ vary significantly for different i.

Finally, by combining (10.24) and (10.27) we get a direct bound for $\|\sin\angle(\widehat{X}_1, X_1)\|$:

$$\|\sin\angle(\widehat{X}_1, X_1)\| \leq \sqrt{1 + \sum_{i=2}^{k} \left(\frac{\|\widetilde{R}_i\|_2}{d_i} \right)^2} \frac{\|\widetilde{R}_1\|}{gap(A_3, \Lambda_1) - \left(\sum_{i=2}^{k} \frac{\|\widetilde{R}_i\|_2^2}{d_i} \right)}.$$

10.5. Singular vectors

In this section we present natural analogues of the above arguments to the projection-based method for computing the SVD. In doing so we rederive the well-known result by Wedin [**163**], and derive extensions of the theorems by Saad and Knyazev to the SVD.

Let A be an m-by-n matrix and $\widehat{U}_i, \widehat{V}_i$ be approximations to the matrices of singular vectors such that

$$(\widetilde{A} =)\ [\widehat{U}_1\ \widehat{U}_2]^* A [\widehat{V}_1\ \widehat{V}_2] = \begin{bmatrix} A_1 & \widetilde{R} \\ \widetilde{S} & A_2 \end{bmatrix}. \tag{10.28}$$

When $\widetilde{R}, \widetilde{S}$ are both small we can expect $\widehat{U}_i, \widehat{V}_i$ to be good approximations to the singular vectors. The goal here is to bound the angles $\angle(\widehat{U}_1, U_1)$ and $\angle(\widehat{V}_1, V_1)$, which measure the quality of the singular subspaces computed by projecting A onto the left-subspace $[\widehat{U}_1\ \widehat{U}_2]$ and right-subspace $[\widehat{V}_1\ \widehat{V}_2]$ (sometimes called the Petrov-Galerkin method [**13**]). Suppose the goal is to compute the largest singular values (care is needed when dealing with the subspace corresponding to the smallest singular values [**163**]), so the singular values of A_1 are larger than those of A_2.

Denoting by $U = [U_1\ U_2], V = [V_1\ V_2]$ the exact singular vectors of $\widetilde{A}$ corresponding to the singular value matrices Σ_1, Σ_2 where Σ_1 contains the largest singular values, the key is to note that the desired angles can be computed via the CS decomposition of U, V.

By the second block of $\begin{bmatrix} A_1 & \widetilde{R} \\ \widetilde{S} & A_2 \end{bmatrix} \begin{bmatrix} V_{11} \\ V_{21} \end{bmatrix} = \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} \Sigma_1$ we get

$$\widetilde{S} V_{11} + A_2 V_{21} = U_{21} \Sigma_1,$$

so

$$U_{21} \Sigma_1 - A_2 V_{21} = \widetilde{S} V_{11},$$

taking norms we get

$$\sigma_{\min}(\Sigma_1)\|U_{21}\| - \|A_2\|_2 \|V_{21}\| \leq \|\widetilde{S}\|, \tag{10.29}$$

Similarly by the second block (column-wise) of $\begin{bmatrix} U_{11}^* & U_{21}^* \end{bmatrix} \begin{bmatrix} A_1 & \widetilde{R} \\ \widetilde{S} & A_2 \end{bmatrix} = \Sigma_1 \begin{bmatrix} V_{11}\ V_{21} \end{bmatrix}$ we get

$$U_{11}^* \widetilde{R} + U_{21}^* A_2 = \Sigma_1 V_{21},$$

so

$$\Sigma_1 V_{21} - U_{21}^* A_2 = U_{11}^* \widetilde{R},$$

so taking norms we get

$$\sigma_{\min}(\Sigma_1)\|V_{21}\| - \|A_2\|_2 \|U_{21}\| \leq \|\widetilde{R}\|, \tag{10.30}$$

Let us eliminate the $\|V_{21}\|$ term from (10.29) and (10.30). We get

$$((\sigma_{\min}(\Sigma_1))^2 - \|A_2\|_2^2)\|U_{21}\| \leq \sigma_{\min}(\Sigma_1)\|\widetilde{R}\| + \|A_2\|_2 \|\widetilde{S}\|,$$

Hence assuming that $\sigma_{\min}(\Sigma_1) > \|A_2\|_2$ we get

$$\|U_{21}\| \leq \frac{\sigma_{\min}(\Sigma_1)\|\widetilde{R}\| + \|A_2\|_2 \|\widetilde{S}\|}{(\sigma_{\min}(\Sigma_1))^2 - \|A_2\|_2^2}.$$

Similarly we can get a bound for $\|V_{21}\|$:

$$\|V_{21}\| \leq \frac{\|A_2\|_2\|\widetilde{R}\| + \sigma_{\min}(\Sigma_1)\|\widetilde{S}\|}{(\sigma_{\min}(\Sigma_1))^2 - \|A_2\|_2^2}.$$

Hence we conclude that

$$\max\{\|V_{21}\|, \|U_{21}\|\} \leq \frac{\max\{\|\widetilde{R}\|, \|\widetilde{S}\|\}}{\sigma_{\min}(\Sigma_1) - \|A_2\|_2}. \tag{10.31}$$

Note that this is the same as the result called generalized $\sin\theta$ theorem in Wedin [**163**], hence above is its new and simpler derivation.

10.5.1. Three blocks case. Now we turn to the case where there are three blocks. Let

$$(\widetilde{A} =)\ [\widehat{U}_1\ \widehat{U}_2\ \widehat{U}_3]^* A[\widehat{V}_1\ \widehat{V}_2\ \widehat{V}_3] = \begin{bmatrix} A_1 & 0 & \widetilde{R}_1 \\ 0 & A_2 & \widetilde{R}_2 \\ \widetilde{S}_1 & \widetilde{S}_2 & A_3 \end{bmatrix}. \tag{10.32}$$

Let (Σ_1, U_1, V_1) be the singular triplets corresponding to the largest singular values of $\tilde{A}$ and let $V_1 = \begin{bmatrix} V_{11} \\ V_{21} \\ V_{31} \end{bmatrix}, U_1 = \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \end{bmatrix}$. Note that by the CS decomposition we have

$$\|\sin\angle(\widehat{U}_1, U_1)\| = \left\| \begin{bmatrix} U_{21} \\ U_{31} \end{bmatrix} \right\|, \quad \|\sin\angle(\widehat{V}_1, V_1)\| = \left\| \begin{bmatrix} V_{21} \\ V_{31} \end{bmatrix} \right\| \tag{10.33}$$

By the second block of $\begin{bmatrix} A_1 & 0 & \widetilde{R}_1 \\ 0 & A_2 & \widetilde{R}_2 \\ \widetilde{S}_1 & \widetilde{S}_2 & A_3 \end{bmatrix} \begin{bmatrix} V_{11} \\ V_{21} \\ V_{31} \end{bmatrix} = \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \end{bmatrix} \Sigma_1$ we get

$$A_2 V_{21} + \widetilde{R}_2 V_{31} = U_{21}\Sigma_1, \tag{10.34}$$

and also by the second block of $\begin{bmatrix} U_{11}^* & U_{21}^* & U_{31}^* \end{bmatrix} \begin{bmatrix} A_1 & 0 & \widetilde{R}_1 \\ 0 & A_2 & \widetilde{R}_2 \\ \widetilde{S}_1 & \widetilde{S}_2 & A_3 \end{bmatrix} = \Sigma_1 \begin{bmatrix} V_{11}^* & V_{21}^* & V_{31}^* \end{bmatrix}$ we get

$$U_{21}^* A_2 + U_{31}^* \widetilde{S}_2 = \Sigma_1 V_{21}^*. \tag{10.35}$$

Taking norms and using the triangular inequality in (10.34) and (10.35) we get

$$\|U_{21}\|\sigma_{\min}(\Sigma_1) - \|V_{21}\|\|A_2\|_2 \leq \|\widetilde{R}_2 V_{31}\|,$$
$$\|V_{21}\|\sigma_{\min}(\Sigma_1) - \|U_{21}\|\|A_2\|_2 \leq \|U_{31}^* \widetilde{S}_2\|.$$

We can eliminate the $\|V_{21}\|$ term, and assuming that $\sigma_{\min}(\Sigma_1) > \|A_2\|_2$ we get

$$\|U_{21}\| \leq \frac{\sigma_{\min}(\Sigma_1)\|\widetilde{R}_2 V_{31}\| + \|A_2\|_2\|U_{31}^* \widetilde{S}_2\|}{(\sigma_{\min}(\Sigma_1))^2 - \|A_2\|_2^2}.$$

We can similarly obtain

$$\|V_{21}\| \leq \frac{\sigma_{\min}(\Sigma_1)\|U_{31}^*\widetilde{S}_2\| + \|A_2\|_2\|\widetilde{R}_2V_{31}\|}{(\sigma_{\min}(\Sigma_1))^2 - \|A_2\|_2^2}.$$

Hence we have

(10.36)

$$\max\{\|U_{21}\|, \|V_{21}\|\} \leq \frac{\max\{\|U_{31}^*\widetilde{S}_2\|, \|\widetilde{R}_2V_{31}\|\}}{\sigma_{\min}(\Sigma_1) - \|A_2\|_2}$$

(10.37)

$$\leq \frac{\max\{\|U_{31}^*\|, \|V_{31}\|\}\max\{\|\widetilde{R}_2\|, \|\widetilde{S}_2\|\}}{\sigma_{\min}(\Sigma_1) - \|A_2\|_2} \equiv \gamma_{23}\max\{\|U_{31}^*\|, \|V_{31}\|\}.$$

Note that this can be regarded as a generalization of Knyazev's theorem [**90**, Thm. 3.2] to the SVD.

Recalling (10.33) and using the facts

$$\|\sin\angle([\widehat{U}_1\ \widehat{U}_2], U_1)\| = \|U_{31}\|, \quad \|\sin\angle([\widehat{V}_1\ \widehat{V}_2], V_1)\| = \|V_{31}\|$$

we obtain

$$\max\{\|\sin\angle(\widehat{U}_1, U_1)\|, \|\sin\angle(\widehat{V}_1, V_1)\|\}$$

(10.38)

$$\leq \sqrt{1+\gamma_{23}^2}\max\{\|\sin\angle([\widehat{U}_1\ \widehat{U}_2], U_1)\|, \|\sin\angle([\widehat{V}_1\ \widehat{V}_2], V_1)\|\},$$

which claims that the computed singular vectors obtained by the projection method is optimal up to a factor $\sqrt{1+\gamma_{23}^2}$. (10.38) can be seen as a direct generalization of (10.18) to the SVD.

We next show that we can get direct bounds for $\max\{\|\sin\angle(\widehat{U}_1, U_1)\|, \|\sin\angle(\widehat{V}_1, V_1)\|\}$ with some more work. Now by the third block of $\begin{bmatrix} A_1 & 0 & \widetilde{R}_1 \\ 0 & A_2 & \widetilde{R}_2 \\ \widetilde{S}_1 & \widetilde{S}_2 & A_3 \end{bmatrix}\begin{bmatrix} V_{11} \\ V_{21} \\ V_{31} \end{bmatrix} = \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \end{bmatrix}\Sigma_1$ we get

$$\widetilde{S}_1V_{11} + \widetilde{S}_2V_{21} + A_3V_{31} = U_{31}\Sigma_1,$$

and also by the third block of $\begin{bmatrix} U_{11}^* & U_{21}^* & U_{31}^* \end{bmatrix}\begin{bmatrix} A_1 & 0 & \widetilde{R}_1 \\ 0 & A_2 & \widetilde{R}_2 \\ \widetilde{S}_1 & \widetilde{S}_2 & A_3 \end{bmatrix} = \Sigma_1\begin{bmatrix} V_{11}^* & V_{21}^* & V_{31}^* \end{bmatrix}$ we get

$$U_{11}^*\widetilde{R}_1 + U_{21}^*\widetilde{R}_2 + U_{31}^*A_3 = \Sigma_1V_{31}^*.$$

Hence

$$\|U_{31}\|\sigma_{\min}(\Sigma_1) - \|A_3\|_2\|V_{31}\| \leq \|\widetilde{S}_1V_{11} + \widetilde{S}_2V_{21}\|,$$

$$\sigma_{\min}(\Sigma_1)\|V_{31}\| - \|A_3\|_2\|U_{31}\| \leq \|U_{11}^*\widetilde{R}_1 + U_{21}^*\widetilde{R}_2\|.$$

We eliminate $\|V_{31}\|$ from the two inequalities to get (assuming that $\sigma_{\min}(\Sigma_1) > \|A_3\|_2$)

$$((\sigma_{\min}(\Sigma_1))^2 - \|A_3\|_2^2)\|U_{31}\| \leq \sigma_{\min}(\Sigma_1)\|\widetilde{S}_1V_{11} + \widetilde{S}_2V_{21}\| + \|A_3\|_2\|U_{11}^*\widetilde{R}_1 + U_{21}^*\widetilde{R}_2\|.$$

Similarly we eliminate $\|U_{31}\|$ to get

$$((\sigma_{\min}(\Sigma_1))^2 - \|A_3\|_2^2)\|V_{31}\| \le \sigma_{\min}(\Sigma_1)\|U_{11}^*\widetilde{R}_1 + U_{21}^*\widetilde{R}_2\| + \|A_3\|_2\|\widetilde{S}_1V_{11} + \widetilde{S}_2V_{21}\|.$$

Combining these we get

$$\begin{aligned}\max\{\|U_{31}\|, \|V_{31}\|\} &\le \frac{(\sigma_{\min}(\Sigma_1) + \|A_3\|_2)\max\{\|U_{11}^*\widetilde{R}_1 + U_{21}^*\widetilde{R}_2\|, \|\widetilde{S}_1V_{11} + \widetilde{S}_2V_{21}\|\}}{(\sigma_{\min}(\Sigma_1))^2 - \|A_3\|_2^2} \\ &= \frac{\max\{\|U_{11}^*\widetilde{R}_1 + U_{21}^*\widetilde{R}_2\|, \|\widetilde{S}_1V_{11} + \widetilde{S}_2V_{21}\|\}}{\sigma_{\min}(\Sigma_1) - \|A_3\|_2}.\end{aligned}$$

The last term can be bounded above by

$$\frac{\max\{\|U_{11}\|_2, \|V_{11}\|_2\}\max\{\|\widetilde{R}_1\|, \|\widetilde{S}_1\|\} + \gamma_{23}\max\{\|U_{31}\|, \|V_{31}\|\}\max\{\|\widetilde{R}_2\|, \|\widetilde{S}_2\|\}}{\sigma_{\min}(\Sigma_1) - \|A_3\|_2},$$

so assuming $\gamma_{23}\max\{\|\widetilde{R}_2\|, \|\widetilde{S}_2\|\}) < \sigma_{\min}(\Sigma_1) - \|A_3\|_2$, we have

$$\max\{\|U_{31}\|, \|V_{31}\|\} \le \frac{\max\{\|\widetilde{R}_1\|, \|\widetilde{S}_1\|\}}{\sigma_{\min}(\Sigma_1) - \|A_3\|_2 - \gamma_{23}\max\{\|\widetilde{R}_2\|, \|\widetilde{S}_2\|\})}\max\{\|U_{11}\|_2, \|V_{11}\|_2\}.$$

Since $\|U_{11}\|_2, \|V_{11}\|_2 \le 1$, recalling (10.33) we conclude that

(10.39)

$$\max\{\|\sin\angle(\widehat{U}_1, U_1)\|, \|\sin\angle(\widehat{V}_1, V_1)\|\} \le \frac{\max\{\|\widetilde{R}_1\|, \|\widetilde{S}_1\|\}}{\sigma_{\min}(\Sigma_1) - \|A_3\|_2 - \gamma_{23}\max\{\|\widetilde{R}_2\|, \|\widetilde{S}_2\|\})}.$$

Note that when the residuals $\|R_i\|, \|S_i\|$ are small enough the bound is approximately $\frac{\max\{\|\widetilde{R}_1\|, \|\widetilde{S}_1\|\}}{\sigma_{\min}(\Sigma_1) - \|A_3\|_2}$. (10.39) can be regarded an extension of the theorems by Saad and Knyazev (on the quality of a Galerkin approximation) to the SVD case, which is new to the author's knowledge.

10.6. The $\cos\theta$ theorem

In this section we first derive what might be called the $\cos\theta$ theorem, which measures the distance (instead of nearness as in the $\sin\theta$ or $\tan\theta$ theorems) between two susbpaces. The essential message is simple: two subspaces X, Y are automatically nearly orthogonal if they are approximate invariant subspaces of disjoint eigenvalues of a Hermitian matrix.

10.6.1. Two vectors. Suppose that x, y are approximate eigenvectors of different eigenvalues so that

$$\begin{aligned}Ax - \lambda_x x &= r_x, \\ Ay - \lambda_y y &= r_y.\end{aligned}$$

Then, left-multiplying the first equality by y^* and the second by x^* yields

$$\begin{aligned}y^*Ax &= \lambda_x y^*x + y^*r_x, \\ x^*Ay &= \lambda_y x^*y + x^*r_y.\end{aligned}$$

Subtracting the second from the first equation we get

$$(\lambda_x - \lambda_y)y^*x = x^*r_y - y^*r_x, \tag{10.40}$$

hence noting that $|y^*x| = \cos\angle(x,y)$, we get the $\cos\theta$ theorem for the case of two vectors:

$$\cos\angle(x,y) = |y^*x| \leq \frac{\|x^*r_y\|_2 + \|y^*r_x\|_2}{|\lambda_x - \lambda_y|} \leq \frac{\|r_y\|_2 + \|r_x\|_2}{|\lambda_x - \lambda_y|}. \tag{10.41}$$

Note that in the above argument (λ_x, x) and (λ_y, y) need not be Ritz pairs.

We next suppose that (λ_x, x) and (λ_y, y) are Ritz pairs, that is, $x^*r_x = y^*r_y = 0$. Then by (10.40) we have

$$(\lambda_x - \lambda_y)y^*x = x^*(I - yy^*)r_y - y^*(I - xx^*)r_x,$$

and noting that $\|x^*(I - yy^*)\| = \|\sin\angle(x,y)\| = \|y^*(I - xx^*)\|$ we get

$$\cos\theta = |y^*x| \leq \frac{\|r_y\|_2 + \|r_x\|_2}{|\lambda_x - \lambda_y|}\sin\theta.,$$

hence

$$\frac{1}{\tan\theta} \leq \frac{\|r_y\|_2 + \|r_x\|_2}{|\lambda_x - \lambda_y|}.$$

To summarize the above, the $1/\tan\theta$ theorem holds when the approximate eigenpairs are Ritz pairs, while the $\cos\theta$ theorem holds without this condition. This is much the same relation between the Davis-Kahan $\sin\theta$ and $\tan\theta$ theorem, the latter of which requires the approximate eigenpairs to be Ritz pairs.

10.6.2. A vector and a subspace. Suppose that x is an approximate eigenvector and Y is an approximate eigenspace of dimension k_y, and

$$Ax - \lambda_x x = r_x,$$
$$AY - Y\Lambda_Y = R_Y,$$

where Λ_Y is a diagonal matrix of approximate eigenvalues corresponding to Y. Then left-multiplying the first equality by Y^* and the second by x^* yields

$$Y^*Ax = Y^*x\lambda_x + Y^*r_x,$$
$$x^*AY = x^*Y\Lambda_Y + x^*R_Y.$$

Subtracting the Hermitian transpose of the second equation from the first equation we get

$$(\lambda_x I - \Lambda_Y)Y^*x = Y^*r_x - R_Y^*x,$$

hence

$$|Y^*x| \leq \frac{\|Y^*r_x - R_Y^*x\|_2}{gap_{x,Y}} \leq \frac{\|R_Y\|_2 + \|r_x\|_2}{gap_{x,Y}}, \tag{10.42}$$

where $gap_{x,Y} = \min(\mathrm{diag}(\Lambda_Y) - \lambda_x)$.

When we are dealing with Ritz pairs we can get bounds in terms of $1/\tan\theta$, similarly to the case of two vectors. Suppose (Λ_x, x) and (Λ_y, Y) are Ritz pairs. Since residuals are orthogonal to the Ritz vector we get

$$(\lambda_x I - \Lambda_Y)Y^*x = Y^*(I - xx^*)r_x - R_Y^*(I - YY^*)x.$$

Therefore, using $\|Y^*(I-xx^*)\| = \|(I-YY^*)x\| = \|\sin\angle(x,Y)\|$ we get

$$\begin{aligned}\|\cos\angle(x,Y)\| &= \|Y^*x\| \\ &\leq \frac{\|Y^*(I-xx^*)r_x - R_Y^*(I-YY^*)x\|}{gap_{x,Y}} \\ &\leq \frac{\|r_x\|_2 + \|R_Y\|_2}{gap_{x,Y}}\|\sin\angle(x,Y)\|.\end{aligned}$$

since there is only one angle we conclude that

$$\frac{1}{\tan\angle(x,Y)} \leq \frac{\|R_x\|_2 + \|R_Y\|_2}{gap_{x,Y}}.$$

10.6.3. Two subspaces. Suppose that X, Y are approximate invariant subspaces of dimension k_x, k_y respectively, and

$$\begin{aligned}AX - X\Lambda_X &= R_X, \\ AY - Y\Lambda_Y &= R_Y,\end{aligned}$$

where Λ_X, Λ_Y are diagonal matrices. Then left-multiplying the first equality by Y^* and the second by X^*, and taking its complex transpose yields

$$\begin{aligned}Y^*AX &= Y^*X\Lambda_X + Y^*R_X, \\ Y^*AX &= \Lambda_Y Y^*X + R_Y^*X.\end{aligned}$$

Subtracting the second equation from the first we get

$$Y^*X\Lambda_X - \Lambda_Y Y^*X = Y^*R_X - R_Y^*X.$$

Then using Lemma 2.1 we conclude that

$$\|Y^*X\| \leq \frac{\|Y^*R_X - R_Y^*X\|}{\delta_{X,Y}} \leq \frac{\|R_X\| + \|R_Y\|}{\delta_{X,Y}}, \tag{10.43}$$

where $gap_{X,Y} = \max\{\min(\mathrm{diag}(\Lambda_X)) - \max(\mathrm{diag}(\Lambda_Y)), \min(\mathrm{diag}(\Lambda_Y)) - \max(\mathrm{diag}(\Lambda_X))\}$. Since $\|Y^*X\| = \|\cos\angle(X,Y)\|$ we have the following result.

THEOREM 10.3. *Let A be an N-by-N matrix and let $X \in \mathbb{C}^{N\times k_x}$, $Y_1 \in \mathbb{C}^{N\times k_y}$ be approximate invariant subspaces such that*

$$\begin{aligned}AX - X\Lambda_X &= R_X, \\ AY - Y\Lambda_Y &= R_Y,\end{aligned}$$

where Λ_X, Λ_Y are diagonal matrices. Then,

$$\|\cos\angle(X,Y)\| = \|Y^*X\| \leq \frac{\|Y^*R_X - R_Y^*X\|}{\delta_{X,Y}} \leq \frac{\|R_X\| + \|R_Y\|}{\delta_{X,Y}}, \tag{10.44}$$

where $\cos\angle(X,Y)$ is a matrix whose singular values are the cosines of the $k = \min\{k_1, k_2\}$ canonical angles between X and Y.

A natural question is to ask whether the same type of a bound holds for the $\|1/\tan\angle(X,Y)\|$, just like in the case of a vector and a subspace? The answer seems nontrivial and is an open problem as of writing.

Future study: efficient execution of Rayleigh-Ritz. We conclude this chapter by suggesting a direction for possible future work. As we have discussed previously, solving a large sparse symmetric eigenvalue problem of matrix size N typically involves forming an approximate subspace and performing the Rayleigh-Ritz process. In the common case where the matrix is highly sparse so that a matrix-vector multiplication can be done in $O(N)$ flops, the computational bottleneck of the eigensolver lies in the orthogonalization step in the Rayleigh-Ritz process, which necessarily requires $O(Nm^2)$ flops where m is the number of eigenpairs desired. This issue is raised clearly in [**135**].

Now, in many applications (for example molecular dynamics simulations) it is required to solve a sequence of eigenproblems, and solving each of which accurately may not be necessary. Instead a solution with accuracy 10^{-4} (say) is sufficient. Recall that the Rayleigh-Ritz process yields a solution that is orthogonal to working accuracy, that is, gives a solution that is numerically on the Grassman manifold [**40**]. On the other hand, the $\cos\theta$ theorem guarantees that an approximate eigenvector is nearly orthogonal to another approximate eigenvector if the corresponding approximate eigenvalues are sufficiently distinct. Our idea is the following: if the required solution is of accuracy $O(10^{-4})$, why not relax the requirement that the solution is nearly orthogonal to orthogonal to within $O(10^{-4})$, thereby reducing the cost involved in the Rayleigh-Ritz process?

Let us give more details. In many eigensolvers (such as the Jacobi-Davidson and LOBPCG algorithms) the Rayleigh-Ritz process computes the eigendecomposition of (Z^*AZ, Z^*BZ) where $Z = [X\ P]$. Here X is the current approximate solution and P is the search space, formed as preconditioned residuals. In view of the $\cos\theta$ theorem, our idea is that to obtain the next approximate solution x_i we need only use the current x_i, some neighboring colums of X, p_i and its neighbors. Other vectors are nearly orthogonal to x_i (and the exact solution that x_i converges to) so we can safely ignore them as far as updating x_i is concerned. Hence our task is basically to compute many smaller pairs $(Z_i^*AZ_i, Z_i^*BZ_i)$ for $i = 1, \ldots,$ where $Z_i = [X_i\ P_i]$ where $X = [X_1\ X_2 \ldots X_s], P = [P_1\ P_2 \ldots P_s]$ (in actual calculations we may need to let Z_i have some overlaps, that is, X_1 and X_2 share some columns).

The process is somewhat similar to the algebraic substructuring method discussed in [**167**]. However here we do not take into accout the effect of the lower-right corner matrix. We are basically approximating the next solution $\widehat{X}$ by first observing that the important components of the ith column of $\widehat{X}$ come from the ith and nearby columns of X and the ith column of P, then performing the (perhaps Harmonic or refined) Rayleigh-Ritz process with the small projected matrix.

This has the potential of significantly reducing the arithmetic cost involved in the Rayleigh-Ritz process. First recall that the ordinary Rayleigh-Ritz process requires $O(Nm^2)$ flops. If we have s partitions, the cost for each harmonic Rayleigh-Ritz process becomes $O(N(m/s)^2)$. Since we compute s of them, the total cost becomes $O(N(m/s)^2 s) = O(Nm^2/s)$, so the flop count is s times smaller. Moreover, the s small Rayleigh-Ritz can be processed independently

on a parallel machine, which is a desirable feature. We intend to further investigate this in a future work.

CHAPTER 11

Gerschgorin theory

In this final chapter of the dissertation we develop variants of Gerschgorin's theorem. The goal is to derive eigenvalue inclusion regions that are ideally tight but computationally inexpensive. We first derive a new Gerschgorin-type region that is tighter than Gerschgorin's theorem. We then present an eigenvalue inclusion set applicable to generalized eigenvalue problems. We argue that the proposed set is computationally more attractive. The underlying idea for both developments is to obtain bound on the eigenvector components, from which estimates of eigenvalues can be obtained.

Introduction. For any square matrix A, Gerschgorin's thoerem gives a set $\Gamma(A)$ that includes all the eigenvalues, as reviewed in Section 2.11. For convenience we restate the set as

$$\{\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)\} \subseteq \Gamma(A) \equiv \bigcup_{i=1}^{n} \Gamma_i(A), \tag{11.1}$$

where $\Gamma_i(A) \equiv \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \in N \setminus \{i\}} |a_{ij}| \right\}$. Since its discovery, a number of generalizations have been reported and discussed [**17, 18, 28, 64, 99**]. A typical extension of Gerschgorin's theorem defines a set (not necessarily a union of circles) in the complex plane that contains all the eigenvalues of a matrix, and which is ideally included in the Gerschgorin set, providing more accurate information on the eigenvalue distribution. The book by Varga [**160**] includes a good coverage of such generalizations.

In the first part of this chapter (Sections 11.1–11.2) we propose a new idea of generalizing Gerschgorin's theorem, after which we present an eigenvalue inclusion set that is always included in two known sets.

The second (main) part deals with generalized eigenproblems $Ax = \lambda Bx$, where A and B are non-Hermitian. Given the simplicity and wide range of applications of Gerschgorin's theorem, it should be useful to have available a similar simple theory to estimate the eigenvalues for generalized eigenproblems as well.

In fact, Stewart and Sun [**140, 146**] provide an eigenvalue inclusion set applicable to generalized eigenvalue problems. The set is the union of n regions defined by

$$G_i(A, B) \equiv \{ z \in \mathbb{C} : \chi(z, a_{i,i}/b_{i,i}) \leq \varrho_i \}, \tag{11.2}$$

where

$$\varrho_i = \sqrt{\frac{\left(\sum_{j\neq i}|a_{i,j}|\right)^2 + \left(\sum_{j\neq i}|b_{i,j}|\right)^2}{|a_{i,i}|^2 + |b_{i,i}|^2}}. \tag{11.3}$$

All the eigenvalues of the pair (A, B) lie in the union of $G_i(A, B)$, i.e., if λ is an eigenvalue, then

$$\lambda \in G(A, B) \equiv \bigcup_{i=1}^{n} G_i(A, B).$$

Note that λ can be infinite. We briefly review the definition of eigenvalues of a pair at the beginning of Section 11.3.

The region (11.2) is defined in terms of the chordal metric χ, defined by [**56**, Ch.7.7]

$$\chi(x, y) = \frac{|x - y|}{\sqrt{1 + |x|^2}\sqrt{1 + |y|^2}}.$$

As we discussed in Chapter 8, despite the benefit of a unifying treatment of finite and infinite eigenvalues, using the chordal metric makes the application of the theory less intuitive and usually more complicated. In particular, interpreting the set G in the Euclidean metric is a difficult task, as opposed to the the Gerschgorin set for standard eigenvalue problems, which is defined as a union of n disks. Another caveat of using G is that it is not clear whether the region G will give a nontrivial estimate of the eigenvalues. Specifically, since any two points in the complex plane have distance smaller than 1 in the chordal metric, if there exists i such that $\varrho_i \geq 1$, then G is the whole complex plane, providing no information. In view of (11.3), it follows that G is useful only when both A and B have small off-diagonal elements.

Another Gerschgorin-type eigenvalue localization theory applicable to generalized eigenvalue problems appear in a recent chapter [**93**] by Kostic et al. Their inclusion set is defined by

$$K_i(A, B) \equiv \left\{ z \in \mathbb{C} : |b_{i,i}z - a_{i,i}| \leq \sum_{j\neq i} |b_{i,j}z - a_{i,j}| \right\}, \tag{11.4}$$

and all the eigenvalues of the pair (A, B) exist in the union $K(A, B) \equiv \bigcup_{i=1}^{n} K_i(A, B)$. This set is defined in the Euclidean metric, and (11.4) shows that $K(A, B)$ is a compact set in the complex plane $\mathbb{C}$ if and only if B is strictly diagonally dominant. However, the set (11.4) is in general a complicated region, which makes its practical application difficult.

The goal of the second part of this chapter is to present a different generalization of Gerschgorin's theorem applicable to generalized eigenvalue problems, which solves the issues mentioned above. In brief, our eigenvalue inclusion sets have the following properties:

- They involve only circles in the Euclidean complex plane, using the same information as (11.2) does. Therefore it is simple to compute and visualize.
- They are defined in the Euclidean metric, but still deal with finite and infinite eigenvalues uniformly.
- One variant $\Gamma^S(A, B)$ is a union of n disks when B is strictly diagonally dominant.

- Comparison with $G(A, B)$: Our results are defined in the Euclidean metric. Tightness is incomparable, but our results are tighter when B is close to a diagonal matrix.
- Comparison with $K(A, B)$: Our results are defined by circles and are much simpler. $K(A, B)$ is always tighter, but our results approach $K(A, B)$ when B is close to a diagonal matrix.

In summary, our results provide a method for estimating eigenvalues of (A, B) in a much cheaper way than the two known results do.

The idea penetrating this entire chapter is to bound eigenvector components in absolute value, from which eigenvalue bounds can be obtained. This is somewhat in the same vein as in Chapter 7 where we gave refined perturbation bounds for Hermitian block tridiagonal matrices, but here we do not assume any matrix structure.

The structure of this chapter is as follows. In Section 11.1 we state the idea of generalizing Gerschgorin's theorem. Based on the idea, a new eigenvalue inclusion set is presented. We show that the new set is always tighter than the Gerschgorin set and Brauer's ovals of Cassini. In Section 11.2 we show numerical examples and specifications to the symmetric tridiagonal case. Section 11.3 embarks on generalized eigenproblems, in which the basic framework for bounding the eigenvalue location is discussed. We then present a simple bound and derive two Gerschgorin-type bounds. In addition, we show that our results can localize a specific number of eigenvalues, a well-known property of G and the Gerschgorin set for standard eigenvalue problems. Section 11.4 shows examples to illustrate the sets. Section 11.5 presents applications of our results, where we develop forward error analyses for the computed eigenvalues of a non-Hermitian generalized eigenvalue problem.

11.1. A new Gerschgorin-type eigenvalue inclusion set

In this section we present our idea of generalizing Gerschgorin's theorem, followed by a definition of a new eigenvalue inclusion set.

11.1.1. Idea. In order to derive the basic idea, we start with reviewing the proof of Gerschgorin's theorem.

Proof of Theorem 2.7. [**160**] Let λ_k be an eigenvalue of A, and let $x(\neq 0) \in \mathbb{C}^n$ be its associated eigenvector, so $Ax = \lambda_k x$. Let x_{k_ℓ} denote the element of x with the ℓth largest absolute value. In other words, $k_1, k_2, \cdots, k_n$ are distinct integers in $N = \{1, 2, \cdots, n\}$ that satisfy

$$|x_{k_1}| \geq |x_{k_2}| \geq \cdots \geq |x_{k_n}|. \tag{11.5}$$

Note that $|x_{k_1}| > 0$ because $x \neq 0$. Then the row of $Ax = \lambda_k x$ corresponding to x_{k_1} yields

$$(\lambda_k - a_{k_1 k_1})x_{k_1} = \sum_{i \in N \setminus \{1\}} a_{k_1 k_i} x_{k_i}\,.$$

Since $|x_{k_1}| > 0$, we obtain

$$|\lambda_k - a_{k_1k_1}| \leq \sum_{i \in N \setminus \{1\}} |a_{k_1k_i}| \frac{|x_{k_i}|}{|x_{k_1}|}. \tag{11.6}$$

From (11.5) we have $\dfrac{|x_{k_i}|}{|x_{k_1}|} \leq 1$ for all $i \in N \setminus \{1\}$. Therefore,

$$|\lambda_k - a_{k_1k_1}| \leq \sum_{i \in N \setminus \{1\}} |a_{k_1k_i}| = R_{k_1},$$

which implies $\lambda_k \in \Gamma_{k_1}(A) \subset \Gamma(A)$. □

In the proof, the terms $|x_{k_i}|/|x_{k_1}|$ $(i \in N \setminus \{1\})$ on the right hand side of (11.6) are bounded from above uniformly by 1, which would be overestimates. Hence if we can obtain tighter upper bounds for these terms, we must be able to define a smaller eigenvalue inclusion set. The following lemma should provide such upper bounds.

LEMMA 11.1. *For $i \in N$ such that $\lambda_k \neq a_{k_ik_i}$,*

$$\frac{|x_{k_i}|}{|x_{k_1}|} \leq \frac{R_{k_i}}{|\lambda_k - a_{k_ik_i}|}.$$

PROOF. The k_ith row in $Ax = \lambda_k x$ yields

$$|\lambda_k - a_{k_ik_i}||x_{k_i}| \leq \sum_{j \in N \setminus \{i\}} |a_{k_ik_j}||x_{k_j}|.$$

From (11.5), we have

$$|\lambda_k - a_{k_ik_i}||x_{k_i}| \leq \sum_{j \in N \setminus \{i\}} |a_{k_ik_j}||x_{k_1}| = R_{k_i}|x_{k_1}|. \tag{11.7}$$

Then, noting that $|\lambda_k - a_{k_ik_i}||x_{k_1}| > 0$ by assumption, we obtain

$$\frac{|x_{k_i}|}{|x_{k_1}|} \leq \frac{R_{k_i}}{|\lambda_k - a_{k_ik_i}|}. \tag{11.8}$$

□

11.1.2. A new eigenvalue inclusion set. Now we can define an eigenvalue inclusion set by substituting the new upper bounds for $|x_{k_i}|/|x_{k_1}|$ into (11.6).

THEOREM 11.1. *For any $n \times n$ complex matrix A, define*

$$E_i(A) \equiv \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \in N \setminus \{i\}} |a_{ij}| \frac{R_j}{|z - a_{jj}|} \text{ and } z \notin O(A) \right\} \cup O(A)$$
$$(i = 1, 2, \cdots, n), \tag{11.9}$$

where $O(A) := \{a_{11}, a_{22}, \cdots, a_{nn}\}$. *Then all the eigenvalues of* A, $\lambda_p(A)$ $(p = 1, 2, \cdots, n)$, *lie within the union of all the sets* $E_i(A)$, *i.e.,*

$$\{\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)\} \subseteq E(A) \equiv \bigcup_{i \in N} E_i(A). \tag{11.10}$$

PROOF. First, if $\lambda_k \in O(A)$, then λ_k is automatically included in $E(A)$. Therefore it suffices to consider the case where $\lambda_k \notin O(A)$.

If $\lambda_k \notin O(A)$, then by Lemma 11.1 we have $\dfrac{|x_{k_i}|}{|x_{k_1}|} \leq \dfrac{R_{k_i}}{|\lambda_k - a_{k_i k_i}|}$ for all $i \in N \setminus \{1\}$. Then from (11.6) we get

$$|\lambda_k - a_{k_1 k_1}| \leq \sum_{i \in N \setminus \{1\}} |a_{k_1 k_i}| \frac{R_{k_i}}{|\lambda_k - a_{k_i k_i}|},$$

which implies $\lambda_k \in E_{k_1}(A) \subset E(A)$. □

We note that this is not the only set we can define using the same idea. Specifically, we can substitute either $R_{k_i}/|\lambda_k - a_{k_i k_i}|$ or 1 into $|x_{k_i}|/|x_{k_1}|$, which yields another eigenvalue inclusion set, and is possibly a set included in $E(A)$. In this manner, the optimal set using the idea can be defined by

$$E_i^{\infty}(A) \equiv \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \in N \setminus \{i\}} |a_{ij}| \cdot \min\left\{ \frac{R_j}{|z - a_{jj}|}, 1 \right\} \text{ and } z \notin O(A) \right\} \cup O(A)$$
$$(i = 1, 2, \cdots, n), \tag{11.11}$$

and it can be shown that

$$\{\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)\} \subseteq E^{\infty}(A) \equiv \bigcup_{i \in N} E_i^{\infty}(A). \tag{11.12}$$

11.1.3. Comparison with known results. Here we compare the new set E in Theorem 11.1 with other well-known eigenvalue inclusion sets to examine its tightness and computational efficiency. We first show a comparison with the original Gerschgorin set.

11.1.3.1. *Comparison with Gerschgorin's theorem.* It can be shown that the set E in Theorem 11.1 is always included in the Gerschgorin set.

THEOREM 11.2. *For any* $n \times n$ *complex matrix* A,

$$E(A) \subseteq \Gamma(A)\,.$$

PROOF. We prove that $z \notin \Gamma(A) \Rightarrow z \notin E(A)$.

If $z \notin \Gamma(A)$, then $|z - a_{ii}| > R_i \left(\Leftrightarrow \dfrac{R_i}{|z - a_{ii}|} < 1 \right)$ for all $i \in N$. Hence

$$|z - a_{ii}| > R_i = \sum_{j \neq i} |a_{ij}| \geq \sum_{j \neq i} |a_{ij}| \frac{R_j}{|z - a_{jj}|} \quad \text{for all } i,$$

which implies $z \notin E(A)$. □

11.1.3.2. *Comparison with Brauer's ovals of Cassini.* Brauer's ovals of Cassini is a well-known Gerschgorin-type eigenvalue inclusion set [**17, 158**].

THEOREM 11.3 (Brauer's ovals of Cassini). *For any $n \times n$ $(n \geq 2)$ complex matrix A , define $n(n-1)/2$ ovals of Cassini by*

$$C_{i,j}(A) := \{z \in \mathbb{C} : |z - a_{ii}||z - a_{jj}| \leq R_i \cdot R_j\} \quad (i, j \in N, i < j).$$

Then all the eigenvalues of A, $\lambda_p(A)\,(p = 1, 2, \cdots, n)$, lie within the union of all of the ovals $C_{i,j}$, i.e.,

$$\{\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)\} \subseteq C(A) \equiv \bigcup_{\substack{i,j \in N \\ i<j}} C_{i,j}(A). \tag{11.13}$$

$C(A)$ is called the Brauer set of A. Its important property is that it is always included in the Gerschgorin set.

THEOREM 11.4. [**17, 158**] *For any $n \times n$ $(n \geq 2)$ complex matrix A,*

$$C(A) \subseteq \Gamma(A)\,.$$

Now we are interested in comparing $C(A)$ with $E(A)$. We prove that the latter is always a subset of the former.

THEOREM 11.5. *For any $n \times n$ $(n \geq 2)$ complex matrix A,*

$$E(A) \subseteq C(A)\,.$$

PROOF. We prove that $z \in E(A) \Rightarrow z \in C(A)$.

Suppose $z \in E_i(A)$. First, if $z \in O(A)$, then obviously $z \in C(A)$. Therefore we need only to consider the case where $z \notin O(A)$. By the definition of $E_i(A)$, we have

$$|z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \frac{R_j}{|z - a_{jj}|}. \tag{11.14}$$

Here, denote by s the integer such that

$$\frac{R_s}{|z - a_{ss}|} = \max_{j \neq i} \frac{R_j}{|z - a_{jj}|}. \tag{11.15}$$

(Note that s depends on z). Combined with (11.14) we get

$$|z - a_{ii}| \leq \left(\sum_{j \neq i} |a_{ij}| \right) \frac{R_s}{|z - a_{ss}|} = R_i \frac{R_s}{|z - a_{ss}|}.$$

Therefore

$$|z - a_{ii}||z - a_{ss}| \leq R_i R_s \quad \Longleftrightarrow \quad z \in C_{i,s}(A).$$

Thus we have proved $z \in E_i(A) \Rightarrow \exists s(\neq i) \in N, z \in C_{i,s}(A)$. Therefore $E(A) \subseteq C(A)$. □

The proof also reveals that if $z \in E_i$, then by finding the integer s such that $\dfrac{R_s}{|z - a_{ss}|} = \max\limits_{j \neq i} \dfrac{R_j}{|z - a_{jj}|}$, we know that z belongs to $C_{i,s}$. This in turn implies $z \in \Gamma_i \bigcup \Gamma_s$.

11.1.3.3. *Comparison with Brualdi's Lemniscate.* Next we compare the set E with another well-known Gerschgorin-type result, the Brualdi set [**18, 160**]. Its description requires the following definitions from graph theory.

For $A \in \mathbb{C}^{n\times n}$, the *vertices* of A are $1, 2, \ldots, n$. There is an *arc* (i, j) from vertices i to j if and only if $i \neq j$ and $a_{ij} \neq 0$. A *directed graph* $\mathbb{G}(A)$ is the set of all the arcs (i, j). A *strong cycle* $\gamma = (i_1, i_2, \cdots, i_p, i_{p+1})$ in $\mathbb{G}(A)$ is a sequence of integers such that $p \geq 2$, $i_{p+1} = i_1$, $\{i_1, i_2, \cdots, i_p\}$ is a set of distinct integers, and $(i_1, i_2), \ldots, (i_p, i_{p+1})$ are arcs. We say γ is of *length* p. If there exists a vertex i such that there is no strong cycle passing through i, then we say $\mathbb{G}(A)$ has a *weak cycle* $\gamma = (i)$. The *cycle set* $C(A)$ is the set of all strong and weak cycles of $\mathbb{G}(A)$.

We also define the reduced row sum $\tilde{R}_i = \sum_{j\in S_i, j\neq i} |a_{ij}|$, where S_i is the set of indices j such that there exists a strong cycle γ in $\mathbb{G}(A)$ that passes through both i and j. We set $\tilde{R}_i = 0$ when $\gamma = (i)$ is a weak cycle.

THEOREM 11.6 (Brualdi's set). *For any $n \times n$ ($n \geq 2$) complex matrix A, suppose $\gamma = (i_1, i_2, \cdots, i_p, i_{p+1})$ is a strong or weak cycle in the cycle set $C(A)$. Define the Brualdi lemniscate by*

$$B_\gamma(A) := \left\{ z \in \mathbb{C} : \prod_{i\in\gamma} |z - a_{ii}| \leq \prod_{i\in\gamma} \tilde{R}_i \right\}.$$

All the eigenvalues of A, $\lambda_p(A)$ ($p = 1, 2, \cdots, n$), lie within the union of all the Brualdi lemniscates, i.e.,

$$\{\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)\} \subseteq B(A) \equiv \bigcup_{\gamma\in C(A)} B_\gamma(A). \tag{11.16}$$

Note that computing $B(A)$ requires computing k sets, where k is the number of strong and weak cycles in $\mathbb{G}(A)$. In the worst case (when A is dense) k becomes as large as $2^n - n - 1$ [**159**], making $B(A)$ an unrealistic set.

$B(A)$ is still a useful set in certain cases, because it simplifies significantly for matrices with some sparsity structures. Furthermore, it is known to be a tighter set than $C(A)$ is.

THEOREM 11.7. [**160**] *For any $n \times n$ ($n \geq 2$) complex matrix A,*

$$B(A) \subseteq C(A).$$

We saw above that $E(A)$ and $B(A)$ are both tighter than $\Gamma(A)$ and $C(A)$. A natural question is to compare $B(A)$ with $E(A)$. The answer to this is that these are incomparable. To see this, first consider a dense matrix A, for which it is known that $B(A) = C(A)$, suggesting $E(A) \subsetneq B(A)$ (e.g., see Figure 2 in Section 11.2.1). Next, let $A \in \mathbb{C}^{n\times n}$ have nonzero off-diagonal elements only in the (i, j)th elements such that $j = i+1$ ($1 \leq i \leq n-1$) and $(i, j) = (n, 1)$. It is known [**160**, Sec. 2.4] that in such cases $B(A)$ is an optimal set, in that all the eigenvalues lie on the boundaries of $B(A)$. By contrast, it is easy to see that $E(A) = \left\{\bigcup_{1\leq i\leq n-1} C_{i,i+1}(A)\right\} \bigcup C_{1,n}(A)$, so that $B(A) \subsetneq E(A)$.

11.1.4. Cost comparison. Our result $E(A)$ may look very attractive in that it is always tighter than $C(A)$ and $\Gamma(A)$, and is a union of only n regions. By contrast, $C(A)$ involves $n(n+1)/2$ sets, and $B(A)$ requires finding the cycle set cycle(A), and can involve as many as $2^n - n - 1$ sets. However, it should be stressed that this does not mean $E(A)$ is computationally cheaper than $C(A)$ and $B(A)$. Computing $E(A)$ is generally very costly, because noting that the set $E_i(A)$ can be rewritten as

$$E_i(A) \equiv \left\{ z \in \mathbb{C} : \prod_{k=1}^{n} |z - a_{ii}| \leq \sum_{j \in N \setminus \{i\}} |a_{ij}| R_j \cdot \prod_{l \neq i, l \neq j}^{n} |z - a_{ll}| \right\},$$

we see that (11.9) is essentially an nth order polynomial in terms of $|z|$. In comparison, $\Gamma_i(A)$ can be computed simply by summing $n-1$ numbers, $C_{i,j}(A)$ is a quadratic inequality, and $B_\gamma(A)$ is a polynomial of order p, the length of the cycle γ. This suggests that $E(A)$, along with $B(A)$, is not a practical eigenvalue bound in the general case.

Nevertheless, there are situations in which $E(A)$ and $B(A)$ can be effectively applied, besides being of theoretical interest. It is known that certain sparsity structures of A simplify $B(A)$ and make it a useful set. As for $E(A)$, by observing (11.9) we notice that when the ith row of A has only p nonzero off-diagonal elements, $E_i(A)$ is a $(p+1)$st-order polynomial in terms of $|z|$. Therefore, for certain types of sparse matrices (tridiagonal, banded, etc), the cost of computing $E(A)$ should be much cheaper than that for dense matrices. In view of this, in Section 11.2.2 we consider the symmetric tridiagonal case, showing how $E(A)$ simplifies and becomes computationally realistic, while being a tighter bound than all the other three sets in this particular case.

11.2. Examples and applications

11.2.1. Examples for specific matrices. In order to illustrate the inclusion properties discussed above, we consider the matrices

$$M_1 = \begin{bmatrix} 2 & 1 & & \\ 1 & 2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 2 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 4i & 1 & 1 \\ 1 & 1 & -4 & 1 \\ 1 & 1 & 1 & -4i \end{bmatrix},$$

$$M_3 = \begin{bmatrix} 2 & 0 & 0 & 0 & 0.2 & 2 \\ 0 & 1+i & 0.02 & 0.2 & 0 & 1 \\ 0 & 0.02 & 0 & 0 & 0.2 & 1.4 \\ 0 & 0.2 & 0 & -6+2i & 0 & 1.4 \\ 0.5 & 0.5 & 0.3 & 0 & -7 & 1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.01 & -3-4i \end{bmatrix}, \quad M_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Figures 1-4 show the sets $\Gamma(M_i), C(M_i), B(M_i)$ and $E(M_i)$ $(0 \leq i \leq 4)$. Here, $B(M_i)$ and $E(M_i)$ are plotted in a naive way, i.e., by dividing the complex plane into small regions and testing for each region whether it is included in each set.

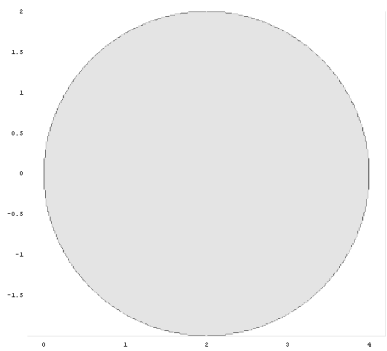

FIGURE 11.2.1. The set $\Gamma(M_1) = C(M_1) = B(M_1) = E(M_1)$.

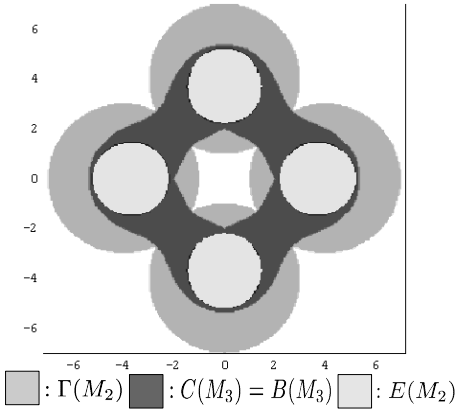

FIGURE 11.2.2. Sets $\Gamma(M_2), C(M_2)$ and $E(M_2)$.

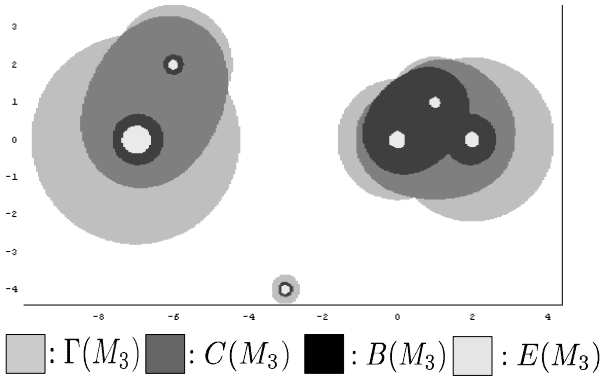

FIGURE 11.2.3. Sets $\Gamma(M_3), C(M_3)$ and $E(M_3)$.

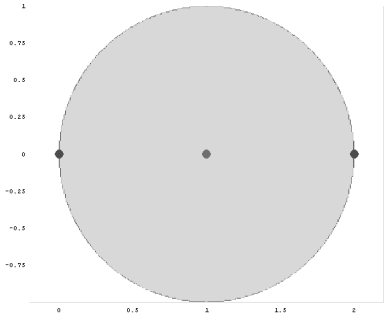

FIGURE 11.2.4. Set $\Gamma(M_4) = C(M_4) = B(M_4) = E(M_4)$ and the lemniscate of order 3 (point (1,0)). Actual eigenvalues: $\lambda = 0, 1, 1, 2$.

The inclusion properties $E(M_i) \subseteq C(M_i) \subseteq \Gamma(M_i)$ and $B(M_i) \subseteq C(M_i)$ are confirmed. For M_1, all the inclusions are equalities. For M_2 and M_3, strict inclusion properties $E(M_i) \subsetneq B(M_i) \subseteq C(M_i) \subsetneq \Gamma(M_i)$ are observed. M_3 is a case where $E(M_i)$ gives a much more accurate estimate than other sets do.

We note that M_4 is an example for which the naive extension of Brauer's ovals of Cassini

$$\bigcup_{i_1,i_2,\cdots,i_m} \left\{ z \in \mathbb{C} : \prod_{j=1}^{m} |z - a_{i_j,i_j}| \leq \prod_{j=1}^{m} R_{i_j}(A) \right\},$$

which is called the lemniscate of order m, fails to capture some of the eigenvalues [**160**, p.44]. Brualdi's set B is known to be a remedy for this problem, and can be considered a higher order extension of C. We claim that the eigenvalue inclusion property of E, together with its form, suggests that E can be regarded as another way of extending the ovals of Cassini to higher order while maintaining the eigenvalue inclusion property.

11.2.2. Application: Extremal eigenvalue bounds of a symmetric tridiagonal matrix. As we saw in Section 11.1.4, E is generally too expensive to compute, albeit being tighter than Γ and C. Here we present a situation where E can be a useful bound, being much less computationally challenging than in the general case.

11.2.2.1. *Problem statement.* Here we consider a n-by-n real symmetric tridiagonal matrix

$$A = \begin{bmatrix} a_1 & b_1 & & \\ b_1 & a_2 & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & b_{n-1} & a_n \end{bmatrix}.$$

Assuming $b_i > 0$ loses no generality. There are several applications in which we need bounds for the extremal eigenvalues of A. Here we consider getting a tight lower bound for the smallest eigenvalue $\lambda_{\min}(A)$.

11.2.2.2. *Applying known bounds.* One way to get a lower bound for $\lambda_{\min}(A)$ is to apply Gerschgorin's theorem and find the smallest intersection of the set with the real axis [**85**], which yields

$$\lambda_{\min}(A) \geq \min_{1 \leq i \leq n} \{a_i - b_{i-1} - b_i\}, \tag{11.17}$$

where for convenience we set $b_0 = b_n = 0$.

An improved bound can be obtained by using Brauer's ovals of Cassini. This amounts to finding the smallest root of the equations

$$|\lambda - a_i||\lambda - a_j| = (b_{i-1} + b_i)(b_{j-1} + b_j)$$

for $i \neq j$. Using the fact that $\lambda_{\min}(A) < a_i$ for all i, the absolute values can be removed, and solving for the smaller root yields the lower bound

$$\lambda_{\min}(A) \geq \min_{1 \leq i < j \leq n} \eta_{i,j}, \tag{11.18}$$

where $\eta_{i,j} = \dfrac{1}{2}\left(a_i + a_j - \sqrt{(a_i - a_j)^2 + 4(b_{i-1} + b_i)(b_{j-1} + b_j)}\right)$.

An even tighter bound is obtained by applying the Brualdi set. Fortunately, for tridiagonal matrices the cycle set cycle(A) consists of only $n-1$ strong cycles of length 2, namely $\gamma = (i, i+1)$ $(1 \leq i \leq n-1)$. Therefore, it follows that the Brualdi set reduces to a union of $n-1$ ovals Cassini, yielding the bound

$$\lambda_{\min}(A) \geq \min_{1 \leq i \leq n-1} \eta_{i,i+1}. \tag{11.19}$$

This bound both improves the estimate and reduces the cost compared with (11.18). We note that this coincides with the bound that is obtained using the set discussed in [**99**].

11.2.2.3. *Applying E.* Now we consider applying the set E. Computing the set $E_i(A)$ for $2 \le i \le n-1$ is equivalent to solving the equation

$$|\lambda - a_i| = \frac{b_{i-1}(b_{i-2} + b_{i-1})}{|\lambda - a_{i-1}|} + \frac{b_i(b_i + b_{i+1})}{|\lambda - a_{i+1}|},$$

which, after removing the absolute values using $\lambda_{\min}(A) < a_j$ $(j = i, i \pm 1)$, can be reduced to the cubic equation

$$g_i(\lambda) \equiv (\lambda - a_{i-1})(\lambda - a_i)(\lambda - a_{i+1}) - b_{i-1}(b_{i-2} + b_{i-1})(\lambda - a_{i+1}) - b_i(b_i + b_{i+1})(\lambda - a_{i-1}) = 0.$$

This yields the lower bound

$$\lambda_{\min}(A) \ge \min_{1 \le i \le n} \xi_i, \tag{11.20}$$

where ξ_i $(2 \le i \le n-1)$ is the smallest real root of $g_i(\lambda) = 0$, and $\xi_1 = \eta_{1,2},\ \xi_n = \eta_{n-1,n}$.

We now consider the analytic expression of ξ_i. First note that $g_i(\lambda)$ always has 3 real roots. This can be proved by the following. Note that $\lim_{\lambda \to -\infty} g_i(\lambda) = -\infty$ and $\lim_{\lambda \to -\infty} g_i(\lambda) = \infty$. If $a_{i-1} < a_{i+1}$, then we see that $g_i(a_{i-1}) > 0$ and $g_i(a_{i+1}) < 0$, so 3 real roots exist. If $a_{i-1} > a_{i+1}$, then similarly $g_i(a_{i+1}) > 0$ and $g_i(a_{i-1}) < 0$, so 3 real roots exist. If $a_{i-1} = a_{i+1}$, then $g_i(a_{i-1}) = 0$ and $g_i'(a_{i-1}) < 0$, so again 3 real roots exist.

Now to find the analytic smallest real root of $g_i(\lambda)$, we apply Cardano's method to get

$$\xi_i = \sqrt{\frac{-4p}{3}} \cos\left\{ \frac{2\pi}{3} + \frac{1}{3} \cos^{-1}\left(\frac{3\sqrt{3}q}{2p\sqrt{-p}} \right) \right\} - r,$$

where writing $g_i(\lambda) = \lambda^3 + \alpha_i \lambda^2 + \beta_i \lambda + \gamma_i$, p, q, r are defined by $p = (-\alpha_i^2 + 3\beta_i)/3,\ q = (2\alpha_i^3 - 9\alpha_i\beta_i)/27 + \gamma_i$ and $r = \alpha_i/3$. This closed form expression makes the computation of $E(A)$ much simpler than in the general case where nth order polynomials are involved.

Comparing its cost with the other bounds discussed above, we see that (11.20) is slightly more expensive than (11.19), because to compute (11.20) we need to solve $n-2$ cubic equations and 2 quadratic equations, while computing (11.19) involves n quadratic equations. Still, it can be the choice of bound because (11.20) provides the tightest bound among the four shown above. To see this, we recall that Theorems 11.2 and 11.4 tell us that the bound (11.20) is tighter than (11.17) and (11.18). In order to prove that (11.20) is tighter than (11.19), we recall that (11.20) is a union of $n-1$ ovals of Cassini, and use a similar argument to the proof of Theorem 11.5, focusing only on nonzero a_{ij}. Specifically, in (11.15) we define s as the integer such that $\dfrac{R_s}{|z - a_{ss}|} = \max\limits_{\substack{j \ne i \\ |a_{ij}| \ne 0}} \dfrac{R_j}{|z - a_{jj}|}$, after which the same argument proves the claim.

In summary, in the symmetric tridiagonal case E provides a tight and computationally realistic eigenvalue bound. In a similar manner, E is a useful eigenvalue bound for banded matrices with small bandwidths.

11.3. Gerschgorin theorems for generalized eigenproblems

We now turn to generalized eigenproblems $Ax = \lambda Bx$, and Recall that λ is a finite eigenvalue of the pair if $\det(A - \lambda B) = 0$, and in this case there exists nonzero $x \in \mathbb{C}^n$ such

that $Ax = \lambda Bx$. If the degree of the characteristic polynomial $\det(A-\lambda B)$ is $d < n$, then we say the pair has $n-d$ infinite eigenvalues. In this case, there exists a nonzero vector $x \in \mathbb{C}^n$ such that $Bx = 0$. When B is nonsingular, the pair has n finite eigenvalues, matching those of $B^{-1}A$.

In what follows we assume that for each $i \in \{1, 2, \cdots, n\}$, the ith row of either A or B is strictly diagonally dominant, unless otherwise mentioned. Although this may seem a rather restrictive assumption, its justification is the observation that the set $G(A, B)$ is always the entire complex plane unless this assumption is true.

11.3.1. Idea. Suppose $Ax = \lambda Bx$ (we consider the case $\lambda = \infty$ later). We write $x = (x_1, x_2, \cdots, x_n)^T$ and denote by $a_{p,q}$ and $b_{p,q}$the (p, q)th element of A and B respectively. Denote by i the integer such that $|x_i| = \max_{1\le j\le n} |x_j|$, so that $x_i \neq 0$. First we consider the case where the ith row of B is strictly diagonally dominant, so $|b_{i,i}| > \sum_{j\neq i} |b_{i,j}|$. From the ith equation of $Ax = \lambda Bx$ we have

$$a_{i,i}x_i + \sum_{j\neq i} a_{i,j}x_j = \lambda(b_{i,i}x_i + \sum_{j\neq i} b_{i,j}x_j). \tag{11.21}$$

Dividing both sides by x_i and rearranging yields

$$\lambda\left(b_{i,i} + \sum_{j\neq i} b_{i,j}\frac{x_j}{x_i}\right) - a_{i,i} = \sum_{j\neq i} a_{i,j}\frac{x_j}{x_i}.$$

Therefore

$$\left|\lambda\left(b_{i,i} + \sum_{j\neq i} b_{i,j}\frac{x_j}{x_i}\right) - a_{i,i}\right| \le \sum_{j\neq i} |a_{i,j}|\frac{|x_j|}{|x_i|} \le R_i, \tag{11.22}$$

where we write $R_i = \sum_{j\neq i} |a_{i,j}|$. The last inequality holds because $|x_j| \le |x_i|$ for all j. Here, using the assumption $|b_{i,i}| > \sum_{j\neq i} |b_{i,j}|$, we have

$$\begin{aligned}\left|b_{i,i} + \sum_{j\neq i} b_{i,j}\frac{x_j}{x_i}\right| &\ge |b_{i,i}| - \sum_{j\neq i} |b_{i,j}|\frac{|x_j|}{|x_i|} \\ &\ge |b_{i,i}| - \sum_{j\neq i} |b_{i,j}| > 0,\end{aligned}$$

where we used $|x_j| \le |x_i|$ again. Hence we can divide (11.22) by $\left|b_{i,i} + \sum_{j\neq i} b_{i,j}\frac{x_j}{x_i}\right|$, which yields

$$\left|\lambda - \frac{a_{i,i}}{\left(b_{i,i} + \sum_{j\neq i} b_{i,j}\frac{x_j}{x_i}\right)}\right| \le \frac{R_i}{\left|b_{i,i} + \sum_{j\neq i} b_{i,j}\frac{x_j}{x_i}\right|}. \tag{11.23}$$

Now, writing $\gamma_i = (\sum_{j\neq i} b_{i,j}\frac{x_j}{x_i})/b_{i,i}$, we have $|\gamma_i| \le \sum_{j\neq i} |b_{i,j}|/|b_{i,i}|$ $(\equiv r_i) < 1$, and (11.23) becomes

$$\left|\lambda - \frac{a_{i,i}}{b_{i,i}} \cdot \frac{1}{1+\gamma_i}\right| \le \frac{R_i}{|b_{i,i}|}\frac{1}{|1+\gamma_i|}. \tag{11.24}$$

Our interpretation of this inequality is as follows: λ lies in the disk of radius $R_i/|b_{i,i}||1+\gamma_i|$ centered at $a_{i,i}/b_{i,i}(1+\gamma_i)$, defined in the complex plane. Unfortunately the exact value of γ_i is unknown, so we cannot specify the disk. Fortunately, we show in Section 11.3.2 that using $|\gamma_i| \le r_i$ we can obtain a region that contains all the disks defined by (11.24) for any γ_i such that $|\gamma_i| \le r_i$.

Before we go on to analyze the inequality (11.24), let us consider the case where the ith row of A is strictly diagonally dominant. As we will see, this also lets us treat infinite eigenvalues.

Recall (11.21). We first note that if $|x_i| = \max_j |x_j|$ and the ith row of A is strictly diagonally dominant, then $\lambda \neq 0$, because $|a_{i,i}x_i + \sum_{j\neq i} a_{i,j}x_j| \ge |a_{i,i}||x_i| - \sum_{j\neq i}|a_{i,j}||x_j| \ge |x_i|(|a_{i,i}| - \sum_{j\neq i}|a_{i,j}|) > 0$. Therefore, in place of (11.21) we start with the equation

$$b_{i,i}x_i + \sum_{j\neq i} b_{i,j}x_j = \frac{1}{\lambda}\left(a_{i,i}x_i + \sum_{j\neq i} a_{i,j}x_j\right).$$

Note that this expression includes the case $\lambda = \infty$, because then the equation becomes $Bx = 0$. Following the same analysis as above, we arrive at the inequality corresponding to (11.24):

$$\left|\frac{1}{\lambda} - \frac{b_{i,i}}{a_{i,i}} \cdot \frac{1}{1+\gamma_i^A}\right| \le \frac{R_i^A}{|a_{i,i}|}\frac{1}{|1+\gamma_i^A|}, \tag{11.25}$$

where we write $R_i^A = \sum_{j\neq i}|b_{i,j}|$ and $\gamma_i^A = (\sum_{j\neq i} a_{i,j}\frac{x_j}{x_i})/a_{i,i}$. Note that $|\gamma_i^A| \le \sum_{j\neq i}|a_{i,j}|/|a_{i,i}|$ $(\equiv r_i^A) < 1$. Therefore we are in an essentially same situation as in (11.24), the only difference being that we are bounding $1/\lambda$ instead of λ.

In summary, in both cases the problem boils down to finding a region that contains all z such that

$$\left|z - \frac{s}{1+\gamma}\right| \le \frac{t}{|1+\gamma|}, \tag{11.26}$$

where $s \in \mathbb{C}, t > 0$ are known and $0 < r < 1$ is known such that $|\gamma| \le r$.

11.3.2. Gerschgorin theorem. First we bound the right-hand side of (11.26). This can be done simply by

$$\frac{t}{|1+\gamma|} \le \frac{t}{1-|\gamma|} \le \frac{t}{1-r}. \tag{11.27}$$

Next we consider a region that contains all the possible centers of the disk (11.24). We use the following result.

LEMMA 11.2. *If $|\gamma| \le r < 1$, then the point $1/(1+\gamma)$ lies in the disk in the complex plane of radius $r/(1-r)$ centered at 1.*

PROOF.

$$\left|\frac{1}{1+\gamma} - 1\right| = \left|\frac{\gamma}{1+\gamma}\right|$$

$$\leq \frac{r}{1-|\gamma|}$$
$$\leq \frac{r}{1-r}.$$

□

In view of (11.26), this means that $s/(1+\gamma)$, the center of the disk (11.26), has to lie in the disk of radius $sr/(1-r)$ centered at s. Combining this and (11.27), we conclude that z that satisfies (11.26) is included in the disk of radius $\dfrac{sr}{1-r}+\dfrac{t}{1-r}$, centered at s.

Using this for (11.24) by letting $s=|a_{i,i}|/|b_{i,i}|, t=R_i/|b_{i,i}|$ and $r=r_i$, we see that λ that satisfies (11.24) is necessarily included in the disk centered at $a_{i,i}/b_{i,i}$, and of radius

$$\rho_i = \frac{|a_{i,i}|}{|b_{i,i}|}\frac{r_i}{1-r_i}+\frac{R_i}{|b_{i,i}|}\frac{1}{1-r_i} = \frac{|a_{i,i}|r_i+R_i}{|b_{i,i}|(1-r_i)}.$$

Similarly, applying the result to (11.25), we see that $1/\lambda$ satisfying (11.25) has to satisfy

$$\left|\frac{1}{\lambda}-\frac{b_{i,i}}{a_{i,i}}\right| \leq \frac{|b_{i,i}|r_i^A+R_i^A}{|a_{i,i}|(1-r_i^A)}. \tag{11.28}$$

This is equivalent to

$$|a_{i,i}-\lambda b_{i,i}| \leq \frac{|b_{i,i}|r_i^A+R_i^A}{(1-r_i^A)}|\lambda|.$$

If $b_{i,i}=0$, this becomes $\dfrac{R_i^A}{(1-r_i^A)}|\lambda| \geq |a_{i,i}|$, which is $|\lambda| \geq \dfrac{|a_{i,i}|}{R_i^A}(1-r_i^A)$ when $R_i^A \neq 0$. If $b_{i,i}=R_i^A=0$, no finite λ satisfies the inequality, so we say the point $\lambda=\infty$ includes the inequality.

If $b_{i,i}\neq 0$, we have

$$\left|\lambda-\frac{a_{i,i}}{b_{i,i}}\right| \leq \frac{|b_{i,i}|r_i^A+R_i^A}{|b_{i,i}|(1-r_i^A)}\cdot|\lambda|. \tag{11.29}$$

For simplicity, we write this inequality as

$$|\lambda-\alpha_i| \leq \beta_i|\lambda|, \tag{11.30}$$

where $\alpha_i=\dfrac{a_{i,i}}{b_{i,i}}$ and $\beta_i=\dfrac{|b_{i,i}|r_i^A+R_i^A}{|b_{i,i}|(1-r_i^A)}>0$. Notice that the equality of (11.30) holds on a certain circle of Apollonius [**122**, sec.2], defined by $|\lambda-\alpha_i|=\beta_i|\lambda|$. It is easy to see that the radius of the Apollonius circle is

$$\rho_i^A = \left|\frac{1}{2}\left(\frac{|\alpha_i|}{1-\beta_i}-\frac{|\alpha_i|}{1+\beta_i}\right)\right| = \frac{|\alpha_i|\beta_i}{|1-\beta_i^2|},$$

and the center is

$$c_i = \frac{1}{2}\left(\frac{\alpha_i}{1+\beta_i}+\frac{\alpha_i}{1-\beta_i}\right) = \frac{\alpha_i}{1-\beta_i^2}.$$

From (11.30) we observe the following. The Apollonius circle divides the complex plane into two regions, and λ exists in the region that contains $\alpha_i=a_{i,i}/b_{i,i}$. Consequently, λ lies

outside the circle of Apollonius when $\beta_i > 1$, and inside it when $\beta_i < 1$. When $\beta_i = 1$, the Apollonius circle is the perpendicular bisector of the line that connects α_i and 0, dividing the complex plane into halves.

The above arguments motivate the following definition.

DEFINITION 11.1. *For $n \times n$ complex matrices A and B, denote by S^B (and S^A) the set of $i \in \{1, 2, \cdots, n\}$ such that the ith row of B (A) is strictly diagonally dominant.*

For $i \in S^B$, define the disk $\Gamma_i^B(A, B)$ by

$$\Gamma_i^B(A,B) \equiv \left\{ z \in \mathbb{C} : \left| z - \frac{a_{i,i}}{b_{i,i}} \right| \le \rho_i \right\} \quad (i = 1, 2, \cdots, n), \tag{11.31}$$

where denoting $r_i = \sum_{j \neq i} \frac{|b_{i,j}|}{|b_{i,i}|} (< 1)$ and $R_i = \sum_{j \neq i} |a_{i,j}|$, the radii ρ_i are defined by

$$\rho_i = \frac{|a_{i,i}| r_i + R_i}{|b_{i,i}|(1 - r_i)}.$$

For $i \notin S^B$, we set $\Gamma_i^B(A, B) = \mathbb{C}$, the whole complex plane.

We also define $\Gamma_i^A(A, B)$ by the following. For $i \in S^A$, denote $r_i^A = \sum_{j \neq i} \frac{|a_{i,j}|}{|a_{i,i}|} (< 1)$ and $R_i^A = \sum_{j \neq i} |b_{i,j}|$.

If $b_{i,i} = R_i^A = 0$, define $\Gamma_i^A(A, B) = \{\infty\}$, the point $z = \infty$. If $b_{i,i} = 0$ and $R_i^A > 0$, define $\Gamma_i^A(A, B) \equiv \left\{ z \in \mathbb{C} : |z| \ge \frac{|a_{i,i}|}{R_i^A}(1 - r_i^A) \right\}$.

For $b_{i,i} \neq 0$, denoting $\alpha_i = \dfrac{a_{i,i}}{b_{i,i}}$ and $\beta_i = \dfrac{|b_{i,i}| r_i^A + R_i^A}{|b_{i,i}|(1 - r_i^A)}$,

- *If $\beta_i < 1$, then define*

$$\Gamma_i^A(A,B) \equiv \left\{ z \in \mathbb{C} : |z - c_i| \le \rho_i^A \right\}, \tag{11.32}$$

  where $c_i = \dfrac{\alpha_i}{1 - \beta_i^2}$ and $\rho_i^A = \dfrac{|\alpha_i| \beta_i}{|1 - \beta_i^2|}$.

- *If $\beta_i > 1$, then define*

$$\Gamma_i^A(A,B) \equiv \left\{ z \in \mathbb{C} : |z - c_i| \ge \rho_i^A \right\}, \tag{11.33}$$

- *If $\beta_i = 1$, then define*

$$\Gamma_i^A(A,B) \equiv \left\{ z \in \mathbb{C} : |z - \alpha_i| \le |z| \right\}. \tag{11.34}$$

Finally for $i \notin S^A$, we set $\Gamma_i^A(A, B) = \mathbb{C}$.

Note that $\Gamma_i^A(A, B)$ in (11.33) and (11.34) contains the point $\{\infty\}$.

We now present our eigenvalue localization theorem.

THEOREM 11.8 (Gerschgorin-type theorem for generalized eigenvalue problems). *Let A, B be $n \times n$ complex matrices.*

All the eigenvalues of the pair (A, B) lie in the union of n regions $\Gamma_i(A, B)$ in the complex plane defined by

$$\Gamma_i(A, B) \equiv \Gamma_i^B(A, B) \cap \Gamma_i^A(A, B). \tag{11.35}$$

In other words, if λ is an eigenvalue of the pair, then

$$\lambda \in \Gamma(A, B) \equiv \bigcup_{1 \le i \le n} \Gamma_i(A, B).$$

PROOF. First consider the case where λ is a finite eigenvalue, so that $Ax = \lambda Bx$. The above arguments show that $\lambda \in \Gamma_i(A, B)$ for i such that $|x_i| = \max_j |x_j|$.

Similarly, in the infinite eigenvalue case $\lambda = \infty$, let $Bx = 0$. Note that the ith row (such that $|x_i| = \max_j |x_j|$) of B cannot be strictly diagonally dominant, because if it is, then $|b_{i,i}x_i + \sum_{j \ne i} b_{i,j}x_j| \ge |b_{i,i}||x_i| - \sum_{j \ne i} |b_{i,j}||x_j| \ge |x_i|(|b_{i,i}| - \sum_{j \ne i} |b_{i,j}|) > 0$. Therefore, $\Gamma_i^B(A, B) = \mathbb{C}$, so $\Gamma_i(A, B) = \Gamma_i^A(A, B)$. Here if $i \notin S^A$, then $\Gamma_i(A, B) = \mathbb{C}$, so $\lambda \in \Gamma(A, B)$ is trivial. Therefore we consider the case $i \in S^A$. Note that the fact that B is not strictly diagonally dominant implies $|b_{i,i}| < R_i^A$, which in turn means $\beta_i > 1$, because recalling that $\beta_i = \dfrac{|b_{i,i}|r_i^A + R_i^A}{|b_{i,i}|(1 - r_i^A)}$, we have

$$|b_{i,i}|r_i^A + R_i^A - |b_{i,i}|(1 - r_i^A) = |b_{i,i}|(2r_i^A - 1) + R_i^A > 2r_i^A|b_{i,i}| > 0.$$

Hence, recalling (11.33) we see that $\infty \in \Gamma_i^A(A, B)$.

Therefore, any eigenvalue of the pair lies in $\Gamma_i(A, B)$ for some i, so all the eigenvalues lie in the union $\bigcup_{1 \le i \le n} \Gamma_i(A, B)$. □

Theorem 11.8 shares the properties with the standard Gerschgorin theorem that it is an eigenvalue inclusion set that is easy to compute, and the boundaries are defined as circles (except for $\Gamma_i^A(A, B)$ for the special case $\beta_i = 1$). One difference between the two is that Theorem 11.8 involves $n + m$ circles, where m is the number of rows for which both A and B are strictly diagonally dominant. By contrast, the standard Gerschgorin always needs n circles. Also, when $B \to I$, the set does not become the standard Gerschgorin set, but rather becomes a slightly tighter set (owing to $\Gamma_i^A(A, B)$). Although these are not serious defects of out set $\Gamma(A, B)$, the following simplified variant solves the two issues.

DEFINITION 11.2. *We use the notations in Definition 11.1. For $i \in S^B$, define $\Gamma_i^S(A, B)$ by $\Gamma_i^S(A, B) = \Gamma_i^B(A, B)$. For $i \notin S^B$, define $\Gamma_i^S(A, B) = \Gamma_i^A(A, B)$.*

COROLLARY 11.3. *Let A, B be $n \times n$ complex matrices. All the eigenvalues of the pair (A, B) lie in $\Gamma^S(A, B) = \bigcup_{1 \le i \le n} \Gamma_i^S(A, B)$.*

PROOF. It is easy to see that $\Gamma_i(A, B) \subseteq \Gamma_i^S(A, B)$ for all i. Using Theorem 11.8 the conclusion follows immediately. □

As a special case, this result becomes a union of n disks when B is strictly diagonally dominant.

COROLLARY 11.4. *Let A, B be $n \times n$ complex matrices, and let B be strictly diagonally dominant. Then, $\Gamma_i^S(A,B) = \Gamma_i^B(A,B)$, and denoting by $\lambda_1, \cdots, \lambda_n$ the n finite eigenvalues of the pair (A,B),*

$$\lambda \in \Gamma^S(A,B) = \bigcup_{1 \le i \le n} \Gamma_i^B(A,B).$$

PROOF. The fact that $\Gamma_i^S(A,B) = \Gamma_i^B(A,B)$ follows immediately from the diagonal dominance of B. The diagonal dominance of B also forces it to be nonsingular, so that the pair (A,B) has n finite eigenvalues. □

Several points are worth noting regarding the above results.

- $\Gamma^S(A,B)$ in Corollaries 11.3 and 11.4 is defined by n circles. Moreover, it is easy to see that $\Gamma^S(A,B)$ reduces to the original Gerschgorin theorem by letting $B = I$. In this respect $\Gamma^S(A,B)$ might be considered a more natural generalization of the standard Gerschgorin theorem than $\Gamma(A,B)$. We note that these properties are shared by $K(A,B)$ in (11.4) but not shared by $G(A,B)$ in (11.2), which is defined by n regions, but not circles in the Euclidean metric, and is not equivalent to (always worse, see below) the standard Gerschgorin set when $B = I$. $\Gamma(A,B)$ also shares with $K(A,B)$ the property that it is a compact set in $\mathbb{C}$ if and only if B is strictly diagonally dominant, as mentioned in Theorem 8 in [**93**].
- $K(A,B)$ is always included in $\Gamma(A,B)$. To see this, suppose that $z \in K_i(A,B)$ so $|b_{i,i}z - a_{i,i}| \le \sum_{j \ne i} |b_{i,j}z - a_{i,j}|$. Then for $b_{i,i} \ne 0$, (note that $\Gamma_i^B(A,B) = \mathbb{C}$ so trivially $z \in \Gamma_i^B(A,B)$ if $b_{i,i} = 0$)

$$\begin{aligned}
&|b_{i,i}z - a_{i,i}| \le \sum_{j \ne i} |b_{i,j}z| + \sum_{j \ne i} |a_{i,j}| \\
\Leftrightarrow &|z - \frac{a_{i,i}}{b_{i,i}}| - \sum_{j \ne i} \frac{|b_{i,j}z|}{|b_{i,i}|} \le \frac{R_i}{|b_{i,i}|} \quad \left(\text{recall } R_i = \sum_{j \ne i} |a_{i,j}| \right) \\
\Rightarrow &|z - \frac{a_{i,i}}{b_{i,i}}| - r_i|z| \le \frac{R_i}{|b_{i,i}|}. \quad \left(\text{recall } r_i = \sum_{j \ne i} \frac{|b_{i,j}|}{|b_{i,i}|} \right)
\end{aligned}$$

Since we can write $|z - a_{i,i}| - r_i|z| = |z - a_{i,i} + r_i e^{i\theta} z|$ for some $\theta \in [0, 2\pi]$, it follows that if $z \in K_i(A,B)$ then

$$\left| z(1 + r_i e^{i\theta}) - \frac{a_{i,i}}{b_{i,i}} \right| \le \frac{R_i}{|b_{i,i}|}.$$

Since $r_i < 1$, we can divide this by $(1 + r_i e^{i\theta})$, which yields

$$\left| z - \frac{a_{i,i}}{b_{i,i}} \frac{1}{1 + r_i e^{i\theta}} \right| \le \frac{R_i}{|b_{i,i}|} \frac{1}{|1 + r_i e^{i\theta}|}. \tag{11.36}$$

Note that this becomes (11.24) if we substitute γ_i into $r_i e^{i\theta}$ and λ into z. Now, since $\Gamma_i^B(A,B)$ is derived from (11.24) by considering a disk that contains λ that satisfies (11.24) for any γ_i such that $|\gamma_i| < r_i$, it follows that z that satisfies (11.36) is included

in $\Gamma_i^B(A,B)$. By a similar argument we can prove $z \in K_i(A,B) \Rightarrow z \in \Gamma_i^A(A,B)$, so the claim is proved.

- Although $K(A,B)$ is always sharper than $\Gamma(A,B)$ is, $\Gamma(A,B)$ has the obvious advantage over $K(A,B)$ in its practicality. $\Gamma(A,B)$ is much easier to compute than $K(A,B)$, which is generally a union of complicated regions. It is also easy to see that $\Gamma(A,B)$ approaches $K(A,B)$ as B approaches a diagonal matrix, see examples in Section 11.4. $\Gamma(A,B)$ sacrifices some tightness for the sake of simplicity. For instance, $K(A,B)$ is difficult to use for the analysis in Section 11.5.2.
- $G(A,B)$ and $\Gamma(A,B)$ are generally not comparable, see the examples in Section 11.4. However, we can see that $\Gamma_i(A,B)$ is a nontrivial set in the complex plane $\mathbb{C}$ whenever $G_i(A,B)$ is, but the contrary does not hold. This can be verified by the following. Suppose $G_i(A,B)$ is a nontrivial set in $\mathbb{C}$, which means $(\sum_{j\neq i}|a_{i,j}|)^2 + (\sum_{j\neq i}|b_{i,j}|)^2 < |a_{i,i}|^2 + |b_{i,i}|^2$. This is true only if $\sum_{j\neq i}|a_{i,j}| < |a_{i,i}|$ or $\sum_{j\neq i}|b_{i,j}| < |b_{i,i}|$, so the ith row of at least one of A and B has to be strictly diagonally dominant. Hence, $\Gamma_i(A,B)$ is a nontrivial subset of $\mathbb{C}$.

  To see the contrary is not true, consider the pair (A_1, B_1) defined by

$$A_1 = \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{11.37}$$

  which has eigenvalues -1 and $5/3$. For this pair $\Gamma(A_1,B_1) = \{z \in \mathbb{C} : |z-1| \le 4\}$. In contrast, we have $G(A_1,B_1) = \left\{z \in \mathbb{C} : \chi(\lambda,1) \le \sqrt{10/8}\right\}$, which is useless because the chordal radius is larger than 1.
- When $B \simeq I$, $\Gamma(A,B)$ is always a tighter region than $G(A,B)$ is, because $G_i(A,I)$ is

$$\frac{|\lambda - a_{i,i}|}{\sqrt{1+|\lambda|^2}\sqrt{1+|a_{i,i}|^2}} \lesssim \sqrt{\frac{\left(\sum_{j\neq i}|a_{i,j}|\right)^2}{1+|a_{i,i}|^2}} = \sqrt{\frac{R_i^2}{1+|a_{i,i}|^2}}.$$

  Hence

$$|\lambda - a_{i,i}| \lesssim \sqrt{1+|\lambda|^2}R_i,$$

  whereas $\Gamma_i^S(A,I)$ is the standard Gerschgorin set

$$|\lambda - a_{i,i}| \le R_i,$$

  from which $\Gamma_i^S(A,B) \subseteq G_i(A,I)$ follows trivially.

11.3.3. A tighter result. Here we show that we can obtain a slightly tighter eigenvalue inclusion set by bounding the center of the disk (11.26) more carefully. Instead of Lemma 11.2, we use the following two results.

LEMMA 11.3. *The point $1/(1+re^{i\theta})$ where $r \ge 0$ and $\theta \in [0, 2\pi]$ lies on a circle of radius $r/(1-r^2)$ centered at $1/(1-r^2)$.*

PROOF.

$$\left|\frac{1}{1+re^{i\theta}} - \frac{1}{1-r^2}\right| = \left|\frac{(1-r^2)-(1+re^{i\theta})}{(1+re^{i\theta})(1-r^2)}\right|$$

$$= \left| \frac{r(r+e^{i\theta})}{(1+re^{i\theta})(1-r^2)} \right|$$
$$= \left| \frac{re^{i\theta}(1+re^{-i\theta})}{(1+re^{i\theta})(1-r^2)} \right|$$
$$= \left| \frac{r}{1-r^2} \right|. \quad \text{(because } \left| \frac{1+re^{-i\theta}}{1+re^{i\theta}} \right| = 1)$$

□

LEMMA 11.4. *Denote by $M(r)$ the disk of radius $r/(1-r^2)$ centered at $1/(1-r^2)$. If $0 \le r' < r < 1$ then $M(r') \subseteq M(r)$.*

PROOF. We prove by showing that $z \in M(r') \Rightarrow z \in M(r)$. Suppose $z \in M(r')$. z satisfies $\left| z - \frac{1}{1-(r')^2} \right| \le \left| \frac{r'}{1-(r')^2} \right|$, so

$$\left| z - \frac{1}{1-r^2} \right| \le \left| z - \frac{1}{1-(r')^2} \right| + \left| \frac{1}{1-(r')^2} - \frac{1}{1-r^2} \right|$$
$$\le \left| \frac{r'}{1-(r')^2} \right| + \left| \frac{r^2-(r')^2}{(1-(r')^2)(1-r^2)} \right|$$
$$= \frac{r'(1-r^2)+r^2-(r')^2}{(1-(r')^2)(1-r^2)}.$$

Here, the right-hand side is smaller than $r/(1-r^2)$, because

$$\frac{r}{1-r^2} - \frac{r'(1-r^2)+r^2-(r')^2}{(1-(r')^2)(1-r^2)} = \frac{r(1-(r')^2)-(r'(1-r^2)+r^2-(r')^2)}{(1-(r')^2)(1-r^2)}$$
$$= \frac{(1-r)(1-r')(r-r')}{(1-(r')^2)(1-r^2)} > 0.$$

Hence $\left| z - \frac{1}{1-r^2} \right| \le \frac{r}{1-r^2}$, so $z \in M(r)$. Since the above argument holds for any $z \in M(r')$, $M(r') \subseteq M(r)$ is proved. □

The implication of these two Lemmas applied to (11.26) is that the center $s/(1+\gamma)$ lies in $sM(r)$. Therefore we conclude that z that satisfies (11.26) is included in the disk centered at $\frac{s}{1-r^2}$, and of radius $\frac{sr}{1-r^2} + \frac{t}{1-r}$.

Therefore, it follows that λ that satisfies (11.24) lies in the disk of radius $\frac{|a_{i,i}|r_i + R_i(1+r_i)}{|b_{i,i}|(1-r_i^2)}$, centered at $\frac{a_{i,i}}{b_{i,i}(1-r_i^2)}$.

Similarly, we can conclude that $1/\lambda$ that satisfies (11.25) has to satisfy

$$\left| \frac{1}{\lambda} - \frac{b_{i,i}}{a_{i,i}} \cdot \frac{1}{1-(r_i^A)^2} \right| \le \frac{|b_{i,i}|r_i^A + R_i^A(1+r_i^A)}{|a_{i,i}|(1-(r_i^A)^2)}. \tag{11.38}$$

Recalling the analysis that derives (11.30), we see that when $b_{i,i} \neq 0$, this inequality is equivalent to

$$|\lambda - \tilde{\alpha}_i| \leq \tilde{\beta}_i |\lambda|, \tag{11.39}$$

where $\tilde{\alpha}_i = a_{i,i}(1-(r_i^A)^2)/b_{i,i}$, $\tilde{\beta}_i = r_i^A + R_i^A(1+r_i^A)/|b_{i,i}|$.

The equality of (11.39) holds on an Apollonius circle, whose radius is $\tilde{\rho}_i^A = \dfrac{|\tilde{\alpha}_i|\tilde{\beta}_i}{|1-\tilde{\beta}_i^2|}$, and center is $c_i = \dfrac{\tilde{\alpha}_i}{1-\tilde{\beta}_i^2}$.

The above analyses leads to the following definition, analogous to that in Definition 11.1.

Definition 11.5. *We use the same notations $S, S^A, r_i, R_i, r_i^A, R_i^A$ as in Definition 11.1. For $i \in S^B$, define the disk $\tilde{\Gamma}_i^B$ by*

$$\tilde{\Gamma}_i^B(A,B) \equiv \left\{ z \in \mathbb{C} : \left| z - \frac{a_{i,i}}{b_{i,i}} \frac{1}{1-(r_i)^2} \right| \leq \tilde{\rho}_i \right\} \quad (i = 1, 2, \cdots, n), \tag{11.40}$$

where the radii $\tilde{\rho}_i$ are defined by

$$\tilde{\rho}_i = \frac{|a_{i,i}| r_i + R_i(1+r_i)}{|b_{i,i}|(1-r_i^2)}.$$

For $i \notin S^B$, we set $\tilde{\Gamma}_i^B(A,B) = \mathbb{C}$.

$\tilde{\Gamma}_i^A(A,B)$ is defined by the following. For $i \in S^A$ and $b_{i,i} \neq 0$, denote

$$\tilde{\alpha}_i = \frac{a_{i,i}}{b_{i,i}}(1-(r_i^A)^2), \tilde{\beta}_i = r_i^A + \frac{R_i^A(1+r_i^A)}{|b_{i,i}|}, \tilde{c}_i = \frac{\alpha_i}{1-\beta_i^2} \text{ and } \tilde{\rho}_i^A = \frac{|\alpha_i|\beta_i}{|1-\beta_i^2|}.$$

Then, $\tilde{\Gamma}_i^A(A,B)$ is defined similarly to $\Gamma_i^A(A,B)$ (by replacing $\alpha_i, \beta_i, c_i, \rho_i^A$ with $\tilde{\alpha}_i, \tilde{\beta}_i, \tilde{c}_i, \tilde{\rho}_i^A$ respectively in (11.32)-(11.34)), depending on whether $\tilde{\beta}_i > 1, \tilde{\beta}_i < 1$ or $\tilde{\beta}_i = 1$.

When $b_{i,i} = 0$ or $i \notin S^A$, $\tilde{\Gamma}_i^A(A,B) = \Gamma_i^A(A,B)$ defined in Definition 11.1.

Thus we arrive at a slightly tighter Gerschgorin theorem.

Theorem 11.9 (Tighter Gerschgorin-type theorem). *Let A, B be $n \times n$ complex matrices.*

All the eigenvalues of the pair (A,B) lie in the union of n regions $\tilde{\Gamma}_i(A,B)$ in the complex plane defined by

$$\tilde{\Gamma}_i(A,B) \equiv \tilde{\Gamma}_i^B(A,B) \cap \tilde{\Gamma}_i^A(A,B). \tag{11.41}$$

In other words, if λ is an eigenvalue of the pair, then

$$\lambda \in \tilde{\Gamma}(A,B) \equiv \bigcup_{1 \leq i \leq n} \tilde{\Gamma}_i(A,B).$$

The proof is the same as the one for Theorem 11.8 and is omitted. The simplified results of Theorem 11.9 analogous to Corollaries 11.3 and 11.4 can also be derived but is omitted.

It is easy to see that $\tilde{\Gamma}_i(A,B) \subseteq \Gamma_i(A,B)$ for all i, so $\tilde{\Gamma}(A,B)$ is a sharper eigenvalue bound than $\Gamma(A,B)$. For example, for the pair (11.37), we have $\tilde{\Gamma}_i(A_1,B_1) = \left\{ z \in \mathbb{C} : \left| z - \frac{4}{3} \right| \leq \frac{11}{3} \right\}$. We can also see that $\tilde{\Gamma}(A,B)$ shares all the properties mentioned at the end of Section 11.3.2.

The reason we presented $\Gamma(A,B)$ although $\tilde{\Gamma}(A,B)$ is always tighter is that $\Gamma(A,B)$ has centers $a_{i,i}/b_{i,i}$, which may make it simpler to apply than $\tilde{\Gamma}(A,B)$. In fact, in the analysis in Section 11.5.2 we only use Theorem 11.8.

11.3.4. Localizing a specific number of eigenvalues. We are sometimes interested not only in the eigenvalue bounds, but also in the number of eigenvalues included in a certain region. The classical Gerschgorin theorem serves this need [**160**], which has the property that if a region contains exactly k Gerschgorin disks and is disjoint from the other disks, then it contains exactly m eigenvalues. This fact is used to derive a perturbation result for simple eigenvalues in [**165**]. An analogous result holds for the set $G(A,B)$ [**146**, Ch.5]. Here we show that our Gerschgorin set also possesses the same property.

THEOREM 11.10. *If a union of k Gerschgorin regions $\Gamma_i(A,B)$ (or $\tilde{\Gamma}_i(A,B)$) in the above Theorems (Theorem 11.8, 11.9 or Corollary 11.3, 11.4) is disjoint from the remaining $n-k$ regions and is not the entire complex plane $\mathbb{C}$, then exactly k eigenvalues of the pair (A,B) lie in the union.*

PROOF. We prove the result for $\Gamma_i(A,B)$. The other sets can be treated in an entirely identical way.

We use the same trick used for proving the analogous result for the set $G(A,B)$, shown in [**146**, Ch.5]. Let $\tilde{A} = \mathrm{diag}(a_{11}, a_{22}, \cdots, a_{nn})$, $\tilde{B} = \mathrm{diag}(b_{11}, b_{22}, \cdots, b_{nn})$ and define

$$A(t) = \tilde{A} + t(A - \tilde{A}), \quad B(t) = \tilde{B} + t(B - \tilde{B}).$$

It is easy to see that the Gerschgorin disks $\Gamma_i(A(t), B(t))$ get enlarged as t increases from 0 to 1.

In [**146**] it is shown in the chordal metric that the eigenvalues of a regular pair (A,B) are continuous functions of the elements provided that the pair is regular.

Note that each of the regions $\Gamma_i(A(t), B(t))$ is a closed and bounded subset of $\mathbb{C}$ in the chordal metric, and that if a union of k regions $\Gamma_i(A(t), B(t))$ is disjoint from the other $n-k$ regions in the Euclidean metric, then this disjointness holds also in the chordal metric. Therefore, if the pair $(A(t), B(t))$ is regular for $0 \le t \le 1$, then an eigenvalue that is included in a certain union of k disks $\bigcup_{1\le i\le k} \Gamma_i(A(t), B(t))$ cannot jump to another disjoint region as t increases, so the claim is proved. Hence it suffices to prove that the pair $(A(t), B(t))$ is regular.

The regularity is proved by contradiction. If $(A(t), B(t))$ is singular for some $0 \le t \le 1$, then any point $z \in \mathbb{C}$ is an eigenvalue of the pair. However, the disjointness assumption implies that there must exist a point $z' \in \mathbb{C}$ such that z' lies in none of the Gerschgorin disks, so z' cannot be an eigenvalue. Therefore, $(A(t), B(t))$ is necessarily regular for $0 \le t \le 1$.

□

11.4. Examples

Here we show some examples to illustrate the regions we discussed above. As test matrices we consider the simple n-by-n pair (A, B) where

$$A = \begin{bmatrix} 4 & a & & \\ a & 4 & \ddots & \\ & \ddots & \ddots & a \\ & & a & 4 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 4 & b & & \\ b & 4 & \ddots & \\ & \ddots & \ddots & b \\ & & b & 4 \end{bmatrix}. \tag{11.42}$$

Note that $\Gamma(A, B)$, $\tilde{\Gamma}(A, B)$ and $K(A, B)$ are nontrivial regions if $b < 2$, and $G(A, B)$ is nontrivial only if $a^2+b^2 < 8$. Figure 11.4.1 shows our results $\Gamma(A, B)$ and $\tilde{\Gamma}(A, B)$ for different parameters (a, b). The two crossed points indicate the smallest and largest eigenvalues of the pair (11.42) when the matrix size is $n = 100$. For $(a, b) = (1, 2)$ the largest eigenvalue (not shown) was $\simeq 1034$. Note that $\Gamma(A, B) = \Gamma^B(A, B)$ when $(a, b) = (2, 1)$ and $\Gamma(A, B) = \Gamma^A(A, B)$ when $(a, b) = (1, 2)$.

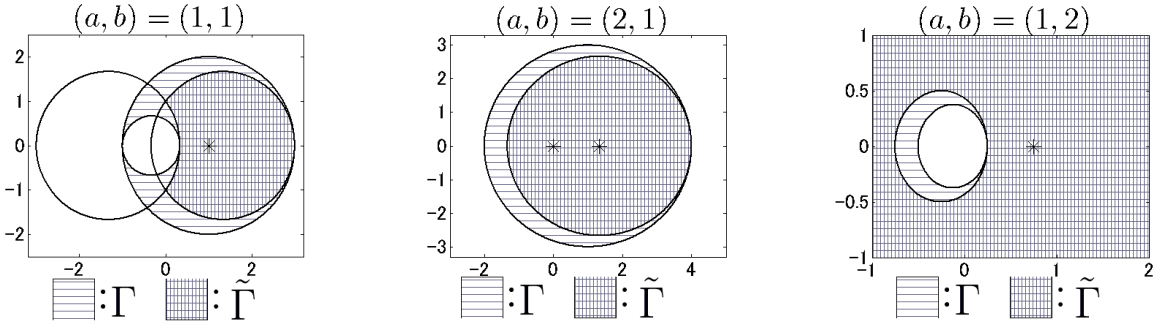

FIGURE 11.4.1. Plots of $\Gamma(A, B)$ and $\tilde{\Gamma}(A, B)$ for matrices (11.42) with different a, b.

The purpose of the figures below is to compare our results with the known results $G(A, B)$ and $K(A, B)$. As for our results we only show $\Gamma^S(A, B)$ for simplicity. Figure 11.4.2 compares $\Gamma^S(A, B)$ with $G(A, B)$. We observe that in the cases $(a, b) = (2, 1), (3, 1)$, $\Gamma(A, B)$ is a much more useful set than $G(A, B)$ is, which in the latter case is the whole complex plane. This reflects the observation given in Section 11.3.2 that $\Gamma(A, B)$ is always tighter when $B \simeq I$.

Figure 11.4.3 compares $\Gamma^S(A, B)$ with $K(A, B)$, in which the boundary of $\tilde{\Gamma}^S(A, B)$ is shown as dashed circles. We verify the relation $K(A, B) \subseteq \tilde{\Gamma}(A, B) \subseteq \Gamma(A, B)$. These three sets become equivalent when B is nearly diagonal, as shown in the middle graph. The right graph shows the regions for the matrix defined in Example 1 in [**93**], in which all the eigenvalues are shown as crossed points.

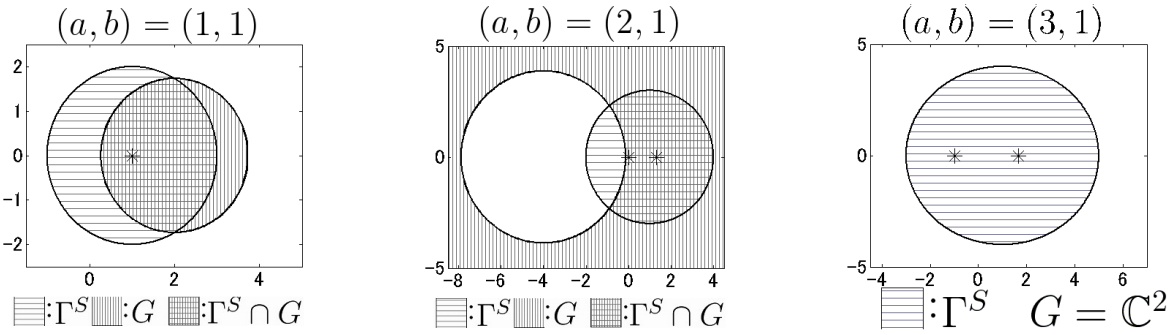

FIGURE 11.4.2. Plots of $\Gamma^S(A,B)$ and $G(A,B)$ for matrices (11.42) with different a, b.

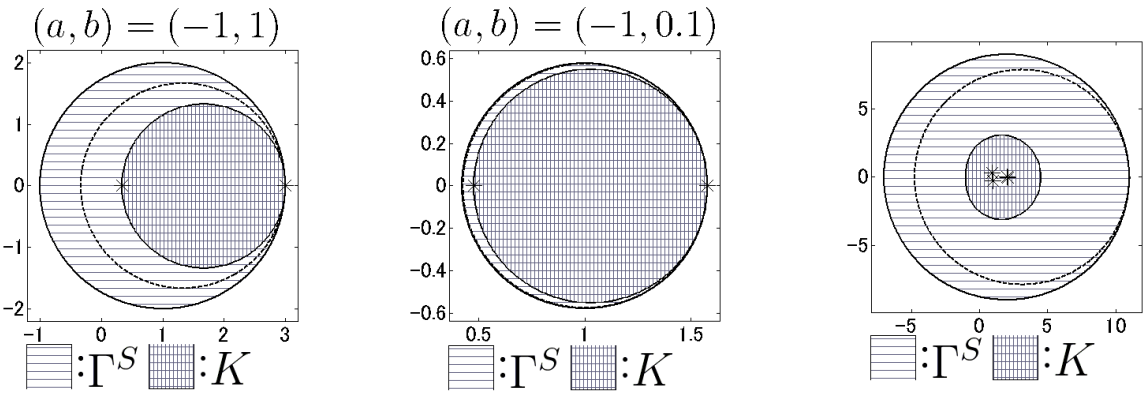

FIGURE 11.4.3. Plots of $\Gamma(A,B)$ and $K(A,B)$

We emphasize that our result $\Gamma(A,B)$ is defined by circles and so is easy to plot, while the regions $K(A,B)$ and $G(A,B)$ are generally complicated regions, and are difficult to plot. In the above figures we obtained $K(A,B)$ and $G(A,B)$ by a very naive method, i.e., by dividing the complex plane into small regions and testing whether the center of the region is contained in each set.

11.5. Applications

In this section we describe two applications for which the proposed Gerschgorin theorem for generalized eigenvalue problems may be useful.

11.5.1. Nonlinear eigenvalue problems. Nonlinear (polynomial) eigenvalue problems of the form $P(\lambda)x = 0$ where $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ arise in many areas of application. The standard eigenvalue problem $Ax = \lambda x$ and generalized eigenvalue problems $Ax = \lambda Bx$, on which we have focused so far in this dissertation, can be considered special cases of polynomial eigenvalue problems. There has been significant recent developments in its theory and numerical solution, see for example [**54, 65, 105, 153**]. Most approaches to solving $P(\lambda)x = 0$ is to first linearize the problem to obtain an nk-by-nk generalized eigenvalue problem $L(\lambda)z = (\lambda B + A)z = 0$ with the same eigenvalues, then solve $L(\lambda)z = 0$, for which

many reliable methods exist. The most well-known linearization is the companion form

$$L(\lambda) = \lambda \begin{bmatrix} A_k & 0 & \cdots & 0 \\ 0 & I_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I_n \end{bmatrix} + \begin{bmatrix} A_{k-1} & A_{k-2} & \cdots & A_0 \\ -I_n & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -I_n & 0 \end{bmatrix}.$$

Hence if our Gerschgorin theory applied to this generalized eigenvalue problem gives a non-trivial set (e.g., when A_k is diagonally dominant) we can obtain a simple bound on the eigenvalues of the polynomial eigenvalue problem $P(\lambda)$, thereby generalizing the bounds in [**79**] which assume $A_k = I_n$ (or involves A_k^{-1}). We can apply the same argument to any other linearization, and in particular we see that a linearization that makes B diagonally dominant is ideal for estimating the eigenvalues by this approach.

11.5.2. Generalized eigenvalue forward error analysis. The Gerschgorin theorems presented in Section 11.3 can be used in a straightforward way for a matrix pair with some diagonal dominance property whenever one wants a simple estimate for the eigenvalues or bounds for the extremal eigenvalues, as the standard Gerschgorin theorem is used for standard eigenvalue problems.

Here we show how our results can also be used to provide a forward error analysis for computed eigenvalues of a n-by-n diagonalizable pair (A, B).

For simplicity we assume only finite eigenvalues exist. After the computation of eigenvalues $\tilde{\lambda}_i$ $(1 \le i \le n)$ and eigenvectors (both left and right) one can normalize the eigenvectors to get $X, Y \in \mathbb{C}^{n\times n}$ such that

$$Y^*AX(\equiv \widehat{A}) = \begin{bmatrix} \tilde{\lambda}_1 & e_{1,2} & \cdots & e_{1,n} \\ e_{2,1} & \tilde{\lambda}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & e_{n-1,n} \\ e_{n,1} & \cdots & e_{n,n-1} & \tilde{\lambda}_n \end{bmatrix} = \operatorname{diag}\{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_n\} + E,$$

$$Y^*BX(\equiv \widehat{B}) = \begin{bmatrix} 1 & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & f_{n-1,n} \\ f_{n,1} & \cdots & f_{n,n-1} & 1 \end{bmatrix} = I + F.$$

The matrices E and F represent the errors, which we expect to be small after a successful computation (note that in practice computing the matrix products Y^*AX, Y^*BX also introduces errors, but here we ignore this effect, to focus on the accuracy of the eigensolver). We denote by $E_j = \sum_l |e_{j,l}|$ and $F_j = \sum_l |f_{j,l}|$ $(1 \le j \le n)$ their absolute jth row sums. We assume that $F_j < 1$ for all j, or equivalently that $I + F$ is strictly diagonally dominant, so in the following we only consider $\Gamma^S(A, B)$ in Corollary 11.3 and refer to it as the Gerschgorin disk.

We note that the assumption that both eigenvalues and eigenvectors are computed restricts the problem size to moderate n. Nonetheless, computation of a full eigendecomposition of a small-sized problem is often necessary in practice. For example, it is the computational kernel of the Rayleigh-Ritz process in a method for computing several eigenpairs of a large-scale problem [**6**, Ch. 5].

Simple bound. For a particular computed eigenvalue $\tilde{\lambda}_i$, we are interested in how close it is to an "exact" eigenvalue of the pair (A, B). We consider the simple and multiple eigenvalue cases separately.

(1) When $\tilde{\lambda}_i$ is a simple eigenvalue. We define $\delta \equiv \min_{j\neq i} |\tilde{\lambda}_i - \tilde{\lambda}_j| > 0$. If E and F are small enough, then $\Gamma_i(\widehat{A}, \widehat{B})$ is disjoint from all the other $n-1$ disks. Specifically, this is true if $\delta > \rho_i + \rho_j$ for all $j \neq i$, where

$$\rho_i = \frac{|\tilde{\lambda}_i|F_i + E_i}{1 - F_i}, \quad \rho_j = \frac{|\tilde{\lambda}_i|F_j + E_j}{1 - F_j} \tag{11.43}$$

are the radii of the ith and jth Gerschgorin disks in Theorem 11.8, respectively. If the inequalities are satisfied for all $j \neq i$, then using Theorem 11.10 we conclude that there exists exactly 1 eigenvalue λ_i of the pair $(\widehat{A}, \widehat{B})$ (which has the same eigenvalues as (A, B)) such that

$$|\lambda_i - \tilde{\lambda}_i| \leq \rho_i. \tag{11.44}$$

(2) When $\tilde{\lambda}_i$ is a multiple eigenvalue of multiplicity k, so that $\tilde{\lambda}_i = \tilde{\lambda}_{i+1} = \cdots = \tilde{\lambda}_{i+k-1}$. It is straightforward to see that a similar argument holds and if the k disks $\Gamma_{i+l}(\widehat{A}, \widehat{B})$ $(0 \leq l \leq k-1)$ are disjoint from the other $n-k$ disks, then there exist exactly k eigenvalues λ_j $(i \leq j \leq i+k-1)$ of the pair (A, B) such that

$$|\lambda_j - \tilde{\lambda}_i| \leq \max_{0\leq l\leq k} \rho_{i+l}. \tag{11.45}$$

Tighter bound. Here we derive another bound that can be much tighter than (11.44) when the error matrices E and F are small. We use the technique of diagonal similarity transformations employed in [**165, 146**], where first-order eigenvalue perturbation results are obtained.

We consider the case where $\tilde{\lambda}_i$ is a simple eigenvalue and denote $\delta \equiv \min_{j\neq i} |\tilde{\lambda}_i - \tilde{\lambda}_j| > 0$, and suppose that the ith Gerschgorin disk of the pair $(\widehat{A}, \widehat{B})$ is disjoint from the others.

Let T be a diagonal matrix whose ith diagonal is τ and 1 otherwise. We consider the Gerschgorin disks $\Gamma_j(T\widehat{A}T^{-1}, T\widehat{B}T^{-1})$, and find the smallest τ such that the ith disk is disjoint from the others. By the assumption, this disjointness holds when $\tau = 1$, so we only consider $\tau < 1$.

The center of $\Gamma_j(T\widehat{A}T^{-1}, T\widehat{B}T^{-1})$ is $\tilde{\lambda}_j$ for all j. As for the radii $\widehat{\rho}_i$ and $\widehat{\rho}_j$, for $\tau < F_i, F_j$ we have

$$\widehat{\rho}_i = \frac{\tau|\tilde{\lambda}_i|F_i + \tau E_i}{1 - \tau F_i} \leq \tau\rho_i,$$

and

$$\widehat{\rho}_j \leq \frac{|\tilde{\lambda}_j|F_j/\tau + E_j/\tau}{1 - F_j/\tau}, \quad \text{for } j \neq i.$$

Since $\tau < 1$, we see that writing $\delta_j = |\tilde{\lambda}_i - \tilde{\lambda}_j|$,

$$\rho_i + \widehat{\rho}_j < \delta_j \tag{11.46}$$

is a sufficient condition to for the disks $\Gamma_i(T\widehat{A}T^{-1}, T\widehat{B}T^{-1})$ and $\Gamma_j(T\widehat{A}T^{-1}, T\widehat{B}T^{-1})$ to be disjoint. (11.46) is satisfied if

$$\begin{aligned}
&\rho_i + \frac{|\tilde{\lambda}_j|F_j/\tau + E_j/\tau}{1 - F_j/\tau} < \delta_j \\
\Leftrightarrow &(\tau - F_j)(\delta_j - \rho_i) > |\tilde{\lambda}_j|F_j + E_j \\
\Leftrightarrow &\tau > F_j + \frac{|\tilde{\lambda}_j|F_j + E_j}{\delta_j - \rho_i},
\end{aligned}$$

where we used $\delta_j - \rho_i > 0$, which follows from the disjointness assumption. Here, since $\delta_j \geq \delta > \rho_i$, we see that (11.46) is true if

$$\tau > F_j + \frac{|\tilde{\lambda}_j|F_j + E_j}{\delta - \rho_i}.$$

Repeating the same argument for all $j \neq i$, we conclude that if

$$\tau > F_j + \frac{\max_{j\neq i}\{|\tilde{\lambda}_j|F_j + E_j\}}{\delta - \rho_i} \ (\equiv \tau_0), \tag{11.47}$$

then the disk $\Gamma_i(T\widehat{A}T^{-1}, T\widehat{B}T^{-1})$ is disjoint from the remaining $n-1$ disks.

Therefore, by letting $\tau = \tau_0$ and using Theorem 11.10 for the pair $(T\widehat{A}T^{-1}, T\widehat{B}T^{-1})$, we conclude that there exists exactly one eigenvalue λ_i of the pair (A, B) such that

$$|\lambda_i - \tilde{\lambda}_i| \leq \frac{\tau_0(|\tilde{\lambda}_i|F_i + E_i)}{1 - \tau_0 F_i} \leq \tau_0 \rho_i. \tag{11.48}$$

Using $\delta \leq |\tilde{\lambda}_i| + |\tilde{\lambda}_j|$, we can bound τ_0 from above by

$$\tau_0 \leq \frac{\max_{j\neq i}\{(2|\tilde{\lambda}_j| + |\tilde{\lambda}_i|)F_j + E_j\}}{\delta - \rho_i} \leq \frac{\max_{j\neq i}\{(2|\tilde{\lambda}_j| + |\tilde{\lambda}_i|)F_j + E_j\}}{(1 - F_i)(\delta - \rho_i)}.$$

Also observe from (11.43) that

$$\rho_i = \frac{|\tilde{\lambda}_i|F_i + E_i}{1 - F_i} \leq \frac{\max_{1\leq j\leq n}\{(2|\tilde{\lambda}_j| + |\tilde{\lambda}_i|)F_j + E_j\}}{1 - F_i}.$$

Therefore, denoting $\delta' = \delta - \rho_i$ and $r = \frac{1}{1-F_i}\max_{1\leq j\leq n}\{(2|\tilde{\lambda}_j| + |\tilde{\lambda}_i|)F_j + E_j\}$, we have $\tau_0 \leq r/\delta'$ and $\rho_i \leq r$. Hence, from (11.48) we conclude that

$$|\lambda_i - \tilde{\lambda}_i| \leq \frac{r^2}{\delta'}. \tag{11.49}$$

Since r is essentially the size of the error, and δ' is essentially the gap between $\tilde{\lambda}_i$ and any other computed eigenvalue, we note that this bound resembles the quadratic bound for the standard Hermitian eigenvalue problem, $|\tilde{\lambda} - \lambda| \leq \|R\|^2/\delta$ [**127**, Ch.11]. Our result (11.49) indicates that this type of quadratic error bound holds also for the non-Hermitian generalized eigenvalue problems.

CHAPTER 12

Summary and future work

Here we give a brief account of the topics and developments treated in this dissertation and mention possible directions for future work.

The efficient, communication-minimizing algorithms proposed in Part 1 are potentially very attractive on emerging computing architectures. Our treatment here focused more on the theoretical aspects, the algorithmic development and the stability. The numerical experiments were all done in MATLAB on a quad-core machine, and did not outperform the best available algorithm (divide-and-conquer) in speed. A natural avenue to pursue further is to implement a Fortran code that take full advantage of the communication-minimizing feature, and test on highly parallel computers.

We showed that with the help of aggressive early deflation the dqds algorithm not only becomes faster, but parallelizable. The next natural step is to test and optimize such a parallel version of dqd(s). In doing so we expect that many details will need to be worked out to get best performance. For example, a shift/split strategy suitable for a parallel version will certainly need to be reconsidered.

The eigenvalue perturbation bounds we derived for Hermitian block tridiagonal matrices can be much tighter than existing results. In particular, using our bounds we can conclude that some of the eigenvalues can be computed accurately from a much smaller submatrix. Such studies have attracted much attention recently and our contribution may be used to assist the progress.

We extended two of the most well-known eigenvalue perturbation bounds for Hermitian eigenproblems to generalized Hermitian eigenproblems. Perturbation theory on Hermitian eigenproblems has numerous other existing results, and a natural next step is to consider extending them as well.

The new eigenvector perturbation bounds suggest that eigenvectors computed via the Rayleigh-Ritz process can be much more accurate than existing results guarantee. We devoted a subsection to describe a possible future work of the efficient execution of an inexact Rayleigh-Ritz process. This has the potential of significantly reducing the cost of an eigensolver for large-scale sparse Hermitian matrices.

Gerschgorin's theorem for standard eigenproblems has found a vast area of applications. Naturally one may consider using our Gerschgorin-type set for generalized eigenproblems. One does need to note that the applicable matrices are necessarily restricted, at least one of A and B being required to be diagonally dominant for each row.

Bibliography

[1] K. Aishima, T. Matsuo, K. Murota, and M. Sugihara. On convergence of the dqds algorithm for singular value computation. *SIAM J. Matrix Anal. Appl.*, 30(2):522–537, 2008.

[2] K. Aishima, T. Matsuo, K. Murota, and M. Sugihara. Superquadratic convergence of DLASQ for computing matrix singular values. *J. Comput. Appl. Math.*, 234(4):1179–1187, 2010.

[3] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Philadelphia, PA, third edition, 1999.

[4] T. A. Arias, M. C. Payne, and J. D. Joannopoulos. Ab initio molecular-dynamics techniques extended to large-length-scale systems. *Physical Review B*, 45(4):1538–1549, 1992.

[5] Z. Bai and J. Demmel. On a block implementation of Hessenberg multishift QR iteration. *Int. J. High Speed Comput.*, 1(1):97–112, 1989.

[6] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, PA, USA, 2000.

[7] Z. Bai, J. Demmel, and M. Gu. An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems. *Numer. Math.*, 76(3):279–308, 1997.

[8] G. Ballard, J. Demmel, and I. Dumitriu. Minimizing communication for eigenproblems and the singular value decomposition. Technical Report 237, LAPACK Working Note, 2010.

[9] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in linear algebra. Technical Report 218, LAPACK Working Note, 2009.

[10] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Communication-optimal parallel and sequential Cholesky decomposition. *SIAM J. Sci. Comp*, 32(6):3495–3523, 2010.

[11] J. Barlow and J. Demmel. Computing accurate eigensystems of scaled diagonally dominant matrices. *SIAM J. Numer. Anal.*, 27(3):762–791, 1990.

[12] J. Barlow and I. Slapnicar. Optimal perturbation bounds for the Hermitian eigenvalue problem. *Linear Algebra Appl.*, 309(19–43):373–382, 2000.

[13] C. Beattie. Galerkin eigenvector approximations. *Math. Comp.*, 69:1409–1434, 2000.

[14] R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics, vol. 169. Springer, New York, 1996.

[15] R. Bhatia, C. Davis, and A. McIntosh. Perturbation of spectral subspaces and solution of linear operator equations. *Linear Algebra Appl.*, 52-53:45–67, 1983.

[16] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. Part II: Aggressive early deflation. *SIAM J. Matrix Anal. Appl.*, 23:948–973, 2002.

[17] A. Brauer. Limits for the characteristic roots of a matrix. *Duke Mathematical Journal*, 13(3):387–395, 1946.

[18] R. A. Brualdi. Matrices, eigenvalues, and directed graphs. *Linear and Multilinear Algebra*, 11:143–165, 1982.

[19] R. Byers and H. Xu. An inverse free method for the polar decomposition. *Unpublished note*, pages 1–38, 2001.

[20] R. Byers and H. Xu. A new scaling for Newton's iteration for the polar decomposition and its backward stability. *SIAM J. Matrix Anal. Appl.*, 30:822–843, 2008.

[21] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20:1956–1982, 2010.

[22] Z.-H. Cao, J.-J. Xie, and R.-C. Li. A sharp version of Kahan's theorem on clustered eigenvalues. *Linear Algebra Appl.*, 245:147–155, 1996.
[23] T. F. Chan. An improved algorithm for computing the singular value decomposition. *ACM Trans. Math. Soft.*, 8:72–83, 1982.
[24] X. Chen. On perturbation bounds of generalized eigenvalues for diagonalizable pairs. *Numerische Mathematik*, 107(1):79–86, 2007.
[25] A. K. Cline, C. B. Moler, G. W. Stewart, and J. H. Wilkinson. Estimate for the condition number of a matrix. *SIAM J. Numer. Anal.*, 16(2):368–375, 1979.
[26] S. Crudge. The QR factorization and its applications. Master's thesis, University of Manchester, 1998.
[27] J. J. M. Cuppen. A divide and conquer method for the symmetric tridiagonal eigenproblem. *Numer. Math.*, 36:177–195, 1981.
[28] L. Cvetkovic, V. Kostic, and R. S. Varga. A new Gersgorin-type eigenvalue inclusion set. *Electronic Transactions on Numerical Analysis*, 18:73–80, 2004.
[29] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.
[30] F. De Teran, F. M. Dopico, and J. Moro. First order spectral perturbation theory of square singular matrix pencils. *Linear Algebra Appl.*, 429(2-3):548–576, 2008.
[31] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, USA, 1997.
[32] J. Demmel, I. Dumitriu, and O. Holtz. Fast linear algebra is stable. *Numer. Math.*, 108(1):59–91, 2007.
[33] J. Demmel, L. Grigori, and M. Hoemmen. Implementing communication-optimal parallel and sequential QR factorizations. arXiv:0809.2407.
[34] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. Communication-avoiding parallel and sequential QR factorization. *Technical Report No. UCB/EECS-2008-74, Electrical Engineering and Computer Science, UC Berkeley*, 2008.
[35] J. Demmel and W. Kahan. Accurate singular values of bidiagonal matrices. *SIAM J. Sci. Stat. Comput.*, 11(2):873–912, 1990.
[36] J. Demmel and K. Veselic. Jacobi's method is more accurate than QR. *SIAM J. Matrix Anal. Appl.*, 13(4):1204–1245, OCT 1992.
[37] I. S. Dhillon. *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem.* PhD thesis, University of California, Berkeley, 1997.
[38] I. S. Dhillon and B. N. Parlett. Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices. *Linear Algebra Appl.*, 387:1–28, 2004.
[39] I. S. Dhillon and B. N. Parlett. Orthogonal eigenvectors and relative gaps. *SIAM J. Matrix Anal. Appl.*, 25:858–899, 2004.
[40] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.
[41] S. C. Eisenstat and I. C. F. Ipsen. Relative perturbation techniques for singular value problems. *SIAM J. Numer. Anal.*, 32:1972–1988, 1995.
[42] J. L. Fattebert. Accelerated block preconditioned gradient method for large scale wave functions calculations in density functional theory. *J. Comput. Phys.*, 229(2):441–452, 2010.
[43] J. L. Fattebert and J. Bernholc. Towards grid-based O(N) density-functional theory methods: Optimized nonorthogonal orbitals and multigrid acceleration. *Physical Review B*, 62(3):1713–1722, 2000.
[44] J. L. Fattebert and F. Gygi. Linear scaling first-principles molecular dynamics with controlled accuracy. *Computer Physics Communications*, 162(1):24–36, 2004.
[45] K. V. Fernando and B. N. Parlett. Accurate singular-values and differential qd-algorithms. *Numer. Math.*, 67(2):191–229, 1994.
[46] J. Francis. QR transformation - A unitary analog to LR transformation 1. *Computer Journal*, 4:265–271, 1961.
[47] J. Francis. The QR transformation 2. *Computer Journal*, 4(4):332–345, 1962.
[48] R. W. Freund. Quasi-kernel polynomials and their use in non-Hermitian matrix iterations. *J. Comput. Appl. Math.*, 43:135–158, 1992.

[49] R. W. Freund. Krylov-subspace methods for reduced-order modeling in circuit simulation. *J. Comput. Appl. Math.*, 123:395–421, 2000.
[50] R. W. Freund. Model reduction methods based on Krylov subspaces. *Acta Numerica*, 12:267–319, 2003.
[51] W. Gander. On Halley iteration method. *American Mathematical Monthly*, 92(2):131–134, 1985.
[52] W. Gander. Algorithms for the polar decomposition. *SIAM J. Sci. Comp*, 11(6):1102–1115, 1990.
[53] S. Geršgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk SSSR Ser. Mat. 1*, 7:749–755, 1931.
[54] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix polynomials*. SIAM, Philadelphia, USA, 2009. Unabridged republication of book first published by Academic Press in 1982.
[55] G. H. Golub and G. Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton Series in Applied Mathematics, 2009.
[56] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996.
[57] G. H. Golub and J. H. Welsch. Calculation of Gauss quadrature rules. *Math. Comp.*, 23(106):221–230, 1969.
[58] G. H. Golub and Q. Ye. New perturbation bounds for the unitary polar factor. *SIAM J. Matrix Anal. Appl.*, 16(1):327–332, 2000.
[59] J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.
[60] W. B. Gragg. The QR algorithm for unitary Hessenberg matrices. *J. Comput. Appl. Math.*, 16:1–8, 1986.
[61] S. L. Graham, M. Snir, and C. A. P. (eds.). *Getting Up to Speed: The Future of Supercomputing*. The National Academies Press, 2005.
[62] M. Gu and S. C. Eisenstat. A divide-and-conquer algorithm for the bidiagonal SVD. *SIAM J. Matrix Anal. Appl.*, 16(1):79–92, 1995.
[63] M. Gu and S. C. Eisenstat. A divide-and-conquer algorithm for the symmetrical tridiagonal eigenproblem. *SIAM J. Matrix Anal. Appl.*, 16(1):172–191, 1995.
[64] V. Gudkov. On a certain test for nonsingularity of matrices. *Latvian Math. Yearbook*, 385-390, 1965.
[65] C.-H. Guo, N. J. Higham, and F. Tisseur. Detecting and solving hyperbolic quadratic eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 30(4):1593–1613, 2009.
[66] F. Gygi. Architecture of Qbox: A scalable first-principles molecular dynamics code. *IBM Journal of Research and Development*, 52(1-2):137–144, 2008.
[67] F. Gygi, E. W. Draeger, M. Schulz, B. R. de Supinski, J. A. Gunnels, V. Austel, J. C. Sexton, F. Franchetti, S. Kral, C. W. Ueberhuber, and J. Lorenz. Large-scale electronic structure calculations of high-Z metals on the BlueGene/L platform. In *SC '06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, page 45, NY, USA, 2006. ACM.
[68] B. Hadri, H. Ltaief, E. Agullo, and J. Dongarra. Tall and skinny QR matrix factorization using tile algorithms on multicore architectures. Technical Report 222, LAPACK Working Note, 2009.
[69] W. W. Hager. Condition estimators. *SIAM Journal on scientific and statistical computing*, 5(2):311–316, 1984.
[70] S. Hakim and M. B. Fuchs. Sensitivities of multiple singular values for optimal geometries of precision structures. *International Journal of Solids and Structures*, 36:2217–2230, 1999.
[71] D. J. Higham. Condition numbers and their condition numbers. *Linear Algebra Appl.*, 214:193–213, 1995.
[72] D. J. Higham and N. J. Higham. Structured backward error and condition of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 20(2):493–512, 1998.
[73] N. J. Higham. Computing the polar decomposition - with applications. *SIAM J. Matrix Anal. Appl.*, 7(4):1160–1174, 1986.
[74] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, USA, second edition, 2002.
[75] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, PA, USA, 2008.

[76] N. J. Higham and P. Papadimitriou. Parallel singular value decomposition via the polar decomposition. Technical report. Numerical Analysis Report No. 239, Manchester Centre for Computational Mathematics, Manchester, England, 1993.
[77] N. J. Higham and P. Papadimitriou. A new parallel algorithm for computing the singular value decomposition. *In Proceedings of the Fifth SIAM Conference on Applied Linear Algebra*, pages 80–84, 1994.
[78] N. J. Higham and P. Papadimitriou. A parallel algorithm for computing the polar decomposition. *Parallel Comput.*, 20(8):1161–1173, 1994.
[79] N. J. Higham and F. Tisseur. Bounds for eigenvalues of matrix polynomials. *Linear Algebra Appl.*, 358:5–22, 2001.
[80] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
[81] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1986.
[82] I. C. F. Ipsen and B. Nadler. Refined perturbation bounds for eigenvalues of Hermitian and non-Hermitian matrices. *SIAM J. Matrix Anal. Appl.*, 31(1):40–53, 2009.
[83] Z. Jia and G. W. Stewart. An analysis of the Rayleigh-Ritz method for approximating eigenspaces. *Math. Comp.*, 70:637–647, 2001.
[84] E.-X. Jiang. Perturbation in eigenvalues of a symmetric tridiagonal matrix. *Linear Algebra Appl.*, 399:91–107, 2005.
[85] C. R. Johnson. A Gersgorin-type lower bound for the smallest singular value. *Linear Algebra Appl.*, 112:1–7, 1989.
[86] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York, second edition, 2002.
[87] C. Kenney and A. J. Laub. On scaling Newton's method for polar decomposition and the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 13(3):688–706, 1992.
[88] A. Kiełbasiński and K. Ziętak. Note on "A new scaling for Newton's iteration for the polar decomposition and its backward stability" by R. Byers and H. Xu. *SIAM J. Matrix Anal. Appl.*, 31(3):1538–1539, 2010.
[89] A. Kielbasinski and K. Zietak. Numerical behaviour of Higham's scaled method for polar decomposition. *Numerical Algorithms*, 32(2-4):105–140, 2003.
[90] A. V. Knyazev. New estimates for Ritz vectors. *Math. Comp.*, 66(219):985–995, 1997.
[91] A. V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comp*, 23(2):517–541, 2001.
[92] E. Kokiopoulou, C. Bekas, and E. Gallopoulos. Computing smallest singular triplets with implicitly restarted Lanczos bidiagonalization. *Appl. Numer. Math.*, 49(1):39–61, 2004.
[93] V. Kostic, L. J. Cvetkovic, and R. S. Varga. Gersgorin-type localizations of generalized eigenvalues. *Numer. Lin. Alg. Appl.*, 16(11), 2009.
[94] D. Kressner. The effect of aggressive early deflation on the convergence of the QR algorithm. *SIAM J. Matrix Anal. Appl.*, 30(2):805–821, 2008.
[95] D. Kressner, M. Jose Pelaez, and J. Moro. Structured Hölder condition numbers for multiple eigenvalues. *SIAM J. Matrix Anal. Appl.*, 31(1):175–201, 2009.
[96] B. Laszkiewicz and K. Zietak. Approximation of matrices and a family of Gander methods for polar decomposition. *BIT*, 46(2):345–366, 2006.
[97] C.-K. Li and R.-C. Li. A note on eigenvalues of perturbed Hermitian matrices. *Linear Algebra Appl.*, 395:183–190, 2005.
[98] C.-K. Li and R. Mathias. The Lidskii-Mirsky-Wielandt theorem - additive and multiplicative versions. *Numer. Math.*, 81(3):377–413, 1999.
[99] L. Li. A simplied Brauer's theorem on matrix eigenvalues. *Applied Mathematics*, 14, No.3:259–264, 1999.
[100] R.-C. Li. A perturbation bound for definite pencils. *Linear Algebra Appl.*, 179:191–202, 1993.
[101] R.-C. Li. Relative perturbation theory: I. Eigenvalue and singular value variations. *SIAM J. Matrix Anal. Appl.*, 19(4):956–982, 1998.

[102] R.-C. Li. On perturbations of matrix pencils with real spectra, a revisit. *Math. Comp.*, 72(242):715–728, 2003.
[103] R.-C. Li. Matrix perturbation theory. In L. Hogben, R. Brualdi, A. Greenbaum, and R. Mathias, editors, *Handbook of Linear Algebra*, chapter 15. CRC Press, Boca Raton, FL, 2006.
[104] R.-C. Li, Y. Nakatsukasa, N. Truhar, and S. Xu. Perturbation of partitioned Hermitian generalized eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 32(2):642–663, 2011.
[105] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Vector spaces of linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 28:971–1004, 2005.
[106] O. A. Marques. Private communication, 2010.
[107] O. A. Marques, C. Voemel, J. W. Demmel, and B. N. Parlett. Algorithm 880: A testing infrastructure for symmetric tridiagonal eigensolvers. *ACM Trans. Math. Softw.*, 35(1), 2008.
[108] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.
[109] R. Mathias. Quadratic residual bounds for the Hermitian eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 19(2):541–550, 1998.
[110] T. Miyata, Y. Yamamoto, and S.-L. Zhang. A fully pipelined multishift QR algorithm for parallel solution of symmetric tridiagonal eigenproblems. *IPSJ Trans. Advanced Computing Systems*, (1):14–27, 2008.
[111] R. B. Morgan. Computing interior eigenvalues of large matrices. *Linear Algebra Appl.*, 154-156:289 – 309, 1991.
[112] J. Moro, J. V. Burke, and M. L. Overton. On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM J. Matrix Anal. Appl.*, 18(4):793–817, 1997.
[113] Y. Nakatsukasa. Absolute and relative Weyl theorems for generalized eigenvalue problems. *Linear Algebra Appl.*, 432:242–248, 2010.
[114] Y. Nakatsukasa. Perturbation behavior of a multiple eigenvalue in generalized Hermitian eigenvalue problems. *BIT Numerical Mathematics*, 50(1):109–121, 2010.
[115] Y. Nakatsukasa. Gerschgorin's theorem for generalized eigenvalue problems in the Euclidean metric. *Math. Comp.*, 80(276):2127–2142, 2011.
[116] Y. Nakatsukasa. Eigenvalue perturbation bounds for Hermitian block tridiagonal matrices. *Appl. Numer. Math.*, 62(1):67 – 78, 2012.
[117] Y. Nakatsukasa. On the condition numbers of a multiple eigenvalue of a generalized eigenvalue problem. *Numer. Math.*, 121(3):531–544, 2012.
[118] Y. Nakatsukasa. The tan θ theorem with relaxed conditions. *Linear Algebra Appl.*, 436(5):1528–1534, 2012.
[119] Y. Nakatsukasa, K. Aishima, and I. Yamazaki. dqds with aggressive early deflation. *SIAM J. Matrix Anal. Appl.*, 33(1):22–51, 2012.
[120] Y. Nakatsukasa, Z. Bai, and F. Gygi. Optimizing Halley's iteration for computing the matrix polar decomposition. *SIAM J. Matrix Anal. Appl.*, 31(5):2700–2720, 2010.
[121] J. Nie. A solution to rational optimization. Private communication, 2009.
[122] S. C. Ogilvy. *Excursions in Geometry*. Dover, 1990.
[123] S. Oliveira. A new parallel chasing algorithm for transforming arrowhead matrices to tridiagonal form. *Math. Comp.*, 67(221):221–235, 1998.
[124] C. Paige and M. Wei. History and generality of the CS decomposition. *Linear Algebra Appl.*, 208-209:303 – 326, 1994.
[125] B. Parlett and C. Vömel. Detecting localization in an invariant subspace. In preparation.
[126] B. N. Parlett. Invariant subspaces for tightly clustered eigenvalues of tridiagonals. 36(3):542–562, 1996.
[127] B. N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, 1998.
[128] B. N. Parlett. A result complementary to Gersgorin's circle theorem. *Linear Algebra Appl.*, 431(1-2):20–27, 2009.
[129] B. N. Parlett. Private communication, 2010.

[130] B. N. Parlett and E. Barszcz. Another orthogonal matrix. *Linear Algebra Appl.*, 417(2-3):342 – 346, 2006.
[131] B. N. Parlett and J. Le. Forward instability of tridiagonal QR . *SIAM J. Matrix Anal. Appl.*, 14:279–316, 1993.
[132] B. N. Parlett and O. A. Marques. An implementation of the dqds algorithm (positive case). *Linear Algebra Appl.*, 309(1-3):217–259, 2000.
[133] M. Petschow. Private communication, 2010.
[134] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, Philadelphia, PA, USA, second edition, 2011.
[135] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical methods for electronic structure calculations of materials. *SIAM Rev.*, 52(1):3–54, 2010.
[136] G. L. G. Sleijpen, J. van den Eshof, and P. Smit. Optimal a priori error bounds for the Rayleigh-Ritz method. *Math. Comput.*, 72(242):677–684, 2003.
[137] G. L. G. Sleijpen and H. A. Van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Rev.*, 42(2):267–293, 2000.
[138] F. Song, A. YarKhan, and J. Dongarra. Dynamic task scheduling for linear algebra algorithms on distributed-memory multicore systems. Technical Report 221, LAPACK Working Note, 2009.
[139] D. Sorensen. Deflation for implicitly restarted Arnoldi methods. Technical Report 98-12, Rice University, CAAM, 1998.
[140] G. W. Stewart. Gershgorin theory for the generalized eigenvalue problem $Ax = \lambda Bx$. *Math. Comput*, 29(130):600–606, 1975.
[141] G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Rev.*, 19(4):634–662, 1977.
[142] G. W. Stewart. *Matrix Algorithms Volume I: Basic decompositions*. SIAM, Philadelphia, 1998.
[143] G. W. Stewart. A generalization of Saad's theorem on Rayleigh-Ritz approximations. *Linear Algebra Appl.*, 327:115–119, 1999.
[144] G. W. Stewart. A Krylov-Schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.*, 23:601–614, 2001.
[145] G. W. Stewart. *Matrix Algorithms Volume II: Eigensystems*. SIAM, Philadelphia, 2001.
[146] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory (Computer Science and Scientific Computing)*. Academic Press, 1990.
[147] G. W. Stewart and G. Zhang. Eigenvalues of graded matrices and the condition numbers of a multiple eigenvalue. *Numer. Math.*, 58(7):703–712, 1991.
[148] J. F. Sturm. SeDuMi 1.02, A MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11&12:625–653, 1999.
[149] J.-G. Sun. Eigenvalues of Rayleigh quotient matrices. *Numer. Math.*, 59:603–614, 1991.
[150] J.-G. Sun. On condition numbers of a nondefective multiple eigenvalue. *Numer. Math.*, 61(2):265–275, 1992.
[151] J.-G. Sun. On worst-case condition numbers of a nondefective multiple eigenvalue. *Numer. Math.*, 69(3):373–382, 1995.
[152] The NAG Toolbox for MATLAB, http://www.nag.co.uk/numeric/MB/start.asp.
[153] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43:235–286, 2001.
[154] L. N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, Philadelphia, 2013.
[155] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.
[156] L. N. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.
[157] R. A. Van De Geijn. Deferred shifting schemes for parallel QR methods. *SIAM J. Matrix Anal. Appl.*, 14(1):180–194, 1993.
[158] R. S. Varga. *Matrix Iterative Analysis*. Springer-Verlag, 2000.
[159] R. S. Varga. Gerschgorin disks, Brauer ovals of Cassini (a vindication), and Brualdi sets. *Information, 14*, 14(2):171–178, 2001.

[160] R. S. Varga. *Geršgorin and His Circles*. Springer-Verlag, 2004.
[161] K. Veselic and I. Slapnicar. Floating-point perturbations of Hermitian matrices. *Linear Algebra Appl.*, 195:81 – 116, 1993.
[162] J. L. Walsh. The existence of rational functions of best approximation. *Trans. AMS*, 33:668–689, 1931.
[163] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.
[164] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen. *Math. Ann.*, 71:441–479, 1912.
[165] J. H. Wilkinson. *The Algebraic Eigenvalue Problem (Numerical Mathematics and Scientific Computation)*. Oxford University Press, USA, 1965.
[166] P. Willems. *On MR^3-type Algorithms for the Tridiagonal Symmetric Eigenproblem and the Bidiagonal SVD*. PhD thesis, University of Wuppertal, 2010.
[167] C. Yang, W. Gao, Z. Bai, X. Li, L. Lee, P. Husbands, and E. Ng. An algebraic substructuring method for large-scale eigenvalue calculation. *SIAM J. Sci. Comp*, 27(3):873–892, 2005.
[168] Q. Ye. On close eigenvalues of tridiagonal matrices. *Numer. Math.*, 70:507–514, 1995.
[169] H. Zha. A two-way chasing scheme for reducing a symmetric arrowhead matrix to tridiagonal form. *J. Numerical Linear Algebra*, 1:494–499, 1993.
[170] H. Zha and Z. Zhang. A cubically convergent parallelizable method for the Hermitian eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 19(2):468–486, 1998.
[171] Z. Zhang, H. Zha, and W. Ying. Fast parallelizable methods for computing invariant subspaces of Hermitian matrices. *Journal of Computational Mathematics*, 25(5):583–594, 2007.