Stability and Accuracy

Perturbation Analysis of Algebraic Eigenproblems

(Revised Version)

Ji-guang Sun

Umeå University
Department of Computing Science
S-901 87 Umeå, Sweden

Contents

Preface

This work covers perturbation analysis of the following three kinds of algebraic eigenproblems: the eigenvalue problem

$$Ax = \lambda x, \tag{I}$$

the singular value decomposition

$$A = U\Sigma V^H, \tag{II}$$

and the generalized eigenvalue problem

$$\beta Ax = \alpha Bx, \tag{III}$$

where the matrices A and B are data.

Perturbation analysis of the algebraic eigenproblems contains forward perturbation analysis and backward perturbation analysis.

Forward perturbation analysis is motivated by the fact that any one of the problems (I), (II) and (III), as the others in matrix computations, is usually subject to perturbations on the data reflecting various errors in the formulation of the problem and in its solution by a computer. When solving a computation problem, it may well be asked: How does a solution change when the data are subject to perturbations?

The result of a forward perturbation analysis may be a perturbation expansion, or a condition number, or a perturbation bound. A perturbation expansion approximates the perturbation in the solution in terms of a known perturbation on the data. A condition number is a measure of the sensitivity of the solution to perturbations on the data. A perturbation bound is used to bound the resulting perturbation in the solution.

Backward perturbation analysis is motivated by the following fact. Let an approximate solution to a computation problem be given. For example, the approximate solution may come from a numerical algorithm for approximating the exact solution. Then there are two important questions associated with the approximate solution: (1) Is the approximate solution the exact solution of a slightly perturbed problem? (2) Is the approximate solution close to the exact solution?

To answer the question (1) we need the notion of backward error of a problem with respect to an approximate solution. In general, an approximate solution of a problem solves many perturbed problems. The backward error of the problem with respect to the approximate solution is a measure of the nearness between the perturbed problems and the original problem. A small backward error means that the approximate solution is the exact solution of a slightly perturbed problem. Consequently, to find a computable formula of the backward error may be very useful for testing the backward stability of practical algorithms.

When computable formulas of the backward error and the associated optimal (minimum) backward perturbation on the data are found, an answer to the question (2) will be obtained by applying an appropriate forward perturbation result to the optimal (minimum) backward perturbation on the data. Generally speaking, the optimal (minimum) backward perturbation is expressed through some residual of the problem with respect to the approximate solution, and so the obtained estimate concerning the accuracy of the approximate solution is usually in the form of residual bound.

The present work is a sequel to the books *Matrix Perturbation Analysis* [104, 1987 and 2001] and *Matrix Perturbation Theory* [97, 1990]. In the two previously published books, we were chiefly concerned with perturbation bounds for linear systems, least squares, eigenvalue problems, the singular value decomposition, and generalized eigenvalue problems. The main object of this work is to describe techniques for deriving perturbation expansions, condition numbers, backward errors, and residual bounds for the problems (I), (II) and (III).

I hope that this work will be useful to graduate students in technical areas and my colleagues in numerical analysis, and also to all computational scientists and engineers who are concerned about the stability and accuracy of their results.

The first chapter of the work reviews and collects necessary background material from matrix algebra and analysis. Chapters 2, 3 and 4 are devoted to the problems (I), (II) and (III), separately. We have supplemented each section with a set of "Notes and References" in which literature citations are given, and other related results are discussed. Besides, a certain number of simple numerical examples are used to illustrate some theoretical results. All computations were performed using MATLAB, version 4.2c. The relative machine precision is 2.2204×10^{-16}.

This work has greatly benefited from the insight and knowledge provided by many friends and colleagues. In particular, the work of Pete Stewart and Nick Higham has strongly influenced my research in perturbation analysis of algebraic eigenproblems.

I would like to thank Bo Kågström for providing a fine working environment necessary for such work. I wish to thank Zhaojun Bai and Anders Barrlund for reading the text and valuable suggestions. Thanks also to my colleagues at the Department of Computing Science of Umeå University for providing a nice atmosphere and good condition. I am indebted to the Swedish National Science Research Council and the Faculty of Science and Technology of Umeå University for their support.

Ji-guang Sun
Umeå

Chapter 1

Preliminaries

This chapter contains necessary background material for the chapters that follow.

The first section introduces some notation. §1.2 – §1.5 are devoted to norms, metrics, matrix orthogonal decompositions, and solutions of some matrix equations.

The implicit function theorem, the Brouwer fixed point theorem and the Schauder fixed point theorem, are cited in §1.6 and §1.7, respectively.

In §1.8 and §1.9 we introduce definitions of normwise condition numbers and normwise backward errors.

1.1 Notation

Throughout this work we shall use the following notational conventions.

The symbol $\mathcal{C}^{m\times n}$ ($\mathcal{R}^{m\times n}$) will denote the set of $m\times n$ complex (real) matrices, $\mathcal{C}^n = \mathcal{C}^{n\times 1}$, $\mathcal{R}^n = \mathcal{R}^{n\times 1}$, $\mathcal{C} = \mathcal{C}^1$, and $\mathcal{R} = \mathcal{R}^1$. As usual $\emptyset$ is the empty set.

The transpose of a matrix A will be written A^T, the conjugat $\overline{A}$, and $A^H = \overline{A}^T$. The trace of a square matrix A will be written $\mathrm{tr}(A)$. The symbol $|A|$ will denote the matrix $(|\alpha_{ij}|)$ for $A = (\alpha_{ij})$. The identity matrix will be wriiten I, e_j is the jth column of I, and $e_j^{(n)}$ stands for the jth column vector of I_n, the identity matrix of order n. The null matrix will be written 0.

The set of $n\times n$ *Hermitian* (*real symmetric*) matrices will be written $\mathcal{H}^{n\times n}$ ($\mathcal{S}^{n\times n}$), and the set of $m\times n$ *unitary* (*real orthogonal*) matrices will be written $\mathcal{U}^{m\times n}$ ($\mathcal{O}^{m\times n}$); i.e.,

$$\mathcal{H}^{n\times n} = \{A \in \mathcal{C}^{n\times n} \ : \ A^H = A\}, \qquad \mathcal{S}^{n\times n} = \{A \in \mathcal{R}^{n\times n} \ : \ A^T = A\},$$

$$\mathcal{U}^{m\times n} = \{A \in \mathcal{C}^{m\times n} \ : \ A^H A = I\}, \qquad \mathcal{O}^{m\times n} = \{A \in \mathcal{R}^{m\times n} \ : \ A^T A = I\}.$$

The positive definiteness (or semi-definiteness) of $A \in \mathcal{H}^{n\times n}$ (or $\mathcal{S}^{n\times n}$) will be denoted by $A > 0$ (or $A \geq 0$).

For $A \in \mathcal{C}^{m\times n}$, $A^\dagger$ stands for the *Moore-Penrose inverse* of A, and $\mathcal{R}(A)$ the column space of A, i.e., $\mathcal{R}(A) = \{Ax \ : \ x \in \mathcal{C}^n\}$. The orthogonal projection onto the subspace $\mathcal{R}(A)$ will be written $P_{\mathcal{R}(A)}$ (or simply, P_A), and $P_A^\perp = I - P_A$. For a subspace $\mathcal{X}$, $\dim(\mathcal{X})$ will denote the dimension of $\mathcal{X}$.

The set of all eigenvalues of A will be wriiten $\lambda(A)$, the set of all singular values of A will be written $\sigma(A)$, and $\sigma_+(A)$ will denote the set of all positive singular values of A. The largest (smallest) singular value of A will be written $\sigma_{\max}(A)$ ($\sigma_{\min}(A)$).

For $A = (\alpha_{jk}) = (a_1, \ldots, a_n) \in \mathcal{C}^{m\times n}$ and a matrix B, $A \otimes B = (\alpha_{jk}B)$ is a *Kronecker product*, and $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A) = (a_1^T, \ldots, a_n^T)^T$. For basic properties of the Kronecker product and vec operator, see Graham [42, Chapters 1 and 2], or Horn and Johnson [56, Chapter 4], or Lancaster and Tismenelsky [67, Chapter 12].

Throughout this work, the symbol $\|\cdot\|$ will be used to denote any *unitarily invariant norm* (see Section 1.2.3) if there is no a special statement.

For linear spaces $\mathcal{A}$ and $\mathcal{B}$, the *product space* $\mathcal{A} \times \mathcal{B}$ is defined by

$$\mathcal{A} \times \mathcal{B} = \{(a, b) \ : \ a \in \mathcal{A}, \ b \in \mathcal{B}\}.$$

The relation "$\equiv$" is used for implicit definitions.

1.2 Norms

Most of the perturbation results presented in the following chapters are on normwise. Therefore, norms have an important role to play in our work. In this section we collect certain basic notion and facts on vector norms and matrix norms.

1.2.1 Vector Norms

A vector norm is a generalization of the modulus of a complex number.

A function $\nu : \mathcal{C}^n \to \mathcal{R}$ is a *norm* on $\mathcal{C}^n$ if ν satisfies the following conditions:

1. $x \neq 0 \implies \nu(x) > 0$,

2. $\nu(\alpha x) = |\alpha|\nu(x)$ for any $\alpha \in \mathcal{C}$,

3. $\nu(x + y) \leq \nu(x) + \nu(y)$.

For any $x = (\xi_1, \ldots, \xi_n)^T \in \mathcal{C}^n$, the p-norm $\|x\|_p$ is defined by

$$\|x\|_p = (|\xi_1|^p + \cdots + |\xi_n|^p)^{\frac{1}{p}}, \quad p \geq 1.$$

The most useful p-norms are the 1-norm, 2-norm (or the Euclidean norm), and the ∞-norm:

$$\|x\|_1 = \sum_{j=1}^{n} |\xi_j|, \qquad \|x\|_2 = \sqrt{x^H x}, \qquad \|x\|_\infty = \max_{1 \leq j \leq n} |\xi_j|.$$

A vector norm ν on $\mathcal{C}^n$ is *absolute* if $\nu(|x|) = \nu(x)$ for all $x \in \mathcal{C}^n$, where $|x|$ denotes the vector whose elements are the absolute values of the elements of x. Any p-norm is obviously an absolute norm.

It is known (see, e.g., Stewart and Sun [97, Chapter II, Theorem 1.3]) that a vector norm ν is absolute if and only if

$$|x| \leq |y| \implies \nu(x) \leq \nu(y).$$

For any vector norm $\nu(\cdot)$ on $\mathcal{C}^n$, the *dual* norm $\nu^D(\cdot)$ is defined by

$$\nu^D(y) = \max_{\nu(x)=1} |y^H x|, \quad y \in \mathcal{C}^n.$$

It can be verified that for any vector x we have

$$\|x\|_2^D = \|x\|_2, \qquad \|x\|_1^D = \|x\|_\infty, \qquad \|x\|_\infty^D = \|x\|_1.$$

1.2.2 Matrix Norms

A function $\nu : \mathcal{C}^{m \times n} \to \mathcal{R}$ is a *norm* on $\mathcal{C}^{m \times n}$ if ν satisfies the following conditions:

1. $A \neq 0 \implies \nu(A) > 0$,

2. $\nu(\alpha A) = |\alpha| \nu(A)$ for any $\alpha \in \mathcal{C}$,

3. $\nu(A + B) \leq \nu(A) + \nu(B)$.

Let ν_1, ν_2 and ν_3 be norms on $\mathcal{C}^{m \times n}$, $\mathcal{C}^{n \times k}$ and $\mathcal{C}^{m \times k}$, respectively. Then ν_1, ν_2 and ν_3 are *mutually consistent* if

$$\nu_3(AB) \leq \nu_1(A)\nu_2(B)$$

whenever $A \in \mathcal{C}^{m \times n}$ and $B \in \mathcal{C}^{n \times k}$. In particular, a matrix norm ν on $\mathcal{C}^{n \times n}$ is *consistent* if

$$\nu(AB) \leq \nu(A)\nu(B)$$

for all $A, B \in \mathcal{C}^{n \times n}$.

The most frequently used matrix norms in matrix perturbation analysis are the Frobenius norm $\|\cdot\|_F$ and the p-norm $\|\cdot\|_p$. For $A = (\alpha_{jk}) \in \mathcal{C}^{m\times n}$, the norms $\|A\|_F$ and $\|A\|_p$ are defined by

$$\|A\|_F = \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{n} |\alpha_{jk}|^2},$$

and

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \qquad p \geq 1.$$

Note that the Frobenius norm and any p-norm are consistent norms.

The most useful p-norms are the 1-norm, 2-norm (i.e., the spectral norm), and the ∞-norm.

If $\sigma_1, \ldots, \sigma_n$ are the singular values of $A \in \mathcal{C}^{m\times n}$, i.e., if $\sigma_1, \ldots, \sigma_n$ are nonnegative scalars and $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of $A^H A$, then the norms $\|A\|_F$ and $\|A\|_2$ can be expressed by

$$\|A\|_F = \sqrt{\sum_{j=1}^{n} \sigma_j^2}, \qquad \|A\|_2 = \sigma_{\max}(A).$$

1.2.3 Unitarily Invariant Norms

For any $A \in \mathcal{C}^{m\times n}$, $U \in \mathcal{U}^{m\times m}$ and $V \in \mathcal{U}^{n\times n}$, we have

$$\|UAV\|_F = \|A\|_F, \qquad \|UAV\|_2 = \|A\|_2,$$

and

$$\|A\|_F = \|A\|_2 \quad \text{if } \operatorname{rank}(A) = 1.$$

These facts suggest the following definition.

A norm $\|\cdot\|$ on $\mathcal{C}^{m\times n}$ is called a *unitarily invariant* norm if it satisfies

4. $\|U^H AV\| = \|A\|$ for any $U \in \mathcal{U}^{m\times m}$ and $V \in \mathcal{U}^{n\times n}$,

5. $\|A\| = \|A\|_2$ if $\operatorname{rank}(A) = 1$.

Note that any unitarily invariant norm is a consistent norm.

By the von Neumann theorem [125] (or see Stewart and Sun [97, Chapter II, Theorem 3.6]), any unitarily invariant norm can be characterized as a symmetric gauge function of singular values.

A function $\phi : \mathcal{R}^n \to \mathcal{R}$ is called a *symmetric gauge function* if it satisfies the following five properties:

$$1.\ x \neq 0 \implies \phi(x) > 0,$$

$$2.\ \phi(\gamma x) = |\gamma|\phi(x) \ \text{ for any } \gamma \in \mathcal{R},$$

$$3.\ \phi(x+y) \leq \phi(x) + \phi(y),$$

$$4.\ \phi(P|x|) = \phi(x) \ \text{ for any permutation matrix } P,$$

$$5.\ \phi\left(e_1^{(n)}\right) = 1.$$

Suppose that ϕ is a symmetric gauge function on $\mathcal{R}^N$, where N is a sufficiently large natural number. Then for any $m, n \leq N$ we may define a unitarily invariant norm $\|\cdot\|$ on $\mathcal{C}^{m\times n}$ $(m, n \leq N)$ by

$$\|A\| = \phi(\sigma_1, \ldots, \sigma_n, 0, \ldots, 0),$$

where $\sigma_1, \ldots, \sigma_n$ are the singular values of A. Consequently, we obtain a family of unitarily invariant norms on $\bigcup\limits_{m,n\leq N} \mathcal{C}^{m\times n}$ generated by ϕ. For simplicity, the symbol $\|\cdot\|$ will also be used to denote a family of unitarily invariant norms generated by any symmetric gauge function.

The following properties possessed by unitarily invariant norms are well known:

$$\|A^H\| = \|A\|,$$

$$\sigma_+(A_1) = \sigma_+(A_2) \implies \|A_1\| = \|A_2\|,$$

$$\|A_2\| = \left\|\begin{pmatrix} 0 \\ A_2 \end{pmatrix}\right\| \leq \left\|\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}\right\|,$$

$$\|AB\| \leq \|A\|_2\|B\|, \quad \|AB\| \leq \|B\|_2\|A\|.$$

Besides, if the singular values of $A, B \in \mathcal{C}^{m\times n}$ are $\sigma_1 \geq \cdots \geq \sigma_n$ and $\tau_1 \geq \cdots \geq \tau_n$, respectively, and $\sigma_j \leq \tau_j$ for $j = 1, \ldots, n$, then $\|A\| \leq \|B\|$.

1.2.4 Some Results on Matrix Norms

The following results on matrix norms will be used in chapters 2–4.

Theorem 1.2.1. *Let* $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ *be a partitioned matrix, and let*

$$B = \begin{pmatrix} 0 & A_{12} \\ A_{21} & 0 \end{pmatrix}, \quad C = \begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \quad A_1 = \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix}.$$

Then

$$\|B\| \leq \|A\|, \quad \|C\| \leq \|A\|, \quad \|D\| \leq \|A\|, \quad \|A_1\| \leq \|A\|.$$

Proof. Let

$$Q = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

Then from

$$B = \frac{1}{2}(A - QAQ), \quad D = \frac{1}{2}(A + QAQ), \quad C = \frac{1}{2}(D + QD)$$

we get

$$\|B\| \leq \|A\|, \quad \|C\| \leq \|D\| \leq \|A\|.$$

Moreover, from

$$A_1 = A \begin{pmatrix} I \\ 0 \end{pmatrix}$$

we get

$$\|A_1\| \leq \|A\| \left\| \begin{pmatrix} I \\ 0 \end{pmatrix} \right\|_2 \leq \|A\|. \qquad \square$$

Theorem 1.2.2. *Let $B \in \mathcal{C}^{m\times m}, C \in \mathcal{C}^{n\times n}$ $(m \geq n)$ be normal matrices, and $\Gamma = \mathrm{diag}(\gamma_i)$ with $\gamma_1 \geq \cdots \geq \gamma_n \geq 0$. Then*

$$\left\| B \begin{pmatrix} \Gamma \\ 0 \end{pmatrix} - \begin{pmatrix} \Gamma \\ 0 \end{pmatrix} C \right\|_F \geq \gamma_n \left\| B \begin{pmatrix} I^{(n)} \\ 0 \end{pmatrix} - \begin{pmatrix} I^{(n)} \\ 0 \end{pmatrix} C \right\|_F. \qquad (1.2.1)$$

Proof. We first prove the inequality (1.2.1) for $m = n$. Let

$$\delta = \|B\Gamma - \Gamma C\|_F^2 - \gamma_n^2 \|B - C\|_F^2$$

and

$$\Omega = \Gamma - \gamma_n I.$$

Obviously, the diagonal elements of Ω are nonnegative. Moreover,

$$\begin{aligned} \delta &= \|B\Omega - \Omega C + \gamma_n(B - C)\|_F^2 - \gamma_n^2\|B - C\|_F^2 \\ &= \|B\Omega - \Omega C\|_F^2 + 2\gamma_n \mathrm{Re}\left(\mathrm{tr}\left[(B\Omega - \Omega C)^H (B - C)\right]\right) \\ &= \|B\Omega - \Omega C\|_F^2 \\ &\quad +\gamma_n \mathrm{tr}\left(\Omega\left[(B - C)^H(B - C) + (B - C)(B - C)^H\right]\right) \\ &\geq \|B\Omega - \Omega C\|_F^2 \geq 0, \end{aligned}$$

which shows

$$\|B\Gamma - \Gamma C\|_F \geq \gamma_n \|B - C\|_F. \qquad (1.2.2)$$

We now prove the inequality (1.2.1) for $m > n$. Let

$$\hat{\Gamma} = \begin{pmatrix} \Gamma & 0 \\ 0 & \gamma_n I_{m-n} \end{pmatrix}, \quad \hat{C} = \begin{pmatrix} C & 0 \\ 0 & I_{m-n} \end{pmatrix}.$$

Then we have

$$\begin{aligned} \left\| B\hat{\Gamma} - \hat{\Gamma}\hat{C} \right\|_F^2 &= \left\| \left(B \begin{pmatrix} \Gamma \\ 0 \end{pmatrix} - \begin{pmatrix} \Gamma \\ 0 \end{pmatrix} C,\ B \begin{pmatrix} 0 \\ \gamma_n I_{m-n} \end{pmatrix} - \begin{pmatrix} 0 \\ \gamma_n I_{m-n} \end{pmatrix} \right) \right\|_F^2 \\ &= \left\| B \begin{pmatrix} \Gamma \\ 0 \end{pmatrix} - \begin{pmatrix} \Gamma \\ 0 \end{pmatrix} C \right\|_F^2 + \gamma_n^2 \left\| B \begin{pmatrix} 0 \\ I_{m-n} \end{pmatrix} - \begin{pmatrix} 0 \\ I_{m-n} \end{pmatrix} \right\|_F^2, \end{aligned} \tag{1.2.3}$$

and

$$\begin{aligned} \gamma_n^2 \left\| B - \hat{C} \right\|_F^2 &= \gamma_n^2 \left\| B \begin{pmatrix} I_n & 0 \\ 0 & I_{m-n} \end{pmatrix} - \begin{pmatrix} C & 0 \\ 0 & I_{m-n} \end{pmatrix} \right\|_F^2 \\ &= \gamma_n^2 \left\| B \begin{pmatrix} I_n \\ 0 \end{pmatrix} - \begin{pmatrix} I_n \\ 0 \end{pmatrix} C \right\|_F^2 + \gamma_n^2 \left\| B \begin{pmatrix} 0 \\ I_{m-n} \end{pmatrix} - \begin{pmatrix} 0 \\ I_{m-n} \end{pmatrix} \right\|_F^2. \end{aligned} \tag{1.2.4}$$

By (1.2.2),

$$\left\| B\hat{\Gamma} - \hat{\Gamma}\hat{C} \right\|_F \geq \gamma_n \left\| B - \hat{C} \right\|_F.$$

Combining it with (1.2.3) and (1.2.4) shows the inequality (1.2.1). □

We now cite two famous results on norm-preserving dilations.

Theorem 1.2.3 (Kreǐn and Kahan). *Let*

$$\Phi(W) = \begin{pmatrix} A & C^H \\ C & W \end{pmatrix}$$

with $A \in \mathcal{H}^{k \times k}$ and $C \in \mathcal{C}^{l \times k}$. Then

$$\min_{W \in \mathcal{H}^{l \times l}} \|\Phi(W)\|_2 = \left\| \begin{pmatrix} A \\ C \end{pmatrix} \right\|_2.$$

Theorem 1.2.4 (Kahan, Weinberger, Davis, and Parrot). *Let*

$$\Psi(Z_{22}) = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}$$

with $Z_{11} \in \mathcal{C}^{k\times k}$ *and* $Z_{21}, Z_{12}^T \in \mathcal{C}^{l\times k}$. *Then*

$$\min_{Z_{22}\in\mathcal{C}^{l\times l}} \|\Psi(Z_{22})\|_2 = \max\left\{\left\|\begin{pmatrix} Z_{11} \\ Z_{21}\end{pmatrix}\right\|_2, \ \|(Z_{11}, Z_{12})\|_2\right\}.$$

Note that Theorems 1.2.3 and 1.2.4 are valid for real matrices.

Notes and References

NR 1.2–1. There is a large literature on vector norms and matrix norms. For deeper issues concerning norms, see Householder [57, Chapter 2], and Horn and Johnson [55, Chapter 5]; for unitarily invariant norms, see von Neumann [125] and Mirsky [78]; for historical comments on the development of norms in numerical analysis, see Stewart and Sun [97, Chapter II].

NR 1.2–2. The inequality (1.2.2) is proved by Sun [100]. Theorem 1.2.2 is proved by Chen and Sun [20].

NR 1.2–3. Theorem 1.2.3, as an important dilation theorem, is discovered by Kreǐn [64] and Kahan [61] (see Parlett [83, 231–233]). Recently, Elsner, He and Mehrmann [36] give a proof in a different way.

NR 1.2–4. Theorem 1.2.4 is a general dilation theorem. The first proof is given by Kahan and Weinberger; later proofs are given by Davis [25] and Parrott [84] (see Davis, Kahan and Weinberger [27]).

1.3 Metrics on Subspaces of $\mathcal{C}^n$

In some applications, the object that is perturbed is not a vector or a matrix, but a subspace, for example, an invariant subspace of a matrix. In this section we shall discuss measures of metrics on subspaces.

The symbol $\mathcal{G}_l^n$ will be used to denote the set of l-dimensional subspaces of $\mathcal{C}^n$. We shall use $\mathcal{X}_1, \mathcal{Y}_1, \mathcal{Z}_1$ for subspaces.

1.3.1 Unitarily Invariant Metrics

A function $d(\cdot,\cdot) : \mathcal{G}_l^n \to \mathcal{R}$ is a *metric* on $\mathcal{G}_l^n$ if it satisfies the following conditions:

1. $d(\mathcal{X}_1,\mathcal{Y}_1) \geq 0, \ \text{ and } \ d(\mathcal{X}_1,\mathcal{Y}_1) = 0 \iff \mathcal{X}_1 = \mathcal{Y}_1,$

2. $d(\mathcal{X}_1,\mathcal{Y}_1) = d(\mathcal{Y}_1,\mathcal{X}_1),$

3. $d(\mathcal{X}_1,\mathcal{Y}_1) \leq d(\mathcal{X}_1,\mathcal{Z}_1) + d(\mathcal{Z}_1,\mathcal{Y}_1).$

A metric $d(\cdot,\cdot)$ on $\mathcal{G}_l^n$ is *unitarily invariant* if it satisfies

4. $d(U\mathcal{X}_1,U\mathcal{Y}_1) = d(\mathcal{X}_1,\mathcal{Y}_1) \quad \text{for any } U \in \mathcal{U}^{n\times n}.$

From the definition it follows that for any unitarily invariant norm $\|\cdot\|$ on $\mathcal{C}^{n\times n}$, $\|P_{\mathcal{X}_1} - P_{\mathcal{Y}_1}\|$ is a unitarily invariant metric on $\mathcal{G}_l^n$.

Let $\mathcal{X}_1$ and $\mathcal{Y}_1$ be l-dimensional subspaces. Take $X_1, Y_1 \in \mathcal{U}^{n\times l}$ such that $\mathcal{R}(X_1) = \mathcal{X}_1, \mathcal{R}(Y_1) = \mathcal{Y}_1$. Define $\Theta(X_1,Y_1) \in \mathcal{H}^{l\times l}$ by

$$\Theta(X_1,Y_1) = \arccos(X_1^H Y_1 Y_1^H X_1)^{\frac{1}{2}} \geq 0. \tag{1.3.1}$$

Then we have the following result.

Theorem 1.3.1. *For any unitarily invariant norm $\|\cdot\|$ on $\mathcal{C}^{l\times l}$, there exists a unitarily invariant norm $\|\cdot\|^*$ on $\mathcal{C}^{n\times n}$ such that*

$$\|P_{X_1} - P_{Y_1}\|^* = \|\sin\Theta(X_1,Y_1)\|.$$

Conversely, for any unitarily invariant norm $\|\cdot\|^*$ on $\mathcal{C}^{n\times n}$, there exists a unitarily invariant norm $\|\cdot\|$ on $\mathcal{C}^{l\times l}$ such that

$$\|\sin\Theta(X_1,Y_1)\| = \|P_{X_1} - P_{Y_1}\|^*.$$

Theorem 1.3.1 shows that for any unitarily invariant norm $\|\cdot\|$ on $\mathcal{C}^{l\times l}$, the quantity $\rho(\mathcal{X}_1,\mathcal{Y}_1)$ defined by

$$\rho(\mathcal{X}_1,\mathcal{Y}_1) = \|\sin\Theta(X_1,Y_1)\| \tag{1.3.2}$$

is a unitarily invariant metric on $\mathcal{G}_l^n$. Particularly, we have

$$\begin{aligned} \rho_2(\mathcal{X}_1,\mathcal{Y}_1) &\equiv \|\sin\Theta(X_1,Y_1)\|_2 = \|P_{X_1} - P_{Y_1}\|_2, \\ \rho_F(\mathcal{X}_1,\mathcal{Y}_1) &\equiv \|\sin\Theta(X_1,Y_1)\|_F = \frac{1}{\sqrt{2}}\|P_{X_1} - P_{Y_1}\|_F. \end{aligned} \tag{1.3.3}$$

We now consider the simplest case: $n = 2$ and $l = 1$. Let $x_1 = (\alpha, \beta)^T$ and $y_1 = (\gamma, \delta)^T$ be nonzero vectors of $\mathcal{C}^2$. Then by (1.3.1) and (1.3.2) we have

$$\begin{aligned} \rho(\mathcal{R}(x_1), \mathcal{R}(y_1)) &= \frac{|\alpha\delta - \beta\gamma|}{\sqrt{(|\alpha|^2 + |\beta|^2)(|\gamma|^2 + |\delta|^2)}} \\ &\equiv \rho((\alpha, \beta), (\gamma, \delta)), \end{aligned} \tag{1.3.4}$$

which is the *chordal metric* on the *complex projective plane* (or the chordal metric on the *Riemann sphere*). Hence, the metric $\rho(\mathcal{X}_1, \mathcal{Y}_1)$ defined by (1.3.2) is usually called the *generalized chordal metric*.

Let $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ be l-dimensional subspaces of $\mathcal{C}^n$, where $X_1, Y_1 \in \mathcal{U}^{n \times l}$. By Stewart [91, Appendix] (or see Stewart and Sun [97, Chapter 1, Theorem 5.2]), there are unitary matrices Q, U_1 and V_1 such that

$$QX_1U_1 = \begin{pmatrix} I_l \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad QY_1V_1 = \begin{pmatrix} \Gamma \\ \Sigma \\ 0 \end{pmatrix} \qquad (\text{when } 2l \leq n)$$

or

$$QX_1U_1 = \begin{pmatrix} I_{n-l} & 0 \\ 0 & I_{2l-n} \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad QY_1V_1 = \begin{pmatrix} \Gamma & 0 \\ 0 & I_{2l-n} \\ \Sigma & 0 \end{pmatrix} \qquad (\text{when } 2l > n),$$

where

$$\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_{n_1}), \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n_1}),$$

$$0 \leq \gamma_1 \leq \cdots \leq \gamma_{n_1}, \quad \sigma_1 \geq \cdots \geq \sigma_{n_1} \geq 0,$$

$$\gamma_j^2 + \sigma_j^2 = 1, \quad j = 1, \ldots, n_1,$$

in which

$$n_1 = \begin{cases} l & \text{if } 2l \leq n, \\ n - l & \text{otherwise.} \end{cases} \tag{1.3.5}$$

The angles $\theta_j \equiv \sin^{-1} \sigma_j \in [0, \pi/2]$ $(j = 1, \ldots, n_1)$ are called the *canonical angles* between $\mathcal{X}_1$ and $\mathcal{Y}_1$.

Using the canonical angles, the metric $\rho(\mathcal{X}_1, \mathcal{Y}_1)$ can be expressed by

$$\rho(\mathcal{X}_1, \mathcal{Y}_1) = \|\operatorname{diag}(\sin \theta_j)\|. \tag{1.3.6}$$

particularly, we have

$$\rho_2(\mathcal{X}_1, \mathcal{Y}_1) = \sin \theta_1, \qquad \rho_F(\mathcal{X}_1, \mathcal{Y}_1) = \sqrt{\sum_j \sin^2 \theta_j}. \tag{1.3.7}$$

1.3.2 Some Estimates of Metrics

The following result reveals a relation between $\rho(\mathcal{X}_1, \mathcal{Y}_1)$ and $Y_1 - X_1$, where $\mathcal{X}_1 = \mathcal{R}(X_1)$, and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$.

Theorem 1.3.2. *Let $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$, where $X_1, Y_1 \in \mathcal{C}^{n\times l}$, and* $\operatorname{rank}(X_1) = \operatorname{rank}(Y_1) = l$. *Then*

$$\begin{aligned} \rho(\mathcal{X}_1, \mathcal{Y}_1) &= \left\| P_{X_1}^{\perp}(Y_1 - X_1)(Y_1^H Y_1)^{-1/2} \right\| \\ &= \left\| P_{Y_1}^{\perp}(Y_1 - X_1)(X_1^H X_1)^{-1/2} \right\|. \end{aligned} \tag{1.3.8}$$

Theorem 1.3.2 implies that for nonzero vectors $x, y \in \mathcal{C}^n$, we have

$$\sin\theta(x,y) = \sin\theta(u,v) \le \min\left\{ \frac{\|y-x\|_2}{\|y\|_2}, \frac{\|y-x\|_2}{\|x\|_2} \right\}, \tag{1.3.9}$$

where $u = x/\|x\|_2, v = y/\|y\|_2$, and $\theta(u,v)$ denotes the angle between the one-dimensional subspaces $\mathcal{R}(x)$ and $\mathcal{R}(y)$.

Proof of Theorem 1.3.2. Define Z_1 and W_1 by

$$Z_1 = X_1(X_1^H X_1)^{-1/2}, \quad W_1 = Y_1(Y_1^H Y_1)^{-1/2}.$$

Then from (1.3.1) and (1.3.2)

$$\rho(\mathcal{X}_1, \mathcal{Y}_1) = \left\| \left(I - Z_1^H W_1 W_1^H Z_1\right)^{1/2} \right\|. \tag{1.3.10}$$

Moreover, we have

$$P_{X_1}^{\perp}(Y_1 - X_1)(Y_1^H Y_1)^{-1/2} = P_{X_1}^{\perp} Y_1 (Y_1^H Y_1)^{-1/2} = P_{Z_1}^{\perp} W_1,$$

and

$$\begin{aligned} \left\| P_{X_1}^{\perp}(Y_1 - X_1)(Y_1^H Y_1)^{-1/2} \right\| &= \left\| \left[\left(P_{Z_1}^{\perp} W_1\right)^H \left(P_{Z_1}^{\perp} W_1\right) \right]^{1/2} \right\| \\ &= \left\| \left(I - W_1^H Z_1 Z_1^H W_1\right)^{1/2} \right\|. \end{aligned} \tag{1.3.11}$$

Observe that

$$\sigma_+\left(\left(I - Z_1^H W_1 W_1^H Z_1\right)^{1/2}\right) = \sigma_+\left(\left(I - W_1^H Z_1 Z_1^H W_1\right)^{1/2}\right).$$

Hence, by a property possessed by unitarily invariant norms (see §1.2.3), we have

$$\left\| \left(I - Z_1^H W_1 W_1^H Z_1\right)^{1/2} \right\| = \left\| \left(I - W_1^H Z_1 Z_1^H W_1\right)^{1/2} \right\|.$$

Combining it with (1.3.10) and (1.3.11) shows the first equality of (1.3.8). Further, interchanging X_1 and Y_1 of the equality yields the second equality of (1.3.8). □

Let $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ be l-dimensional subspaces of $\mathcal{C}^n$, where $X_1, Y_1 \in \mathcal{U}^{n\times l}$. Moreover, let $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_{n_1}$ be the canonical angles between $\mathcal{X}_1$ and $\mathcal{Y}_1$, where n_1 is defined by (1.3.5). Then from (1.3.6), (1.3.7) and

$$\sin\theta_j \leq \tan\theta_j, \qquad \tan\theta_j \leq \frac{\sin\theta_j}{\sqrt{1-\sin^2\theta_1}} \quad \text{if} \quad \sin\theta_1 < 1,$$

we get

$$\rho(\mathcal{X}_1, \mathcal{Y}_1) \leq \|\tan\Theta(X_1, Y_1)\|, \tag{1.3.12}$$

and

$$\|\tan\Theta(X_1, Y_1)\| \leq \frac{\rho(\mathcal{X}_1, \mathcal{Y}_1)}{\sqrt{1-\rho_2^2(\mathcal{X}_1, \mathcal{Y}_1)}} \quad \text{if} \quad \rho_2(\mathcal{X}_1, \mathcal{Y}_1) < 1. \tag{1.3.13}$$

The following result gives some estimates of the distance between the subspaces $\mathcal{R}(X_1)$ and $\mathcal{R}(X_1 + X_2 Z)$, where $(X_1, X_2) \in \mathcal{U}^{n\times n}$ with $X_1 \in \mathcal{U}^{n\times l}$.

Theorem 1.3.3. *Let $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ with $X_1 \in \mathcal{U}^{n\times l}$. Let*

$$\tilde{X}_1 = X \begin{pmatrix} I_l \\ Z \end{pmatrix}, \tag{1.3.14}$$

and

$$Y_1 = \tilde{X}_1 (\tilde{X}_1^H \tilde{X}_1)^{-\frac{1}{2}}. \tag{1.3.15}$$

Then

$$\|Z\| = \|\tan\Theta(X_1, Y_1)\|, \tag{1.3.16}$$

and

$$\rho(\mathcal{X}_1, \mathcal{Y}_1) = \|Z\| + O(\|Z\|^3) \qquad \text{as } Z \to 0, \tag{1.3.17}$$

where $\Theta(X_1, Y_1)$ is defined by (1.3.1), $\mathcal{X}_1 = \mathcal{R}(X_1)$, and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$.

Proof. Let

$$Z = PTQ^H \tag{1.3.18}$$

be the singular value decomposition of Z, where $P \in \mathcal{U}^{(n-l)\times(n-l)}, Q \in \mathcal{U}^{l\times l}$, and $T = \text{diag}(\tau_j)$. Then by (1.3.1), (1.3.14), (1.3.15) and (1.3.18), we have

$$\sin\Theta(X_1, Y_1) = Q\text{diag}\left(\frac{\tau_1}{\sqrt{1+\tau_1^2}}, \ldots, \frac{\tau_l}{\sqrt{1+\tau_l^2}}\right) Q^H,$$

which shows that $\tau_j/\sqrt{1+\tau_j^2} = \sin\theta_j$, and so we have $\tau_j = \tan\theta_j$ for $j = 1, \ldots, n_1$, where n_1 is defined by (1.3.5). Combining this fact with (1.3.18) shows (1.3.16).

By the first equality of (1.3.8), we have

$$\rho(\mathcal{X}_1, \mathcal{Y}_1) = \left\| P_{X_1}^{\perp} Y_1 (Y_1^H Y_1)^{-1/2} \right\| = \left\| Z(I + Z^H Z)^{-1/2} \right\|. \qquad (1.3.19)$$

Substituting the expansion

$$(I + Z^H Z)^{-1/2} = I - \frac{1}{2} Z^H Z + \frac{3}{8}(Z^H Z)^2 - \cdots \qquad (\|Z\|_2 < 1)$$

into (1.3.19) gives the relation (1.3.17). □

Let $\mathcal{X}_1$ and $\mathcal{Y}_1$ be the subspaces of Theorem 1.3.3. From (1.3.12), (1.3.16) and (1.3.17) we see that a sharper upper bound ξ^* for $\|Z\|$ is also a sharper upper bound for $\rho(\mathcal{X}_1, \mathcal{Y}_1)$ if ξ^* is very small.

Notes and References

NR 1.3–1. Usually, $\mathcal{G}_l^n$ (the set of l-dimensional subspaces of $\mathcal{C}^n$) is called a *complex projective space*, or a *Grassmann manifold* (see, e.g., Hirsch [53, Chapter 1]). There are various approaches to introduce metrics on $\mathcal{G}_l^n$ (see Berkson [4], and Stewart and Sun [97, Chapter II]). According to Berkson [4], the metric $\rho_2(\mathcal{X}_1, \mathcal{Y}_1)$ of (1.3.3) was first defined on Hilbert space by Kreȋn and Krasnoselsky [65].

NR 1.3–2. The functions $\arccos(X_1^H Y_1 Y_1^H X_1)^{\frac{1}{2}}$, $\sin\Theta(X_1, Y_1)$ and $\tan\Theta(X_1, Y_1)$ of (1.3.1), (1.3.2) and (1.3.16) are matrix functions. For definitions and approximation methods of matrix functions, see Golub and Van Loan [41, Chapter 11].

NR 1.3–3. Theorem 1.3.1 is cited from Sun [99, Theorem 3.1], a proof of the result is given by Sun [104, Chapter 2, Theorem 4.4]. From (1.3.6) we see that the canonical angles between two subspaces play important role for measuring the distance between the subspaces. Consequently, it may well be asked: Can we characterize any unitarily invariant metric on $\mathcal{G}_l^n$ as some special kind of functions of the canonical angles? This is a research problem.

NR 1.3–4. Theorem 1.3.2 is a generalization of a result given by Sun [104, Chapter 4, Theorem 4.5], where only $\rho_2(\mathcal{X}_1, \mathcal{Y}_1)$ and $\rho_F(\mathcal{X}_1, \mathcal{Y}_1)$ are considered.

NR 1.3–5. For more results on the generalized chordal metric $\rho(\mathcal{X}_1, \mathcal{Y}_1)$ and the canonical angles, see Kato [63, Chapter IV], Davis and Kahan [26], Stewart [91], [93], and Stewart and Sun [97, Chapter II]. For numerical methods for computing the canonical angles, see Björck and Golub [7].

1.4 Matrix Orthogonal Decompositions

Let $X \in \mathcal{C}^{n\times l}$ with $\text{rank}(X) = l$. The QR factorization $X = QR$, the singular value decomposition $X = U\Sigma V^H$, and the polar decomposition $X = PH$ are important orthogonal decompositions of X, where $Q, U, P \in \mathcal{U}^{n\times l}$, $V \in \mathcal{U}^{l\times l}$, $R \in \mathcal{C}^{l\times l}$ is upper triangular with positive diagonal elements, $\Sigma \in \mathcal{R}^{l\times l}$ is diagonal with positive diagonal elements, and $H \in \mathcal{H}^{l\times l}$ is positive definite. The matrix Q is called the *unitary* QR *factor*, and P the *unitary polar factor* of X.

It is known that the unitary polar factor P of X possess the best approximation property:

$$\min_{U\in\mathcal{U}^{n\times l}} \|X - U\|_F = \|X - P\|_F.$$

In this section we shall show the following fact:

$$\|X^H X - I\|_F \ll 1 \quad \Longrightarrow \quad \|X - P\|_F \ll 1 \quad \text{and} \quad \|X - Q\|_F \ll 1, \tag{1.4.1}$$

where Q is the unitary QR factor of X.

We first prove a result on perturbations of the *Cholesky factor* of the identity matrix.

Theorem 1.4.1. *If $H \in \mathcal{H}^{n\times n}$ satisfies $\|H\|_2 < 1$, then $I + H$ has a unique Cholesky factorization*

$$I + H = LL^H, \tag{1.4.2}$$

where $L = I + G$ is a lower triangular matrix with positive diagonal elements, and

$$\|G\|_F \le \frac{\sqrt{2}\|H\|_F}{1 - \|H\|_2 + \sqrt{1 - \|H\|_2}}. \tag{1.4.3}$$

Proof. The assumption $\|H\|_2 < 1$ implies that the Hermitian matrix $I + H$ is positive definite, and so there is a unique Cholesky factorization (1.4.2). We now prove the estimate (1.4.3).

The elements of L are obviously differentiable functions of the elements of H. Differentiating (1.4.2) gives

$$dH = dLL^H + LdL^H$$

and

$$L^{-1}dL + (L^{-1}dL)^H = L^{-1}dHL^{-H}.$$

Combining it with

$$\sqrt{2}\|L\|_2^{-1}\|dL\|_F \le \sqrt{2}\|L^{-1}dL\|_F \le \|L^{-1}dL + (L^{-1}dL)^H\|_F$$

and

$$\|L^{-1}dHL^{-H}\|_F \leq \|L^{-1}\|_2^2\|dH\|_F$$

shows

$$\|dL\|_F \leq \frac{1}{\sqrt{2}}\|L\|_2\|L^{-1}\|_2^2\|dH\|_F. \tag{1.4.4}$$

Let

$$A(t) = I + tH, \quad -1 \leq t \leq 1.$$

From $\|H\|_2 < 1$ we see that $A(t)$ is positive definite, and there is a unique Cholesky factorization

$$A(t) = L(t)L(t)^H \text{ with } L(0) = I \text{ and } L(1) = I + G.$$

By (1.4.4), we have

$$\|dL(t)\|_F \leq \frac{1}{\sqrt{2}}\|H\|_F\|L(t)\|_2\|L(t)^{-1}\|_2^2 dt. \tag{1.4.5}$$

Let $\lambda_1(t) \geq \cdots \geq \lambda_n(t)$ be the eigenvalues of $A(t)$. Obviously, $\lambda_j(0) = 1$ for all j. From (1.4.5)

$$\begin{aligned}\|G\|_F &= \|L(1) - L(0)\|_F = \left\|\int_0^1 dL(t)\right\|_F \\ &\leq \int_0^1 \|dL(t)\|_F \leq \frac{\|H\|_F}{\sqrt{2}}\int_0^1 \frac{\sqrt{\lambda_1(t)}}{\lambda_n(t)}dt.\end{aligned}$$

Observe that by the Weyl theorem [128] (or see Stewart and Sun [97, p.203]),

$$\lambda_1(t) \leq 1 + \|H\|_2 t, \qquad \lambda_n(t) \geq 1 - \|H\|_2 t.$$

Hence, we have

$$\begin{aligned}\|G\|_F &\leq \frac{\|H\|_F}{\sqrt{2}}\int_0^1 \frac{\sqrt{1+\|H\|_2 t}}{1-\|H\|_2 t}dt \leq \frac{\|H\|_F}{\sqrt{2}}\int_0^1 \frac{dt}{(1-\|H\|_2 t)^{3/2}} \\ &= \frac{\sqrt{2}\|H\|_F}{1 - \|H\|_2 + \sqrt{1-\|H\|_2}}.\end{aligned}$$

The proof is completed. □

The following result gives upper bounds for $\|X - P\|_F$ and $\|X - Q\|_F$, where P and Q are the unitary polar factor and the unitary QR factor of X, respectively.

Theorem 1.4.2. *Let $X = PH$ and $X = QR$ be the polar decomposition and the QR factorization of a full column rank matrix X, respectively. Then if*

$$\|X^H X - I\|_2 < 1,$$

we have

$$\|X - P\|_F \leq \frac{\|X^H X - I\|_F}{1 + \sigma_{\min}(X)}, \tag{1.4.6}$$

and

$$\|X - Q\|_F \leq \frac{\sqrt{2}[1 + \sigma_{\max}(X)]}{1 - \|X^H X - I\|_2 + \sqrt{1 - \|X^H X - I\|_2}} \|X - P\|_F. \tag{1.4.7}$$

Proof. By the decomposition $X = PH$, we have

$$\begin{aligned} \|X - P\|_F &= \|H - I\|_F = \|(H + I)^{-1}(H^2 - I)\|_F \\ &\leq \|H^2 - I\|_F/[1 + \lambda_{min}(H)] = \|X^H X - I\|_F/[1 + \sigma_{\min}(X)]. \end{aligned}$$

The estimate (1.4.6) is proved.

Observe that R^H is the Cholesky factor of the Hermitian positive definite matrix H^2. Moreover, H^2 and R^H can be regarded as perturbations of the identity matrix I and its Cholesky factor I. Hence, by Theorem 1.4.1, if $\|H^2 - I\|_2 < 1$ then

$$\|R - I\|_F \leq \frac{\sqrt{2}\|H^2 - I\|_F}{1 - \|H^2 - I\|_2 + \sqrt{1 - \|H^2 - I\|_2}}.$$

This together with

$$\|R - I\|_F = \|X - Q\|_F, \qquad H^2 = X^H X,$$

$$\|H^2 - I\|_F \leq [1 + \lambda_{\max}(H)]\|H - I\|_F = [1 + \sigma_{\max}(X)]\|X - P\|_F,$$

shows (1.4.7). □

Substituting (1.4.6) into (1.4.7) gives

$$\|X - Q\|_F \leq \frac{\sqrt{2}[1 + \sigma_{\max}(X)]\|X^H X - I\|_F}{[1 + \sigma_{\min}(X)]\left(1 - \|X^H X - I\|_2 + \sqrt{1 - \|X^H X - I\|_2}\right)}. \tag{1.4.8}$$

It is evident that the estimates (1.4.6) and (1.4.8) imply the fact (1.4.1).

Note that for the unitary factor U of the singular value decomposition $X = U\Sigma V^H$, the assumption $\|X^H X - I\|_F \ll 1$ doesn't guarantee $\|X - U\|_F \ll 1$.

Notes and References

NR 1.4–1. This section is based on Sun [110, Theorem 1.4] and Sun [116, Lemma 2.4].

NR 1.4–2. For the QR factorization and the polar decomposition, as well as the best approximation property of the unitary polar factor, see Fan and Hoffman [37], Golub and Van Loan [41, Chapters 5 and 12], and Higham [47].

NR 1.4–3. Let X, P, Q be as in Theorem 1.4.2, where $X = (x_1, \ldots, x_l)$. Chandrasekaran and Ipsen [18] prove that if $\|x_i\|_2 = 1$ for all i, then

$$\|X - Q\|_F \le 5\sqrt{l}\|X - P\|_2. \tag{1.4.9}$$

Obviously, the estimates (1.4.7) and (1.4.9) require different conditions. Note that if X satisfies $\|X^H X - I\|_2 \le 0.6755 \equiv \epsilon$, then the estimate (1.4.7) implies

$$\|X - Q\|_F < \frac{1 + \sqrt{1+\epsilon}}{\sqrt{2}(1-\epsilon)}\|X - P\|_F < 5\|X - P\|_F \le 5\sqrt{l}\|X - P\|_2.$$

NR 1.4–4. Theorem 1.4.1 gives a perturbation bound for the Cholesky factor of the identity matrix. A nice perturbation analysis of the Cholesky factorization is given by Chang, Paige and Stewart [19].

1.5 Solutions of Some Matrix Equations

In this section we consider two kinds of matrix equations. The first one is

$$AEB = C, \tag{1.5.1}$$

where $A \in \mathcal{C}^{p\times m}, B \in \mathcal{C}^{n\times q}, C \in \mathcal{C}^{p\times q}$, and $E \in \mathcal{C}^{m\times n}$ is the unknown matrix. The second one is

$$HB = C, \tag{1.5.2}$$

where $B, C \in \mathcal{C}^{n\times l}$, and $H \in \mathcal{H}^{n\times n}$ is the unknown matrix.

The following result gives explicit expressions of the solutions to the equation (1.5.1).

Theorem 1.5.1. *Let $A \in \mathcal{C}^{p\times m}, B \in \mathcal{C}^{n\times q}$ and $C \in \mathcal{C}^{p\times q}$ be given. Define the sets $\mathcal{E}$ and $\mathcal{F}$ by*

$$\mathcal{E} = \{E \in \mathcal{C}^{m\times n} \;:\; AEB = C\}$$

and

$$\mathcal{F} = \{A^\dagger C B^\dagger + Z - P_{A^H} Z P_B \;:\; Z \in \mathcal{C}^{m\times n}\},$$

respectively. Then $\mathcal{E} \neq \emptyset$ if and only if A, B and C satisfy

$$P_A C P_{B^H} = C, \tag{1.5.3}$$

and in the case of $\mathcal{E} \neq \emptyset$, we have $\mathcal{E} = \mathcal{F}$.

Proof. The relation (1.5.3) is obviously a necessary condition for $\mathcal{E} \neq \emptyset$. We now prove $\mathcal{E} = \mathcal{F}$ under the condition (1.5.3).

Assume $E \in \mathcal{E}$. Then we may represent the matrix E as

$$E = A^\dagger C B^\dagger + E - P_{A^H} E P_B.$$

This means that there exists a matrix Z $(= E) \in \mathcal{C}^{m\times n}$ such that the matrix $E \in \mathcal{E}$ may be expressed by

$$E = A^\dagger C B^\dagger + Z - P_{A^H} Z P_B \in \mathcal{F}. \tag{1.5.4}$$

Thus, $\mathcal{E} \subset \mathcal{F}$.

Conversely, assume $E \in \mathcal{F}$, and let E be expressed by (1.5.4) with some $Z \in \mathcal{C}^{m\times n}$. Then the expression (1.5.4) and the condition (1.5.3) imply $AEB = C$, i.e., $E \in \mathcal{E}$. Thus, $\mathcal{F} \subset \mathcal{E}$. Consequently, we have $\mathcal{E} = \mathcal{F}$. □

The following result gives explicit expressions of the solutions to the equation (1.5.2).

Theorem 1.5.2. *Let $B, C \in \mathcal{C}^{n\times l}$ be given. Define the sets $\mathcal{H}$ and $\mathcal{G}$ by*

$$\mathcal{H} = \{H \in \mathcal{H}^{n\times n} \ : \ HB = C\}$$

and

$$\mathcal{G} = \{CB^\dagger + {B^\dagger}^H C^H - {B^\dagger}^H C^H P_B + P_B^\perp T P_B^\perp \ : \ T \in \mathcal{H}^{n\times n}\},$$

respectively. Then $\mathcal{H} \neq \emptyset$ if and only if B and C satisfy

$$CP_{B^H} = C \quad \text{and} \quad P_B C B^\dagger \in \mathcal{H}^{n\times n}, \tag{1.5.5}$$

and in the case of $\mathcal{H} \neq \emptyset$, we have $\mathcal{H} = \mathcal{G}$.

Proof. It can be verified that the relations (1.5.5) are necessary conditions for $\mathcal{H} \neq \emptyset$. We now prove $\mathcal{H} = \mathcal{G}$ under the conditions (1.5.5).

Assume $H \in \mathcal{H}$. Then we may represent the matrix H as

$$H = CB^\dagger + {B^\dagger}^H C^H - {B^\dagger}^H C^H P_B + P_B^\perp H P_B^\perp.$$

This means that there exists a matrix $T \ (= H) \in \mathcal{H}^{n\times n}$ such that the matrix $H \in \mathcal{H}$ may be expressed by

$$H = CB^{\dagger} + {B^{\dagger}}^{H} C^{H} - {B^{\dagger}}^{H} C^{H} P_B + P_B^{\perp} T P_B^{\perp} \in \mathcal{G}. \tag{1.5.6}$$

Thus, $\mathcal{H} \subset \mathcal{G}$.

Conversely, assume $H \in \mathcal{G}$, and let H be expressed by (1.5.6) with some $T \in \mathcal{H}^{n\times n}$. Then the expression (1.5.6) and the condition (1.5.5) imply $H \in \mathcal{H}^{n\times n}$ and $HB = C$, i.e., $H \in \mathcal{H}$. Thus, $\mathcal{G} \subset \mathcal{H}$. Consequently, we have $\mathcal{H} = \mathcal{G}$. □

Notes and References

NR 1.5–1. This section is based on Sun [115, Lemmas 1.3 and 1.4]. The proofs given in this section are simpler.

NR 1.5–2. The following results are known (see Dennis and Moré [30], Higham [49], and Bunch, Demmel, and Van Loan [11]):

Proposition 1.5.3. *let b, c be real vectors, and $b \neq 0$. Then*

$$E = \frac{cb^T}{b^T b}$$

is the smallest real matrix in the spectral norm and Frobenius norm for which the vector b, c satisfy $Eb = c$.

Proposition 1.5.4. *Let b, c be real vectors, and $b \neq 0$. Then*

$$H = \frac{cb^T + bc^T}{b^T b} - \frac{b^T c}{(b^T b)^2} bb^T$$

is the smallest real symmetric matrix in the Frobenius norm for which the vector b, c satisfy $Hb = c$.

Propositions 1.5.3 and 1.5.4 can be obtained by applying Theorems 1.5.1 and 1.5.2, respectively.

NR 1.5–3. Let $B, C \in \mathcal{C}^{n\times l}$ be given. Define the set $\mathcal{U}$ by

$$\mathcal{U} = \{U \in \mathcal{C}^{n\times n} \ : \ UB = C\}.$$

Explicit expressions of the elements of $\mathcal{U}$ are discussed by Sun [118].

1.6 The Implicit Function Theorem

The implicit function theorem is an important existence theorem in analysis. In this section we cite the implicit function theorem on analytic functions.

We first introduce the definition of an analytic function.

Definition 1.6.1. Let $x = (\xi_1, \ldots, \xi_n)^T \in \mathcal{C}^n$, and let $f(x)$ be a complex-valued function defined in an open set $\mathcal{D} \subset \mathcal{C}^n$. The function $f(x)$ is said to be *analytic* at a point $a = (\alpha_1, \ldots, \alpha_n)^T \in \mathcal{D}$ if there is a neighborhood $\mathcal{B}(a) \subset \mathcal{D}$ of a such that $f(x)$ can be expressed as a convergent *power series*

$$f(x) = \sum_{m_1,\ldots m_n=0}^{\infty} c_{m_1\cdots m_n}(\xi_1-\alpha_1)^{m_1}\cdots(\xi_n-\alpha_n)^{m_n}, \qquad (\xi_1,\ldots,\xi_n)^T \in \mathcal{B}(a). \tag{1.6.1}$$

If $f(x)$ is analytic at any point $a \in \mathcal{D}$, then the function $f(x)$ is said to be *analytic* in $\mathcal{D}$.

In the same way we can define an analytic real-valued function $f(x)$ in an open set $\mathcal{D} \subset \mathcal{R}^n$.

Note that a complex-valued function $f(x) = u(x)+iv(x)$ of $x \in \mathcal{R}^n$ with $i = \sqrt{-1}$ is said to be an *analytic function* of x if the real-valued functions $u(x)$ and $v(x)$ are analytic functions of x.

A basic fact about analytic functions is that if the complex-valued (or real-valued) function $f(x)$ is analytic at $a \in \mathcal{C}^n$ (or $\mathcal{R}^n$), then there is a neighborhood $\mathcal{B}(a)$ of a such that $f(x)$ has continuous partial derivatives

$$\frac{\partial^{m_1+\cdots+m_n} f(x)}{\partial\xi_1^{m_1}\cdots\partial\xi_n^{m_n}} \quad \text{for} \quad m_1,\ldots,m_n \geq 0, \quad x \in \mathcal{B}(a),$$

and the coefficients $c_{m_1\cdots m_n}$ of the power series expansion (1.6.1) can be expressed by

$$c_{m_1\cdots m_n} = \frac{1}{m_1!\cdots m_n!}\left[\frac{\partial^{m_1+\cdots+m_n} f(x)}{\partial\xi_1^{m_1}\cdots\partial\xi_n^{m_n}}\right]_{x=a}.$$

Suppose that the function

$$f : \mathcal{D} \subset \mathcal{C}^n \to \mathcal{C}^m \text{ (or } \mathcal{D} \subset \mathcal{R}^n \to \mathcal{R}^m\text{)},$$

with

$$f(x) = (f_1(x),\ldots,f_m(x))^T, \quad x = (x_1,\ldots,x_n)^T,$$

is defined in an open subset $\mathcal{D}$ of $\mathcal{C}^n$ (or $\mathcal{R}^n$), and that its component functions $f_i, i = 1,\ldots,m$, have continuous first order partial derivatives on $\mathcal{D}$. Then we define the *Jacobian matrix* f'_x by

$$f'_x = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix},$$

and in the case of $m = n$, we define the *Jacobian* $\frac{\partial(f_1,\ldots,f_n)}{\partial(x_1,\ldots,x_n)}$ (or simply, $\frac{\partial f}{\partial x}$) by

$$\frac{\partial(f_1,\ldots,f_n)}{\partial(x_1,\ldots,x_n)} = \det f'_x.$$

The following implicit function theorem is a basic tool of this work for deriving perturbation expansions for eigenvalues, singular values, generalized eigenvalues, and certain characteristic subspaces.

Theorem 1.6.2. *If the complex-valued (or real-valued) functions*

$$f_j(\xi_1,\ldots,\xi_k;\eta_1,\ldots,\eta_l), \quad j = 1,\ldots,k$$

are analytic functions of $k+l$ complex (or real) variables in some neighborhood of the origin of $\mathcal{C}^{k+l}$ (or $\mathcal{R}^{k+l}$) if $f_j(0;0) = 0$, $j = 1,\ldots,k$, and if the Jacobian

$$\frac{\partial(f_1,\ldots,f_k)}{\partial(\xi_1,\ldots,\xi_k)} \neq 0 \quad \text{for} \quad \xi_1 = \cdots = \xi_k = \eta_1 = \cdots = \eta_l = 0,$$

then the equations

$$f_j(\xi_1,\ldots,\xi_k;\eta_1,\ldots,\eta_l) = 0, \quad j = 1,\ldots,k$$

have a unique solution

$$\xi_j = g_j(\eta_1,\ldots,\eta_l), \quad j = 1,\ldots,k$$

vanishing for $\eta_1 = \cdots = \eta_l = 0$ and analytic in some neighborhood of the origin of $\mathcal{C}^l$ (or $\mathcal{R}^l$).

The following result is about the Jacobian.

Theorem 1.6.3. *If $f_j(z_1,\ldots,z_k)$, $j = 1,\ldots,k$, are analytic functions of complex variables $z_1,\ldots,z_k$, and if $f_j = u_j + iv_j$ and $z_j = x_j + iy_j$ with $i = \sqrt{-1}$, then*

$$\frac{\partial(u_1,v_1,\ldots,u_k,v_k)}{\partial(x_1,y_1,\ldots,x_k,y_k)} = \left|\frac{\partial(f_1,\ldots,f_k)}{\partial(z_1,\ldots,z_k)}\right|^2.$$

Notes and References

NR 1.6–1. Most of the materials of this section are cited from Bochner and Martin [8, Chapter II]. Theorem 1.6.2 for real-valued function is cited from Dieudonné [32, p.277].

1.7 Fixed Point Theorems

Fixed point theorems are also important existence theorems in analysis. In this section we cite two results, the Brouwer fixed point theorem and Schauder fixed point theorem, from the fixed point theory.

Theorem 1.7.1 (The Brouwer Fixed Point Theorem). *Let $\mathcal{S}$ be a compact convex set in $\mathcal{R}^n$, and $\mathcal{M}$ be a continuous mapping on $\mathcal{S}$ which maps $\mathcal{S}$ into $\mathcal{S}$. Then $\mathcal{M}$ has a fixed point in $\mathcal{S}$.*

The Brouwer fixed point theorem extends to an arbitrary Banach space; this is the Schauder fixed point theorem.

Theorem 1.7.2 (The Schauder Fixed Point Theorem). *Let $\mathcal{S}$ be a compact convex set in a Banach space $\mathcal{B}$, and $\mathcal{M}$ be a continuous mapping on $\mathcal{S}$ which maps $\mathcal{S}$ into $\mathcal{S}$. Then $\mathcal{M}$ has a fixed point in $\mathcal{S}$.*

Notes and References

NR 1.7–1. The Brouwer fixed point theorem and the Schauder fixed point theorem are well known results of functional analysis (see, e.g., Ortega and Rheinboldt [81, §6.3], or E. Zeidler [137, §2.3 and §2.6]).

1.8 Condition Numbers

In this section, we shall introduce definitions of *normwise* condition numbers.

Let $x = \phi(a)$ be a solution of a matrix problem, where $a \in \mathcal{A}$ and $x \in \mathcal{X}$, $\mathcal{A}$ and $\mathcal{X}$ are finite dimensional normed linear spaces with the norms $\nu(\cdot)$ and $\mu(\cdot)$, respectively. A condition number of x is a measure of the sensitivity of the solution x to small changes in a.

Let Δa be any perturbation in a, and Δx be the corresponding perturbation in the solution x. Then by Rice [88], the *condition number* $c(x)$ of x can be defined by

$$c(x) = \lim_{\delta \to 0} \sup_{\frac{\nu(\Delta a)}{\alpha} \leq \delta} \frac{\mu(\Delta x)}{\xi \delta}, \tag{1.8.1}$$

where α, ξ are positive parameters. For instance, taking $\alpha = \xi = 1$ we get the *absolute* condition number $c_{\text{abs}}(x)$, and taking $\alpha = \nu(a)$ and $\xi = \mu(x)$ (if $\nu(a) \neq 0$ and $\mu(x) \neq 0$) we get the *relative* condition number $c_{\text{rel}}(x)$.

It is known that the *conditioning* of a problem is the sensitivity of the solution to perturbations on the data. Consequently, the relative (absolute) condition number $c_{\text{rel}}(x)$ $(c_{\text{abs}}(x))$ is a measure of the relative (absolute) conditioning of the problem.

From the definition (1.8.1) it follows that in first order approximation the inequality

$$\frac{\mu(\Delta x)}{\xi} \leq c(x)\frac{\nu(\Delta a)}{\alpha} \tag{1.8.2}$$

holds.

More general, assume that $\mathcal{A}$ and $\mathcal{X}$ are finite-dimensional metric spaces with the metrics $d_{\mathcal{A}}(\cdot,\cdot)$ and $d_{\mathcal{X}}(\cdot,\cdot)$. The condition number $c(x)$ of x can be defined by

$$c(x) = \lim_{\delta\to 0} \sup_{\frac{d_{\mathcal{A}}(a,\tilde{a})}{\alpha}\leq\delta} \frac{d_{\mathcal{X}}(x,\tilde{x})}{\xi\delta}, \tag{1.8.3}$$

where $\tilde{x} = \phi(\tilde{a})$, and α, ξ are positive parameters.

From the definition (1.8.3) it follows that in first order approximation the inequality

$$\frac{d_{\mathcal{X}}(x,\tilde{x})}{\xi} \leq c(x)\frac{d_{\mathcal{A}}(a,\tilde{a})}{\alpha}$$

holds.

If the data a have some special structure (i.e., $a \in \mathcal{A}_s$, a subset of $\mathcal{A}$), and if we are interested in the requirement that the perturbed elements $\tilde{a}$ have the same special structure (i.e., $\tilde{a} \in \mathcal{A}_s$) too, then we may define the *structured* condition number $c_s(x)$ of x by

$$c_s(x) = \lim_{\delta\to 0} \sup_{\substack{\tilde{a}\in\mathcal{A}_s \\ \frac{d_{\mathcal{A}}(a,\tilde{a})}{\alpha}\leq\delta}} \frac{d_{\mathcal{X}}(x,\tilde{x})}{\xi\delta},$$

where α, ξ are positive parameters.

If one is interested in the sensitivity of the solution $x = \phi(a_1, a_2)$ to perturbations in each individual member of a_1 and a_2, then we may define the *partial* condition numbers $c_{a_1}(x)$ and $c_{a_2}(x)$ of x by

$$\begin{aligned} c_{a_1}(x) &= \lim_{\delta\to 0} \sup_{\frac{d_{\mathcal{A}}(a,\tilde{a})}{\alpha_1}\leq\delta,\ \tilde{a}_2=a_2} \frac{d_{\mathcal{X}}(x,\tilde{x})}{\xi\delta}, \\ c_{a_2}(x) &= \lim_{\delta\to 0} \sup_{\frac{d_{\mathcal{A}}(a,\tilde{a})}{\alpha_2}\leq\delta,\ \tilde{a}_1=a_1} \frac{d_{\mathcal{X}}(x,\tilde{x})}{\xi\delta}. \end{aligned} \tag{1.8.4}$$

where α_1, α_2 and ξ are positive parameters.

We now consider the case of $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$, a product space. Assume that $\mathcal{A}_1 \times \mathcal{A}_2$ and $\mathcal{X}$ are finite dimensional normed linear spaces with the norm $\nu(\cdot)$ and $\mu(\cdot)$, respectively. Moreover, assume that the norms $\nu_i(\cdot)$ are the restrictions of $\nu(\cdot)$ on $\mathcal{A}_i$ for $i = 1, 2$, and write $\nu_i(\cdot)$ as $\nu(\cdot)$, i.e., $\nu(a_1) = \nu(a_1, 0)$ and $\nu(a_2) = \nu(0, a_2)$.

Let Δa_1 and Δa_2 be any perturbations in a_1 and a_2, respectively, and Δx be the corresponding perturbation in the solution x. As a generalization of the definition (1.8.1), we can define the condition number $c(x)$ of x by the following approach. First, we define the vector $v \in \mathcal{R}^2$ by

$$v = \left(\frac{\nu(\Delta a_1)}{\alpha_1}, \frac{\nu(\Delta a_2)}{\alpha_2} \right)^T,$$

and then define the condition number $c(x)$ of x as

$$c(x) = \lim_{\delta \to 0} \sup_{\|v\| \leq \delta} \frac{\mu(\Delta x)}{\xi \delta}, \tag{1.8.5}$$

where $\| \cdot \|$ denotes any norm on $\mathcal{R}^2$, and α_1, α_2, ξ are positive parameters.

As another generalization of the definition (1.8.1), we can define the condition number $c^*(x)$ of x by

$$c^*(x) = \lim_{\delta \to 0} \sup_{\frac{\nu(\Delta a_1, \Delta a_2)}{\alpha} \leq \delta} \frac{\mu(\Delta x)}{\xi \delta}, \tag{1.8.6}$$

where α, ξ are positive parameters.

The definitions (1.8.5) and (1.8.6) imply that in first order approximation the inequalities

$$\frac{\mu(\Delta x)}{\xi} \leq c(x) \left\| \left(\frac{\nu(\Delta a_1)}{\alpha_1}, \frac{\nu(\Delta a_2)}{\alpha_2} \right)^T \right\| \tag{1.8.7}$$

and

$$\frac{\mu(\Delta x)}{\xi} \leq c^*(x) \frac{\nu(\Delta a_1, \Delta a_2)}{\alpha} \tag{1.8.8}$$

hold.

From the definitions (1.8.5) and (1.8.6) we see that every condition number of x is defined with respect to a particular class of perturbations in a_1 and a_2. Therefore, different condition numbers have different meanings, and the values of two different condition numbers of the solution with the same data may be quite different.

Notes and References

NR 1.8–1. The study of conditioning in matrix computations is an important subject of matrix perturbation theory, on which there is a very large literature. The first general theory of condition was developed by Rice [88].

NR 1.8–2. For the study of *componentwise* condition numbers and structured condition numbers of some numerical linear algebra problems, see D. Higham and N. Higham [45], [46], N. Higham [52, §7.2], and Chaitin-Chatelin and Frayssé [17, Chapter 3].

1.9 Backward Errors

A matrix problem may be cast in the form of solving an equation $r(a;x) = 0$, where $a \in \mathcal{A}$, and the solution $x \in \mathcal{X}$. For example, $r(A;x,\lambda) = \lambda x - Ax$ for the eigenvalue problem $Ax = \lambda x$, where $A \in \mathcal{C}^{n\times n}$, and the solution $(x,\lambda) \in \mathcal{C}^n \times \mathcal{C}$, the product space of $\mathcal{C}^n$ and $\mathcal{C}$.

Let $\tilde{x}$ be an approximate solution of the problem $r(a;x) = 0$. For example, $\tilde{x}$ may come from a numerical algorithm for approximating the solution. Then it may well be asked: Is $\tilde{x}$ the exact solution of a slightly perturbed problem?

For answering the question, we need the notion of backward error of the problem $r(a;x) = 0$ with respect to the approximate solution $\tilde{x}$. In this section we shall introduce definitions of *normwise* backward errors.

Let

$$\mathcal{E} = \{\Delta a \ : \ a + \Delta a \in \mathcal{A} \ \text{ and } \ r(a + \Delta a; \tilde{x}) = 0\}.$$

In general, the set $\mathcal{E}$ has many (even an infinity of) elements. The *backward error* $\eta(\tilde{x})$ is defined by

$$\eta(\tilde{x}) = \min_{\Delta a \in \mathcal{E}} \frac{\nu(\Delta a)}{\alpha}, \tag{1.9.1}$$

where $\nu(\cdot)$ is a norm on $\mathcal{A}$, and α is a positive parameter. For instance, taking $\alpha = 1$ yields the *absolute* backward error, and taking $\alpha = \nu(a)$ (if $\nu(a) \neq 0$) yields the *relative* backward error. A small $\eta(\tilde{x})$ means that the approximate solution $\tilde{x}$ is the exact solution of a slightly perturbed problem.

An algorithm for approximating the solution of the problem $r(a;x) = 0$ is defined to be *backward stable* if, for any $a \in \mathcal{A}$, it produces a computed $\tilde{x}$ with a small $\eta(\tilde{x})$. Consequently, a computable formula of the backward error $\eta(\tilde{x})$ may be useful for testing the *stability* of practical algorithms.

For the problem $r(a;x) = 0$, any element $\Delta a \in \mathcal{E}$ is called a *backward perturbation* of a associated with $\tilde{x}$, and the element $\Delta a_{\rm opt} \in \mathcal{E}$ satisfying $\eta(\tilde{x}) = \nu(\Delta a_{\rm opt})/\alpha$ is called the *optimal* (minimum) backward perturbation. Therefore, the backward error $\eta(\tilde{x})$ is also known as the *optimal backward perturbation bound.*

If the data a of the problem $r(a;x) = 0$ have some special structure (i.e., $a \in \mathcal{A}_s$, a subset of $\mathcal{A}$), and if we are interested in the requirement that the perturbed elements $a + \Delta a$ have the same special structure (i.e., $a + \Delta a \in \mathcal{A}_s$) too, then we may define a structured backward error. Let

$$\mathcal{E}_s \equiv \{\Delta a \ : \ a + \Delta a \in \mathcal{A}_s \ \text{ and } \ r(a + \Delta a; \tilde{x}) = 0\}.$$

In general, the set $\mathcal{E}_s$ has many (even an infinity of) elements. The *structured* backward error $\eta_s(\tilde{x})$ is defined by

$$\eta_s(\tilde{x}) = \min_{\Delta a \in \mathcal{E}_s} \frac{\nu(\Delta a)}{\alpha},$$

where $\nu(\cdot)$ is any norm on $\mathcal{A}$, and α is a positive parameter.

Bunch [10] defines that an algorithm for solving $r(a;x) = 0$ is *strongly* backward stable if, for any $a \in \mathcal{A}_s$, it produces a computed $\tilde{x}$ with a small $\eta_s(\tilde{x})$. Consequently, a computable formula of the structured backward error $\eta_s(\tilde{x})$ may be useful for testing the *strong stability* of practical algorithms.

We now consider the case of $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$, a product space. Let $\nu(\cdot)$ be a norm on $\mathcal{A}_1 \times \mathcal{A}_2$. Assume that the norms $\nu_i(\cdot)$ are the restrictions of $\nu(\cdot)$ on $\mathcal{A}_i$ for $i = 1, 2$, and write $\nu_i(\cdot)$ as $\nu(\cdot)$, i.e., $\nu(a_1) = \nu(a_1, 0)$ and $\nu(a_2) = \nu(0, a_2)$. In such a case, there are various ways to define normwise backward errors. For instance, the following definitions are advisable:

(i) Define the backward error $\eta_\infty(\tilde{x})$ by

$$\eta_\infty(\tilde{x}) = \min \left\{ \epsilon \ : \ \begin{array}{l} (a_1 + \Delta a_1, a_2 + \Delta a_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \\ \\ r(a_1 + \Delta a_1, a_2 + \Delta a_2; \tilde{x}) = 0, \\ \\ \nu_i(\Delta a_i) \le \epsilon \alpha_i, \ i = 1, 2 \end{array} \right\}, \tag{1.9.2}$$

where $\nu_1(\cdot)$ and $\nu_2(\cdot)$ are any norms on $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, and α_1, α_2 are positive parameters.

(ii) Define the backward error $\beta^{(\omega)}(\tilde{x})$ by

$$\beta^{(\omega)}(\tilde{x}) = \min \left\{ \mu \begin{pmatrix} \nu_1(\Delta a_1) \\ \omega \nu_2(\Delta a_2) \end{pmatrix} \ : \ \begin{array}{l} (a_1 + \Delta a_1, a_2 + \Delta a_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \\ \\ r(a_1 + \Delta a_1, a_2 + \Delta a_2; \tilde{x}) = 0 \end{array} \right\}, \tag{1.9.3}$$

where $\nu_1(\cdot)$ and $\nu_2(\cdot)$ are any norms on $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, $\mu(\cdot)$ is any norm on $\mathcal{R}^2$, and ω is a positive parameter.

(iii) Define the backward error $\eta^{(\theta)}(\tilde{x})$ by

$$\eta^{(\theta)}(\tilde{x}) = \min\left\{ \nu(\Delta a_1, \theta\Delta a_2) \ : \begin{array}{c} (a_1+\Delta a_1, a_2+\Delta a_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \\ r(a_1+\Delta a_1, a_2+\Delta a_2; \tilde{x}) = 0 \end{array} \right\}, \qquad (1.9.4)$$

where $\nu(\cdot)$ is any norm on $\mathcal{A}_1 \times \mathcal{A}_2$, and θ is a positive parameter.

It is worth pointing out that the parameters ω and θ in (1.9.3) and (1.9.4) allow us some flexibility. We now note some examples:

Example 1.9.1. Let α_1, α_2 be any positive scalars (for instance, $\alpha_1 = \alpha_2 = 1$, or $\alpha_i = \nu_i(a_i)$ if $\nu(a_i) \neq 0$ for $i = 1, 2$). Taking $\omega = \alpha_1/\alpha_2$ and $\mu(\cdot) = \|\cdot\|$ (any norm on $\mathcal{R}^2$) in (1.9.3), and multiplying $\beta^{(\omega)}(\tilde{x})$ by $1/\alpha_1$, yields the backward error

$$\begin{aligned} \eta(\tilde{x}) &\equiv \frac{1}{\alpha_1}\beta^{(\alpha_1/\alpha_2)}(\tilde{x}) \\ &= \min\left\{ \left\| \begin{pmatrix} \nu_1(\Delta a_1)/\alpha_1 \\ \nu_2(\Delta a_2)/\alpha_2 \end{pmatrix} \right\| \ : \begin{array}{c} (a_1+\Delta a_1, a_2+\Delta a_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \\ r(a_1+\Delta a_1, a_2+\Delta a_2; \tilde{x}) = 0 \end{array} \right\}. \end{aligned} \qquad (1.9.5)$$

Particularly, taking $\|\cdot\| = \|\cdot\|_p$ with $p = 1, 2, \infty$ in (1.9.5), yields the backward errors $\eta_1(\tilde{x})$, $\eta_2(\tilde{x})$, and $\eta_\infty(\tilde{x})$, respectively, where $\eta_\infty(\tilde{x})$ coincides with (1.9.2).

Example 1.9.2. Taking $\theta = 1$ in (1.9.4), and multiplying $\eta^{(\theta)}(\tilde{x})$ by $1/\alpha$, yields the backward error

$$\begin{aligned} \eta^*(\tilde{x}) &\equiv \frac{1}{\alpha}\eta^{(1)}(\tilde{x}) \\ &= \min\left\{ \frac{\nu(\Delta a_1, \Delta a_2)}{\alpha} \ : \begin{array}{c} (a_1+\Delta a_1, a_2+\Delta a_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \\ r(a_1+\Delta a_1, a_2+\Delta a_2; \tilde{x}) = 0 \end{array} \right\}, \end{aligned} \qquad (1.9.6)$$

where α is a positive parameter. For instance, $\alpha = 1$, or $\alpha = \nu(a_1, a_2)$ if $\nu(a_1, a_2) \neq 0$.

Example 1.9.3. Let α_1, α_2 be any positive scalars (for instance, $\alpha_1 = \alpha_2 = 1$, or $\alpha_i = \nu_i(a_i)$ if $\nu(a_i) \neq 0$ for $i = 1, 2$). Taking $\theta = \alpha_1/\alpha_2$ in (1.9.4), and multiplying $\eta^{(\theta)}(\tilde{x})$ by $1/\alpha_1$, yields the backward error

$$\begin{aligned} \hat{\eta}(\tilde{x}) &\equiv \frac{1}{\alpha_1}\eta^{(\alpha_1/\alpha_2)}(\tilde{x}) \\ &= \min\left\{ \nu\left(\frac{\Delta a_1}{\alpha_1}, \frac{\Delta a_2}{\alpha_2}\right) \ : \begin{array}{c} (a_1+\Delta a_1, a_2+\Delta a_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \\ r(a_1+\Delta a_1, a_2+\Delta a_2; \tilde{x}) = 0 \end{array} \right\}, \end{aligned} \qquad (1.9.7)$$

Example 1.9.4. Taking $\theta \to \infty$ forces $\Delta a_2 = 0$ in (1.9.4), yields the backward error where only a_1 is perturbed.

We now assume the norm $\nu(\cdot)$ on $\mathcal{A}_1 \times \mathcal{A}_2$ has the property that if $\nu(a_i) \le \nu(\hat{a}_i)$ $(i = 1, 2)$ then $\nu(a_1, a_2) \le \nu(\hat{a}_1, \hat{a}_2)$. The following result reveals the relations between $\eta(\tilde{x})$ and $\eta_p(\tilde{x})$ for $p = 1, 2, \infty$. The proof is left as an exercise.

Theorem 1.9.5. *Let $\eta_p(\tilde{x})$ $(p = 1, 2, \infty)$ be the backward errors defined by (1.9.5) with $\|\cdot\| = \|\cdot\|_p$, and $\eta^*(\tilde{x})$ be the backward error defined by (1.9.6), where we take $\alpha_1 = \nu(a_1)$, $\alpha_2 = \nu(a_2)$, $\alpha = \nu(a_1, a_2)$, and assume that $\nu(a_1) \neq 0$ and $\nu(a_2) \neq 0$. Then*

$$\begin{aligned} &\eta_\infty(\tilde{x}) \le \eta_1(\tilde{x}) \le 2\eta_\infty(\tilde{x}), \\ &\frac{1}{\sqrt{2}}\eta_1(\tilde{x}) \le \eta_2(\tilde{x}) \le \eta_1(\tilde{x}), \\ &\frac{1}{\sqrt{2}}\eta_2(\tilde{x}) \le \eta_\infty(\tilde{x}) \le \eta_2(\tilde{x}), \end{aligned} \tag{1.9.8}$$

and

$$\frac{\min\{\nu(a_1), \nu(a_2)\}}{\nu(a_1, a_2)}\eta_\infty(\tilde{x}) \le \eta^*(\tilde{x}) \le \frac{\max\{\nu(a_1), \nu(a_2)\}}{\nu(a_1, a_2)}\eta_1(\tilde{x}). \tag{1.9.9}$$

Remark 1.9.6. From the definitions (1.9.5)–(1.9.7) we see that every backward error of the problem $r(a_1, a_2; x) = 0$ with respect to $\tilde{x}$ is defined with respect to a particular class of backward perturbations in a_1 and a_2. Therefore, different backward errors have different meanings, and the values of two different backward errors of the problem $r(a_1, a_2; x) = 0$ with respect to the same $\tilde{x}$ may be quite different. For example, $\eta^*(\tilde{x})$ may be quite different from $\eta_p(\tilde{x})$ for $p = 1, 2, \infty$. In fact, the first inequality of (1.9.9) implies that if any one of $\nu(a_1)$ and $\nu(a_2)$ is much smaller than the other, then $\eta^*(\tilde{x})$ is bounded from below by $\epsilon\eta_\infty(\tilde{x})$, where $\epsilon > 0$ is a very small positive scalar. This means that in some cases the quantity $\eta^*(\tilde{x})$ may be much smaller than $\eta_p(\tilde{x})$ for $p = 1, 2, \infty$. Note that a very small backward error $\eta^*(\tilde{x})$ may be uninformative for the following reason: In the case that there is a great disparity between $\nu(a_1)$ and $\nu(a_2)$, while the optimal backward perturbation $(\Delta a_{1*}, \Delta a_{2*})$ is very small compared with (a_1, a_2), it may be making a large relative perturbation in the small one of a_1 and a_2.

Notes and References

NR 1.9–1. This section is based on the author's Technical Report " Optimal backward perturbation bounds for linear systems and linear least squares problems", UMINF 96.15, ISSN-0348-0542, Department of Computing Science, Umeå University, 1996.

NR 1.9–2. The earlest results on computable formulas of backward errors for linear systems are given by Oettli and Prager [80], and Rigal and Gaches [89].

NR 1.9–3. For historical comments on the development of backward error analysis and backward errors in numerical analysis, see Higham [52, §1.21 and §19.7]. For the importance of the study of computable formulas of backward errors, see Stewart [94] and Higham [50].

NR 1.9–4. Let $x = \phi(a)$ be a solution of a matrix problem $r(a; x) = 0$ with $a \in \mathcal{A}$ (or $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$), and let $\tilde{x}$ be an approximation of x. Moreover, let $c(x)$ be the condition number defined by (1.8.1) (or (1.8.5)), and $\eta(\tilde{x})$ be the backward error defined by (1.9.1) (or (1.9.5)). Then the relation (1.8.2) (or (1.8.7)) shows that in first order approximation we have

$$\frac{\mu(\tilde{x} - x)}{\xi} \leq c(x)\eta(\tilde{x}). \tag{1.9.10}$$

Similarly, from (1.8.6), (1.8.8) and (1.9.6) it follows that in first order approximation we have

$$\frac{\mu(\tilde{x} - x)}{\xi} \leq c^*(x)\eta^*(\tilde{x}). \tag{1.9.11}$$

One way to interpret the relation (1.9.10) (or (1.9.11)) is to say that the approximation $\tilde{x}$ may not be close to x if the condition number $c(x)$ (or $c^*(x)$) is very large, even if the approximate solution $\tilde{x}$ has a small backward error $\eta(\tilde{x})$ (or $\eta^*(\tilde{x})$).

NR 1.9–5. For the study of componentwise backward errors and structured backward errors of some numerical linear algebra problems, see D. Higham and N. Higham [45], [46], N. Higham [52, §7.2], and Chaitin-Chatelin and Frayssé [17, Chapter 5].

NR 1.9–6. Let $r(a; x) = 0$ be a matrix computation problem, and let $\tilde{x}$ be an approximation of the exact solution x to the problem. If the optimal backward perturbation Δa_{opt} is found, then we can apply an appropriate forward perturbation result to the perturbation Δa_{opt}, and obtain an upper bound for $\nu(\tilde{x} - x)$. Generally speaking, the optimal backward perturbation Δa_{opt} can be expressed by the *residual* $r(a; \tilde{x})$, so the obtained upper bound for $\nu(\tilde{x} - x)$ is usually in the form of *residual bound*, and the upper bound is called a residual bound. Note that there are different ways to obtain upper bounds for $\nu(\tilde{x} - x)$; but usually, the upper bounds are dependent on the residual $r(a; \tilde{x})$.

Chapter 2

Eigenvalue Problems

This chapter is devoted to the eigenvalue problem $Ax = \lambda x$, where $A \in \mathcal{C}^{n\times n}$. We begin in §2.1 with perturbation expansions for eigenvalues and invariant subspaces. On the basis of the results of §2.1 we derive explicit expressions of condition numbers for eigenvalues and invariant subspaces in §2.2. In §2.3 we present perturbation bounds for invariant subspaces. In §2.4 we treat backward errors and residual bounds. The chapter concludes with a section on Hermitian matrices.

2.1 Perturbation Expansions

2.1.1 Simple Eigenvalues

Let $A \in \mathcal{C}^{n\times n}$. If

$$Ax = \lambda x$$

for $\lambda \in \mathcal{C}$ and a nonzero $x \in \mathcal{C}^n$, then λ is called an *eigenvalue* of A, and x a *right eigenvector* of A associated with λ. Usually, we call x an *eigenvector* of A associated with λ. The corresponding nonzero solution $y \in \mathcal{C}^n$ of the equation

$$y^H A = \lambda y^H$$

is called a *left eigenvector* of A associated with λ.

Let $p = (p_1, \ldots, p_N)^T \in \mathcal{C}^N$ (or $\mathcal{R}^N$), and let $A(p) = (\alpha_{jk}(p)) \in \mathcal{C}^{n\times n}$ (or $\mathcal{R}^{n\times n}$) be an analytic matrix-valued function in some neighborhood $\mathcal{B}(p^*)$ of the point p^*. For simplicity, we assume $p^* = 0$, the origin of $\mathcal{C}^N$ (or $\mathcal{R}^N$). By Definition 1.6.1, the function $A(p)$ can be expressed by

$$A(p) = A(0) + E(p), \qquad E(p) = (\epsilon_{jk}(p)),$$

where

$$\epsilon_{jk}(p) = \sum_{r=1}^{\infty} \sum_{\sum t_i = r} \alpha^{(jk)}_{t_1\cdots t_N} p_1^{t_1} \cdots p_N^{t_N}, \quad 1 \le j, k \le N, \quad p \in \mathcal{B}(0),$$

and $\sum t_i = t_1 + \cdots + t_N$.

Let λ be a *simple* eigenvalue of $A(0)$, and x, y be associated right and left eigenvectors satisfying $y^H x = 1$. Then, as a consequence, there are $X_2, Y_2 \in \mathcal{C}^{n\times(n-1)}$ such that the matrices

$$X = (x, X_2), \quad Y = (y, Y_2) \tag{2.1.1}$$

satisfy

$$Y^H X = I \tag{2.1.2}$$

and

$$Y^H A(0) X = \begin{pmatrix} \lambda & 0 \\ 0 & A_2 \end{pmatrix}, \quad \lambda \notin \lambda(A_2). \tag{2.1.3}$$

First applying the implicit function theorem we prove the following result.

Theorem 2.1.1 *Let $p \in \mathcal{C}^N$, and let $A(p) \in \mathcal{C}^{n\times n}$ be an analytic function of p in some neighborhood $\mathcal{B}(0)$ of the origin. Assume that λ is a simple eigenvalue of $A(0)$, and x, y are associated right and left eigenvectors satisfying $y^H x = 1$. Moreover, assume that the relation (2.1.3) holds, in which X and Y are the matrices of (2.1.1) and satisfy (2.1.2)-(2.1.3). Then*

1) there exists a simple eigenvalue $\lambda(p)$ of $A(p)$ which is an analytic function of p in some neighborhood $\mathcal{B}_0$ of the origin, and $\lambda(0) = \lambda$;

2) the function $\lambda(p)$ has a power series expansion at $p = 0$ of the form

$$\lambda(p) = \lambda + \sum_{j=1}^{N} \left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} p_j + \frac{1}{2}\sum_{j,k=1}^{N}\left(\frac{\partial^2 \lambda(p)}{\partial p_j \partial p_k}\right)_{p=0} p_j p_k + \cdots, \qquad p \in \mathcal{B}_0,$$

where

$$\left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} = y^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x, \tag{2.1.4}$$

and

$$\begin{aligned}\left(\frac{\partial^2 \lambda(p)}{\partial p_j \partial p_k}\right)_{p=0} = y^H \left(\frac{\partial^2 A(p)}{\partial p_j \partial p_k}\right)_{p=0} x + y^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} \Omega \left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} x \\ + y^H \left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} \Omega \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x,\end{aligned} \tag{2.1.5}$$

in which

$$\Omega = X_2(\lambda I - A_2)^{-1} Y_2^H. \tag{2.1.6}$$

Proof. 1) By the hypotheses there are $X, Y \in \mathcal{C}^{n\times n}$ such that the relations (2.1.1)–(2.1.3) hold. For $p \in \mathcal{B}(0)$ we set

$$\tilde{A}(p) = Y^H A(p) X = \begin{pmatrix} \tilde{a}_{11}(p) & \tilde{a}_{12}(p) \\ \tilde{a}_{21}(p) & \tilde{A}_{22}(p) \end{pmatrix}, \quad \tilde{a}_{11}(p) \in \mathcal{C}, \tag{2.1.7}$$

and introduce a vector-valued function

$$f(z,p) = \tilde{a}_{21}(p) - \tilde{a}_{11}(p)z + \tilde{A}_{22}(p)z - z\tilde{a}_{12}(p)z, \tag{2.1.8}$$

where

$$f = (f_1, \dots, f_{n-1})^T, \quad z = (\zeta_1, \dots, \zeta_{n-1})^T \in \mathcal{C}^{n-1}, \quad p \in \mathcal{B}(0).$$

Observe that the vector-valued function $f(z,p)$ is analytic for $z \in \mathcal{C}^{n-1}$ and $p \in \mathcal{B}(0)$, $f_j(0,0) = 0$ for $j = 1, \dots, n-1$, and

$$\left(\frac{\partial(f_1, \dots, f_{n-1})}{\partial(\zeta_1, \dots, \zeta_{n-1})}\right)_{z=0,\, p=0} = \det(A_2 - \lambda I) \neq 0.$$

Hence, by the implicit function theorem (Theorem 1.6.2) the equation

$$(f_1(z,p), \dots, f_{n-1}(z,p)) = (0, \dots, 0) \tag{2.1.9}$$

has a unique analytic solution $z = z(p) \in \mathcal{C}^{n-1}$ in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin, and $z(0) = 0$.

From (2.1.7)–(2.1.9) it follows that for $p \in \mathcal{B}_0$ we have

$$\begin{pmatrix} 1 & 0 \\ z(p) & I \end{pmatrix}^{-1} \tilde{A}(p) \begin{pmatrix} 1 & 0 \\ z(p) & I \end{pmatrix} = \begin{pmatrix} \lambda(p) & * \\ 0 & * \end{pmatrix}, \tag{2.1.10}$$

where

$$\lambda(p) = \tilde{a}_{11}(p) + \tilde{a}_{12}(p)z(p). \tag{2.1.11}$$

The relation (2.1.10) shows that $\lambda(p)$ is an eigenvalue of $A(p)$, and the eigenvalue is simple provided that the neighborhood $\mathcal{B}_0$ is sufficiently small. Moreover, the analyticity of the functions $\tilde{a}_{11}(p)$, $\tilde{a}_{12}(p)$ and $z(p)$ implies that $\lambda(p)$ is an analytic function of $p \in \mathcal{B}_0$, and from (2.1.11) it follows that $\lambda(0) = \lambda$.

2) From (2.1.11) and $\tilde{a}_{12}(0)^T = z(0) = 0$ we obtain

$$\left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} = \left(\frac{\partial \tilde{a}_{11}(p)}{\partial p_j}\right)_{p=0}, \tag{2.1.12}$$

$$\begin{aligned} \left(\frac{\partial^2 \lambda(p)}{\partial p_j \partial p_k}\right)_{p=0} &= \left(\frac{\partial^2 \tilde{a}_{11}(p)}{\partial p_j \partial p_k}\right)_{p=0} + \left(\frac{\partial \tilde{a}_{12}(p)}{\partial p_j}\right)_{p=0} \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0} \\ &+ \left(\frac{\partial \tilde{a}_{12}(p)}{\partial p_k}\right)_{p=0} \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}. \end{aligned} \tag{2.1.13}$$

Moreover, from (2.1.7) we obtain

$$\begin{aligned}\left(\frac{\partial \tilde{a}_{11}(p)}{\partial p_j}\right)_{p=0} &= y^H \left(\frac{\partial\ A(p)}{\partial p_j}\right)_{p=0} x,\\ \left(\frac{\partial^2 \tilde{a}_{11}(p)}{\partial p_j \partial p_k}\right)_{p=0} &= y^H \left(\frac{\partial^2 A(p)}{\partial p_j \partial p_k}\right)_{p=0} x, \qquad (2.1.14)\\ \left(\frac{\partial \tilde{a}_{12}(p)}{\partial p_j}\right)_{p=0} &= y^H \left(\frac{\partial\ A(p)}{\partial p_j}\right)_{p=0} X_2.\end{aligned}$$

Combining (2.1.12) with the first formula of (2.1.14) shows (2.1.4). From (2.1.13) and (2.1.14) we see that for obtaining the formula (2.1.5) we only need to find an explicit expression of $\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}$.

It is known that $z(p)$ is the unique analytic solution of $f(z,p)=0$ in $\mathcal{B}_0$, where $f(z,p)$ is defined by (2.1.8); i.e., $z(p)$ satisfies the equation

$$\tilde{a}_{21}(p) - \tilde{a}_{11}(p)z(p) + \tilde{A}_{22}(p)z(p) - \tilde{a}_{12}(p)z(p)z(p) = 0, \quad p \in \mathcal{B}_0. \qquad (2.1.15)$$

Differentiating (2.1.15) at $p=0$ gives

$$\begin{aligned}\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} &= (\lambda I - A_2)^{-1} \left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_j}\right)_{p=0}\\ &= (\lambda I - A_2)^{-1} Y_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x.\end{aligned} \qquad (2.1.16)$$

Substituting (2.1.14) and (2.1.16) into (2.1.13) shows the formula (2.1.5). □

Remark 2.1.2. From (2.1.10), (2.1.7) and (2.1.2) we get

$$A(p)x(p) = \lambda(p)x(p), \quad p \in \mathcal{B}_0, \qquad (2.1.17)$$

where $x(p)$ is defined by

$$x(p) = X \begin{pmatrix} 1 \\ z(p) \end{pmatrix}. \qquad (2.1.18)$$

The relation (2.1.17) shows that the vector $x(p)$ is an eigenvector of $A(p)$ associated with $\lambda(p)$, and the expression (2.1.18) shows that the eigenvector is an analytic function of $p \in \mathcal{B}_0$ satisfying $x(0) = x$. Moreover, the relations (2.1.18) and (2.1.16) imply that the eigenvector $x(p)$ has the expansion of the form

$$x(p) = x + \sum_{j=1}^{N} \left(\frac{\partial x(p)}{\partial p_j}\right)_{p=0} p_j + \cdots, \quad p \in \mathcal{B}_0,$$

where

$$\left(\frac{\partial x(p)}{\partial p_j}\right)_{p=0} = \Omega \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x.$$

Remark 2.1.3. Without the assumption of $y^H x = 1$ in Theorem 2.1.1, the formula (2.1.4) becomes

$$\left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} = \frac{y^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x}{y^H x}. \tag{2.1.19}$$

Example 2.1.4. Consider the matrix

$$A(p) = \begin{pmatrix} 3 & \frac{2}{1+p_1+p_2} \\ -\frac{4}{1+p_1-p_2} & -3 \end{pmatrix}, \quad (p_1, p_2)^T = p \in \mathcal{R}^2.$$

Obviously, $A(p)$ is an analytic matrix-valued function of p in a small neighborhood of the origin of $\mathcal{R}^2$. Moreover,

$$A(0) = \begin{pmatrix} 3 & 2 \\ -4 & -3 \end{pmatrix},$$

and the real matrices

$$X = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \equiv (x_1, x_2) \quad \text{and} \quad Y = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \equiv (y_1, y_2)$$

satisfy

$$Y^T A(0) X = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \equiv \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

and

$$Y^T X = I.$$

Observe that

$$\left(\frac{\partial A(p)}{\partial p_1}\right)_{p=0} = \begin{pmatrix} 0 & -2 \\ 4 & 0 \end{pmatrix}, \quad \left(\frac{\partial A(p)}{\partial p_2}\right)_{p=0} = \begin{pmatrix} 0 & -2 \\ -4 & 0 \end{pmatrix},$$

$$\left(\frac{\partial^2 A(p)}{\partial p_1^2}\right)_{p=0} = \begin{pmatrix} 0 & 4 \\ -8 & 0 \end{pmatrix}, \quad \left(\frac{\partial^2 A(p)}{\partial p_2^2}\right)_{p=0} = \begin{pmatrix} 0 & 4 \\ -8 & 0 \end{pmatrix},$$

and

$$\left(\frac{\partial^2 A(p)}{\partial p_1 \partial p_2}\right)_{p=0} = \begin{pmatrix} 0 & 4 \\ 8 & 0 \end{pmatrix}.$$

Hence, if $\lambda_1(p)$ and $\lambda_2(p)$ denote the eigenvalues of $A(p)$, then by the formulas (2.1.4)–(2.1.6) we have

$$\left(\frac{\partial \lambda_1(p)}{\partial p_1}\right)_{p=0} = 8, \quad \left(\frac{\partial \lambda_1(p)}{\partial p_2}\right)_{p=0} = 0,$$

$$\left(\frac{\partial \lambda_2(p)}{\partial p_1}\right)_{p=0} = -8, \quad \left(\frac{\partial \lambda_2(p)}{\partial p_2}\right)_{p=0} = 0,$$

and

$$\left(\frac{\partial^2\lambda_1(p)}{\partial p_1^2}\right)_{p=0} = -88, \quad \left(\frac{\partial^2\lambda_1(p)}{\partial p_1 \partial p_2}\right)_{p=0} = 0, \quad \left(\frac{\partial^2\lambda_1(p)}{\partial p_2^2}\right)_{p=0} = -8,$$

$$\left(\frac{\partial^2\lambda_2(p)}{\partial p_1^2}\right)_{p=0} = 88, \quad \left(\frac{\partial^2\lambda_2(p)}{\partial p_1 \partial p_2}\right)_{p=0} = 0, \quad \left(\frac{\partial^2\lambda_2(p)}{\partial p_2^2}\right)_{p=0} = 8.$$

Consequently, $\lambda_1(p)$ and $\lambda_2(p)$ have the expansions

$$\lambda_1(p) = 1 + 8p_1 - 44p_1^2 - 4p_2^2 + O(\|p\|_2^3) \tag{2.1.20}$$

and

$$\lambda_2(p) = -1 - 8p_1 + 44p_1^2 + 4p_2^2 + O(\|p\|_2^3) \tag{2.1.21}$$

as $p \to 0$.

Note that the eigenvalues $\lambda_1(p)$ and $\lambda_2(p)$ have the explicit expressions

$$\lambda_1(p) = \sqrt{9 - \frac{8}{(1+p_1)^2 - p_2^2}}, \qquad \lambda_2(p) = -\sqrt{9 - \frac{8}{(1+p_1)^2 - p_2^2}}.$$

From the expressions we can also obtain the second order perturbation expansions (2.1.20) and (2.1.21).

2.1.2 Simple Invariant Subspaces

Let λ be an eigenvalue of $A \in \mathcal{C}^{n\times n}$, and $x \in \mathcal{C}^n$ be an associated eigenvector. Then the one-dimensional subspace $\mathcal{R}(x)$ satisfies $A\mathcal{R}(x) \subset \mathcal{R}(x)$, and which is called a one-dimensional invariant subspace of A. This definition extends in a natural way to higher dimensions.

Let $A \in \mathcal{C}^{n\times n}$ and let $\mathcal{X}_1 \subset \mathcal{C}^n$. If

$$\dim(\mathcal{X}_1) = l \quad \text{and} \quad A\mathcal{X}_1 \subset \mathcal{X}_1,$$

then $\mathcal{X}_1$ is said to be an l-dimensional *(right) invariant subspace* of A.

The invariant subspace $\mathcal{X}_1$ may be equivalently defined by $\mathcal{X}_1 = \mathcal{R}(X_1)$ with $X_1 \in \mathcal{C}^{n\times l}$ satisfying

$$\operatorname{rank}(X_1) = l \quad \text{and} \quad AX_1 = X_1A_1$$

for some $A_1 \in \mathcal{C}^{l\times l}$. The matrix A_1 may be called the *(right) eigenmatrix* of A associated with X_1.

Let $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ with $X_1 \in \mathcal{U}^{n\times l}$ such that

$$X^HAX = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad A_{11} \in \mathcal{C}^{l\times l}. \tag{2.1.22}$$

Then the invariant subspace $\mathcal{X}_1 = \mathcal{R}(X_1)$ is called a *simple* invariant subspace of A if $\lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset$. In this chapter we only consider simple invariant subspaces. We now prove the following perturbation expansion theorem.

Theorem 2.1.5. *Let $A \in \mathcal{C}^{n\times n}$, and let $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ with $X_1 \in \mathcal{U}^{n\times l}$ such that the relation (2.1.22) holds, and $\lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset$. Moreover, let $\mathcal{X}_1 = \mathcal{R}(X_1)$, for $M \in \mathcal{C}^{n\times n}$ let*

$$X^H M X = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

and define the linear operator $\mathbf{T} : \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{T}P = PA_{11} - A_{22}P, \quad P \in \mathcal{C}^{(n-l)\times l}. \tag{2.1.23}$$

Then

1) there exists a unique l-dimensional simple invariant subspace $\mathcal{X}_1(\tau)$ of $A + \tau M$ such that $\mathcal{X}_1(0) = \mathcal{X}_1$, and the basis vectors $x_1(\tau), \ldots, x_l(\tau)$ of $\mathcal{X}_1(\tau)$ may be defined to be analytic functions of τ in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{C}$;

2) the analytic matrix-valued function $X_1(\tau) = (x_1(\tau), \ldots, x_l(\tau))$ has the perturbation expansion

$$X_1(\tau) = X_1 + X_2 \sum_{j=1}^{\infty} K_j \tau^j, \quad \tau \in \mathcal{B}(0), \tag{2.1.24}$$

in which

$$K_1 = \mathbf{T}^{-1} M_{21},$$

$$K_2 = \mathbf{T}^{-1}[M_{22}K_1 - K_1 M_{11} - K_1 A_{12} K_1],$$

$$K_j = \mathbf{T}^{-1}\left[M_{22}K_{j-1} - K_{j-1}M_{11} - \sum_{k=1}^{j-2} K_{j-1-k} M_{12} K_k - \sum_{k=1}^{j-1} K_{j-k} A_{12} K_k\right],$$

$$j \geq 3. \tag{2.1.25}$$

Proof. 1) Let

$$A(\tau) = A + \tau M$$

and

$$\tilde{A}(\tau) = X^H A(\tau) X = \begin{pmatrix} \tilde{A}_{11}(\tau) & \tilde{A}_{12}(\tau) \\ \tilde{A}_{21}(\tau) & \tilde{A}_{22}(\tau) \end{pmatrix}, \quad \tilde{A}_{11}(\tau) \in \mathcal{C}^{l\times l}, \tag{2.1.26}$$

where

$$\tilde{A}_{jk}(\tau) = A_{jk} + \tau M_{jk}, \quad 1 \le j,k \le 2, \qquad A_{21} = 0. \tag{2.1.27}$$

For $Z \in \mathcal{C}^{(n-l)\times l}$ and $\tau \in \mathcal{C}$ define the function F by

$$F(Z,\tau) = \tilde{A}_{21}(\tau) - Z\tilde{A}_{11}(\tau) + \tilde{A}_{22}(\tau)Z - Z\tilde{A}_{12}(\tau)Z, \tag{2.1.28}$$

and let

$$f = \mathrm{vec}(F), \qquad z = \mathrm{vec}(Z).$$

Observe that by (2.1.28) and the hypothesis (2.1.22), we have

$$\left(\frac{\partial f}{\partial z}\right)_{z=0,\ \tau=0} = \det(I \otimes A_{22} - A_{11}^T \otimes I) \ne 0.$$

Hence, by the implicit function theorem (Theorem 1.6.2) the equation

$$F(Z,\tau) = 0$$

has a unique analytic solution $Z = Z(\tau)$ of τ in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{C}$ satisfying $Z(0) = 0$, or equivalently, we have

$$\begin{pmatrix} I & 0 \\ Z(\tau) & I \end{pmatrix}^{-1} \tilde{A}(\tau) \begin{pmatrix} I & 0 \\ Z(\tau) & I \end{pmatrix} = \begin{pmatrix} A_1(\tau) & \tilde{A}_{12}(\tau) \\ 0 & A_2(\tau) \end{pmatrix}, \tag{2.1.29}$$

where

$$A_1(\tau) = \tilde{A}_{11}(\tau) + \tilde{A}_{12}(\tau)Z(\tau), \qquad A_2(\tau) = \tilde{A}_{22}(\tau) - Z(\tau)\tilde{A}_{12}(\tau),$$

and $\lambda(A_1(\tau)) \bigcap \lambda(A_2(\tau)) = \emptyset$ provided that the neighborhood $\mathcal{B}(0)$ is sufficiently small.

Define

$$X_1(\tau) = X \begin{pmatrix} I \\ Z(\tau) \end{pmatrix}. \tag{2.1.30}$$

Then from (2.1.29) and (2.1.26)

$$A(\tau)X_1(\tau) = X_1(\tau)A_1(\tau).$$

Thus, we have proved that $\mathcal{X}_1(\tau) \equiv \mathcal{R}(X_1(\tau))$ is the unique l-dimensional simple invariant subspace of $A(\tau)$ in $\mathcal{B}(0)$ satisfying $\mathcal{X}_1(0) = \mathcal{X}_1$, and $X_1(\tau)$ is an analytic matrix-valued function of $\tau \in \mathcal{B}(0)$.

2) Substituting the relations of (2.1.27) into $F(Z(\tau),\tau) = 0$, we get the basic equation for $Z(\tau)$:

$$Z(\tau)(A_{12}+\tau M_{12})Z(\tau)+Z(\tau)(A_{11}+\tau M_{11})-(A_{22}+\tau M_{22})Z(\tau)-\tau M_{21} = 0, \tag{2.1.31}$$

where $\tau \in \mathcal{B}(0)$.

Differentiating (2.1.31) at $\tau = 0$, and writing

$$Z^{(j)} = \left(\frac{d^j Z(\tau)}{d\tau^j}\right)_{\tau=0}, \quad j = 1, 2, \ldots,$$

we get

$$\begin{aligned}
&\mathbf{T}Z^{(1)} = M_{21}, \\
&\mathbf{T}Z^{(2)} = 2\left[M_{22}Z^{(1)} - Z^{(1)}M_{11} - Z^{(1)}A_{12}Z^{(1)}\right], \\
&\mathbf{T}Z^{(j)} = j\left[M_{22}Z^{(j-1)} - Z^{(j-1)}M_{11} - \sum_{k=1}^{j-2}\begin{pmatrix} j-1 \\ k \end{pmatrix} Z^{(j-1-k)}M_{12}Z^{(k)}\right] \qquad (2.1.32) \\
&\qquad -\sum_{k=1}^{j-1}\begin{pmatrix} j \\ k \end{pmatrix} Z^{(j-k)}A_{12}Z^{(k)}, \qquad j \geq 3,
\end{aligned}$$

where $\mathbf{T}$ is the linear operator defined by (2.1.23), and $\begin{pmatrix} j \\ k \end{pmatrix}$ are binomial coefficients.

Since $\lambda(A_{11} \bigcap \lambda(A_{22}) = \emptyset$, the operator $\mathbf{T}$ is invertible. Define

$$K_k = \frac{1}{k!}Z^{(k)}, \qquad k = 1, 2, \ldots.$$

Then from (2.1.32) we get the relations (2.1.25) and the power series expansion of $Z(\tau)$ at $\tau = 0$:

$$Z(\tau) = \sum_{j=1}^{\infty}\frac{1}{j!}Z^{(j)}\tau^j = \sum_{j=1}^{\infty}K_j\tau^j.$$

Substituting it into (2.1.30) shows (2.1.24). □

The following result, as a corollary of Theorem 2.1.5, gives a modified form of the first order perturbation expansion of a simple invariant subspace.

Corollary 2.1.6. *Let $A, X, A_{11}, A_{22}, \mathcal{X}_1$ and $\mathbf{T}$ be as in Theorem 2.1.5, and for $E \in \mathcal{C}^{n\times n}$ let*

$$X^H E X = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \qquad E_{11} \in \mathcal{C}^{l\times l}.$$

If $\|E\|_F$ is sufficiently small, then there exists a unique l-dimensional simple invariant subspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A + E$ such that $\tilde{X}_1$ has the expansion

$$\tilde{X}_1 = X_1 + X_2 Z_1 + O(\|E\|_F^2), \qquad (2.1.33)$$

where $E \to 0$, and $Z_1 \in \mathcal{C}^{(n-l)\times l}$ is defined by

$$Z_1 = \mathbf{T}^{-1}E_{21}. \tag{2.1.34}$$

Let $\mathbf{T}$ be the linear operator defined by (2.1.23). Using the Kronecker product and vec operator, the matrix representation T of the operator $\mathbf{T}$ can be expressed by

$$T = A_{11}^T \otimes I_{n-l} - I_l \otimes A_{22}, \tag{2.1.35}$$

and the relation (2.1.34) can be written

$$\mathrm{vec}(Z_1) = T^{-1}\mathrm{vec}(E_{21}). \tag{2.1.36}$$

Example 2.1.7. Let $A \in \mathcal{C}^{n\times n}$ be a *normal matrix*, i.e., A satisfies $A^HA = AA^H$. Then there is a matrix $X = (x_1, X_2) \in \mathcal{U}^{n\times n}$ with $X_2 = (x_2, \ldots, x_n)$ such that

$$X^HAX = \mathrm{diag}(\lambda_1, \Lambda_2), \qquad \Lambda_2 = \mathrm{diag}(\lambda_2, \ldots, \lambda_n).$$

Let λ_1 be a simple eigenvalue of A, and for $M \in \mathcal{C}^{n\times n}$ let

$$X^HMX = \begin{pmatrix} \mu_{11} & m_2^T \\ m_1 & M_{22} \end{pmatrix} = (\mu_{ij}), \qquad \mu_{11} \in \mathcal{C}.$$

Then by Theorem 2.1.1 and Theorem 2.1.5, we have the following conclusions:

i) There is a simple eigenvalue $\lambda_1(\tau)$ of $A + \tau M$ which is an analytic function of τ in some neighborhood $\mathcal{B}_0$ of the origin of $\mathcal{C}$, and $\lambda_1(0) = \lambda_1$;

ii) The function $\lambda_1(\tau)$ has a power series expansion at $\tau = 0$ of the form

$$\lambda_1(\tau) = \lambda_1 + x_1^HMx_1\tau + x_1^HMX_2(\lambda_1I - \Lambda_2)^{-1}X_2^HMx_1\tau^2 + \cdots, \quad \tau \in \mathcal{B}_0;$$

iii) There exists a unique 1-dimensional simple invariant subspace $\mathcal{X}_1(\tau)$ of $A+\tau M$ such that $\mathcal{X}_1(0) = \mathcal{R}(x_1)$, and the basis vector $x_1(\tau)$ of $\mathcal{X}_1(\tau)$ may be defined to be an analytic function of τ in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{C}$;

iv) The analytic vector-valued function $x_1(\tau)$ has the perturbation expansion

$$x_1(\tau) = x_1 + X_2\sum_{j=1}^{\infty} K_j\tau^j, \qquad \tau \in \mathcal{B}(0),$$

where the vectors $K_j \in \mathcal{C}^{n-1}$ are defined by (2.1.25). In particular, the first-order term of the perturbation in x_1 is given by

$$\begin{aligned} X_2K_1\tau &= X_2(\lambda_1I - \Lambda_2)^{-1}m_1\tau \\ &= \left(\frac{\mu_{21}x_2}{\lambda_1 - \lambda_2} + \frac{\mu_{31}x_3}{\lambda_1 - \lambda_3} + \cdots + \frac{\mu_{n1}x_n}{\lambda_1 - \lambda_n}\right)\tau, \end{aligned}$$

and the second-order term is given by

$$\begin{aligned} X_2K_2\tau^2 &= X_2(\lambda_1 I - \Lambda_2)^{-1}(M_{22} - \mu_{11}I)(\lambda_1 I - \Lambda_2)^{-1}m_1\tau^2 \\ &= \left(\sum_{j=2}^{n} \frac{(\mu_{2j} - \mu_{11})\mu_{j1}}{\lambda_1 - \lambda_j} \cdot \frac{x_2}{\lambda_1 - \lambda_2} + \sum_{j=2}^{n} \frac{(\mu_{3j} - \mu_{11})\mu_{j1}}{\lambda_1 - \lambda_j} \cdot \frac{x_3}{\lambda_1 - \lambda_3} \right. \\ &\qquad \left. + \cdots + \sum_{j=2}^{n} \frac{(\mu_{nj} - \mu_{11})\mu_{j1}}{\lambda_1 - \lambda_j} \cdot \frac{x_n}{\lambda_1 - \lambda_n}\right)\tau^2. \end{aligned}$$

Notes and References

NR 2.1–1. This section is based on Sun [102] and [109]. The basic tool is the implicit function theorem (Theorem 1.6.2). The approach of setting the (2,1) submatrices of the equations (2.1.10) and (2.1.29) equal to zero to obtain nonlinear equations for a basis is due to Stewart [91]. The technique described in this section will be used in chapters 3 and 4 for deriving perturbation expansions of singular values, singular subspaces, generalized eigenvalues, and deflating subspaces. This technique is also used by Chu [21] for studying the mean of *multiple* eigenvalues and the corresponding invariant subspaces, and by Andrew, Chu and Lancaster [1] for studying eigenvalues and eigenvectors of *matrix functions*. Recently, Lin and Sun [71] study the eigenproblem of *periodic matrix pairs* by using the same technique.

NR 2.1–2. Let $\lambda_1(p)$ and $x_1(p)$ be as in Theorem 2.1.1. By using (2.1.11), (2.1.7) and (2.1.15), we can derive formulas of

$$\left(\frac{\partial^k \lambda_1(p)}{\partial p_1^{k_1} \cdots \partial p_N^{k_N}}\right)_{p=0}, \qquad k_1 + \cdots + k_N = k$$

for $k \geq 3$. Moreover, by using (2.1.8), (2.1.7) and (2.1.15), we can derive formulas of

$$\left(\frac{\partial^k x_1(p)}{\partial p_1^{k_1} \cdots \partial p_N^{k_N}}\right)_{p=0}, \qquad k_1 + \cdots + k_N = k$$

for $k \geq 2$ (see Sun [102] and Chu [21]).

NR 2.1–3. A general matrix may only depend analytically on one parameter. For such matrices, the derivatives of eigenvalues and associated eigenvectors are studied by Rellich [86], Lancaster [66], and perturbation expansions of eigenvalues and associated *total projections* are studied by Kato [63] by using different techniques from that described in this section.

NR 2.1–4. Explicit expressions of the *derivatives* of the eigenvalues and eigenvectors of a general matrix depending analytically on one or several parameters have important practical and theoretical applications, such as in the perturbation theory of eigenvalue problems (see Kato [63]), *inverse eigenvalue problems* (see Sun [103], Xu [134]) and *control system* and *engineering designs* (see Crossley and Porter [24]). Numerical methods for computing the derivatives are studied by many authors, see, e.g., Andrew, Hoog and Tan [2] and the references contained therein.

NR 2.1–5. There are some situations in which the use of the *group inverse* is natural in the formulation of explicit formulas for derivatives of the eigenvectors. This is demonstrated by Deutsch and Neumann [31], and Meyer and Stewart [77].

NR 2.1–6. Large vibration systems and control systems are frequently dependent on many physical and geometrical parameters $p_1, \ldots, p_N$, and it will generally happen that several eigenvalues overlap at some points $(p_1, \ldots, p_N)$. It is worth pointing out that the local behavior of the eigenvalues dependent on *several* parameters at the overlapping points is different from that of the eigenvalues only dependent on *one* parameter. Lancaster and Tismenetsky [67, Chapter 11, Theorem 1] show that if $A(\xi)$ is an analytic function of one parameter ξ in a neighborhood of $\xi = 0$, and λ is a *nondefective multiple eigenvalue* of $A(0)$ with multiplicity r, then $A(\xi)$ has r eigenvalues $\lambda_1(\xi)$, ..., $\lambda_r(\xi)$ such that $\lambda_j(0) = 0$ for all j, and each $\lambda_j(\xi)$ has derivative $\frac{d\lambda_j(\xi)}{d\xi}$ at $\xi = 0$. However, if $A(p)$ is an analytic function of several parameters, then the situation becomes complicated. For example, consider the matrix

$$A(p) = \begin{pmatrix} 1 + 2p_1 + p_2 & -2p_1 \\ p_2 & 1 - p_2 \end{pmatrix}, \qquad p = (p_1, p_2)^T \in \mathcal{R}^2.$$

The elements of $A(p)$ are real analytic functions of $p \in \mathcal{R}^2$, the matrix $A(0)$ has the nondefective multiple eigenvalue 1 with multiplicity 2, and the eigenvalues of $A(p)$ are

$$\lambda_1(p) = 1 + p_1 + \sqrt{p_1^2 + p_2^2}, \qquad \lambda_2(p) = 1 + p_1 - \sqrt{p_1^2 + p_2^2}.$$

Obviously no arrangement of these eigenvalues could make them that the rearranged eigenvalues have partial derivatives at $p = 0$. Sun [113] studies the existence and expressions of the *directional derivatives* of nondefective multiple eigenvalues of a general matrix depending analytically on several parameters. The result of [113] are used to discuss the condition numbers of nondefective multiple eigenvalues by Sun [114] and [117].

NR 2.1–7. Lidskii [70] establishes a perturbation theory for eigenvalues of matrices with arbitrary *Jordan structure* (see Moro, Burke and Overton [79] for a nice review and an alternative proof of Lidskii's main theorem). Let A be a complex matrix with arbitrary Jordan structure, and λ_1 be an eigenvalue of A whose largest *Jordan block* has size m. Lidskii [70] shows that if A is perturbed to $A + \xi B$ with a small parameter ξ then the splitting of λ_1 is, generally, of order $\xi^{1/m}$, and obtains explicit formulas for the leading coefficients which involves the perturbation matrix and the eigenvectors of A.

2.2 Condition Numbers

2.2.1 Simple Eigenvalues

Let $A \in \mathcal{C}^{n\times n}$, and λ be a simple eigenvalue of A. Let $\tilde{A} = A + E$ be a perturbation of A, and $\tilde{\lambda}$ be the corresponding perturbation of λ. Then by (1.8.1) we define the condition number $c(\lambda)$ for λ as

$$c(\lambda) = \lim_{\delta\to 0} \sup_{\frac{\|E\|}{\alpha}\le\delta} \frac{|\tilde{\lambda}-\lambda|}{\xi\delta}, \tag{2.2.1}$$

where α and ξ are positive parameters.

From the definition (2.2.1) we see that in first order approximation the inequality

$$\frac{|\tilde{\lambda}-\lambda|}{\xi} \le c(\lambda)\frac{\|E\|}{\alpha}$$

holds.

Let $x, y \in \mathcal{C}^n$ be right and left eigenvectors of A associated with λ. Then by Theorem 2.1.1 and (2.1.19) we have

$$\tilde{\lambda} = \lambda + \frac{y^H E x}{y^H x} + O(\|E\|^2), \qquad E \to 0.$$

Substituting it into (2.2.1) gives

$$c(\lambda) = \alpha \sup_{\|E\|\le 1} \frac{\left|y^H E x\right|}{\xi|y^H x|} = \frac{\alpha\|x\|_2\|y\|_2}{\xi|y^H x|}. \tag{2.2.2}$$

Taking $\alpha = \xi = 1$ yields the absolute condition number

$$c_{\rm abs}(\lambda) = \frac{\|x\|_2\|y\|_2}{|y^H x|}, \tag{2.2.3}$$

and taking $\alpha = \|A\|_2$ and $\xi = |\lambda|$ (if $\lambda \neq 0$) yields the relative condition number

$$c_{\rm rel}(\lambda) = \frac{\|A\|_2\|x\|_2\|y\|_2}{|\lambda||y^H x|}. \tag{2.2.4}$$

The following result shows an important fact that if λ is a simple eigenvalue of A, then the shortest distance from A to a matrix which has an eigenvalue λ of multiplicity at least two is approximately bounded by $\|A\|_2/c_{\text{abs}}(\lambda)$ for large $c_{\text{abs}}(\lambda)$.

Theorem 2.2.1 (Wilkinson). *Let λ be a simple eigenvalue of $A \in \mathcal{C}^{n\times n}$ with right eigenvector x and left eigenvector y. If $c_{\text{abs}}(\lambda) > 1$ then there exists a matrix $E \in \mathcal{C}^{n\times n}$ such that $A+E$ has λ as an eigenvalue of multiplicity at least two and*

$$\|E\|_2 \le \frac{\|A\|_2}{\sqrt{c_{\text{abs}}^2(\lambda) - 1}}. \tag{2.2.5}$$

Proof. By the hypothesis, A has the Schur decomposition

$$A = U \begin{pmatrix} \lambda & a^H \\ 0 & A_2 \end{pmatrix} U^H, \tag{2.2.6}$$

where $U = (u_1, U_2) \in \mathcal{U}^{n\times n}$ with $u_1 \in \mathcal{C}^n$, and $\lambda \notin \lambda(A_2)$. Thus, there is a $w \in \mathcal{C}^{n-1}$ such that

$$\begin{pmatrix} 1 & -w^H \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} \lambda & a^H \\ 0 & A_2 \end{pmatrix} \begin{pmatrix} 1 & -w^H \\ 0 & I \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & A_2 \end{pmatrix}.$$

Combining it with (2.2.6) implies that

$$\begin{pmatrix} u_1^H + w^H U_2^H \\ * \end{pmatrix} A\,(u_1,\, *) = \begin{pmatrix} \lambda & 0 \\ 0 & A_2 \end{pmatrix},$$

which shows that the vectors

$$x = u_1 \quad \text{and} \quad y = u_1 + U_2 w \tag{2.2.7}$$

are right and left eigenvectors of A belonging to λ. Consequently, by (2.2.3) and (2.2.7) we have

$$c_{\text{abs}}(\lambda) = \sqrt{1 + \|w\|_2^2}. \tag{2.2.8}$$

Moreover, combining $y^H A = \lambda y^H$ with (2.2.6) gives

$$(1,\ w^H) \begin{pmatrix} \lambda & a^H \\ 0 & A_2 \end{pmatrix} = \lambda(1,\ w^H),$$

or equivalently,

$$w^H \left(A_2 + \frac{w a^H}{\|w\|_2^2} \right) = \lambda w^H.$$

Take

$$E = U \begin{pmatrix} 0 & 0 \\ 0 & \frac{w a^H}{\|w\|_2^2} \end{pmatrix} U^H. \tag{2.2.9}$$

Then λ is an eigenvalue of $A+E$ of multiplicity at least two, and from (2.2.8) and (2.2.9) we get the estimate (2.2.5). □

2.2.2 Invariant Subspaces

Let $A \in \mathcal{C}^{n\times n}$, and $\mathcal{X}_1$ be a simple invariant subspace of A. Let $\tilde{A} = A + E$ be a perturbation of A, and $\tilde{\mathcal{X}}_1$ be the corresponding perturbation of $\mathcal{X}_1$. Then by (1.8.3) we define the condition number $c(\mathcal{X}_1)$ for $\mathcal{X}_1$ as

$$c(\mathcal{X}_1) = \lim_{\delta\to 0} \sup_{\frac{\|E\|_F}{\alpha}\le\delta} \frac{\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1)}{\delta}, \tag{2.2.10}$$

where α is a positive parameter, and $\rho_F(\cdot,\cdot)$ is the generalized chordal metric defined by (1.3.3).

From (2.2.10) we see that in first order approximation the inequality

$$\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \le c(\mathcal{X}_1)\frac{\|E\|_F}{\alpha}$$

holds.

By (2.2.10), (2.1.33) and Theorem 1.3.3 (see (1.3.17)), we have

$$c(\mathcal{X}_1) = \lim_{\delta\to 0} \sup_{\frac{\|E\|_F}{\alpha}\le\delta} \frac{\|Z_1\|_F}{\delta}, \tag{2.2.11}$$

where Z_1, as a function of E, is defined by (2.1.34). Combining (2.2.11) with (2.1.36) gives

$$\begin{aligned} c(\mathcal{X}_1) &= \lim_{\delta\to 0} \sup_{\frac{\|\mathrm{vec}(E)\|_2}{\alpha}\le\delta} \frac{\|\mathrm{vec}(Z_1)\|_2}{\delta} = \alpha \sup_{\|\mathrm{vec}(E)\|_2\le 1} \|T^{-1}\mathrm{vec}(E_{21})\|_2 \\ &= \alpha \sup_{\|\mathrm{vec}(E_{21})\|_2\le 1} \|T^{-1}\mathrm{vec}(E_{21})\|_2 = \alpha\|T^{-1}\|_2, \end{aligned} \tag{2.2.12}$$

where T is the matrix of (2.1.35).

Taking $\alpha = 1$ yields the absolute condition number

$$c_{\mathrm{abs}}(\mathcal{X}_1) = \|T^{-1}\|_2, \tag{2.2.13}$$

and taking $\alpha = \|A\|_F$ yields the relative condition number

$$c_{\mathrm{rel}}(\mathcal{X}_1) = \|A\|_F\|T^{-1}\|_2. \tag{2.2.14}$$

Using the function sep introduced by Stewart [91], (2.2.13) can be written

$$c_{\mathrm{abs}}(\mathcal{X}_1) = \mathrm{sep}_F^{-1}(A_{11}, A_{22}),$$

where the function $\mathrm{sep}_F(A_{11}, A_{22})$ is defined by

$$\mathrm{sep}_F(A_{11}, A_{22}) = \begin{cases} \|\mathbf{T}^{-1}\|^{-1} & \text{if } 0 \notin \lambda(\mathbf{T}), \\ 0 & \text{if } 0 \in \lambda(\mathbf{T}), \end{cases}$$

in which $\mathbf{T}$ is the operator defined by (2.1.23), $\|\cdot\|$ denotes the operator norm induced by the Frobenius norm, and $\lambda(\mathbf{T})$ denotes the *spectrum* of $\mathbf{T}$.

Notes and References

NR 2.2–1. The expression (2.2.3) of the absolute condition number $c_{\mathrm{abs}}(\lambda)$ is a well known result (see Wilkinson [130]). The expression (2.2.4) of the relative condition number $c_{\mathrm{rel}}(\lambda)$ is obtained by Geurts [40]. Moreover, Geurts [40] gives the componentwise relative condition number $c_{\mathrm{rel}}^{(c)}(\lambda)$ for the non-zero simple eigenvalue λ_1:

$$c_{\mathrm{rel}}^{(c)}(\lambda) = \frac{|y^H||A||x|}{|\lambda||y^H x|},$$

where x and y are right and left eigenvectors of A associated with λ, respectively.

Let $\nu(\cdot)$ be any vector norm, and $\nu^D(\cdot)$ be the dual norm of $\nu(\cdot)$. Then we can obtain a more general expression of the condition number $c(\lambda)$:

$$c(\lambda) = \frac{\alpha\nu(x)\nu^D(y)}{\xi|y^H x|}.$$

Particularly, taking $\alpha = \xi = 1$ yields the absolute condition number

$$c_{\mathrm{abs}}(\lambda) = \frac{\nu(x)\nu^D(y)}{|y^H x|},$$

and taking $\alpha = \|A\|$ and $\xi = \lambda$ (if $\lambda \neq 0$) yields the relative condition number

$$c_{\mathrm{rel}}(\lambda) = \frac{\|A\|\nu(x)\nu^D(y)}{|\lambda||y^H x|},$$

where $\|\cdot\|$ is a matrix norm consistent with $\nu(\cdot)$.

NR 2.2–2. If an eigenvalue of a matrix has multiplicity at least two, then the corresponding eigenvalue problem is called *ill-posed* for the eigenvalue. Theorem 2.2.1 (Wilkinson [131]) shows that if λ is a simple eigenvalue of A, then the shortest distance from the eigenvalue problem to an ill-posed one is bounded by the reciprocal of the condition number of λ. Demmel [29] gives and compares various bounds on the distance from a matrix to the nearest one with a multiple eigenvalue, and he shows that for many problems of numerical analysis, there is the same relationship as for the eigenvalue problem between the condition number of a problem and the shortest distance from that problem to an ill-posed one.

NR 2.2–3. Assume that λ is a nondefective multiple eigenvalue of $A \in \mathcal{C}^{n\times n}$ with *multiplicity* r, i.e., there are matrices $X, Y \in \mathcal{C}^{n\times n}$ such that

$$Y^H A X = \begin{pmatrix} \lambda I_r & 0 \\ 0 & A_2 \end{pmatrix}, \quad Y^H X = I, \quad \lambda \notin \lambda(A_2).$$

Generally speaking, the multiple eigenvalue λ of A will split into r simple eigenvalues $\tilde{\lambda}_j$ $(j = 1, \ldots, r)$ when A is slightly perturbed to $\tilde{A}$. Hence, a multiple eigenvalue of multiplicity r can have r condition numbers that reflect the different sensitivities of its progeny. The typical behavior of the eigenvalues and corresponding condition numbers are studied by Stewart and Zhang [98], and Sun [114], [117].

NR 2.2–4. On the basis of Lidskii's perturbation theory for eigenvalues of matrices with arbitrary Jordan structure (see NR 2.1–7), Moro, Burke and Overton [79] suggest a notion of Hölder condition number for multiple eigenvalues, depending only on the conditioning of the associated eigenvectors.

NR 2.2–5. The condition number $c_{\rm abs}(\mathcal{X}_1) = {\rm sep}_F^{-1}(A_{11}, A_{22})$ is given by Stewart [91]. In §2.2.2 we present a proof by applying Rice's theory of condition and using the generalized chordal metric.

NR 2.2–6. Varah [124] discusses some properties of ${\rm sep}_F(A_{11}, A_{22})$, and gives some examples to show how very small it can be for seemingly harmless problems. Sun [101] and Xu [133] give some theoretical estimates on lower bounds for ${\rm sep}_F(A_{11}, A_{22})$.

NR 2.2–7. Byers [12] proposes an algorithm for estimating ${\rm sep}_F(A_{11}, A_{22})$ in the style of the LINPACK condition number estimator. Kågström and Poromaa [58] present estimators for ${\rm sep}_F(A_{11}, A_{22})$ by using distributed and shared memory block algorithms. For a nice survey on condition estimation in general, see Higham [48].

NR 2.2–8. Let $A, A_{11}, A_{22}, \mathcal{X}_1$ be as in Theorem 2.1.5, where $\lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset$; the corresponding eigenvalue problem is called *well-posed* for the invariant subspace $\mathcal{X}_1$. If $\lambda(A_{11}) \bigcap \lambda(A_{22}) \neq \emptyset$, then the corresponding eigenvalue problem is called ill-posed for the invariant subspace. Demmel [29] gives a lower bound on the distance from a well-posed eigenvalue problem for an invariant subspace to the nearest ill-posed one by means of the reciprocal of the condition number $c_{\rm abs}(\mathcal{X}_1)$

NR 2.2–9. Bai, Demmel and McKenney [3] review the theory of condition numbers for the eigenvalue problem and give a tabular summary of bounds for eigenvalues, means of clusters of eigenvalues, eigenvectors, invariant subspaces, and related quantities. They describe the design of algorithms for estimating these condition numbers.

2.3 Perturbation Bounds for Invariant Subspaces

We first prove a forward perturbation theorem for simple invariant subspaces. The proof is based on the use of Stewart's technique [91] and Theorem 2.3.4 at the end of this subsection.

Theorem 2.3.1. *Let $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$, and suppose that $\mathcal{X}_1 = \mathcal{R}(X_1)$ is an l-dimensional simple invariant subspace of A and (2.1.22) holds. For a perturbation E we let*

$$X^H E X = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \tag{2.3.1}$$

and assume that $\lambda(A_{11}+E_{11}) \bigcap \lambda(A_{22}+E_{22}) = \emptyset$. Define the linear operator $\mathbf{L} : \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{L}Z = Z(A_{11}+E_{11}) - (A_{22}+E_{22})Z, \qquad Z \in \mathcal{C}^{(n-l)\times l}, \tag{2.3.2}$$

and set

$$b = \|\mathbf{L}^{-1}E_{21}\|, \qquad c = \|\mathbf{L}^{-1}\|, \qquad \eta = \|A_{12}+E_{12}\|_2. \tag{2.3.3}$$

If

$$4bc\eta < 1, \tag{2.3.4}$$

then there is a unique l-dimensional invariant subspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A+E$ satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\| \le \frac{2b}{1+\sqrt{1-4bc\eta}}, \tag{2.3.5}$$

where $\tilde{X}_1 \in \mathcal{U}^{n\times l}$.

Proof. Consider the equation

$$\mathbf{L}Z = E_{21} + \phi(Z), \tag{2.3.6}$$

where $\mathbf{L}$ is the operator defined by (2.3.2), and the function ϕ is defined by

$$\phi(Z) = -Z(A_{12}+E_{12})Z, \qquad Z \in \mathcal{C}^{(n-l)\times l}.$$

Observe that the function ϕ satisfies

$$\|\phi(Z)\| \le \eta\|Z\|^2, \quad \|\phi(\tilde{Z}) - \phi(Z)\| \le 2\eta\max\{\|\tilde{Z}\|, \|Z\|\}\|\tilde{Z} - Z\|,$$

and the scalars b, c, η defined by (2.3.3) satisfy (2.3.4). Hence, by Theorem 2.3.4 at the end of this subsection there is a unique solution Z^* of the equation (2.3.6) satisfying

$$\|Z^*\| \le \frac{2b}{1+\sqrt{1-4bc\eta}}. \tag{2.3.7}$$

It can be verified that the relation

$$\mathbf{L}Z^* = E_{21} + \phi(Z^*)$$

is equivalent to

$$\begin{pmatrix} I & 0 \\ Z^* & I \end{pmatrix}^{-1} (A+E) \begin{pmatrix} I & 0 \\ Z^* & I \end{pmatrix} = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}.$$

Combining it with (2.1.22), (2.3.1), (2.3.7) and (1.3.16) shows that the subspace $\tilde{\mathcal{X}}_1$ defined by $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ with

$$\tilde{X}_1 = X \begin{pmatrix} I \\ Z^* \end{pmatrix} \left(I + Z^{*^H} Z^*\right)^{-\frac{1}{2}} \in \mathcal{U}^{n\times l}$$

is the unique invariant subspace of $A+E$ satisfying (2.3.5). □

From Theorem 2.3.1 we get the following corollary.

Corollary 2.3.2. *Let $A, E, X, \mathcal{X}_1$ be as in Theorem 2.3.1, and assume that*

$$E_{11} = 0, \qquad E_{12} = 0, \qquad E_{22} = 0.$$

Define the linear operator $\mathbf{T} : \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{T}Z = ZA_{11} - A_{22}Z, \qquad Z \in \mathcal{C}^{(n-l)\times l}, \tag{2.3.8}$$

and assume $\lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset$. If

$$4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}E_{21}\|\|A_{12}\|_2 < 1,$$

then there is a unique l-dimensional invariant subspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A+E$ satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\| \le \frac{2\|\mathbf{T}^{-1}E_{21}\|}{1 + \sqrt{1 - 4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}E_{21}\|\|A_{12}\|_2}},$$

where $\tilde{X}_1 \in \mathcal{U}^{n\times l}$.

Moreover, if the perturbation matrix E itself is unknown but some upper bounds for $\|E_{jk}\|$ are known, then we have the following well known result.

Theorem 2.3.3 (Stewart). *Let $A, E, X, \mathcal{X}_1$ be as in Theorem 2.3.1, and let $\mathbf{T}$ be the linear operator defined by (2.3.8). Assume that*

$$\lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset \quad \text{and} \quad \|\mathbf{T}^{-1}\|(\|E_{11}\| + \|E_{22}\|) < 1,$$

and set

$$\tilde{c} = \frac{\|\mathbf{T}^{-1}\|}{1 - \|\mathbf{T}^{-1}\|(\|E_{11}\| + \|E_{22}\|)}, \qquad \gamma = \|E_{21}\|, \qquad \tilde{\eta} = \|A_{12}\|_2 + \|E_{12}\|_2.$$

If

$$4\tilde{c}^2\gamma\tilde{\eta} < 1,$$

then there is a unique l-dimensional invariant subspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A + E$ satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\| \leq \frac{2\tilde{c}\gamma}{1 + \sqrt{1 - 4\tilde{c}^2\gamma\tilde{\eta}}},$$

where $\tilde{X}_1 \in \mathcal{U}^{n\times l}$.

We now prove a general result on solution of a nonlinear equation, which can be used to establish the existence of Z^* in Theorem 2.3.1. We state and prove it for a Banach space, which the reader may take to be a finite dimensional normed linear space.

Theorem 2.3.4. *Let $\mathbf{T}$ be a bounded linear operator on a Banach space $\mathcal{B}$, and let $\|\cdot\|$ be a norm on $\mathcal{B}$ and the induced operator norm. Assume that $\mathbf{T}$ has a bounded inverse, and set*

$$c = \|\mathbf{T}^{-1}\|. \tag{2.3.9}$$

Let $\phi : \mathcal{B} \to \mathcal{B}$ be a function that satisfies

$$\|\phi(x)\| \leq \eta\|x\|^2, \quad \|\phi(\tilde{x}) - \phi(x)\| \leq 2\eta\max\{\|\tilde{x}\|, \|x\|\}\|\tilde{x} - x\| \tag{2.3.10}$$

for any $x, \tilde{x} \in \mathcal{B}$ and some $\eta \geq 0$. For any $g \in \mathcal{B}$, let

$$b = \|\mathbf{T}^{-1}g\|. \tag{2.3.11}$$

If

$$4bc\eta < 1, \tag{2.3.12}$$

then there is a unique solution x^* of the nonlinear equation

$$\mathbf{T}x = g + \phi(x) \tag{2.3.13}$$

that satisfies

$$\|x^*\| \leq \frac{2b}{1 + \sqrt{1 - 4bc\eta}} \equiv \xi^*. \tag{2.3.14}$$

Proof. Define

$$\mathcal{S}_{\xi^*} = \{x \in \mathcal{B} \;:\; \|x\| \leq \xi^*\}.$$

We first prove that if there is a solution of (2.3.13) in $\mathcal{S}_{\xi^*}$, then it is unique.

Assume that the equation (2.3.13) has different solutions $x^*, \hat{x} \in \mathcal{S}_{\xi^*}$. Then by (2.3.9), (2.3.10), and (2.3.14), we have

$$\begin{aligned} \|x^* - \hat{x}\| &\leq \|\mathbf{T}^{-1}\| \|\phi(x^*) - \phi(\hat{x})\| \\ &\leq 2c\eta \max\{\|x^*\|, \|\hat{x}\|\} \|x^* - \hat{x}\| \\ &\leq 2c\eta \cdot \frac{2b}{1 + \sqrt{1 - 4bc\eta}} \|x^* - \hat{x}\| \\ &< 4bc\eta \|x^* - \hat{x}\| < \|x^* - \hat{x}\|. \end{aligned}$$

This contradiction shows that there is at most one solution of the equation (2.3.13) in $\mathcal{S}_{\xi^*}$.

Now we prove the existence of a solution of (2.3.13) in $\mathcal{S}_{\xi^*}$.

Consider the continuous mapping $\mathcal{M} : \mathcal{B} \to \mathcal{B}$ defined by

$$y = \mathbf{T}^{-1}[g + \phi(x)]. \tag{2.3.15}$$

Since any fixed point of the mapping $\mathcal{M}$ is a solution of the equation (2.3.13), the problem of finding a solution of (2.3.13) satisfying (2.3.14) reduces to the problem of showing that there is a fixed point of the mapping $\mathcal{M}$ in $\mathcal{S}_{\xi^*}$.

It is easy to verify that the scalar ξ^* defined by (2.3.14) is a solution of the equation

$$c\eta\xi^2 - \xi + b = 0. \tag{2.3.16}$$

From (2.3.15) we see that if $x \in \mathcal{B}$ satisfies $\|x\| \leq \xi^*$ then y satisfies

$$\begin{aligned} \|y\| &\leq \|\mathbf{T}^{-1}g\| + \|\mathbf{T}^{-1}\| \|\phi(x)\| \\ &\leq b + c\eta\|x\|^2 \qquad \text{(by (2.3.9) – (2.3.11))} \\ &\leq b + c\eta\xi^{*2} \\ &= \xi^*, \qquad \text{(by (2.3.16))} \end{aligned}$$

which means that for the mapping $\mathcal{M}$ defined by (2.3.15) we have

$$x \in \mathcal{S}_{\xi^*} \implies y \in \mathcal{S}_{\xi^*}. \tag{2.3.17}$$

Observe that $\mathcal{S}_{\xi^*}$ is a bounded closed convex set of $\mathcal{B}$, and (2.3.17) shows that the continuous mapping $\mathcal{M}$ maps $\mathcal{S}_{\xi^*}$ into $\mathcal{S}_{\xi^*}$. Hence, by the Schauder fixed-point theorem (Theorem 1.7.2) the mapping $\mathcal{M}$ has a fixed point in $\mathcal{S}_{\xi^*}$, and thus the equation (2.3.13) has a solution in $\mathcal{S}_{\xi^*}$. □

Notes and References

NR 2.3–1. Theorem 2.3.3 is given by Stewart [91]. Theorem 2.3.1 and Corollary 2.3.2 are new results, which give perturbation bounds for invariant subspaces when the perturbation matrix E itself is known.

NR 2.3–2. Theorem 2.3.4 is proved by Sun [121].

NR 2.3–3. A note on Theorem 2.3.4. Let $\gamma = \|g\|$. Stewart [91, Theorem 3.1] shows that if the function ϕ satisfies (2.3.10), and

$$4c^2\gamma\eta < 1,$$

then there is a unique solution x^* of the equation (2.3.13) that satisfies

$$\|x^*\| \le \frac{2c\gamma}{1+\sqrt{1-4c^2\gamma\eta}}. \tag{2.3.18}$$

We now compare the estimates (2.3.18) and (2.3.14). Assume that $\mathcal{B}$ is a finite dimensional Banach space. Let T be the matrix representation of $\mathbf{T}$, and let v_{x^*}, v_g be the vector representations of x^*, g, respectively. Then in first order approximation the estimates (2.3.14) and (2.3.18) can be written

$$\|v_{x^*}\| \le \|T^{-1}v_g\|, \qquad \|v_{x^*}\| \le \|T^{-1}\|\|v_g\|, \tag{2.3.19}$$

respectively, where $\|\cdot\|$ denotes any consistent matrix norm and associated vector norm. The attraction of the first bound of (2.3.19) is that if v_g is known then large elements in the jth column of T^{-1} may be countered by a small jth element of v_g (or a large jth element of v_g may be countered by small elements in the jth column of T^{-1}), making the bound much smaller than the second bound of (2.3.19). This fact is pointed out by Higham [51, section 5]. Note that if the vector v_g itself is unknown but some upper bound for $\|v_g\|$ is known, then we are forced to use the second bound of (2.3.19), i.e., if the g itself is unknown but some upper bound for $\|g\|$ is known, then we are forced to use the bound (2.3.18).

2.4 Backward Errors and Residual Bounds

2.4.1 Backward Errors

In this subsection we discuss several kinds of normwise backward errors which are defined by using some information of approximate simple invariant subspaces and associated eigenmatrices of a matrix A. An approximate invariant subspace may come from a numerical algorithm (see, e.g., Dongarra, Hammarling, and Wilkinson [33] for methods for computing invariant subspaces).

2.4.1.1 The Backward Error $\eta(\tilde{\mathcal{X}}_1)$

Let $\tilde{\mathcal{X}}_1$ approximate an l-dimensional simple invariant subspace of $A \in \mathcal{C}^{n\times n}$. By §1.9, we define the backward error $\eta(\tilde{\mathcal{X}}_1)$ of A with respect to $\tilde{\mathcal{X}}_1$ by

$$\eta(\tilde{\mathcal{X}}_1) = \min_{E\in\mathcal{E}} \|E\|, \tag{2.4.1}$$

where the set $\mathcal{E}$ is defined by

$$\mathcal{E} = \left\{ E \in \mathcal{C}^{n\times n} \ : \ (A+E)\tilde{\mathcal{X}}_1 \subset \tilde{\mathcal{X}}_1 \right\}. \tag{2.4.2}$$

The following result gives a computable formula of $\eta(\tilde{\mathcal{X}}_1)$.

Theorem 2.4.1. *Choose $\tilde{U}_1 \in \mathcal{U}^{n\times l}$ so that $\mathcal{R}(\tilde{U}_1) = \tilde{\mathcal{X}}_1$. Let*

$$R = \tilde{U}_1(\tilde{U}_1^H A\tilde{U}_1) - A\tilde{U}_1 \tag{2.4.3}$$

be the residual of A with respect to $\tilde{U}_1$. Then the backward error $\eta(\tilde{\mathcal{X}}_1)$ can be expressed by

$$\eta(\tilde{\mathcal{X}}_1) = \|R\|. \tag{2.4.4}$$

The expressions (2.4.3) and (2.4.4) imply that the backward error $\eta(\tilde{\mathcal{X}}_1)$ defined by (2.4.1)–(2.4.2) is independent of the choice of the matrix $\tilde{U}_1$ whose column vectors form an orthonormal basis of $\tilde{\mathcal{X}}_1$.

Proof of Theorem 2.4.1. From (2.4.2) it follows that a matrix $E \in \mathcal{E}$ if and only if E is a solution of the equation

$$(A+E)\tilde{U}_1 = \tilde{U}_1 A_1$$

for some $A_1 \in \mathcal{C}^{l\times l}$; or equivalently, E satisfies

$$E\tilde{U}_1 = \tilde{U}_1 A_1 - A\tilde{U}_1. \tag{2.4.5}$$

Applying Theorem 1.5.1 to the equation (2.4.5) we see that the equation is solvable, and any solution E of the equation can be expressed by

$$E = (\tilde{U}_1 A_1 - A\tilde{U}_1)\tilde{U}_1^H + Z(I - \tilde{U}_1\tilde{U}_1^H), \tag{2.4.6}$$

where $Z \in \mathcal{C}^{n\times n}$.

Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Then from (2.4.6)

$$\tilde{U}^H E\tilde{U} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{U}_1 & \tilde{U}_1^H Z\tilde{U}_2 \\ -\tilde{U}_2^H A\tilde{U}_1 & \tilde{U}_2^H Z\tilde{U}_2 \end{pmatrix} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{U}_1 & \tilde{U}_1^H Z\tilde{U}_2 \\ \tilde{U}_2^H R & \tilde{U}_2^H Z\tilde{U}_2 \end{pmatrix}.$$

By the definition (2.4.1) and Theorem 1.2.1 we have

$$\eta(\tilde{\mathcal{X}}_1) = \|E_{\text{opt}}\| \quad \text{with} \quad E_{\text{opt}} = \tilde{U} \begin{pmatrix} 0 & 0 \\ \tilde{U}_2^H R & 0 \end{pmatrix} \tilde{U}^H. \tag{2.4.7}$$

Observe that the relation

$$\tilde{U}^H R = \begin{pmatrix} 0 \\ \tilde{U}_2^H R \end{pmatrix}$$

implies

$$\sigma_+(\tilde{U}_2^H R) = \sigma_+(R).$$

Hence, we have

$$\sigma_+ \begin{pmatrix} 0 & 0 \\ \tilde{U}_2^H R & 0 \end{pmatrix} = \sigma_+(R).$$

Combining it with (2.4.7) shows (2.4.4). □

2.4.1.2 The Backward Error $\eta(\tilde{X}_1, \tilde{A}_1)$

Let $A \in \mathcal{C}^{n\times n}$, and let $\mathcal{X}_1$ be an l-dimensional subspace of $\mathcal{C}^n$. By the definition, $\mathcal{X}_1$ is an invariant subspace of A if and only if there are matrices $X_1 \in \mathcal{C}^{n\times l}$ and $A_1 \in \mathcal{C}^{l\times l}$ such that

$$\mathcal{X}_1 = \mathcal{R}(X_1) \quad \text{and} \quad AX_1 = X_1 A_1.$$

The matrix A_1 may be called the *(right) eigenmatrix* of A associated with X_1.

Suppose that the column vectors of $\tilde{X}_1 \in \mathcal{C}^{n\times l}$ form a basis of an approximate invariant subspace of A, and $\tilde{A}_1 \in \mathcal{C}^{l\times l}$ is the associated eigenmatrix. By §1.9, we define the backward error $\eta(\tilde{X}_1, \tilde{A}_1)$ of A with respect to $\tilde{X}_1$ and $\tilde{A}_1$ by

$$\eta(\tilde{X}_1, \tilde{A}_1) = \min_{E\in\mathcal{E}} \|E\|, \tag{2.4.8}$$

where the set $\mathcal{E}$ is defined by

$$\mathcal{E} = \left\{ E \in \mathcal{C}^{n\times n} \ : \ (A+E)\tilde{X}_1 = \tilde{X}_1\tilde{A}_1 \right\}. \tag{2.4.9}$$

The following result gives a computable formula of $\eta(\tilde{X}_1, \tilde{A}_1)$.

Theorem 2.4.2. *Let*

$$R = \tilde{X}_1\tilde{A}_1 - A\tilde{X}_1 \tag{2.4.10}$$

be the residual of A with respect to $\tilde{X}_1$ and $\tilde{A}_1$. Then the backward error $\eta(\tilde{X}_1, \tilde{A}_1)$ can be expressed by

$$\eta(\tilde{X}_1, \tilde{A}_1) = \left\| R\tilde{X}_1^\dagger \right\|. \tag{2.4.11}$$

Proof. From (2.4.9) it follows that a matrix $E \in \mathcal{E}$ if and only if E satisfies

$$E\tilde{X}_1 = R, \tag{2.4.12}$$

where R is the residual defined by (2.4.10).

Applying Theorem 1.5.1 to the equation (2.4.12) we see that the equation is solvable, and any solution E to the equation can be expressed by

$$E = R\tilde{X}^{\dagger} + Z(I - \tilde{X}\tilde{X}^{\dagger}), \tag{2.4.13}$$

where $Z \in \mathcal{C}^{n\times n}$.

Take an orthogonal decomposition $\tilde{X}_1 = \tilde{U}_1 L$ of $\tilde{X}_1$, where $\tilde{U}_1 \in \mathcal{U}^{n\times l}$ and $L \in \mathcal{C}^{l\times l}$. Further, choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Then from (2.4.13)

$$E = RL^{-1}\tilde{U}_1^H + Z(I - \tilde{U}_1\tilde{U}_1^H) = (RL^{-1},\ Z\tilde{U}_2)\tilde{U}^H.$$

By the definition (2.4.8) and Theorem 1.2.1, we have

$$\eta(\tilde{X}_1, \tilde{A}_1) = \|E_{\text{opt}}\| \quad \text{with} \quad E_{\text{opt}} = RL_1^{-1}\tilde{U}_1^H = R\tilde{X}_1^{\dagger},$$

which shows (2.4.11). □

Remark 2.4.3. Let $\tilde{\lambda}_1 \in \mathcal{C}$ be an approximate eigenvalue of $A \in \mathcal{C}^{n\times n}$, and $\tilde{x}_1 \in \mathcal{C}^n$ be an associated eigenvector. Then by Theorem 2.4.2, the backward error $\eta(\tilde{x}_1, \tilde{\lambda}_1)$ of A with respect to $\tilde{x}_1$ and $\tilde{\lambda}_1$ can be expressed by

$$\eta(\tilde{x}_1, \tilde{\lambda}_1) = \frac{\|r\|_2}{\|\tilde{x}_1\|_2}, \tag{2.4.14}$$

where

$$r = \tilde{\lambda}_1\tilde{x}_1 - A\tilde{x}_1$$

be the residual of A with respect to $\tilde{x}_1$ and $\tilde{\lambda}_1$. Moreover, the optimal backward perturbation E_{opt} in A is expressed by

$$E_{\text{opt}} = r\tilde{x}_1^{\dagger},$$

which is the smallest perturbation of A (in any unitarily invariant norm) such that $\tilde{\lambda}_1$ is an eigenvalue of $A + E_{\text{opt}}$, and $\tilde{x}_1$ is an associated eigenvector.

Example 2.4.4 (Yamamoto [135, Example 2]). Consider the matrix

$$A = \begin{pmatrix} 14 & 9 & 6 & 4 & 2 \\ -9 & -4 & -3 & -2 & -1 \\ -2 & -2 & 0 & -1 & -1 \\ 3 & 3 & 3 & 5 & 3 \\ -9 & -9 & -9 & -9 & -4 \end{pmatrix},$$

which has the eigenvalues

$$\lambda_1 = 1 + \sqrt{2}i, \quad \lambda_2 = 1 - \sqrt{2}i, \quad \lambda_3 = 5, \quad \lambda_4 = \lambda_5 = 2,$$

and associated eigenvectors

$$x_1 = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 2-\sqrt{2}i \\ -1+2\sqrt{2}i \end{pmatrix}, \quad x_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 2+\sqrt{2}i \\ -1-2\sqrt{2}i \end{pmatrix}, \quad x_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad x_4 = \begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \\ 0 \end{pmatrix}.$$

Using the MATLAB file "eig" (which is an implementation of the QR method) to the matrix A, we get the computed eigenvalues $\tilde{\lambda}_j$ and associated eigenvectors $\tilde{x}_j$ for $j = 1, 2, 3, 4, 5$, among which $\tilde{x}_4$ and $\tilde{x}_5$ are approximately linearly dependent. Applying (2.4.14) we get

$$\eta(\tilde{x}_1, \tilde{\lambda}_1) \approx 6.94 \times 10^{-15}, \quad \eta(\tilde{x}_2, \tilde{\lambda}_2) \approx 6.94 \times 10^{-15}, \quad \eta(\tilde{x}_3, \tilde{\lambda}_3) \approx 3.38 \times 10^{-15},$$

$$\eta(\tilde{x}_4, \tilde{\lambda}_4) \approx 4.22 \times 10^{-15}, \quad \eta(\tilde{x}_5, \tilde{\lambda}_5) \approx 5.05 \times 10^{-15},$$

which show that each computed eigenvalue $\tilde{\lambda}_j$ and associated eigenvector $\tilde{x}_j$ are an exact eigenvalue and an associated eigenvector of a very slightly perturbed matrix of A; in other words, the computation has proceeded quite stably.

2.4.1.3 The Backward Errors $\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ and $\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$

Let $A \in \mathcal{C}^{n\times n}$. If $Y_1 \in \mathcal{C}^{n\times l}$ and $C_1 \in \mathcal{C}^{l\times l}$ satisfy

$$\operatorname{rank}(Y_1) = l \quad \text{and} \quad Y_1^H A = C_1 Y_1^H,$$

then $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ is called a *left* invariant subspace of A, and the matrix C_1 may be called the *left* eigenmatrix of A associated with Y_1.

Let the column vectors of $\tilde{X}_1 \in \mathcal{C}^{n\times l}$ form a basis of a subspace $\tilde{\mathcal{X}}_1$ which approximates an invariant subspace $\mathcal{X}_1$ of A, and let $\tilde{A}_1 \in \mathcal{C}^{l\times l}$ be the associated approximate (right) eigenmatrix. Moreover, let the column vectors of $\tilde{Y}_1 \in \mathcal{C}^{n\times l}$ form a basis of a subspace $\tilde{\mathcal{Y}}_1$ which approximates a left invariant subspace $\mathcal{Y}_1$ of A, and $\tilde{C}_1 \in \mathcal{C}^{l\times l}$ is the associated approximate left eigenmatrix. Suppose that the invariant subspaces $\mathcal{X}_1$ and $\mathcal{Y}_1$ correspond to the same eigenvalues of A. Therefore, we may assume that

$$\operatorname{rank}(\tilde{Y}_1^H \tilde{X}_1) = l. \tag{2.4.15}$$

By §1.9, we define the backward errors $\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ and $\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ of A with respect to $\tilde{X}_1$, $\tilde{Y}_1$, $\tilde{A}_1$, $\tilde{C}_1$ by

$$\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \min_{E\in\mathcal{E}} \|E\|_F, \quad \eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \min_{E\in\mathcal{E}} \|E\|_2, \tag{2.4.16}$$

where the set $\mathcal{E}$ is defined by

$$\mathcal{E} = \left\{ E \in \mathcal{C}^{n\times n} \ : \ (A+E)\tilde{X}_1 = \tilde{X}_1\tilde{A}_1, \ \ \tilde{Y}_1^H(A+E) = \tilde{C}_1\tilde{Y}_1^H \right\}. \tag{2.4.17}$$

If the matrices $\tilde{X}_1$ and $\tilde{Y}_1$ satisfy $\tilde{X}_1^H\tilde{X}_1 = \tilde{Y}_1^H\tilde{Y}_1 = I$, then the following result gives computable formulas of $\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ and $\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ in the case of $\mathcal{E} \neq \emptyset$.

Theorem 2.4.5 (Kahan, Parlett, and Jiang). *If $\tilde{X}_1, \tilde{Y}_1 \in \mathcal{U}^{n\times l}$, and if*

$$\operatorname{rank}(\tilde{Y}_1^H\tilde{X}_1) = l, \qquad \tilde{C}_1 = \left(\tilde{Y}_1^H\tilde{X}_1\right)\tilde{A}_1\left(\tilde{Y}_1^H\tilde{X}_1\right)^{-1}. \tag{2.4.18}$$

Then

$$\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \sqrt{\|R\|_F^2 + \|S\|_F^2 - \|S\tilde{X}_1\|_F^2}, \tag{2.4.19}$$

and

$$\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \max\{\|R\|_2,\ \|S\|_2\}, \tag{2.4.20}$$

where R, S are the residuals defined by

$$R = \tilde{X}_1\tilde{A}_1 - A\tilde{X}_1, \qquad S = \tilde{C}_1\tilde{Y}_1^H - \tilde{Y}_1^H A. \tag{2.4.21}$$

Proof. From (2.4.17) it follows that a matrix $E \in \mathcal{E}$ if and only if E is a solution of the equations

$$E\tilde{X}_1 = R, \qquad \tilde{Y}_1^H E = S, \tag{2.4.22}$$

where R and S are the residuals defined by (2.4.21).

Applying Theorem 1.5.1 to the first equation of (2.4.22) we see that the equation is solvable, and any solution E to the equation can be expressed by

$$E = R\tilde{X}_1^H + Z(I - \tilde{X}_1\tilde{X}_1^H), \tag{2.4.23}$$

where $Z \in \mathcal{C}^{n\times n}$. Combining (2.4.23) with (2.4.21) and the second equation of (2.4.22), shows that the matrix Z of (2.4.23) satisfies

$$\tilde{Y}_1^H Z(I - \tilde{X}_1\tilde{X}_1^H) = \tilde{C}_1\tilde{Y}_1^H - \tilde{Y}_1^H A - \tilde{Y}_1^H\tilde{X}_1\tilde{A}_1\tilde{X}_1^H + \tilde{Y}_1^H A\tilde{X}_1\tilde{X}_1^H. \tag{2.4.24}$$

Applying Theorem 1.5.1 to the equation (2.4.24) we see that the equation is solvable if and only if $\tilde{A}_1$ and $\tilde{C}_1$ satisfy (2.4.18), and under the condition (2.4.18) any solution Z to the equation (2.4.24) can be expressed by

$$Z = \tilde{Y}_1 S(I - \tilde{X}_1\tilde{X}_1^H) + W - \tilde{Y}_1\tilde{Y}_1^H W(I - \tilde{X}_1\tilde{X}_1^H).$$

Substituting it into (2.4.23) gives

$$E = R\tilde{X}_1^H + \tilde{Y}_1 S(I - \tilde{X}_1\tilde{X}_1^H) + (I - \tilde{Y}_1\tilde{Y}_1^H)W(I - \tilde{X}_1\tilde{X}_1^H). \tag{2.4.25}$$

Choose $\tilde{X}_2, \tilde{Y}_2$ so that $\tilde{X} = (\tilde{X}_1, \tilde{X}_2), \tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2) \in \mathcal{U}^{n\times n}$. Then any matrix E of (2.4.25) satisfies

$$\tilde{Y}^H E\tilde{X} = \begin{pmatrix} \tilde{Y}_1^H R & S\tilde{X}_2 \\ \tilde{Y}_2^H R & \tilde{Y}_2^H W\tilde{X}_2 \end{pmatrix}. \tag{2.4.26}$$

By the definition (2.4.16) and Theorem 1.2.1, we have

$$\left[\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)\right]^2 = \|R\|_F^2 + \left\|S\tilde{X}_2\right\|_F^2.$$

Combining it with

$$\|S\|_F^2 = \left\|S\tilde{X}_1\right\|_F^2 + \left\|S\tilde{X}_2\right\|_F^2$$

shows (2.4.19).

Moreover, by the definition (2.4.16) and Theorem 1.2.4, from (2.4.26) we get

$$\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \max\left\{\|R\|_2,\ \left\|\left(\tilde{Y}_1^H R,\ S\tilde{X}_2\right)\right\|_2\right\}.$$

Combining it with

$$\left\|\left(\tilde{Y}_1^H R,\ S\tilde{X}_2\right)\right\|_2 = \left\|S\left(\tilde{X}_1, \tilde{X}_2\right)\right\|_2 = \|S\|_2,$$

shows (2.4.20). □

2.4.2 Residual Bounds

Let an approximate invariant subspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$ of $A \in \mathcal{C}^{n\times n}$ be given, where $\tilde{U}_1 \in \mathcal{U}^{n\times l}$. Then by using Theorem 2.4.1 and an appropriate forward perturbation result we can determine the accuracy of the approximate invariant subspace $\tilde{\mathcal{X}}_1$.

Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. By the proof of Theorem 2.4.1, the optimal backward perturbation E_{opt} of (2.4.7) satisfies

$$\tilde{U}^H(A + E_{\text{opt}})\tilde{U} = \begin{pmatrix} \tilde{U}_1^H A\tilde{U}_1 & \tilde{U}_1^H A\tilde{U}_2 \\ 0 & \tilde{U}_2^H A\tilde{U}_2 \end{pmatrix} \equiv \begin{pmatrix} \tilde{A}_{11} & -S\tilde{U}_2 \\ 0 & \tilde{A}_{22} \end{pmatrix}, \tag{2.4.27}$$

and

$$\tilde{U}^H E_{\text{opt}}\tilde{U} = \begin{pmatrix} 0 & 0 \\ \tilde{U}_2^H R & 0 \end{pmatrix}, \tag{2.4.28}$$

where R is the residual defined by (2.4.3), and S is the residual of A with respect to $\tilde{U}_1^H$ defined by

$$S = (\tilde{U}_1^H A\tilde{U}_1)\tilde{U}_1^H - \tilde{U}_1^H A.$$

The relation (2.4.27) implies that the subspace $\tilde{\mathcal{X}}_1$ is an invariant subspace of $A + E_{\text{opt}}$.

Applying Corollary 2.3.2 to the matrices $A + E_{\text{opt}}$ and A, and using the relations $S\tilde{U}_1 = 0$ and $\|S\tilde{U}_2\| = \|S\tilde{U}\| = \|S\|$, we obtain the following result which gives a residual bound for the approximate invariant subspace $\tilde{\mathcal{X}}_1$ of A.

Theorem 2.4.6. *Let $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$ be an approximate invariant subspace of $A \in \mathcal{C}^{n\times n}$, where $\tilde{U}_1 \in \mathcal{U}^{n\times l}$. Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Define the matrices $\tilde{A}_{11}, \tilde{A}_{22}$ by (2.4.27), and define the linear operator $\mathbf{T}$ by*

$$\mathbf{T}Z = Z\tilde{A}_{11} - \tilde{A}_{22}Z, \qquad Z \in \mathcal{C}^{(n-l)\times l}.$$

Moreover, define the residuals R and S by

$$R = \tilde{U}_1\tilde{A}_{11} - A\tilde{U}_1, \qquad S = \tilde{A}_{11}\tilde{U}_1^H - \tilde{U}_1^H A.$$

If

$$\lambda(\tilde{A}_{11}) \bigcap \lambda(\tilde{A}_{22}) = \emptyset \quad \text{and} \quad 4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}(\tilde{U}_2^H R)\|\|S\|_2 < 1,$$

then there is a unique invariant subspace $\mathcal{X}_1 = \mathcal{R}(U_1)$ of A with $U_1 \in \mathcal{U}^{n\times l}$ such that

$$\rho(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \le \|\tan\Theta(U_1, \tilde{U}_1)\| \le \frac{2\|\mathbf{T}^{-1}(\tilde{U}_2^H R)\|}{1 + \sqrt{1 - 4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}(\tilde{U}_2^H R)\|\|S\|_2}}. \qquad (2.4.29)$$

From the relation (2.4.27) we see that the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ of $\tilde{U}_1^H A\tilde{U}_1$, as l approximate eigenvalues of A, are l eigenvalues of $A + E_{\text{opt}}$. Applying the Henrici theorem on perturbations of eigenvalues [44] (or see Stewart and Sun [97, Chapter IV, Theorem 1.9]) to the matrices A and $A + E_{\text{opt}}$, we can obtain a residual bound for the approximate eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ of A. Before the statement of the result on a residual bound for $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$, we first define the 2-departure of a matrix from normality by using the Schur decomposition.

It is well known that for any $A \in \mathcal{C}^{n\times n}$, there is the Schur decomposition $A = UTU^H$, where $U \in \mathcal{U}^{n\times n}$, and $T \in \mathcal{C}^{n\times n}$ is upper triangular. Let $A \in \mathcal{C}^{n\times n}$, and let $\mathcal{U}_A$ be the set defined by

$$\mathcal{U}_A = \left\{ U \in \mathcal{U}^{n\times n} \ : \ U^H AU \ \text{ is upper triangular} \right\}.$$

For each $U \in \mathcal{U}_A$ write $U^H AU = \Lambda_U + R_U$, where Λ_U is diagonal, and R_U is strictly upper triangular. Then by Henrici [52], the 2-*departure from normality* of A is the number

$$\Delta_2(A) \equiv \min_{U \in \mathcal{U}_A} \|R_U\|_2,$$

and by the Henrici theorem [44, Theorem 4], for any eigenvalue $\tilde{\lambda}$ of $A + E$ with $E \neq 0$, there is an eigenvalue λ of A such that

$$\left|\tilde{\lambda} - \lambda\right| \le \frac{\eta}{g(\eta)}\|E\|_2, \qquad \eta \equiv \frac{\Delta_2(A)}{\|E\|_2}, \qquad (2.4.30)$$

where $g(\eta)$ is the unique nonnegative solution of the equation

$$g + g^2 + \cdots + g^n = \eta \quad (\eta \ge 0).$$

Applying Henrici's estimate (2.4.30) to the matrices A and $A+E_{\text{opt}}$ of (2.4.27)–(2.4.28) yields the following result.

Theorem 2.4.7. *Let $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$ be an approximate invariant subspace of $A \in \mathcal{C}^{n\times n}$, and R be the residual defined by (2.4.3), where $\tilde{U}_1 \in \mathcal{U}^{n\times l}$. Moreover, let $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ be the eigenvalues of $\tilde{U}_1^H A \tilde{U}_1$. Then for any $\tilde{\lambda}_k$ $(1 \le k \le l)$ there is an eigenvalue λ_{j_k} of A such that*

$$\left|\tilde{\lambda}_k - \lambda_{j_k}\right| \le \frac{\eta}{g(\eta)} \left\|\tilde{U}_2^H R\right\|_2, \qquad \eta \equiv \frac{\Delta_2(A)}{\left\|\tilde{U}_2^H R\right\|_2},$$

where $g(\eta)$ is the unique nonnegative solution of the equation

$$g + g^2 + \cdots + g^n = \eta \qquad (\eta \ge 0).$$

Example 2.4.8. Consider the matrix A of Example 2.4.4. The vector

$$x_1 = (0,\ 0,\ -1,\ 2-\sqrt{2}i,\ -1+2\sqrt{2}i)^T$$

is an eigenvector associated with the eigenvalue $\lambda_1 = 1+\sqrt{2}i$ of A. Suppose that we have an approximate eigenvector

$$\tilde{x}_1 = (0.0001,\ 0.0000,\ -0.9999,\ 1.9999-1.4142i,\ -1.0001+2.8284i)^T,$$

and let $u_1 = x_1/\|x_1\|_2$, $\tilde{u}_1 = \tilde{x}_1/\|\tilde{x}_1\|_2$. A calculation gives

$$\sin\theta(u_1,\tilde{u}_1) \approx 4.5453\times 10^{-5}, \qquad \tan\theta(u_1,\tilde{u}_1) \approx 4.5453\times 10^{-5}, \tag{2.4.31}$$

and

$$|\tilde{u}_1^T A \tilde{u}_1 - \lambda_1| \approx 9.2454\times 10^{-5}, \tag{2.4.32}$$

where $\theta(u_1,\tilde{u}_1)$ denotes the angle between the two 1-dimensional subspaces $\mathcal{R}(u_1)$ and $\mathcal{R}(\tilde{u}_1)$.

Choose $\tilde{U}_2$ so that $(\tilde{u}_1,\tilde{U}_2) \in \mathcal{U}^{5\times 5}$. (See NR 2.4–4 for a simple algorithm for determining such a matrix $\tilde{U}_2$.) Compute

$$\tilde{A}_{11} = \tilde{u}_1^T A \tilde{u}_1, \qquad \tilde{A}_{22} = \tilde{U}_2^T A \tilde{U}_2,$$

and

$$r = \tilde{A}_{11}\tilde{u}_1 - A\tilde{u}_1, \qquad s = \tilde{A}_{11}\tilde{u}_1^H - \tilde{u}_1^H A, \qquad T = \tilde{A}_{11} I - \tilde{A}_{22}.$$

A calculation shows that $\tilde{A}_{11} \notin \lambda(\tilde{A}_{22})$, and

$$4\|T^{-1}\|_2 \|T^{-1}(\tilde{U}_2^T r)\|_2 \|s\|_2 \approx 1.3021\times 10^{-2} < 1.$$

Consequently, applying Theorem 2.4.6, there is a unit eigenvector u of A such that

$$\tan\theta(u,\tilde{u}_1) \le \frac{2\|T^{-1}(\tilde{U}_2^T r)\|_2}{1+\sqrt{1-4\|T^{-1}\|_2\|T^{-1}(\tilde{U}_2^T r)\|_2\|s\|_2}} \approx 4.5601\times 10^{-5}. \tag{2.4.33}$$

Comparing (2.4.33) with (2.4.31) shows that the estimate (2.4.33) is fairly sharp.

Observe that

$$4\|T^{-1}\|_2^2\|\tilde{U}_2^T r\|_2\|s\|_2 \approx 4.6181\times 10^{-1} < 1.$$

Hence, applying Theorem 2.3.3, there is a unit eigenvector u of A such that

$$\tan\theta(u,\tilde{u}_1) \le \frac{2\|T^{-1}\|_2\|\tilde{U}_2^T r\|_2}{1+\sqrt{1-4\|T^{-1}\|_2^2\|\tilde{U}_2^T r\|_2\|s\|_2}} \approx 1.8597\times 10^{-3},$$

which is weaker than the estimate (2.4.33).

Moreover, by Theorem 2.4.7, we have

$$\min_{\lambda_j\in\lambda(A)} |\tilde{u}_1^T A\tilde{u}_1 - \lambda_j| \lesssim 6.8769\times 10^{-1}. \tag{2.4.34}$$

Comparing (2.4.34) with (2.4.32) shows that the estimate (2.4.34) obtained by applying Theorem 2.4.7 is a severe overestimate.

For improving the estimate (2.4.34), we first prove a lemma.

Lemma 2.4.9. *Let λ be an eigenvalue of $A\in\mathcal{C}^{n\times n}$ and x be an associated eigenvector. Define $G(x,\lambda)$ by*

$$G(x,\lambda) = (\lambda I - A,\ x).$$

Then λ is simple if and only if $\operatorname{rank}(G(x,\lambda)) = n$.

Proof. Without loss of generality we may assume that the matrix $A = J$, the *Jordan canonical form* of A.

If λ is simple, then

$$J = \operatorname{diag}(\lambda, J_1), \quad \lambda\notin\lambda(J_1),$$

and

$$x = \alpha e_1^{(n)} \text{ with nonzero } \alpha\in\mathcal{C}.$$

Thus, we have

$$G(x,\lambda) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & \lambda I_{n-1} - J_1 & 0 \end{pmatrix},$$

and obviously, $\operatorname{rank}(G(x,\lambda)) = n$.

On the other hand, if λ is a multiple eigenvalue of A with multiplicity $m > 1$, and x is an associated eigenvector, then

$$J = \operatorname{diag}(J_\lambda,\ J_1),$$

where $J_\lambda \in \mathcal{C}^{m\times m}$ is in Jordan canonical form and whose eigenvalues are only $\lambda \notin \lambda(J_1)$. If J_λ only contains one Jordan block, then $x = \alpha e_1^{(n)}$ with a nonzero $\alpha \in \mathcal{C}$, and the mth row of $G(x,\lambda)$ is $(0,\ldots,0)$, and so we have $\operatorname{rank}(G(x,\lambda)) < n$. If J_λ contains at least two Jordan blocks, then there are at least two zero columns among the first m columns of $G(x,\lambda)$, and so we have $\operatorname{rank}(G(x,\lambda)) < n$, too. □

Let λ be a simple eigenvalue of $A \in \mathcal{C}^{n\times n}$, and x be an associated eigenvector. Then from Lemma 2.4.8 we see that if $(\tilde{u},\tilde{\lambda})$ is a good approximation of (x,λ), then

$$\operatorname{rank}\left(\tilde{\lambda}I - A,\ \tilde{u}\right) = n.$$

The following result gives a residual bound for an approximate eigenvalue and associated eigenvector of a matrix, in which the approximate eigenvalue approximates a simple eigenvalue of the matrix.

Theorem 2.4.10. *Let $\tilde{\lambda} \in \mathcal{C}$ be an approximate eigenvalue of $A \in \mathcal{C}^n$, and $\tilde{u} \in \mathcal{C}^n$ be an associated unit eigenvector. Define the residual r by*

$$r = \tilde{\lambda}\tilde{u} - A\tilde{u}, \tag{2.4.35}$$

and let

$$T = \left(\tilde{\lambda}I - A,\ \tilde{u}\right), \tag{2.4.36}$$

$$T^\dagger r = \begin{pmatrix} c \\ d \end{pmatrix}, \quad T^\dagger = \begin{pmatrix} W \\ z \end{pmatrix}, \quad c, z^T \in \mathcal{C}^n, \tag{2.4.37}$$

and

$$\gamma = \|c\|_2, \quad \delta = |d|, \quad \omega = \|W\|_2, \quad \zeta = \|z\|_2. \tag{2.4.38}$$

If

$$\operatorname{rank}(T) = n \tag{2.4.39}$$

and

$$(1 + \gamma\zeta - \delta\omega)^2 - 4\gamma\zeta > 0, \tag{2.4.40}$$

then there exist an eigenvalue λ of A and an associated unit eigenvector u such that

$$\sin\theta(u,\tilde{u}) \le \frac{2\gamma}{1 + \gamma\zeta - \delta\omega + \sqrt{(1 + \gamma\zeta - \delta\omega)^2 - 4\gamma\zeta}} \equiv \xi_1^* \tag{2.4.41}$$

and

$$|\tilde{\lambda} - \lambda| \le \frac{2\delta}{1 - \gamma\zeta + \delta\omega + \sqrt{(1 - \gamma\zeta + \delta\omega)^2 - 4\delta\omega}} \equiv \xi_2^*. \tag{2.4.42}$$

Proof. By theorem 1.3.2 (see (1.3.9)), we only need to prove the following conclusion: Under the conditions (2.4.39) and (2.4.40), there exist $\lambda \in \mathcal{C}$ and $x \in \mathcal{C}^n$ such that

$$Ax = \lambda x, \tag{2.4.43}$$

and

$$\|\tilde{u} - x\|_2 \le \xi_1^*, \qquad |\tilde{\lambda} - \lambda| \le \xi_2^*, \tag{2.4.44}$$

where ξ_1^* and ξ_2^* are defined by (2.4.41) and (2.4.42), respectively.

Suppose that x and λ satisfy (2.4.43). Let

$$\Delta x = \tilde{u} - x, \qquad \Delta\lambda = \tilde{\lambda} - \lambda.$$

Combining it with (2.4.35) and (2.4.43) shows that Δx and $\Delta\lambda$ satisfy

$$T \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = r + \Delta\lambda\Delta x, \tag{2.4.45}$$

where T is the matrix defined by (2.4.36).

Consider the nonlinear equation

$$\begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = T^\dagger \left(r + \Delta\lambda\Delta x\right), \tag{2.4.46}$$

where $\Delta\lambda$ and the elements of Δx are independent variables. By (2.4.39), we have $TT^\dagger = I$, so multiplying the equation (2.4.46) on the left by T yields (2.4.45). This shows that any solution of (2.4.46) is a solution of (2.4.45). Define the function f by

$$f(\Delta x, \Delta\lambda) = T^\dagger \left(r + \Delta\lambda\Delta x\right), \tag{2.4.47}$$

which can be regarded as a continuous mapping $\mathcal{M} : \mathcal{C}^{n+1} \to \mathcal{C}^{n+1}$. Since any fixed point of the mapping $\mathcal{M}$ is a solution of (2.4.46), the problem of proving (2.4.44) reduces to the problem of showing the existence of a fixed point of the continuous mapping $\mathcal{M}$ and then determining an upper bound on its size.

Let $f = (g^T, h)^T$, where $g \in \mathcal{C}^n$. Then by using (2.4.37), the mapping (2.4.47) can be expressed by

$$\begin{cases} g(\Delta x, \Delta\lambda) = c + W\Delta\lambda\Delta x, \\ \\ h(\Delta x, \Delta\lambda) = d + z\Delta\lambda\Delta x. \end{cases} \tag{2.4.48}$$

Consider the nonlinear system

$$\begin{cases} \xi_1 = \gamma + \omega\xi_1\xi_2, \\ \\ \xi_2 = \delta + \zeta\xi_1\xi_2. \end{cases} \tag{2.4.49}$$

It is easy to verify that under the condition (2.4.40), (ξ_1^*, ξ_2^*) is a solution of (2.4.49). We now define

$$\mathcal{S}_{\xi_1^*, \xi_2^*} = \left\{ \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \ : \ \|\Delta x\|_2 \le \xi_1^*, \ |\Delta\lambda| \le \xi_2^* \right\}.$$

$\mathcal{S}_{\xi_1^*,\xi_2^*}$ is obviously a bounded closed convex set of $\mathcal{C}^{n+1}$. Moreover, (2.4.48) and (2.4.49) imply

$$\begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \in \mathcal{S}_{\xi_1^*,\xi_2^*} \implies \begin{pmatrix} g(\Delta x,\Delta\lambda) \\ h(\Delta x,\Delta\lambda) \end{pmatrix} \in \mathcal{S}_{\xi_1^*,\xi_2^*},$$

which shows that the continuous mapping $\mathcal{M}$ expressed by (2.4.48) maps $\mathcal{S}_{\xi_1^*,\xi_2^*}$ into $\mathcal{S}_{\xi_1^*,\xi_2^*}$. Therefore, by the Brouwer fixed-point theorem (Theorem 1.7.1), the mapping $\mathcal{M}$ has a fixed point in $\mathcal{S}_{\xi_1^*,\xi_2^*}$. Thus, we have proved that under the conditions (2.4.39) and (2.4.40) the equation (2.4.45) has a solution $\begin{pmatrix} \Delta x_* \\ \Delta\lambda_* \end{pmatrix}$ satisfying $\|\Delta x_*\|_2 \le \xi_1^*$ and $|\Delta\lambda_*| \le \xi_2^*$. Let $x = \tilde{u} - \Delta x_*$ and $\lambda = \tilde{\lambda} - \Delta\lambda_*$, then x and λ satisfy (2.4.43) and (2.4.44). □

Since ξ_1^* and ξ_2^* satisfy (2.4.49), we can first compute ξ_1^* by (2.4.41), and then compute ξ_2^* by

$$\xi_2^* = \frac{\delta}{1-\zeta\xi_1^*}.$$

Example 2.4.11. Let $A, u_1, \lambda_1, \tilde{x}_1, \tilde{u}_1$ be as in Example 2.4.8, and let

$$\tilde{\lambda}_1 = \tilde{u}_1^T A \tilde{u}_1.$$

Suppose that we have the approximation $(\tilde{u}_1, \tilde{\lambda}_1)$. Applying Theorem 2.4.10, there exist an eigenvalue λ of A and an associated unit eigenvector u such that

$$\sin\theta(u,\tilde{u}_1) \le 5.4826 \times 10^{-5}, \qquad |\tilde{\lambda}_1 - \lambda| \le 9.2495 \times 10^{-5}. \tag{2.4.50}$$

Comparing (2.4.50) with (2.4.31) and (2.4.32) shows that the error bounds obtained by applying Theorem 2.4.10 are fairly sharp.

Example 2.4.12. Consider the matrix A of Example 2.4.4. Let $\tilde{\lambda}_j$ and $\tilde{x}_j$ be the computed eigenvalues and associated eigenvectors of A by using the MATLAB file "eig", and let $\tilde{u}_j = \tilde{x}_j/\|\tilde{x}_j\|_2$. Applying Theorem 2.4.10, there exist simple eigenvalues λ_j $(j = 1, 2, 3)$ and associated unit eigenvectors u_j of A such that

$$\sin\theta(u_1,\tilde{u}_1) \le 4.26 \times 10^{-15}, \qquad |\tilde{\lambda}_1 - \lambda_1| \le 1.02 \times 10^{-14},$$

$$\sin\theta(u_2,\tilde{u}_2) \le 4.26 \times 10^{-15}, \qquad |\tilde{\lambda}_2 - \lambda_2| \le 1.02 \times 10^{-14},$$

$$\sin\theta(u_3,\tilde{u}_3) \le 7.25 \times 10^{-16}, \qquad |\tilde{\lambda}_3 - \lambda_3| \le 3.02 \times 10^{-15},$$

which mean that the computed simple eigenvalues and associated eigenvectors of A by using the MATLAB file "eig" have very high precision. Note that $\lambda_4 = 2$ is a multiple eigenvalue of A, we cannot use (2.4.41) and (2.4.42) to give appropriate estimates of error bounds for the computed eigenvalue $\tilde{\lambda}_4$ and associated eigenvector $\tilde{x}_4$. (In fact, the condition (2.4.40) is violated for $\tilde{\lambda}_4$ and $\tilde{x}_4$.)

Remark 2.4.13. An obvious drawback of Theorem 2.4.10 is that it needs to compute the Moore-Penrose inverse of an $n \times (n+1)$ matrix. The problem of how to find nearly optimal residual bounds with less effort for computed eigenvalues and associated eigenvectors is worth studying.

Notes and References

NR 2.4–1. Theorem 2.4.1 is proved by Sun [115].

NR 2.4–2. Theorem 2.4.5 is established by Kahan, Parlett and Jiang [62, Main Theorem], and the result is used to derive a useful set of criteria for terminating the two-sided Lanczos algorithm.

NR 2.4–3. If the assumption $\tilde{X}_1^H \tilde{X}_1 = \tilde{Y}_1^H \tilde{Y}_1 = I$ of Theorem 2.4.5 is removed, then we have the following result. (The proof is left as an exercise.)

Theorem 2.4.14. *Let $\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ and $\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1)$ be the backward errors defined by (2.4.16), where $\mathcal{E}$ is the set defined by (2.4.17), and assume that the condition (2.4.15) is satisfied. Then $\mathcal{E} \neq \emptyset$ if and only if the matrices $\tilde{A}_1$ and $\tilde{C}_1$ satisfy*

$$\tilde{C}_1 = \left(\tilde{Y}_1^H \tilde{X}_1\right) \tilde{A}_1 \left(\tilde{Y}_1^H \tilde{X}_1\right)^{-1},$$

and in the case of $\mathcal{E} \neq \emptyset$, we have the formulas

$$\eta_F(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \sqrt{\left\|R\tilde{X}_1^\dagger\right\|_F^2 + \left\|\tilde{Y}_1^{\dagger^H} S\right\|_F^2 - \left\|\tilde{Y}_1^{\dagger^H} S P_{\tilde{X}_1}\right\|_F^2},$$

and

$$\eta_2(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{C}_1) = \max\left\{\left\|R\tilde{X}_1^\dagger\right\|_2, \left\|\tilde{Y}_1^{\dagger^H} S\right\|_2\right\},$$

where

$$R = \tilde{X}_1\tilde{A}_1 - A\tilde{X}_1, \qquad S = \tilde{C}_1\tilde{Y}_1^H - \tilde{Y}_1^H A$$

are the residuals.

If $\tilde{X}_1$ and $\tilde{Y}_1$ satisfy $\tilde{X}_1^H \tilde{X}_1 = \tilde{Y}_1^H \tilde{Y}_1 = I$, then Theorem 2.4.14 is reduced to 2.4.5.

Theorem 2.4.2 can be regarded as a one-sided version of Theorem 2.4.14. Whether one needs to use the one-sided or the two-sided result depends on whether one is interested in the left and right invariant subspaces simultaneously or in the right invariant subspace only.

NR 2.4–4. Suppose that $\tilde{u}_1 \in \mathcal{R}^n$ is a unit vector. We now present a simple algorithm for determining a matrix $\tilde{U}_2$ such that $(\tilde{u}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$: Let

$$v = \tilde{u}_1 - e_1^{(n)},$$

and

$$\tilde{U} = I_n - \frac{vv^T}{v^Tv}.$$

Then

$$\tilde{U} = (\tilde{u}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}.$$

In fact, the matrix $\tilde{U}$ is a Householder reflection which satisfies

$$\tilde{U}^T\tilde{U} = I_n, \qquad \tilde{U}^T = \tilde{U},$$

and

$$\tilde{U}^T\tilde{u}_1 = e_1^{(n)}.$$

(See, e.g., [48, §5.1].)

NR 2.4–5. A componentwise error bound for computed eigenvalues and associated eigenvectors is given by Yamamoto [135] and [136]. For applying Yamamoto's result it needs to compute the inverse of an $(n+1)\times(n+1)$ matrix.

NR 2.4–6. Suppose that one has an approximation for an invariant subspace of a matrix, and one also has an approximation for the corresponding eigenvalues. Haviv and Ritov [43] develop bounds on the angle between the approximating subspace and the invariant subspace itself. These bounds are functions of the following three terms: (1) the residual of the approximations; (2) singular value separation in an associated matrix; and (3) the goodness of the approximations to the eigenvalues.

2.5 Hermitian Matrices

In this section we treat perturbation analysis of the Hermitian eigenvalue problem $Ax = \lambda x$, where $A \in \mathcal{H}^{n\times n}$.

2.5.1 Perturbation Expansions

Let $p = (p_1, \ldots, p_N)^T \in \mathcal{R}^N$, and let $A(p) \in \mathcal{H}^{n\times n}$ be an analytic matrix-valued function in some neighborhood $\mathcal{B}(0)$ of the origin. It is well known that the eigenvalues of $A(p)$ are real.

Let $\lambda \in \mathcal{R}$ be a simple eigenvalue of $A(0)$, and $x \in \mathcal{C}^n$ be an associated unit eigenvector. Then there is a matrix X_2 such that

$$X = (x, X_2) \in \mathcal{U}^{n\times n},$$

and

$$X^HA(0)X = \begin{pmatrix} \lambda & 0 \\ 0 & A_2 \end{pmatrix}, \qquad \lambda_1 \notin \lambda(A_2). \qquad (2.5.1)$$

In this subsection we first apply the implicit function theorem to prove the following result which gives perturbation expansions for simple eigenvalues of a Hermitian matrix..

Theorem 2.5.1. *Let $p \in \mathcal{R}^N$ and let $A(p) \in \mathcal{H}^{n\times n}$ be an analytic matrix-valued function of p in some neighborhood $\mathcal{B}(0)$ of the origin. Suppose that λ is a simple eigenvalue of $A(0)$, and x is an associated unit eigenvector. Then*

1) there exists a simple eigenvalue $\lambda(p)$ of $A(p)$ which is a real analytic function of p in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin, and $\lambda(0) = \lambda$;

2) the function $\lambda(p)$ has a power series expansion at $p = 0$ of the form

$$\lambda(p) = \lambda + \sum_{j=1}^{N} \left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} p_j + \frac{1}{2}\sum_{j,k=1}^{N}\left(\frac{\partial^2 \lambda(p)}{\partial p_j \partial p_k}\right)_{p=0} p_j p_k + \cdots, \quad p \in \mathcal{B}_0,$$

where

$$\left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} = x^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x, \tag{2.5.2}$$

and

$$\left(\frac{\partial^2 \lambda(p)}{\partial p_j \partial p_k}\right)_{p=0} = x^H \left(\frac{\partial^2 A(p)}{\partial p_j \partial p_k}\right)_{p=0} x + x^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} \Omega_H \left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} x$$
$$+ x^H \left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} \Omega_H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x, \tag{2.5.3}$$

in which

$$\Omega_H = X_2(\lambda I - A_2)^{-1} X_2^H. \tag{2.5.4}$$

Proof. 1) By the hypotheses there is a matrix $X = (x, X_2) \in \mathcal{U}^{n\times n}$ such that the relation (2.5.1) holds. For $p \in \mathcal{B}(0)$ we set

$$\tilde{A}(p) = X^H A(p) X = \begin{pmatrix} \tilde{a}_{11}(p) & \tilde{a}_{21}(p)^H \\ \tilde{a}_{21}(p) & \tilde{A}_{22}(p) \end{pmatrix}, \quad \tilde{a}_{11}(p) \in \mathcal{R}, \tag{2.5.5}$$

and introduce a vector-valued function

$$f(z,p) = \tilde{a}_{21}(p) - \tilde{a}_{11}(p) z + \tilde{A}_{22}(p) z - z \tilde{a}_{21}(p)^H z, \tag{2.5.6}$$

where

$$f = (f_1, \ldots, f_{n-1})^T, \quad z = (\zeta_1, \ldots, \zeta_{n-1})^T \in \mathcal{C}^{n-1}, \quad p \in \mathcal{B}(0).$$

Let

$$f_j = \phi_j + i\psi_j, \quad \zeta_j = \mu_j + i\nu_j, \quad i = \sqrt{-1}, \quad j = 1, \ldots, n-1,$$

and

$$u = (\mu_1, \ldots, \mu_{n-1})^T, \; v = (\nu_1, \ldots, \nu_{n-1})^T \in \mathcal{R}^{n-1}.$$

Obviously, $\phi_j(u, v, p)$ and $\psi_j(u, v, p)$ are real analytic functions of the real variables $u, v \in \mathcal{R}^{n-1}$ and $p \in \mathcal{B}(0)$, and the functions satisfy

$$\phi(0, 0, 0) = 0, \quad \psi(0, 0, 0) = 0, \quad j = 1, \ldots, n-1.$$

Since $f_1, \ldots, f_{n-1}$ are complex analytic functions of the complex variables $\zeta_1, \ldots, \zeta_{n-1}$ for any $p \in \mathcal{B}(0)$, by Theorem 1.6.3 we have

$$\left(\frac{\partial(\phi_1, \ldots, \phi_{n-1}, \psi_1, \ldots, \psi_{n-1})}{\partial(\mu_1, \ldots, \mu_{n-1}, \nu_1, \ldots, \nu_{n-1})} \right)_{u=v=0,\, p=0}$$

$$= \left| \frac{\partial(f_1, \ldots, f_{n-1})}{\partial(\zeta_1, \ldots, \zeta_{n-1})} \right|^2_{z=0,\, p=0} = |\det(A_2 - \lambda I)|^2 > 0.$$

Therefore, by the implicit function theorem (Theorem 1.6.2) the system of equations

$$\phi_j(u, v, p) = 0, \quad \psi_j(u, v, p) = 0, \quad j = 1, \ldots, n-1$$

has a unique real analytic solution $u = u(p)$, $v = v(p)$ in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin, and $u(0) = v(0) = 0$. In other words, the equation $f(z, p) = 0$ has a unique analytic solution $z = z(p)$ in $\mathcal{B}_0$, and $z(0) = 0$. Moreover, we may choose $\mathcal{B}_0$ so small that $1 + z(p)^H z(p) > 0$ for any $p \in \mathcal{B}_0$.

Define

$$Q(p) = \begin{pmatrix} 1 & -z(p)^H \\ z(p) & I \end{pmatrix} \begin{pmatrix} \left(1 + z(p)^H z(p)\right)^{-\frac{1}{2}} & 0 \\ 0 & \left(I + z(p) z(p)^H\right)^{-\frac{1}{2}} \end{pmatrix} \in \mathcal{U}^{n \times n}.$$

Then we have

$$Q(p)^H \tilde{A}(p) Q(p) = \begin{pmatrix} \lambda(p) & 0 \\ 0 & * \end{pmatrix}, \quad p \in \mathcal{B}_0.$$

Combining it with (2.5.5) gives

$$A(p) x(p) = \lambda(p) x(p), \tag{2.5.7}$$

where

$$\begin{aligned} \lambda(p) \;&= \left(\tilde{a}_{11}(p) + z(p)^H \tilde{a}_{21}(p) + \tilde{a}_{21}(p)^H z(p) + z(p)^H \tilde{A}_{22}(p) z(p)\right) \\ &\quad \times \left(1 + z(p)^H z(p)\right)^{-1}, \end{aligned} \tag{2.5.8}$$

and

$$x(p) = X \begin{pmatrix} 1 \\ z(p) \end{pmatrix} \left(1 + z(p)^H z(p)\right)^{-\frac{1}{2}}, \tag{2.5.9}$$

in which the analytic vector-valued function $z(p)$ satisfies the equation

$$\tilde{a}_{21}(p) - \tilde{a}_{11}(p)z(p) + \tilde{A}_{22}(p)z(p) - z(p)\tilde{a}_{21}(p)^H z(p) = 0, \quad p \in \mathcal{B}_0. \tag{2.5.10}$$

From (2.5.7)–(2.5.9) we see the following facts: (i) $\lambda(p)$ is an eigenvalue of $A(p)$ and $x(p)$ is an associated eigenvector; (ii) $\lambda(p)$ and $x(p)$ are analytic functions of p in $\mathcal{B}_0$, and $\lambda(0) = \lambda$, $x(0) = x$; (iii) the eigenvalue $\lambda(p)$ of $A(p)$ is simple in $\mathcal{B}_0$ provided that $\mathcal{B}_0$ is sufficiently small.

2) From (2.5.8), (2.5.5) and $\tilde{a}_{21}(0) = z(0) = 0$, we get

$$\left(\frac{\partial \lambda(p)}{\partial p_j}\right)_{p=0} = \left(\frac{\partial \tilde{a}_{11}(p)}{\partial p_j}\right)_{p=0}, \tag{2.5.11}$$

and

$$\begin{aligned}\left(\frac{\partial^2 \lambda(p)}{\partial p_j \partial p_k}\right)_{p=0} &= \left(\frac{\partial^2 \tilde{a}_{11}(p)}{\partial p_j \partial p_k}\right)_{p=0} \\ &+ \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}^H \left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_k}\right)_{p=0} + \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0}^H \left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_j}\right)_{p=0} \\ &+ \left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_j}\right)_{p=0}^H \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0} + \left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_k}\right)_{p=0}^H \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} \\ &+ \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}^H A_2 \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0} + \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0}^H A_2 \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} \\ &- \lambda_1 \left[\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}^H \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0} + \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0}^H \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}\right].\end{aligned} \tag{2.5.12}$$

Moreover, from (2.5.10)

$$\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} = (\lambda I - A_2)^{-1} X_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x. \tag{2.5.13}$$

Combining (2.5.11)–(2.5.13) with

$$\left(\frac{\partial \tilde{a}_{11}(p)}{\partial p_j}\right)_{p=0} = x^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x,$$

$$\left(\frac{\partial^2 \tilde{a}_{11}(p)}{\partial p_j \partial p_k}\right)_{p=0} = x^H \left(\frac{\partial^2 A(p)}{\partial p_j \partial p_k}\right)_{p=0} x,$$

$$\left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_j}\right)_{p=0} = X_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x,$$

and using the relation

$$X_2(\lambda I - A_2)^{-1} X_2^H + X_2(\lambda I - A_2)^{-1} A_2 (\lambda I - A_2)^{-1} X_2^H$$

$$= \lambda X_2 (\lambda I - A_2)^{-2} X_2^H,$$

we obtain the formulas (2.5.2)–(2.5.4). □

Note that the relations (2.5.9) and (2.5.13) imply that the eigenvector $x(p)$ has the expansion of the form

$$x(p) = x + \Omega_H \sum_{j=1}^{N} \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x p_j + \cdots, \quad p \in \mathcal{B}_0,$$

where Ω_H is the matrix defined by (2.5.4).

Example 2.5.2. Consider the Hermitian matrix

$$A(p) = \begin{pmatrix} 2 & \frac{1}{p_1 - ip_2 + 1} \\ \frac{1}{p_1 + ip_2 + 1} & 2 \end{pmatrix}, \quad (p_1, p_2)^T = p \in \mathcal{R}^2, \quad i = \sqrt{-1}.$$

Obviously, $A(p)$ is an analytic matrix-valued function of p in a small neighborhood of the origin of $\mathcal{R}^2$. Moreover,

$$A(0) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

and the real orthogonal matrix

$$X = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \equiv (x_1, x_2)$$

satisfies

$$X^T A(0) X = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \equiv \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Observe that

$$\left(\frac{\partial A(p)}{\partial p_1}\right)_{p=0} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \quad \left(\frac{\partial A(p)}{\partial p_2}\right)_{p=0} = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix},$$

$$\left(\frac{\partial^2 A(p)}{\partial p_1^2}\right)_{p=0} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}, \qquad \left(\frac{\partial^2 A(p)}{\partial p_2^2}\right)_{p=0} = \begin{pmatrix} 0 & -2 \\ -2 & 0 \end{pmatrix},$$

and

$$\left(\frac{\partial^2 A(p)}{\partial p_1 \partial p_2}\right)_{p=0} = \begin{pmatrix} 0 & -2i \\ 2i & 0 \end{pmatrix}.$$

Hence, if $\lambda_1(p)$ and $\lambda_2(p)$ denote the eigenvalues of $A(p)$, then by the formulas (2.5.2)–(2.5.4) we have

$$\left(\frac{\partial \lambda_1(p)}{\partial p_1}\right)_{p=0} = -1, \qquad \left(\frac{\partial \lambda_1(p)}{\partial p_2}\right)_{p=0} = 0,$$

$$\left(\frac{\partial \lambda_2(p)}{\partial p_1}\right)_{p=0} = 1, \qquad \left(\frac{\partial \lambda_2(p)}{\partial p_2}\right)_{p=0} = 0,$$

and

$$\left(\frac{\partial^2 \lambda_1(p)}{\partial p_1^2}\right)_{p=0} = 2, \quad \left(\frac{\partial^2 \lambda_1(p)}{\partial p_1 \partial p_2}\right)_{p=0} = 0, \quad \left(\frac{\partial^2 \lambda_1(p)}{\partial p_2^2}\right)_{p=0} = -1,$$

$$\left(\frac{\partial^2 \lambda_2(p)}{\partial p_1^2}\right)_{p=0} = -2, \quad \left(\frac{\partial^2 \lambda_2(p)}{\partial p_1 \partial p_2}\right)_{p=0} = 0, \quad \left(\frac{\partial^2 \lambda_2(p)}{\partial p_2^2}\right)_{p=0} = 1.$$

Consequently, $\lambda_1(p)$ and $\lambda_2(p)$ have the expansions

$$\lambda_1(p) = 3 - p_1 + p_1^2 - \frac{1}{2}p_2^2 + O(\|p\|_2^3) \tag{2.5.14}$$

and

$$\lambda_2(p) = 1 + p_1 - p_1^2 + \frac{1}{2}p_2^2 + O(\|p\|_2^3) \tag{2.5.15}$$

as $p \to 0$.

Note that the eigenvalues $\lambda_1(p)$ and $\lambda_2(p)$ have the explicit expressions

$$\lambda_1(p) = 2 + \frac{1}{\sqrt{(1+p_1)^2 + p_2^2}}, \qquad \lambda_2(p) = 2 - \frac{1}{\sqrt{(1+p_1)^2 + p_2^2}}.$$

From the expressions we can also obtain the second order perturbation expansions (2.5.14) and (2.5.15).

Let $A \in \mathcal{H}^{n\times n}$ and let $\mathcal{X}_1 \subset \mathcal{C}^n$. If

$$\dim(\mathcal{X}_1) = l \quad \text{and} \quad A\mathcal{X}_1 \subset \mathcal{X}_1,$$

then $\mathcal{X}_1$ is said to be an l-dimensional eigenspace of A.

The eigenspace $\mathcal{X}_1$ can be equivalently defined by $\mathcal{X}_1 = \mathcal{R}(X_1)$ with X_1 satisfying

$$X_1 \in \mathcal{U}^{n\times l}, \quad \text{and} \quad AX_1 = X_1 A_1$$

for some $A_1 \in \mathcal{H}^{l\times l}$.

Let $X_1 \in \mathcal{U}^{n\times l}$. It can be verified that the subspace $\mathcal{X}_1 = \mathcal{R}(X_1)$ is an *eigenspace* of $A \in \mathcal{H}^{n\times n}$ if and only if there exists a matrix $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ such that

$$X^H A X = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \qquad A_{11} \in \mathcal{H}^{l\times l}. \tag{2.5.16}$$

If $\lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset$, then the eigenspace $\mathcal{X}_1$ is called a *simple* eigenspace of A. In this section we only consider simple eigenspaces.

Using the same technique described in the proofs of Theorems 2.1.5 and 2.5.1, we obtain the following result which gives perturbation expansions for eigenspaces.

Theorem 2.5.3. *Let $A \in \mathcal{H}^{n\times n}$, and let $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ with $X_1 \in \mathcal{U}^{n\times l}$ such that*

$$X^H A X = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \qquad A_{11} \in \mathcal{H}^{l\times l}, \qquad \lambda(A_{11}) \bigcap \lambda(A_{22}) = \emptyset.$$

Moreover, let $\mathcal{X}_1 = \mathcal{R}(X_1)$, for $H \in \mathcal{H}^{n\times n}$ let

$$X^H H X = \begin{pmatrix} H_{11} & H_{21}^H \\ H_{21} & H_{22} \end{pmatrix},$$

and define a linear operator $\mathbf{T} : \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{T}P = PA_{11} - A_{22}P, \qquad P \in \mathcal{C}^{(n-l)\times l}. \tag{2.5.17}$$

Then

1) there exists a unique l-dimensional simple eigenspace $\mathcal{X}_1(\tau)$ of $A + \tau H$ such that $\mathcal{X}_1(0) = \mathcal{X}_1$, and the basis vectors $x_1(\tau), \ldots, x_l(\tau)$ of $\mathcal{X}_1(\tau)$ may be defined to be analytic functions of τ in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{R}$;

2) the analytic matrix-valued function $X_1(\tau) = (x_1(\tau), \ldots, x_l(\tau))$ has the perturbation expansion

$$X_1(\tau) = X_1 + X_2 \sum_{j=1}^{\infty} K_j \tau^j, \qquad \tau \in \mathcal{B}(0),$$

in which

$$K_1 = \mathbf{T}^{-1} H_{21},$$

$$K_2 = \mathbf{T}^{-1}(H_{22}K_1 - K_1 H_{11}),$$

$$K_j = \mathbf{T}^{-1}\left[H_{22}K_{j-1} - K_{j-1}H_{11} - \sum_{k=1}^{j-2} K_{j-1-k} H_{12} K_k\right], \quad j \geq 3.$$

2.5.2 Structured Condition Numbers

Let $A \in \mathcal{H}^{n\times n}$, and λ be a simple eigenvalue of A. Let $\tilde{A} = A + H \in \mathcal{H}^{n\times n}$ be a perturbation of A, and $\tilde{\lambda}$ be the corresponding perturbation of λ. Then by §1.8 we define the structured condition number $c(\lambda)$ for λ as (2.2.1), but the perturbation matrices $E \in \mathcal{C}^{n\times n}$ are replaced by $H \in \mathcal{H}^{n\times n}$.

Let $x \in \mathcal{C}^n$ be the unit eigenvector of A associated with λ. Then by Theorem 2.5.1 we have

$$\tilde{\lambda} = \lambda + x^H Hx + O(\|H\|^2).$$

Combining it with the definition (2.2.1) yields

$$c(\lambda) = \frac{\alpha}{\xi},$$

where α and ξ are positive parameters. Obviously, we have the absolute condition number $c_{\text{abs}}(\lambda) = 1$, and the relative condition number $c_{\text{rel}}(\lambda) = \|A\|/|\lambda|$ if $\lambda \neq 0$.

Let $\mathcal{X}_1$ be a simple eigenspace of $A \in \mathcal{H}^{n\times n}$. Let $\tilde{A} = A + H \in \mathcal{H}^{n\times n}$ be a perturbation of A, and $\tilde{\mathcal{X}}_1$ be the corresponding perturbation of $\mathcal{X}_1$. Then by §1.8 we define the structured condition number $c(\mathcal{X}_1)$ for $\mathcal{X}_1$ as (2.2.10), but the perturbation matrices $E \in \mathcal{C}^{n\times n}$ are replaced by $H \in \mathcal{H}^{n\times n}$. By the same argument as in §2.2,2, we obtain

$$c(\mathcal{X}_1) = \alpha\|T^{-1}\|_2,$$

where α is a positive parameter, and T is the matrix representation of the linear operator $\mathbf{T}$ defined by (2.5.17).

Let

$$\lambda(A_{11}) = \{\lambda_1, \ldots, \lambda_l\}, \qquad \lambda(A_{22}) = \{\lambda_{l+1}, \ldots, \lambda_n\}.$$

Then $c(\mathcal{X}_1)$ has the expression

$$c(\mathcal{X}_1) = \frac{\alpha}{\min\limits_{\substack{1 \leq j \leq l \\ l+1 \leq k \leq n}} |\lambda_j - \lambda_k|}.$$

2.5.3 Perturbation Bounds for Eigenspaces

In this subsection we give perturbation bounds for eigenspaces. The proofs of the following three results are similar to those of Theorem 2.3.1, Corollary 2.3.2 and Theorem 2.3.3, and left as exercises.

Theorem 2.5.4. *Let $A \in \mathcal{H}^{n\times n}$. Let $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$, and suppose that $\mathcal{X}_1 = \mathcal{R}(X_1)$ is an l-dimensional simple eigenspace of A and (2.5.16) holds. For a*

Hermitian perturbation H we let

$$X^H H X = \begin{pmatrix} H_{11} & H_{21}^H \\ H_{21} & H_{22} \end{pmatrix},$$

and assume that $\lambda(A_{11} + H_{11}) \bigcap \lambda(A_{22} + H_{22}) = \emptyset$. Define the linear operator $\mathbf{L} : \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{L}Z = Z(A_{11} + H_{11}) - (A_{22} + H_{22})Z, \qquad Z \in \mathcal{C}^{(n-l)\times l},$$

and set

$$b = \|\mathbf{L}^{-1} H_{21}\|, \qquad c = \|\mathbf{L}^{-1}\|, \qquad \eta = \|H_{21}\|_2.$$

If

$$4bc\eta < 1,$$

then there is a unique l-dimensional eigenspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A + H$ satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\| \leq \frac{2b}{1 + \sqrt{1 - 4bc\eta}},$$

where $\tilde{X}_1 \in \mathcal{U}^{n\times l}$.

From Theorem 2.5.4 we get the following corollary.

Corollary 2.5.5. *Let $A, H, X, \mathcal{X}_1$ be as in Theorem 2.5.4, and assume that*

$$H_{11} = 0, \qquad H_{22} = 0.$$

Define the linear operator $\mathbf{T} : \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{T}Z = ZA_{11} - A_{22}Z, \qquad Z \in \mathcal{C}^{(n-l)\times l}.$$

If

$$4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}H_{21}\|\|H_{21}\|_2 < 1,$$

then there is a unique l-dimensional eigenspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A + H$ satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\| \leq \frac{2\|\mathbf{T}^{-1}H_{21}\|}{1 + \sqrt{1 - 4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}H_{21}\|\|H_{21}\|_2}},$$

where $\tilde{X}_1 \in \mathcal{U}^{n\times l}$.

Moreover, if the perturbation matrix H itself is unknown but some upper bounds for $\|H_{jk}\|$ are known, then we have the following well known result.

Theorem 2.5.6 (Stewart). *Let $A, H, X, \mathcal{X}_1$ be as in Theorem 2.5.4, and let* $\mathbf{T}$ *be the linear operator as in Corollary 2.5.5. Set*

$$\tilde{c} = \frac{\|\mathbf{T}^{-1}\|}{1 - \|\mathbf{T}^{-1}\|(\|H_{11}\| + \|H_{22}\|)}, \qquad \gamma = \|H_{21}\|.$$

If

$$2\tilde{c}\gamma < 1,$$

then there is a unique l-dimensional eigenspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ of $A + H$ satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\| \le \frac{2\tilde{c}\gamma}{1 + \sqrt{1 - (2\tilde{c}\gamma)^2}},$$

where $\tilde{X}_1 \in \mathcal{U}^{n\times l}$.

2.5.4 Structured Backward Errors

2.5.4.1 The Backward Error $\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1)$

Let $\tilde{\mathcal{X}}_1$ approximate an l-dimensional eigenspace of $A \in \mathcal{H}^{n\times n}$. Then there are various ways to define backward errors of A with respect to $\tilde{\mathcal{X}}_1$. For example, we may define the backward error $\eta(\tilde{\mathcal{X}}_1)$ by (2.4.1), in which the set $\mathcal{E}$ consists of backward general perturbations $E \in \mathcal{C}^{n\times n}$ of A. An explicit expression of $\eta(\tilde{\mathcal{X}}_1)$ is given by (2.4.4).

The expression (2.4.4) gives a distance from the Hermitian matrix A to the nearest matrix $A + E_{\mathrm{opt}}$ for which the given approximate eigenspace $\tilde{\mathcal{X}}_1$ of A is an exact invariant subspace of $A + E_{\mathrm{opt}}$. However, from the expression (2.4.7) of the optimal backward (general) perturbation E_{opt} we see that the perturbed matrix $A+E_{\mathrm{opt}}$ may not be Hermitian. Consequently, if we are interested in the requirement that the perturbed matrices are Hermitian too, then the definition (2.4.1) has to be modified.

We now define the structured backward error $\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1)$ of A with respect to $\tilde{\mathcal{X}}_1$ by

$$\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1) = \min_{H\in\mathcal{H}} \|H\|, \tag{2.5.18}$$

where the set $\mathcal{H}$ is defined by

$$\mathcal{H} = \left\{ H \in \mathcal{H}^{n\times n} \ : \ (A + H)\tilde{\mathcal{X}}_1 \subset \tilde{\mathcal{X}}_1 \right\}. \tag{2.5.19}$$

The following result gives a computable formula of $\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1)$.

Theorem 2.5.7. *Choose $\tilde{U}_1 \in \mathcal{U}^{n\times l}$ so that $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$. Let*

$$R = \tilde{U}_1(\tilde{U}_1^H A\tilde{U}_1) - A\tilde{U}_1 \tag{2.5.20}$$

be the residual of A with respect to $\tilde{U}_1$. Then the backward error $\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1)$ can be expressed by

$$\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1) = \left\| \begin{pmatrix} 0 & R^H \\ R & 0 \end{pmatrix} \right\|. \tag{2.5.21}$$

The expressions (2.5.20) and (2.5.21) imply that the backward error $\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1)$ defined by (2.5.18) is independent of the choice of the matrix $\tilde{U}_1$ whose column vectors form an orthonormal basis of $\tilde{\mathcal{X}}_1$.

Proof of Theorem 2.5.7. From (2.5.19) it follows that a matrix $H \in \mathcal{H}$ if and only if H is a solution of the equation

$$(A+H)\tilde{U}_1 = \tilde{U}_1 A_1$$

for some $A_1 \in \mathcal{H}^{l\times l}$, or equivalently, H satisfies

$$H\tilde{U}_1 = \tilde{U}_1 A_1 - A\tilde{U}_1. \tag{2.5.22}$$

Applying Theorem 1.5.2 to the equation (2.5.22) we see that the equation is solvable, and any solution H to the equation can be expressed by

$$\begin{aligned} H \;&= (\tilde{U}_1 A_1 - A\tilde{U}_1)\tilde{U}_1^H + \tilde{U}_1(\tilde{U}_1 A_1 - A\tilde{U}_1)^H \\ &-\tilde{U}_1(A_1 - \tilde{U}_1^H A\tilde{U}_1)\tilde{U}_1^H + P_{\tilde{U}_1}^{\perp} T P_{\tilde{U}_1}^{\perp}, \end{aligned} \tag{2.5.23}$$

where $T \in \mathcal{H}^{n\times n}$.

Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Then from (2.5.23)

$$\tilde{U}^H H \tilde{U} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{U}_1 & -\tilde{U}_1^H A\tilde{U}_2 \\ -\tilde{U}_2^H A\tilde{U}_1 & \tilde{U}_2^H T\tilde{U}_2 \end{pmatrix} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{U}_1 & R^H\tilde{U}_2 \\ \tilde{U}_2^H R & \tilde{U}_2 T\tilde{U}_2 \end{pmatrix}.$$

By the definition (2.5.18) and Theorem 1.2.1, we have

$$\eta_{\mathrm{H}}(\tilde{\mathcal{X}}_1) = \|H_{\mathrm{opt}}\| \quad \text{with} \quad H_{\mathrm{opt}} = \tilde{U}\begin{pmatrix} 0 & R^H\tilde{U}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix}\tilde{U}^H. \tag{2.5.24}$$

Observe that the relation

$$\tilde{U}^H R = \begin{pmatrix} 0 \\ \tilde{U}_2^H R \end{pmatrix}$$

implies

$$\sigma_+(\tilde{U}_2^H R) = \sigma_+(R), \qquad \sigma_+(R^H\tilde{U}_2) = \sigma_+(R).$$

Hence, we have

$$\sigma_+\begin{pmatrix} 0 & R^H\tilde{U}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix} = \sigma_+\begin{pmatrix} 0 & R^H \\ R & 0 \end{pmatrix}$$

Combining it with (2.5.24) shows (2.5.21). □

Comparing the optimal backward Hermitian perturbation $H_{\rm opt}$ of (2.5.24) with the optimal backward general perturbation $E_{\rm opt}$ of (2.4.7) we obtain the following Corollary.

Corollary 2.5.8. *Let $A \in \mathcal{H}^{n\times n}$, and $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$ approximate an eigenspace of A. Moreover, let $H_{\rm opt}$ be the optimal backward Hermitian perturbation , and $E_{\rm opt}$ the optimal backward general perturbation. Then we have*

$$\|E_{\rm opt}\| \le \|H_{\rm opt}\| \le 2\|E_{\rm opt}\|;$$

and particularly,

$$\|H_{\rm opt}\|_F = \sqrt{2}\|E_{\rm opt}\|_F, \qquad \|H_{\rm opt}\|_2 = \|E_{\rm opt}\|_2.$$

2.5.4.2 The Backward Errors $\eta_{\rm H,F}(\tilde{U}_1, \tilde{\Lambda}_1)$ and $\eta_{\rm H,2}(\tilde{U}_1, \tilde{\Lambda}_1)$

Let $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l \in \mathcal{R}$ $(l \le n)$ be approximate eigenvalues of $A \in \mathcal{H}^{n\times n}$, and $\tilde{x}_1, \ldots, \tilde{x}_l$ be associated approximate eigenvectors. Generally speaking, the approximate eigenvectors are linearly independent but not necessarily orthonormal. An important question associated with the approximate eigensystem is: How can we define a backward error of A with respect to the approximate eigensystem, and how can we obtain a computable formula of the backward error? In this subsection we discuss the question.

Assume that the vectors $\tilde{x}_1, \ldots, \tilde{x}_l$ are close to orthonormal, i.e., the matrix $\tilde{X}_1 = (\tilde{x}_1, \ldots, \tilde{x}_l)$ satisfies

$$\epsilon \equiv \|\tilde{X}_1^H \tilde{X}_1 - I\|_F \ll 1. \tag{2.5.25}$$

Let

$$\tilde{X}_1 = \tilde{U}_1 \tilde{L}_1 \tag{2.5.26}$$

be an orthogonal decomposition of $\tilde{X}_1$, i.e., $\tilde{U}_1 \in \mathcal{U}^{n\times l}$, and $\tilde{L}_1 \in \mathcal{C}^{l\times l}$ is nonsingular. If the column vectors of $\tilde{U}_1$ are approximate eigenvectors of A associated with $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$, then through the orthogonal decomposition (2.5.26) we may define the backward errors $\eta_{\rm H,F}(\tilde{U}_1, \tilde{\Lambda}_1)$ and $\eta_{\rm H,2}(\tilde{U}_1, \tilde{\Lambda}_1)$ of A with respect to $\tilde{U}_1$ and $\tilde{\Lambda}_1$ by

$$\eta_{\rm H,F}(\tilde{U}_1, \tilde{\Lambda}_1) = \min_{H\in\mathcal{H}} \|H\|_F, \qquad \eta_{\rm H,2}(\tilde{U}_1, \tilde{\Lambda}_1) = \min_{H\in\mathcal{H}} \|H\|_2, \tag{2.5.27}$$

where the set $\mathcal{H}$ is defined by

$$\mathcal{H} = \left\{ H \in \mathcal{H}^{n\times n} \ : \ (A+H)\tilde{U}_1 = \tilde{U}_1 \tilde{\Lambda}_1 \right\}. \tag{2.5.28}$$

Note that there exist orthogonal decompositions (2.5.26) of $\tilde{X}_1$ such that under the hypothesis (2.5.25) the column vectors of $\tilde{U}_1$ are approximate eigenvectors of

A associated with $\tilde{\lambda}_1, \dots, \tilde{\lambda}_l$. For example, the polar decomposition and the QR factorization of $\tilde{X}_1$ are such decompositions. In fact, if

$$\tilde{X}_1 = P_1 H_1, \qquad \tilde{X}_1 = Q_1 R_1$$

are the polar decomposition and the QR factorization of $\tilde{X}_1$, respectively, where $P_1, Q_1 \in \mathcal{U}^{n\times l}$, $H_1 \in \mathcal{C}^{l\times l}$ is Hermitian positive definite, and $R_1 \in \mathcal{C}^{l\times l}$ is upper triangular with positive diagonal elements, then by Theorem 1.4.2 we have

$$\|\tilde{X} - P_1\|_F \le \frac{\epsilon}{1+\sigma_{\min}(\tilde{X}_1)} \le \frac{\epsilon}{1+\sqrt{1-\epsilon}} \ll 1, \tag{2.5.29}$$

and

$$\begin{aligned}\|\tilde{X}_1 - Q_1\|_F \quad &\le \frac{\sqrt{2}\left(1+\|\tilde{X}_1\|_2\right)\epsilon}{\left(1+\sigma_{\min}(\tilde{X}_1)\right)\left(1-\|\tilde{X}_1^H\tilde{X}_1 - I\|_2 + \sqrt{1-\|\tilde{X}_1^H\tilde{X}_1 - I\|_2}\right)} \\ &\le \frac{\sqrt{2}(1+\sqrt{1+\epsilon})\epsilon}{(1+\sqrt{1-\epsilon})(1-\epsilon+\sqrt{1-\epsilon})} \ll 1.\end{aligned} \tag{2.5.30}$$

The relations (2.5.29) and (2.5.30) show that if the approximate eigenvectors $\tilde{x}_1, \dots, \tilde{x}_l$ of A are close to orthonormal, then the column vectors of P_1 and Q_1 are also approximate eigenvectors of A associated with the same eigenvalues.

Let $\eta_{\mathrm{H,F}}(\tilde{U}_1, \tilde{\Lambda}_1)$ and $\eta_{\mathrm{H,2}}(\tilde{U}_1, \tilde{\Lambda}_1)$ be the backward errors defined by (2.5.27). The following result gives computable formulas of the backward errors $\eta_{\mathrm{H,F}}(\tilde{U}_1, \tilde{\Lambda}_1)$ and $\eta_{\mathrm{H,2}}(\tilde{U}_1, \tilde{\Lambda}_1)$.

Theorem 2.5.9. *Let $A \in \mathcal{H}^{n\times n}$, and let $\tilde{U}_1 \in \mathcal{U}^{n\times l}$ and $\tilde{\Lambda}_1 = \mathrm{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_l)$ be given, where $\tilde{\lambda}_j \in \mathcal{R}$ $(1 \le j \le l)$. Moreover, let*

$$R = \tilde{U}_1\tilde{\Lambda}_1 - A\tilde{U}_1 \tag{2.5.31}$$

be the residuals of A with respect to $\tilde{U}_1$ and $\tilde{\Lambda}_1$. Then the backward errors $\eta_{\mathrm{H,F}}(\tilde{U}_1, \tilde{\Lambda}_1)$ and $\eta_{\mathrm{H,2}}(\tilde{U}_1, \tilde{\Lambda}_1)$ can be expressed by

$$\eta_{\mathrm{H,F}}(\tilde{U}_1, \tilde{\Lambda}_1) = \sqrt{2\|R\|_F^2 - \|\tilde{U}_1^H R\|_F^2}, \tag{2.5.32}$$

and

$$\eta_{\mathrm{H,2}}(\tilde{U}_1, \tilde{\Lambda}_1) = \|R\|_2. \tag{2.5.33}$$

Proof. From (2.5.28) it follows that a matrix $H \in \mathcal{H}$ if and only if H satisfies

$$H\tilde{U}_1 = R, \tag{2.5.34}$$

where R is the residual defined by (2.5.31).

Applying Theorem 1.5.2 to the equation (2.5.34) we see that the equation is solvable, and any solution H of the equation can be expressed by

$$H = R\tilde{U}_1^H + \tilde{U}_1 R^H - \tilde{U}_1 R^H \tilde{U}_1 \tilde{U}_1^H + P_{\tilde{U}_1}^{\perp} T P_{\tilde{U}_1}^{\perp}, \tag{2.5.35}$$

where $T \in \mathcal{H}^{n\times n}$.

Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Then from (2.5.35)

$$\tilde{U}^H H \tilde{U} = \begin{pmatrix} \tilde{U}_1^H R & R^H \tilde{U}_2 \\ \tilde{U}_2^H R & \tilde{U}_2^H T \tilde{U}_2 \end{pmatrix}. \tag{2.5.36}$$

Consequently, by the definition (2.5.27) we have

$$\eta_{\mathrm{H,F}}(\tilde{U}_1, \tilde{\Lambda}_1) = \|H_{\mathrm{opt}}\| \quad \text{with} \quad H_{\mathrm{opt}} = \tilde{U} \begin{pmatrix} \tilde{U}_1^H R & R^H \tilde{U}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix} \tilde{U}^H,$$

which shows (2.5.32).

Moreover, by (2.5.27), (2.5.36) and Theorem 1.2.3, we have

$$\eta_{\mathrm{H},2}(\tilde{U}_1, \tilde{\Lambda}_1) = \max\{\|R\|_2,\ \|(\tilde{U}_1^H R,\ R^H \tilde{U}_2)\|\}.$$

Combining it with

$$\|(\tilde{U}_1^H R,\ R^H \tilde{U}_2)\| = \|R^H (\tilde{U}_1, \tilde{U}_2)\|_2 = \|R\|_2,$$

shows (2.5.33). □

Remark 2.5.10. Let $\tilde{\lambda}_1 \in \mathcal{R}$ be an approximate eigenvalue of $A \in \mathcal{H}^{n\times n}$, and $\tilde{u}_1 \in \mathcal{C}^n$ be an associated unit eigenvector. Then the backward errors $\eta_{\mathrm{H,F}}(\tilde{u}_1, \tilde{\lambda}_1)$ and $\eta_{\mathrm{H},2}(\tilde{u}_1, \tilde{\lambda}_1)$ of A with respect to $\tilde{x}_1$ and $\tilde{\lambda}_1$ has the expressions

$$\eta_{\mathrm{H,F}}(\tilde{u}_1, \tilde{\lambda}_1) = \sqrt{2\|r\|_2^2 - |\tilde{u}_1^H r|^2}, \qquad \eta_{\mathrm{H},2}(\tilde{u}_1, \tilde{\lambda}_1) = \|r\|_2,$$

where

$$r = \tilde{\lambda}_1 \tilde{u}_1 - A\tilde{u}_1$$

is the residual of A with respect to $\tilde{x}_1$ and $\tilde{\lambda}_1$. Moreover, the optimal backward Hermitian perturbation H_{opt} of A associated with $\eta_{\mathrm{H,F}}(\tilde{u}_1, \tilde{\lambda}_1)$ has the expression

$$H_{\mathrm{opt}} = r\tilde{u}_1^H + \tilde{u}_1 r^H - \left(\tilde{\lambda}_1 - \tilde{u}_1^H A \tilde{u}_1\right) \tilde{u}_1 \tilde{u}_1^H.$$

Let P_1 be the unitary polar factor of $\tilde{X}_1$. Take $\tilde{U}_1 = P_1$ in (2.5.28). We now discuss the backward error $\eta_{\mathrm{H,F}}(P_1, \tilde{\Lambda}_1)$.

The following result shows that the Frobenius norm of the residual $\tilde{X}_1\tilde{\Lambda}_1 - A\tilde{X}_1$ can be used to bound the backward error $\eta_{\mathrm{H,F}}(P_1, \tilde{\Lambda}_1)$.

Theorem 2.5.11. *Let $A, \tilde{\Lambda}_1$ be as in Theorem 2.5.9, and let P_1 be the unitary polar factor of $\tilde{X}_1$. Then*

$$\eta_{\mathrm{H,F}}(P_1, \tilde{\Lambda}_1) \leq \frac{\sqrt{2}\|R_{\tilde{X}_1}\|_F}{\sigma_{\min}(\tilde{X}_1)}, \tag{2.5.37}$$

where

$$R_{\tilde{X}_1} = \tilde{X}_1\tilde{\Lambda}_1 - A\tilde{X}_1$$

is the residual of A with respect to $\tilde{X}_1$ and $\tilde{\Lambda}_1$.

The estimate (2.5.37) shows that if $\|R_{\tilde{X}_1}\|_F$ is small and if $\tilde{x}_1, \ldots, \tilde{x}_l$ are close to orthonormal, then there is a Hermitian matrix $A + H_{\mathrm{opt}}$ with small $\|H_{\mathrm{opt}}\|_F$ such that $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ and the column vectors of P_1 (the unitary polar factor of $\tilde{X}_1$) are l exact eigenvalues and associated eigenvectors of $A + H_{\mathrm{opt}}$.

Proof of Theorem 2.5.11. By (2.5.32) we only need to prove the inequality

$$\|R\|_F \leq \|R_{\tilde{X}_1}\|_F / \sigma_{\min}(\tilde{X}_1), \tag{2.5.38}$$

where

$$R = P_1\tilde{\Lambda}_1 - AP_1.$$

Let $\tilde{X}_1 = U \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} V^H$ be the singular value decomposition of $\tilde{X}_1$, where $U = (U_1, U_2) \in \mathcal{U}^{n\times n}$ with $U_1 \in \mathcal{U}^{n\times l}$, $V \in \mathcal{U}^{l\times l}$, and $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_l)$ with $\sigma_1 \geq \cdots \geq \sigma_l > 0$. Then the unitary polar factor P_1 of $\tilde{X}_1$ can be expressed by $P_1 = U_1V^H$. Thus, we have

$$\begin{aligned}
\|R_{\tilde{X}_1}\|_F &= \|A\tilde{X}_1 - \tilde{X}_1\tilde{\Lambda}_1\|_F = \left\| U^HAU \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} - \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} V^H\tilde{\Lambda}_1 V \right\|_F \\
&\geq \sigma_l \left\| U^HAU \begin{pmatrix} I_1 \\ 0 \end{pmatrix} - \begin{pmatrix} I_1 \\ 0 \end{pmatrix} V^H\tilde{\Lambda}_1 V \right\|_F \quad \text{(by Theorem 1.2.2)} \\
&= \sigma_{\min}(\tilde{X}_1)\|AP_1 - P_1\tilde{\Lambda}_1\|_F,
\end{aligned}$$

which shows the inequality (2.5.38). □

2.5.5 Residual Bounds

Let an approximate eigenspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$ of $A \in \mathcal{H}^{n\times n}$ be given, where $\tilde{U}_1 \in \mathcal{U}^{n\times l}$. Then by using Theorem 2.5.7 and appropriate forward perturbation results we can determine how the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ of $\tilde{U}_1^H A \tilde{U}_1$ relate to those of A, and determine the accuracy of the approximate eigenspace $\tilde{\mathcal{X}}_1$.

Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n \times n}$. By the proof of Theorem 2.5.7, the optimal backward perturbation H_{opt} of (2.5.27) satisfies

$$\tilde{U}^H (A + H_{\text{opt}}) \tilde{U} = \begin{pmatrix} \tilde{U}_1^H A \tilde{U}_1 & 0 \\ 0 & \tilde{U}_2^H A \tilde{U}_2 \end{pmatrix} \equiv \begin{pmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix}, \tag{2.5.39}$$

and

$$\tilde{U}^H H_{\text{opt}} \tilde{U} = \begin{pmatrix} 0 & R^H \tilde{U}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix}. \tag{2.5.40}$$

where R is the residual defined by (2.5.20). The relation (2.5.39) implies that the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ of $\tilde{U}_1^H A \tilde{U}_1$, as l approximate eigenvalues of A, are l eigenvalues of $A + H_{\text{opt}}$, and the subspace $\tilde{\mathcal{X}}_1$ is an eigenspace of $A + H_{\text{opt}}$.

Applying the Mirsky theorem [78] (see below NR 2.5–8) to the Hermitian matrices $A + H_{\text{opt}}$ and A, we obtain the following result which gives a residual bound for the approximate eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$.

Theorem 2.5.12. *Let $A \in \mathcal{H}^{n \times n}$, and let $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{U}_1)$ be an approximate eigenspace of A, where $\tilde{U}_1 \in \mathcal{U}^{n \times l}$. If the eigenvalues of A are $\lambda_1 \geq \cdots \geq \lambda_n$, and the eigenvalues of $\tilde{U}_1^H A \tilde{U}_1$ are $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_l$, then there are integers $j_1 < j_2 < \cdots < j_l$ such that*

$$\|\text{diag}(\tilde{\lambda}_1 - \lambda_{j_1}, \ldots, \tilde{\lambda}_l - \lambda_{j_l})\| \leq \left\| \begin{pmatrix} 0 & R^H \\ R & 0 \end{pmatrix} \right\|, \tag{2.5.41}$$

where R is the residual defined by

$$R = \tilde{U}_1 \tilde{A}_{11} - A \tilde{U}_1. \tag{2.5.42}$$

Applying Corollary 2.5.5 to the Hermitian matrices $A + H_{\text{opt}}$ and A shows the following result which gives a residual bound for the approximate eigenspace $\tilde{\mathcal{X}}_1$ of A.

Theorem 2.5.13. *Let $A, \tilde{\mathcal{X}}_1, \tilde{U}_1$ be as in Theorem 2.5.12. Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n \times n}$. Define the matrices $\tilde{A}_{11}, \tilde{A}_{22}$ by (2.5.39), define the residual R by (2.5.42), and define the linear operator $\mathbf{T}$ by*

$$\mathbf{T}Z = Z\tilde{A}_{11} - \tilde{A}_{22}Z, \qquad Z \in \mathcal{C}^{(n-l) \times l}.$$

If

$$\lambda(\tilde{A}_{11}) \bigcap \lambda(\tilde{A}_{22}) = \emptyset \quad \text{and} \quad 4\|\mathbf{T}^{-1}\| \|\mathbf{T}^{-1}(\tilde{U}_2^H R)\| \|R\|_2 < 1,$$

then there is a unique eigenspace $\mathcal{X}_1 = \mathcal{R}(U_1)$ *of* A *with* $U_1 \in \mathcal{U}^{n\times l}$ *such that*

$$\rho(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \le \|\tan\Theta(U_1, \tilde{U}_1)\| \le \frac{2\|\mathbf{T}^{-1}(\tilde{U}_2^H R)\|}{1+\sqrt{1-4\|\mathbf{T}^{-1}\|\|\mathbf{T}^{-1}(\tilde{U}_2^H R)\|\|R\|_2}} \equiv \tau(R). \tag{2.5.43}$$

It is worth noting that by using Theorem 2.5.13 and Theorem 2.5.17 of the next subsection (§2.5.6), we obtain the following result on residual bounds for eigenvalues which may be sharper than the estimate (2.5.41).

Theorem 2.5.14. *Let* $A, \tilde{\mathcal{X}}_1, \tilde{U}_1, \tilde{U}_2, \mathbf{T}$ *and* R *be as in Theorem 2.5.13. Let the eigenvalues of* A *be* $\lambda_1 \ge \cdots \ge \lambda_n$, *and the eigenvalues of* $\tilde{U}_1^H A \tilde{U}_1$ *be* $\tilde{\lambda}_1 \ge \cdots \ge \tilde{\lambda}_l$. *If the scalar* $\tau_2(R)$ *defined by (2.5.43) with* $\|\cdot\| = \|\cdot\|_2$ *satisfies*

$$\tau_2(R) < 1, \tag{2.5.44}$$

then there are integers $j_1 < j_2 < \cdots < j_l$ *such that*

$$\|\mathrm{diag}(\tilde{\lambda}_1 - \lambda_{j_1}, \ldots, \tilde{\lambda}_l - \lambda_{j_l})\| \le \frac{\tau_2(R)\|R\|}{\sqrt{1-(\tau_2(R))^2}}. \tag{2.5.45}$$

Proof. By Theorem 2.5.17 of the next subsection (§2.5.6), there are integers $j_1 < j_2 < \cdots < j_l$ such that

$$\|\mathrm{diag}(\tilde{\lambda}_1 - \lambda_{j_1}, \ldots, \tilde{\lambda}_l - \lambda_{j_l})\| \le \frac{\rho_2(\mathcal{X}_1, \tilde{\mathcal{X}}_1)\|R\|}{\sqrt{1-\rho_2^2(\mathcal{X}_1, \tilde{\mathcal{X}}_1)}}, \tag{2.5.46}$$

where $\rho_2(\cdot,\cdot)$ is the generalized chordal metric defined by (1.3.3). Substituting (2.5.43) into (2.5.46) shows (2.5.45). □

Define δ_2 by

$$\frac{1}{\delta_2} = \sup_{\substack{W \in \mathcal{C}^{(n-l)\times l} \\ W \ne 0}} \frac{\|\mathbf{T}^{-1}W\|_2}{\|W\|_2}.$$

Then from (2.5.43)

$$\rho_2(\mathcal{X}_1, \tilde{\mathcal{X}}_1) < 2\|\mathbf{T}^{-1}(\tilde{U}_2^H R)\|_2 \le \frac{2\|R\|_2}{\delta_2}.$$

Consequently, if

$$\omega \equiv \frac{2\|R\|_2}{\delta_2} < 1,$$

then from (2.5.46) we get a weaker estimate

$$\|\mathrm{diag}(\tilde{\lambda}_1 - \lambda_{J_1}, \ldots, \tilde{\lambda}_l - \lambda_{J_l})\| \le \frac{2}{\sqrt{1-\omega^2}} \frac{\|R\|_2\|R\|}{\delta_2}. \tag{2.5.47}$$

Example 2.5.15. Consider the real symmetric matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 10^5 & 1 \\ 0 & 10^5 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

The vector $u_1 = (1, 0, 0, 0)^T$ is clearly a unit eigenvector of A associated with the eigenvalue $\lambda_1 = 1$. Suppose that we have an approximate eigenvector

$$\tilde{x}_1 = (1,\ 10^{-6},\ 10^{-7},\ 10^{-8})^T,$$

and let $\tilde{u}_1 = \tilde{x}_1/\|\tilde{x}_1\|_2$. A calculation gives

$$\sin\theta(u_1, \tilde{u}_1) \approx 1.0049 \times 10^{-6}, \qquad \tan\theta(u_1, \tilde{u}_1) \approx 1.0049 \times 10^{-6}, \tag{2.5.48}$$

and

$$\begin{aligned} &\lambda(A) = \{1.0,\ -1.0 \times 10^5,\ 1.0 \times 10^5,\ 2.0\}, \\ &|\tilde{u}_1^T A\tilde{u}_1 - \lambda_1| \approx 2.0000 \times 10^{-8}, \end{aligned} \tag{2.5.49}$$

where $\theta(u_1, \tilde{u}_1)$ denotes the angle between the two 1-dimensional subspaces $\mathcal{R}(u_1)$ and $\mathcal{R}(\tilde{u}_1)$.

Choose $\tilde{U}_2$ so that $(\tilde{u}_1, \tilde{U}_2) \in \mathcal{O}^{4\times 4}$. Compute $\tilde{A}_{11}$, $\tilde{A}_{22}$, r and T by

$$\tilde{A}_{11} = \tilde{u}_1^T A\tilde{u}_1, \qquad \tilde{A}_{22} = \tilde{U}_2^T A\tilde{U}_2,$$

and

$$r = \tilde{A}_{11}\tilde{u}_1 - A\tilde{u}_1, \qquad T = \tilde{A}_{11}I - \tilde{A}_{22}.$$

calculation shows that $\tilde{A}_{11} \notin \lambda(\tilde{A}_{22})$, and

$$4\|T^{-1}\|_2\|T^{-1}(\tilde{U}_2^T r)\|_2\|r\|_2 \approx 4.0403 \times 10^{-7} < 1,$$

$$\tau_2(r) \equiv \frac{2\|T^{-1}(\tilde{U}_2^T r)\|_2}{1 + \sqrt{1 - 4\|T^{-1}\|_2\|T^{-1}(\tilde{U}_2^T r)\|_2\|r\|_2}} \approx 1.0100 \times 10^{-6} < 1.$$

Consequently, applying Theorem 2.5.13, there is a unit eigenvector u of A such that

$$\tan\theta(u, \tilde{u}_1) \le \tau_2(r) \approx 1.0100 \times 10^{-6}, \tag{2.5.50}$$

and applying Theorem 2.5.14 (see (2.5.45)), there is an eigenvalue $\lambda(= u^H Au)$ of A such that

$$|\tilde{u}_1^T A\tilde{u}_1 - \lambda| \le \frac{\tau_2(r)\|r\|_2}{\sqrt{1 - (\tau_2(r))^2}} \approx 1.0100 \times 10^{-7}. \tag{2.5.51}$$

Comparing (2.5.50) and (2.5.51) with (2.5.48) and (2.5.49) we see that the estimates obtained by applying Theorems 2.5.13 and 2.5.14 are fairly sharp.

It is worth pointing out that by Theorem 2.5.12 (see (2.5.41)) there is an eigenvalue λ of A such that

$$|\tilde{u}_1^T A\tilde{u}_1 - \lambda| \le \|r\|_2 \approx 1.0050 \times 10^{-1},$$

and by (2.5.47) there is an eigenvalue λ of A such that

$$|\tilde{u}_1^T A\tilde{u}_1 - \lambda| \leq \frac{2\|T^{-1}\|_2\|r\|_2^2}{\sqrt{1-(2\|T^{-1}\|_2\|r\|_2)^2}} \approx 2.0621 \times 10^{-2}.$$

The estimates obtained by (2.5.41) and (2.5.47) are obviously severe overestimates.

Example 2.5.16 (Saad [90, Example 3.4]). Consider the real symmetric matrix

$$A = \begin{pmatrix} 1.00 & 0.0055 & 0.10 & 0.10 & 0.00 \\ 0.0055 & 2.00 & -0.05 & 0.00 & -0.10 \\ 0.10 & -0.05 & 3.00 & 0.10 & 0.05 \\ 0.10 & 0.00 & 0.10 & 4.00 & 0.00 \\ 0.00 & -0.10 & 0.05 & 0.00 & 500 \end{pmatrix}.$$

Since the off-diagonal elements of A are small, the diagonal elements can be considered approximations to the eigenvalues of A. The question is: How good accuracy can be expected? We now apply Theorem 2.5.12, (2.5.47), and Theorem 2.5.14, to give estimates on the accuracy.

For any fixed integer $k \in [1,5]$ let $\tilde{u}_1 = e_k^{(5)}$, the kth column of the identity matrix I_5, and then compute $\tilde{A}_{11} = \tilde{u}_1^T A\tilde{u}_1 = k$. Let λ_k $(k = 1,\ldots,5)$ be the eigenvalues of A. By Theorem 2.5.12 (see (2.5.41)) we get the estimates

$$|\lambda_1 - 1.0| \leq 0.14153, \qquad |\lambda_2 - 2.0| \leq 0.11194, \qquad |\lambda_3 - 3.0| \leq 0.15811,$$

$$|\lambda_4 - 4.0| \leq 0.14142, \qquad |\lambda_5 - 5.0| \leq 0.11180.$$

By (2.5.47) we get the estimates

$$|\lambda_1 - 1.0| \leq 0.04203, \qquad |\lambda_2 - 2.0| \leq 0.02591, \qquad |\lambda_3 - 3.0| \leq 0.05251,$$

$$|\lambda_4 - 4.0| \leq 0.04197, \qquad |\lambda_5 - 5.0| \leq 0.02603.$$

By Theorem 2.5.14 (see (2.5.45)) we get the estimates

$$|\lambda_1 - 1.0| \leq 0.00838, \qquad |\lambda_2 - 2.0| \leq 0.00662, \qquad |\lambda_3 - 3.0| \leq 0.02050,$$

$$|\lambda_4 - 4.0| \leq 0.01591, \qquad |\lambda_5 - 5.0| \leq 0.00480.$$

Note that the actual errors are

$$|\lambda_1 - 1.0| \approx 0.00805, \qquad |\lambda_2 - 2.0| \approx 0.00556, \qquad |\lambda_3 - 3.0| \approx 0.00492,$$

$$|\lambda_4 - 4.0| \approx 0.01386, \qquad |\lambda_5 - 5.0| \approx 0.00467.$$

Obviously, the estimates obtained by applying Theorem 2.5.14 are fairly sharp.

By the way, the vectors $e_1^{(5)},\ldots,e_5^{(5)}$ can be considered approximations to the eigenvectors of A. Let u_k be an eigenvector of A associated with λ_k, $k = 1,\ldots,5$. Then by Theorem 2.5.13 (see (2.5.43)) we get the estimates

$$\tan\theta(u_1, e_1^{(5)}) \leq 0.05908, \qquad \tan\theta(u_2, e_2^{(5)}) \leq 0.05903, \qquad \tan\theta(u_3, e_3^{(5)}) \leq 0.12856,$$

$$\tan\theta(u_4, e_4^{(5)}) \leq 0.11179, \qquad \tan\theta(u_5, e_5^{(5)}) \leq 0.04292.$$

2.5.6 Eigenvalues of Rayleigh Quotient Matrices

In this subsection we shall reveal an approximation property of the eigenvalues of a Rayleigh quotient matrix that can be used to establish the estimate (2.5.45) of Theorem 2.5.14.

Let $\mathcal{X}_1 = \mathcal{R}(X_1)$ be an eigenspace of $A \in \mathcal{H}^{n\times n}$, and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ approximate $\mathcal{X}_1$, where $X_1, Y_1 \in \mathcal{U}^{n\times l}$. Let

$$A_1 = X_1^H A X_1, \qquad H_1 = Y_1^H A Y_1. \tag{2.5.52}$$

The matrices A_1 and H_1 are called the *Rayleigh quotient matrices* of A with respect to X_1 and Y_1, respectively.

It is easy to see that

$$AX_1 = X_1 A_1, \qquad \lambda(A_1) \subset \lambda(A).$$

However, in general, $AY_1 \neq Y_1 H_1$ and $\lambda(H_1) \not\subset \lambda(A)$. In such a case, we introduce the residual R of A with respect to Y_1 defined by

$$R = Y_1 H_1 - AY_1. \tag{2.5.53}$$

In this subsection we prove the following result (Theorem 2.5.17) which gives an upper bound for the distance between the sets $\lambda(H_1)$ and $\lambda(A_1)$ in terms of $\|R\|$ and $\rho_2(\mathcal{X}_1, \mathcal{Y}_1)$. Here $\rho_2(\cdot,\cdot)$ is the generalized chordal metric defined by (1.3.3), i.e.,

$$\rho_2(\mathcal{X}_1, \mathcal{Y}_1) = \|\sin\Theta\|_2,$$

in which the matrix Θ is defined by

$$\Theta = \Theta(X_1, Y_1) = \arccos(X_1^H Y_1 Y_1^H X_1)^{\frac{1}{2}} \geq 0.$$

Theorem 2.5.17. *Let A, X_1, Y_1, A_1, H_1, R and $\mathcal{X}_1, \mathcal{Y}_1$ be the above-mentioned matrices and subspaces. Let*

$$\lambda(A_1) = \{\lambda_j\}_{j=1}^l, \qquad \lambda_1 \geq \cdots \geq \lambda_l,$$

$$\lambda(H_1) = \{\mu_j\}_{j=1}^l, \qquad \mu_1 \geq \cdots \geq \mu_l,$$

and

$$\Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_l), \qquad M_1 = \operatorname{diag}(\mu_1, \ldots, \mu_l).$$

If $\rho_2(\mathcal{X}_1, \mathcal{Y}_1) < 1$, then

$$\|\Lambda_1 - M_1\| \leq \frac{\rho_2(\mathcal{X}_1, \mathcal{Y}_1)\|R\|}{\sqrt{1 - \rho_2^2(\mathcal{X}_1, \mathcal{Y}_1)}}. \tag{2.5.54}$$

Proof. 1) By Stewart [93, Appendix] (or see Stewart and Sun [97, Chapter I, Theorem 5.2]), there are unitary matrices Q, U_1 and V_1 such that

$$QX_1U_1 = \begin{pmatrix} I_l \\ 0 \end{pmatrix} \quad \text{and} \quad QY_1V_1 = \begin{pmatrix} \Gamma \\ \Sigma \end{pmatrix},$$

where

$$\Gamma = \begin{cases} \Gamma_1 = \mathrm{diag}(\gamma_1, \ldots, \gamma_l) & \text{if } 2l \le n \\ \\ \mathrm{diag}(\Gamma_1,\ I_{2l-n}),\ \ \Gamma_1 = \mathrm{diag}(\gamma_1, \ldots, \gamma_{n-l}) & \text{if } 2l > n, \end{cases} \tag{2.5.55}$$

$$\Sigma = \begin{cases} \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} \in \mathcal{R}^{(n-l)\times l},\ \ \Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_l) & \text{if } 2l \le n, \\ \\ (\Sigma_1, 0) \in \mathcal{R}^{(n-l)\times l},\ \ \Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n-l}) & \text{if } 2l > n, \end{cases} \tag{2.5.56}$$

and

$$0 \le \gamma_1 \le \gamma_2 \le \cdots \le 1, \quad 1 \ge \sigma_1 \ge \sigma_2 \ge \cdots \ge 0, \quad \gamma_j^2 + \sigma_j^2 = 1 \ \ \forall j. \tag{2.5.57}$$

Without loss of generality we may assume that the matrices A, X_1 and Y_1 have the following reduced forms:

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \quad X_1 = \begin{pmatrix} I_l \\ 0 \end{pmatrix}, \quad Y_1 = \begin{pmatrix} \Gamma \\ \Sigma \end{pmatrix}, \tag{2.5.58}$$

where Γ, Σ are the matrices of (2.5.55) and (2.5.56). Thus, we have

$$\rho_2(\mathcal{X}_1, \mathcal{Y}_1) = \|\sin\Theta\|_2 = \|\Sigma\|_2, \tag{2.5.59}$$

and

$$R = \begin{pmatrix} \Gamma H_1 - A_1\Gamma \\ \Sigma H_1 - A_2\Sigma \end{pmatrix}. \tag{2.5.60}$$

2) Let

$$\hat{\Gamma} = \begin{cases} \mathrm{diag}(\Gamma_1, I_{n-2l}) & \text{if } 2l \le n, \\ \\ \Gamma_1 & \text{if } 2l > n. \end{cases} \tag{2.5.61}$$

Combining it with (2.5.55) and (2.5.56) shows

$$\Sigma\Gamma = \hat{\Gamma}\Sigma. \tag{2.5.62}$$

Moreover, let

$$Y_2 = \begin{pmatrix} -\Sigma^T \\ \hat{\Gamma} \end{pmatrix}, \quad Y = (Y_1, Y_2). \tag{2.5.63}$$

Then from (2.5.62) we see that $Y \in \mathcal{U}^{n\times n}$, and the relations (2.5.60), (2.5.63) and $H_1 = Y_1^H A Y_1$ imply

$$Y^H R = \begin{pmatrix} 0 \\ B \end{pmatrix}. \tag{2.5.64}$$

Thus,

$$\|R\| = \|B\|. \tag{2.5.65}$$

From (2.5.60), (2.5.63) and (2.5.64)

$$\begin{aligned}\Gamma H_1 - A_1\Gamma &= (I_l, 0)R = (I_l, 0)Y \begin{pmatrix} 0 \\ B \end{pmatrix} \\ &= (I_l, 0)Y_2 B = -\Sigma^T B.\end{aligned}$$

Combining it with (2.5.65) gives

$$\|\Gamma H_1 - A_1\Gamma\| \leq \|\Sigma\|_2\|B\| = \|\Sigma\|_2\|R\|. \tag{2.5.66}$$

3) Applying a result due to Bhatia, Davis and Kittaneh [9] (see below NR 2.5–5), and using the expressions (2.5.55)–(2.5.57), for the Hermitian matrices H_1 and A_1 we have

$$\|\Gamma H_1 - A_1\Gamma\| \geq \gamma_1\|H_1 - A_1\| = \sqrt{1 - \|\Sigma\|_2^2}\|A_1 - H_1\|. \tag{2.5.67}$$

Moreover, by the Mirsky theorem [78] (see below NR 2.5–8), we have

$$\|A_1 - H_1\| \geq \|\Lambda_1 - M_1\|.$$

Substituting it into (2.5.67) gives

$$\|\Gamma H_1 - A_1\Gamma\| \geq \sqrt{1 - \|\Sigma\|_2^2}\|\Lambda_1 - M_1\|. \tag{2.5.68}$$

Combining (2.5.68) with (2.5.66), (2.5.59) and the assumption $\rho_2(\mathcal{X}_1, \mathcal{Y}_1) < 1$, shows (2.5.54). □

Notes and References

NR 2.5–1. §2.5.1 is based on Sun [102]. Theorem 2.5.6 is proved by Stewart [91].

NR 2.5–2. §2.5.4 and §2.5.5 are based on Sun [115] and [116].

NR 2.5–3. For the spectral norm, the residual bound (2.5.41) of Theorem 2.5.12 is due to Kahan [61] (or see Parlett [83, p.219–220]). If $l = 1$ and if we write $\tilde{U}_1 = \tilde{u}_1$ and $R = r$, then (2.5.41) becomes

$$|\tilde{\lambda}_1 - \lambda_{j_1}| \leq \|r\|_2,$$

which is a well known inclusion theorem (see, e.g., Wielandt [129]).

NR 2.5–4. Theorem 2.5.17 is proved by Sun [111].

NR 2.5–5. Bhatia, Davis, and Kittaneh [6] prove the following result: *Let $A, H, Z \in \mathcal{C}^{n\times n}$, in which A and H are Hermitian. Then for any unitarily invariant norm $\|\cdot\|$*

$$\|AZ - ZH\| \geq \sigma_{\min}(Z)\|A - H\|.$$

NR 2.5–6. Let A, X_1, Y_1, A_1, H_1, R be the matrices as in Theorem 2.5.17. It is easy to see that for any $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ the matrix A has the spectral resolution

$$X^H AX = \mathrm{diag}(A_1, A_2).$$

Davis and Kahan [26] show that if

$$\lambda(H_1) \subset [\alpha, \beta] \tag{2.5.69}$$

and for some $\delta > 0$,

$$\lambda(A_2) \subset \mathcal{R}\backslash[\alpha - \delta, \beta + \delta], \tag{2.5.70}$$

then

$$\|\sin\Theta(X_1, Y_1)\| \leq \frac{\|R\|}{\delta}.$$

This is the Davis-Kahan $\sin\theta$ theorem.

Combining the Davis-Kahan $\sin\theta$ theorem with Theorem 2.5.17 we see that under the assumptions (2.5.69) and (2.5.70) we have the following corollary: If

$$\epsilon \equiv \frac{\|R\|_2}{\delta} < 1,$$

then

$$\|\Lambda_1 - M_1\| \leq \frac{\|R\|_2\|R\|}{\delta\sqrt{1-\epsilon^2}}, \tag{2.5.71}$$

where Λ_1 and M_1 are the diagonal matrices of Theorem 2.5.17.

An estimate similar to (2.5.71) is first given by Stewart [96]. Recently, Mathias [76] obtains stronger and more general $O(\|R\|^2)$ bounds for the Hermitian eigenvalue problem, and the results are extended to singular values, eigenvalues of non-Hermitian matrices, and generalized eigenvalues.

NR 2.5–7. More results on Rayleigh quotients and eigenvalues of Rayleigh quotient matrices are given by Kahan [61], Paige [82], Parlett [83], Chatelin [16], Li [68], Liu and Xu [73], Sun [111], and Cao, Xie and Li [14].

NR 2.5–8. Mirsky Theorem [78]. *Let A and $\tilde{A}$ be Hermitian matrices of the same dimension with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \qquad \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n.$$

Then for any unitarily invariant norm $\|\cdot\|$,

$$\|\mathrm{diag}(\tilde{\lambda}_i - \lambda_i)\| \leq \|\tilde{A} - A\|.$$

(See, e.g., Stewart and Sun [97, Chapter IV, Corollary 4.12.)

NR 2.5–9. In recent years, perturbation theory for the unitary eigenproblem has been developed; see, e.g., Bhatia and Davis [5], Elsner and He [35], and Bohnhorst, Bunse-Gerstner and Fassbender [9]. Backward errors and residual bounds are discussed by Sun [118] and [120].

Chapter 3

The Singular Value Decomposition

In this chapter we will be concerned with perturbation analysis of the *singular value decomposition* of $A \in \mathcal{C}^{m\times n}$: $A = U\Sigma V^H$, where U and V are unitary matrices and Σ is a diagonal matrix with nonnegative diagonal elements. Perturbation expansions and condition numbers of singular values and singular subspaces, perturbation bounds for singular subspaces, and backward errors and residual bounds, will be studied in §3.1 – §3.4, separately.

3.1 Perturbation Expansions

3.1.1 Simple Non-Zero Singular Values

Let $A \in \mathcal{C}^{m\times n}$. If

$$Av = \sigma u \quad \text{and} \quad A^H u = \sigma v$$

for $\sigma \geq 0$ and unit vectors $v \in \mathcal{C}^n$ and $u \in \mathcal{C}^m$, then σ is called a *singular value* of A, and v and u are called *unit right* and *left singular vectors* of A associated with σ. Without loss of generality we may assume that $m \geq n$.

Let $A \in \mathcal{C}^{m\times n}$, and let

$$A = U\Sigma V^H$$

be an singular value decomposition of $A \in \mathcal{C}^{m\times n}$, where

$$V = (v_1, \ldots, v_n) \in \mathcal{U}^{n\times n}, \qquad U = (u_1, \ldots, u_m) \in \mathcal{U}^{m\times m},$$

and

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots) \in \mathcal{R}^{m\times n} \quad \text{with} \quad \sigma_1, \ldots, \sigma_n \geq 0.$$

Then $\sigma_1, \ldots, \sigma_n$ are the singular values of A, and v_j and u_j are unit right and left singular vectors of A associated with σ_j, $j = 1, \ldots, n$.

Let $p = (p_1, \ldots, p_N)^T$, and let $A(p) \in \mathcal{C}^{m\times n}$ be a matrix-valued function in some neighborhood $\mathcal{B}(p^*)$ of the point p^*. For simplicity, we assume that $p^* = 0$, and $p_1, \ldots, p_N$ are real parameters.

Let $\sigma > 0$ be a *simple* singular value of $A(0)$, and v and u be the associated unit right and left singular vectors. Then, as a consequence, there are matrices

$$U = (u, U_2) \in \mathcal{U}^{m\times m}, \quad V = (v, V_2) \in \mathcal{U}^{n\times n}, \tag{3.1.1}$$

and

$$\Sigma = \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_2 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} \sigma_2 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{pmatrix} \in \mathcal{R}^{(m-1)\times(n-1)} \tag{3.1.2}$$

with $\sigma, \sigma_2, \ldots, \sigma_n \geq 0$ and $\sigma_j \neq \sigma > 0$ for $j = 2, \ldots, n$, such that $A(0)$ has the singular value decomposition

$$A(0) = U\Sigma V^H.$$

First applying the implicit function theorem we prove the following result.

Theorem 3.1.1. *Let $p \in \mathcal{R}^N$ and $A(p) \in \mathcal{C}^{m\times n}$. Suppose that* $\mathrm{Re}[A(p)]$ *and* $\mathrm{Im}[A(p)]$ *are real analytic matrix-valued functions of p in some neighborhood $\mathcal{B}(0)$ of the origin. If $A(0)$ has a singular value decomposition $A(0) = U\Sigma V^H$, where U, V and Σ are the matrices of (3.1.1) and (3.1.2). Then*

1) there exists a simple singular value $\sigma(p)$ of $A(p)$ which is a real analytic function of p in some neighborhood $\mathcal{B}_0$ of the origin, and $\sigma(0) = \sigma$; the unit right singular vector $v(p)$ and the unit left singular vector $u(p)$ of $A(p)$ associated with $\sigma(p)$ may be so defined that $\mathsf{Re}[v(p)]$, $\mathsf{Im}[v(p)]$, $\mathsf{Re}[u(p)]$ *and* $\mathsf{Im}[u(p)]$ *are real analytic functions of p in $\mathcal{B}_0$, $v(0) = v$ and $u(0) = u$;*

2) the function $\sigma(p)$ has a power series expansion at $p = 0$ of the form

$$\sigma(p) = \sigma + \sum_{j=1}^{N} \left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} p_j + \frac{1}{2} \sum_{j,k=1}^{N} \left(\frac{\partial^2 \sigma(p)}{\partial p_j \partial p_k}\right)_{p=0} p_j p_k + \cdots, \qquad p \in \mathcal{B}_0,$$

where

$$\left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} = \mathsf{Re}\left[u^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v\right], \tag{3.1.3}$$

and

$$\left(\frac{\partial^2\sigma(p)}{\partial p_j\partial p_k}\right)_{p=0} = \mathsf{Re}\left[u^H\left(\frac{\partial^2 A(p)}{\partial p_j\partial p_k}\right)_{p=0} v + \left(\begin{array}{c} u \\ v \end{array}\right)^H D_k^H S D_j \left(\begin{array}{c} u \\ v \end{array}\right)\right]$$
$$+\frac{1}{\sigma}\mathsf{Im}\left[u^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v\right]\mathsf{Im}\left[u^H\left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} v\right], \tag{3.1.4}$$

where

$$D_j = \mathrm{diag}\left(\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}^H, \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}\right), \tag{3.1.5}$$

and

$$S = \left(\begin{array}{cc} V_2\Phi_1 V_2^H & V_2\Omega^T U_2^H \\ U_2\Omega V_2^H & U_2\Phi_2 U_2^H \end{array}\right), \tag{3.1.6}$$

in which

$$\Phi_1 = \sigma(\sigma^2 I - \Sigma_2^T\Sigma_2)^{-1}, \qquad \Phi_2 = \sigma(\sigma^2 I - \Sigma_2\Sigma_2^T)^{-1},$$
$$\Omega = \Sigma_2(\sigma^2 I - \Sigma_2^T\Sigma_2)^{-1}, \tag{3.1.7}$$

and u, v, U_2, V_2, σ and Σ_2 are defined by (3.1.1) and (3.1.2).

Proof. 1) Define $\tilde{A}(p)$ by

$$\tilde{A}(p) = V^H A(p)^H A(p) V = \left(\begin{array}{cc} \tilde{a}_{11}(p) & \tilde{a}_{21}(p)^H \\ \tilde{a}_{21}(p) & \tilde{A}_{22}(p) \end{array}\right), \quad \tilde{a}_{11}(p) \in \mathcal{R},$$

and introduce a vector-valued function

$$f(z,p) = \tilde{a}_{21}(p) + [\tilde{A}_{22}(p) - \tilde{a}_{11}(p)I]z - z\tilde{a}_{21}(p)^H z,$$

where

$$f = (f_1, \ldots, f_{n-1})^T, \qquad z = (\zeta_1, \ldots, \zeta_{n-1})^T \in \mathcal{C}^{n-1}, \qquad p \in \mathcal{R}^N.$$

Let

$$f_j = \phi_j + i\psi_j, \quad \zeta_j = \xi_j + i\eta_j, \quad i = \sqrt{-1}, \quad j = 1, \ldots, n-1,$$

and

$$x = (\xi_1, \ldots, \xi_{n-1})^T,\ y = (\eta_1, \ldots, \eta_{n-1})^T \in \mathcal{R}^{n-1}.$$

Obviously, $\phi_j(x,y,p)$ and $\psi_j(x,y,p)$ are real analytic functions of the real variables $x, y \in \mathcal{R}^{n-1}$ and $p \in \mathcal{B}(0)$, and the functions satisfy

$$\phi(0,0,0) = 0, \qquad \psi(0,0,0) = 0, \qquad j = 1, \ldots, n-1.$$

Since $f_1, \ldots, f_{n-1}$ are complex analytic functions of the complex variables $\zeta_1, \ldots, \zeta_{n-1}$ for any $p \in \mathcal{B}(0)$, by Theorem 1.6.3 we have

$$\left(\frac{\partial(\phi_1, \ldots, \phi_{n-1}, \psi_1, \ldots, \psi_{n-1})}{\partial(\xi_1, \ldots, \xi_{n-1}, \eta_1, \ldots, \eta_{n-1})} \right)_{x=y=0,\ p=0} = \left| \frac{\partial(f_1, \ldots, f_{n-1})}{\partial(\zeta_1, \ldots, \zeta_{n-1})} \right|^2_{z=0,\ p=0}$$

$$= \left| \det(\tilde{A}_{22}(0) - \tilde{a}_{11}(0)I) \right|^2 = \left| \det(\Sigma_2^T \Sigma_2 - \sigma^2 I) \right|^2$$

$$= \prod_{l=2}^{n} (\sigma_l^2 - \sigma^2)^2 > 0.$$

Therefore, by the implicit function theorem (Theorem 1.6.2) the system of equations

$$\phi_j(x, y, p) = 0, \qquad \psi_j(x, y, p) = 0, \qquad j = 1, \ldots, n-1$$

has a unique real analytic solution $x = x(p), y = y(p)$ in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin, and $x(0) = y(0) = 0$. In other words, the equation $f(z, p) = 0$ has a unique analytic solution $z = z(p)$ in $\mathcal{B}_0$, and $z(0) = 0$. Moreover, we may choose $\mathcal{B}_0$ so small that $1 + z(p)^H z(p) > 0$ for any $p \in \mathcal{B}_0$. As a result, we have

$$\begin{pmatrix} 1 & -z(p)^H \\ z(p) & I \end{pmatrix}^H \tilde{A}(p) \begin{pmatrix} 1 & -z(p)^H \\ z(p) & I \end{pmatrix}$$

$$= \begin{pmatrix} \tilde{a}_{11}(p) + z(p)^H \tilde{a}_{21}(p) + \tilde{a}_{21}(p)^H z(p) + z(p)^H \tilde{A}_{22}(p) z(p) & 0 \\ 0 & * \end{pmatrix}, \quad p \in \mathcal{B}_0;$$

and from this relation we get

$$V^H A(p)^H A(p) V \begin{pmatrix} 1 \\ z(p) \end{pmatrix} \left(1 + z(p)^H z(p)\right)^{-\frac{1}{2}}$$

$$= \left(\tilde{a}_{11}(p) + z(p)^H \tilde{a}_{21}(p) + \tilde{a}_{21}(p)^H z(p) + z(p)^H \tilde{A}_{22}(p) z(p)\right) \left(1 + z(p)^H z(p)\right)^{-1}$$

$$\times \begin{pmatrix} 1 \\ z(p) \end{pmatrix} \left(1 + z(p)^H z(p)\right)^{-\frac{1}{2}}. \tag{3.1.8}$$

Since

$$\tilde{a}_{11}(p) + z(p)^H \tilde{a}_{21}(p) + \tilde{a}_{21}(p)^H z(p) + z(p)^H \tilde{A}_{22}(p) z(p) > 0$$

for $p \in \mathcal{B}_0$ provided that $\mathcal{B}_0$ is sufficiently small, we may define a positive valued analytic function $\sigma_1(p)$ by

$$\begin{aligned} \sigma(p) \quad &= \Big[\Big(\tilde{a}_{11}(p) + z(p)^H \tilde{a}_{21}(p) + \tilde{a}_{21}(p)^H z(p) \\ &\quad + z(p)^H \tilde{A}_{22}(p) z(p)\Big) \left(1 + z(p)^H z(p)\right)^{-1} \Big]^{1/2}, \quad p \in \mathcal{B}_0. \end{aligned}$$

Further, for $p \in \mathcal{B}_0$ we define two vector-valued analytic functions $v(p)$ and $u(p)$ by

$$v(p) = V \begin{pmatrix} 1 \\ z(p) \end{pmatrix} \left(1 + z(p)^H z(p)\right)^{-1/2}, \qquad u(p) = A(p)v(p)/\sigma(p). \tag{3.1.9}$$

Then the relation (3.1.8) implies that for $p \in \mathcal{B}_0$ the functions $\sigma(p), v(p)$ and $u(p)$ satisfy

$$\begin{gathered} A(p)v(p) = \sigma(p)u(p), \qquad A(p)^H u(p) = \sigma(p)v(p), \\ \|u(p)\|_2 = \|v(p)\|_2 = 1, \end{gathered} \tag{3.1.10}$$

which means that $\sigma(p)$ is a singular value of $A(p)$, and $u(p)$ and $v(p)$ are associated unit right and unit left singular vectors. Moreover, we have

$$\sigma(0) = \sigma, \qquad v(0) = v, \qquad u(0) = u, \tag{3.1.11}$$

and the singular value $\sigma(p)$ is simple provided that the neighborhood $\mathcal{B}_0$ is sufficiently small.

2-1) By (3.1.10) and (3.1.11),

$$\sigma(p) = u(p)^H A(p)v(p) = v(p)^H A(p)^H u(p).$$

Thus, we have

$$\frac{\partial \sigma(p)}{\partial p_j} = \sigma(p) \left(\frac{\partial u(p)}{\partial p_j}\right)^H u(p) + u(p)^H \frac{\partial A(p)}{\partial p_j} v(p) + \sigma(p)v(p)^H \frac{\partial v(p)}{\partial p_j}, \tag{3.1.12}$$

and

$$\frac{\partial \sigma(p)}{\partial p_j} = \sigma(p) \left(\frac{\partial v(p)}{\partial p_j}\right)^H v(p) + v(p)^H \left(\frac{\partial A(p)}{\partial p_j}\right)^H u(p) + \sigma(p)u(p)^H \frac{\partial u(p)}{\partial p_j}. \tag{3.1.13}$$

From (3.1.12), (3.1.13) and

$$u(p)^H u(p) = v(p)^H v(p) = 1,$$

we get

$$\frac{\partial \sigma(p)}{\partial p_j} = \frac{1}{2} \left[u(p)^H \frac{\partial A(p)}{\partial p_j} v(p) + v(p)^H \left(\frac{\partial A(p)}{\partial p_j}\right)^H u(p) \right]. \tag{3.1.14}$$

Substituting $p = 0$ into (3.1.14) gives (3.1.3).

2-2) From

$$A(p)^H A(p)v(p) = \sigma(p)^2 v(p)$$

it follows that

$$\left(\sigma^2 I - A(0)^H A(0)\right)\left(\frac{\partial v(p)}{\partial p_j}\right)_{p=0}$$

$$= \left[\left(\frac{\partial A(p)}{\partial p_j}\right)^H_{p=0} A(0) + A(0)^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} - 2\sigma \left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} I\right] v.$$

Combining it with (3.1.2), (3.1.9) and $z(0) = 0$ gives

$$\begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 I - \Sigma_2^T \Sigma_2 \end{pmatrix} \begin{pmatrix} 0 \\ \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} \end{pmatrix} = \sigma V^H \left(\frac{\partial A(p)}{\partial p_j}\right)^H_{p=0} u$$

$$+ \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_2^T \end{pmatrix} U^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v - 2\sigma \left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} e_1^{(n)}$$

and

$$\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} = \left(\sigma^2 I - \Sigma_2^T \Sigma_2\right)^{-1} \left[\sigma V_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)^H_{p=0} u + \Sigma_2^T U_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v\right]. \tag{3.1.15}$$

Substituting (3.1.15) into

$$\left(\frac{\partial v(p)}{\partial p_j}\right)_{p=0} = V_2 \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0},$$

and using the matrices Φ_1 and Ω defined by (3.1.7), we get

$$\left(\frac{\partial v(p)}{\partial p_j}\right)_{p=0} = V_2 \left(\Phi_1 V_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)^H_{p=0},\ \Omega^T U_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}\right) \begin{pmatrix} u \\ v \end{pmatrix}. \tag{3.1.16}$$

2-3) From $A(p)v(p) = \sigma(p)u(p)$ we obtain

$$\left(\frac{\partial u(p)}{\partial p_j}\right)_{p=0} = \frac{1}{\sigma}\left[\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v + A(0)\left(\frac{\partial v(p)}{\partial p_j}\right)_{p=0} - \left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} u\right].$$

Combining it with (3.1.3), (3.1.16), and using the relations

$$\frac{1}{\sigma} A(0) V_2 \Phi_1 V_2^H = U_2 \Omega V_2^H, \quad \frac{1}{\sigma} A(0) V_2 \Omega^T U_2^H = U_2 \Phi_2 U_2^H - \frac{1}{\sigma_1} U_2 U_2^H,$$

we get

$$\left(\frac{\partial u(p)}{\partial p_j}\right)_{p=0} = U_2\left(\Omega V_2^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}^H,\ \Phi_2 U_2^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}\right)\begin{pmatrix} u \\ v \end{pmatrix}$$

$$+\frac{i}{\sigma}\mathsf{Im}\left[u^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v\right]u, \qquad i=\sqrt{-1}. \tag{3.1.17}$$

2-4) From (3.1.14) it follows that

$$\left(\frac{\partial^2\sigma(p)}{\partial p_j\partial p_k}\right)_{p=0} = \mathsf{Re}\left[u^H\left(\frac{\partial^2 A(p)}{\partial p_j\partial p_k}\right)_{p=0} v\right] + \mathsf{Re}\left[\left(\frac{\partial u(p)}{\partial p_k}\right)_{p=0}^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v\right.$$

$$\left. + \left(\frac{\partial v(p)}{\partial p_k}\right)_{p=0}^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}^H u\right].$$

Combining it with (3.1.16) and (3.1.17) shows

$$\left(\frac{\partial^2\sigma(p)}{\partial p_j\partial p_k}\right)_{p=0} = \mathsf{Re}\left[u^H\left(\frac{\partial^2 A(p)}{\partial p_j\partial p_k}\right)_{p=0} v\right]$$

$$+\mathsf{Re}\left[\begin{pmatrix} u \\ v \end{pmatrix}^H\begin{pmatrix} \left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0}^H & 0 \\ 0 & \left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0}\end{pmatrix}^H\right. \tag{3.1.18}$$

$$\left.\times\begin{pmatrix} V_2\Phi_1V_2^H & V_2\Omega^TU_2^H \\ U_2\Omega V_2^H & U_2\Phi_2U_2^H\end{pmatrix}\begin{pmatrix} \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}^H & 0 \\ 0 & \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}\end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}\right]$$

$$+\frac{1}{\sigma}\mathsf{Im}\left[u^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v\right]\mathsf{Im}\left[u^H\left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} v\right].$$

Using the matrices D_j and S defined by (3.1.5) and (3.1.6), the formula (3.1.18) can be written as (3.1.4). □

Example 3.1.2 [74]. Consider the matrix

$$A(p) = \begin{pmatrix} 1 & -1 \\ \frac{1}{p_1+ip_2+2} & 1 \end{pmatrix}, \quad p=(p_1,p_2)^T\in\mathcal{R}^2, \quad i=\sqrt{-1}.$$

Obviously, $A(p)$ is an analytic matrix-valued function of p in a neighborhood of the origin. Moreover, the matrix $A(0)=\begin{pmatrix} 1 & -1 \\ \frac{1}{2} & 1 \end{pmatrix}$ has a singular value decomposition $A(0)=U\Sigma V^H$

with

$$U = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix} = (u_1, u_2), \quad V = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} = (v_1, v_2), \quad \Sigma = \begin{pmatrix} \frac{3}{2} & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus, we have

$$\sigma_1 = \frac{3}{2}, \qquad \sigma_2 = 1, \qquad \Phi_1 = \Phi_2 = \frac{6}{5}, \qquad \Omega = \frac{4}{5},$$

$$\left(\frac{\partial A(p)}{\partial p_1}\right)_{p=0} = \begin{pmatrix} 0 & 0 \\ -\frac{1}{4} & 0 \end{pmatrix}, \qquad \left(\frac{\partial A(p)}{\partial p_2}\right)_{p=0} = \begin{pmatrix} 0 & 0 \\ -\frac{i}{4} & 0 \end{pmatrix},$$

$$\left(\frac{\partial^2 A(p)}{\partial p_1^2}\right)_{p=0} = \begin{pmatrix} 0 & 0 \\ \frac{1}{4} & 0 \end{pmatrix}, \qquad \left(\frac{\partial^2 A(p)}{\partial p_2^2}\right)_{p=0} = \begin{pmatrix} 0 & 0 \\ -\frac{1}{4} & 0 \end{pmatrix},$$

and

$$\left(\frac{\partial^2 A(p)}{\partial p_1 \partial p_2}\right)_{p=0} = \begin{pmatrix} 0 & 0 \\ \frac{i}{4} & 0 \end{pmatrix}.$$

Using the formulas (3.1.3) and (3.1.4), we get

$$\left(\frac{\partial \sigma_1(p)}{\partial p_1}\right)_{p=0} = 0.05, \qquad \left(\frac{\partial \sigma_1(p)}{\partial p_2}\right)_{p=0} = 0,$$

$$\left(\frac{\partial \sigma_2(p)}{\partial p_1}\right)_{p=0} = -0.2, \qquad \left(\frac{\partial \sigma_2(p)}{\partial p_2}\right)_{p=0} = 0,$$

and

$$\left(\frac{\partial^2 \sigma_1(p)}{\partial p_1^2}\right)_{p=0} = -0.042, \quad \left(\frac{\partial^2 \sigma_1(p)}{\partial p_1 \partial p_2}\right)_{p=0} = 0, \quad \left(\frac{\partial^2 \sigma_1(p)}{\partial p_2^2}\right)_{p=0} = 0.098\dot{3},$$

$$\left(\frac{\partial^2 \sigma_2(p)}{\partial p_1^2}\right)_{p=0} = 0.208, \quad \left(\frac{\partial^2 \sigma_2(p)}{\partial p_1 \partial p_2}\right)_{p=0} = 0, \quad \left(\frac{\partial^2 \sigma_2(p)}{\partial p_2^2}\right)_{p=0} = -0.12.$$

Consequently, $\sigma_1(p)$ and $\sigma_2(p)$ have the expansions

$$\sigma_1(p) = 1.5 + 0.05p_1 - 0.021p_1^2 + 0.0491\dot{6}p_2^2 + O(\|p\|_2^3),$$

and

$$\sigma_2(p) = 1.0 - 0.2p_1 + 0.104p_1^2 - 0.06p_2^2 + O(\|p\|_2^3),$$

where $p \in \mathcal{B}_0$, a neighborhood of the origin.

From Theorem 3.1.1 we obtain the following result.

Theorem 3.1.3. *Let $A(p) \in \mathcal{R}^{m\times n}$ with $p \in \mathcal{R}^N$. Suppose that $A(p)$ is a real analytic matrix-valued function of p in some neighborhood $\mathcal{B}(0)$ of the origin. If $A(0)$ has a singular value decomposition $A(0) = U\Sigma V^T$, where U, V and Σ are expressed by (3.1.1) and (3.1.2), among which U and V are real orthogonal matrices, and $\sigma, \sigma_2, \ldots, \sigma_n \geq 0$ with $\sigma_j \neq \sigma > 0$ for $j = 2, \ldots, n$. Then*

1) there exists a simple singular value $\sigma(p)$ of $A(p)$ which is a real analytic function of p in some neighborhood $\mathcal{B}_0$ of the origin, and $\sigma(0) = \sigma$; the unit right singular vector $v(p)$ and the unit left singular vector $u(p)$ of $A(p)$ associated with $\sigma(p)$ may be so defined that $v(p)$ and $u(p)$ are real analytic functions of p in $\mathcal{B}_0$, $v(0) = v$ and $u(0) = u$;

2) the function $\sigma(p)$ has a power series expansion at $p = 0$ of the form

$$\sigma(p) = \sigma + \sum_{j=1}^{N} \left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} p_j + \frac{1}{2} \sum_{j,k=1}^{N} \left(\frac{\partial^2 \sigma(p)}{\partial p_j \partial p_k}\right)_{p=0} p_j p_k + \cdots, \quad p \in \mathcal{B}_0,$$

where

$$\left(\frac{\partial \sigma(p)}{\partial p_j}\right)_{p=0} = u^T \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} v, \tag{3.1.19}$$

and

$$\left(\frac{\partial^2 \sigma(p)}{\partial p_j \partial p_k}\right)_{p=0} = u^T \left(\frac{\partial^2 A(p)}{\partial p_j \partial p_k}\right)_{p=0} v + \begin{pmatrix} u \\ v \end{pmatrix}^T D_k^T S D_j \begin{pmatrix} u \\ v \end{pmatrix}, \tag{3.1.20}$$

in which

$$S = \begin{pmatrix} V_2 \Phi_1 V_2^T & V_2 \Omega^T U_2^T \\ U_2 \Omega V_2^T & U_2 \Phi_2 U_2^T \end{pmatrix},$$

$$D_j = \operatorname{diag}\left(\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0}^T, \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} \right),$$

and Φ_1, Φ_2, Ω are defined by (3.1.7).

Note that for the singular vectors $v(p)$ and $u(p)$ we have the formula

$$\begin{pmatrix} \left(\frac{\partial v(p)}{\partial p_j}\right)_{p=0} \\ \left(\frac{\partial u(p)}{\partial p_j}\right)_{p=0} \end{pmatrix} = S D_j \begin{pmatrix} u \\ v \end{pmatrix}.$$

The following two results, as corollaries of Theorems 3.1.1 and 3.1.3, give second order perturbation expansions of any non-zero simple singular value. The proofs are left as exercises.

Corollary 3.1.4. *Let $A = U \Sigma V^H$ be a singular value decomposition of $A \in \mathcal{C}^{m \times n}$, where the unitary matrices U, V and the diagonal matrix Σ are expressed by (3.1.1) and (3.1.2), in which $\sigma, \sigma_2, \ldots, \sigma_n \geq 0$ and $\sigma_j \neq \sigma > 0$ for $j = 2, \ldots, n$. Moreover, let $E \in \mathcal{C}^{m \times n}$. Then as $E \to 0$ the matrix $A + E$ has a simple singular* *value $\tilde{\sigma}$ satisfying*

$$\begin{aligned}\tilde{\sigma} \quad &= \sigma + \mathsf{Re}(u^H E v) \\ &+\frac{1}{2}\mathsf{Re}\begin{pmatrix} u \\ v \end{pmatrix}^H \begin{pmatrix} EV_2\Phi_1 V_2^H E^H & EV_2\Omega^T U_2^H E \\ E^H U_2 \Omega V_2^H E^H & E^H U_2 \Phi_2 U_2^H E \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \\ &+\frac{1}{2\sigma}\left[\mathsf{Im}(u^H E v)\right]^2 + O(\|E\|_F^3),\end{aligned} \tag{3.1.21}$$

and the associated right and left singular vectors $\tilde{v}$ and $\tilde{u}$ satisfy

$$\begin{aligned}\tilde{v} &= v + V_2(\Phi_1 V_2^H E^H,\ \Omega^T U_2^H E)\begin{pmatrix} u \\ v \end{pmatrix} + O(\|E\|_F^2), \\ \tilde{u} &= u + U_2(\Omega V_2^H E^H,\ \Phi_2 U_2^H E)\begin{pmatrix} u \\ v \end{pmatrix} + \frac{i}{\sigma}\mathrm{Im}(u^H E v)u + O(\|E\|_F^2),\end{aligned} \tag{3.1.22}$$

where Φ_1, Φ_2, Ω are the matrices defined by (3.1.7).

Corollary 3.1.5. *Let $A = U\Sigma V^T$ be a singular value decomposition of $A \in \mathcal{R}^{m\times n}$, where the real orthogonal matrices U, V and the diagonal matrix Σ are expressed by (3.1.1) and (3.1.2), in which $\sigma, \sigma_2, \ldots, \sigma_n \geq 0$ and $\sigma_j \neq \sigma > 0$ for $j = 2, \ldots, n$. Moreover, let $E \in \mathcal{R}^{m\times n}$. Then as $E \to 0$ the matrix $A + E$ has a simple singular value $\tilde{\sigma}$ satisfying*

$$\begin{aligned}\tilde{\sigma} \quad &= \sigma + u^T E v \\ &+\frac{1}{2}\begin{pmatrix} u \\ v \end{pmatrix}^T \begin{pmatrix} EV_2\Phi_1 V_2^T E^T & EV_2\Omega^T U_2^T E \\ E^T U_2 \Omega V_2^T E^T & E^T U_2 \Phi_2 U_2^T E \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \\ &+O(\|E\|_F^3),\end{aligned}$$

and the associated right and left singular vectors $\tilde{v}$ and $\tilde{u}$ satisfy

$$\tilde{v} = v + V_2(\Phi_1 V_2^T E^T,\ \Omega^T U_2^T E)\begin{pmatrix} u \\ v \end{pmatrix} + O(\|E\|_F^2),$$

$$\tilde{u} = u + U_2(\Omega V_2^T E^T,\ \Phi_2 U_2^T E)\begin{pmatrix} u \\ v \end{pmatrix} + O(\|E\|_F^2),$$

where Φ_1, Φ_2, Ω are the matrices defined by (3.1.7).

3.1.2 Singular Subspaces

Let $A \in \mathcal{C}^{m\times n}$, and let $v \in \mathcal{C}^n$ and $u \in \mathcal{C}^m$ be unit right and unit left singular vectors of A associated with the same singular value. Then the pair of one-dimensional subspaces $\mathcal{R}(v)$, $\mathcal{R}(u)$ satisfy $A\mathcal{R}(v) \subset \mathcal{R}(u)$ and $A^H\mathcal{R}(u) \subset \mathcal{R}(v)$, and the pair $\{\mathcal{R}(v), \mathcal{R}(u)\}$ is called a pair of one-dimensional singular subspaces of A. This definition extends in a natural way to higher dimensions.

Let $\mathcal{X}_1$ be a subspace of $\mathcal{C}^n$, $\mathcal{Y}_1$ be a subspace of $\mathcal{C}^m$ with the same dimension as $\mathcal{X}_1$. The pair of subspaces $\mathcal{X}_1, \mathcal{Y}_1$ is called a pair of *singular subspaces* of $A \in \mathcal{C}^{m\times n}$ if $A\mathcal{X}_1 \subset \mathcal{Y}_1$ and $A^H\mathcal{Y}_1 \subset \mathcal{X}_1$.

Let $X_1 \in \mathcal{U}^{n\times l}, Y_1 \in \mathcal{U}^{m\times l}$, and let the columns of X_1, Y_1 form bases for subspaces $\mathcal{X}_1, \mathcal{Y}_1$, respectively. Then it is easy to see that the pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is a pair of singular subspaces of A if and only if there is a matrix $A_1 \in \mathcal{C}^{l\times l}$ such that $AX_1 = Y_1A_1$ and $A^HY_1 = X_1A_1^{\mathrm{H}}$.

Let $X_1 \in \mathcal{U}^{n\times l}$, $Y_1 \in \mathcal{U}^{m\times l}$. It can be proved that the pair of the subspaces $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ is a singular subspace pair of $A \in \mathcal{C}^{m\times n}$ if and only if there are matrices $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ and $Y = (Y_1, Y_2) \in \mathcal{U}^{m\times m}$ such that

$$Y^HAX = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad A_1 \in \mathcal{C}^{l\times l}. \tag{3.1.23}$$

Define the matrix $\hat{A}_2$ by

$$\hat{A}_2 = \begin{cases} A_2 & \text{if } m = n, \\ \\ (A_2, 0) \in \mathcal{C}^{(m-l)\times(m-l)} & \text{if } m > n. \end{cases} \tag{3.1.24}$$

If $\sigma(A_1) \bigcap \sigma(\hat{A}_2) = \emptyset$, then the singular subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is called a *simple* singular subspace pair. The condition $\sigma(A_1) \bigcap \sigma(\hat{A}_2) = \emptyset$ means that

$$\sigma(A_1) \bigcap \sigma(A_2) = \emptyset \qquad \text{if } m = n,$$

$$\sigma(A_1) \bigcap \sigma(A_2) = \emptyset \text{ and } 0 \notin \sigma(A_1) \qquad \text{if } m > n.$$

In this chapter we only consider simple singular subspaces.

In this subsection we prove the following perturbation expansion theorem.

Theorem 3.1.6. *Let $A \in \mathcal{C}^{m\times n}$ $(m \geq n)$ have the decomposition (3.1.23), where $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ and $Y = (Y_1, Y_2) \in \mathcal{U}^{m\times m}$ with $X_1 \in \mathcal{U}^{n\times l}$ and $Y_1 \in \mathcal{U}^{m\times l}$, and*

$$\sigma(A_1) \bigcap \sigma(\hat{A}_2) = \emptyset, \tag{3.1.25}$$

in which the matrix $\hat{A}_2$ is defined by (3.1.24). Moreover, let $\mathcal{X}_1 = \mathcal{R}(X_1), \mathcal{Y}_1 = \mathcal{R}(Y_1)$, for $M \in \mathcal{C}^{m\times n}$ let

$$Y^HMX = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}, \quad M_{11} \in \mathcal{C}^{l\times l}, \tag{3.1.26}$$

and define the linear operator $\mathbf{T} : \mathcal{C}^{(n-l)\times l} \times \mathcal{C}^{(m-l)\times l} \to \mathcal{C}^{(n-l)\times l} \times \mathcal{C}^{(m-l)\times l}$ *by*

$$\mathbf{T}\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} ZA_1^H - A_2^H W \\ -A_2 Z + W A_1 \end{pmatrix}, \quad Z \in \mathcal{C}^{(n-l)\times l}, \quad W \in \mathcal{C}^{(m-l)\times l}. \tag{3.1.27}$$

Then

(1) there is a unique l*-dimensional simple singular subspace pair* $\{\mathcal{X}_1(\tau), \mathcal{Y}_1(\tau)\}$ *of* $A + \tau M$ $(\tau \in \mathcal{R})$ *such that* $\mathcal{X}_1(0) = \mathcal{X}_1$, $\mathcal{Y}_1(0) = \mathcal{Y}_1$, *and the basis vectors* $x_1(\tau), \ldots, x_l(\tau)$ *of* $\mathcal{X}_1(\tau)$ *and the basis vectors* $y_1(\tau), \ldots, y_l(\tau)$ *of* $\mathcal{Y}_1(\tau)$ *may be chosen to be analytic functions of* $\tau \in (-\delta, \delta)$ *for some* $\delta > 0$*;*

(2) the analytic matrix-valued functions

$$X_1(\tau) = (x_1(\tau), \ldots, x_l(\tau)), \qquad Y_1(\tau) = (y_1(\tau), \ldots, y_l(\tau))$$

have the perturbation expansions

$$X_1(\tau) = X_1 + X_2 \sum_{j=1}^{\infty} K_j \tau^j, \quad Y_1(\tau) = Y_1 + Y_2 \sum_{j=1}^{\infty} L_j \tau^j, \quad \tau \in (-\delta, \delta), \tag{3.1.28}$$

in which

$$\begin{pmatrix} K_1 \\ L_1 \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} M_{12}^H \\ M_{21} \end{pmatrix},$$

$$\begin{pmatrix} K_2 \\ L_2 \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} -K_1 M_{11}^H + M_{22}^H L_1 \\ M_{22} K_1 - L_1 M_{11} \end{pmatrix},$$

$$\begin{pmatrix} K_j \\ L_j \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} -K_{j-1} M_{11}^H + M_{22}^H L_{j-1} - \sum_{k=1}^{j-2} K_k M_{21}^H L_{j-1-k} \\ M_{22} K_{j-1} - L_{j-1} M_{11} - \sum_{k=1}^{j-2} L_k M_{12} K_{j-1-k} \end{pmatrix}, \quad j \geq 3. \tag{3.1.29}$$

Proof. Let

$$A(\tau) = A + \tau M, \quad \tilde{A}(\tau) = Y^H A(\tau) X = \begin{pmatrix} \tilde{A}_{11}(\tau) & \tilde{A}_{12}(\tau) \\ \tilde{A}_{21}(\tau) & \tilde{A}_{22}(\tau) \end{pmatrix}, \tag{3.1.30}$$

where $\tilde{A}_{11}(\tau) \in \mathcal{C}^{l\times l}$, and

$$\tilde{A}_{jj}(\tau) = A_j + \tau M_{jj}, \quad j = 1, 2; \qquad \tilde{A}_{jk}(\tau) = \tau M_{jk}, \quad j \neq k. \tag{3.1.31}$$

For $Z \in \mathcal{C}^{(n-l)\times l}, W \in \mathcal{C}^{(m-l)\times l}$ and $\tau \in \mathcal{R}$ define the functions Φ and Ψ by

$$\begin{aligned} \Phi(Z, W, \tau) &= \tilde{A}_{12}(\tau)^H - Z\tilde{A}_{11}(\tau)^H + \tilde{A}_{22}(\tau)^H W - Z\tilde{A}_{21}(\tau)^H W, \\ \Psi(Z, W, \tau) &= \tilde{A}_{21}(\tau) + \tilde{A}_{22}(\tau) Z - W\tilde{A}_{11}(\tau) - W\tilde{A}_{12}(\tau) Z; \end{aligned} \tag{3.1.32}$$

and let

$$\phi = \mathrm{vec}(\Phi) = f + ig, \quad \psi = \mathrm{vec}(\Psi) = p + iq,$$

$$z = \mathrm{vec}(Z) = x + iy, \quad w = \mathrm{vec}(W) = u + iv,$$

where f, g, p, q are real analytic vector-valued functions of x, y, u, v, τ. Applying Theorem 1.6.3 and using the expressions (3.1.31) and (3.1.32) we get

$$\left(\frac{\partial(f,g,p,q)}{\partial(x,y,u,v)}\right)_{\substack{x=y=0\\ u=v=0\\ \tau=0}} = \left|\frac{\partial(\phi,\psi)}{\partial(z,w)}\right|^2_{z=0,\,w=0}$$

$$= \left|\det\begin{pmatrix} -\overline{A}_1 \otimes I_{n-l} & I_l \otimes A_2^H \\ I_l \otimes A_2 & -A_1^T \otimes I_{m-l}\end{pmatrix}\right|^2.$$

Let $A_j = U_j \Sigma_j V_j^H$ be singular value decompositions of A_j for $j = 1, 2$, where U_j, V_j are unitary matrices, and

$$\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_l), \quad \Sigma_2 = \mathrm{diag}(\sigma_{l+1}, \sigma_{l+2}, \ldots).$$

Then

$$\left(\frac{\partial(f,g,p,q)}{\partial(x,y,u,v)}\right)_{\substack{x=y=0\\ u=v=0\\ \tau=0}} = \left|\det\begin{pmatrix} -\Sigma_1 \otimes I_{n-l} & I_l \otimes \Sigma_2^T \\ I_l \otimes \Sigma_2 & -\Sigma_1 \otimes I_{m-l}\end{pmatrix}\right|^2$$

$$= \left[\left(\prod_{j=1}^{l} \sigma_j\right)^{m-n} \prod_{j=1}^{l}\prod_{k=l+1}^{n} (\sigma_j^2 - \sigma_k^2)\right]^2 > 0,$$

where we have used the assumption (3.1.25). Therefore, by the implicit function theorem (Theorem 1.6.2) the equations

$$\Phi(Z, W, \tau) = 0, \quad \Psi(Z, W, \tau) = 0$$

have a unique analytic solution $Z = Z(\tau), W = W(\tau)$ of $\tau \in (-\delta, \delta)$ for some $\delta > 0$ satisfying $Z(0) = 0$ and $W(0) = 0$. Moreover, we may choose δ so small that the matrices $I + Z(\tau)^H Z(\tau)$ and $I + W(\tau)^H W(\tau)$ are nonsingular. Thus, we have

$$\begin{pmatrix} I & W(\tau)^H \\ -W(\tau) & I \end{pmatrix} \tilde{A}(\tau) \begin{pmatrix} I & -Z(\tau)^H \\ Z(\tau) & I \end{pmatrix} = \begin{pmatrix} A_1(\tau) & 0 \\ 0 & A_2(\tau)\end{pmatrix}, \quad (3.1.33)$$

where

$$A_1(\tau) = \tilde{A}_{11}(\tau) + \tilde{A}_{12}(\tau)Z(\tau) + W(\tau)^H \tilde{A}_{21}(\tau) + W(\tau)^H \tilde{A}_{22}(\tau) Z(\tau),$$

$$A_2(\tau) = \tilde{A}_{22}(\tau) - \tilde{A}_{21}(\tau)Z(\tau)^H - W(\tau)\tilde{A}_{12}(\tau) + W(\tau)\tilde{A}_{11}(\tau)Z(\tau)^H,$$

and $\sigma(A_1(\tau))\bigcap\sigma(\hat{A}_2(\tau)) = \emptyset$ for $\tau \in (-\delta,\delta)$ provided that the positive scalar δ is sufficiently small, in which

$$\hat{A}_2(\tau) \equiv \begin{cases} A_2(\tau) & \text{if } m = n, \\ (A_2(\tau), 0) \in \mathcal{C}^{(m-l)\times(m-l)} & \text{if } m > n. \end{cases}$$

From the relations (3.1.33) and (3.1.30) it follows that if we define

$$X_1(\tau) = X\begin{pmatrix} I \\ Z(\tau) \end{pmatrix}, \quad Y_1(\tau) = Y\begin{pmatrix} I \\ W(\tau) \end{pmatrix} \tag{3.1.34}$$

and

$$\mathcal{X}_1(\tau) = \mathcal{R}(X_1(\tau)), \quad \mathcal{Y}_1(\tau) = \mathcal{R}(Y_1(\tau)),$$

then the pair $\{\mathcal{X}_1(\tau),\mathcal{Y}_1(\tau)\}$ is the unique l-dimensional simple singular subspace pair of $A(\tau)$ in $(-\delta,\delta)$ satisfying $\mathcal{X}_1(0) = \mathcal{X}_1$ and $\mathcal{Y}_1(0) = \mathcal{Y}_1$, and $X_1(\tau), Y_1(\tau)$ are analytic matrix-valued functions of $\tau \in (-\delta,\delta)$.

Observe that $Z(\tau), W(\tau)$ satisfy

$$\Phi(Z(\tau), W(\tau), \tau) = 0, \quad \Psi(Z(\tau), W(\tau), \tau) = 0,$$

where $\Phi(Z,W,\tau)$ and $\Psi(Z,W,\tau)$ are defined by (3.1.32), in which $\tilde{A}_{jk}(\tau)$ are expressed by (3.1.31). Hence, we get the basic equations for $Z(\tau), W(\tau)$:

$$\begin{cases} \tau Z(\tau)M_{21}^H W(\tau) + Z(\tau)(A_1 + \tau M_{11})^H - (A_2 + \tau M_{22})^H W(\tau) - \tau M_{12}^H = 0, \\ \tau W(\tau)M_{12}Z(\tau) - (A_2 + \tau M_{22})Z(\tau) + W(\tau)(A_1 + \tau M_{11}) - \tau M_{21} = 0, \end{cases} \tag{3.1.35}$$

where $\tau \in (-\delta,\delta)$.

Differentiating (3.1.35) at $\tau = 0$, and writing

$$Z^{(j)} = \left(\frac{\mathrm{d}^j Z(\tau)}{\mathrm{d}\tau^j}\right)_{\tau=0}, \quad W^{(j)} = \left(\frac{\mathrm{d}^j W(\tau)}{\mathrm{d}\tau^j}\right)_{\tau=0}, \quad j = 1, 2, \ldots,$$

we get

$$\mathbf{T}\begin{pmatrix} Z^{(1)} \\ W^{(1)} \end{pmatrix} = \begin{pmatrix} M_{12}^H \\ M_{21} \end{pmatrix},$$

$$\mathbf{T}\begin{pmatrix} Z^{(2)} \\ W^{(2)} \end{pmatrix} = 2\begin{pmatrix} -Z^{(1)}M_{11}^H + M_{22}^H W^{(1)} \\ M_{22}Z^{(1)} - W^{(1)}M_{11} \end{pmatrix},$$

$$\mathbf{T}\begin{pmatrix} Z^{(j)} \\ W^{(j)} \end{pmatrix} = j\begin{pmatrix} -Z^{(j-1)}M_{11}^H + M_{22}^H W^{(j-1)} - \sum_{k=1}^{j-2}\begin{pmatrix} j-1 \\ k \end{pmatrix} Z^{(k)}M_{21}^H W^{(j-1-k)} \\ M_{22}Z^{(j-1)} - W^{(j-1)}M_{11} - \sum_{k=1}^{j-2}\begin{pmatrix} j-1 \\ k \end{pmatrix} W^{(k)}M_{12}Z^{(j-1-k)} \end{pmatrix}, \tag{3.1.36}$$

where $j \geq 3$, and $\mathbf{T}$ is the linear operator defined by (3.1.27).

The assumption (3.1.25) implies that the operator $\mathbf{T}$ is invertible. Define

$$K_j = \frac{1}{j!}Z^{(j)}, \quad L_j = \frac{1}{j!}W^{(j)}, \quad j = 1, 2, \ldots .$$

Then from (3.1.36) we get the relations (3.1.29) and the power series expansions of $Z(\tau), W(\tau)$ at $\tau = 0$:

$$Z(\tau) = \sum_{j=1}^{\infty}\frac{1}{j!}Z^{(j)}\tau^j = \sum_{j=1}^{\infty}K_j\tau^j, \qquad W(\tau) = \sum_{j=1}^{\infty}\frac{1}{j!}W^{(j)}\tau^j = \sum_{j=1}^{\infty}L_j\tau^j.$$

This together with (3.1.34) gives (3.1.28). □

We now consider a special case where

$$A_1 = \Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_l) \quad \text{with} \quad \sigma_1, \ldots, \sigma_l > 0, \quad \text{and} \quad A_2 = 0. \tag{3.1.37}$$

In such a case, the operator $\mathbf{T}$ defined by (3.1.27) can be written

$$\mathbf{T}\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} Z \\ W \end{pmatrix}\Sigma_1, \quad Z \in \mathcal{C}^{(n-l)\times l}, \ W \in \mathcal{C}^{(m-l)\times l},$$

and the relations of (3.1.29) become

$$\begin{pmatrix} K_1 \\ L_1 \end{pmatrix} = \begin{pmatrix} M_{12}^H \\ M_{21} \end{pmatrix} \Sigma_1^{-1},$$

$$\begin{pmatrix} K_2 \\ L_2 \end{pmatrix} = \begin{pmatrix} -K_1 M_{11}^H + M_{22}^H L_1 \\ M_{22}K_1 - L_1 M_{11} \end{pmatrix} \Sigma_1^{-1},$$

$$\begin{pmatrix} K_j \\ L_j \end{pmatrix} = \begin{pmatrix} -K_{j-1} M_{11}^H + M_{22}^H L_{j-1} - \sum\limits_{k=1}^{j-2} K_k M_{21}^H L_{j-1-k} \\ M_{22}K_{j-1} - L_{j-1} M_{11} - \sum\limits_{k=1}^{j-2} L_k M_{12} K_{j-1-k} \end{pmatrix} \Sigma_1^{-1}, \quad j \geq 3. \tag{3.1.38}$$

Observe that from (3.1.26) $M_{jk} = Y_j^H M X_k$ for $j, k = 1, 2$. Hence, if we let $E = \tau M, \tilde{X}_1 = X_1(\tau)$ and $\tilde{Y}_1 = Y_1(\tau)$, then from (3.1.28), (3.1.29) and (3.1.38) we get the following corollary.

Corollary 3.1.7 (Vaccaro). *Let $A, X, Y, \mathcal{X}_1, \mathcal{Y}_1$ be as in Theorem 3.1.6, and let A_1 and A_2 be the matrices of (3.1.37). If $\|E\|$ is sufficiently small, then there is a unique l-dimensional singular subspace pair $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ of $A + E$ with $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ that $\tilde{X}_1$ and $\tilde{Y}_1$ have the second-order perturbation expansions*

$$\begin{aligned} \tilde{X}_1 = \ & X_1 + X_2 X_2^H E^H Y_1 \Sigma_1^{-1} \\ & -X_2 X_2^H E^H \left(Y_1 \Sigma_1^{-1} X_1^H E^H Y_1 - Y_2 Y_2^H E X_1 \Sigma_1^{-1} \right) \Sigma_1^{-1} + O(\|E\|_F^3), \end{aligned} \tag{3.1.39}$$

and

$$\begin{aligned} \tilde{Y}_1 = \ & Y_1 + Y_2 Y_2^H E X_1 \Sigma_1^{-1} \\ & -Y_2 Y_2^H E \left(X_1 \Sigma_1^{-1} Y_1^H E X_1 - X_2 X_2^H E^H Y_1 \Sigma_1^{-1} \right) \Sigma_1^{-1} + O(\|E\|_F^3), \end{aligned} \tag{3.1.40}$$

where $E \to 0$.

The following result, as a corollary of Theorem 3.1.6, gives modified forms of the first order perturbation expansions of $X_1(\tau)$ and $Y_1(\tau)$.

Corollary 3.1.8. *Let A, X, Y and $\mathbf{T}$ be as in Theorem 3.1.6, and let $\mathcal{X}_1 = \mathcal{R}(X_1), \mathcal{Y}_1 = \mathcal{R}(Y_1)$. Moreover, for $E \in \mathcal{C}^{m \times n}$ let*

$$Y^H E X = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \qquad E_{11} \in \mathcal{C}^{l \times l}.$$

If $\|E\|_F$ is sufficiently small, then there exists a unique l-dimensional singular subspace pair $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ of $A + E$ with $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ that $\tilde{X}_1$ and $\tilde{Y}_1$ have *the expansions*

$$\tilde{X}_1 = X_1 + X_2 Z_1 + O(\|E\|_F^2), \quad \tilde{Y}_1 = Y_1 + X_2 W_1 + O(\|E\|_F^2), \tag{3.1.41}$$

where $E \to 0$, *and* $Z_1, W_1 \in \mathcal{C}^{(n-l)\times l}$ *are defined by*

$$\begin{pmatrix} Z_1 \\ W_1 \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} E_{12}^H \\ E_{21} \end{pmatrix}. \tag{3.1.42}$$

Observe that by using the Kronecker product and vec operator, the matrix representation T of the linear operator $\mathbf{T}$ defined by (3.1.27) can be expressed by

$$T = \begin{pmatrix} \overline{A}_1 \otimes I_{n-l} & -I_l \otimes A_2^H \\ -I_l \otimes A_2 & A_1^T \otimes I_{m-l} \end{pmatrix}.$$

Hence, the relation (3.1.42) can be written

$$\begin{pmatrix} \mathrm{vec}(Z_1) \\ \mathrm{vec}(W_1) \end{pmatrix} = C \begin{pmatrix} \mathrm{vec}(E_{12}^H) \\ \mathrm{vec}(E_{21}) \end{pmatrix}, \tag{3.1.43}$$

where

$$C \equiv T^{-1} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \tag{3.1.44}$$

in which

$$\begin{aligned} C_1 &= \Big((A_1^T \otimes I_{n-l})K^{-1},\, (I_l \otimes A_2^H)L^{-1}\Big), \\ C_2 &= \Big((I_l \otimes A_2)K^{-1},\, (\overline{A}_1 \otimes I_{m-l})L^{-1}\Big), \end{aligned} \tag{3.1.45}$$

and

$$K = \overline{A}_1 A_1^T \otimes I_{n-l} - I_l \otimes A_2^H A_2, \quad L = A_1^T \overline{A}_1 \otimes I_{m-l} - I_l \otimes A_2 A_2^H. \tag{3.1.46}$$

Notes and References

NR 3.1–1. This section is based on Sun [105] and [119].

NR 3.1–2. MacFarlane and Hung [74] consider the singular values of a rational matrix-valued function of a complex variable. Analytic properties and Taylor series expansions of the singular values are studied. The technique used in [74] is different from that in §3.2.1.

NR 3.1–3. A second order perturbation expansion for small singular values of a matrix A is derived by Stewart [95]. The key step is to work with the cross-product matrix $A^H A$ and to get a second order perturbation expansion of the corresponding small eigenvalues of $A^H A$.

NR 3.1–4. Elsner and He [34] consider the matrix $G(s) = G_1 + sG_2$, where s is a real parameter, G_1 and G_2 are complex matrices. The smallest singular value $\sigma(s)$ of $G(s)$ is assumed positive and simple. Explicit expressions of the first and second order derivatives of $\sigma(s)$ are obtained in [34], which coincide with the results of Theorem 3.1.1 with $N = 1$. The explicit expressions of derivatives serve as a basis for an algorithm to compute the distance to *uncontrollability.*

NR 3.1–5. Let $A(p)$ and $\mathcal{B}(0)$ be as in Theorem 3.1.1. If σ_* is a *zero* singular value or a *multiple* singular value of $A(0)$, then, in general, there is no a real differentiable function $\sigma(p) \geq 0$ defined in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin such that $\sigma(p)$ is a singular value of $A(p)$ in $\mathcal{B}_0$, and $\sigma(0) = \sigma_*$. Sun [106] studies the existence and expressions of the directional derivatives of zero singular values and multiple singular values.

NR 3.1–6. The second order perturbation expansions (3.1.39) and (3.1.40) are derived by Vaccaro [123] in another way. Vaccaro [123] points out that the expressions can be used to analyze the performance of *direction-finding algorithms* in *array signal processing.*

3.2 Condition Numbers

3.2.1 Simple Non-Zero Singular Values

Let $A \in \mathcal{C}^{m\times n}$, and $\sigma > 0$ be a simple singular value of A. Let $\tilde{A} = A + E$ be a perturbation of A, and $\tilde{\sigma}$ be the corresponding perturbation of σ. Then by (1.8.1) we define the condition number $c(\sigma)$ for σ as

$$c(\sigma) = \lim_{\delta\to 0} \sup_{\frac{\|E\|}{\alpha}\leq\delta} \frac{|\tilde{\sigma} - \sigma|}{\xi\delta},$$

where α and ξ are positive parameters.

From the definition of $c(\sigma)$ we see that in first order approximation the inequality

$$\frac{|\tilde{\sigma} - \sigma|}{\xi} \leq c(\sigma)\frac{\|E\|}{\alpha}$$

holds.

Let $v \in \mathcal{C}^n$ and $u \in \mathcal{C}^m$ be the unit right and unit left singular vectors of A associated with σ, respectively. Then from Corollary 3.1.4

$$\tilde{\sigma} = \sigma + \mathsf{Re}(u^H E v) + O(\|E\|^2).$$

Consequently,

$$c(\sigma) = \alpha \sup_{\|E\| \leq 1} \frac{\left|\mathsf{Re}(u^H E v)\right|}{\xi} = \frac{\alpha}{\xi}.$$

Taking $\alpha = \xi = 1$ yields the absolute condition number $c_{\text{abs}}(\sigma) = 1$, and taking $\alpha = \|A\|$ and $\xi = \sigma$ yields the relative condition number $c_{\text{rel}}(\sigma) = \|A\|/\sigma$.

Note that the formula of the condition number $c(\sigma)$ is generalized to simple generalized singular values by Sun [122].

3.2.2 Singular Subspaces

Let $A \in \mathcal{C}^{m\times n}$, and $\{\mathcal{X}_1, \mathcal{Y}_1\}$ be a simple singular subspace pair of A. Let $\tilde{A} = A + E$ be a perturbation of A, and $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ be the corresponding perturbation of $\{\mathcal{X}_1, \mathcal{Y}_1\}$. Then by (1.8.3) we define the condition numbers $c(\mathcal{X}_1), c(\mathcal{Y}_1)$ for $\mathcal{X}_1, \mathcal{Y}_1$ as

$$c(\mathcal{X}_1) = \lim_{\delta\to 0} \sup_{\frac{\|E\|_F}{\alpha} \leq \delta} \frac{\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1)}{\delta}, \qquad c(\mathcal{Y}_1) = \lim_{\delta\to 0} \sup_{\frac{\|E\|_F}{\alpha} \leq \delta} \frac{\rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1)}{\delta}, \tag{3.2.1}$$

where γ is a positive parameter, and $\rho_F(\cdot,\cdot)$ is the generalized chordal metric defined by (1.3.3).

From the definition (3.2.1) we see that in first order approximation the inequalities

$$\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \leq c(\mathcal{X}_1)\frac{\|E\|_F}{\alpha}, \qquad \rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) \leq c(\mathcal{Y}_1)\frac{\|E\|_F}{\alpha}$$

hold.

By (3.2.1), (3.1.41) and Theorem 1.3.3 (see (1.3.17)),

$$c(\mathcal{X}_1) = \lim_{\delta\to 0} \sup_{\frac{\|E\|_F}{\alpha} \leq \delta} \frac{\|Z_1\|_F}{\delta}, \tag{3.2.2}$$

where Z_1 is defined by (3.1.42). Combining (3.2.2) with (3.1.43)–(3.1.46) gives

$$\begin{aligned} c(\mathcal{X}_1) &= \lim_{\delta\to 0} \sup_{\frac{\|\text{vec}(E)\|_2}{\alpha} \leq \delta} \frac{\|\text{vec}(Z_1)\|_2}{\delta} = \alpha \sup_{\|\text{vec}(E)\|_2 \leq 1} \left\| C_1 \begin{pmatrix} \text{vec}(E_{12}^H) \\ \text{vec}(E_{21}) \end{pmatrix} \right\|_2 \\ &= \alpha \sup_{\left\| \begin{pmatrix} \text{vec}(E_{12}^H) \\ \text{vec}(E_{21}) \end{pmatrix} \right\|_2 \leq 1} \left\| C_1 \begin{pmatrix} \text{vec}(E_{12}^H) \\ \text{vec}(E_{21}) \end{pmatrix} \right\|_2 = \alpha \|C_1\|_2. \end{aligned}$$

Similarly, we have $c(\mathcal{Y}_1) = \alpha\|C_2\|_2$.

Let $A_j = U_j \Sigma_j V_j^H$ be the singular value decomposition of A_j for $j = 1, 2$, and let

$$C_{10} = \left((\Sigma_1 \otimes I_{n-l}) K_0^{-1},\ (I_l \otimes \Sigma_2^T) L_0^{-1} \right),$$

$$C_{20} = \left((I_l \otimes \Sigma_2) K_0^{-1},\ (\Sigma_1 \otimes I_{m-l}) L_0^{-1} \right),$$

where

$$\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_l), \qquad \Sigma_2 = \operatorname{diag}(\sigma_{l+1}, \sigma_{l+2}, \ldots),$$

$$K_0 = \Sigma_1^2 \otimes I_{n-l} - I_l \otimes \Sigma_2^T \Sigma_2, \qquad L_0 = \Sigma_1^2 \otimes I_{m-l} - I_l \otimes \Sigma_2 \Sigma_2^T.$$

Then we have

$$c(\mathcal{X}_1) = \alpha \|C_1\|_2 = \alpha \|C_{10}\|_2 = \alpha \max_{\substack{1 \le j \le l \\ l+1 \le k \le n}} \frac{\sqrt{\sigma_j^2 + \sigma_k^2}}{|\sigma_j^2 - \sigma_k^2|}, \tag{3.2.3}$$

and

$$c(\mathcal{Y}_1) = \alpha \|C_2\|_2 = \alpha \|C_{20}\|_2 = \alpha \max_{\substack{1 \le j \le l \\ l+1 \le k \le m}} \frac{\sqrt{\sigma_j^2 + \sigma_k^2}}{|\sigma_j^2 - \sigma_k^2|}, \tag{3.2.4}$$

where we define

$$\sigma_{n+1} = \cdots = \sigma_m = 0 \quad \text{if } m > n.$$

From (3.2.3) and (3.2.4) we see that

$$c(\mathcal{Y}_1) = \begin{cases} c(\mathcal{X}_1) & \text{if } m = n, \\ \max\left\{ c(\mathcal{X}_1),\ \alpha \max\limits_{1 \le j \le l} \frac{1}{\sigma_j} \right\} & \text{if } m > n. \end{cases}$$

Consequently, the expressions (3.2.3) and (3.2.4) reveal an important fact: If $m > n$ then, in general, the singular subspaces $\mathcal{X}_1$ and $\mathcal{Y}_1$ have different condition numbers $c(\mathcal{X}_1)$ and $c(\mathcal{Y}_1)$, respectively.

Taking $\alpha = 1$ yields the absolute condition numbers

$$c_{\text{abs}}(\mathcal{X}_1) = \max_{\substack{1 \le j \le l \\ l+1 \le k \le n}} \frac{\sqrt{\sigma_j^2 + \sigma_k^2}}{|\sigma_j^2 - \sigma_k^2|} \tag{3.2.5}$$

and

$$c_{\text{abs}}(\mathcal{Y}_1) = \max_{\substack{1 \le j \le l \\ l+1 \le k \le m}} \frac{\sqrt{\sigma_j^2 + \sigma_k^2}}{|\sigma_j^2 - \sigma_k^2|}, \tag{3.2.6}$$

and taking $\alpha = \|A\|_F$ yields the relative condition numbers

$$c_{\rm rel}(\mathcal{X}_1) = \|A\|_F c_{\rm abs}(\mathcal{X}_1), \quad c_{\rm rel}(\mathcal{Y}_1) = \|A\|_F c_{\rm abs}(\mathcal{Y}_1). \tag{3.2.7}$$

Remark 3.2.1. Let $\mathbf{T}$ be the operator defined by (3.1.27), and define

$$\|\mathbf{T}\| = \max_{\substack{Z \in \mathcal{C}^{(n-l)\times l}, W \in \mathcal{C}^{(m-l)\times l} \\ \left\| \begin{pmatrix} Z \\ W \end{pmatrix} \right\|_F = 1}} \left\| \mathbf{T} \begin{pmatrix} Z \\ W \end{pmatrix} \right\|_F .$$

Then

$$\|\mathbf{T}^{-1}\| = \|C\|_2 = \max_{\substack{1 \le j \le l \\ l+1 \le k \le m}} \frac{1}{|\sigma_j - \sigma_k|} \equiv c(\mathcal{X}_1, \mathcal{Y}_1), \tag{3.2.8}$$

where C, the matrix representation of $\mathbf{T}^{-1}$, is expressed by (3.1.44)–(3.1.46), and $\sigma_{n+1} = \cdots = \sigma_m = 0$ if $m > n$. Usually, by Stewart [96, Theorems 6.3 and 6.4], the quantity $c(\mathcal{X}_1, \mathcal{Y}_1)$ defined by (3.2.8) is regarded as the (absolute) condition number of the singular subspaces $\mathcal{X}_1, \mathcal{Y}_1$ of A. However, observe that the expressions (3.2.5), (3.2.6) and (3.2.8) imply that

$$\frac{1}{\sqrt{2}} c(\mathcal{X}_1, \mathcal{Y}_1) \le c_{\rm abs}(\mathcal{X}_1) = c_{\rm abs}(\mathcal{Y}_1) \le c(\mathcal{X}_1, \mathcal{Y}_1) \quad \text{if } m = n, \tag{3.2.9}$$

and in the case of $m > n$,

$$\begin{gathered} c_{\rm abs}(\mathcal{X}_1) \le c_{\rm abs}(\mathcal{Y}_1) \le c(\mathcal{X}_1, \mathcal{Y}_1), \\ c_{\rm abs}(\mathcal{X}_1) \ll c_{\rm abs}(\mathcal{Y}_1) \le c(\mathcal{X}_1, \mathcal{Y}_1) \quad \text{if } c_{\rm abs}(\mathcal{X}_1) \ll \max_{1 \le j \le l} \tfrac{1}{\sigma_j}, \\ \tfrac{1}{\sqrt{2}} c(\mathcal{X}_1, \mathcal{Y}_1) \le c_{\rm abs}(\mathcal{Y}_1) \le c(\mathcal{X}_1, \mathcal{Y}_1). \end{gathered} \tag{3.2.10}$$

Hence, the condition number $c(\mathcal{X}_1, \mathcal{Y}_1)$ may be a severe overestimate of the sensitivity of the right singular subspace $\mathcal{X}_1$ in some cases.

Example 3.2.2. Consider the matrix

$$A = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{pmatrix} \quad \text{with } \sigma_1 = 10^{-8}, \ \sigma_2 = 1.$$

Let

$$x_1 = (1,\ 0)^T, \quad y_1 = (1,\ 0,\ 0)^T.$$

Then the pair of the subspaces $\mathcal{X}_1 = \mathcal{R}(x_1)$ and $\mathcal{Y}_1 = \mathcal{R}(y_1)$ is the singular subspace pair of A associated with σ_1. By (3.2.5) and (3.2.6) we have

$$c_{\rm abs}(\mathcal{X}_1) = \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{|\sigma_1^2-\sigma_2^2|} \approx 1, \qquad c_{\rm abs}(\mathcal{Y}_1) = \max\left\{ \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{|\sigma_1^2-\sigma_2^2|},\ \frac{1}{\sigma_1} \right\} = 10^8,$$

$$c(\mathcal{X}_1,\mathcal{Y}_1) = \max\left\{ \frac{1}{|\sigma_1-\sigma_2|},\ \frac{1}{\sigma_1} \right\} = 10^8,$$

and

$$c_{\rm abs}(\mathcal{X}_1) \ll c_{\rm abs}(\mathcal{Y}_1) = c(\mathcal{X}_1,\mathcal{Y}_1).$$

Obviously, $c(\mathcal{X}_1,\mathcal{Y}_1)$ is a severe overestimate of the sensitivity of the right singular subspace $\mathcal{X}_1$.

Notes and References

NR 3.2–1. The condition numbers $c_{\rm abs}(\mathcal{X}_1)$ and $c_{\rm abs}(\mathcal{Y}_1)$ are given by Sun [119]. From the analysis of Remark 3.2.1 we see that $c(\mathcal{X}_1,\mathcal{Y}_1)$ and $c_{\rm abs}(\mathcal{Y}_1)$ are qualitatively the same; but $c_{\rm abs}(\mathcal{X}_1) \le c(\mathcal{X}_1,\mathcal{Y}_1)$, and in some cases $c_{\rm abs}(\mathcal{X}_1) \ll c(\mathcal{X}_1,\mathcal{Y}_1)$. The drawback of the condition number $c(\mathcal{X}_1,\mathcal{Y}_1)$ is that it is governed by the ill-conditioning of the most sensitive subspace of a singular subspace pair.

3.3 Perturbation Bounds for Singular Subspaces

A perturbation bound for a pair of simple singular subspaces has been obtained by Stewart [96, Theorem 6.4]. We now apply Theorem 3.3.5 at the end of this subsection to derive a new result. The difference between the new result and Stewart's result is that the new result gives an individual perturbation bound for each subspace in a pair of singular subspaces, separately.

Theorem 3.3.1. *Let $A, X, Y, \mathcal{X}_1, \mathcal{Y}_1$ be as in Theorem 3.1.6. For $E \in \mathcal{C}^{m\times n}$, let*

$$Y^H E X = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \qquad E_{11} \in \mathcal{C}^{l\times l}. \tag{3.3.1}$$

Moreover, let $c_{\rm abs}(\mathcal{X}_1), c_{\rm abs}(\mathcal{Y}_1)$ be the condition numbers expressed by (3.2.5) and (3.2.6), and let

$$c_* = \sqrt{[c_{\rm abs}(\mathcal{X}_1)]^2 + [c_{\rm abs}(\mathcal{Y}_1)]^2}, \qquad \epsilon = \|E_{11}\|_2 + \|E_{22}\|_2, \tag{3.3.2}$$

and

$$\gamma = \sqrt{\|E_{12}\|_F^2 + \|E_{21}\|_F^2}, \qquad \eta = \max\{\|E_{12}\|_2,\ \|E_{21}\|_2\}. \tag{3.3.3}$$

If

$$c_*(2\sqrt{\gamma\eta} + \epsilon) < 1, \tag{3.3.4}$$

then there is a unique pair of l-dimensional singular subspaces $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1), \tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ of $A+E$ such that $\tilde{X}_1 \in \mathcal{U}^{n\times l}$, $\tilde{Y}_1 \in \mathcal{U}^{m\times l}$, and

$$\begin{aligned} \rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) &\le \|\tan\Theta(X_1, \tilde{X}_1)\|_F \le \frac{2c_{\text{abs}}(\mathcal{X}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}}, \\ \rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) &\le \|\tan\Theta(Y_1, \tilde{Y}_1)\|_F \le \frac{2c_{\text{abs}}(\mathcal{Y}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}}, \end{aligned} \tag{3.3.5}$$

where $\Theta(\cdot,\cdot)$ is defined by (1.3.1).

Proof. Let $\mathbf{T}$ be the linear operator defined by (3.1.27). It is easy to verify that $\begin{pmatrix} Z \\ W \end{pmatrix}$ is a solution of the equation

$$\mathbf{T}\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} E_{12}^H \\ E_{21} \end{pmatrix} + \begin{pmatrix} -ZE_{11}^H + E_{22}^H W \\ E_{22}Z - WE_{11} \end{pmatrix} - \begin{pmatrix} ZE_{21}^H W \\ WE_{12}Z \end{pmatrix} \tag{3.3.6}$$

if and only if Z and W satisfy

$$\begin{aligned} \begin{pmatrix} I & 0 \\ -W & I \end{pmatrix} \begin{pmatrix} A_1 + E_{11} & E_{12} \\ E_{21} & A_2 + E_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ Z & I \end{pmatrix} &= \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}, \\ \begin{pmatrix} I & 0 \\ -Z & I \end{pmatrix} \begin{pmatrix} A_1 + E_{11} & E_{12} \\ E_{21} & A_2 + E_{22} \end{pmatrix}^H \begin{pmatrix} I & 0 \\ W & I \end{pmatrix} &= \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}. \end{aligned} \tag{3.3.7}$$

The relations of (3.3.7) imply that the pair of the subspaces

$$\tilde{\mathcal{X}}_1 = \mathcal{R}\left(X\begin{pmatrix} I \\ Z \end{pmatrix}\right), \quad \tilde{\mathcal{Y}}_1 = \mathcal{R}\left(Y\begin{pmatrix} I \\ W \end{pmatrix}\right)$$

is a pair of l-dimensional singular subspaces of $A+E$. Consequently, by (1.3.12) and (1.3.16), the problem of proving (3.3.5) is reduced to the problem of finding a solution $\begin{pmatrix} Z^* \\ W^* \end{pmatrix}$ of (3.3.6) in a certain neighborhood of the origin.

Let C_1, C_2 be the matrices defined by (3.1.45) and (3.1.46), and let

$$z = \text{vec}(Z), \quad w = \text{vec}(W), \quad e_{12} = \text{vec}(E_{12}^H), \quad e_{21} = \text{vec}(E_{21}),$$

$$x(z,w) = \text{vec}(-ZE_{11}^H + E_{22}^H W), \quad y(z,w) = \text{vec}(E_{22}Z - WE_{11}) \tag{3.3.8}$$

and

$$u(z,w) = \text{vec}(ZE_{21}^H W), \quad v(z,w) = \text{vec}(WE_{12}Z). \tag{3.3.9}$$

Then the equation (3.3.6) can be written in an equivalent form

$$\begin{cases} z = C_1 \left[\begin{pmatrix} e_{12} \\ e_{21} \end{pmatrix} + \begin{pmatrix} x(z,w) \\ y(z,w) \end{pmatrix} - \begin{pmatrix} u(z,w) \\ v(z,w) \end{pmatrix} \right], \\ \\ w = C_2 \left[\begin{pmatrix} e_{12} \\ e_{21} \end{pmatrix} + \begin{pmatrix} x(z,w) \\ y(z,w) \end{pmatrix} - \begin{pmatrix} u(z,w) \\ v(z,w) \end{pmatrix} \right]. \end{cases} \tag{3.3.10}$$

Define the functions f and h by

$$f = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad h = \begin{pmatrix} u \\ v \end{pmatrix}.$$

Observe that f and h satisfy the conditions (3.3.21) and (3.3.22) (see below Theorem 3.3.4), where ϵ and η are the scalars defined by (3.3.2) and (3.3.3), respectively. Hence, by Theorem 3.3.5 at the end of this subsection, if

$$c_*\epsilon < 1 \quad \text{and} \quad \frac{4c_*^2\gamma\eta}{(1-c_*\epsilon)^2} < 1,$$

or equivalently, if $c_*, \epsilon, \gamma, \eta$ satisfy (3.3.4), then the system of equations (3.3.10) has a unique solution $\begin{pmatrix} z^* \\ w^* \end{pmatrix}$ (or equivalently, the equation (3.3.6) has a unique solution $\begin{pmatrix} Z^* \\ W^* \end{pmatrix}$) satisfying

$$\|Z^*\|_F = \|z^*\|_2 \le \frac{2c_{\text{abs}}(\mathcal{X}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}},$$

$$\|W^*\|_F = \|w^*\|_2 \le \frac{2c_{abs}(\mathcal{Y}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}}.$$

Combining it with (1.3.12) and (1.3.16) shows the inequalities of (3.3.5). □

Remark 3.3.2. The estimates (3.3.5) imply that if $c_*(2\sqrt{\gamma\eta} + \epsilon)$ is sufficiently small, or more intuitively, if $\|E\|$ is sufficiently small, then

$$\|\tan\Theta(X_1, \tilde{X}_1)\|_F \lesssim c_{\text{abs}}(\mathcal{X}_1)\gamma, \qquad \|\tan\Theta(Y_1, \tilde{Y}_1)\|_F \lesssim c_{\text{abs}}(\mathcal{Y}_1)\gamma. \tag{3.3.11}$$

Note that by Stewart [114, Theorem 6.4], we have

$$\left\| \begin{pmatrix} \tan\Theta(X_1, \tilde{X}_1) \\ \tan\Theta(Y_1, \tilde{Y}_1) \end{pmatrix} \right\|_F \lesssim c(\mathcal{X}_1, \mathcal{Y}_1)\gamma \tag{3.3.12}$$

when $\|E\|$ is sufficiently small, where $c(\mathcal{X}_1, \mathcal{Y}_1)$ is defined by (3.2.8). The relations (3.2.9) and (3.2.10) show that the bounds of (3.3.11) and (3.3.12) are, in general, qualitatively the same, but in some cases the result (3.3.11) is better (even much better) than (3.3.12) if one needs to bound perturbations of each subspace of the pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$, separately. The drawback of the bound (3.3.12) is that it is governed by the *ill-conditioning* of the most sensitive subspace of the singular subspace pair.

If the matrices E_{jk} of (3.3.1) are known, then we can apply Theorem 3.3.4 (which is given below) to derive the following result on perturbation bounds for singular subspaces which will be used in §3.4.2.

Theorem 3.3.3. *Let $A, X, Y, A_1, A_2, \mathcal{X}_1, \mathcal{Y}_1, E$ and E_{jk} $(j,k = 1,2)$ be as in Theorem 3.3.1, and C_1, C_2 be the matrices defined by (3.1.45) and (3.1.46). Moreover, let*

$$\begin{aligned} b_1 &= \left\| C_1 \begin{pmatrix} \mathrm{vec}(E_{12}^H) \\ \mathrm{vec}(E_{21}) \end{pmatrix} \right\|_2, \qquad c_1 = \|C_1\|_2, \\ b_2 &= \left\| C_2 \begin{pmatrix} \mathrm{vec}(E_{12}^H) \\ \mathrm{vec}(E_{21}) \end{pmatrix} \right\|_2, \qquad c_2 = \|C_2\|_2, \\ b &= b_1 + b_2, \qquad c = c_1 + c_2, \end{aligned} \tag{3.3.13}$$

and let

$$\eta = \max\{\|E_{12}\|_2,\ \|E_{21}\|_2\}, \qquad \epsilon = \|E_{11}\|_2 + \|E_{22}\|_2. \tag{3.3.14}$$

If

$$c\epsilon + 2\sqrt{bc\eta} < 1, \tag{3.3.15}$$

then there is a unique pair of singular subspaces $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$, $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ of $A + E$ such that $\tilde{X}_1 \in \mathcal{U}^{n\times l}$, $\tilde{Y}_1 \in \mathcal{U}^{m\times l}$, and

$$\begin{aligned} \rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) &\le \|\tan\Theta(X_1, \tilde{X}_1)\|_F \le b_1 + c_1(\epsilon\beta + \eta\beta^2), \\ \rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) &\le \|\tan\Theta(Y_1, \tilde{Y}_1)\|_F \le b_2 + c_2(\epsilon\beta + \eta\beta^2), \end{aligned} \tag{3.3.16}$$

where

$$\beta = \frac{2b}{1 - c\epsilon + \sqrt{(1-c\epsilon)^2 - 4bc\eta}}. \tag{3.3.17}$$

Proof. From the proof of Theorem 3.3.1 we see that it only needs to show the following fact: Under the assumptions (3.3.15) the system (3.3.10) has a unique solution $\begin{pmatrix} z^* \\ w^* \end{pmatrix}$ satisfying

$$\begin{aligned} \|z^*\|_2 &\le b_1 + c_1(\epsilon\beta + \eta\beta^2), \\ \|w^*\|_2 &\le b_2 + c_2(\epsilon\beta + \eta\beta^2), \end{aligned} \tag{3.3.18}$$

where β is the scalar defined by (3.3.17).

Applying Theorem 3.3.4 (see below) to the system (3.3.10), and using the fact that the assumption (3.3.15) is equivalent to

$$c\epsilon < 1 \quad \text{and} \quad \frac{4bc\eta}{(1-c\epsilon)^2} < 1,$$

we get the estimates (3.3.18) immediately. □

We now prove two general results on solutions of some nonlinear equations. The first one, Theorem 3.3.4, is an extension of Theorem 2.3.4 for finite-dimensional spaces, which can be used to establish the existence of $\begin{pmatrix} z^* \\ w^* \end{pmatrix}$ in Theorem 3.3.3.

Theorem 3.3.4. *Let* $x = (x_1^T, x_2^T)^T \in \mathcal{C}^m$ *with* $x_j \in \mathcal{C}^{m_j}$ *for* $j = 1, 2$, $g \in \mathcal{C}^n$, *and*

$$C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \quad \text{with} \quad C_j \in \mathcal{C}^{m_j \times n}, \;\; j = 1, 2. \tag{3.3.19}$$

Let

$$\begin{aligned} b &= b_1 + b_2 \quad \text{with} \quad b_j = \|C_j g\|_2, \\ c &= c_1 + c_2 \quad \text{with} \quad c_j = \|C_j\|_2, \end{aligned} \tag{3.3.20}$$

and let $f, h : \mathcal{C}^m \to \mathcal{C}^n$ *be two continuous mappings satisfying*

$$\|f(x)\|_2 \le \epsilon \|x\|_2, \qquad \|f(\tilde{x}) - f(x)\|_2 \le \epsilon \|\tilde{x} - x\|_2 \tag{3.3.21}$$

and

$$\|h(x)\|_2 \le \eta \|x\|_2^2, \qquad \|h(\tilde{x}) - h(x)\|_2 \le 2\eta \max\{\|\tilde{x}\|_2, \|x\|_2\} \|\tilde{x} - x\|_2 \tag{3.3.22}$$

for some $\epsilon, \eta \ge 0$. *If*

$$c\epsilon < 1 \quad \text{and} \quad \frac{4bc\eta}{(1-c\epsilon)^2} < 1, \tag{3.3.23}$$

then there is a unique solution x^* *of the nonlinear equation*

$$x = C[g + f(x) + h(x)] \tag{3.3.24}$$

that satisfies

$$\|x_j^*\|_2 \le b_j + c_j(\epsilon\beta + \eta\beta^2) \equiv \xi_j^*, \qquad j = 1, 2, \tag{3.3.25}$$

where

$$\beta = \frac{2b}{1 - c\epsilon + \sqrt{(1-c\epsilon)^2 - 4bc\eta}}. \tag{3.3.26}$$

Proof. Define

$$\mathcal{S}_{\xi_1^*, \xi_2^*} = \left\{ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \; : \; x_j \in \mathcal{C}^{m_j}, \; \|x_j\|_2 \le \xi_j^*, \;\; j = 1, 2 \right\}. \tag{3.3.27}$$

We first prove that if there is a solution of (3.3.24) in $\mathcal{S}_{\xi_1^*,\xi_2^*}$, then it is unique.

Assume that the equation (3.3.24) has different solutions $x^*, \hat{x} \in \mathcal{S}_{\xi_1^*,\xi_2^*}$. Then by (3.3.20)–(3.3.24) we have

$$\begin{aligned} \|\hat{x} - x^*\|_2 &\leq c\left(\epsilon\|\hat{x} - x^*\|_2 + 2\eta \max\{\|\hat{x}\|_2, \|x^*\|_2\}\|\hat{x} - x^*\|_2\right) \\ &\leq c\left(\epsilon + 2\eta[b + c(\epsilon\beta + \eta\beta^2)]\right)\|\hat{x} - x^*\|_2 \\ &< c\left(\epsilon + 2\eta\left[b + c\epsilon \cdot \frac{2b}{1 - c\epsilon} + c\eta \cdot \left(\frac{2b}{1 - c\epsilon}\right)^2\right]\right)\|\hat{x} - x^*\|_2 \qquad (3.3.28) \\ &= \left(c\epsilon + 2bc\eta + \frac{4bc^2\epsilon\eta}{1 - c\epsilon} + \frac{8b^2c^2\eta^2}{(1 - c\epsilon)^2}\right)\|\hat{x} - x^*\|_2. \end{aligned}$$

Observe that the assumptions (3.3.23) imply

$$2bc\eta < \frac{1}{2}(1 - c\epsilon)^2, \qquad \frac{4bc^2\epsilon\eta}{1 - c\epsilon} < c\epsilon(1 - c\epsilon), \qquad \frac{8b^2c^2\eta^2}{(1 - c\epsilon)^2} < \frac{1}{2}(1 - c\epsilon)^2.$$

Hence, from (3.3.28)

$$\|\hat{x} - x^*\|_2 < \|\hat{x} - x^*\|_2.$$

This contradiction shows that there is at most one solution of the equation (3.3.24) in $\mathcal{S}_{\xi_1^*,\xi_2^*}$.

We now prove the existence of a solution of (3.3.24) in $\mathcal{S}_{\xi_1^*,\xi_2^*}$.

Consider the continuous mapping $\mathcal{M} : \mathcal{C}^{m_1} \times \mathcal{C}^{m_2} \to \mathcal{C}^{m_1} \times \mathcal{C}^{m_2}$ defined by

$$y = C[g + f(x) + h(x)]. \qquad (3.3.29)$$

Since any fixed point of the mapping $\mathcal{M}$ is a solution of the equation (3.3.24), the problem of finding a solution of (3.3.24) satisfying (3.3.25) reduces to the problem of showing that there is a fixed point of the mapping $\mathcal{M}$ in $\mathcal{S}_{\xi_1^*,\xi_2^*}$.

Let ξ_1^* and ξ_2^* be the scalars defined by (3.3.25), in which β is defined by (3.3.26). It can be verified that β is a solution of the equation

$$c\eta\beta^2 - (1 - c\epsilon)\beta + b = 0.$$

Combining this fact with (3.3.20) and (3.3.25) shows that ξ_1^* and ξ_2^* satisfy the relations $\xi_1 + \xi_2 = \beta$ and

$$\xi_j = b_j + c_j\left[\epsilon(\xi_1 + \xi_2) + \eta(\xi_1 + \xi_2)^2\right], \quad j = 1, 2. \qquad (3.3.30)$$

Let $x \in \mathcal{S}_{\xi_1^*,\xi_2^*}$. Then by (3.3.29), y satisfies

$$\begin{aligned}
\|y_j\|_2 &\le b_j + c_j \left(\epsilon\|x\|_2 + \eta\|x\|_2^2\right) \qquad \text{(by (3.3.20) – (3.3.22))} \\
&\le b_j + c_j \left[\epsilon\sqrt{\xi_1^{*2} + \xi_2^{*2}} + \eta\left(\xi_1^{*2} + \xi_2^{*2}\right)\right] \qquad \text{(by (3.3.25))} \\
&\le b_j + c_j \left[\epsilon(\xi_1^* + \xi_2^*) + \eta(\xi_1^* + \xi_2^*)^2\right] \\
&= \xi_j^*, \qquad j = 1, 2, \qquad \text{(by (3.3.30))}
\end{aligned}$$

which means that for the mapping $\mathcal{M}$ defined by (3.3.29) we have

$$x \in \mathcal{S}_{\xi_1^*,\xi_2^*} \implies y \in \mathcal{S}_{\xi_1^*,\xi_2^*}. \tag{3.3.31}$$

Observe that $\mathcal{S}_{\xi_1^*,\xi_2^*}$ is a bounded closed convex set of $\mathcal{C}^{m_1} \times \mathcal{C}^{m_2}$, and (3.3.31) shows that the continuous mapping $\mathcal{M}$ maps $\mathcal{S}_{\xi_1^*,\xi_2^*}$ into $\mathcal{S}_{\xi_1^*,\xi_2^*}$. By the Schauder fixed-point theorem (Theorem 1.7.2), the mapping $\mathcal{M}$ has a fixed point in $\mathcal{S}_{\xi_1^*,\xi_2^*}$, and thus the equation (3.3.24) has a solution in $\mathcal{S}_{\xi_1^*,\xi_2^*}$. □.

Theorem 3.3.4 shows the existence and uniqueness of a solution x^* of the equation (3.3.24) in $\mathcal{S}_{\xi_1^*,\xi_2^*}$ under the assumption that the vector g itself is known. However, in some applications, the vector g itself is unknown but some upper bound for $\|g\|_2$ is known. In such a case, we have the following result on the existence of some solution to the equation (3.3.24), which can be used to establish the existence of $\begin{pmatrix} z^* \\ w^* \end{pmatrix}$ in Theorem 3.3.1.

Theorem 3.3.5. *Let $x, g, C, c_1, c_2, f(x), h(x), \epsilon, \eta$ be as in Theorem 3.3.4, and let*

$$\gamma = \|g\|_2, \qquad c_* = \sqrt{c_1^2 + c_2^2}. \tag{3.3.32}$$

If

$$c_*\epsilon < 1 \qquad \text{and} \qquad \frac{4c_*^2\gamma\eta}{(1 - c_*\epsilon)^2} < 1, \tag{3.3.33}$$

then there is a unique solution of the nonlinear equation (3.3.24) that satisfies

$$\|x_j\|_2 \le \frac{2c_j\gamma}{1 - c_*\epsilon + \sqrt{(1 - c_*\epsilon)^2 - 4c_*^2\gamma\eta}} \equiv \xi_{j*}, \qquad j = 1, 2. \tag{3.3.34}$$

Proof. Define

$$\mathcal{S}_{\xi_{1*},\xi_{2*}} = \left\{x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;:\; x_j \in \mathcal{C}^{m_j},\ \|x_j\|_2 \le \xi_{j*}\ \ j = 1, 2\right\}. \tag{3.3.35}$$

We first prove that if there is a solution of (3.3.24) in $\mathcal{S}_{\xi_{1*},\xi_{2*}}$, then it is unique.

Assume that the equation (3.3.24) has different solutions $x_*, \hat{x} \in \mathcal{S}_{\xi_{1*},\xi_{2*}}$. Then by (3.3.20)–(3.3.22), (3.3.24), (3.3.32)–(3.3.34) we have

$$\begin{aligned}\|\hat{x} - x_*\|_2 &\leq c_* \left(\epsilon \|\hat{x} - x_*\|_2 + 2\eta \max\{\|\hat{x}\|_2, \|x_*\|_2\} \|\hat{x} - x_*\|_2\right) \\ &\leq c_* \left(\epsilon + \frac{4c_*\gamma\eta}{1 - c_*\epsilon + \sqrt{(1 - c_*\epsilon)^2 - 4c_*^2\gamma\eta}}\right) \|\hat{x} - x_*\|_2 \\ &< c_* \left(\epsilon + \frac{4c_*\gamma\eta}{1 - c_*\epsilon}\right) \|\hat{x} - x_*\|_2 \\ &= \left(c_*\epsilon + \frac{4c_*^2\gamma\eta}{(1 - c_*\epsilon)^2}(1 - c_*\epsilon)\right) \|\hat{x} - x_*\|_2 \\ &< \|\hat{x} - x_*\|_2.\end{aligned}$$

This contradiction shows that there is at most one solution of the equation (3.3.24) in $\mathcal{S}_{\xi_{1*},\xi_{2*}}$.

We now prove the existence of a solution of (3.3.24) in $\mathcal{S}_{\xi_{1*},\xi_{2*}}$.

Let $\mathcal{M}$ be the mapping defined by (3.3.29). Then the problem of finding a solution of (3.3.24) satisfying (3.3.34) reduces to the problem of showing that there is a fixed point of the mapping $\mathcal{M}$ in $\mathcal{S}_{\xi_{1*},\xi_{2*}}$.

It can be verified that the scalars ξ_{1*} and ξ_{2*} defined by (3.3.34) satisfy the equations

$$\xi_j = c_j \left[\gamma + \epsilon\sqrt{\xi_1^2 + \xi_2^2} + \eta \left(\xi_1^2 + \xi_2^2\right)\right], \quad j = 1, 2. \tag{3.3.36}$$

From (3.3.29) we see that if $x \in \mathcal{S}_{\xi_{1*},\xi_{2*}}$ then y satisfies

$$\begin{aligned}\|y_j\|_2 &\leq c_j(\gamma + \epsilon\|x\|_2 + \eta\|x\|_2^2) \qquad \text{(by (3.3.20) – (3.3.22), (3.3.24), (3.3.32))} \\ &\leq c_j \left[\gamma + \epsilon\sqrt{\xi_{1*}^2 + \xi_{2*}^2} + \eta \left(\xi_{1*}^2 + \xi_{2*}^2\right)\right] \qquad \text{(by (3.3.34))} \\ &= \xi_{j*}, \quad j = 1, 2, \qquad \text{(by (3.3.36))}\end{aligned}$$

which means that for the mapping $\mathcal{M}$ defined by (3.3.29) we have

$$x \in \mathcal{S}_{\xi_{1*},\xi_{2*}} \implies y \in \mathcal{S}_{\xi_{1*},\xi_{2*}}.$$

By the same argument as above in the proof of Theorem 3.3.4, the mapping $\mathcal{M}$ has a fixed point in $\mathcal{S}_{\xi_{1*},\xi_{2*}}$, and thus the equation (3.3.24) has a solution in $\mathcal{S}_{\xi_{1*},\xi_{2*}}$. $\square$.

Notes and References

NR 3.3–1. The first perturbation bound for singular subspace pair was obtained by Stewart [91, Theorem 6.4]. Theorem 3.3.1 is proved by Sun [119, Theorem 2.5.1].

NR 3.3–2. Theorem 3.3.5 is proved by Sun [119, Theorem 1.3.1].

3.4 Backward Errors and Residual Bounds

3.4.1 Backward Errors

In this subsection we discuss several kinds of normwise backward errors which are defined by using some information of approximate singular subspaces and associated singular values of a matrix A.

3.4.1.1 The Backward Error $\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$

Let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ approximate an l-dimensional simple singular subspace pair of $A \in \mathcal{C}^{m\times n}$. By §1.9, we define the backward error $\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ of A with respect to $\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1$ by

$$\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \min_{E\in\mathcal{E}} \|E\|, \tag{3.4.1}$$

where the set $\mathcal{E}$ is defined by

$$\mathcal{E} = \left\{E \in \mathcal{C}^{m\times n} \ : \ (A+E)\tilde{\mathcal{X}}_1 \subset \tilde{\mathcal{Y}}_1, \ (A+E)^H\tilde{\mathcal{Y}}_1 \subset \tilde{\mathcal{X}}_1\right\}. \tag{3.4.2}$$

The following result gives a computable formula of $\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$.

Theorem 3.4.1. *Let $A \in \mathcal{C}^{m\times n}$. Let $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$ with $\tilde{V}_1 \in \mathcal{U}^{n\times l}$ and $\tilde{U}_1 \in \mathcal{U}^{m\times l}$, and let*

$$R = \tilde{U}_1(\tilde{U}_1^H A\tilde{V}_1) - A\tilde{V}_1, \quad S = \tilde{V}_1(\tilde{V}_1^H A^H\tilde{U}_1) - A^H\tilde{U}_1 \tag{3.4.3}$$

be the residuals of A and A^H with respect to $\tilde{V}_1, \tilde{U}_1$, respectively. Then the backward error $\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ can be expressed by

$$\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \left\| \begin{pmatrix} 0 & S^H \\ R & 0 \end{pmatrix} \right\|. \tag{3.4.4}$$

The expressions (3.4.3) and (3.4.4) imply that the backward error $\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ defined by (3.4.1) is independent of the choice of the matrices $\tilde{V}_1$ and $\tilde{U}_1$ whose column vectors form orthonormal bases of $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{Y}}_1$, respectively.

Proof of Theorem 3.4.1. From (3.4.2) it follows that a matrix $E \in \mathcal{E}$ if and only if E is a solution to the equations

$$(A+E)\tilde{V}_1 = \tilde{U}_1 A_1, \quad (A+E)^H \tilde{U}_1 = \tilde{V}_1 A_1^H$$

for some $A_1 \in \mathcal{C}^{l\times l}$; or equivalently, E satisfies

$$E\tilde{V}_1 = \tilde{U}_1 A_1 - A\tilde{V}_1, \quad E^H \tilde{U}_1 = \tilde{V}_1 A_1^H - A^H \tilde{U}_1. \tag{3.4.5}$$

Applying Theorem 1.5.1 to the first equation of (3.4.5) we see that the equation is solvable, and any solution E of the equation can be expressed by

$$E = (\tilde{U}_1 A_1 - A\tilde{V}_1)\tilde{V}_1^H + Z(I - \tilde{V}_1\tilde{V}_1^H), \qquad Z \in \mathcal{C}^{m\times n}. \tag{3.4.6}$$

Choose $\tilde{V}_2$ so that $\tilde{V} = (\tilde{V}_1, \tilde{V}_2) \in \mathcal{U}^{n\times n}$. Then (3.4.6) can be written

$$E = (\tilde{U}_1 A_1 - A\tilde{V}_1)\tilde{V}_1^H + Z\tilde{V}_2\tilde{V}_2^H, \qquad Z \in \mathcal{C}^{m\times n}. \tag{3.4.7}$$

Combining it with the second equation of (3.4.5) shows that the matrix Z of (3.4.7) satisfies

$$\tilde{U}_1^H Z\tilde{V}_2 = -\tilde{U}_1^H A\tilde{V}_2. \tag{3.4.8}$$

By Theorem 1.5.1, the equation (3.4.8) is solvable, and any solution Z can be expressed by

$$Z = -\tilde{U}_1\tilde{U}_1^H A\tilde{V}_2\tilde{V}_2^H + W - \tilde{U}_1\tilde{U}_1^H W\tilde{V}_2\tilde{V}_2^H, \qquad W \in \mathcal{C}^{m\times n}. \tag{3.4.9}$$

Choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{m\times m}$. Then from (3.4.9)

$$Z\tilde{V}_2 = -\tilde{U}_1\tilde{U}_1^H A\tilde{V}_2 + \tilde{U}_2\tilde{U}_2^H W\tilde{V}_2.$$

Substituting it into (3.4.7) gives

$$E = (\tilde{U}_1 A_1 - A\tilde{V}_1)\tilde{V}_1^H - \tilde{U}_1\tilde{U}_1^H A\tilde{V}_2\tilde{V}_2^H + \tilde{U}_2\tilde{U}_2^H W\tilde{V}_2\tilde{V}_2^H,$$

and

$$\tilde{U}^H E\tilde{V} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{V}_1 & -\tilde{U}_1^H A\tilde{V}_2 \\ -\tilde{U}_2^H A\tilde{V}_1 & \tilde{U}_2^H W\tilde{V}_2 \end{pmatrix} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{V}_1 & S^H\tilde{V}_2 \\ \tilde{U}_2^H R & \tilde{U}_2 W\tilde{V}_2 \end{pmatrix}.$$

Consequently, By Theorem 1.2.1 and the definition (3.4.1) we have

$$\eta(\tilde{\mathcal{X}}_1, \tilde{Y}_1) = \|E_{\rm opt}\| \quad \text{with} \quad E_{\rm opt} = \tilde{U}\begin{pmatrix} 0 & S^H\tilde{V}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix}\tilde{V}^H. \tag{3.4.10}$$

Observe that the relations

$$\tilde{U}^H R = \begin{pmatrix} 0 \\ \tilde{U}_2^H R \end{pmatrix}, \qquad S^H\tilde{V} = (0,\ S^H\tilde{V}_2)$$

imply

$$\sigma_+(\tilde{U}_2^H R) = \sigma_+(R), \quad \sigma_+(S^H \tilde{V}_2) = \sigma_+(S^H).$$

Hence, we have

$$\sigma_+ \begin{pmatrix} 0 & S^H \tilde{V}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix} = \sigma_+ \begin{pmatrix} 0 & S^H \\ R & 0 \end{pmatrix}.$$

Combining it with (3.4.10) shows (3.4.4). □

3.4.1.2 The Backward Errors $\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ and $\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$

Let $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_l$ $(l \le n)$ be approximate singular values of $A \in \mathcal{C}^{m\times n}$, and $\tilde{x}_1, \ldots, \tilde{x}_l$ and $\tilde{y}_1, \ldots, \tilde{y}_l$ be associated right and left singular vectors, respectively. Generally speaking, the vectors are linearly independent but not necessarily orthonormal. This subsection is devoted to some backward errors of A with respect to $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_l$, $\tilde{x}_1, \ldots, \tilde{x}_l$, and $\tilde{y}_1, \ldots, \tilde{y}_l$.

Let

$$\tilde{\Sigma}_1 = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_l), \quad \tilde{X}_1 = (\tilde{x}_1, \ldots, \tilde{x}_l), \quad \tilde{Y}_1 = (\tilde{y}_1, \ldots, \tilde{y}_l). \tag{3.4.11}$$

Take orthogonal decompositions of $\tilde{X}_1$ and $\tilde{Y}_1$:

$$\tilde{X}_1 = \tilde{V}_1 F_1, \quad \tilde{Y}_1 = \tilde{U}_1 G_1, \tag{3.4.12}$$

where $\tilde{V}_1 \in \mathcal{U}^{n\times l}, \tilde{U}_1 \in \mathcal{U}^{m\times l}$, and $F_1, G_1 \in \mathcal{C}^{l\times l}$. Then by §1.9, we define the backward errors $\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ and $\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ by

$$\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1) = \min_{E\in\mathcal{E}} \|E\|_F, \quad \eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1) = \min_{E\in\mathcal{E}} \|E\|_2, \tag{3.4.13}$$

where the set $\mathcal{E}$ is defined by

$$\mathcal{E} = \left\{ E \in \mathcal{C}^{m\times n} \ : \ (A+E)\tilde{V}_1 = \tilde{U}_1\tilde{\Sigma}_1, \ (A+E)^H \tilde{U}_1 = \tilde{V}_1 \tilde{\Sigma}_1 \right\}. \tag{3.4.14}$$

Computable formulas of $\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ and $\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ are given by the following result.

Theorem 3.4.2. *Let*

$$R = \tilde{U}_1\tilde{\Sigma}_1 - A\tilde{V}_1, \quad S = \tilde{V}_1\tilde{\Sigma}_1 - A^H\tilde{U}_1 \tag{3.4.15}$$

be the residuals of A and A^H with respect to $\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1$, respectively, where $\tilde{V}_1, \tilde{U}_1$ and $\tilde{\Sigma}_1$ are defined by (3.4.12) and (3.4.11). Then the backward errors $\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ and $\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ can be expressed by

$$\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1) = \sqrt{\|R\|_F^2 + \|S\|_F^2 - \|\tilde{V}_1^H S\|_F^2}, \tag{3.4.16}$$

and

$$\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1) = \max\{\|R\|_2,\ \|S\|_2\}. \tag{3.4.17}$$

Proof. From (3.4.14) and (3.4.15) it follows that a matrix $E \in \mathcal{E}$ if and only if E is a solution of the equations

$$E\tilde{V}_1 = R, \qquad E^H\tilde{U}_1 = S. \tag{3.4.18}$$

Choose $\tilde{V}_2$ and $\tilde{U}_2$ so that $\tilde{V} = (\tilde{V}_1, \tilde{V}_2) \in \mathcal{U}^{n\times n}$ and $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{m\times m}$. Then by the same argument as in the proof of Theorem 3.4.1 we can show that any solution E of the equations (3.4.18) can be expressed by

$$E = \tilde{U}\begin{pmatrix} \tilde{U}_1^H R & S^H\tilde{V}_2 \\ \tilde{U}_2^H R & \tilde{U}_2^H W\tilde{V}_2 \end{pmatrix}\tilde{V}^H, \qquad W \in \mathcal{C}^{m\times n}. \tag{3.4.19}$$

Thus, applying Theorem 1.2.1 we obtain

$$\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1) = \|E_{\rm opt}\|_F \quad \text{with} \quad E_{\rm opt} = \tilde{U}\begin{pmatrix} \tilde{U}_1^H R & S^H\tilde{V}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix}\tilde{V}^H, \tag{3.4.20}$$

which gives the expression (3.4.16).

Moreover, by (3.4.19) and Theorem 1.2.4 we have

$$\begin{aligned}\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1) &= \max\{\|R\|_2,\ \|(\tilde{U}_1^H R,\ S^H\tilde{V}_2)\|_2\} \\ &= \max\{\|R\|_2,\ \|S^H(\tilde{V}_1,\ \tilde{V}_2)\|_2\},\end{aligned}$$

which gives (3.4.17). □

Remark 3.4.3. Let $\tilde{\sigma}_1$ be an approximate singular value of $A \in \mathcal{C}^{m\times n}$, and $\tilde{v}_1 \in \mathcal{C}^n$ and $\tilde{u}_1 \in \mathcal{C}^m$ be associated unit right and left singular vectors. Then by Theorem 3.4.2, the backward errors $\eta_F(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1)$ and $\eta_2(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1)$ of A with respect to $\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1$ can be expressed by

$$\eta_F(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1) = \sqrt{\|r\|_2^2 + \|s\|_2^2 - |\tilde{v}_1^H s|^2}$$

and

$$\eta_2(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1) = \max\{\|r\|_2,\ \|s\|_2\},$$

where r and s are the residuals defined by

$$r = \tilde{\sigma}_1\tilde{u}_1 - A\tilde{v}_1, \qquad s = \tilde{\sigma}_1\tilde{v}_1 - A^H\tilde{u}_1.$$

Moreover, by (3.4.20), the matrix

$$E_{\rm opt} = \tilde{U}\begin{pmatrix} \tilde{u}_1^H r & s^H\tilde{V}_2 \\ \tilde{U}_2^H r & 0 \end{pmatrix}\tilde{V}^H = r\tilde{v}_1^H + \tilde{u}_1 s^H - s^H\tilde{v}_1\tilde{u}_1\tilde{v}_1^H$$

is the smallest perturbation of A such that $\tilde{\sigma}_1$ is a singular value of $A+E_{\rm opt}$, and $\tilde{v}_1, \tilde{u}_1$ are associated unit right and unit left singular vectors. Note that the formula of $\eta_2(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1)$ is generalized to the generalized singular value decomposition by Sun [122].

From (3.4.15)–(3.4.17) we see that $\eta_F(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ and $\eta_2(\tilde{V}_1, \tilde{U}_1, \tilde{\Sigma}_1)$ are dependent on the orthogonal decompositions (3.4.12) of $\tilde{X}_1$ and $\tilde{Y}_1$. In view of a best approximation property possessed by the unitary polar factor, we take the polar decompositions of $\tilde{X}_1$ and $\tilde{Y}_1$:

$$\tilde{X}_1 = P_1 H_1, \quad \tilde{Y}_1 = \Pi_1 K_1, \tag{3.4.21}$$

where $P_1 \in \mathcal{U}^{n\times l}, \Pi_1 \in \mathcal{U}^{m\times l}$, and $H_1, K_1 \in \mathcal{H}^{l\times l}$ are positive definite. By (3.4.16) and (3.4.15), we have

$$\eta_F(P_1, \Pi_1, \tilde{\Sigma}_1) \leq \sqrt{\|\Pi_1\tilde{\Sigma}_1 - AP_1\|_F^2 + \|P_1\tilde{\Sigma}_1 - A^H\Pi_1\|_F^2}. \tag{3.4.22}$$

The following result presents an upper bound for $\eta_F(P_1, \Pi_1, \tilde{\Sigma}_1)$ by using $\tilde{\Sigma}_1$ and the residuals $\tilde{Y}_1\tilde{\Sigma}_1 - A\tilde{X}_1$ and $\tilde{X}_1\tilde{\Sigma}_1 - A^H\tilde{Y}_1$.

Theorem 3.4.4. *Let $\tilde{\Sigma}_1, \tilde{X}_1, \tilde{Y}_1$ be the matrices of (3.4.11). Define $\delta, \epsilon_1, \epsilon_2$ by*

$$\delta = \max\{|\sigma_{\max}(\tilde{X}_1) - \sigma_{\min}(\tilde{Y}_1)|, \ |\sigma_{\max}(\tilde{Y}_1) - \sigma_{\min}(\tilde{X}_1)|\}, \tag{3.4.23}$$

and

$$\begin{aligned} \epsilon_1 &= \frac{\|\tilde{Y}_1\tilde{\Sigma}_1 - A\tilde{X}_1\|_F + \delta\|AP_{\tilde{X}_1}\|_F}{\sigma_{\min}(\tilde{Y}_1)}, \\ \epsilon_2 &= \frac{\|\tilde{X}_1\tilde{\Sigma}_1 - A^H\tilde{Y}_1\|_F + \delta\|A^HP_{\tilde{Y}_1}\|_F}{\sigma_{\min}(\tilde{X}_1)}. \end{aligned} \tag{3.4.24}$$

Then for the backward error $\eta_F(P_1, \Pi_1, \tilde{\Sigma}_1)$ defined by (3.4.13) and (3.4.14) with $\tilde{V}_1 = P_1$ and $\tilde{U}_1 = \Pi_1$, we have the estimate

$$\eta_F(P_1, \Pi_1, \tilde{\Sigma}_1) \leq \sqrt{\epsilon_1^2 + \epsilon_2^2}. \tag{3.4.25}$$

The estimate (3.4.25) shows that if the residuals $\tilde{Y}_1\tilde{\Sigma}_1 - A\tilde{X}_1$ and $\tilde{X}_1\tilde{\Sigma}_1 - A^H\tilde{Y}_1$ are small, and if $\tilde{x}_1, \ldots, \tilde{x}_l$ and $\tilde{y}_1, \ldots, \tilde{y}_l$ are close to orthonormal, then there is a matrix $A + E_{\rm opt}$ with small $E_{\rm opt}$ of (3.4.20) (in which $\tilde{V}_1 = P_1$ and $\tilde{U}_1 = \Pi_1$) such that $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_l$ are exact singular values of $A + E_{\rm opt}$, and the column vectors of P_1 (the unitary polar factor of $\tilde{X}_1$) and Π_1 (the unitary polar factor of $\tilde{Y}_1$) are associated unit right and left singular vectors.

Proof of Theorem 3.4.4. Take the singular value decompositions of $\tilde{X}_1$ and $\tilde{Y}_1$:

$$\tilde{X}_1 = U\begin{pmatrix} M_1 \\ 0 \end{pmatrix}V^H, \quad \tilde{Y}_1 = Z\begin{pmatrix} N_1 \\ 0 \end{pmatrix}W^H, \tag{3.4.26}$$

where $U = (U_1, U_2) \in \mathcal{U}^{n\times n}$ with $U_1 \in \mathcal{U}^{n\times l}$, $Z = (Z_1, Z_2) \in \mathcal{U}^{m\times m}$ with $Z_1 \in \mathcal{U}^{m\times l}$, $V, W \in \mathcal{U}^{l\times l}$, and

$$M_1 = \mathrm{diag}(\mu_j), \quad \mu_1 \geq \cdots \geq \mu_l > 0, \quad N_1 = \mathrm{diag}(\nu_j), \quad \nu_1 \geq \cdots \geq \nu_l > 0.$$

Then by the uniqueness of the polar decomposition, we have

$$P_1 = U_1 V^H, \qquad \Pi_1 = Z_1 W^H. \tag{3.4.27}$$

By (3.4.26), we have

$$\|A\tilde{X}_1 - \tilde{Y}_1\tilde{\Sigma}_1\|_F = \left\| Z^H AU \begin{pmatrix} M_1 \\ 0 \end{pmatrix} - \begin{pmatrix} N_1 \\ 0 \end{pmatrix} W^H \tilde{\Sigma}_1 V \right\|_F. \tag{3.4.28}$$

Let

$$B = Z^H AU = (B_1, B_2), \quad B_1 = \begin{pmatrix} B_{11} \\ B_{21} \end{pmatrix}, \quad B_{11} \in \mathcal{C}^{l\times l}, \quad C = W^H \tilde{\Sigma}_1 V. \tag{3.4.29}$$

Then from (3.4.28)

$$\begin{aligned} \|A\tilde{X}_1 - \tilde{Y}_1\tilde{\Sigma}_1\|_F &= \left\| B_1 M_1 - \begin{pmatrix} N_1 C \\ 0 \end{pmatrix} \right\|_F \\ &= \left\| \left(B_1 M_1 - \begin{pmatrix} N_1 & 0 \\ 0 & \nu_l I \end{pmatrix} B_1 \right) + \begin{pmatrix} N_1(B_{11} - C) \\ \nu_l B_{21} \end{pmatrix} \right\|_F \\ &\geq \left\| \begin{pmatrix} N_1(B_{11} - C) \\ \nu_l B_{21} \end{pmatrix} \right\|_F - \left\| B_1 M_1 - \begin{pmatrix} N_1 & 0 \\ 0 & \nu_l I \end{pmatrix} B_1 \right\|_F \\ &\geq \nu_l \left\| B_1 - \begin{pmatrix} C \\ 0 \end{pmatrix} \right\|_F - \max_{j,k} |\mu_j - \nu_k| \|B_1\|_F. \end{aligned} \tag{3.4.30}$$

Observe that

$$\begin{aligned} \left\| B_1 - \begin{pmatrix} C \\ 0 \end{pmatrix} \right\|_F &= \left\| B \begin{pmatrix} I^{(l)} \\ 0 \end{pmatrix} - \begin{pmatrix} I^{(l)} \\ 0 \end{pmatrix} C \right\|_F \\ &= \left\| Z^H AU \begin{pmatrix} I^{(l)} \\ 0 \end{pmatrix} - \begin{pmatrix} I^{(l)} \\ 0 \end{pmatrix} W^H \tilde{\Sigma}_1 V \right\|_F \quad \text{(by (3.4.29))} \\ &= \|AP_1 - \Pi_1 \tilde{\Sigma}_1\|_F, \quad \text{(by (3.4.27))} \end{aligned}$$

$$\max_{j,k} |\mu_j - \nu_k| = \delta, \qquad \text{(by (3.4.23))}$$

and

$$\|B_1\|_F = \|Z^H A U_1\|_F = \|A U_1 V^H\|_F = \|A P_1\|_F = \|A P_1 P_1^H\|_F = \|A P_{\tilde{X}_1}\|_F.$$

Hence, (3.4.30) implies

$$\|\Pi_1 \tilde{\Sigma}_1 - A P_1\|_F \le \epsilon_1. \tag{3.4.31}$$

Similarly, we have

$$\|P_1 \tilde{\Sigma}_1 - A^H \Pi_1\|_F \le \epsilon_2. \tag{3.4.32}$$

The scalars ϵ_1 and ϵ_2 are defined by (3.4.24).

Combining (3.4.31) and (3.4.32) with (3.4.22) yields (3.4.25). □

3.4.2 Residual Bounds

Let an approximate simple singular subspace pair $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ of A be given, where $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$ with $\tilde{V}_1 \in \mathcal{U}^{n\times l}$ and $\tilde{U}_1 \in \mathcal{U}^{m\times l}$. Then by using Theorem 3.4.1 and appropriate forward perturbation results, such as the Mirsky theorem on perturbations of singular values [81] (see below NR 3.4–3) and Theorem 3.3.3 on perturbation bounds for singular subspaces, we can determine how the singular values $\tilde{\sigma}_1, \dots, \tilde{\sigma}_l$ of $\tilde{U}_1^H A \tilde{V}_1$ relate to those of A, and determine the accuracy of the approximate singular subspaces $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{Y}}_1$.

By the proof of Theorem 3.4.1, the optimal backward perturbation E_{opt} of (3.4.10) satisfies

$$(A + E_{\text{opt}})\tilde{V}_1 = \tilde{U}_1(\tilde{U}_1^H A \tilde{V}_1), \quad (A + E_{\text{opt}})^H \tilde{U}_1 = \tilde{V}_1(\tilde{U}_1^H A \tilde{V}_1)^H.$$

These relations imply that the singular values of $\tilde{U}_1^H A \tilde{V}_1$, as l approximate singular values of A, are l singular values of $A + E_{\text{opt}}$. Combining this fact with the Mirsky theorem (see below NR 3.4–3) shows the following result which gives a residual bound for the l approximate singular values $\tilde{\sigma}_1, \dots, \tilde{\sigma}_l$ of A.

Theorem 3.4.5. *Let $A, \tilde{V}_1, \tilde{U}_1, R, S$ be as in Theorem 3.4.1, and let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ with $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$ be an approximate singular subspace pair of A. If $\sigma_1 \ge \cdots \ge \sigma_n$ are the singular values of A, and $\tilde{\sigma}_1 \ge \cdots \ge \tilde{\sigma}_l$ are the singular values of $\tilde{U}_1^H A \tilde{V}_1$, then there are integers $j_1 < j_2 < \cdots < j_l$ such that*

$$\|\text{diag}\,(\tilde{\sigma}_1 - \sigma_{j_1}, \dots, \tilde{\sigma}_l - \sigma_{j_l})\| \le \left\| \begin{pmatrix} 0 & S^H \\ R & 0 \end{pmatrix} \right\|. \tag{3.4.33}$$

From (3.4.3) we see that the optimal backward perturbation E_{opt} of (3.4.10) satisfies

$$\tilde{U}^H (A + E_{\text{opt}}) \tilde{V} = \begin{pmatrix} \tilde{U}_1^H A \tilde{V}_1 & 0 \\ 0 & \tilde{U}_2^H A \tilde{V}_2 \end{pmatrix} \equiv \begin{pmatrix} \tilde{A}_1 & 0 \\ 0 & \tilde{A}_2 \end{pmatrix}, \tag{3.4.34}$$

and

$$\tilde{U}^H E_{\rm opt} \tilde{V} = \begin{pmatrix} 0 & S^H \tilde{V}_2 \\ \tilde{U}_2^H R & 0 \end{pmatrix}. \tag{3.4.35}$$

The relation (3.4.34) implies that $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ is a singular subspace pair of $A + E_{\rm opt}$. Moreover, if $\tilde{A}_1$ and the matrix $\widehat{\tilde{A}}_2$ defined by

$$\widehat{\tilde{A}}_2 = \begin{cases} \tilde{A}_2 & \text{if } m = n \\ \\ (\tilde{A}_2, 0) \in \mathcal{C}^{(m-l)\times(m-l)} & \text{if } m > n \end{cases}$$

satisfy

$$\sigma(\tilde{A}_1) \bigcap \sigma(\widehat{\tilde{A}}_2) = \emptyset, \tag{3.4.36}$$

then the singular subspace pair $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ of $A + E_{\rm opt}$ is a simple singular subspace pair.

Applying Theorem 3.3.3 to the matrices $A + E_{\rm opt}$ and A of (3.4.34) and (3.4.35) shows the following result which gives residual bounds for the approximate singular subspaces $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{Y}}_1$.

Theorem 3.4.6. *Let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ be an l-dimensional approximate singular subspace pair of $A \in \mathcal{C}^{m\times n}$, where $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$, $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$, $\tilde{V}_1 \in \mathcal{U}^{n\times l}$ and $\tilde{U}_1 \in \mathcal{U}^{m\times l}$. Define the matrices $\tilde{A}_1$ and $\tilde{A}_2$ by (3.4.34), and assume (3.4.36) is satisfied. Define the residuals R and S by*

$$R = \tilde{U}_1 \tilde{A}_1 - A\tilde{V}_1, \qquad S = \tilde{V}_1 \tilde{A}_1^H - A^H \tilde{U}_1,$$

and define the matrices $\tilde{C}_1$, $\tilde{C}_2$ by

$$\begin{aligned} \tilde{C}_1 &= \left((\tilde{A}_1^T \otimes I_{n-l})\tilde{K}^{-1},\ (I_l \otimes \tilde{A}_2^H)\tilde{L}^{-1} \right), \\ \tilde{C}_2 &= \left((I_l \otimes \tilde{A}_2)\tilde{K}^{-1},\ (\overline{\tilde{A}}_1 \otimes I_{m-l})\tilde{L}^{-1} \right), \end{aligned} \tag{3.4.37}$$

where

$$\tilde{K} = \overline{\tilde{A}}_1 \tilde{A}_1^T \otimes I_{n-l} - I_l \otimes \tilde{A}_2^H \tilde{A}_2, \qquad \tilde{L} = \tilde{A}_1^T \overline{\tilde{A}}_1 \otimes I_{m-l} - I_l \otimes \tilde{A}_2 \tilde{A}_2^H.$$

Moreover, let

$$\begin{aligned} \tilde{b}_1 &= \left\| \tilde{C}_1 \begin{pmatrix} {\rm vec}(\tilde{V}_2^H S) \\ {\rm vec}(\tilde{U}_2^H R) \end{pmatrix} \right\|_2, \qquad \tilde{c}_1 = \|\tilde{C}_1\|_2, \\ \tilde{b}_2 &= \left\| \tilde{C}_2 \begin{pmatrix} {\rm vec}(\tilde{V}_2^H S) \\ {\rm vec}(\tilde{U}_2^H R) \end{pmatrix} \right\|_2, \qquad \tilde{c}_2 = \|\tilde{C}_2\|_2, \\ \tilde{b} &= \tilde{b}_1 + \tilde{b}_2, \qquad \tilde{c} = \tilde{c}_1 + \tilde{c}_2, \end{aligned} \tag{3.4.38}$$

and define $\tilde{\eta}$ *by*

$$\tilde{\eta} = \max\{\|R\|_2,\ \|S\|_2\}. \tag{3.4.39}$$

If

$$4\tilde{b}\tilde{c}\tilde{\eta} < 1,$$

then there is a unique pair of singular subspaces $\mathcal{X}_1 = \mathcal{R}(V_1)$, $\mathcal{Y}_1 = \mathcal{R}(U_1)$ *of* A *such that* $V_1 \in \mathcal{U}^{n\times l}$, $U_1 \in \mathcal{U}^{m\times l}$, *and*

$$\begin{aligned} \rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) &\le \|\tan\Theta(V_1, \tilde{V}_1)\|_F \le \tilde{b}_1 + \tilde{c}_1\tilde{\eta}\tilde{\beta}^2 \equiv \tau_{\mathcal{X}_1}, \\ \rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) &\le \|\tan\Theta(U_1, \tilde{U}_1)\|_F \le \tilde{b}_2 + \tilde{c}_2\tilde{\eta}\tilde{\beta}^2 \equiv \tau_{\mathcal{Y}_1}, \end{aligned} \tag{3.4.40}$$

where

$$\tilde{\beta} \equiv \frac{2\tilde{b}}{1 + \sqrt{1 - 4\tilde{b}\tilde{c}\tilde{\eta}}}. \tag{3.4.41}$$

It is worth noting that by using Theorem 3.4.6 and Theorem 3.4.9 of the next subsection (§3.4.3), we obtain the following result on residual bounds for singular values which may be sharper than the estimate (3.4.33).

Theorem 3.4.7. *Let* $A, \mathcal{X}_1, V_1, \mathcal{Y}_1, U_1,\ \tilde{\mathcal{X}}_1, \tilde{V}_1,\ \tilde{\mathcal{Y}}_1, \tilde{U}_1,\ R, S,\ \tau_{\mathcal{X}_1}$ *and* $\tau_{\mathcal{Y}_1}$ *be as in Theorem 3.4.6. If the singular values of* A *are* $\sigma_1 \ge \cdots \ge \sigma_n$, *the singular values of* $\tilde{U}_1^H A\tilde{V}_1$ *are* $\tilde{\sigma}_1 \ge \cdots \ge \tilde{\sigma}_l$, *and if the scalars* $\tau_{\mathcal{X}_1}$ *and* $\tau_{\mathcal{Y}_1}$ *satisfy*

$$\tau_{\mathcal{X}_1} < 1 \quad \text{and} \quad \tau_{\mathcal{Y}_1} < 1,$$

then there are integers $j_1 < j_2 < \cdots < j_l$ *such that*

$$|\tilde{\sigma}_i - \sigma_{j_i}| \le \frac{\max\{\tau_{\mathcal{X}_1}\|S\|_2,\ \tau_{\mathcal{Y}_1}\|R\|_2\}}{\min\left\{\sqrt{1-\tau_{\mathcal{X}_1}^2},\ \sqrt{1-\tau_{\mathcal{Y}_1}^2}\right\}}, \qquad i = 1, \ldots, l. \tag{3.4.42}$$

Proof. By Theorem 3.4.9 of the next subsection (§3.4.3) there are integers $j_1 < j_2 < \cdots < j_l$ such that

$$|\tilde{\sigma}_i - \sigma_{j_i}| \le \frac{\max\left\{\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1)\|S\|_2,\ \rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1)\|R\|_2\right\}}{\min\left\{\sqrt{1-\rho_F^2(\mathcal{X}_1, \tilde{\mathcal{X}}_1)},\ \sqrt{1-\rho_F^2(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1)}\right\}}, \qquad i = 1, \ldots, l, \tag{3.4.43}$$

where $\rho_F(\cdot,\cdot)$ is the generalized chordal metric defined by (1.3.3). Substituting (3.4.40) into (3.4.43) shows (3.4.42). □

Example 3.4.8. Consider the matrix

$$A = \begin{pmatrix} 5/\sqrt{6} & 10^3/\sqrt{3} \\ -10/\sqrt{6} & 10^3/\sqrt{3} \\ 5/\sqrt{6} & 10^3/\sqrt{3} \end{pmatrix}.$$

The vectors

$$v_1 = (1,\ 0)^T, \qquad u_1 = \left(\frac{1}{\sqrt{6}},\ -\frac{2}{\sqrt{6}},\ \frac{1}{\sqrt{6}}\right)^T$$

are unit right and left singular vectors of A belonging to the singular value $\sigma_1 = 5$. (Note. The other singular value is $\sigma_2 = 10^3$.) Suppose that we have approximate right and left singular vectors

$$\tilde{x}_1 = (1,\ 10^{-6})^T, \qquad \tilde{y}_1 = (0.40825,\ -0.81496,\ 0.40824)^T,$$

and let

$$\tilde{v}_1 = \tilde{x}_1/\|\tilde{x}_1\|_2, \qquad \tilde{u}_1 = \tilde{y}_1/\|\tilde{y}_1\|_2, \qquad \tilde{\sigma}_1 = \tilde{u}_1^T A\tilde{v}_1.$$

A calculation gives

$$\sin\theta(v_1,\tilde{v}_1) \approx 1.0000 \times 10^{-6}, \qquad \sin\theta(u_1,\tilde{u}_1) \approx 8.8449 \times 10^{-4}, \tag{3.4.44}$$

and

$$|\tilde{\sigma}_1 - \sigma_1| \approx 1.0713 \times 10^{-6}. \tag{3.4.45}$$

Choose $\tilde{v}_2$ and $\tilde{U}_2$ so that $(\tilde{v}_1, \tilde{v}_2) \in \mathcal{O}^{2\times 2}$ and $(\tilde{u}_1, \tilde{U}_2) \in \mathcal{O}^{3\times 3}$. Compute

$$\tilde{A}_1 = \tilde{u}_1^T A\tilde{v}_1\ (= \tilde{\sigma}_1), \quad \tilde{A}_2 = \tilde{U}_2^T A\tilde{v}_2, \quad r = \tilde{A}_1\tilde{u}_1 - A\tilde{v}_1, \quad s = \tilde{A}_1\tilde{v}_1 - A^T\tilde{u}_1,$$

and compute $\tilde{C}_1, \tilde{C}_2,\ \tilde{b}_1, \tilde{c}_1, \tilde{b}_2, \tilde{c}_2,\ \tilde{b}, \tilde{c}$ and $\tilde{\eta}$ by (3.4.37)–(3.4.39). A calculation shows that

$$4\tilde{b}\tilde{c}\tilde{\eta} \approx 6.2967 \times 10^{-4} < 1.$$

Consequently, applying Theorem 3.4.6, there are unit right and left singular vectors v and u of A corresponding to the same singular value, such that

$$\tan\theta(v,\tilde{v}_1) \le \tau_{\mathcal{X}_1} \approx 1.0007 \times 10^{-6}, \quad \tan\theta(u,\tilde{u}_1) \le \tau_{\mathcal{Y}_1} \approx 8.8463 \times 10^{-4}. \tag{3.4.46}$$

Moreover, applying Theorem 3.4.7, there is a singular value σ of A such that

$$|\tilde{\sigma}_1 - \sigma| \le \frac{\max\left\{\tau_{\mathcal{X}_1}\|s\|_2,\ \tau_{\mathcal{Y}_1}\|r\|_2\right\}}{\min\left\{\sqrt{1-\tau_{\mathcal{X}_1}^2},\ \sqrt{1-\tau_{\mathcal{Y}_1}^2}\right\}} \approx 3.0276 \times 10^{-6}. \tag{3.4.47}$$

Comparing (3.4.46) with (3.4.44) and comparing (3.4.47) with (3.4.45) we see that the estimates obtained by applying Theorems 3.4.6 and 3.4.7 are fairly sharp.

Applying Theorem 3.4.5, there is a singular value σ of A such that

$$|\tilde{\sigma}_1 - \sigma| \le \max\{\|r\|_2,\ \|s\|_2\} \approx 8.8445 \times 10^{-1}. \tag{3.4.48}$$

Comparing it with (3.4.45) shows that the estimate obtained by applying Theorem 3.4.5 is a severe overestimate.

Note that by Theorem 3.4.2 (or Remark 3.4.3) we have

$$\eta_2(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1) = \max\{\|r\|_2,\ \|s\|_2\} \approx 8.8445 \times 10^{-1},$$

which means that $\tilde{\sigma}_1$ is an exact singular value and $\{\tilde{v}_1, \tilde{u}_1\}$ is an associated unit singular vector pair of a perturbed matrix $A + E_*$ with

$$\|E_*\|_2 = \eta_2(\tilde{v}_1, \tilde{u}_1, \tilde{\sigma}_1) \approx 8.8445 \times 10^{-1}.$$

Combining this fact with the Mirsky theorem (see below NR 3.4–3) we get the same estimate as that of (3.4.48).

3.4.3 An Approximation Theorem on Singular Values

In this subsection we shall prove an approximation theorem on singular values that can be used to establish the estimate (3.4.42) of Theorem 3.4.7.

Let $\{\mathcal{X}_1, \mathcal{Y}_1\}$ with $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ be a singular subspace pair of $A \in \mathcal{C}^{m\times n}$, and let $\{\mathcal{Z}_1, \mathcal{W}_1\}$ with $\mathcal{Z}_1 = \mathcal{R}(Z_1)$ and $\mathcal{W}_1 = \mathcal{R}(W_1)$ approximate $\{\mathcal{X}_1, \mathcal{Y}_1\}$, where $X_1, Z_1 \in \mathcal{U}^{n\times l}$, and $Y_1, W_1 \in \mathcal{U}^{m\times l}$. Let

$$A_1 = Y_1^H A X_1, \qquad K_1 = W_1^H A Z_1.$$

It is easy to see that

$$Y_1 A_1 = A X_1, \qquad X_1 A_1^H = A^H Y_1, \qquad \sigma(A_1) \subset \sigma(A).$$

However, in general, $W_1 K_1 \neq A Z_1$ and/or $Z_1 K_1^H \neq A^H W_1$, and $\sigma(K_1) \not\subset \sigma(A)$. In such a case, we define the matrices R and S by

$$R = W_1 K_1 - A Z_1, \qquad S = Z_1 K_1^H - A^H W_1,$$

which are the residuals of A and A^H with respect to $\{Z_1, W_1\}$, respectively.

The following result gives an upper bound for the distance between the sets $\sigma(K_1)$ and $\sigma(A_1)$ in terms of $\|R\|_2, \|S\|_2, \rho_2(\mathcal{X}_1, \mathcal{Z}_1)$ and $\rho_2(\mathcal{Y}_1, \mathcal{W}_1)$, where $\rho_2(\cdot,\cdot)$ is the generalized chordal metric defined by (1.3.3).

Theorem 3.4.9. *Let $A, X_1, Y_1, Z_1, W_1, A_1, K_1$, R, S and $\mathcal{X}_1$, $\mathcal{Y}_1$, $\mathcal{Z}_1$, $\mathcal{W}_1$ be the above-mentioned matrices and subspaces. Let*

$$\sigma(A_1) = \{\alpha_j\}_{j=1}^l, \qquad \alpha_1 \geq \cdots \geq \alpha_l,$$

$$\sigma(K_1) = \{\kappa_j\}_{j=1}^l, \qquad \kappa_1 \geq \cdots \geq \kappa_l.$$

If $\rho_2(\mathcal{X}_1, \mathcal{Z}_1) < 1$ and $\rho_2(\mathcal{Y}_1, \mathcal{W}_1) < 1$, then

$$|\alpha_j - \kappa_j| \leq \frac{\max\{\rho_2(\mathcal{X}_1, \mathcal{Z}_1)\|S\|_2,\ \rho_2(\mathcal{Y}_1, \mathcal{W}_1)\|R\|_2\}}{\min\left\{\sqrt{1-\rho_2^2(\mathcal{X}_1, \mathcal{Z}_1)},\ \sqrt{1-\rho_2^2(\mathcal{Y}_1, \mathcal{W}_1)}\right\}}, \qquad j = 1, \ldots, l. \tag{3.4.49}$$

Proof. 1) By Stewart [93, Appendix] (or see Stewart and Sun [97, Chapter I, Theorem 5.2]), there are unitary matrices Q, U_1, V_1 such that

$$Q X_1 U_1 = \begin{pmatrix} I_l \\ 0 \end{pmatrix} \qquad \text{and} \qquad Q Z_1 V_1 = \begin{pmatrix} \Gamma \\ \Sigma \end{pmatrix},$$

where $\Gamma = \mathrm{diag}(\gamma_j)$ and $\Sigma = \mathrm{diag}(\sigma_j)$ have the expressions (2.5.58)–(2.5.60). Similarly, there are unitary matrices P, F_1, G_1 such that

$$P Y_1 F_1 = \begin{pmatrix} I_l \\ 0 \end{pmatrix} \qquad \text{and} \qquad P W_1 G_1 = \begin{pmatrix} M \\ N \end{pmatrix},$$

where $M = \text{diag}(\mu_j)$ and $N = \text{diag}(\nu_j)$ have similar expressions as Γ and Σ in (2.5.58)–(2.5.60). Without loss of generality we may assume that the matrices A, X_1, Y_1, Z_1, W_1 have the following reduced forms:

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \quad X_1 = Y_1 = \begin{pmatrix} I_l \\ 0 \end{pmatrix}, \quad Z_1 = \begin{pmatrix} \Gamma \\ \Sigma \end{pmatrix}, \quad W_1 = \begin{pmatrix} M \\ N \end{pmatrix}.$$

Thus, we have

$$\rho_2(\mathcal{X}_1, \mathcal{Z}_1) = \|\Sigma\|_2, \qquad \rho_2(\mathcal{Y}_1, \mathcal{W}_1) = \|N\|_2, \tag{3.4.50}$$

and

$$R = \begin{pmatrix} MK_1 - A_1\Gamma \\ NK_1 - A_2\Sigma \end{pmatrix}, \qquad S = \begin{pmatrix} \Gamma K_1^H - A_1^H M \\ \Sigma K_1^H - A_2^H N \end{pmatrix}. \tag{3.4.51}$$

2) For Γ define the diagonal matrix $\hat{\Gamma}$ by (2.5.64), and for M define the diagonal matrix $\hat{M}$ similarly. Then there are the relations

$$\Sigma\Gamma = \hat{\Gamma}\Sigma, \qquad NM = \hat{M}N. \tag{3.4.52}$$

Moreover, let

$$Z_2 = \begin{pmatrix} -\Sigma^T \\ \hat{\Gamma} \end{pmatrix}, \qquad Z = (Z_1,\ Z_2), \tag{3.4.53}$$

and

$$W_2 = \begin{pmatrix} -N^T \\ \hat{M} \end{pmatrix}, \qquad W = (W_1,\ W_2). \tag{3.4.54}$$

Then the relations (3.4.52)–(3.4.54) imply that $Y, W \in \mathcal{U}^{m\times m}$, and from (3.4.51), (3.4.53), (3.4.54) and $K_1 = W_1^H A Z_1$

$$W^H R = \begin{pmatrix} 0 \\ B \end{pmatrix}, \qquad Z^H S = \begin{pmatrix} 0 \\ C \end{pmatrix}. \tag{3.4.55}$$

Thus,

$$\|R\| = \|B\|, \qquad \|S\| = \|C\|. \tag{3.4.56}$$

From (3.4.51) and (3.4.53)–(3.4.55) we get

$$\begin{aligned} MK_1 - A_1\Gamma &= (I_l,\ 0)R = (I_l,\ 0)W \begin{pmatrix} 0 \\ B \end{pmatrix} \\ &= (I_l,\ 0)W_2 B = -N^T B, \end{aligned}$$

and

$$\begin{aligned} \Gamma K_1^H - A_1^H M &= (I_l,\ 0)S = (I_l,\ 0)Z \begin{pmatrix} 0 \\ C \end{pmatrix} \\ &= (I_l,\ 0)Z_2 C = -\Sigma^T C; \end{aligned}$$

or equivalently,

$$\begin{pmatrix} \Gamma & 0 \\ 0 & M \end{pmatrix}\begin{pmatrix} 0 & K_1^H \\ K_1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & A_1^H \\ A_1 & 0 \end{pmatrix}\begin{pmatrix} \Gamma & 0 \\ 0 & M \end{pmatrix} = \begin{pmatrix} 0 & -\Sigma^T C \\ -N^T B & 0 \end{pmatrix}. \tag{3.4.57}$$

3) Taking the spectral norm $\|\cdot\|_2$ on the two sides of (3.4.57), applying a result due to Bhatia, Davis and Kittaneh [6] (see NR 2.5–5), and by the Mirsky theorem [78] (see below NR 3.4–3) we obtain

$$\begin{aligned} &\left\| \begin{pmatrix} \Gamma & 0 \\ 0 & M \end{pmatrix}\begin{pmatrix} 0 & K_1^H \\ K_1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & A_1^H \\ A_1 & 0 \end{pmatrix}\begin{pmatrix} \Gamma & 0 \\ 0 & M \end{pmatrix} \right\|_2 \\ &\geq \sigma_{\min}\begin{pmatrix} \Gamma & 0 \\ 0 & M \end{pmatrix} \left\| \begin{pmatrix} 0 & (K_1 - A_1)^H \\ K_1 - A_1 & 0 \end{pmatrix} \right\|_2 \\ &= \min\left\{ \sqrt{1 - \|\Sigma\|_2^2},\ \sqrt{1 - \|N\|_2^2} \right\} \|K_1 - A_1\|_2 \\ &\geq \min\left\{ \sqrt{1 - \rho_2^2(\mathcal{X}_1, \mathcal{Z}_1)},\ \sqrt{1 - \rho_2^2(\mathcal{Y}_1, \mathcal{W}_1)} \right\} |\kappa_j - \alpha_j|, \\ &j = 1, \ldots, l, \end{aligned} \tag{3.4.58}$$

where the relations of (3.4.50) are used.

On the other hand, using the relations of (3.4.50) and (3.4.56) we obtain

$$\begin{aligned} \left\| \begin{pmatrix} 0 & -\Sigma^T C \\ -N^T B & 0 \end{pmatrix} \right\|_2 &= \max\left\{ \|\Sigma^T C\|_2,\ \|N^T B\|_2 \right\} \\ &\leq \max\left\{ \|\Sigma\|_2 \|C\|_2,\ \|N\|_2 \|B\|_2 \right\} \\ &= \max\left\{ \rho_2(\mathcal{X}_1, \mathcal{Z}_1)\|S\|_2,\ \rho_2(\mathcal{Y}_1, \mathcal{W}_1)\|R\|_2 \right\}. \end{aligned}$$

Combining it with (3.4.57) and (3.4.58) shows the estimate (3.4.49). □

Notes and References

NR 3.4–1. Theorems 3.4.1 and 3.4.5 are proved by Sun [115].

NR 3.4–2. Let $A, X_1, Y_1, Z_1, W_1, A_1, K_1, R, S$ be the matrices as in Theorem 3.4.9. It is easy to see that for any $X = (X_1, X_2) \in \mathcal{U}^{n\times n}$ and $Y = (Y_1, Y_2) \in \mathcal{U}^{m\times m}$ we have

$$Y^H A X = \operatorname{diag}(A_1, A_2).$$

Wedin [127] shows that if for some $\delta > 0$

$$\sigma(K_1) \subset [\alpha + \delta,\ +\infty) \quad \text{and} \quad \sigma(A_2) \subset (-\infty,\ \alpha], \tag{3.4.59}$$

then

$$\max\{\|\sin\Theta(X_1, Z_1)\|_2,\ \|\sin\Theta(Y_1, W_1)\|_2\} \le \frac{\max\{\|R\|_2,\ \|S\|_2\}}{\delta}, \tag{3.4.60}$$

which gives a residual bound for an approximate singular subspace pair. Combining (3.4.60) with Theorem 3.4.9 we see that under the assumption (3.4.59) we have the following corollary: *If*

$$\epsilon \equiv \frac{\max\{\|R\|_2,\ \|S\|_2\}}{\delta} < 1,$$

then

$$|\alpha_j - \kappa_j| \le \frac{(\max\{\|R\|_2,\ \|S\|_2\})^2}{\delta\sqrt{1-\epsilon^2}}, \qquad j = 1, \ldots, l, \tag{3.4.61}$$

where α_j and κ_j are the singular values of A_1 and K_1, respectively. (3.4.61) gives a residual bound for approximate singular values.

NR 3.4–3. Mirsky Theorem [78]. *Let A and $\tilde{A}$ be matrices of the same dimensions with singular values*

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n, \qquad \tilde{\sigma}_1 \ge \tilde{\sigma}_2 \ge \cdots \ge \tilde{\sigma}_n.$$

Then for any unitarily invariant norm $\|\cdot\|$,

$$\|\mathrm{diag}(\tilde{\sigma}_i - \sigma_i)\| \le \|\tilde{A} - A\|.$$

(See, e.g., Stewart and Sun [97, Chapter IV, Theorem 4.11].)

Chapter 4

Generalized Eigenvalue Problems

This chapter is devoted to the generalized eigenvalue problem $\beta Ax = \alpha Bx$, where (A, B) is an $n \times n$ *regular pair*, i.e., $A, B \in \mathcal{C}^{n\times n}$ and there are $\alpha, \beta \in \mathcal{C}$ such that $\det(\beta A - \alpha B) \neq 0$. Perturbation expansions and condition numbers of eigenvalues and deflating subspaces, perturbation bounds for deflating subspaces, and backward errors and residual bounds, are discussed in §4.1 – §4.4, separately. The chapter concludes with a section on symmetric-definite generalized eigenproblems.

4.1 Perturbation Expansions

4.1.1 Simple Eigenvalues

Let (A, B) be an $n \times n$ regular pair. If

$$\beta Ax = \alpha Bx$$

for $(\alpha, \beta) \neq (0, 0)$ and a nonzero $x \in \mathcal{C}^n$, then (α, β) is called an *eigenvalue* of (A, B), and x a *right eigenvector* associated with (α, β). Usually, x is called an *eigenvector* of (A, B) associated with (α, β). The corresponding nonzero solution $y \in \mathcal{C}^n$ of the equation

$$\beta y^H A = \alpha y^H B$$

is called a *left* eigenvector associated with (α, β).

A basic fact of the generalized eigenvalue problem is that any eigenvalue (α, β) lies on the complex projective plane, or equivalently, any eigenvalue (α, β) lies on the Riemann sphere; i.e., (α, β) and $(z\alpha, z\beta)$ for any nonzero $z \in \mathcal{C}$ represent the same eigenvalue. If an eigenvalue (α, β) satisfies $\beta \neq 0$, then $\lambda = \alpha/\beta$ is a *finite* eigenvalue; otherwise, (α, β) is an *infinite* eigenvalue.

The set of the eigenvalues of a regular pair (A, B) is denoted by $\lambda(A, B)$.

Let $p = (p_1, \ldots, p_N)^T \in \mathcal{C}^N$, and $\mathcal{B}(0) \subset \mathcal{C}^N$ be a neighborhood of the origin. Let $A(p), B(p) \in \mathcal{C}^{n\times n}$ be analytic functions of p and $(A(p), B(p))$ be a regular pair for $p \in \mathcal{B}(0)$. Assume that (α, β) is a *simple* eigenvalue of $(A(0), B(0))$, and x, y are associated right and left eigenvectors satisfying

$$y^H A(0)x = \alpha, \qquad y^H B(0)x = \beta.$$

Then, as a consequence, there are $X_2, Y_2 \in \mathcal{C}^{n\times(n-1)}$ such that

$$X = (x, X_2) \quad \text{and} \quad Y = (y, Y_2) \quad \text{are nonsingular,} \tag{4.1.1}$$

and

$$Y^H A(0)X = \begin{pmatrix} \alpha & 0 \\ 0 & A_2 \end{pmatrix}, \qquad Y^H B(0)X = \begin{pmatrix} \beta & 0 \\ 0 & B_2 \end{pmatrix}, \tag{4.1.2}$$

where the pair (A_2, B_2) is regular, and

$$(\alpha, \beta) \notin \lambda(A_2, B_2). \tag{4.1.3}$$

First applying the implicit function theorem we prove the following result.

Theorem 4.1.1. *Let $A(p), B(p) \in \mathcal{C}^{n\times n}$ be analytic matrix-valued functions of p, and the matrix pair $(A(p), B(p))$ be a regular pair for $p \in \mathcal{B}(0)$, a neighborhood of the origin in $\mathcal{C}^N$. Assume that (α, β) is a simple eigenvalue of $(A(0), B(0))$, and x, y are associated right and left eigenvectors. Then*

1) there exists a simple eigenvalue $(\alpha(p), \beta(p))$ of the regular pair $(A(p), B(p))$, where $\alpha(p)$ and $\beta(p)$ are analytic functions of p in some neighborhood $\mathcal{B}_0$ of the origin, and $\alpha(0) = \alpha$, $\beta(0) = \beta$;

2) the functions $\alpha(p)$ and $\beta(p)$ have power series expansions at $p = 0$ of the forms

$$\alpha(p) = \alpha + \sum_{j=1}^{N} \left(\frac{\partial \alpha(p)}{\partial p_j}\right)_{p=0} p_j + \frac{1}{2}\sum_{j,k=1}^{N} \left(\frac{\partial^2 \alpha(p)}{\partial p_j \partial p_k}\right)_{p=0} p_j p_k + \cdots,$$

and

$$\beta(p) = \beta + \sum_{j=1}^{N} \left(\frac{\partial \beta(p)}{\partial p_j}\right)_{p=0} p_j + \frac{1}{2}\sum_{j,k=1}^{N} \left(\frac{\partial^2 \beta(p)}{\partial p_j \partial p_k}\right)_{p=0} p_j p_k + \cdots,$$

where $p \in \mathcal{B}_0$, and

$$\begin{aligned} \left(\frac{\partial \alpha(p)}{\partial p_j}\right)_{p=0} &= y^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x, \\ \left(\frac{\partial \beta(p)}{\partial p_j}\right)_{p=0} &= y^H \left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0} x, \end{aligned} \tag{4.1.4}$$

$$\left(\frac{\partial^2\alpha(p)}{\partial p_j\partial p_k}\right)_{p=0} = y^H\left(\frac{\partial^2 A(p)}{\partial p_j\partial p_k}\right)_{p=0} x + y^H\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} X_2G^{-1}Y_2^HD_kx$$
$$+y^H\left(\frac{\partial A(p)}{\partial p_k}\right)_{p=0} X_2G^{-1}Y_2^HD_jx, \tag{4.1.5}$$

$$\left(\frac{\partial^2\beta(p)}{\partial p_j\partial p_k}\right)_{p=0} = y^H\left(\frac{\partial^2 B(p)}{\partial p_j\partial p_k}\right)_{p=0} x + y^H\left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0} X_2G^{-1}Y_2^HD_kx$$
$$+y^H\left(\frac{\partial B(p)}{\partial p_k}\right)_{p=0} X_2G^{-1}Y_2^HD_jx, \tag{4.1.6}$$

in which the matrices G and D_j are defined by

$$G = \alpha B_2 - \beta A_2 \tag{4.1.7}$$

and

$$D_j = \beta\left(\frac{\partial A(p)}{\partial p_j}\right)_{p=1} - \alpha\left(\frac{\partial B(p)}{\partial p_j}\right)_{p=1}, \quad j = 1,\ldots,N. \tag{4.1.8}$$

Proof. 1) By the hypotheses there are matrices $X = (x, X_2), Y = (y, Y_2) \in \mathcal{C}^{n\times n}$ and $A_2, B_2 \in \mathcal{C}^{(n-1)\times(n-1)}$ such that the relations (4.1.1)–(4.1.3) hold. For $p \in \mathcal{B}(0)$ we set

$$\tilde{A}(p) = Y^HA(p)X = \begin{pmatrix} \tilde{a}_{11}(p) & \tilde{a}_{12}(p) \\ \tilde{a}_{21}(p) & \tilde{A}_{22}(p)\end{pmatrix}, \quad \tilde{a}_{11}(p) \in \mathcal{C},$$
$$\tilde{B}(p) = Y^HB(p)X = \begin{pmatrix} \tilde{b}_{11}(p) & \tilde{b}_{12}(p) \\ \tilde{b}_{21}(p) & \tilde{B}_{22}(p)\end{pmatrix}, \quad \tilde{b}_{11}(p) \in \mathcal{C}. \tag{4.1.9}$$

Using the same technique described by the proof of Theorem 2.1.1 we can show that there are analytic functions $z(p), w(p) \in \mathcal{C}^{n-1}$ of p in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin such that

$$\begin{pmatrix} 1 & 0 \\ -w(p) & I\end{pmatrix}\tilde{A}(p)\begin{pmatrix} 1 & 0 \\ z(p) & I\end{pmatrix} = \begin{pmatrix} \alpha(p) & * \\ 0 & *\end{pmatrix},$$
$$\begin{pmatrix} 1 & 0 \\ -w(p) & I\end{pmatrix}\tilde{B}(p)\begin{pmatrix} 1 & 0 \\ z(p) & I\end{pmatrix} = \begin{pmatrix} \beta(p) & * \\ 0 & *\end{pmatrix}, \tag{4.1.10}$$

and $z(0) = w(0) = 0$. Moreover, the functions $z(p)$ and $w(p)$ are uniquely determined.

From (4.1.10)

$$\alpha(p) = \tilde{a}_{11}(p) + \tilde{a}_{12}(p)z(p), \qquad \beta(p) = \tilde{b}_{11}(p) + \tilde{b}_{12}(p)z(p). \tag{4.1.11}$$

The relations of (4.1.10) show that $(\alpha(p), \beta(p))$ is an eigenvalue of $(A(p), B(p))$, and the eigenvalue is simple provided that the neighborhood $\mathcal{B}_0$ is sufficiently small. Moreover, the analyticity of the functions $\tilde{a}_{11}(p)$, $\tilde{a}_{12}(p)$, $\tilde{b}_{11}(p)$, $\tilde{b}_{12}(p)$ and $z(p)$ implies that $\alpha(p)$ and $\beta(p)$ are analytic functions of $p \in \mathcal{B}_0$, and $\alpha(0) = \alpha$, $\beta(0) = \beta$.

2) From (4.1.11) and $\tilde{a}_{12}(0)^T = z(0) = 0$ we obtain

$$\left(\frac{\partial \alpha(p)}{\partial p_j}\right)_{p=0} = \left(\frac{\partial \tilde{a}_{11}(p)}{\partial p_j}\right)_{p=0}, \tag{4.1.12}$$

and

$$\begin{aligned}\left(\frac{\partial^2 \alpha(p)}{\partial p_j \partial p_k}\right)_{p=0} = \left(\frac{\partial^2 \tilde{a}_{11}(p)}{\partial p_j \partial p_k}\right)_{p=0} + \left(\frac{\partial \tilde{a}_{12}(p)}{\partial p_j}\right)_{p=0} \left(\frac{\partial z(p)}{\partial p_k}\right)_{p=0} \\ + \left(\frac{\partial \tilde{a}_{12}(p)}{\partial p_k}\right)_{p=0} \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}.\end{aligned} \tag{4.1.13}$$

Moreover, from (4.1.9) we obtain

$$\begin{aligned}\left(\frac{\partial \tilde{a}_{11}(p)}{\partial p_j}\right)_{p=0} &= y^H \left(\frac{\partial\ A(p)}{\partial p_j}\right)_{p=0} x, \\ \left(\frac{\partial^2 \tilde{a}_{11}(p)}{\partial p_j \partial p_k}\right)_{p=0} &= y^H \left(\frac{\partial^2 A(p)}{\partial p_j \partial p_k}\right)_{p=0} x, \\ \left(\frac{\partial \tilde{a}_{12}(p)}{\partial p_j}\right)_{p=0} &= y^H \left(\frac{\partial\ A(p)}{\partial p_j}\right)_{p=0} X_2.\end{aligned} \tag{4.1.14}$$

Combining (4.1.12) with the first formula of (4.1.14) shows the first formula of (4.1.4).

From (4.1.13) and (4.1.14) it follows that for obtaining the formula (4.1.5) we only need to find an explicit expression of $\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0}$. By (4.1.10), the functions $z(p)$ and $w(p)$ satisfy the equations

$$\begin{aligned}&\tilde{a}_{21}(p) - \tilde{a}_{11}(p)w(p) + \tilde{A}_{22}(p)z(p) - \tilde{a}_{12}(p)z(p)w(p) = 0, \\ &\tilde{b}_{21}(p) - \tilde{b}_{11}(p)w(p) + \tilde{B}_{22}(p)z(p) - \tilde{b}_{12}(p)z(p)w(p) = 0,\end{aligned} \tag{4.1.15}$$

where $p \in \mathcal{B}_0$. Differentiating (4.1.15) at $p = 0$, we get

$$\begin{aligned}
\begin{pmatrix} \left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} \\ \left(\frac{\partial w(p)}{\partial p_j}\right)_{p=0} \end{pmatrix} &= \begin{pmatrix} -A_2 & \alpha I \\ -B_2 & \beta I \end{pmatrix}^{-1} \begin{pmatrix} \left(\frac{\partial \tilde{a}_{21}(p)}{\partial p_j}\right)_{p=0} \\ \left(\frac{\partial \tilde{b}_{21}(p)}{\partial p_j}\right)_{p=0} \end{pmatrix} \\
&= \begin{pmatrix} \beta I & -\alpha I \\ B_2 & -A_2 \end{pmatrix} \begin{pmatrix} G^{-1} Y_2^H \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} x \\ G^{-1} Y_2^H \left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0} x \end{pmatrix},
\end{aligned}$$

which gives

$$\left(\frac{\partial z(p)}{\partial p_j}\right)_{p=0} = G^{-1} Y_2^H \left[\beta \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} - \alpha \left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0} \right] x, \qquad (4.1.16)$$

where G is the matrix defined by (4.1.7). Substituting (4.1.14) and (4.1.16) into (4.1.13) shows the formula (4.1.5).

Similarly, we obtain the second formula of (4.1.4) and the formula (4.1.6). □

Remark 4.1.2. From (4.1.9) and (4.1.10)

$$\beta(p) A(p) x(p) = \alpha(p) B(p) x(p), \quad p \in \mathcal{B}_0, \qquad (4.1.17)$$

where $x(p)$ is defined by

$$x(p) = X \begin{pmatrix} 1 \\ z(p) \end{pmatrix}. \qquad (4.1.18)$$

The relation (4.1.17) shows that the vector $x(p)$ is an eigenvector of $(A(p), B(p))$ associated with $(\alpha(p), \beta(p))$, and the expression (4.1.18) shows that the eigenvector is an analytic function of $p \in \mathcal{B}_0$ satisfying $x(0) = x$. Moreover, (4.1.16) and (4.1.18) imply that the eigenvector $x(p)$ has the expansion of the form

$$x(p) = x + \sum_{j=1}^{N} \left(\frac{\partial x(p)}{\partial p_j}\right)_{p=0} p_j + \cdots, \quad p \in \mathcal{B}_0,$$

where

$$\left(\frac{\partial x(p)}{\partial p_j}\right)_{p=0} = X_2 G^{-1} Y_2^H \left[\beta \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} - \alpha \left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0} \right] x$$

for $j = 1, \ldots, N$.

The following result, as a corollary of Theorem 4.1.1, gives modified forms of the first order perturbation expansions of simple eigenvalues.

Corollary 4.1.3. *Let (A,B) be an $n \times n$ regular pair. Assume that (α,β) is a simple eigenvalue of (A,B), and x,y are associated right and left eigenvectors satisfying*

$$y^H Ax = \alpha, \qquad y^H Bx = \beta.$$

If $E,F \in \mathcal{C}^{n\times n}$ and $\|(E,F)\|_F$ is sufficiently small, then there exists a simple eigenvalue $(\tilde{\alpha},\tilde{\beta})$ of the regular pair $(A+E,B+F)$, and $\tilde{\alpha},\tilde{\beta}$ have the expansions

$$\begin{aligned}\tilde{\alpha} &= \alpha + y^H Ex + O(\|(E,F)\|_F^2),\\ \tilde{\beta} &= \beta + y^H Fx + O(\|(E,F)\|_F^2),\end{aligned} \tag{4.1.19}$$

where $(E,F) \to 0$.

Let $\rho(\cdot,\cdot)$ be the chordal metric defined by (1.3.4). Then the expansions of (4.1.19) give

$$\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta)) = \frac{\left|y^H Ex\cdot y^H Bx - y^H Fx \cdot y^H Ax\right|}{|y^H Ax|^2 + |y^H Bx|^2} + O\left(\|(E,F)\|_F^2\right), \tag{4.1.20}$$

where $(E,F) \to 0$.

Remark 4.1.4 (Definite Pairs). Let $A,B \in \mathcal{H}^{n\times n}$. The pair (A,B) is called a *definite pair* if

$$c(A,B) \equiv \min_{\substack{x \in \mathcal{C}^n\\ \|x\|_2 = 1}} \left|x^H(A+iB)x\right| > 0.$$

It is known (see, e.g., Crawford [23], Stewart and Sun [97, Chapter VI]) that any eigenvalue of an $n \times n$ definite pair (A,B) can be expressed by $(\alpha,\beta) \neq (0,0)$ with $\alpha,\beta \in \mathcal{R}$, and there is a nonsingular matrix $X \in \mathcal{C}^{n\times n}$ such that

$$X^H AX = \operatorname{diag}(\alpha_1,\dots,\alpha_n), \qquad X^H BX = \operatorname{diag}(\beta_1,\dots,\beta_n).$$

Let $A(p)$ and $B(p)$ be analytic matrix-valued functions, and $(A(p),B(p))$ be an $n \times n$ definite pair for $p \in \mathcal{B}(0)$, a neighborhood of the origin of $\mathcal{R}^N$. Assume that (α,β) is a simple eigenvalue of $(A(0),B(0))$, and x is an associated eigenvector. Then by using the same techniques described by the proofs of Theorem 3.1.1 and Theorem 4.1.1 we can obtain the same results as in (4.1.4)–(4.1.8) with $y = x$ and $Y_2 = X_2$, where $X = (x, X_2)$ is a nonsingular matrix that

$$X^H AX = \begin{pmatrix} \alpha & 0\\ 0 & A_2\end{pmatrix}, \quad X^H BX = \begin{pmatrix} \beta & 0\\ 0 & B_2\end{pmatrix},$$

in which (A_2,B_2) is an definite pair, and $(\alpha,\beta) \notin \lambda(A_2,B_2)$. Note that in the proof we only need to replace the transformation matrices

$$\begin{pmatrix} 1 & 0\\ -w(p) & I\end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0\\ z(p) & I\end{pmatrix}$$

of (4.1.10) by

$$\begin{pmatrix} 1 & z(p)^H \\ -w(p) & I \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & -w(p)^H \\ z(p) & I \end{pmatrix}.$$

Similar to Corollary 4.1.3, we have the following result for definite pairs.

Corollary 4.1.5. *Let (A, B) be an $n \times n$ definite pair. Assume that (α, β) is a simple eigenvalue of (A, B), and x is an associated eigenvector satisfying*

$$x^H Ax = \alpha, \qquad x^H Bx = \beta.$$

If $E, F \in \mathcal{H}^{n\times n}$ and $\|(E, F)\|_F$ is sufficiently small, then there exists a simple eigenvalue $(\tilde{\alpha}, \tilde{\beta})$ of the definite pair $(A + E, B + F)$, and $\tilde{\alpha}, \tilde{\beta}$ have the expansions

$$\tilde{\alpha} = \alpha + x^H Ex + O(\|(E, F)\|_F^2),$$

$$\tilde{\beta} = \beta + x^H Fx + O(\|(E, F)\|_F^2),$$

where $(E, F) \to 0$.

4.1.2 Deflating Subspaces

Let (A, B) be an $n \times n$ regular pair, (α, β) be an eigenvalue of (A, B), and $x \in \mathcal{C}^n$ be an associated eigenvector. By the definition of eigenvalue and eigenvector, there is a one-dimensional subspace $\mathcal{Y}_1 \subset \mathcal{C}^n$ such that

$$A\mathcal{R}(x) \subset \mathcal{Y}_1 \quad \text{and} \quad B\mathcal{R}(x) \subset \mathcal{Y}_1.$$

The pair $\{\mathcal{R}(x), \mathcal{Y}_1\}$ is called a pair of one-dimensional deflating subspaces of (A, B). Moreover, $\mathcal{R}(x)$ is called a one-dimensional eigenspace of (A, B). These definitions extend in a natural way to higher dimensions.

Let $\mathcal{X}_1, \mathcal{Y}_1$ be subspaces of $\mathcal{C}^n$ with the same dimension. The pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is called a pair of *deflating subspaces* of (A, B) if

$$A\mathcal{X}_1 \subset \mathcal{Y}_1 \quad \text{and} \quad B\mathcal{X}_1 \subset \mathcal{Y}_1.$$

The subspace $\mathcal{X}_1$ in the deflating subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is called an *eigenspace* (or a *generalized invariant subspace*) of (A, B). (If $B = I$, then $\mathcal{X}_1$ is an invariant subspace of A.)

The deflating subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ may be equivalently defined by $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$, in which $X_1, Y_1 \in \mathcal{C}^{n\times l}$ satisfy

$$\operatorname{rank}(X_1) = \operatorname{rank}(Y_1) = l \quad \text{and} \quad AX_1 = Y_1A_1, \;\; BX_1 = Y_1B_1$$

for some $l \times l$ regular pair (A_1, B_1).

Let $X_1, Y_1 \in \mathcal{U}^{n\times l}$, and let $\mathcal{X}_1 = \mathcal{R}(X_1)$, $\mathcal{Y}_1 = \mathcal{R}(Y_1)$. It can be verified that the subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is a deflating subspace pair of a regular matrix pair (A, B) if and only if there are matrices $X = (X_1, X_2)$, $Y = (Y_1, Y_2) \in \mathcal{U}^{n\times n}$ such that

$$Y^H AX = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Y^H BX = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}, \quad A_{11}, B_{11} \in \mathcal{C}^{l\times l}.$$

Moreover, (A_{11}, B_{11}) and (A_{22}, B_{22}) are regular pairs.

If $\lambda(A_{11}, B_{11}) \bigcap \lambda(A_{22}, B_{22}) = \emptyset$, then the deflating subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is called a *simple* deflating subspace pair. In this chapter we only consider simple deflating subspace pairs.

The main result of this subsection is the following perturbation expansion theorem.

Theorem 4.1.6. *Let (A, B) be an $n \times n$ regular matrix pair, and let $X = (X_1, X_2), Y = (Y_1, Y_2) \in \mathcal{U}^{n\times n}$ with $X_1, Y_1 \in \mathcal{U}^{n\times l}$ such that*

$$Y^H AX = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Y^H BX = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}, \tag{4.1.21}$$

where $A_{11}, B_{11} \in \mathcal{C}^{l\times l}$, and

$$\lambda(A_{11}, B_{11}) \bigcap \lambda(A_{22}, B_{22}) = \emptyset. \tag{4.1.22}$$

Moreover, let $\mathcal{X}_1 = \mathcal{R}(X_1), \mathcal{Y}_1 = \mathcal{R}(Y_1)$, for $M, N \in \mathcal{C}^{n\times n}$ let

$$Y^H MX = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}, \quad Y^H NX = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix}, \quad M_{11}, N_{11} \in \mathcal{C}^{l\times l}, \tag{4.1.23}$$

and define the linear operator $\mathbf{T} : \mathcal{C}^{(n-l)\times l} \times \mathcal{C}^{(n-l)\times l} \to \mathcal{C}^{(n-l)\times l} \times \mathcal{C}^{(n-l)\times l}$ by

$$\mathbf{T}\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} WA_{11} - A_{22}Z \\ WB_{11} - B_{22}Z \end{pmatrix}, \quad Z, W \in \mathcal{C}^{(n-l)\times l}. \tag{4.1.24}$$

Then

(1) there is a unique l-dimensional simple deflating subspace pair $\{\mathcal{X}_1(\tau), \mathcal{Y}_1(\tau)\}$ of $(A + \tau M, B + \tau N)$ $(\tau \in \mathcal{C})$ such that $\mathcal{X}_1(0) = \mathcal{X}_1, \mathcal{Y}_1(0) = \mathcal{Y}_1$, and the basis vectors $x_1(\tau), \ldots, x_l(\tau)$ of $\mathcal{X}_1(\tau)$ and the basis vectors $y_1(\tau), \ldots, y_l(\tau)$ of $\mathcal{Y}_1(\tau)$ may be chosen to be analytic functions of τ in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{C}$;

(2) the analytic matrix-valued functions

$$X_1(\tau) = (x_1(\tau), \ldots, x_l(\tau)), \qquad Y_1(\tau) = (y_1(\tau), \ldots, y_l(\tau))$$

have the perturbation expansions

$$X_1(\tau) = X_1 + X_2 \sum_{j=1}^{\infty} K_j \tau^j, \qquad Y_1(\tau) = Y_1 + Y_2 \sum_{j=1}^{\infty} L_j \tau^j \tag{4.1.25}$$

for $\tau \in \mathcal{B}(0)$, in which

$$\begin{pmatrix} K_1 \\ L_1 \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} M_{21} \\ N_{21} \end{pmatrix},$$

$$\begin{pmatrix} K_2 \\ L_2 \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} M_{22}K_1 - L_1 M_{11} - L_1 A_{12} K_1 \\ N_{22}K_1 - L_1 N_{11} - L_1 B_{12} K_1 \end{pmatrix},$$

$$\begin{pmatrix} K_j \\ L_j \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} M_{22}K_{j-1} - L_{j-1}M_{11} - \sum\limits_{k=1}^{j-2} L_{j-1-k} M_{12} K_k - \sum\limits_{k=1}^{j-1} L_{j-k} A_{12} K_k \\ N_{22}K_{j-1} - L_{j-1}N_{11} - \sum\limits_{k=1}^{j-2} L_{j-1-k} N_{12} K_k - \sum\limits_{k=1}^{j-1} L_{j-k} B_{12} K_k \end{pmatrix},$$

$$j \geq 3. \tag{4.1.26}$$

Proof. 1) Let

$$A(\tau) = A + \tau M, \quad B(\tau) = B + \tau N$$

and

$$\tilde{A}(\tau) = Y^H A(\tau) X = \begin{pmatrix} \tilde{A}_{11}(\tau) & \tilde{A}_{12}(\tau) \\ \tilde{A}_{21}(\tau) & \tilde{A}_{22}(\tau) \end{pmatrix},$$

$$\tilde{B}(\tau) = Y^H B(\tau) X = \begin{pmatrix} \tilde{B}_{11}(\tau) & \tilde{B}_{12}(\tau) \\ \tilde{B}_{21}(\tau) & \tilde{B}_{22}(\tau) \end{pmatrix}, \tag{4.1.27}$$

where $\tilde{A}_{11}(\tau), \tilde{B}_{11}(\tau) \in \mathcal{C}^{l \times l}$, and

$$\tilde{A}_{jk}(\tau) = A_{jk} + \tau M_{jk}, \quad \tilde{B}_{jk}(\tau) = B_{jk} + \tau N_{jk}, \quad A_{21} = B_{21} = 0. \tag{4.1.28}$$

Using the same technique described by the proof of Theorem 2.1.5 we can show that there are analytic matrix-valued functions $Z(\tau)$ and $W(\tau)$ of τ in some neighborhood $\mathcal{B}(0)$ of the origin such that

$$\begin{pmatrix} I & 0 \\ -W(\tau) & I \end{pmatrix} \tilde{A}(\tau) \begin{pmatrix} I & 0 \\ Z(\tau) & I \end{pmatrix} = \begin{pmatrix} A_1(\tau) & \tilde{A}_{12}(\tau) \\ 0 & A_2(\tau) \end{pmatrix},$$

$$\begin{pmatrix} I & 0 \\ -W(\tau) & I \end{pmatrix} \tilde{B}(\tau) \begin{pmatrix} I & 0 \\ Z(\tau) & I \end{pmatrix} = \begin{pmatrix} B_1(\tau) & \tilde{B}_{12}(\tau) \\ 0 & B_2(\tau) \end{pmatrix}, \tag{4.1.29}$$

and $Z(0) = W(0) = 0$. Moreover, the functions $Z(\tau)$ and $W(\tau)$ are uniquely determined; and $\lambda(A_1(\tau), B_1(\tau)) \bigcap \lambda(A_2(\tau), B_2(\tau)) = \emptyset$ provided that the neighborhood $\mathcal{B}(0)$ is sufficiently small.

Define

$$X_1(\tau) = X \begin{pmatrix} I \\ Z(\tau) \end{pmatrix}, \quad Y_1(\tau) = Y \begin{pmatrix} I \\ W(\tau) \end{pmatrix}. \tag{4.1.30}$$

Then from (4.1.27) and (4.1.29)

$$A(\tau)X_1(\tau) = Y_1(\tau)A_1(\tau), \quad B(\tau)X_1(\tau) = Y_1(\tau)B_1(\tau).$$

Consequently, we have proved that the pair $\{\mathcal{X}_1(\tau), \mathcal{Y}_1(\tau)\}$ with

$$\mathcal{X}_1(\tau) = \mathcal{R}(X_1(\tau)), \quad \mathcal{Y}_1(\tau) = \mathcal{R}(Y_1(\tau))$$

is the unique pair of deflating subspaces of $(A(\tau), B(\tau))$ in $\mathcal{B}(0)$ satisfying $\mathcal{X}_1(0) = \mathcal{X}_1, \mathcal{Y}_1(0) = \mathcal{Y}_1$, and $X_1(\tau), Y_1(\tau)$ are analytic matrix-valued functions of $\tau \in \mathcal{B}(0)$.

2) From (4.1.27)–(4.1.29) we get the basic equations for $Z(\tau)$ and $W(\tau)$:

$$\begin{aligned} &W(\tau)(A_{12} + \tau M_{12})Z(\tau) + W(\tau)(A_{11} + \tau M_{11}) - (A_{22} + \tau M_{22})Z(\tau) - \tau M_{21} = 0, \\ &W(\tau)(B_{12} + \tau N_{12})Z(\tau) + W(\tau)(B_{11} + \tau N_{11}) - (B_{22} + \tau N_{22})Z(\tau) - \tau N_{21} = 0, \end{aligned} \tag{4.1.31}$$

where $\tau \in \mathcal{B}(0)$.

Differentiating (4.1.31) at $\tau = 0$, and writing

$$Z^{(j)} = \left(\frac{\mathrm{d}^j Z(\tau)}{\mathrm{d}\tau^j} \right)_{\tau=0}, \quad W^{(j)} = \left(\frac{\mathrm{d}^j W(\tau)}{\mathrm{d}\tau^j} \right)_{\tau=0}, \quad j = 1, 2, \ldots,$$

we get

$$\mathbf{T}\begin{pmatrix} Z^{(1)} \\ W^{(1)} \end{pmatrix} = \begin{pmatrix} M_{21} \\ N_{21} \end{pmatrix},$$

$$\mathbf{T}\begin{pmatrix} Z^{(2)} \\ W^{(2)} \end{pmatrix} = 2\begin{pmatrix} M_{22}Z^{(1)} - W^{(1)}M_{11} - W^{(1)}A_{12}Z^{(1)} \\ N_{22}Z^{(1)} - W^{(1)}N_{11} - W^{(1)}B_{12}Z^{(1)} \end{pmatrix},$$

$$\mathbf{T}\begin{pmatrix} Z^{(j)} \\ W^{(j)} \end{pmatrix} = j\begin{pmatrix} M_{22}Z^{(j-1)} - W^{(j-1)}M_{11} - \sum\limits_{k=1}^{j-2}\begin{pmatrix} j-1 \\ k \end{pmatrix} W^{(j-1-k)}M_{12}Z^{(k)} \\ N_{22}Z^{(j-1)} - W^{(j-1)}N_{11} - \sum\limits_{k=1}^{j-2}\begin{pmatrix} j-1 \\ k \end{pmatrix} W^{(j-1-k)}N_{12}Z^{(k)} \end{pmatrix}$$

$$-\begin{pmatrix} \sum\limits_{k=1}^{j-1}\begin{pmatrix} j \\ k \end{pmatrix} W^{(j-k)}A_{12}Z^{(k)} \\ \sum\limits_{k=1}^{j-1}\begin{pmatrix} j \\ k \end{pmatrix} W^{(j-k)}B_{12}Z^{(k)} \end{pmatrix}, \quad j \geq 3, \tag{4.1.32}$$

where $\mathbf{T}$ is the linear operator defined by (4.1.24).

The hypothesis (4.1.22) implies that the operator $\mathbf{T}$ is invertible. Define

$$K_j = \frac{1}{j!}Z^{(j)}, \quad L_j = \frac{1}{j!}W^{(j)}, \quad j = 1, 2, \ldots.$$

Then from (4.1.32) we get the relations (4.1.26) and the power series expansions of $Z(\tau)$ and $W(\tau)$ at $\tau = 0$:

$$Z(\tau) = \sum_{j=1}^{\infty}\frac{1}{j!}Z^{(j)}\tau^j = \sum_{j=1}^{\infty}K_j\tau^j, \qquad W(\tau) = \sum_{j=1}^{\infty}\frac{1}{j!}W^{(j)}\tau^j = \sum_{j=1}^{\infty}L_j\tau^j.$$

This together with (4.1.30) shows (4.1.25). □

The following result, as a corollary of Theorem 4.1.6, gives modified forms of the first order perturbation expansions of simple deflating subspaces.

Corollary 4.1.7. *Let $(A, B), X, Y, A_{11}, A_{22}, B_{11}, B_{22}$ and $\mathbf{T}$ be as in Theorem 4.1.6, and let $\mathcal{X}_1 = \mathcal{R}(X_1)$, $\mathcal{Y}_1 = \mathcal{R}(Y_1)$. Moreover, for $E, F \in \mathcal{C}^{n\times n}$ let*

$$Y^HEX = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \quad Y^HFX = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}, \quad E_{11}, F_{11} \in \mathcal{C}^{l\times l}.$$

If $\|(E, F)\|_F$ is sufficiently small, then there exists a unique l-dimensional pair of deflating subspaces $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$, $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ of the pair $(A + E, B + F)$, and $\tilde{X}_1$,

$\tilde{Y}_1$ *have the expansions*

$$\begin{aligned}\tilde{X}_1 &= X_1 + X_2 Z_1 + O(\|(E,F)\|_F^2),\\ \tilde{Y}_1 &= Y_1 + Y_2 W_1 + O(\|(E,F)\|_F^2),\end{aligned} \tag{4.1.33}$$

where $(E,F) \to 0$, *and* $Z_1, W_1 \in \mathcal{C}^{(n-l)\times l}$ *are defined by*

$$\begin{pmatrix} Z_1 \\ W_1 \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} E_{21} \\ F_{21} \end{pmatrix}. \tag{4.1.34}$$

Observe that by using the Kronecker product and vec operator, the matrix representation T of the linear operator $\mathbf{T}$ defined by (4.1.24) can be expressed by

$$T = \begin{pmatrix} -I_l \otimes A_{22} & A_{11}^T \otimes I_{n-l} \\ -I_l \otimes B_{22} & B_{11}^T \otimes I_{n-l} \end{pmatrix}. \tag{4.1.35}$$

Consequently, the relation (4.1.34) can be written in an equivalent form:

$$\begin{pmatrix} \mathrm{vec}(Z_1) \\ \mathrm{vec}(W_1) \end{pmatrix} = C \begin{pmatrix} \mathrm{vec}(E_{21}) \\ \mathrm{vec}(F_{21}) \end{pmatrix}, \tag{4.1.36}$$

where

$$\begin{aligned} &C \equiv T^{-1} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \quad C_1 = (C_{11},\, C_{12}), \quad C_2 = (C_{21},\, C_{22}),\\ &C_{11} = (B_{11}^T \otimes I_{n-l})M^{-1}, \quad C_{12} = (-A_{11}^T \otimes I_{n-l})M^{-1},\\ &C_{21} = (I_l \otimes B_{22})M^{-1}, \quad C_{22} = (-I_l \otimes A_{22})M^{-1},\\ &M = A_{11}^T \otimes B_{22} - B_{11}^T \otimes A_{22}. \end{aligned} \tag{4.1.37}$$

Notes and References

NR 4.1–1. The formulas of (4.1.4) are given by Liu [72, Theorem 3.5].

NR 4.1–2. The notion of the deflating subspaces of a regular pair is introduced by Stewart [91].

NR 4.1–3. Theorem 4.1.6 is proved by Sun [119, Theorem 3.1.1].

NR 4.1–4. Crawford [23] first discusses perturbation properties of eigenvalues of definite pairs. For basic perturbation results, see Stewart and Sun [97, Chapter VI].

4.2 Condition Numbers

4.2.1 Simple Eigenvalues

Let (A, B) be a regular pair, and (α, β) be a simple eigenvalue of (A, B). Let $(\tilde{A}, \tilde{B}) = (A+E, B+F)$ be a perturbation of (A, B), and $(\tilde{\alpha}, \tilde{\beta})$ be the corresponding perturbation of (α, β). Then by §1.8 we define the condition number $c(\alpha, \beta)$ for (α, β) by the following approach: Define the vector v by

$$v = \left(\frac{\|E\|_F}{\gamma_{\rm A}}, \frac{\|F\|_F}{\gamma_{\rm B}} \right)^T, \tag{4.2.1}$$

and then define $c(\alpha, \beta)$ as

$$c(\alpha, \beta) = \lim_{\delta \to 0} \sup_{\|v\|_2 \le \delta} \frac{\rho((\tilde{\alpha}, \tilde{\beta}), (\alpha, \beta))}{\delta}, \tag{4.2.2}$$

where $\gamma_{\rm A}$ and $\gamma_{\rm B}$ are positive parameters, and $\rho(\cdot, \cdot)$ denotes the chordal metric defined by (1.3.4).

From the definition (4.2.2) it follows that in first order approximation the inequality

$$\rho((\tilde{\alpha}, \tilde{\beta}), (\alpha, \beta)) \le c(\alpha, \beta) \left\| \left(\frac{\|E\|_F}{\gamma_{\rm A}}, \frac{\|F\|_F}{\gamma_{\rm B}} \right)^T \right\|_2$$

holds.

If one is interested in the sensitivity of (α, β) to perturbations in each individual member of A and B, then by §1.8 we define the partial condition numbers $c_A(\alpha, \beta)$ and $c_B(\alpha, \beta)$ for (α, β) as

$$\begin{aligned} c_A(\alpha, \beta) &= \lim_{\delta \to 0} \sup_{\frac{\|E\|}{\gamma_{\rm A}} \le \delta,\ F=0} \frac{\rho((\tilde{\alpha}, \tilde{\beta}), (\alpha, \beta))}{\delta}, \\ c_B(\alpha, \beta) &= \lim_{\delta \to 0} \sup_{E=0,\ \frac{\|F\|}{\gamma_{\rm B}} \le \delta} \frac{\rho((\tilde{\alpha}, \tilde{\beta}), (\alpha, \beta))}{\delta}, \end{aligned} \tag{4.2.3}$$

where $\gamma_{\rm A}$ and $\gamma_{\rm B}$ are positive parameters.

Let (A, B) be a regular pair, and (α, β) be a simple eigenvalue of (A, B) with right eigenvector x and left eigenvector y. The following results (Theorems 4.2.1 and 4.2.2) give explicit expressions of the condition numbers $c(\alpha, \beta)$, $c_A(\alpha, \beta)$ and $c_B(\alpha, \beta)$.

Theorem 4.2.1. *The condition number $c(\alpha,\beta)$ can be expressed by*

$$c(\alpha,\beta) = \frac{\left\| \left(\gamma_{\mathrm{B}} y^H Ax,\ \gamma_{\mathrm{A}} y^H Bx \right) \right\|_2 \|x\|_2 \|y\|_2}{|y^H Ax|^2 + |y^H Bx|^2}. \tag{4.2.4}$$

Proof. By Corollary 4.1.3 and (4.1.20), we have

$$\begin{aligned} \frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta} &\leq \frac{\left\| \left(\gamma_{\mathrm{B}} y^H Ax,\ \gamma_{\mathrm{A}} y^H Bx \right) \right\|_2 \|x\|_2 \|y\|_2}{|y^H Ax|^2 + |y^H Bx|^2} + O(\delta) \\ &\text{if} \quad \left\| \left(\frac{\|E\|_F}{\gamma_{\mathrm{A}}},\ \frac{\|F\|_F}{\gamma_{\mathrm{B}}} \right) \right\|_2 \leq \delta; \end{aligned} \tag{4.2.5}$$

and the equalities in (4.2.5) are achieved for the specific perturbations

$$\widehat{E} = -\frac{\delta \gamma_{\mathrm{A}} \sigma y x^H}{\|x\|_2 \|y\|_2} \quad \text{and} \quad \widehat{F} = \frac{\delta \gamma_{\mathrm{B}} \tau y x^H}{\|x\|_2 \|y\|_2}$$

with

$$\sigma = \gamma_{\mathrm{A}} \overline{y^H Bx} \left\| \left(\gamma_{\mathrm{A}} y^H Bx,\ \gamma_{\mathrm{B}} y^H Ax \right) \right\|_2^{-1/2},$$

$$\tau = \gamma_{\mathrm{B}} \overline{y^H Ax} \left\| \left(\gamma_{\mathrm{A}} y^H Bx,\ \gamma_{\mathrm{B}} y^H Ax \right) \right\|_2^{-1/2}.$$

Combining these facts with the definition (4.2.2) shows the expression (4.2.4). □

Theorem 4.2.2. *The partial condition numbers $c_A(\alpha,\beta)$ and $c_B(\alpha,\beta)$ can be expressed by*

$$\begin{aligned} c_A(\alpha,\beta) &= \frac{\gamma_{\mathrm{A}} |y^H Bx| \|x\|_2 \|y\|_2}{|y^H Ax|^2 + |y^H Bx|^2}, \\ c_B(\alpha,\beta) &= \frac{\gamma_{\mathrm{B}} |y^H Ax| \|x\|_2 \|y\|_2}{|y^H Ax|^2 + |y^H Bx|^2}. \end{aligned} \tag{4.2.6}$$

Proof. By Corollary 4.1.3 and (4.1.20), we have

$$\begin{aligned} \frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta} &\leq \frac{\gamma_{\mathrm{A}} |y^H Bx| \|x\|_2 \|y\|_2}{|y^H Ax|^2 + |y^H Bx|^2} + O(\delta) \\ &\text{if} \quad \frac{\|E\|}{\gamma_{\mathrm{A}}} \leq \delta \ll 1 \quad \text{and} \quad F = 0, \end{aligned} \tag{4.2.7}$$

and the equalities in (4.2.7) are achieved for the specific perturbations

$$\widehat{E} = -\frac{\delta \gamma_{\mathrm{A}} y x^H}{\|x\|_2 \|y\|_2} \quad \text{and} \quad \widehat{F} = 0.$$

Moreover, we have

$$\frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta} \le \frac{\gamma_{\mathrm{B}}|y^H Ax|\|x\|_2\|y\|_2}{|y^H Ax|^2+|y^H Bx|^2}+O(\delta) \qquad \text{if} \quad E=0 \quad \text{and} \quad \frac{\|F\|}{\gamma_{\mathrm{B}}} \le \delta \ll 1, \tag{4.2.8}$$

and the equalities in (4.2.8) are achieved for the specific perturbations

$$\widehat{E}=0 \qquad \text{and} \qquad \widehat{F}=\frac{\delta\gamma_{\mathrm{B}}yx^H}{\|x\|_2\|y\|_2}.$$

Combining these facts with the definition (4.2.3) shows the expressions of (4.2.6). □

Taking $\gamma_{\mathrm{A}}=\gamma_{\mathrm{B}}=1$ in (4.2.1)–(4.2.2) and (4.2.4) yields the absolute condition number

$$c_{\mathrm{abs}}(\alpha,\beta)=\frac{\|x\|_2\|y\|_2}{\sqrt{|y^H Ax|^2+|y^H Bx|^2}}; \tag{4.2.9}$$

and taking $\gamma_{\mathrm{A}}=\|A\|_F$ and $\gamma_{\mathrm{B}}=\|B\|_F$ in (4.2.1)–(4.2.2) and (4.2.4) yields the relative condition number

$$c_{\mathrm{rel}}(\alpha,\beta)=\frac{\left\|\Big(\|B\|_F y^H Ax,\ \|A\|_F y^H Bx\Big)\right\|_2\|x\|_2\|y\|_2}{|y^H Ax|^2+|y^H Bx|^2}. \tag{4.2.10}$$

Moreover, taking $\gamma_{\mathrm{A}}=\gamma_{\mathrm{B}}=1$ in (4.2.3) and (4.2.6) yields the *absolute partial* condition numbers $c_A^{(\mathrm{abs})}(\alpha,\beta)$ and $c_B^{(\mathrm{abs})}(\alpha,\beta)$, and taking $\gamma_{\mathrm{A}}=\|A\|_F$ and $\gamma_{\mathrm{B}}=\|B\|_F$ in (4.2.3) and (4.2.6) yields the *relative partial* condition numbers $c_A^{(\mathrm{rel})}(\alpha,\beta)$ and $c_B^{(\mathrm{rel})}(\alpha,\beta)$, respectively.

Example 4.2.3 (Parlett [83, p.304–305]). Consider the regular pair (A,B) with

$$A=\begin{pmatrix}1 & 0\\ 0 & 10^{-8}\end{pmatrix}, \qquad B=\begin{pmatrix}1 & 0\\ 0 & 2\times 10^{-8}\end{pmatrix}.$$

The matrix pair has the eigenvalues $(\alpha_1,\beta_1)=(1,1)$ and $(\alpha_2,\beta_2)=(10^{-8},2\times 10^{-8})$, or equivalently, $\lambda_1=\alpha_1/\beta_1=1$ and $\lambda_2=\alpha_2/\beta_2=1/2$. It is easy to see that a change of 10^{-8} in A and/or in B changes the eigenvalue 1 by 10^{-4}, while the eigenvalue 1/2 changes completely. In other words, the eigenvalue 1 is well-conditioned, and the eigenvalue 1/2 is violently ill-conditioned. By (4.2.9), (4.2.10) and (4.2.6), we have

$$c_{\mathrm{abs}}(\alpha_1,\beta_1)=1/\sqrt{2}, \qquad c_{\mathrm{rel}}(\alpha_1,\beta_1)\approx 1/\sqrt{2},$$

$$c_A^{(\mathrm{abs})}(\alpha_1,\beta_1)=1/2, \qquad c_B^{(\mathrm{abs})}(\alpha_1,\beta_1)=1/2,$$

$$c_A^{(\mathrm{rel})}(\alpha_1,\beta_1)\approx 1/2, \qquad c_B^{(\mathrm{rel})}(\alpha_1,\beta_1)\approx 1/2,$$

and

$$c_{\rm abs}(\alpha_2,\beta_2) = \tfrac{1}{\sqrt{5}} \times 10^8, \qquad c_{\rm rel}(\alpha_2,\beta_2) \approx \tfrac{1}{\sqrt{5}} \times 10^8,$$

$$c_A^{(\rm abs)}(\alpha_2,\beta_2) = \tfrac{2}{5} \times 10^8, \qquad c_B^{(\rm abs)}(\alpha_2,\beta_2) = \tfrac{1}{5} \times 10^8,$$

$$c_A^{(\rm rel)}(\alpha_2,\beta_2) \approx \tfrac{2}{5} \times 10^8, \qquad c_B^{(\rm rel)}(\alpha_2,\beta_2) \approx \tfrac{1}{5} \times 10^8.$$

Obviously, for this example the condition numbers defined in this subsection reflect the sensitivity of the eigenvalues.

The following result shows an important fact that if (α,β) is a simple eigenvalue of a regular pair (A,B), then the distance from (A,B) to a matrix pair which has an eigenvalue (α,β) of multiplicity at least two is approximately bounded by the scalar

$$\frac{\|(A,B)\|_2}{c_{\rm abs}(\alpha,\beta)\sqrt{|\alpha|^2+|\beta|^2}}.$$

Theorem 4.2.4. *Let (A,B) be an $n\times n$ regular pair with the generalized Schur decomposition*

$$A = Q\begin{pmatrix} \alpha & a^H \\ 0 & A_2 \end{pmatrix} Z^H, \qquad B = Q\begin{pmatrix} \beta & b^H \\ 0 & B_2 \end{pmatrix} Z^H, \tag{4.2.11}$$

where $Q,Z \in \mathcal{U}^{n\times n}$, and (α,β) is a simple eigenvalue of (A,B). If the condition number $c_{\rm abs}(\alpha,\beta)$ satisfies $\sqrt{|\alpha|^2+|\beta|^2}c_{\rm abs}(\alpha,\beta) > 1$, then there exist $E,F \in \mathcal{C}^{n\times n}$ such that the pair $(A+E,B+F)$ has (α,β) as an eigenvalue of multiplicity at least two and

$$\begin{aligned} \|(E,F)\|_2 &= \frac{\|(a^H,b^H)\|_2}{\sqrt{(|\alpha|^2+|\beta|^2)[c_{\rm abs}(\alpha,\beta)]^2-1}} \\ &< \frac{\|(A,B)\|_2}{\sqrt{(|\alpha|^2+|\beta|^2)[c_{\rm abs}(\alpha,\beta)]^2-1}}. \end{aligned} \tag{4.2.12}$$

Proof. Since (α,β) is a simple eigenvalue of (A,B), there are $v,w \in \mathcal{C}^{n-1}$ such that

$$\begin{aligned} \begin{pmatrix} 1 & w^H \\ 0 & I_{n-1} \end{pmatrix}\begin{pmatrix} \alpha & a^H \\ 0 & A_2 \end{pmatrix}\begin{pmatrix} 1 & v^H \\ 0 & I_{n-1} \end{pmatrix} &= \begin{pmatrix} \alpha & 0 \\ 0 & A_2 \end{pmatrix}, \\ \begin{pmatrix} 1 & w^H \\ 0 & I_{n-1} \end{pmatrix}\begin{pmatrix} \beta & b^H \\ 0 & B_2 \end{pmatrix}\begin{pmatrix} 1 & v^H \\ 0 & I_{n-1} \end{pmatrix} &= \begin{pmatrix} \beta & 0 \\ 0 & B_2 \end{pmatrix}. \end{aligned} \tag{4.2.13}$$

Write $Q = (q_1, Q_2)$ and $Z = (z_1, Z_2)$, where $q_1, z_1 \in \mathcal{C}^n$. Then by (4.2.11) and (4.2.13) we have the relations

$$\begin{pmatrix} q_1^H + w^H Q_2^H \\ * \end{pmatrix} A\,(z_1,\ *) = \begin{pmatrix} \alpha & 0 \\ 0 & A_2 \end{pmatrix},$$

$$\begin{pmatrix} q_1^H + w^H Q_2^H \\ * \end{pmatrix} B\,(z_1,\ *) = \begin{pmatrix} \beta & 0 \\ 0 & B_2 \end{pmatrix},$$

which show that the vectors

$$x = z_1 \quad \text{and} \quad y = q_1 + Q_2 w \tag{4.2.14}$$

are right and left eigenvectors of (A, B) belonging to the simple eigenvalue (α, β). Thus, by (4.2.9) and (4.2.14), the condition number $c_{\text{abs}}(\alpha, \beta)$ can be expressed by

$$c_{\text{abs}}(\alpha, \beta) = \sqrt{(1 + \|w\|_2^2)/(|\alpha|^2 + |\beta|^2)}. \tag{4.2.15}$$

Moreover, the relations (4.2.11), (4.2.14) and $\beta y^H A = \alpha y^H B$ imply

$$\beta(1, w^H) \begin{pmatrix} \alpha & a^H \\ 0 & A_2 \end{pmatrix} = \alpha(1, w^H) \begin{pmatrix} \beta & b^H \\ 0 & B_2 \end{pmatrix},$$

or equivalently,

$$\beta w^H \left(A_2 + \frac{w a^H}{\|w\|_2^2} \right) = \alpha w^H \left(B_2 + \frac{w b^H}{\|w\|_2^2} \right).$$

Take

$$E = Q \begin{pmatrix} 0 & 0 \\ 0 & \frac{w a^H}{\|w\|_2^2} \end{pmatrix} Z^H, \qquad F = Q \begin{pmatrix} 0 & 0 \\ 0 & \frac{w b^H}{\|w\|_2^2} \end{pmatrix} Z^H. \tag{4.2.16}$$

Then (α, β) is an eigenvalue of $(A + E, B + F)$ of multiplicity at least two, and from (4.2.15) and (4.2.16) we get the estimates (4.2.12). □

Remark 4.2.5 (Definite Pairs). Let (A, B) be a definite pair, and (α, β) be a simple eigenvalue of (A, B). Then by (4.2.1)–(4.2.3) we may define the structured condition number $c(\alpha, \beta)$ and the *structured partial* condition numbers $c_A(\alpha, \beta)$ and $c_B(\alpha, \beta)$ by using Hermitian perturbations E and F. Using the same technique described in the proof of Theorems 4.2.1 and 4.2.2, and applying Corollary 4.1.5, we obtain the same formulas as (4.2.4) and (4.2.6), where $y = x$; i.e.,

$$c(\alpha, \beta) = \frac{\left\| \left(\gamma_{\text{B}} x^H A x, \ \gamma_{\text{A}} x^H B x \right) \right\|_2 \|x\|_2^2}{(x^H A x)^2 + (x^H B x)^2},$$

and

$$c_A(\alpha, \beta) = \frac{\gamma_{\text{A}} |x^H B x| \|x\|_2^2}{(x^H A x)^2 + (x^H B x)^2}, \qquad c_B(\alpha, \beta) = \frac{\gamma_{\text{B}} |x^H A x| \|x\|_2^2}{(x^H A x)^2 + (x^H B x)^2}.$$

4.2.2 Deflating Subspaces

Let (A, B) be a regular pair, and $\{\mathcal{X}_1, \mathcal{Y}_1\}$ be a simple deflating subspace pair of (A, B). Let $(\tilde{A}, \tilde{B}) = (A + E, B + F)$ be a perturbation of (A, B), and $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ be the corresponding perturbation of $\{\mathcal{X}_1, \mathcal{Y}_1\}$. Then by §1.8 we define the condition number $c(\mathcal{X}_1)$ for $\mathcal{X}_1$ by the following approach: Define the vector v by (4.2.1), and then define $c(\mathcal{X}_1)$ as

$$c(\mathcal{X}_1) = \lim_{\delta\to 0} \sup_{\|v\|_2\le\delta} \frac{\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1)}{\delta}. \tag{4.2.17}$$

The condition number $c(\mathcal{Y}_1)$ for $\mathcal{Y}_1$ can be defined in the same way.

From the definition (4.2.17) we see that in first order approximation the inequality

$$\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \le c(\mathcal{X}_1) \left\| \left(\frac{\|E\|_F}{\gamma_{\rm A}}, \frac{\|F\|_F}{\gamma_{\rm B}} \right) \right\|_2$$

holds. For $\rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1)$ we have a similar estimate.

Moreover, we may define the partial condition numbers $c_A(\mathcal{X}_1)$ and $c_B(\mathcal{X}_1)$ for $\mathcal{X}_1$ as

$$\begin{aligned} c_A(\mathcal{X}_1) &= \lim_{\delta\to 0} \sup_{\frac{\|E\|_F}{\gamma_{\rm A}}\le\delta,\, F=0} \tfrac{\rho_F(\mathcal{X}_1,\tilde{\mathcal{X}}_1)}{\delta}, \\ c_B(\mathcal{X}_1) &= \lim_{\delta\to 0} \sup_{E=0,\, \frac{\|F\|_F}{\gamma_{\rm B}}\le\delta} \tfrac{\rho_F(\mathcal{X}_1,\tilde{\mathcal{X}}_1)}{\delta}, \end{aligned} \tag{4.2.18}$$

where $\gamma_{\rm A}$ and $\gamma_{\rm B}$ are positive parameters. The partial condition numbers $c_A(\mathcal{Y}_1)$ and $c_B(\mathcal{Y}_1)$ for $\mathcal{Y}_1$ can be defined in the same way.

Take $X = (X_1, X_2), Y = (Y_1, Y_2) \in \mathcal{U}^{n\times n}$ with $X_1, Y_1 \in \mathcal{U}^{n\times l}$ so that $\mathcal{X}_1 = \mathcal{R}(X_1)$, $\mathcal{Y}_1 = \mathcal{R}(Y_1)$, and

$$Y^H AX = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Y^H BX = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}, \tag{4.2.19}$$

where $A_{11}, B_{11} \in \mathcal{C}^{l\times l}$, and $\lambda(A_{11}, B_{11}) \bigcap \lambda(A_{22}, B_{22}) = \emptyset$. For $E, F \in \mathcal{C}^{n\times n}$ let

$$Y^H EX = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \quad Y^H FX = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix},$$

where $E_{11}, F_{11} \in \mathcal{C}^{n\times n}$. Moreover, define the linear operator $\mathbf{T}$ by (4.1.24). It is noted in §4.1.2 that the linear operator $\mathbf{T}$ has the matrix representation T of (4.1.35), $\mathbf{T}$ is invertible, and the inverse of T is expressed by (4.1.37). We now use the expressions of (4.1.37) to give computable formulas of the condition numbers $c(\mathcal{X}_1)$, $c(\mathcal{Y}_1)$, $c_A(\mathcal{X}_1)$, $c_B(\mathcal{X}_1)$, $c_A(\mathcal{Y}_1)$ and $c_B(\mathcal{Y}_1)$.

Theorem 4.2.6. *The condition numbers $c(\mathcal{X}_1)$ and $c(\mathcal{Y}_1)$ can be expressed by*

$$c(\mathcal{X}_1) = \left\| C_1 \begin{pmatrix} \gamma_{\rm A} I & 0 \\ 0 & \gamma_{\rm B} I \end{pmatrix} \right\|_2, \quad c(\mathcal{Y}_1) = \left\| C_2 \begin{pmatrix} \gamma_{\rm A} I & 0 \\ 0 & \gamma_{\rm B} I \end{pmatrix} \right\|_2, \tag{4.2.20}$$

where the matrices C_1 and C_2 are defined by (4.1.37).

Proof. Let A, B have the decomposition (4.2.19), $(A+E, B+F)$ be a perturbation of (A, B) with sufficiently small $\|(E,F)\|$, and let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ be the corresponding perturbation of $\{\mathcal{X}_1, \mathcal{Y}_1\}$. By (4.2.17), Theorem 1.3.3 (see the relation (1.3.17)) and Corollary 4.1.7 (see the expansions of (4.1.33)), we have

$$c(\mathcal{X}_1) = \lim_{\delta \to 0} \sup_{\|v\|_2 \le \delta} \frac{\|Z_1\|_F}{\delta},$$

where v is the vector defined by (4.2.1), and Z_1 by (4.1.34). Thus,

$$\begin{aligned} c(\mathcal{X}_1) &= \sup_{\left\| \begin{pmatrix} \mathrm{vec}(E)/\gamma_{\mathrm{A}} \\ \mathrm{vec}(F)/\gamma_{\mathrm{B}} \end{pmatrix} \right\|_2 \le 1} \left\| C_1 \begin{pmatrix} \gamma_{\mathrm{A}} I & 0 \\ 0 & \gamma_{\mathrm{B}} I \end{pmatrix} \begin{pmatrix} \mathrm{vec}(E_{21})/\gamma_{\mathrm{A}} \\ \mathrm{vec}(F_{21})/\gamma_{\mathrm{B}} \end{pmatrix} \right\|_2 \\ &= \left\| C_1 \begin{pmatrix} \gamma_{\mathrm{A}} I & 0 \\ 0 & \gamma_{\mathrm{B}} I \end{pmatrix} \right\|_2 . \end{aligned}$$

Similarly, we obtain the computable formula of $c(\mathcal{Y}_1)$. □

Theorem 4.2.7. *The partial condition numbers $c_A(\mathcal{X}_1)$, $c_B(\mathcal{X}_1)$, $c_A(\mathcal{Y}_1)$ and $c_B(\mathcal{Y}_1)$ can be expressed by*

$$\begin{aligned} c_A(\mathcal{X}_1) &= \gamma_{\mathrm{A}} \|C_{11}\|_2, \qquad c_B(\mathcal{X}_1) = \gamma_{\mathrm{B}} \|C_{12}\|_2, \\ c_A(\mathcal{Y}_1) &= \gamma_{\mathrm{A}} \|C_{21}\|_2, \qquad c_B(\mathcal{Y}_1) = \gamma_{\mathrm{B}} \|C_{22}\|_2, \end{aligned} \tag{4.2.21}$$

where the matrices C_{11}, C_{12}, C_{21} and C_{22} are defined by (4.1.37).

Proof. By (4.2.18) and the same argument described in the proof of Theorem 4.2.6, we have

$$\begin{aligned} c_A(\mathcal{X}_1) &= \lim_{\delta \to 0} \sup_{\frac{\|E\|_F}{\gamma_{\mathrm{A}}} \le \delta, \, F=0} \frac{\|\mathrm{vec}(Z_1)\|_2}{\delta} \\ &= \gamma_{\mathrm{A}} \sup_{\|\mathrm{vec}(E_{21})\|_2 \le 1} \|C_{11} \mathrm{vec}(E_{21})\|_2 = \gamma_{\mathrm{A}} \|C_{11}\|_2 . \end{aligned}$$

Similarly, we obtain the other formulas of (4.2.21). □

Remark 4.2.8. By Stewart [91], the simple deflating subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ has the (absolute) condition number $c(\mathcal{X}_1, \mathcal{Y}_1)$ which can be expressed by

$$c(\mathcal{X}_1, \mathcal{Y}_1) = \|C\|_2, \tag{4.2.22}$$

where the matrix C is defined by (4.1.37). However, the conditioning of the two subspaces $\mathcal{X}_1$ and $\mathcal{Y}_1$ may be quite different, and the condition number $c(\mathcal{X}_1, \mathcal{Y}_1)$ is governed by the ill-conditioning of the most sensitive subspace of the deflating subspace pair. By (4.1.37), we have

$$\|C_1\|_2 \leq \|C\|_2 \quad \text{and} \quad \|C_2\|_2 \leq \|C\|_2; \tag{4.2.23}$$

and in some cases

$$\|C_1\|_2 \ll \|C\|_2 \quad \text{or} \quad \|C_2\|_2 \ll \|C\|_2, \tag{4.2.24}$$

which means that in some cases $c(\mathcal{X}_1, \mathcal{Y}_1)$ may be a severe overestimate of the sensitivity of $\mathcal{X}_1$ or $\mathcal{Y}_1$ (see Example 4.2.10 below).

Remark 4.2.9. Taking $\gamma_{\rm A} = \gamma_{\rm B} = 1$ in (4.2.17), (4.2.18), (4.2.20) and (4.2.21) yields the absolute condition numbers $c_{\rm abs}(\mathcal{X}_1)$, $c_{\rm abs}(\mathcal{Y}_1)$, and $c_A^{(\rm abs)}(\mathcal{X}_1)$, $c_B^{(\rm abs)}(\mathcal{X}_1)$, $c_A^{(\rm abs)}(\mathcal{Y}_1)$, $c_B^{(\rm abs)}(\mathcal{Y}_1)$. For example, we have

$$c_{\rm abs}(\mathcal{X}_1) = \|C_1\|_2, \qquad c_{\rm abs}(\mathcal{Y}_1) = \|C_2\|_2. \tag{4.2.25}$$

Example 4.2.10. Let (A, B) be a 4×4 regular pair having the generalized Schur decomposition (4.1.21) with

$$A_{11} = 10^{-5} \times I_2, \qquad B_{11} = \begin{pmatrix} 10^{-4} & 0 \\ 10^{-4} & 10^{-4} \end{pmatrix}, \qquad A_{22} = B_{22} = I_2.$$

According to Remark 4.2.9, we have

$$c_{\rm abs}(\mathcal{X}_1) \approx 1.8960, \qquad c_A^{(\rm abs)}(\mathcal{X}_1) \approx 1.8929, \qquad c_B^{(\rm abs)}(\mathcal{X}_1) \approx 1.8883 \times 10^{-1},$$

$$c_{\rm abs}(\mathcal{Y}_1) \approx 2.6705 \times 10^4, \quad c_A^{(\rm abs)}(\mathcal{Y}_1) \approx 1.8883 \times 10^4, \quad c_B^{(\rm abs)}(\mathcal{Y}_1) \approx 1.8883 \times 10^4;$$

and by (4.2.22),

$$c(\mathcal{X}_1, \mathcal{Y}_1) \approx 2.6705 \times 10^4.$$

Obviously, for this example the condition number $c(\mathcal{X}_1, \mathcal{Y}_1)$ is a severe overestimate of the (absolute) sensitivity of the eigenspace $\mathcal{X}_1$.

Notes and References

NR 4.2–1. The chordal metric $\rho((\tilde{\alpha}, \tilde{\beta}), (\alpha, \beta))$ is first used in the perturbation theory for matrix pairs by Stewart [92]. The condition number $c_{\rm abs}(\alpha, \beta)$ of (4.2.9) is given by Stewart and Sun [97, Chapter VI]; and it is also proved by Dedieu [28, Corollary 7.3].

NR 4.2–2. Let (A, B) be a regular pair, and (α, β) be a simple eigenvalue of (A, B). Let $(\tilde{A}, \tilde{B}) = (A + E, B + F)$ be a perturbation of (A, B), and $(\tilde{\alpha}, \tilde{\beta})$ be the corresponding perturbation of (α,β). Define v by (4.2.1), and then define the condition number $c_p(\alpha,\beta)$ of a simple eigenvalue (α,β) by

$$c_p(\alpha,\beta) = \lim_{\delta\to 0} \sup_{\|v\|_p\le\delta} \frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta}, \tag{4.2.26}$$

where $\|\cdot\|_p$ is the p-norm ($p \ge 1$), and $\gamma_{\mathrm{A}}, \gamma_{\mathrm{B}}$ are positive parameters. The following result gives a computable formula of $c_p(\alpha,\beta)$.

Theorem 4.2.11. *The condition number $c_p(\alpha,\beta)$ can be expressed by*

$$c_p(\alpha,\beta) = \frac{\left\|\left(\gamma_{\mathrm{B}} y^H Ax,\ \gamma_{\mathrm{A}} y^H Bx\right)^T\right\|_q \|x\|_2\|y\|_2}{|y^H Ax|^2 + |y^H Bx|^2}, \tag{4.2.27}$$

where q satisfies $1/p+1/q=1$, and x and y are right and left eigenvectors of (A,B) associated with λ.

Proof. The proof is completed by the following three steps.

1. On $c_p(\alpha,\beta)$ for $p=1$. By Corollary 4.1.3 and (4.1.20), we have

$$\begin{aligned}\frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta} &\le \frac{\left(|y^H Bx|\|E\| + |y^H Ax|\|F\|\right)\|x\|_2\|y\|_2}{\delta\left(|y^H Ax|^2 + |y^H Bx|^2\right)} + O(\delta)\\ &\le \frac{\max\left\{\gamma_{\mathrm{B}}|y^H Ax|,\ \gamma_{\mathrm{A}}|y^H Bx|\right\}}{|y^H Ax|^2+|y^H Bx|^2} + O(\delta) \qquad \text{if} \quad \left\|\left(\frac{\|E\|}{\gamma_{\mathrm{A}}}, \frac{\|F\|}{\gamma_{\mathrm{B}}}\right)^T\right\|_1 \le \delta.\end{aligned} \tag{4.2.28}$$

On the other hand, if $\gamma_{\mathrm{A}}|y^H Bx| \ge \gamma_{\mathrm{B}}|y^H Ax|$ then the equalities in (4.2.28) are achieved for the specific perturbations

$$\widehat{E} = \frac{\delta\gamma_{\mathrm{A}} yx^H}{\|x\|_2\|y\|_2} \quad \text{and} \quad \widehat{F} = 0;$$

if $\gamma_{\mathrm{A}}|y^H Bx| \le \gamma_{\mathrm{B}}|y^H Ax|$ then the equalities in (4.2.28) are achieved for the specific perturbations

$$\widehat{E} = 0 \quad \text{and} \quad \widehat{F} = \frac{\delta\gamma_{\mathrm{B}} yx^H}{\|x\|_2\|y\|_2}.$$

Consequently, using the definition (4.2.26) with $p=1$ we derive the computable formula of $c_1(\alpha,\beta)$.

2. On $c_p(\alpha,\beta)$ for $1 < p < \infty$. By Corollary 4.1.3 and (4.1.20), we have

$$\frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta} \leq \frac{\left\|\left(\gamma_{\mathrm{B}}\cdot y^H Ax,\ \gamma_{\mathrm{A}}\cdot y^H Bx\right)^T\right\|_q \|x\|_2\|y\|_2}{|y^H Ax|^2+|y^H Bx|^2} + O(\delta)$$
$$\text{if}\quad \left\|\left(\frac{\|E\|}{\gamma_{\mathrm{A}}},\frac{\|F\|}{\gamma_{\mathrm{B}}}\right)^T\right\|_p \leq \delta. \tag{4.2.29}$$

On the other hand, the equalities in (4.2.29) are achieved for the specific perturbations

$$\widehat{E} = -\frac{\delta\gamma_{\mathrm{A}}\sigma y x^H}{\|x\|_2\|y\|_2} \quad\text{and}\quad \widehat{F} = \frac{\delta\gamma_{\mathrm{B}}\tau y x^H}{\|x\|_2\|y\|_2}$$

with

$$\sigma = \gamma_{\mathrm{A}}\cdot\overline{y^H Bx}\cdot(\gamma_{\mathrm{A}}|y^H Bx|)^{\frac{q}{p}-1}\cdot\left\|\left(\gamma_{\mathrm{A}}\cdot y^H Bx,\ \gamma_{\mathrm{B}}\cdot y^H Ax\right)^T\right\|_q^{\frac{1}{q}-1},$$

$$\tau = \gamma_{\mathrm{B}}\cdot\overline{y^H Ax}\cdot(\gamma_{\mathrm{B}}|y^H Ax|)^{\frac{q}{p}-1}\cdot\left\|\left(\gamma_{\mathrm{A}}\cdot y^H Bx,\ \gamma_{\mathrm{B}}\cdot y^H Ax\right)^T\right\|_q^{\frac{1}{q}-1},$$

Consequently, using the definition (4.2.26) we derive the computable formula of $c_p(\alpha,\beta)$ for $1 < p < \infty$.

3. On $c_\infty(\alpha,\beta)$. By Corollary 4.1.3 and (4.1.20), we have

$$\frac{\rho((\tilde{\alpha},\tilde{\beta}),(\alpha,\beta))}{\delta} \leq \frac{\left(\gamma_{\mathrm{B}}|y^H Ax|+\gamma_{\mathrm{A}}|y^H Bx|\right)\|x\|_2\|y\|_2}{(|y^H Ax|^2+|y^H Bx|^2} + O(\delta)$$
$$\text{if}\quad \left\|\left(\frac{\|E\|}{\gamma_{\mathrm{A}}},\frac{\|F\|}{\gamma_{\mathrm{B}}}\right)^T\right\|_\infty \leq \delta. \tag{4.2.30}$$

On the other hand, the equalities in (4.2.30) are achieved for the specific perturbations

$$\widehat{E} = -\frac{\delta\gamma_{\mathrm{A}}\mathrm{e}^{-i\arg(y^H Bx)}y x^H}{\|x\|_2\|y\|_2} \quad\text{and}\quad \widehat{F} = \frac{\delta\gamma_{\mathrm{B}}\mathrm{e}^{-i\arg(y^H Ax)}y x^H}{\|x\|_2\|y\|_2}.$$

Consequently, using the definition (4.2.26) with $p = \infty$ we derive the computable formula of $c_\infty(\alpha,\beta)$. □

NR 4.2–3. Frayssé and Toumazou [39], and D. Higham and N. Higham [46] consider finite non-zero simple eigenvalues of a regular pair (A,B). Let (λ,β) and $(\tilde{\alpha},\tilde{\beta})$ be as in NR 4.2–2 with $\beta\tilde{\beta}\neq 0$, and let

$$\lambda = \frac{\alpha}{\beta}, \qquad \tilde{\lambda} = \frac{\tilde{\alpha}}{\tilde{\beta}}.$$

Frayssé and Toumazou [39] define the relative condition number $K(\lambda)$ by

$$K(\lambda) = \lim_{\delta\to 0} \sup_{\|v\|_\infty \le \delta} \frac{|\tilde{\lambda} - \lambda|}{|\lambda|\delta}, \tag{4.2.31}$$

where v is the vector defined by (4.2.26), and $\|\cdot\|$ denotes any vector norm and subordinate matrix norm. Frayssé and Toumazou [39, Lemma 3.1] prove that $K(\lambda)$ can be expressed by

$$K(\lambda) = \frac{(\gamma_{\text{A}} + |\lambda|\gamma_{\text{B}})\|x\|\|y\|^D}{|\lambda||y^H Bx|}, \tag{4.2.32}$$

where x and y are right and left eigenvectors of (A, B) associated with λ, and $\|\cdot\|^D$ denotes the dual norm of $\|\cdot\|$.

D. Higham and N. Higham [46] define the componentwise condition number $\text{cond}(\lambda)$ for a finite non-zero simple eigenvalue $\lambda = \alpha/\beta$ by

$$\text{cond}(\lambda) = \lim_{\delta\to 0} \sup_{\substack{|E| \le \delta\Phi \\ |F| \le \delta\Psi}} \frac{|\tilde{\lambda} - \lambda|}{|\lambda|\delta}, \tag{4.2.33}$$

where Φ and Ψ are two proper matrices. D. Higham and N. Higham [46, Theorem 3.2] prove that $\text{cond}(\lambda)$ can be expressed by

$$\text{cond}(\lambda) = \frac{|y^H||E|x| + |\lambda||y^H|F|x|}{|\lambda||y^H Bx|}. \tag{4.2.34}$$

Moreover, structured condition numbers of simple eigenvalues of some special matrix pairs (for example, Hermitian matrix pairs, Toeplitz matrix pairs, or banded matrix pairs) are also studied by D. Higham and N. Higham [46, §4].

Note that the computable formulas (4.2.32) and (4.2.34) can be proved by applying Corollary 4.1.3. In fact, by (4.1.19), we have

$$|\tilde{\lambda} - \lambda| = \frac{\left|y^H Ex - \lambda y^H Fx\right|}{|\beta|} + O\left(\|(E, F)\|^2\right), \tag{4.2.35}$$

where $(E, F) \to 0$. Combining (4.2.31) with (4.2.35) shows (4.2.32), and combining (4.2.33) with (4.2.35) shows (4.2.34).

NR 4.2–4. Consider the regular pair (A, B) with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 10^{-8} & 0 \\ 0 & 1 \end{pmatrix}.$$

It is easy to see that $(\alpha_1, \beta_1) = (1, 10^{-8})$ and $(\alpha_2, \beta_2) = (2, 1)$, or equivalently, $\lambda_1 = \alpha_1/\beta_1 = 10^8$ and $\lambda_2 = \alpha_2/\beta_2 = 2$, are simple eigenvalues of the matrix pair.

Taking $\gamma_{\rm A} = \|A\|_2$, $\gamma_{\rm B} = \|B\|_2$ and $\|\cdot\| = \|\cdot\|_2$ in (4.2.32) gives the condition numbers

$$K(\lambda_1) \approx 10^8, \qquad K(\lambda_2) \approx 2,$$

which mean that the eigenvalue λ_2 is well-conditioned but λ_1 is ill-conditioned according to the prevailing point of view. Observe that the generalized eigenvalues of a matrix pair lie on the Riemann sphere. Hence, to use the chordal metric is more appropriate for investigating perturbation behavior of generalized eigenvalues. By (4.2.10), we have the condition numbers

$$c_{\rm rel}(\alpha_1, \beta_1) \approx 1, \qquad c_{\rm rel}(\alpha_2, \beta_2) \approx 0.6,$$

which show that both the eigenvalues (α_1, β_1) and (α_2, β_2) are well behaved in the chordal metric sense.

NR 4.2–5. Let $\{\mathcal{X}_1, \mathcal{Y}_1\}$ be a simple deflating subspace pair. For estimates of the condition number $c(\mathcal{X}_1, \mathcal{Y}_1)$ of (4.2.22), see Kågström and Poromaa [59] and [60]. The problem of how to compute or estimate the condition numbers $c(\mathcal{X}_1)$, $c(\mathcal{Y}_1)$ of (4.2.20), as well as the partial condition numbers $c_A(\mathcal{X}_1)$, $c_B(\mathcal{X}_1)$, $c_A(\mathcal{Y}_1)$ and $c_B(\mathcal{Y}_1)$ of (4.2.21), efficiently, is a research problem.

4.3 Perturbation Bounds for Deflating Subspaces

A perturbation bound for simple deflating subspace pairs has been obtained by Stewart [91, Theorem 5.7]. We now apply Theorem 3.3.5 to derive a new result. The difference between the new result and Stewart's result is that the new result gives an individual perturbation bound for each subspace in a deflating subspace pair, separately.

Theorem 4.3.1. *Let $(A, B), X = (X_1, X_2), Y = (Y_1, Y_2), A_{ij}, B_{ij}, \mathcal{X}_1, \mathcal{Y}_1$ be as in Theorem 4.1.6. For $E, F \in \mathcal{C}^{n\times n}$ let*

$$Y^H E X = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \quad Y^H F X = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}, \tag{4.3.1}$$

where $E_{11}, F_{11} \in \mathcal{C}^{l\times l}$. Moreover, let $c_{\rm abs}(\mathcal{X}_1), c_{\rm abs}(\mathcal{Y}_1)$ be the condition numbers expressed by (4.2.25), and let

$$c_* = \sqrt{[c_{\rm abs}(\mathcal{X}_1)]^2 + [c_{\rm abs}(\mathcal{Y}_1)]^2}, \qquad \epsilon = \|(E_{11}, F_{11})\|_2 + \left\| \begin{pmatrix} E_{22} \\ F_{22} \end{pmatrix} \right\|_2, \tag{4.3.2}$$

and

$$\gamma = \left\| \begin{pmatrix} E_{21} \\ F_{21} \end{pmatrix} \right\|_F, \qquad \eta = \max\{\|A_{12}\|_2 + \|E_{12}\|_2,\ \|B_{12}\|_2 + \|F_{12}\|_2\}. \tag{4.3.3}$$

If

$$c_* \left(2\sqrt{\gamma\eta} + \epsilon\right) < 1, \tag{4.3.4}$$

then there is a unique pair of l-dimensional deflating subspaces $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ of $(A+E, B+F)$ such that

$$\begin{aligned} \rho(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \le \|\tan\Theta(X_1, \tilde{X}_1)\|_F &\le \frac{2c_{\text{abs}}(\mathcal{X}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}}, \\ \rho(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) \le \|\tan\Theta(Y_1, \tilde{Y}_1)\|_F &\le \frac{2c_{\text{abs}}(\mathcal{Y}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}}, \end{aligned} \tag{4.3.5}$$

where $\Theta(\cdot,\cdot)$ is defined by (1.3.1).

Proof. Let $\mathbf{T}$ be the linear operator defined by (4.1.24). It is easy to verify that $\begin{pmatrix} Z \\ W \end{pmatrix}$ is a solution of the equation

$$\mathbf{T}\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} E_{21} \\ F_{21} \end{pmatrix} + \begin{pmatrix} -WE_{11} + E_{22}Z \\ -WF_{11} + F_{22}Z \end{pmatrix} - \begin{pmatrix} W(A_{12} + E_{12})Z \\ W(B_{12} + F_{12})Z \end{pmatrix} \tag{4.3.6}$$

if and only if Z and W satisfy

$$\begin{aligned} \begin{pmatrix} I & 0 \\ -W & I \end{pmatrix} \begin{pmatrix} A_{11} + E_{11} & A_{12} + E_{12} \\ E_{21} & A_{22} + E_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ Z & I \end{pmatrix} &= \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}, \\ \begin{pmatrix} I & 0 \\ -W & I \end{pmatrix} \begin{pmatrix} B_{11} + F_{11} & B_{12} + F_{12} \\ F_{21} & B_{22} + F_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ Z & I \end{pmatrix} &= \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}. \end{aligned} \tag{4.3.7}$$

The relations of (4.3.7) imply that the pair of the subspaces

$$\tilde{\mathcal{X}}_1 = \mathcal{R}\left(X\begin{pmatrix} I \\ Z \end{pmatrix}\right), \quad \tilde{\mathcal{Y}}_1 = \mathcal{R}\left(Y\begin{pmatrix} I \\ W \end{pmatrix}\right)$$

is an l-dimensional deflating subspace pair of $(A+E, B+F)$. Consequently, by Theorem 1.3.3 (see the relation (1.3.16)), the problem of proving (4.3.5) is reduced to the problem of finding a solution $\begin{pmatrix} Z^* \\ W^* \end{pmatrix}$ of (4.3.6) in a certain neighborhood of the origin.

Let C_1, C_2 be the matrices defined by (4.1.37), and let

$$z = \operatorname{vec}(Z), \quad w = \operatorname{vec}(W), \quad e_{21} = \operatorname{vec}(E_{21}), \quad f_{21} = \operatorname{vec}(F_{21}),$$

$$x(z,w) = \operatorname{vec}(-WE_{11} + E_{22}Z), \quad y(z,w) = \operatorname{vec}(-WF_{11} + F_{22}Z), \tag{4.3.8}$$

and

$$u(z,w) = \mathrm{vec}(W(A_{12}+E_{12})Z), \quad v(z,w) = \mathrm{vec}(W(B_{12}+F_{12})Z). \tag{4.3.9}$$

Then the equation (4.3.6) can be written in an equivalent form

$$\left\{ \begin{aligned} z &= C_1 \left[\begin{pmatrix} e_{21} \\ f_{21} \end{pmatrix} + \begin{pmatrix} x(z,w) \\ y(z,w) \end{pmatrix} - \begin{pmatrix} u(z,w) \\ v(z,w) \end{pmatrix} \right], \\ w &= C_2 \left[\begin{pmatrix} e_{21} \\ f_{21} \end{pmatrix} + \begin{pmatrix} x(z,w) \\ y(z,w) \end{pmatrix} - \begin{pmatrix} u(z,w) \\ v(z,w) \end{pmatrix} \right]. \end{aligned} \right. \tag{4.3.10}$$

Define the functions f and h by

$$f = \begin{pmatrix} x \\ y \end{pmatrix}, \quad h = \begin{pmatrix} u \\ v \end{pmatrix}.$$

Observe that f and h satisfy the conditions (3.3.21) and (3.3.22), where the scalars ϵ and η are defined by (4.3.2) and (4.3.3), respectively. Hence, by Theorem 3.3.5, if

$$c_*\epsilon < 1 \quad \text{and} \quad \frac{4c_*^2\gamma\eta}{(1-c_*\epsilon)^2} < 1,$$

or equivalently, if $c_*, \gamma, \eta, \epsilon$ satisfy the condition (4.3.4), then the equation (4.3.10) has a unique solution $\begin{pmatrix} z^* \\ w^* \end{pmatrix}$ (or equivalently, the equation (4.3.6) has a unique solution $\begin{pmatrix} Z^* \\ W^* \end{pmatrix}$) satisfying

$$\|Z^*\|_F = \|z^*\|_2 \le \frac{2c_{\text{abs}}(\mathcal{X}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}},$$

$$\|W^*\|_F = \|w^*\|_2 \le \frac{2c_{\text{abs}}(\mathcal{Y}_1)\gamma}{1 - c_*\epsilon + \sqrt{(1-c_*\epsilon)^2 - 4c_*^2\gamma\eta}}.$$

Combining it with (1.3.12) and (1.3.16) shows the inequalities of (4.3.5). □

Remark 4.3.2. The estimates (4.3.5) imply that if $c_*\left(2\sqrt{\gamma\eta}+\epsilon\right)$ is sufficiently small, or more intuitively, if $\|(E,F)\|$ is sufficiently small, then

$$\|\tan\Theta(X_1,\tilde{X}_1)\|_F \lesssim c_{\text{abs}}(\mathcal{X}_1)\gamma, \quad \|\tan\Theta(Y_1,\tilde{Y}_1)\|_F \lesssim c_{\text{abs}}(\mathcal{Y}_1)\gamma. \tag{4.3.11}$$

Note that by Stewart [96, Theorem 5.7], we have

$$\left\| \begin{pmatrix} \tan\Theta(X_1\tilde{X}_1) \\ \tan\Theta(Y_1,\tilde{Y}_1) \end{pmatrix} \right\|_F \lesssim c(\mathcal{X}_1,\mathcal{Y}_1)\gamma \tag{4.3.12}$$

when $\|(E,F)\|$ is sufficiently small, where $c(\mathcal{X}_1,\mathcal{Y}_1)$ is defined by (4.2.22). From (4.2.22)–(4.2.25) we see that the bounds (4.3.11) and (4.3.12) are, in general, qualitatively the same, but in some cases the result (4.3.11) is better (even much better) than (4.3.12) if one needs to bound perturbations of each subspace of the pair $\{\mathcal{X}_1,\mathcal{Y}_1\}$, separately. The drawback of the bound (4.3.12) is that it is governed by the ill-conditioning of the most sensitive subspace of the deflating subspace pair.

If the matrices E_{jk} and F_{jk} of (4.3.1) are known, then we can apply Theorem 3.3.4 to derive the following result on perturbation bounds for deflating subspaces which will be used in §4.4.2.

Theorem 4.3.3. *Let $(A,B), X, Y, A_{jk}, B_{jk}, \mathcal{X}_1, \mathcal{Y}_1, E, F$ and E_{jk}, F_{jk} $(j,k = 1,2)$ be as in Theorem 4.3.1, and C_1, C_2 be the matrices of (4.1.37). Moreover, let*

$$b_1 = \left\| C_1 \begin{pmatrix} \mathrm{vec}E_{21} \\ \mathrm{vec}F_{21} \end{pmatrix} \right\|_2, \qquad c_1 = \|C_1\|_2,$$

$$b_2 = \left\| C_2 \begin{pmatrix} \mathrm{vec}E_{21} \\ \mathrm{vec}F_{21} \end{pmatrix} \right\|_2, \qquad c_2 = \|C_2\|_2, \tag{4.3.13}$$

$$b = b_1 + b_2, \qquad c = c_1 + c_2,$$

and let

$$\eta = \max\left\{ \|A_{12} + E_{12}\|_2,\ \|B_{12} + F_{12}\|_2 \right\},$$

$$\epsilon = \left\| \begin{pmatrix} \|E_{11}\|_2 & \|E_{22}\|_2 \\ \|F_{11}\|_2 & \|F_{22}\|_2 \end{pmatrix} \right\|_2. \tag{4.3.14}$$

If

$$c\epsilon < 1 \quad \text{and} \quad \frac{4bc\eta}{(1-c\epsilon)^2} < 1, \tag{4.3.15}$$

then there is a unique pair of deflating subspaces $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$, $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ of the matrix pair $(A+E, B+F)$ such that $\tilde{X}_1 \in \mathcal{U}^{n\times l}$, $\tilde{Y}_1 \in \mathcal{U}^{m\times l}$, and

$$\rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) \le \|\tan\Theta(X_1,\tilde{X}_1)\|_F \le b_1 + c_1(\epsilon\beta + \eta\beta^2),$$

$$\rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) \le \|\tan\Theta(Y_1,\tilde{Y}_1)\|_F \le b_2 + c_2(\epsilon\beta + \eta\beta^2), \tag{4.3.16}$$

where

$$\beta = \frac{2b}{1 - c\epsilon + \sqrt{(1-c\epsilon)^2 - 4bc\eta}}. \tag{4.3.17}$$

Proof. From the proof of Theorem 4.3.1 we see that it only needs to show the following fact: Under the assumptions (4.3.15) the system (4.3.10) has a unique solution $\begin{pmatrix} z^* \\ w^* \end{pmatrix}$ satisfying

$$\begin{aligned} \|z^*\|_2 &\leq b_1 + c_1(\epsilon\beta + \gamma\beta^2), \\ \|w^*\|_2 &\leq b_2 + c_2(\epsilon\beta + \gamma\beta^2), \end{aligned} \tag{4.3.18}$$

where β is the scalar defined by (4.3.17).

Applying Theorem 3.3.4 to the system (4.3.10), and using the assumptions (4.3.15), we get the estimates (4.3.18) immediately. □

Notes and References

NR 4.3–1. The first perturbation bound for deflating subspace pair is obtained by Stewart [91, Theorem 5.7]. Theorem 4.3.1 is proved by Sun [119, Theorem 3.4.1].

4.4 Backward Errors and Residual Bounds

4.4.1 Backward Errors

In this subsection we discuss several kinds of normwise backward errors which are defined by using some information of approximate deflating subspaces and associated eigenmatrices of a matrix pair (A, B).

4.4.1.1 The Backward Errors $\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ and $\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$

Let (A, B) be an $n\times n$ regular pair, and let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ approximate an l-dimensional deflating subspace pair of (A, B). By §1.9, we define the backward errors $\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ and $\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ of (A, B) with respect to $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ by

$$\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{G}} \left\| \begin{pmatrix} E \\ \theta F \end{pmatrix} \right\|, \tag{4.4.1}$$

and

$$\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{G}} \mu \begin{pmatrix} \|E\| \\ \omega\|F\| \end{pmatrix}, \tag{4.4.2}$$

where θ, ω are positive parameters, $\mu(\cdot)$ is any absolute norm on $\mathcal{R}^2$, and the set $\mathcal{G}$ is defined by

$$\mathcal{G} = \left\{ \begin{pmatrix} E \\ F \end{pmatrix} \; : \; E, F \in \mathcal{C}^{n\times n}, \; (A+E)\tilde{\mathcal{X}}_1 \subset \tilde{\mathcal{Y}}_1, \; (B+F)\tilde{\mathcal{X}}_1 \subset \tilde{\mathcal{Y}}_1 \right\}. \tag{4.4.3}$$

The following result gives a computable formula of $\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$.

Theorem 4.4.1. *Let (A, B) be an $n \times n$ regular pair. Let $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$ with $\tilde{V}_1, \tilde{U}_1 \in \mathcal{U}^{n\times l}$, and let*

$$R_A = \tilde{U}_1(\tilde{U}_1^H A\tilde{V}_1) - A\tilde{V}_1, \quad R_B = \tilde{U}_1(\tilde{U}_1^H B\tilde{V}_1) - B\tilde{V}_1 \tag{4.4.4}$$

be the residuals of (A, B) with respect to $\tilde{V}_1$ and $\tilde{U}_1$. Then the backward error $\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ can be expressed by

$$\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \left\| \begin{pmatrix} R_A \\ \theta R_B \end{pmatrix} \right\|. \tag{4.4.5}$$

The expressions (4.4.4) and (4.4.5) imply that the backward error $\eta^{(\theta)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ defined by (4.4.1) is independent of the choice of the matrices $\tilde{V}_1$ and $\tilde{U}_1$ whose column vectors form orthonormal bases of $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{Y}}_1$, respectively.

Proof of Theorem 4.4.1. From (4.4.3) it follows that a matrix $\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{G}$ if and only if $\begin{pmatrix} E \\ F \end{pmatrix}$ is a solution to the equation

$$\begin{pmatrix} A + E \\ B + F \end{pmatrix} \tilde{V}_1 = \begin{pmatrix} \tilde{U}_1 A_1 \\ \tilde{U}_1 B_1 \end{pmatrix}$$

for some $A_1, B_1 \in \mathcal{C}^{l\times l}$, or equivalently, $\begin{pmatrix} E \\ F \end{pmatrix}$ satisfies

$$\begin{pmatrix} E \\ F \end{pmatrix} \tilde{V}_1 = \begin{pmatrix} \tilde{U}_1 A_1 - A\tilde{V}_1 \\ \tilde{U}_1 B_1 - B\tilde{V}_1 \end{pmatrix}. \tag{4.4.6}$$

Applying Theorem 1.5.1 to the equation (4.4.6) we see that the equation is solvable, and any solution $\begin{pmatrix} E \\ F \end{pmatrix}$ of the equation can be expressed by

$$\begin{pmatrix} E \\ F \end{pmatrix} = \begin{pmatrix} \tilde{U}_1 A_1 - A\tilde{V}_1 \\ \tilde{U}_1 B_1 - B\tilde{V}_1 \end{pmatrix} \tilde{V}_1^H + \begin{pmatrix} Z \\ W \end{pmatrix} (I - \tilde{V}_1 \tilde{V}_1^H), \tag{4.4.7}$$

where $Z, W \in \mathcal{C}^{n\times n}$.

Choose $\tilde{V}_2, \tilde{U}_2$ so that $\tilde{V} = (\tilde{V}_1, \tilde{V}_2), \tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Then from (4.4.7)

$$\begin{pmatrix} \tilde{U}^H & 0 \\ 0 & \tilde{U}^H \end{pmatrix} \begin{pmatrix} E \\ \theta F \end{pmatrix} \tilde{V} = \begin{pmatrix} A_1 - \tilde{U}_1^H A\tilde{V}_1 & \tilde{U}_1^H Z\tilde{V}_2 \\ -\tilde{U}_2^H A\tilde{V}_1 & \tilde{U}_2^H Z\tilde{V}_2 \\ \theta(B_1 - \tilde{U}_1^H B\tilde{V}_1) & \theta\tilde{U}_1^H W\tilde{V}_2 \\ -\theta\tilde{U}_2^H B\tilde{V}_1 & \theta\tilde{U}_2^H W\tilde{V}_2 \end{pmatrix}.$$

By the definition (4.4.1) and Theorem 1.2.1, we have

$$\eta(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \left\| \begin{pmatrix} E_{\rm opt} \\ \theta F_{\rm opt} \end{pmatrix} \right\| \quad \text{with} \quad \begin{pmatrix} E_{\rm opt} \\ F_{\rm opt} \end{pmatrix} = \begin{pmatrix} R_A \\ R_B \end{pmatrix} \tilde{V}_1^H, \tag{4.4.8}$$

where R_A, R_B are the residuals defined by (4.4.4).

Combining (4.4.8) with

$$\sigma_+\left(\begin{pmatrix} R_A \\ \theta R_B \end{pmatrix} \tilde{V}_1^H \right) = \sigma_+ \begin{pmatrix} R_A \\ \theta R_B \end{pmatrix}$$

shows (4.4.5). □

The following result gives a computable formula of $\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$.

Theorem 4.4.2. *Let $(A,B), \tilde{\mathcal{X}}_1, \tilde{V}_1, \tilde{\mathcal{Y}}_1, \tilde{U}_1$, R_A, R_B be as in Theorem 4.4.1. Then the backward error $\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ can be expressed by*

$$\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \mu \begin{pmatrix} \|R_A\| \\ \omega \|R_B\| \end{pmatrix}. \tag{4.4.9}$$

The expression (4.4.9) shows that the backward error $\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ defined by (4.4.2) is independent of the choice of the matrices $\tilde{V}_1$ and $\tilde{U}_1$ whose column vectors form orthonormal bases of $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{Y}}_1$, respectively.

Proof of Theorem 4.4.2. From the definition (4.4.2) and the proof of Theorem 4.4.1 we see that

$$\beta^{(\omega)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \min_{\substack{A_1, B_1 \in \mathcal{C}^{l \times l} \\ Z, W \in \mathcal{C}^{n \times n}}} \mu \begin{pmatrix} \|(\tilde{U}_1 A_1 - A\tilde{V}_1)\tilde{V}_1^H + Z(I - \tilde{V}_1 \tilde{V}_1^H)\| \\ \omega \|(\tilde{U}_1 B_1 - B\tilde{V}_1)\tilde{V}_1^H + W(I - \tilde{V}_1 \tilde{V}_1^H)\| \end{pmatrix}. \tag{4.4.10}$$

Observe the following facts: (i) By the proof of Theorem 4.4.1, we have

$$\begin{aligned} \|(\tilde{U}_1 A_1 - A\tilde{V}_1)\tilde{V}_1^H + Z(I - \tilde{V}_1 \tilde{V}_1^H)\| &\geq \|R_A \tilde{V}_1^H\|, \\ \|(\tilde{U}_1 B_1 - B\tilde{V}_1)\tilde{V}_1^H + W(I - \tilde{V}_1 \tilde{V}_1^H)\| &\geq \|R_B \tilde{V}_1^H\|; \end{aligned} \tag{4.4.11}$$

(ii) The equalities in (4.4.11) are achieved when A_1, B_1, Z, W satisfy

$$A_1 = \tilde{U}_1^H A \tilde{V}_1, \quad B_1 = \tilde{U}_1^H B \tilde{V}_1, \quad Z\tilde{V}_2 = W\tilde{V}_2 = 0;$$

(iii) By the hypothesis $\mu(\cdot)$ is an absolute norm. (iv) From

$$\sigma_+(R_A \tilde{V}_1^H) = \sigma_+(R_A), \quad \sigma_+(R_B \tilde{V}_1^H) = \sigma_+(R_B)$$

it follows that

$$\|R_A\tilde{V}_1^H\| = \|R_A\|, \quad \|R_B\tilde{V}_1^H\| = \|R_B\|.$$

Hence, from (4.4.10) we obtain (4.4.9). □.

We now define the relative backward errors $\eta^*_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ and $\eta_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ of (A, B) with respect to $\tilde{\mathcal{X}}_1$, $\tilde{\mathcal{Y}}_1$ by

$$\eta^*_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{G}} \frac{\left\| \begin{pmatrix} E \\ F \end{pmatrix} \right\|_F}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F},$$

and

$$\eta_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{G}} \left\| \begin{pmatrix} \|E\|_F/\|A\|_F \\ \|F\|_F/\|B\|_F \end{pmatrix} \right\|_2,$$

where $\mathcal{G}$ is the set defined by (4.4.3). From (4.4.1), (4.4.2), (4.4.5) and (4.4.9) we get the computable formulas

$$\eta^*_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \frac{1}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F} \eta^{(1)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \frac{\left\| \begin{pmatrix} R_A \\ R_B \end{pmatrix} \right\|_F}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F}, \tag{4.4.12}$$

and

$$\eta_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \frac{1}{\|A\|_F} \beta^{(\|A\|_F/\|B\|_F)}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \left\| \begin{pmatrix} \|R_A\|_F/\|A\|_F \\ \|R_B\|_F/\|B\|_F \end{pmatrix} \right\|_2. \tag{4.4.13}$$

Remark 4.4.3. Let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ approximate an 1-dimensional deflating subspace pair of (A, B), where $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{v}_1), \tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{u}_1)$, and $\tilde{v}_1, \tilde{u}_1$ are unit vectors. By the formulas (4.4.12) and (4.4.13), the relative backward errors $\eta^*_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ and $\eta_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1)$ of (A, B) with respect to $\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1$ can be expressed by

$$\eta^*_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \frac{\sqrt{\|r_A\|_2^2 + \|r_B\|_2^2}}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F}, \quad \eta_{\mathrm{rel}}(\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1) = \left\| \begin{pmatrix} \|r_A\|_2/\|A\|_F \\ \|r_B\|_2/\|B\|_F \end{pmatrix} \right\|_2,$$

where

$$r_A = (\tilde{u}_1^H A \tilde{v}_1)\tilde{u}_1 - A\tilde{v}_1, \quad r_B = (\tilde{u}_1^H B \tilde{v}_1)\tilde{u}_1 - B\tilde{v}_1$$

are the residuals. Moreover, the optimal backward perturbation is

$$\begin{pmatrix} E_{\mathrm{opt}} \\ F_{\mathrm{opt}} \end{pmatrix} = \begin{pmatrix} r_A \\ r_B \end{pmatrix} \tilde{v}_1^H.$$

4.4.1.2 The Backward Errors $\eta^{(\theta)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ and $\beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$

Let (A, B) be an $n \times n$ regular pair, and let $\mathcal{X}_1, \mathcal{Y}_1$ be l-dimensional subspaces of $\mathcal{C}^n$. It is known that the pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is a deflating subspace pair of (A, B) if and only if there are matrices $X_1, Y_1 \in \mathcal{C}^{n\times l}$ and $A_1, B_1 \in \mathcal{C}^{l\times l}$ such that

$$\mathcal{X}_1 = \mathcal{R}(X_1), \ \ \mathcal{Y}_1 = \mathcal{R}(Y_1), \quad \text{and} \quad AX_1 = Y_1A_1, \ \ BX_1 = Y_1B_1, \tag{4.4.14}$$

where (A_1, B_1) is a regular pair.

The matrix pair (A_1, B_1) may be called the *eigenmatrix pair* of (A, B) associated with X_1, Y_1.

Let $\tilde{X}_1, \tilde{Y}_1 \in \mathcal{C}^{n\times l}$, $\text{rank}(\tilde{X}_1) = \text{rank}(\tilde{Y}_1) = l$, and let $(\tilde{A}_1, \tilde{B}_1)$ be an $l \times l$ regular pair. Moreover, let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ with $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{Y}_1)$ approximate a deflating subspace pair of (A, B), and $(\tilde{A}_1, \tilde{B}_1)$ be the associated eigenmatrix pair. By §1.9, we define the backward errors $\eta^{(\theta)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ and $\beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ of (A, B) with respect to $\tilde{X}_1, \tilde{Y}_1$ and $(\tilde{A}_1, \tilde{B}_1)$ by

$$\eta^{(\theta)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{K}} \left\| \begin{pmatrix} E \\ \theta F \end{pmatrix} \right\|, \tag{4.4.15}$$

and

$$\beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{K}} \mu \begin{pmatrix} \|E\| \\ \omega \|F\| \end{pmatrix}, \tag{4.4.16}$$

where θ, ω are positive parameters, $\mu(\cdot)$ is any absolute norm on $\mathcal{R}^2$, and the set $\mathcal{K}$ is defined by

$$\mathcal{K} = \left\{ \begin{pmatrix} E \\ F \end{pmatrix} \ : \ E, F \in \mathcal{C}^{n\times n}, \ (A + E)\tilde{X}_1 = \tilde{Y}_1\tilde{A}_1, \ (B + F)\tilde{X}_1 = \tilde{Y}_1\tilde{B}_1 \right\}. \tag{4.4.17}$$

The following result gives a computable formula of $\eta^{(\theta)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$.

Theorem 4.4.4. *Let*

$$R_A = \tilde{Y}_1\tilde{A}_1 - A\tilde{X}_1, \qquad R_B = \tilde{Y}_1\tilde{B}_1 - B\tilde{X}_1 \tag{4.4.18}$$

be the residual of (A, B) with respect to $\tilde{X}_1, \tilde{Y}_1$ and $(\tilde{A}_1, \tilde{B}_1)$. Then the backward error $\eta^{(\theta)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ can be expressed by

$$\eta^{(\theta)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1) = \left\| \begin{pmatrix} R_A \\ \theta R_B \end{pmatrix} \tilde{X}_1^{\dagger} \right\|. \tag{4.4.19}$$

Proof. From (4.4.17) it follows that a matrix $\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{K}$ if and only if $\begin{pmatrix} E \\ F \end{pmatrix}$ satisfies

$$\begin{pmatrix} E \\ F \end{pmatrix} \tilde{X}_1 = \begin{pmatrix} R_A \\ R_B \end{pmatrix}, \tag{4.4.20}$$

where R_A and R_B are the residuals defined by (4.4.18).

Applying Theorem 1.5.1 to the equation (4.4.20) we see that the equation is solvable, and any solution $\begin{pmatrix} E \\ F \end{pmatrix}$ of the equation can be expressed by

$$\begin{pmatrix} E \\ F \end{pmatrix} = \begin{pmatrix} R_A \\ R_B \end{pmatrix} \tilde{X}_1^\dagger + \begin{pmatrix} Z \\ W \end{pmatrix} (I - \tilde{X}_1 \tilde{X}_1^\dagger), \tag{4.4.21}$$

where $Z, W \in \mathcal{C}^{n\times n}$.

Take an orthogonal decomposition $\tilde{X}_1 = \tilde{U}_1 L$, where $\tilde{U}_1 \in \mathcal{U}^{n\times l}$ and $L \in \mathcal{C}^{l\times l}$. Further, choose $\tilde{U}_2$ so that $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. Then from (4.4.21)

$$\begin{pmatrix} E \\ \theta F \end{pmatrix} = \left(\begin{pmatrix} R_A \\ \theta R_B \end{pmatrix} L^{-1}, \begin{pmatrix} Z \\ \theta W \end{pmatrix} \tilde{U}_2 \right) \tilde{U}^H.$$

By the definition (4.4.15) and Theorem 1.2.1, we have

$$\eta(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1) = \left\| \begin{pmatrix} E_{\text{opt}} \\ \theta F_{\text{opt}} \end{pmatrix} \right\|$$

with

$$\begin{pmatrix} E_{\text{opt}} \\ F_{\text{opt}} \end{pmatrix} = \begin{pmatrix} R_A \\ R_B \end{pmatrix} L^{-1} \tilde{U}_1^H = \begin{pmatrix} R_A \\ R_B \end{pmatrix} \tilde{X}_1^\dagger,$$

which shows (4.4.19). □

The following result gives a computable formula of $\beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$.

Theorem 4.4.5. *Let $(A, B), \tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1$, R_A, R_B be as in Theorem 4.4.4. Then the backward error $\beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ can be expressed by*

$$\beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1) = \mu \begin{pmatrix} \|R_A \tilde{X}_1^\dagger\| \\ \omega \|R_B \tilde{X}_1^\dagger\| \end{pmatrix}. \tag{4.4.22}$$

Proof. From the definition (4.4.16) and the proofs of Theorems 4.4.4 and 4.4.2 we get

$$\begin{aligned} \beta^{(\omega)}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1) &= \min_{Z, W \in \mathcal{C}^{n\times n}} \mu \begin{pmatrix} \|(R_A L^{-1},\ Z\tilde{U}_2)\| \\ \omega \|(R_B L^{-1},\ W\tilde{U}_2)\| \end{pmatrix} \\ &= \mu \begin{pmatrix} \|R_A L^{-1}\| \\ \omega \|R_B L^{-1}\| \end{pmatrix} = \mu \begin{pmatrix} \|R_A \tilde{X}_1^\dagger\| \\ \omega \|R_B \tilde{X}_1^\dagger\| \end{pmatrix}. \end{aligned}$$

The proof is completed. $\square$

We now define the relative backward errors $\eta^*_{\mathrm{rel}}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1)$ and $\eta_{\mathrm{rel}}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1)$ of (A, B) with respect to $\tilde X_1$, $\tilde Y_1$ and $(\tilde A_1, \tilde B_1)$ by

$$\eta^*_{\mathrm{rel}}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{K}} \frac{\left\| \begin{pmatrix} E \\ F \end{pmatrix} \right\|_F}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F},$$

and

$$\eta_{\mathrm{rel}}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1) = \min_{\begin{pmatrix} E \\ F \end{pmatrix} \in \mathcal{K}} \left\| \begin{pmatrix} \|E\|_F/\|A\|_F \\ \|F\|_F/\|B\|_F \end{pmatrix} \right\|_2,$$

where $\mathcal{K}$ is the set defined by (4.4.17). From (4.4.19) and (4.4.22) we get the computable formulas

$$\eta^*_{\mathrm{rel}}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1) = \frac{1}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F} \eta^{(1)}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1) = \frac{\left\| \begin{pmatrix} R_A \\ R_B \end{pmatrix} \tilde X_1^\dagger \right\|_F}{\left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_F}, \quad (4.4.23)$$

and

$$\eta_{\mathrm{rel}}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1) = \frac{1}{\|A\|_F} \beta^{(\|A\|_F/\|B\|_F)}(\tilde X_1, \tilde Y_1, \tilde A_1, \tilde B_1) = \left\| \begin{pmatrix} \|R_A \tilde X_1^\dagger\|_F/\|A\|_F \\ \|R_B \tilde X_1^\dagger\|_F/\|B\|_F \end{pmatrix} \right\|_2. \quad (4.4.24)$$

Example 4.4.6. Consider the regular pair (A, B) with

$$A = \begin{pmatrix} 15 & 70.0 & 79.96 & -20.001 & -60.0000 \\ 0 & 0.7 & 39.92 & 0.000 & 19.9998 \\ 30 & 139.3 & 120.00 & -40.001 & -120.0006 \\ -15 & -70.0 & -79.92 & 9.999 & -20.0002 \\ 300 & 2.1 & 120.04 & -9.998 & -0.0002 \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0.1 & 1.00 & 1.999 & 2.0001 & 3.00000 \\ 0.0 & 0.01 & 0.998 & 0.0000 & -0.99999 \\ 0.2 & 1.99 & 3.000 & 4.0001 & 6.00003 \\ -0.1 & -1.00 & -1.998 & -0.9999 & 1.00001 \\ 2.0 & 0.03 & 3.001 & 0.9998 & 0.00001 \end{pmatrix},$$

where B is nonsingular. Using the MATLAB file "qz" (which is an implementation of the QZ method) to the pair (A, B), we get the computed results:

$$A\tilde X_1 \approx \tilde Y_1 \tilde A_1, \qquad B\tilde X_1 \approx \tilde Y_1 \tilde B_1,$$

where $\tilde{X}_1, \tilde{Y}_1 \in \mathcal{R}^{5\times k}$, $\tilde{A}_1, \tilde{B}_1 \in \mathcal{R}^{k\times k}$, $k = 1,2,3,4,5$. By (4.4.23) and (4.4.24) we compute $\eta^*_{\rm rel}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ and $\eta_{\rm rel}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$ which are listed in Table 4.1.

Table 4.1

k	$\eta^*_{\rm rel}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$	$\eta_{\rm rel}(\tilde{X}_1, \tilde{Y}_1, \tilde{A}_1, \tilde{B}_1)$
1	1.34×10^{-16}	1.35×10^{-16}
2	3.39×10^{-16}	3.49×10^{-15}
3	3.88×10^{-16}	4.51×10^{-16}
4	3.94×10^{-16}	4.97×10^{-16}
5	4.41×10^{-16}	6.46×10^{-16}

The results listed in Table 4.1 show that each computed $\{\mathcal{R}(\tilde{X}_1), \mathcal{R}(\tilde{Y}_1)\}$ and associated $(\tilde{A}_1, \tilde{B}_1)$ by applying the MATLAB file "qz" are an exact deflating subspace pair and an associated eigenmatrix pair of a very slightly perturbed matrix pair of (A, B); in other words, the computation has proceeded quite stably.

4.4.1.3 The Backward Error $\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1)$

Let (A, B) be an $n \times n$ regular pair. By the definition introduced in §4.1.2, an l-dimensional subspace $\mathcal{X}_1$ is called an *eigenspace* of (A, B) if there is an l-dimensional subspace $\mathcal{Y}_1$ such that

$$A\mathcal{X}_1 \subset \mathcal{Y}_1, \qquad B\mathcal{X}_1 \subset \mathcal{Y}_1.$$

Let $\mathcal{X}_1 = \mathcal{R}(X_1) \subset \mathcal{C}^n$, where $X_1 \in \mathcal{C}^{n\times l}$ and $\text{rank}(X_1) = l$. It is known (see Stewart and Sun [125, Chapter VI, Theorem 2.10]) that the subspace $\mathcal{X}_1$ is an eigenspace of (A, B) if and only if there is an $l \times l$ regular pair (A_1, B_1) such that

$$AX_1B_1 = BX_1A_1. \tag{4.4.25}$$

The matrix pair (A_1, B_1) may be called an *eigenmatrix pair* of (A, B) associated with X_1.

Let $\tilde{X}_1 \in \mathcal{C}^{n\times l}$ and $\tilde{A}_1, \tilde{B}_1 \in \mathcal{C}^{l\times l}$ be given, where $\text{rank}(\tilde{X}_1) = l$, and the pair $(\tilde{A}_1, \tilde{B}_1)$ is regular. Moreover, let $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ approximate an eigenspace of (A, B), and $(\tilde{A}_1, \tilde{B}_1)$ be an associated eigenmatrix pair. By §1.9, we define the backward error $\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1)$ of (A, B) with respect to $\tilde{X}_1$ and $(\tilde{A}_1, \tilde{B}_1)$ by

$$\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1) = \min_{(E,F)\in\mathcal{L}} \|(E, \theta F)\|, \tag{4.4.26}$$

where the set $\mathcal{L}$ is defined by

$$\mathcal{L} = \left\{(E, F) \ : \ E, F \in \mathcal{C}^{n\times n}, \ (A + E)\tilde{X}_1\tilde{B}_1 = (B + F)\tilde{X}_1\tilde{A}_1\right\}. \tag{4.4.27}$$

The following result gives a computable formula of $\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1)$.

Theorem 4.4.7. *Let*

$$R = B\tilde{X}_1\tilde{A}_1 - A\tilde{X}_1\tilde{B}_1 \tag{4.4.28}$$

be the residual of (A, B) *with respect to* $\tilde{X}_1$ *and* $(\tilde{A}_1, \tilde{B}_1)$. *Then the backward error* $\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1)$ *can be expressed by*

$$\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1) = \left\| R \begin{pmatrix} \tilde{X}_1\tilde{B}_1 \\ -\frac{1}{\theta}\tilde{X}_1\tilde{A}_1 \end{pmatrix}^{\dagger} \right\|. \tag{4.4.29}$$

Proof. From (4.4.27) it follows that a matrix pair $(E, F) \in \mathcal{L}$ if and only if (E, F) satisfies

$$(E, \theta F) \begin{pmatrix} \tilde{X}_1\tilde{B}_1 \\ -\frac{1}{\theta}\tilde{X}_1\tilde{A}_1 \end{pmatrix} = R, \tag{4.4.30}$$

where R is the residual defined by (4.4.28).

By the hypothesis the matrix pair $(\tilde{A}_1, \tilde{B}_1)$ is regular, so we have rank $\begin{pmatrix} \tilde{A}_1 \\ \tilde{B}_1 \end{pmatrix} = l$. Applying Theorem 1.5.1 to the equation (4.4.30) we see that the equation is solvable, and any solution $(E, \theta F)$ to the equation can be expressed by

$$(E, \theta F) = R \begin{pmatrix} \tilde{X}_1\tilde{B}_1 \\ -\frac{1}{\theta}\tilde{X}_1\tilde{A}_1 \end{pmatrix}^{\dagger} + Z\left(I - \begin{pmatrix} \tilde{X}_1\tilde{B}_1 \\ -\frac{1}{\theta}\tilde{X}_1\tilde{A}_1 \end{pmatrix} \begin{pmatrix} \tilde{X}_1\tilde{B}_1 \\ -\frac{1}{\theta}\tilde{X}_1\tilde{A}_1 \end{pmatrix}^{\dagger}\right), \tag{4.4.31}$$

where $Z \in \mathcal{C}^{n\times 2n}$.

By the definition (4.4.26) and Theorem 1.2.1, from (4.4.31) we obtain

$$\eta^{(\theta)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1) = \|(E_{\text{opt}},\ \theta F_{\text{opt}})\|$$

with

$$(E_{\text{opt}},\ F_{\text{opt}}) = R \begin{pmatrix} \tilde{X}_1\tilde{B}_1 \\ -\frac{1}{\theta}\tilde{X}_1\tilde{A}_1 \end{pmatrix}^{\dagger} \begin{pmatrix} I & 0 \\ 0 & \frac{1}{\theta} \end{pmatrix},$$

which shows (4.4.29). □.

We now define the relative backward error $\eta_{\text{rel}}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1)$ of (A, B) with respect to $\tilde{X}_1$ and $(\tilde{A}_1, \tilde{B}_1)$ by

$$\eta_{\text{rel}}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1) = \min_{(E,F)\in\mathcal{L}} \left\| \left(\frac{\|E\|_F}{\|A\|_F}, \frac{\|F\|_F}{\|B\|_F} \right) \right\|_2,$$

where $\mathcal{L}$ is the set defined by (4.4.27). Obviously, if we take $\|\cdot\| = \|\cdot\|_F$ in (4.4.26), then from (4.4.29) we get a computable formula of $\eta_{\text{rel}}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1)$:

$$\eta_{\text{rel}}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1) = \frac{1}{\|A\|_F}\eta^{(\|A\|_F/\|B\|_F)}(\tilde{X}_1, \tilde{A}_1, \tilde{B}_1) = \left\| R \begin{pmatrix} \|A\|_F\tilde{X}_1\tilde{B}_1 \\ -\|B\|_F\tilde{X}_1\tilde{A}_1 \end{pmatrix}^{\dagger} \right\|_F. \tag{4.4.32}$$

Remark 4.4.8. Let $(\tilde{\alpha}_1, \tilde{\beta}_1)$ be an approximate eigenvalue of (A, B), and $\tilde{x}_1$ be an associated eigenvector. Then by (4.4.32), the relative backward error $\eta_{\rm rel}(\tilde{x}_1, \tilde{\alpha}_1, \tilde{\beta}_1)$ of (A, B) with respect to $\tilde{x}_1$ and $(\tilde{\alpha}_1, \tilde{\beta}_1)$ can be expressed by

$$\eta_{\rm rel}(\tilde{x}_1, \tilde{\alpha}_1, \tilde{\beta}_1) = \frac{1}{\sqrt{|\tilde{\alpha}_1|^2\|B\|_F^2 + |\tilde{\beta}_1|^2\|A\|_F^2}} \frac{\|r\|_2}{\|\tilde{x}_1\|_2}, \tag{4.4.33}$$

where

$$r = \tilde{\alpha}_1 B\tilde{x}_1 - \tilde{\beta}_1 A\tilde{x}_1$$

is the residual. Moreover, the optimal backward perturbation $(E_{\rm opt}, F_{\rm opt})$ in (A, B) is expressed by

$$(E_{\rm opt}, F_{\rm opt}) = \frac{r\left(\|A\|_F^2(\tilde{\beta}_1\tilde{x}_1)^H,\ -\|B\|_F^2(\tilde{\alpha}_1\tilde{x}_1)^H\right)}{\left(|\tilde{\alpha}_1|^2\|B\|_F^2 + |\tilde{\beta}_1|^2\|A\|_F^2\right)\|\tilde{x}_1\|_2^2}.$$

The formula (4.4.33) will be illustrated by the following example.

Example 4.4.9. Consider the regular pair (A, B) of Example 4.4.6. The eigenvalues of (A, B) are $150, 70, 40, -10, -20$, and the associated eigenvectors are $e_1^{(5)}, e_2^{(5)}, e_3^{(5)}, e_4^{(5)}, e_5^{(5)}$, the columns of the identity matrix I_5, respectively. Using the MATLAB file "qz" to the pair (A, B), we obtain the computed eigenvalues $(\tilde{\alpha}_j, \tilde{\beta}_j)$ and associated eigenvectors $\tilde{x}_j$; and then applying (4.4.33) we get

$$\begin{aligned} &\eta_{\rm rel}(\tilde{x}_1, \tilde{\alpha}_1, \tilde{\beta}_1) \approx 2.69 \times 10^{-18}, \qquad \eta_{\rm rel}(\tilde{x}_2, \tilde{\alpha}_2, \tilde{\beta}_2) \approx 7.72 \times 10^{-17}, \\ &\eta_{\rm rel}(\tilde{x}_3, \tilde{\alpha}_3, \tilde{\beta}_3) \approx 9.70 \times 10^{-17}, \qquad \eta_{\rm rel}(\tilde{x}_4, \tilde{\alpha}_4, \tilde{\beta}_4) \approx 6.30 \times 10^{-17}, \\ &\eta_{\rm rel}(\tilde{x}_5, \tilde{\alpha}_5, \tilde{\beta}_5) \approx 1.55 \times 10^{-16}. \end{aligned} \tag{4.4.34}$$

From (4.4.34) we see that each computed eigenvalue $(\tilde{\alpha}_j, \tilde{\beta}_j)$ and associated eigenvector $\tilde{x}_j$ are an exact eigenvalue and an associated eigenvector of a very slightly perturbed matrix pair of (A, B); in other words, the computation has proceeded quite stably.

4.4.2 Residual Bounds

Let an l-dimensional simple approximate deflating subspace pair $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$, $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$ of (A, B) be given, where $\tilde{V}_1, \tilde{U}_1 \in \mathcal{U}^{n\times l}$. Then by using Theorem 4.4.1 and an appropriate forward perturbation result we can determine the accuracy of the approximate deflating subspaces $\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1$.

Choose $\tilde{V}_2, \tilde{U}_2$ so that $\tilde{V} = (\tilde{V}_1, \tilde{V}_2)$, $\tilde{U} = (\tilde{U}_1, \tilde{U}_2) \in \mathcal{U}^{n\times n}$. By the proof of Theorem 4.4.1, the optimal backward perturbation $(E_{\rm opt}, F_{\rm opt})$ of (4.4.8) satisfies

$$\begin{aligned} \tilde{U}^H(A + E_{\rm opt})\tilde{V} &= \begin{pmatrix} \tilde{U}_1^H A\tilde{V}_1 & \tilde{U}_1^H A\tilde{V}_2 \\ 0 & \tilde{U}_2^H A\tilde{V}_2 \end{pmatrix} \equiv \begin{pmatrix} \tilde{A}_{11} & -S_A\tilde{V}_2 \\ 0 & \tilde{A}_{22} \end{pmatrix}, \\ \tilde{U}^H(B + F_{\rm opt})\tilde{V} &= \begin{pmatrix} \tilde{U}_1^H B\tilde{V}_1 & \tilde{U}_1^H B\tilde{V}_2 \\ 0 & \tilde{U}_2^H B\tilde{V}_2 \end{pmatrix} \equiv \begin{pmatrix} \tilde{B}_{11} & -S_B\tilde{V}_2 \\ 0 & \tilde{B}_{22} \end{pmatrix}, \end{aligned} \tag{4.4.35}$$

and

$$\tilde{U}^H E_{\text{opt}} \tilde{V} = \begin{pmatrix} 0 & 0 \\ \tilde{U}_2^H R_A & 0 \end{pmatrix}, \quad \tilde{U}^H F_{\text{opt}} \tilde{V} = \begin{pmatrix} 0 & 0 \\ \tilde{U}_2^H R_B & 0 \end{pmatrix}, \tag{4.4.36}$$

where R_A and R_B are the residuals defined by (4.4.4), and S_A and S_B are the residuals of (A, B) with respect to $\tilde{V}_1^H$ and $\tilde{U}_1^H$ defined by

$$S_A = (\tilde{U}_1^H A \tilde{V}_1)\tilde{V}_1^H - \tilde{U}_1^H A, \qquad S_B = (\tilde{U}_1^H B \tilde{V}_1)\tilde{V}_1^H - \tilde{U}_1^H B.$$

The relation (4.4.35) shows that the subspace pair $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ is a deflating subspace pair of $(A + E_{\text{opt}}, B + F_{\text{opt}})$. Moreover, if

$$\lambda(\tilde{A}_{11}, \tilde{B}_{11}) \bigcap \lambda(\tilde{A}_{22}, \tilde{B}_{22}) = \emptyset, \tag{4.4.37}$$

then $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ is a *simple* deflating subspace pair of $(A + E_{\text{opt}}, B + F_{\text{opt}})$.

The following result gives residual bounds for the approximate deflating subspaces $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{Y}}_1$. The result is obtained by applying Theorem 4.3.3 to the matrix pairs $(A + E_{\text{opt}}, B + F_{\text{opt}})$ and (A, B) of (4.4.35) and (4.4.36).

Theorem 4.4.10. *Let $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ be an approximate simple deflating subspace pair of (A, B), where $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{V}_1)$, $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{U}_1)$, and $\tilde{V}_1, \tilde{U}_1 \in \mathcal{U}^{n\times l}$. Define the matrices $\tilde{A}_{jj}$ and $\tilde{B}_{jj}$ by (4.4.35), and assume (4.4.37) is satisfied. Define the residuals R_A, R_B, S_A, S_B by*

$$R_A = \tilde{U}_1 \tilde{A}_{11} - A\tilde{V}_1, \qquad R_B = \tilde{U}_1 \tilde{B}_{11} - B\tilde{V}_1,$$

$$S_A = \tilde{A}_{11}\tilde{V}_1^H - \tilde{U}_1^H A, \qquad S_B = \tilde{B}_{11}\tilde{V}_1^H - \tilde{U}_1^H B,$$

and define the matrices $\tilde{C}_1, \tilde{C}_2$ by

$$\begin{aligned} \tilde{C}_1 &= \left((\tilde{B}_{11}^T \otimes I_{n-l})\tilde{M}^{-1},\ (-\tilde{A}_{11}^T \otimes I_{n-l})\tilde{M}^{-1}\right), \\ \tilde{C}_2 &= \left((I_l \otimes \tilde{B}_{22})\tilde{M}^{-1},\ (-I_l \otimes \tilde{A}_{22})\tilde{M}^{-1}\right), \end{aligned} \tag{4.4.38}$$

where

$$\tilde{M} = \tilde{A}_{11}^T \otimes \tilde{B}_{22} - \tilde{B}_{11}^T \otimes \tilde{A}_{22}.$$

Moreover, let

$$\begin{aligned} \tilde{b}_1 &= \left\| \tilde{C}_1 \begin{pmatrix} \text{vec}(\tilde{U}_2^H R_A) \\ \text{vec}(\tilde{U}_2^H R_B) \end{pmatrix} \right\|_2, \qquad \tilde{c}_1 = \|\tilde{C}_1\|_2, \\ \tilde{b}_2 &= \left\| \tilde{C}_2 \begin{pmatrix} \text{vec}(\tilde{U}_2^H R_A) \\ \text{vec}(\tilde{U}_2^H R_B) \end{pmatrix} \right\|_2, \qquad \tilde{c}_2 = \|\tilde{C}_2\|_2, \\ \tilde{b} &= \tilde{b}_1 + \tilde{b}_2, \qquad \tilde{c} = \tilde{c}_1 + \tilde{c}_2, \end{aligned} \tag{4.4.39}$$

and define $\tilde{\eta}$ *by*

$$\tilde{\eta} = \max\{\|S_A\|_2,\ \|S_B\|_2\}. \tag{4.4.40}$$

If

$$4\tilde{b}\tilde{c}\tilde{\eta} < 1,$$

Then there is a unique pair of deflating subspaces $\mathcal{X}_1 = \mathcal{R}(V_1)$ *and* $\mathcal{Y}_1 = \mathcal{R}(U_1)$ *of* (A, B) *such that* $\tilde{V}_1, \tilde{U}_1 \in \mathcal{U}^{n\times l}$*, and*

$$\begin{aligned} \rho_F(\mathcal{X}_1, \tilde{\mathcal{X}}_1) &\le \|\tan\Theta(V_1, \tilde{V}_1)\|_F \le \tilde{b}_1 + \tilde{c}_1\tilde{\eta}\tilde{\beta}^2 \equiv \tau_{\mathcal{X}_1}, \\ \rho_F(\mathcal{Y}_1, \tilde{\mathcal{Y}}_1) &\le \|\tan\Theta(U_1, \tilde{U}_1)\|_F \le \tilde{b}_2 + \tilde{c}_2\tilde{\eta}\tilde{\beta}^2 \equiv \tau_{\mathcal{Y}_1}, \end{aligned} \tag{4.4.41}$$

where

$$\tilde{\beta} = \frac{2\tilde{b}}{1 + \sqrt{1 - 4\tilde{b}\tilde{c}\tilde{\eta}}}. \tag{4.4.42}$$

By the way, the relation (4.4.35) shows that the eigenvalues $(\tilde{\alpha}_1, \tilde{\beta}_1), \ldots, (\tilde{\alpha}_l, \tilde{\beta}_l)$ of $(\tilde{A}_{11}, \tilde{B}_{11})$, as l approximate eigenvalues of (A, B), are l eigenvalues of $(A + E_{\text{opt}}, B + F_{\text{opt}})$. How to obtain a sharp error bound for the approximate eigenvalues $(\tilde{\alpha}_1, \tilde{\beta}_1), \ldots, (\tilde{\alpha}_l, \tilde{\beta}_l)$ is a research problem.

Example 4.4.11. Consider the matrix pair (A, B) with

$$A = \begin{pmatrix} 0 & 17 & 20 & -15 & 18 \\ -6 & -2 & 17 & 7 & 4 \\ 0 & 3 & -1 & -11 & 5 \\ 0 & 0 & 2 & -1 & 7 \\ 0 & 0 & 0 & -4 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 & -1 & 4 & 2 \\ -3 & 2 & 5 & -6 & -2 \\ 0 & 1 & 5 & 1 & 13 \\ 0 & 0 & 2 & -6 & 12 \\ 0 & 0 & 0 & -4 & 4 \end{pmatrix},$$

and let

$$x_1 = (1.75,\ -0.25,\ -0.75,\ -0.25,\ -0.25)^T, \quad y_1 = (1,\ 0,\ 0,\ 0,\ 0)^T,$$

$$v_1 = x_1/\|x_1\|_2, \quad u_1 = y_1, \quad \mathcal{X}_1 = \mathcal{R}(v_1), \quad \mathcal{Y}_1 = \mathcal{R}(u_1).$$

The 1-dimensional subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ is a deflating subspace pair of (A, B) corresponding to the eigenvalue $\lambda_1 = 10$. Suppose that we have an approximate deflating subspace pair $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{Y}}_1\}$ with $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{x}_1)$ and $\tilde{\mathcal{Y}}_1 = \mathcal{R}(\tilde{y}_1)$, where

$$\tilde{x}_1 = (1.74999,\ -0.24999,\ 0.7499999,\ -0.24999,\ -0.250001)^T,$$

$$\tilde{y}_1 = (1,\ -1.0 \times 10^{-7},\ 1.0 \times 10^{-5},\ -1.0 \times 10^{-6},\ 1.0 \times 10^{-5})^T.$$

Let

$$\tilde{v}_1 = \tilde{x}_1/\|\tilde{x}_1\|_2, \quad \tilde{u}_1 = \tilde{y}_1/\|\tilde{y}_1\|_2.$$

A calculation gives

$$\sin\theta(v_1, \tilde{v}_1) \approx 5.0260 \times 10^{-6}, \quad \sin\theta(u_1, \tilde{u}_1) \approx 1.4178 \times 10^{-5}. \tag{4.4.43}$$

Choose $\tilde{V}_2$ and $\tilde{U}_2$ so that $(\tilde{v}_1, \tilde{V}_2), (\tilde{u}_1, \tilde{U}_2) \in \mathcal{O}^{5\times 5}$. Compute

$$\tilde{A}_{11} = \tilde{u}_1^T A \tilde{v}_1, \qquad \tilde{A}_{22} = \tilde{U}_2^T A \tilde{V}_2,$$

$$\tilde{B}_{11} = \tilde{u}_1^T B \tilde{v}_1, \qquad \tilde{B}_{22} = \tilde{U}_2^T B \tilde{V}_2,$$

and

$$r_A = \tilde{u}_1 \tilde{A}_{11} - A\tilde{v}_1, \qquad r_B = \tilde{u}_1 \tilde{B}_{11} - B\tilde{v}_1,$$

$$s_A = \tilde{A}_{11}\tilde{v}_1^T - \tilde{u}_1^T A, \qquad s_B = \tilde{B}_{11}\tilde{v}_1^T - \tilde{u}_1^T B,$$

and compute $\tilde{C}_1, \tilde{C}_2$, $\tilde{b}_1, \tilde{c}_1, \tilde{b}_2, \tilde{c}_2$, $\tilde{b}, \tilde{c}$ and $\tilde{\eta}$ by (4.4.38)–(4.4.40). A calculation shows that

$$4\tilde{b}\tilde{c}\tilde{\eta} \approx 7.1745 \times 10^{-2} < 1.$$

Consequently, applying Theorem 4.4.10, there are unit vectors v and u such that $\mathcal{R}(v)$ and $\mathcal{R}(u)$ are deflating subspaces of (A, B) corresponding to the same eigenvalue, and

$$\begin{aligned} \tan\theta(v, \tilde{v}_1) &\le \tau_{\mathcal{X}_1} \approx 5.1609 \times 10^{-6}, \\ \tan\theta(u, \tilde{u}_1) &\le \tau_{\mathcal{Y}_1} \approx 1.4401 \times 10^{-5}. \end{aligned} \tag{4.4.44}$$

Comparing (4.4.44) with (4.4.43) shows that the estimates obtained by applying Theorem 4.4.10 are fairly sharp.

Remark 4.4.12. Let $(\tilde{\alpha}, \tilde{\beta})$ be an approximate eigenvalue of (A, B), and $\tilde{x}$ be an associated eigenvector; i.e., $\tilde{\beta} A\tilde{x} \approx \tilde{\alpha} B\tilde{x}$. It may well be asked: How to determine the accuracy of the approximate solution? A similar result to Theorem 2.4.10 can be derived, but there is the same drawback as Theorem 2.4.10 that it needs to compute the Moore-Penrose inverse of an $n \times (n+1)$ matrix. Therefore, the problem of how to find nearly optimal residual bounds with less effort for computed generalized eigenvalues and eigenvectors is worth studying.

Notes and References

NR 4.4–1. Theorem 4.4.1 is proved by Sun [115].

NR 4.4–2. Cao [13] generalizes Theorem 2.4.5 to matrix pairs. Let (A, B) be an $n \times n$ regular pair. By [13], a subspace pair $\{\mathcal{X}_1, \mathcal{Y}_1\}$ with $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\mathcal{Y}_1 = \mathcal{R}(Y_1)$ is called an l-dimensional right deflation pair of (A, B) if there is an $l \times l$ regular pair (A_{11}, B_{11}) such that

$$AX_1 = Y_1 A_{11} \quad \text{and} \quad BX_1 = Y_1 B_{11};$$

a subspace pair $\{\mathcal{Z}_1, \mathcal{W}_1\}$ with $\mathcal{Z}_1 = \mathcal{R}(Z_1)$ and $\mathcal{W}_1 = \mathcal{R}(W_1)$ is called an l-dimensional left deflation pair of (A, B) if

$$Z_1^H A = A_{11} W_1^H \quad \text{and} \quad Z_1^H B = B_{11} W_1^H.$$

Here A_{11} and B_{11} are called Rayleigh components of (A, B). For a given regular pair (A, B), consider approximate right and left deflation pairs and the corresponding Rayleigh components. Cao [13] shows that under certain hypothesis these approximate quantities for (A, B) are accurate ones for a perturbation matrix pair $(A-E, B-F)$. Furthermore, bounds for $\|E\|_F$ and $\|F\|_F$ as well as $\|E\|_2$ and $\|F\|_2$ can be expressed in terms of the corresponding norms of residual matrices.

NR 4.4–3. Let (A, B) be a definite pair of order n, and Z_1 be an $n \times l$ matrix with full column rank whose column vectors span an approximate eigenspace of (A, B). Some relations between the eigenvalues of the Rayleigh quotient matrix pair $(Z_1^H A Z_1, Z_1^H B Z_1)$ and those of (A, B) are given by Li [69]. Residual bounds for the eigenvalues of $(Z_1^H A Z_1, Z_1^H B Z_1)$ and for the approximate eigenspace $\mathcal{R}(Z_1)$ are given by Sun [112].

NR 4.4–4. Let (A, B) be a regular pair, and let $\tilde{\lambda}$ and $\tilde{x}$ approximate a finite eigenvalue and associated eigenvector of (A, B). Frayssé and Toumazou [39] define the normwise backward error $\eta(\tilde{x}, \tilde{\lambda})$ and the *optimal* backward error $\eta_{\text{opt}}(\tilde{\lambda})$ by

$$\eta(\tilde{x}, \tilde{\lambda}) = \min\left\{\epsilon \ : \begin{array}{c} (A+E)\tilde{x} = \tilde{\lambda}(B+F)\tilde{x}, \\ \\ \|E\| \le \epsilon\alpha, \ \|F\| \le \epsilon\beta \end{array}\right\},$$

and

$$\eta_{\text{opt}}(\tilde{\lambda}) = \min\left\{\epsilon \ : \begin{array}{c} \exists u \ne 0, \ (A+E)u = \tilde{\lambda}(B+F)u, \\ \\ \|E\| \le \epsilon\alpha, \ \|F\| \le \epsilon\beta \end{array}\right\},$$

respectively, where α and β are positive parameters, and $\|\cdot\|$ is any vector norm and subordinate matrix norm. By [39], $\eta(\tilde{x}, \tilde{\lambda})$ can be expressed by

$$\eta(\tilde{x}, \tilde{\lambda}) = \frac{\|\tilde{\lambda}B\tilde{x} - A\tilde{x}\|}{(\alpha + |\tilde{\lambda}|\beta)\|\tilde{x}\|},$$

and $\eta_{\text{opt}}(\tilde{\lambda})$ can be expressed by

$$\eta_{\text{opt}}(\tilde{\lambda}) = \frac{1}{(\alpha + |\tilde{\lambda}|\beta)\left\|(A - \tilde{\lambda}B)^{-1}\right\|}.$$

These results are generalized by D. Higham and N. Higham [46] to any mixed subordinate matrix norm. Moreover, D. Higham and N. Higham [46] give some results on componentwise backward error and structured backward error for the generalized eigenvalue problem.

4.5 Symmetric-Definite Generalized Eigenproblems

In the generalized eigenvalue problem $\beta Ax = \alpha Bx$ it is frequently the case that $A, B \in \mathcal{S}^{n \times n}$ and B is positive definite. By Golub and Van Loan [52, Chapter 8], this problem is called the *symmetric-definite generalized eigenproblem.* Since the eigenvalues of this eigenproblem are finite, we can write the problem as

$$Ax = \lambda Bx \quad \text{with} \quad A, B \in \mathcal{S}^{n\times n} \text{ and } B > 0. \tag{4.5.1}$$

It is known that the eigenvalues of the problem (4.5.1) are real, and there is a matrix $X \in \mathcal{R}^{n\times n}$ such that

$$X^T AX = \Lambda, \qquad X^T BX = I, \tag{4.5.2}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the eigenvalue problem (4.5.1). (See Golub and Van Loan [41, §8.7.2] for numerical methods for computing the matrices X and Λ.)

In this section we investigate perturbation properties of multiple eigenvalues and associated eigenspaces of the symmetric-definite generalized eigenproblem (4.5.1).

4.5.1 Local Behavior of Multiple Eigenvalues

Let $p = (p_1, \dots, p_N)^T \in \mathcal{R}^N$. Suppose that $A(p), B(p) \in \mathcal{S}^{n\times n}$ are real analytic matrix-valued functions of p in some neighborhood $\mathcal{B}(p^*)$ of the point $p^* \in \mathcal{R}^N$ and $B(p) > 0$ for any $p \in \mathcal{B}(p^*)$. Without loss of generality we may assume that the point p^* is the origin of $\mathcal{R}^N$. The eigenproblem

$$A(p)x(p) = \lambda(p)B(p)x(p), \qquad p \in \mathcal{B}(0) \tag{4.5.3}$$

arises often in structural design, and it is often desirable to be able to estimate the sensitivity of the available designs $\lambda(p)$ to changes in the system parameters $p_1, \dots, p_N$.

If $\lambda_1 \in \mathcal{R}$ is a simple eigenvalue of the matrix pair $(A(0), B(0))$, then by using the same technique described in §4.1.1 we can prove that there is an analytic function $\lambda_1(p)$ in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin that is a simple eigenvalue of the matrix pair $(A(p), B(p))$, and $\lambda_1(0) = \lambda_1$. Moreover, we can derive the formulas of the partial derivatives of $\lambda_1(p)$ with respect to each p_j at $p = 0$.

However, if λ_1 is an eigenvalue of $(A(0), B(0))$ with multiplicity $r > 1$, then the situation becomes complicated. Rellich [86] first gives an example to show that the local behavior of a multiple eigenvalue is different from that of a simple eigenvalue for a symmetric eigenvalue problem depending analytically on several parameters.

The following example is a slight modification of the example given by Rellich [86] (or see Rellich [87, p.37]).

Example 4.5.1. Consider the matrix

$$A(p) = \begin{pmatrix} 1+2p_1+2p_2 & p_2 \\ p_2 & 1+2p_2 \end{pmatrix}, \qquad p = (p_1, p_2)^T \in \mathcal{R}^2.$$

Here we assume $B(p) = I$. It is easy to see that the elements of $A(p)$ are real analytic functions of $p \in \mathcal{R}^2$, the matrix $A(0)$ has the eigenvalue 1 with multiplicity 2, and the eigenvalues of $A(p)$ are

$$\lambda_1(p) = 1 + p_1 + 2p_2 + \sqrt{p_1^2 + p_2^2}, \qquad \lambda_2(p) = 1 + p_1 + 2p_2 - \sqrt{p_1^2 + p_2^2}.$$

Obviously no arrangement of these eigenvalues could make them analytic functions of p in some neighborhood of the origin, even no arrangement of these eigenvalues could make them differentiable at $p = 0$.

Let λ_1 be an eigenvalue of $(A(0), B(0))$ with multiplicity r. We shall prove in this subsection that there are r continuous functions $\lambda_1(p), \ldots, \lambda_r(p)$ in some neighborhood of the origin that are the eigenvalues of $(A(p), B(p))$ satisfying $\lambda_s(0) = \lambda_1$ for $s = 1, \ldots, r$, and every $\lambda_s(p)$ has directional derivatives at each point of the neighborhood. Moreover, we shall derive expressions of the directional derivatives.

Before the statement of our result (Theorem 4.5.2) we introduce the definition of directional derivatives. Let $\lambda(p)$ be a function defined in an open set $\mathcal{S} \subset \mathcal{R}^N$. The *directional derivative* $D_v\lambda(p^*)$ of $\lambda(p)$ at $p^* \in \mathcal{S}$ in the direction v is defined by

$$D_v\lambda(p^*) = \lim_{\tau \to 0} \frac{\lambda(p^* + \tau v) - \lambda(p^*)}{\tau}, \tag{4.5.4}$$

where $v \in \mathcal{R}^N$ with $\|v\|_2 = 1$, and τ is a positive scalar.

Theorem 4.5.2. *Let $p = (p_1, \ldots, p_N)^T \in \mathcal{R}^N$, and let $A(p), B(p) \in \mathcal{S}^{n\times n}$ be real analytic functions of p in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{R}^N$, where $B(p) > 0$ for $p \in \mathcal{B}(0)$. Suppose that there is a matrix $X = (X_1, X_2) \in \mathcal{R}^{n\times n}$ with $X_1 \in \mathcal{R}^{n\times r}$ such that*

$$X^T A(0) X = \begin{pmatrix} \lambda_1 I_r & 0 \\ 0 & A_2 \end{pmatrix}, \qquad X^T B(0) X = I, \qquad \lambda_1 \notin \lambda(A_2). \tag{4.5.5}$$

Then there exist r continuous functions $\lambda_1(p), \ldots, \lambda_r(p)$ in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin that are the eigenvalues of the eigenproblem (4.5.3) satisfying

$$\lambda_s(0) = \lambda_1, \qquad s = 1, \ldots, r,$$

and for any $v = (\nu_1, \ldots, \nu_N)^T \in \mathcal{R}^N$ with $\|v\|_2 = 1$ there is a permutation π of $\{1, \ldots, r\}$ dependent on v such that

$$D_v\lambda_s(0) = \lambda_{\pi(s)}\left(\sum_{j=1}^N \nu_j X_1^T S_j(\lambda_1) X_1\right), \qquad s = 1, \ldots, r, \tag{4.5.6}$$

where $\lambda_1(\cdot), \ldots, \lambda_r(\cdot)$ *denote the eigenvalues of an* $r \times r$ *matrix,* $D_v\lambda_s(0)$ *denote the directional derivatives of* $\lambda_s(p)$ *at* $p = 0$ *in the direction* v*, and the matrices* $S_j(\lambda_1)$ *are defined by*

$$S_j(\lambda_1) = \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} - \lambda_1 \left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0}, \qquad j = 1, \ldots, N. \tag{4.5.7}$$

Proof. The proof consists of the following three steps.

1) Let

$$\begin{aligned} \tilde{A}(p) &= X^T A(p) X = \begin{pmatrix} \tilde{A}_{11}(p) & \tilde{A}_{21}(p)^T \\ \tilde{A}_{21}(p) & \tilde{A}_{22}(p) \end{pmatrix}, \\ \tilde{B}(p) &= X^T B(p) X = \begin{pmatrix} \tilde{B}_{11}(p) & \tilde{B}_{21}(p)^T \\ \tilde{B}_{21}(p) & \tilde{B}_{22}(p) \end{pmatrix}. \end{aligned} \tag{4.5.8}$$

By using the implicit function theorem (Theorem 1.6.2) and the same technique described by the proof of Theorems 4.1.1 and 4.1.6 we can prove that there exists a unique pair of real analytic matrix-valued functions $Z(p), W(p) \in \mathcal{R}^{(n-r)\times r}$ in some neighborhood $\mathcal{B}_0 \subset \mathcal{B}(0)$ of the origin of $\mathcal{R}^N$ satisfying $Z(0) = W(0) = 0$ such that the matrix

$$\begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix}$$

is nonsingular for $p \in \mathcal{B}_0$, and

$$\begin{aligned} \begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix}^T \tilde{A}(p) \begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix} &= \begin{pmatrix} A_1(p) & 0 \\ 0 & A_2(p) \end{pmatrix}, \\ \begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix}^T \tilde{B}(p) \begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix} &= \begin{pmatrix} B_1(p) & 0 \\ 0 & B_2(p) \end{pmatrix}, \end{aligned} \tag{4.5.9}$$

where $A_1(p), B_1(p) \in \mathcal{S}^{r\times r}$, $B_1(p) > 0$ for $p \in \mathcal{B}_0$, and

$$\begin{aligned} A_1(p) &= \tilde{A}_{11}(p) + Z(p)^T \tilde{A}_{21}(p) + \tilde{A}_{21}(p)^T Z(p) + Z(p)^T \tilde{A}_{22}(p) Z(p), \\ B_1(p) &= \tilde{B}_{11}(p) + Z(p)^T \tilde{B}_{21}(p) + \tilde{B}_{21}(p)^T Z(p) + Z(p)^T \tilde{B}_{22}(p) Z(p). \end{aligned} \tag{4.5.10}$$

From (4.5.9)

$$\tilde{A}(p) \begin{pmatrix} I \\ Z(p) \end{pmatrix} = \tilde{B}(p) \begin{pmatrix} I \\ Z(p) \end{pmatrix} B_1(p)^{-1} A_1(p).$$

Combining it with (4.5.8) and writing

$$X_1(p) = X \begin{pmatrix} I \\ Z(p) \end{pmatrix}, \tag{4.5.11}$$

we get

$$A(p)X_1(p) = B(p)X_1(p)B_1(p)^{-1}A_1(p) \tag{4.5.12}$$

and

$$A_1(0) = \lambda_1 I_r, \quad B_1(0) = I_r, \quad X_1(0) = X_1. \tag{4.5.13}$$

From (4.5.12)

$$B_1(p)^{-1}A_1(p) = \left[X_1(p)^T B(p) X_1(p)\right]^{-1}\left[X_1(p)^T A(p) X_1(p)\right]. \tag{4.5.14}$$

Let

$$\lambda\left(B_1(p)^{-1}A_1(p)\right) = \{\lambda_s(p)\}_{s=1}^r, \quad p \in \mathcal{B}_0.$$

Then the relations (4.5.5), (4.5.8) and (4.5.9) show that

$$\lambda_s(p) \in \lambda(A(p), B(p)), \quad \lambda_s(0) = \lambda_1, \quad s = 1, \ldots, r,$$

and $\lambda_1(p), \ldots, \lambda_r(p)$ are near λ_1 provided that $\mathcal{B}_0$ is sufficiently small.

2) Let $v \in \mathcal{R}^N$ be any fixed direction. Take $p = \tau v$ in which $\tau \in [-\epsilon, \epsilon]$ and ϵ is a small positive scalar such that $\tau v \in \mathcal{B}_0$ for $\tau \in [-\epsilon, \epsilon]$. Let

$$\mu_s(\tau) = \lambda_s(\tau v), \quad s = 1, \ldots, r \tag{4.5.15}$$

and

$$H_1(p) = B_1(p)^{-1/2}A_1(p)B_1(p)^{-1/2}, \quad \hat{H}_1(\tau) = H_1(\tau v). \tag{4.5.16}$$

Then clearly

$$\lambda(\hat{H}_1(\tau)) = \{\mu_s(\tau)\}_{s=1}^r, \quad \tau \in [-\epsilon, \epsilon], \quad \mu_s(0) = \lambda_1 \ \forall s.$$

But, on the other hand, since $\hat{H}_1(\tau) \in \mathcal{S}^{r\times r}$ is real analytic on $[-\epsilon, \epsilon]$ and $\hat{H}_1(0) = \lambda_1 I_r$, by the Rellich theorem (see below NR 4.5–3) there is a positive scalar $\epsilon_1 \le \epsilon$ and real analytic functions $\hat{\lambda}_1(\tau), \ldots, \hat{\lambda}_r(\tau)$ on $[-\epsilon_1, \epsilon_1]$, such that

$$\lambda(\hat{H}_1(\tau)) = \{\hat{\lambda}_t(\tau)\}_{t=1}^r, \quad \tau \in [-\epsilon_1, \epsilon_1], \quad \hat{\lambda}_t(0) = \lambda_1 \ \forall t.$$

Observe the following facts: (i) Since the zeros of a real analytic function of one real variable are isolated (see, e.g., Cartan [15, p.41]), we have

$$\hat{\lambda}_i(\tau) \ne \hat{\lambda}_j(\tau) \quad \forall \tau \in (0, \epsilon_1], \ i \ne j$$

provided that $\hat{\lambda}_i(\tau) \not\equiv \hat{\lambda}_j(\tau)$ for $\tau \in (0, \epsilon_1]$ and the positive scalar ϵ_1 is sufficiently small; (ii) The functions $\mu_1(\tau), \ldots, \mu_r(\tau)$ are continuous on $[0, \epsilon_1]$; (iii) The sets $\{\mu_s(\tau)\}_{s=1}^r$ and $\{\hat{\lambda}_t(\tau)\}_{t=1}^r$ are just the same for any point $\tau \in [0, \epsilon_1]$, and there is a one-to-one correspondence between the elements of the two sets. Hence, there is a permutation π of $\{1, \ldots, r\}$ depending on the direction v such that

$$\mu_s(\tau) = \hat{\lambda}_{\pi(s)}(\tau) \quad \forall s, \ \tau \in [0, \epsilon_1]. \tag{4.5.17}$$

Consequently, from (4.5.4), (4.5.15) and (4.5.17), we get

$$\begin{aligned} D_v\lambda_s(0) &= \lim_{\tau\to 0}\frac{\lambda_s(\tau v)-\lambda_s(0)}{\tau} = \lim_{\tau\to 0}\frac{\mu_s(\tau)-\mu_s(0)}{\tau} \\ &= \lim_{\tau\to 0}\frac{\hat\lambda_{\pi(s)}(\tau)-\hat\lambda_{\pi(s)}(0)}{\tau} = \left(\frac{d\hat\lambda_{\pi(s)}(\tau)}{d\tau}\right)_{\tau=0}, \qquad s=1,\ldots,r. \end{aligned} \tag{4.5.18}$$

3) Let

$$G_1(p) = B_1(p)^{-1}A_1(p), \qquad \hat G_1(\tau) = G_1(\tau v). \tag{4.5.19}$$

Combining it with (4.5.16) shows

$$\lambda(\hat G_1(\tau)) = \lambda(\hat H_1(\tau)) \quad \forall \tau\in[0,\epsilon].$$

By (4.5.19), (4.5.5), (4.5.7), (4.5.11), (4.5.13) and (4.5.14), we have

$$\left(\frac{d\hat G_1(\tau)}{d\tau}\right)_{\tau=0} = \left(\frac{dG_1(\tau v)}{d\tau}\right)_{\tau=0} = \sum_{j=1}^{N}\nu_j\left(\frac{\partial G_1(p)}{\partial p_j}\right)_{p=0} = \sum_{j=1}^{N}\nu_j X_1^T S_j(\lambda_1)X_1, \tag{4.5.20}$$

which shows

$$\left(\frac{d\hat G_1(\tau)}{d\tau}\right)_{\tau=0}\in\mathcal{S}^{r\times r},$$

and hence there is a matrix $W_1\in\mathcal{O}^{r\times r}$ such that

$$W_1^T\left(\frac{d\hat G_1(\tau)}{d\tau}\right)_{\tau=0}W_1 = \operatorname{diag}(\delta_1,\ldots,\delta_r), \qquad \delta_1\le\cdots\le\delta_r. \tag{4.5.21}$$

We now write

$$W_1^T\hat G_1(\tau)W_1 = (\gamma_{kl}(\tau))_{1\le k,l\le r},$$

in which the functions $\gamma_{kl}(\tau)$ are real analytic and so may be written as the following convergent power series:

$$\gamma_{kl}(\tau) = \gamma_{kl}^{(0)} + \gamma_{kl}^{(1)}\tau + \gamma_{kl}^{(2)}\tau^2 + \cdots, \qquad k,l=1,\ldots,r.$$

From

$$\left(W_1^T\hat G_1(\tau)W_1\right)_{\tau=0} = \lambda_1 I_r$$

and

$$\left[\frac{d\left(W_1^T\hat G_1(\tau)W_1\right)}{d\tau}\right]_{\tau=0} = W_1^T\left(\frac{d\hat G_1(\tau)}{d\tau}\right)_{\tau=0}W_1$$

as well as (4.5.21), it follows that

$$\gamma_{kl}^{(0)} = \begin{cases}\lambda_1 & \text{if } k=l,\\ 0 & \text{otherwise,}\end{cases} \qquad \gamma_{kl}^{(1)} = \begin{cases}\delta_k & \text{if } k=l,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore

$$\gamma_{kl}(\tau) = \begin{cases} \lambda_1 + \delta_k \tau + \gamma_{kl}^{(2)} \tau^2 + \gamma_{kl}^{(3)} \tau^3 + \cdots & \text{if } k = l, \\ \\ \gamma_{kl}^{(2)} \tau^2 + \gamma_{kl}^{(3)} \tau^3 + \cdots & \text{otherwise.} \end{cases} \tag{4.5.22}$$

Assume that

$$\begin{aligned} \delta_1 \quad &= \cdots = \delta_{r_1} < \delta_{r_1+1} = \cdots = \delta_{r_1+r_2} < \cdots \\ &< \delta_{r_1+\cdots+r_{q-1}+1} = \cdots = \delta_{r_1+\cdots+r_{q-1}+r_q}, \qquad r_1 + \cdots + r_q = r, \end{aligned} \tag{4.5.23}$$

and write

$$\delta_{r_1} = \omega_1, \;\; \delta_{r_1+r_2} = \omega_2, \ldots, \delta_{r_1+\cdots+r_q} = \omega_q; \tag{4.5.24}$$

then by the Gerschgorin theorem (see below NR 4.5–5) from (4.5.22)–(4.5.24) we see that there are precisely q circular disks $\mathcal{D}_1, \ldots, \mathcal{D}_q$ with centers

$$\lambda_1 + \omega_1 \tau, \quad \ldots, \quad \lambda_1 + \omega_q \tau$$

and with radii of magnitude $O(\tau^2)$ such that the union $\bigcup\limits_{j=1}^{q} \mathcal{D}_j$ contains all of the eigenvalues $\hat{\lambda}_1(\tau), \ldots, \hat{\lambda}_r(\tau)$. Besides, the disks $\mathcal{D}_1, \ldots, \mathcal{D}_q$ are mutually disjoint provided that τ belongs to a sufficiently small segment $[-\epsilon_1, \epsilon_1]$, and in such a case every disk $\mathcal{D}_j$ contains exactly r_j eigenvalues which may be written as the following convergent power series:

$$\lambda_1 + \omega_j \tau + g_{r_1+\cdots+r_{j-1}+k}^{(2)} \tau^2 + g_{r_1+\cdots+r_{j-1}+k}^{(3)} \tau^3 + \cdots, \qquad k = 1, \ldots, r_j, \tag{4.5.25}$$

where $\tau \in [-\epsilon_1, \epsilon_1]$, $j = 1, \ldots, q$, and $r_0 = 0$.

Combining (4.5.25) with (4.5.23) and (4.5.24), we may rewrite the expressions of (4.5.25) as

$$\hat{\lambda}_t = \lambda_1 + \delta_t \tau + g_t^{(2)} \tau^2 + g_t^{(3)} \tau^3 + \cdots, \qquad t = 1, \ldots, r.$$

Consequently, we obtain

$$\left(\frac{d\hat{\lambda}_t(\tau)}{d\tau} \right)_{\tau=0} = \delta_t, \qquad t = 1, \ldots, r. \tag{4.5.26}$$

Combining (4.5.18) with (4.5.26), (4.5.21) and (4.5.20), shows the formulas (4.5.6). □

From Theorem 4.5.2 we get the following corollary.

Corollary 4.5.3. *Under the hypotheses of Theorem 4.5.2, there are permutations π and π' of $\{1, \dots, r\}$ such that the relations*

$$D_{e_j^{(N)}} \lambda_s(0) = \lambda_{\pi(s)}\left(X_1^T S_j(\lambda_1) X_1\right), \qquad D_{-e_j^{(N)}} \lambda_s(0) = \lambda_{\pi'(s)}\left(X_1^T S_j(\lambda_1) X_1\right)$$

are valid for $j = 1, \dots, N$ and $s = 1, \dots, r$, where the functions $\lambda_1(p), \dots, \lambda_r(p)$ and $S_j(\lambda_1)$ are described in Theorem 4.5.2. Especially, if $r = 1$ then the eigenvalue $\lambda_1(p)$ has the partial derivatives with respect to p_j at the origin

$$\left(\frac{\partial \lambda_1(p)}{\partial p_j}\right)_{p=0} = x_1^T S_j(\lambda_1) x_1, \qquad j = 1, \dots, N, \tag{4.5.27}$$

where x_1 is the associated eigenvector with λ_1 satisfying (4.5.5).

Let $A(p), \lambda_1(p), \lambda_2(p)$ be as in Example 4.5.1. Straightforward calculations show that, for any direction $v = (\cos\theta, \sin\theta)^T \in \mathcal{R}^2$ with $\theta \in [0, 2\pi)$, the functions $\lambda_1(p)$ and $\lambda_2(p)$ have the directional derivatives at $p = 0$:

$$D_v \lambda_1(0) = \cos\theta + 2\sin\theta + 1, \qquad D_v \lambda_2(0) = \cos\theta + 2\sin\theta - 1. \tag{4.5.28}$$

On the other hand, applying Theorem 4.5.2 we have

$$\begin{aligned} \{D_v \lambda_s(0)\}_{s=1}^2 &= \lambda\left(\cos\theta \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} + \sin\theta \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right) \\ &= \{\cos\theta + 2\sin\theta + 1,\ \cos\theta + 2\sin\theta - 1\}, \end{aligned}$$

which coincides with (4.5.28).

4.5.2 Structured Condition Numbers

Let the symmetric-definite generalized eigenproblem (4.5.1) have a multiple eigenvalue λ_1 of multiplicity r, and let (A, B) be slightly perturbed to a symmetric pair $(\tilde{A}, \tilde{B})$. Then, in general, λ_1 will spawn r simple eigenvalues, and the new eigenvalues will be found at varying distance from the original eigenvalue. For example, the eigenproblem (4.5.1) with $A = \text{diag}(2, 2000)$ and $B = \text{diag}(1, 1000)$ has a double eigenvalue $\lambda_1 = 2$. But one of the eigenvalues of the matrix pair (A, B) is usually much more sensitive than the other. Therefore, it may well be asked: How to make this observation precise?

In this subsection we shall define r condition numbers of the multiple eigenvalue λ_1 that measure the sensitivity of λ_1 to small perturbations in A and B, and derive explicit expressions of the r condition numbers.

By the hypothesis the eigenproblem (4.5.1) has a multiple eigenvalue λ_1 of multiplicity r. As a consequence, there is a matrix $X \in \mathcal{R}^{n\times n}$ such that

$$X^TAX = \begin{pmatrix} \lambda_1 I_r & 0 \\ 0 & A_2 \end{pmatrix}, \qquad X^TBX = I, \qquad \lambda_1 \notin \lambda(A_2). \tag{4.5.29}$$

Let $\Phi, \Psi \in \mathcal{S}^{n\times n}$, and let

$$A(p) = A + p\Phi, \quad B(p) = B + p\Psi, \quad p \in \mathcal{R}. \tag{4.5.30}$$

Moreover, let $\tilde{A}(p), \tilde{B}(p)$ be the matrices of (4.5.8). Then from the proof of Theorem 4.5.2 we see that there exist real analytic matrix-valued functions $Z(p), W(p) \in \mathcal{R}^{(n-r)\times r}$ in some neighborhood $\mathcal{B}_0$ of the origin of $\mathcal{R}$ such that $Z(0) = W(0) = 0$ and the relations of (4.5.9) hold, in which $A_1(p)$ and $B_1(p)$ are expressed by (4.5.10), and $B_1(p) > 0$ for $p \in \mathcal{B}_0$. Observe that

$$A_1(0) = \lambda_1 I_r, \quad B_1(0) = I_r, \quad \lambda\left(B_1(p)^{-1}A_1(p)\right) \subset \lambda(A(p), B(p)).$$

Hence, the multiple eigenvalue λ_1 of (A, B) will be perturbed to the eigenvalues of $B_1(p)^{-1}A_1(p)$ as the real symmetric pair (A, B) is perturbed to $(A(p), B(p))$. From (4.5.8), (4.5.10), (4.5.29) and (4.5.30) it follows that

$$A_1(p) = \lambda_1 I_r + pX_1^T\Phi X_1 + O(p^2), \quad B_1(p) = I_r + pX_1^T\Psi X_1 + O(p^2),$$

which imply

$$B_1(p)^{-1}A_1(p) = \lambda_1 I_r + p\left(X_1^T\Phi X_1 - \lambda_1 X_1^T\Psi X_1\right) + O(p^2), \quad p \to 0.$$

We now assume that the real symmetric pair (A, B) is slightly perturbed to a real symmetric pair $(A + E, B + F)$. Let

$$p = \|(E, F)\|_2, \quad \Phi = E/p, \quad \Psi = F/p.$$

Then the above discussion shows that under sufficiently small symmetric perturbations E and F in A and B, the multiple eigenvalue λ_1 of (A, B) will be perturbed to

$$\lambda_1 + \lambda_1(H), \ldots, \lambda_1 + \lambda_r(H),$$

where

$$H = X_1^TEX_1 - \lambda_1 X_1^TFX_1 + O\left(\|(E, F)\|_2^2\right) \in \mathcal{S}^{r\times r}, \tag{4.5.31}$$

and $\lambda_j(H)$ are the eigenvalues of H satisfying

$$|\lambda_1(H)| \geq \cdots \geq |\lambda_r(H)|. \tag{4.5.32}$$

Hence, the multiple eigenvalue λ_1 of multiplicity r can have r condition numbers that reflect the different sensitivities of its progeny.

Referring to §1.8, we define the condition numbers $c_j(\lambda_1)$ of λ_1 by

$$c_j(\lambda_1) = \lim_{\delta\to 0} \sup_{\left\| \begin{pmatrix} \frac{\|E\|_2}{\alpha} \\ \frac{\|F\|_2}{\beta} \end{pmatrix} \right\|_2 \le \delta} \frac{|\lambda_j(H)|}{\gamma\delta}, \quad j = 1, \ldots, r, \tag{4.5.33}$$

where α, β, γ are positive parameters, and the eigenvalues $\lambda_j(H)$ satisfy (4.5.32).

By (4.5.31), the definition (4.5.33) can be equivalently stated as

$$c_j(\lambda_1) = \sup_{\left\| \begin{pmatrix} \frac{\|E\|_2}{\alpha} \\ \frac{\|F\|_2}{\beta} \end{pmatrix} \right\|_2 \le 1} \frac{|\lambda_j(X_1^T(E - \lambda_1 F)X_1)|}{\gamma}, \quad j = 1, \ldots, r, \tag{4.5.34}$$

where $\lambda_j(X_1^T(E-\lambda_1 F)X_1)$ are the eigenvalues of $X_1^T(E-\lambda_1 F)X_1$ satisfying

$$|\lambda_1(X_1^T(E-\lambda_1 F)X_1)| \ge \cdots \ge |\lambda_r(X_1^T(E-\lambda_1 F)X_1)|.$$

The following result gives another characterization of the condition numbers $c_j(\lambda_1)$.

Theorem 4.5.4. *Let $c_j(\lambda_1)$ be the condition numbers of the multiple eigenvalue λ_1 defined by (4.5.34), and let X_1 have the singular value decomposition*

$$X_1 = U \begin{pmatrix} X_0 \\ 0 \end{pmatrix} V^T \quad \text{with} \quad X_0 = \begin{pmatrix} \xi_1 & & \\ & \ddots & \\ & & \xi_r \end{pmatrix}, \quad \xi_1 \ge \cdots \ge \xi_r > 0, \tag{4.5.35}$$

where the matrices $U = (U_1, U_2)$ and V are real orthogonal, and $U_1 \in \mathcal{O}^{n\times r}$. Then $c_j(\lambda_1)$ can be expressed by

$$c_j(\lambda_1) = \frac{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}{\gamma} \sup_{\substack{W \in \mathcal{S}^{r\times r} \\ \|W\|_2 \le 1}} \sigma_j(X_0 W X_0), \quad j = 1, \ldots, r, \tag{4.5.36}$$

where $\sigma_j(X_0WX_0)$ are the singular values of X_0WX_0 satisfying

$$\sigma_1(X_0WX_0) \ge \cdots \ge \sigma_r(X_0WX_0).$$

Proof. Define $d_j(\lambda_1)$ by

$$d_j(\lambda_1) = \frac{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}{\gamma} \sup_{\substack{W \in \mathcal{S}^{r\times r} \\ \|W\|_2 \le 1}} \sigma_j(X_0 W X_0), \quad j = 1, \ldots, r. \tag{4.5.37}$$

Then we only need to prove $c_j(\lambda_1) = d_j(\lambda_1)$ for $j = 1, \ldots, r$.

We first note that by (4.5.35) we have

$$\lambda_j(X_1^T(E - \lambda_1 F)X_1) = \lambda_j(X_0 U_1^T(E - \lambda_1 F)U_1 X_0). \tag{4.5.38}$$

Suppose that

$$\left\| \left(\frac{\|E\|_2}{\alpha}, \frac{\|F\|_2}{\beta} \right)^T \right\|_2 \leq 1.$$

Then the matrix $W \in \mathcal{S}^{r\times r}$ defined by

$$W = \frac{U_1^T(E - \lambda_1 F)U_1}{\sqrt{\alpha^2 + \lambda_1^2\beta^2}} \tag{4.5.39}$$

satisfies

$$\|W\|_2 \leq \frac{1}{\sqrt{\alpha^2 + \lambda_1^2\beta^2}} \begin{pmatrix} \alpha \\ |\lambda_1|\beta \end{pmatrix}^T \begin{pmatrix} \frac{\|E\|_2}{\alpha} \\ \frac{\|F\|_2}{\beta} \end{pmatrix} \leq \left\| \begin{pmatrix} \frac{\|E\|_2}{\alpha} \\ \frac{\|F\|_2}{\beta} \end{pmatrix} \right\|_2 \leq 1,$$

and

$$\begin{aligned} |\lambda_j(X_1^T(E - \lambda_1 F)X_1)| &= \sqrt{\alpha^2 + \lambda_1^2\beta^2}|\lambda_j(X_0 W X_0)| \quad \text{(by (4.5.38) and (4.5.39))} \\ &= \sqrt{\alpha^2 + \lambda_1^2\beta^2}\sigma_j(X_0 W X_0), \quad j = 1, \ldots, r. \end{aligned}$$

Combining this fact with (4.5.34) and (4.5.37) shows $c_j(\lambda_1) \leq d_j(\lambda_1)$.

Conversely, for any $W \in \mathcal{S}^{r\times r}$ satisfying $\|W\|_2 \leq 1$, the $n \times n$ real symmetric matrices E, F defined by

$$E = \phi U \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix} U^T, \quad F = \psi U \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix} U^T$$

with

$$\phi = \frac{\alpha^2}{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}, \quad \psi = \frac{-\text{sign}(\lambda_1)|\lambda_1|\beta^2}{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}$$

satisfy

$$\left\| \begin{pmatrix} \frac{\|E\|_2}{\alpha} \\ \frac{\|F\|_2}{\beta} \end{pmatrix} \right\|_2 = \|W\|_2 \leq 1 \quad \text{and} \quad U_1^T(E - \lambda_1 F)U_1 = \sqrt{\alpha^2 + \lambda_1^2\beta^2}W.$$

Combining this fact with (4.5.34), (4.5.37) and (4.5.38) shows $d_j(\lambda_1) \leq c_j(\lambda_1)$. Consequently, $c_j(\lambda_1) = d_j(\lambda_1)$. □

The following result gives computable formulas of the condition numbers $c_j(\lambda_1)$.

Theorem 4.5.5. *Let (A, B), $X = (X_1, X_2)$ and λ_1 be as in (4.5.29), and let X_1 have the singular value decomposition (4.5.35). Define π_j by*

$$\pi_j = \min_{1\le k\le\left[\frac{j+1}{2}\right]} \xi_k\xi_{j-k+1}, \qquad j = 1, \dots, r, \tag{4.5.40}$$

where $\left[\frac{j+1}{2}\right]$ denotes the greatest integer not greater than $\frac{j+1}{2}$. Then the condition numbers $c_j(\lambda_1)$ defined by (4.5.34) can be expressed by

$$c_j(\lambda_1) = \frac{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}{\gamma}\pi_j, \qquad j = 1, \dots, r. \tag{4.5.41}$$

Proof. For an arbitrarily fixed integer j on $[1, r]$, define ω_j by

$$\omega_j = \sup_{\substack{W \in \mathcal{S}^{r\times r}\\ \|W\|_2 \le 1}} \sigma_j(X_0WX_0). \tag{4.5.42}$$

Then by (4.5.36) we only need to prove $\omega_j = \pi_j$.

Let $W \in \mathcal{S}^{r\times r}$, and $\|W\|_2 \le 1$. Then by Theorem 4.5.16 (see below NR 4.5–9), we have

$$\begin{aligned}
\sigma_j(X_0WX_0) &\le \min_{1\le k\le j}\{\sigma_k(X_0W)\xi_{j-k+1}\}\\
&\le \min_{1\le k\le j}\left\{\left(\min_{1\le l\le k}\xi_l\sigma_{k-l+1}(W)\right)\xi_{j-k+1}\right\}\\
&\le \min_{1\le k\le j}\{\xi_k\xi_{j-k+1}\} = \pi_j.
\end{aligned}$$

Combining it with (4.5.42) shows $\omega_j \le \pi_j$.

On the other hand, the $r \times r$ matrix $W_j = \mathrm{diag}(W^{(j)}, 0)$ with

$$W^{(j)} = \begin{pmatrix} & & & & 1\\ & & & 1 & \\ & & \iddots & & \\ & 1 & & & \\ 1 & & & & \end{pmatrix} \in \mathcal{S}^{j\times j}$$

satisfies $W_j \in \mathcal{S}^{r\times r}$, $\|W_j\|_2 = 1$, and

$$X_0W_jX_0 = \mathrm{diag}\left(\begin{pmatrix} & & & & \xi_1\xi_j\\ & & & \xi_2\xi_{j-1} & \\ & & \iddots & & \\ & \xi_{j-1}\xi_2 & & & \\ \xi_j\xi_1 & & & & \end{pmatrix}, 0\right),$$

which implies $\sigma_j(X_0W_jX_0) = \pi_j$. Combining this fact with (4.5.42) shows $\omega_j \geq \pi_j$. Consequently, $\omega_j = \pi_j$. □

Remark 4.5.6. Theorem 4.5.5 implies that if λ_1 is a simple eigenvalue of the eigenproblem (4.5.1), and if x_1 is an associated eigenvector satisfying $x_1^H Bx_1 = 1$, then the simple eigenvalue λ_1 has the condition number $c(\lambda_1)$, and

$$c(\lambda_1) = c_1(\lambda_1) = \frac{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}{\gamma}\|x_1\|_2^2.$$

Remark 4.5.7. Taking $\alpha = \beta = \gamma = 1$ in (4.5.34) and (4.5.41), we get the absolute condition numbers $c_j^{(\text{abs})}(\lambda_1)$ of the multiple eigenvalue λ_1, which can be expressed by

$$c_j^{(\text{abs})}(\lambda_1) = \sqrt{1 + \lambda_1^2}\pi_j, \quad j = 1, \ldots, r, \tag{4.5.43}$$

where π_j are defined by (4.5.40). Taking $\alpha = \|A\|_2, \beta = \|B\|_2$ and $\gamma = |\lambda_1|$ in (4.5.34) and (4.5.41), we get the relative condition numbers $c_j^{(\text{rel})}(\lambda_1)$ of the multiple eigenvalue λ_1 (if $\lambda_1 \neq 0$), which can be expressed by

$$c_j^{(\text{rel})}(\lambda_1) = \frac{\sqrt{\|A\|_2^2 + \lambda_1^2\|B\|_2^2}}{|\lambda_1|}\pi_j, \quad j = 1, \ldots, r. \tag{4.5.44}$$

Moreover, from the definition (4.5.34) it follows that for sufficiently small perturbations $E, F \in \mathcal{S}^{n\times n}$, the matrix pair $(A + E, B + F)$ has the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r$ such that

$$|\tilde{\lambda}_j - \lambda_1| \lesssim c_j^{(\text{abs})}(\lambda_1)\sqrt{\|E\|_2^2 + \|F\|_2^2} \equiv \beta_j^{(\text{abs})}, \quad j = 1, \ldots, r, \tag{4.5.45}$$

and

$$\frac{|\tilde{\lambda}_j - \lambda_1|}{|\lambda_1|} \lesssim c_j^{(\text{rel})}(\lambda_1)\sqrt{\left(\frac{\|E\|_2}{\|A\|_2}\right)^2 + \left(\frac{\|F\|_2}{\|B\|_2}\right)^2} \equiv \beta_j^{(\text{rel})}, \quad j = 1, \ldots, r, \tag{4.5.46}$$

where it is assumed that $\lambda_1 \neq 0$ in (4.5.46). The scalars $\beta_j^{(\text{abs})}$ and $\beta_j^{(\text{rel})}$ are the first order absolute and relative perturbation bounds for the multiple eigenvalue λ_1, respectively.

Remark 4.5.8. Let π_j be the scalars defined by (4.5.40). Then it follows from $\xi_1 \geq \cdots \geq \xi_r > 0$ that $\pi_1 \geq \cdots \geq \pi_r > 0$. Moreover, it can be proved that if

$$\xi_1 = \cdots = \xi_m > \xi_{m+1} > \cdots > \xi_{m+l}$$

for some integer l satisfying $m + l \leq r$, then

$$\pi_1 = \cdots = \pi_m > \pi_{m+1} > \cdots > \pi_{m+l},$$

and by (4.5.41)–(4.5.46) we have

$$c_1(\lambda_1) = \cdots = c_m(\lambda_1) > c_{m+1}(\lambda_1) > \cdots > c_{m+l}(\lambda_1),$$

$$\beta_1^{(\text{abs})} = \cdots = \beta_m^{(\text{abs})} > \beta_{m+1}^{(\text{abs})} > \cdots > \beta_{m+l}^{(\text{abs})},$$

$$\beta_1^{(\text{rel})} = \cdots = \beta_m^{(\text{rel})} > \beta_{m+1}^{(\text{rel})} > \cdots > \beta_{m+l}^{(\text{rel})}.$$

Remark 4.5.9. From the proof of Theorem 4.5.5 we see that for every integer j on $[1, r]$ there is a matrix $W_j \in \mathcal{S}^{r\times r}$ with $\|W_j\|_2 \leq 1$ such that $\sigma_j(X_0 W_j X_0) = \pi_j$, where the scalars π_j are defined by (4.5.40). It is worth pointing out the following facts:

(i) If the singular values ξ_j of X_1 (see (4.5.35)) satisfy $\xi_1 = \cdots = \xi_m$ for some integer m on $[2, r]$, then there is a matrix $W \in \mathcal{S}^{r\times r}$ (e.g., $W = \text{diag}(I_m, 0)$) with $\|W\|_2 \leq 1$ such that $\sigma_j(X_0 W X_0) = \pi_j = \xi_1^2$ for all $j = 1, \ldots, m$.

(ii) If $\xi_1 = \cdots = \xi_m > \xi_{m+1}$ for some integer m on $[1, r-1]$, then there is no a single $W \in \mathcal{S}^{r\times r}$ with $\|W\|_2 \leq 1$ such that

$$\sigma_j(X_0 W X_0) = \pi_j \quad \text{for} \quad j = 1, \ldots, m+1. \tag{4.5.47}$$

We now prove the fact by contradiction. Assume that the relation (4.5.47) holds for some $W \in \mathcal{S}^{r\times r}$ with $\|W\|_2 \leq 1$. Then by (4.5.40) and $\xi_1 = \cdots = \xi_m$, we have

$$\prod_{j=1}^{m+1} \sigma_j(X_0 W X_0) = \prod_{j=1}^{m+1} \pi_j = \xi_1^{2m+1}\xi_{m+1}. \tag{4.5.48}$$

On the other hand, by Theorem 4.5.17 (see below NR 4.5–9) $\sigma_j(X_0) = \xi_j$, $\sigma_j(W) \leq 1$, and $\xi_1 = \cdots = \xi_m > \xi_{m+1}$, we have

$$\prod_{j=1}^{m+1} \sigma_j(X_0 W X_0) \leq \prod_{j=1}^{m+1} \sigma_j(X_0)\sigma_j(W)\sigma_j(X_0) \leq \xi_1^{2m}\xi_{m+1}^2 < \xi_1^{2m+1}\xi_{m+1},$$

which contradicts the equality (4.5.48). The proof is completed. □

Combining the above-mentioned fact (ii) with Theorems 4.5.4 and 4.5.5 shows that if the singular values $\xi_1, \ldots, \xi_r$ of X_1 are not mutually equal, then there is no a single $W \in \mathcal{S}^{r\times r}$ with $\|W\|_2 \leq 1$ such that

$$c_j(\lambda_1) = \frac{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}{\gamma}\sigma_j(X_0 W X_0) \qquad \text{for } j = 1, \ldots, r.$$

Consequently, the condition numbers $c_1(\lambda_1), \ldots, c_r(\lambda_1)$ may be called the *worst-case* condition numbers of the multiple eigenvalue λ_1.

We now use a simple numerical example cited from [97, p.300] to test our results.

Example 4.5.10. Consider the eigenproblem (4.5.1) with

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2000 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1000 \end{pmatrix}. \tag{4.5.49}$$

We have

$$X^T AX = \mathrm{diag}(2,2), \qquad X^T BX = I,$$

where $X = \mathrm{diag}(\xi_1, \xi_2)$ with the singular values

$$\xi_1 = 1, \qquad \xi_2 = 1/\sqrt{1000} \approx 0.0316.$$

Obviously, the real symmetric pair (A, B) has a double eigenvalue $\lambda_1 = 2$. It is known (see Stewart and Sun [97, p.300]) that one of the eigenvalues is very sensitive to perturbations of order 0.1, whereas the other is not. We now use the results of this subsection to analyze the phenomenon.

By (4.5.40), (4.5.43) and (4.5.44) we have

$$c_1^{(\mathrm{abs})}(\lambda_1) = \sqrt{1+\lambda_1^2}\xi_1^2 \approx 2.2361,$$

$$c_2^{(\mathrm{abs})}(\lambda_1) = \sqrt{1+\lambda_1^2}\xi_1\xi_2 \approx 7.0711 \times 10^{-2},$$

$$c_1^{(\mathrm{rel})}(\lambda_1) = \frac{\sqrt{\|A\|_2^2 + \lambda_1^2\|B\|_2^2}}{|\lambda_1|}\xi_1^2 \approx 1.4142 \times 10^3,$$

$$c_2^{(\mathrm{rel})}(\lambda_1) = \frac{\sqrt{\|A\|_2^2 + \lambda_1^2\|B\|_2^2}}{|\lambda_1|}\xi_1\xi_2 \approx 4.4721 \times 10.$$

Let (E, F) be any symmetric perturbation satisfying $\|E\|_2 = \|F\|_2 = 0.1$, and let $\tilde{\lambda}_1, \tilde{\lambda}_2$ be the eigenvalues of $(A + E, B + F)$. Then by (4.5.45) and (4.5.46) we have the first order perturbation estimates

$$\begin{gathered} |\tilde{\lambda}_1 - \lambda_1| \lesssim \beta_1^{(\mathrm{abs})} \approx 3.1623 \times 10^{-1}, \qquad |\tilde{\lambda}_2 - \lambda_1| \lesssim \beta_2^{(\mathrm{abs})} \approx 1.0000 \times 10^{-2}, \\ \frac{|\tilde{\lambda}_1 - \lambda_1|}{|\lambda_1|} \lesssim \beta_1^{(\mathrm{rel})} \approx 1.5811 \times 10^{-1}, \qquad \frac{|\tilde{\lambda}_2 - \lambda_1|}{|\lambda_1|} \lesssim \beta_2^{(\mathrm{rel})} \approx 5.0000 \times 10^{-3}. \end{gathered} \tag{4.5.50}$$

From the estimates of (4.5.50) we can understand why one of the eigenvalues of the matrix pair (4.5.49) is usually much more sensitive than the other. Hence, the results of this subsection give an answer to the open research problem proposed in [97, p.300].

Remark 4.5.11. For the multiple eigenvalue λ_1 we can use the Frobenius norm $\|\cdot\|_F$ to define the condition numbers $\hat{c}_j(\lambda_1)$ by

$$\hat{c}_j(\lambda_1) = \lim_{\delta\to 0} \sup_{\left\| \begin{pmatrix} \frac{\|E\|_F}{\alpha} \\ \frac{\|F\|_F}{\beta} \end{pmatrix} \right\|_2 \le \delta} \frac{|\lambda_j(H)|}{\gamma\delta}, \quad j = 1, \ldots, r,$$

where H is the matrix of (4.5.31), the eigenvalues $\lambda_j(H)$ satisfy (4.5.32), and α, β, γ are positive parameters. **A conjecture:** *The condition numbers $\hat{c}_j(\lambda_1)$ can be expressed by*

$$\hat{c}_j(\lambda_1) = \frac{\sqrt{\alpha^2 + \lambda_1^2\beta^2}}{\gamma\sqrt{j}}\pi_j, \qquad j = 1, \ldots, r,$$

where the scalars π_j are defined by (4.5.40), in which ξ_k for $k = 1, \ldots, r$ are the singular values of X_1 (see (4.5.35).

4.5.3 Structured Backward Errors

Let λ_1 be a nonzero eigenvalue of the eigenproblem (4.5.1) with multiplicity $r \geq 1$, and $x_1, \ldots, x_l$ be associated eigenvectors. Suppose that $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r$ are approximations of λ_1, and $\tilde{x}_1, \ldots, \tilde{x}_r$ are associated approximate eigenvectors, among which $\tilde{x}_1, \ldots, \tilde{x}_r$ are linearly independent. Then there is a question: Are $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r$ and $\tilde{x}_1, \ldots, \tilde{x}_r$ the eigenvalues and associated eigenvectors of a "nearby" generalized symmetric eigenvalue problem?

In this subsection we suggest a measure for appraising the quality of the approximate solution $\{\tilde{x}_1, \ldots, \tilde{x}_r; \tilde{\lambda}\}$, where $\tilde{\lambda}$ is defined by

$$\tilde{\lambda} = (\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_r)/r,$$

and assume $\tilde{\lambda} \neq 0$.

Let

$$\tilde{X}_1 = (\tilde{x}_1, \ldots, \tilde{x}_r).$$

By §1.9, we define the backward error $\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda})$ by

$$\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda}) = \min\left\{\left\|\begin{pmatrix} \|E\|_2 \\ \omega\|F\|_2 \end{pmatrix}\right\|_2 : \begin{array}{l} E,\, F \in \mathcal{S}^{n\times n}, \\ (A+E)\tilde{X}_1 = \tilde{\lambda}(B+F)\tilde{X}_1 \end{array}\right\}, \qquad (4.5.51)$$

where ω is a positive parameter.

For deriving a computable formula of $\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda})$, we first consider a special case where only the matrix A is perturbed. Define the corresponding backward error $\beta_0(\tilde{X}_1, \tilde{\lambda})$ by

$$\beta_0(\tilde{X}_1, \tilde{\lambda}) = \min\{\|E\|_2 \,:\, E \in \mathcal{S}^{n\times n},\ (A+E)\tilde{X}_1 = \tilde{\lambda}B\tilde{X}_1\}. \qquad (4.5.52)$$

Take the QR factorization of $\tilde{X}_1$:

$$\tilde{X}_1 = Q_1 R_1, \qquad (4.5.53)$$

where $Q_1 \in \mathcal{O}^{n\times r}$, and $R_1 \in \mathcal{R}^{r\times r}$ is upper triangular and nonsingular. The following result gives a computable formula of $\beta_0(\tilde{X}_1, \tilde{\lambda})$.

Theorem 4.5.12. *Let $\beta_0(\tilde{X}_1, \tilde{\lambda})$ be the backward error defined by (4.5.52), and let R be the residual defined by*

$$R = (\tilde{\lambda} B - A)Q_1. \tag{4.5.54}$$

Then

$$\beta_0(\tilde{X}_1, \tilde{\lambda}) = \|R\|_2. \tag{4.5.55}$$

Proof. Using the QR factorization (4.5.53), the constraint $(A+E)\tilde{X}_1 = \tilde{\lambda} B \tilde{X}_1$ in (4.5.52) is equivalent to

$$EQ_1 = R.$$

Consequently, the definition (4.5.52) can be written

$$\beta_0(\tilde{X}_1, \tilde{\lambda}) = \min_{E \in \mathcal{G}_0} \|E\|_2, \tag{4.5.56}$$

where the set $\mathcal{G}_0$ is defined by

$$\mathcal{G}_0 = \left\{ E \in \mathcal{S}^{n \times n} \ : \ EQ_1 = R \right\}.$$

Choose Q_2 so that $Q = (Q_1, Q_2) \in \mathcal{O}^{n \times n}$. By Theorem 1.5.2, $\mathcal{G}_0 \neq \emptyset$, and any $E \in \mathcal{G}_0$ can be expressed by

$$E = RQ_1^T + Q_1 R^T - Q_1 R^T Q_1 Q_1^T + Q_2 Q_2^T T Q_2 Q_2^T,$$

where $T \in \mathcal{S}^{n \times n}$. Thus, from (4.5.56)

$$\beta_0(\tilde{X}_1, \tilde{\lambda}) = \min_{E \in \mathcal{G}_0} \|Q^T E Q\|_2 = \min_{T \in \mathcal{S}^{n \times n}} \left\| \begin{pmatrix} Q_1^T R & R^T Q_2 \\ Q_2^T R & Q_2^T T Q_2 \end{pmatrix} \right\|_2 = \|R\|_2, \tag{4.5.57}$$

where we have applied Theorem 1.2.3, and used the fact that the matrix $T = Q_2 W Q_2^T$ satisfies $T \in \mathcal{S}^{n \times n}$ and $Q_2^T T Q_2 = W$ for any $W \in \mathcal{S}^{(n-r)\times(n-r)}$. The proof is completed. □

Define the set $\mathcal{G}$ by

$$\mathcal{G} = \{(E, F) \ : \ E, F \in \mathcal{S}^{n \times n}, \ EQ_1 = \tilde{\lambda} F Q_1 + R\}, \tag{4.5.58}$$

and for an arbitrarily fixed $F \in \mathcal{S}^{n \times n}$, define the set $\mathcal{G}_F$ by

$$\mathcal{G}_F = \{E \in \mathcal{S}^{n \times n} \ : \ EQ_1 = \tilde{\lambda} F Q_1 + R\}, \tag{4.5.59}$$

where Q_1 and R are defined by (4.5.53) and (4.5.54), respectively. The following result gives a computable formula of $\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda})$.

Theorem 4.5.13. *The backward error $\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda})$ defined by (4.5.51) has the expression*

$$\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda}) = \frac{\omega}{\sqrt{\tilde{\lambda}^2 + \omega^2}} \|R\|_2, \tag{4.5.60}$$

where R is the residual defined by (4.5.54).

Before we give a proof of Theorem 4.5.13, we first prove the following lemma.

Lemma 4.5.14. *Let $G \in \mathcal{R}^{m\times n}$ and $\gamma > 0$ be given. Then for any matrix norm $\|\cdot\|$ we have*

$$\min_{Z\in\mathcal{R}^{m\times n}} \left(\|Z+G\|^2 + \gamma\|Z\|^2\right) = \frac{\gamma}{1+\gamma}\|G\|^2. \tag{4.5.61}$$

Proof. It is easy to verify that the inequality

$$\frac{\gamma}{1+\gamma}(\alpha+\beta)^2 \le \alpha^2 + \gamma\beta^2 \tag{4.5.62}$$

holds for any $\alpha, \beta, \gamma \ge 0$. We now use the inequality (4.5.62) to prove (4.5.61).

For any $Z \in \mathcal{R}^{m\times n}$ we have

$$\begin{aligned}\frac{\gamma}{1+\gamma}\|G\|^2 &\le \frac{\gamma}{1+\gamma}(\|Z+G\| + \|Z\|)^2 \\ &\le \|Z+G\|^2 + \gamma\|Z\|^2. \quad \text{(by (4.5.62))}\end{aligned} \tag{4.5.63}$$

Moreover, the matrix

$$\widehat{Z} = -\frac{1}{1+\gamma}G \tag{4.5.64}$$

satisfies

$$\|\widehat{Z}+G\|^2 + \gamma\|\widehat{Z}\|^2 = \frac{\gamma}{1+\gamma}\|G\|^2.$$

Combining it with (4.5.63) shows (4.5.61). □

Proof of Theorem 4.5.13. From (4.5.51), (4.5.58) and (4.5.59) it follows that

$$\begin{aligned}\left[\beta^{(\omega)}(\tilde{X}_1, \tilde{\lambda})\right]^2 &= \min_{(E,F)\in\mathcal{G}} \left(\|E\|_2^2 + \omega^2\|F\|_2^2\right) \\ &= \min_{F\in\mathcal{S}^{n\times n}} \left(\omega^2\|F\|_2^2 + \min_{E\in\mathcal{G}_F} \|E\|_2^2\right).\end{aligned}$$

Observe the following facts: (i) Using the QR factorization (4.5.53) of $\tilde{X}_1$, the constraint $(A+E)\tilde{X}_1 = \tilde{\lambda}(B+F)\tilde{X}_1$ in (4.5.51) is equivalent to

$$EQ_1 = \tilde{\lambda}FQ_1 + R,$$

where R is the residual defined by (4.5.54); (ii) By Theorem 4.5.11, we have

$$\min_{E\in\mathcal{G}_F} \|E\|_2 = \|\tilde{\lambda}FQ_1 + R\|_2;$$

(iii) The spectral norm is a unitarily invariant norm. Hence,

$$\left[\beta^{(\omega)}(\tilde{X}_1,\tilde{\lambda})\right]^2 = \min_{F\in\mathcal{S}^{n\times n}} \left(\omega^2\|Q^TFQ\|_2^2 + \|Q^T(\tilde{\lambda}FQ_1+R)\|_2^2\right)$$

$$= \min_{H\in\mathcal{S}^{n\times n}} \left(\omega^2\|H\|_2^2 + \left\|\begin{pmatrix} \tilde{\lambda}H_{11}+Q_1^TR \\ \tilde{\lambda}H_{21}+Q_2^TR \end{pmatrix}\right\|_2^2\right)$$

$$= \min_{\substack{H_{11}\in\mathcal{S}^{r\times r} \\ H_{21}\in\mathcal{R}^{(n-r)\times r}}} \left(\omega^2 \min_{H_{22}\in\mathcal{S}^{(n-r)\times(n-r)}} \|H\|_2^2 + \left\|\tilde{\lambda}\begin{pmatrix} H_{11} \\ H_{21}\end{pmatrix} + Q^TR\right\|_2^2\right), \tag{4.5.65}$$

where $H = \begin{pmatrix} H_{11} & H_{21}^T \\ H_{21} & H_{22}\end{pmatrix}$. By Theorem 1.2.3, we have

$$\min_{H_{22}\in\mathcal{S}^{(n-r)\times(n-r)}} \left\|\begin{pmatrix} H_{11} & H_{21}^T \\ H_{21} & H_{22}\end{pmatrix}\right\|_2 = \left\|\begin{pmatrix} H_{11} \\ H_{21}\end{pmatrix}\right\|_2.$$

Substituting it into (4.5.65) shows

$$\left[\beta^{(\omega)}(\tilde{X}_1,\tilde{\lambda})\right]^2 = \min_{\substack{H_{11}\in\mathcal{S}^{r\times r} \\ H_{21}\in\mathcal{R}^{(n-r)\times r}}} \left(\left\|\tilde{\lambda}\begin{pmatrix} H_{11} \\ H_{21}\end{pmatrix} + Q^TR\right\|_2^2 + \omega^2\left\|\begin{pmatrix} H_{11} \\ H_{21}\end{pmatrix}\right\|_2^2\right). \tag{4.5.66}$$

Further, by Lemma 4.5.14 and (4.5.64)), the minimum in (4.5.66) is achieved for

$$\begin{pmatrix} \widehat{H}_{11} \\ \widehat{H}_{21}\end{pmatrix} = -\frac{\tilde{\lambda}}{\tilde{\lambda}^2+\omega^2}Q^TR,$$

where

$$\widehat{H}_{11} = Q_1^TR = Q_1^T(\tilde{\lambda}B-A)Q_1 \in \mathcal{S}^{r\times r};$$

and we have

$$\left[\beta^{(\omega)}(\tilde{X}_1,\tilde{\lambda})\right]^2 = \frac{\omega^2}{\tilde{\lambda}^2+\omega^2}\|Q^TR\|_2^2,$$

which gives (4.5.60). □

Remark 4.5.15. Taking $\omega\to\infty$ forces $F=0$ in (4.5.51) and (4.5.60), we get the expression (4.5.55) of the backward error $\beta_0(\tilde{X}_1,\tilde{\lambda})$ defined by (4.5.52).

Remark 4.5.16. Let $\tilde{\lambda}$ and $\tilde{X}_1$ be as in (4.5.51). We now define the absolute backward error $\eta_{\rm abs}(\tilde{X}_1,\tilde{\lambda}_1)$ by

$$\eta_{\rm abs}(\tilde{X}_1,\tilde{\lambda}_1) = \min\left\{\left\|\begin{pmatrix} \|E\|_2 \\ \|F\|_2\end{pmatrix}\right\|_2 \;:\; \begin{array}{l} E,F\in\mathcal{S}^{n\times n}, \\ (A+E)\tilde{X}_1 = \tilde{\lambda}_1(B+F)\tilde{X}_1, \end{array}\right\},$$

and define the relative backward error $\eta_{\rm rel}(\tilde{X}_1, \tilde{\lambda}_1)$ by

$$\eta_{\rm rel}(\tilde{X}_1, \tilde{\lambda}_1) = \min \left\{ \left\| \begin{pmatrix} \|E\|_2/\|A\|_2 \\ \|F\|_2/\|B\|_2 \end{pmatrix} \right\|_2 : \begin{array}{c} E, F \in \mathcal{S}^{n\times n}, \\ (A+E)\tilde{X}_1 = \tilde{\lambda}_1(B+F)\tilde{X}_1, \end{array} \right\},$$

then by (4.5.60) we have the computable formulas

$$\eta_{\rm abs}(\tilde{X}_1, \tilde{\lambda}_1) = \beta^{(1)}(\tilde{X}_1, \tilde{\lambda}_1) = \frac{\|R\|_2}{\sqrt{1+\tilde{\lambda}_1^2}}, \tag{4.5.67}$$

and

$$\eta_{\rm rel}(\tilde{X}_1, \tilde{\lambda}_1) = \frac{1}{\|A\|_2}\beta^{(\|A\|_2/\|B\|_2)}(\tilde{X}_1, \tilde{\lambda}_1) = \frac{\|R\|_2}{\sqrt{\|A\|_2^2 + \tilde{\lambda}_1^2\|B\|_2^2}}, \tag{4.5.68}$$

where R is the residual defined by (4.5.54).

Example 4.5.17. Consider the symmetric-definite generalized eigenproblem (4.5.1) with

$$A = \begin{pmatrix} 9 & -19 & -23 & 33 & -20 & -2 & -5 \\ -19 & 56 & 5 & 0 & 6 & -2 & 0 \\ -23 & 5 & 26 & 5 & 41 & 0 & 5 \\ 33 & 0 & 5 & 94 & 30 & 0 & -20 \\ 20 & 6 & 41 & 30 & 65 & -2 & 20 \\ -2 & -2 & 0 & 0 & -2 & 4 & 0 \\ -5 & 0 & 5 & -20 & 20 & 0 & 10 \end{pmatrix}$$

and

$$B = \begin{pmatrix} 8 & -3 & -7 & -1 & -4 & -2 & -1 \\ -3 & 12 & 1 & 0 & 2 & -2 & 0 \\ -7 & 1 & 11 & 1 & 5 & 0 & 1 \\ -1 & 0 & 1 & 34 & 6 & 0 & -4 \\ -4 & 2 & 5 & 6 & 23 & -2 & 4 \\ -2 & -2 & 0 & 0 & -2 & 4 & 0 \\ -1 & 0 & 1 & -4 & 4 & 0 & 2 \end{pmatrix}.$$

The eigenvalues of the eigenproblem are

$$\lambda_1 = \lambda_2 = \lambda_3 = 5, \quad \lambda_4 = 3, \quad \lambda_5 = 1, \quad \lambda_6 = -6, \quad \lambda_7 = -14.$$

We compute $X \in \mathcal{R}^{7\times 7}$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_7)$ of (4.5.2) by the following steps (see Golub and Van Loan [41, Algorithm 8.7.1]):

> Compute the Cholesky factorization $B = GG^T$, where G is a lower triangular matrix with positive diagonal elements.
> Compute $C = G^{-1}AG^{-T}$.
> Use the symmetric QR algorithm to compute the Schur decomposition $Q^TCQ = \mathrm{diag}(\lambda_1, \ldots, \lambda_7)$.
> Set $X = G^{-T}Q$.

Write the computed X and λ_j as $\tilde{X}$ and $\tilde{\lambda}_j$ $(j = 1, \ldots, 7)$, respectively. By (4.5.53) and (4.5.54) we compute the residual R, and then applying (4.5.67) and (4.5.68) we get

$$\eta_{\text{abs}}(\tilde{X}_1, \tilde{\lambda}) = 7.63 \times 10^{-15}, \quad \eta_{\text{rel}}(\tilde{X}_1, \tilde{\lambda}) = 1.74 \times 10^{-16},$$

where $\tilde{\lambda} = (\tilde{\lambda}_1 + \tilde{\lambda}_2 + \tilde{\lambda}_3)/3$ is an approximation of λ_1, and $\tilde{X}_1$ consists of the associated eigenvectors of $\tilde{\lambda}_1$, $\tilde{\lambda}_2$, $\tilde{\lambda}_3$.

Similarly, for $\tilde{\lambda}_4, \tilde{\lambda}_5, \tilde{\lambda}_6, \tilde{\lambda}_7$ and associated $\tilde{x}_4, \tilde{x}_5, \tilde{x}_6, \tilde{x}_7$, we get

$$\eta_{\text{abs}}(\tilde{x}_4, \tilde{\lambda}_4) = 3.27 \times 10^{-15}, \quad \eta_{\text{rel}}(\tilde{x}_4, \tilde{\lambda}_4) = 6.22 \times 10^{-17},$$

$$\eta_{\text{abs}}(\tilde{x}_5, \tilde{\lambda}_5) = 8.95 \times 10^{-15}, \quad \eta_{\text{rel}}(\tilde{x}_5, \tilde{\lambda}_5) = 9.93 \times 10^{-17},$$

$$\eta_{\text{abs}}(\tilde{x}_6, \tilde{\lambda}_6) = 1.99 \times 10^{-15}, \quad \eta_{\text{rel}}(\tilde{x}_6, \tilde{\lambda}_6) = 4.71 \times 10^{-17},$$

$$\eta_{\text{abs}}(\tilde{x}_7, \tilde{\lambda}_7) = 2.80 \times 10^{-15}, \quad \eta_{\text{rel}}(\tilde{x}_7, \tilde{\lambda}_7) = 7.28 \times 10^{-17}.$$

The results show that each computed eigenvalue and associated eigenvector are an exact eigenvalue and an associated eigenvector of a very slightly perturbed symmetric generalized eigenproblem; in other words, the computation has proceeded quite stably.

Remark 4.5.18. For the approximate solution $\{\tilde{x}_1, \ldots, \tilde{x}_r; \tilde{\lambda}\}$ given at the beginning of this subsection we can use the Frobenius norm $\|\cdot\|_F$ to define the backward error $\hat{\beta}^{(\omega)}(\tilde{X}_1, \tilde{\lambda})$ by

$$\hat{\beta}^{(\omega)}(\tilde{X}_1, \tilde{\lambda}) = \min \left\{ \left\| \begin{pmatrix} \|E\|_F \\ \omega \|F\|_F \end{pmatrix} \right\|_2 : \begin{array}{c} E,\, F \in \mathcal{S}^{n\times n}, \\ (A+E)\tilde{X}_1 = \tilde{\lambda}(B+F)\tilde{X}_1 \end{array} \right\},$$

where $\tilde{X}_1 = (\tilde{x}_1, \ldots, \tilde{x}_r)$, and ω is a positive parameter. It can be proved that *there is a computable formula for the backward error* $\hat{\beta}^{(\omega)}(\tilde{X}_1, \tilde{\lambda})$:

$$\hat{\beta}^{(\omega)}(\tilde{X}_1, \tilde{\lambda}) = \frac{\omega}{\sqrt{\tilde{\lambda}^2 + \omega^2}} \sqrt{2\|R\|_F^2 - \|Q_1^T R\|_F^2},$$

where R is the residual defined by (4.5.54), and Q_1 is the orthogonal factor of $\tilde{X}_1$ in its QR factorization (see (4.5.53)). The proof is left as an exercise.

Notes and References

NR 4.5–1. §4.5.1 and §4.5.2 are based on Sun [107].

NR 4.5–2. Example 4.5.1 is cited from Wang and Garbow [126, p.606].

NR 4.5–3. Rellich Theorem [87]. *Let $A(\xi) \in \mathcal{S}^{n\times n}$ be an analytic matrix-valued function of a single real variable ξ in a neighborhood of the origin, and let*

λ_1 be an eigenvalue of $A(0)$ with multiplicity r. Then there exist r real analytic functions $\lambda_1(\xi), \dots, \lambda_r(\xi)$ in a neighborhood of the origin, such that $\lambda_1(\xi), \dots, \lambda_r(\xi)$ are eigenvalues of $A(\xi)$ and

$$\lambda_s(0) = \lambda_1, \quad s = 1, \dots, r.$$

NR 4.5–4. The Rellich theorem stated in NR 4.5–3 is only a real form of Rellich's result on the eigenvalues of a Hermitian matrix-valued function of a single real variable. By Rellich [87, p.31], the general result can be stated as follows: *Let $A(\xi) \in \mathcal{H}^{n\times n}$ for real ξ with small $|\xi|$, and let the elements of $A(\xi)$ be convergent power series for small $|\xi|$. Then the eigenvalues of $A(\xi)$ can be considered as power series in ξ convergent for small $|\xi|$.* Wimmer [132] gives a short proof of the Rellich theorem based on the fact that the ring $H(\Omega)$ of complex functions which are holomorphic in a region Ω is an elementary divisor domain. Besides, the Rellich theorem is extended to normal matrices by Lancaster and Tismenetsky [67, Chapter 11, Theorem 2].

NR 4.5–5. Gerschgorin Theorem. *For $A = (\alpha_{ij}) \in \mathcal{C}^{n\times n}$ let*

$$\mathcal{G}_i(A) = \{z \in \mathcal{C} : |z - \alpha_{ii}| \le \sum_{j\ne i} |\alpha_{ij}|\}.$$

Then

$$\lambda(A) \subset \bigcup_{i=1}^{n} \mathcal{G}_i(A).$$

Moreover, if m of the Gerschgorin discs $\mathcal{G}_i(A)$ are isolated from the other $n - m$ discs, then there are precisely m eigenvalues of A in their union. (See, e.g., Stewart and Sun [97, Chapter IV, Theorem 2.1].)

NR 4.5–6. The relations of (4.5.27) are obtained by Fox and Kapoor [38].

NR 4.5–7. The perturbation analyses of eigenvalues of real symmetric positive definite matrices are made by Polak and Wardi [85], and these analyses can be carried over to the cases of symmetric matrices and bilinear forms (e.g., the *local Lipschitz continuity* of the eigenvalues and the generalized gradient introduced by Clarke [22] of multiple eigenvalues).

NR 4.5–8. Some results on the generalized gradients of multiple eigenvalues of the symmetric-definite generalized eigenproblem (4.5.3) are given by Sun [108].

NR 4.5–9. The following two theorems on singular values are cited from the literature. The first one is used to prove Theorem 4.5.5, and the second one is used to show the fact (ii) in Remark 4.5.9.

Theorem 4.5.16. *Let $K, L \in \mathcal{C}^{m\times n}$ $(m \geq n)$ be given, let the ordered singular values of K, L and KL^H be*

$$\sigma_1(K) \geq \cdots \geq \sigma_n(K), \qquad \sigma_1(L) \geq \cdots \geq \sigma_n(L),$$

and

$$\sigma_1(KL^H) \geq \cdots \geq \sigma_n(KL^H) \geq \sigma_{n+1}(KL^H) = \cdots = \sigma_m(KL^H) = 0,$$

respectively. Then

$$\sigma_j(KL^H) \leq \min_{1\leq k\leq j}\{\sigma_k(K)\sigma_{j-k+1}(L)\}, \quad j = 1, \ldots, n.$$

See Horn and Johnson [55, p.423] for the proof of Theorem 4.5.16.

Theorem 4.5.17. *Let $K, L \in \mathcal{C}^{n\times n}$ be given, let the ordered singular values of K, L and KL be*

$$\sigma_1(K) \geq \cdots \geq \sigma_n(K), \qquad \sigma_1(L) \geq \cdots \geq \sigma_n(L),$$

and

$$\sigma_1(KL) \geq \cdots \geq \sigma_n(KL),$$

respectively. Then

$$\textstyle\prod_{k=1}^{j} \sigma_k(KL) \leq \prod_{k=1}^{j} \sigma_k(K)\sigma_k(L), \quad j = 1, \ldots, n-1,$$

$$\textstyle\prod_{k=1}^{n} \sigma_k(KL) = \prod_{k=1}^{n} \sigma_k(K)\sigma_k(L).$$

This result is proved by Horn [54]. An alternative proof can be found in Marshall and Olkin's book [75].

Bibliography

[1] A. L. Andrew, K.-W. E. Chu, and P. Lancaster, Derivatives of eigenvalues and eigenvectors of matrix functions, *SIAM J. Matrix Anal. Appl.,* 14 (1993), 903–926.

[2] A. L. Andrew, F. R. de Hoog, and R. C. E. Tan, Direct computation of derivatives of eigenvalues and eigenvectors, *Internat. J. Comput. Math.,* 36 (1990), 251–255.

[3] Z. Bai, J. Demmel and A. McKenney, On computing condition numbers for the nonsymmetric eigenproblem, *ACM Trans. Math. Softw.* 19 (1993), 202-223.

[4] E. Berkson, Some metrics on the subspaces of a Banach space, *Pacific J. Math.,* 13 (1963), 7–22.

[5] R. Bhatia and C. Davis, A bound for the spectral variation of a unitary operator, *Linear and Multilinear Algebra,* 15 (1984), 71-76.

[6] R. Bhatia, C. Davis and F. Kittaneh, Some inequalities for commutators and an application to spectral variation, *Aequationes Mathematicae,* 41 (1991), 70-78.

[7] Å. Björck and G. Golub, Numerical methods for computing angles between linear subspaces, *Math. Comput.,* 27 (1973), 579-594.

[8] S. Bochner and W. T. Martin, *Several Complex Variables,* Princeton U. P., 1948.

[9] B. Bohnhorst, A. Bunse-Gerstner, and H. Fassbender, On the perturbation theory for unitary eigenvalue problems, *SIAM J. Matrix Anal. Appl.,* 21 (2000), 809–824.

[10] J. R. Bunch, The weak and strong stability of algorithms in numerical linear algebra, *Linear Algebra Appl.,* 88/89 (1987), 49–66.

[11] J. R. Bunch, J. W. Demmel, and C. F. Van Loan, The strong stability of algorithms for solving symmetric linear systems, *SIAM J. Matrix Anal. Appl.,* 10 (1989), 494-499.

[12] R. Byers, A LINPACK-style condition estimator for the equation $AX - XB^T = C$, *IEEE Trans. Autom. Control,* AC-29 (1984), 926–928.

[13] Z. -H. Cao, Residual error bounds of generalized eigenvalue systems, *Linear Algebra Appl.,* 93 (1987), 113–120.

[14] Z. -H. Cao, J. -J. Xie, and R. -C. Li, A sharp version of Kahan's theorem on clustered eigenvalues, *Linear Algebra Appl.,* 245 (1996), 147–155.

[15] H. Cartan, *Elementary Theory of Analytic Functions of One or Several Complex Variables,* Hermann, Paris, 1963.

[16] F. Chaitin-Chatelin, *Spectral Approximation of Linear Operators,* Academic Press, New York, 1983.

[17] F. Chaitin-Chatelin and V. Frayssé, *Lectures on Finite Precision Computations,* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

[18] S. Chandrasekaran and I. C. F. Ipsen, Backward errors for eigenvalue and singular value decompositions, *Numer. Math.,* 68 (1994), 215–223.

[19] X. -W. Chang, C. Paige, and G. W. Stewart, New perturbation analyses for the Cholesky factorization, *IMA J. Numer. Anal.,* 16 (1996), 457–484.

[20] C. -H. Chen and J. -G. Sun, Perturbation bounds for the polar factors, *J. Comput. Math.,* 7 (1989), 397–401.

[21] K. -w. E. Chu, On multiple eigenvalues of matrices depending on several parameters, *SIAM J. Numer. Anal.,* 27 (1990), 1368–1385.

[22] F. H. Clarke, *Optimization and Nonsmooth Analysis,* John Wiley, New York, 1983.

[23] C. R. Crawford, A stable generalized eigenvalue problem, *SIAM J. Numer. Anal.,* 6 (1976), 854-860.

[24] T. R. Crossley and B. Porter, Eigenvalue and eigenvector sensitivities in linear system theory, *Internat. J. Control,* 10 (1969), 167–170.

[25] C. Davis, An extremal problem for extensions of a sesquilinear form, *Linear Algebra Appl.,* 13 (1976), 91-102.

[26] C. Davis and W. M. Kahan, The rotation of eigenvectors by a perturbation, III. *SIAM J. Numer. Anal.,* 7 (1970), 1–46.

[27] C. Davis, W. M. Kahan, and H. Weinberger, Norm-preserving dilations and their applications to optimal error bounds, *SIAM J. Numer. Anal.,* 19 (1982), 445–469.

[28] J. P. Dedieu, Condition operators, condition numbers and condition number theorem for the generalized eigenvalue problem, *Linear Algebra Appl.,* 263 (1997), 1-24.

[29] J. W. Demmel, On condition numbers and the distance to the nearest ill-posed problem, *Numer. Math.,* 51 (1987), 251–289.

[30] J. E. Dennis and J. J. Moré, Quasi-Newton methods, motivations and theory, *SIAM Rev.,* 19 (1977), 46–89.

[31] E. Deutsch and M. Neumann, On the first and second order derivatives of the Perron vector, *Linear Algebra Appl.,* 71 (1985), 57–76.

[32] J. Dieudonné, *Élements d'Analyse, 1. Fondements de l'Analyse Moderne,* Gauthier-Villars, Paris, 1968.

[33] J. J. Dongarra, S. Hammarling, and J. H. Wilkinson, Numerical considerations in computing invariant subspaces, *SIAM J. Matrix Anal. Appl.,* 13 (1992), 145-161.

[34] L. Elsner and C. He, An algorithm for computing the distance to uncontrollability, *Systems & Control Letters,* 7 (1991), 453–464.

[35] L. Elsner and C. He, Perturbation and interlace theorems for the unitary eigenvalue problem, *Linear Algebra Appl.,* 188/189 (1993), 207–229.

[36] L. Elsner, C. He, and V. Mehrmann, Minimization of the norm, the norm of the inverse and the condition number of a matrix by completion, *Numerical Linear*

Algebra with Applications, 2 (1995), 155–171.

[37] K. Fan and J. Hoffman, Some metric inequalities in the space of matrices, *Proc. Amer. Math. Soc.*, 6 (1955), 111–116.

[38] R. L. Fox and M. P. Kapoor, Rates of change of eigenvalues and eigenvectors, *AIAA J.*, 6 (1968), 2426–2429.

[39] V. Frayssé and V. Toumazou, A note on the normwise perturbation theory for the regular generalized eigenproblem $Ax = \lambda Bx$, *Numer. Linear Algebra Appl.*, 5 (1998), 1–10.

[40] A. J. Geurts, A contribution to the theory of condition, *Numer. Math.*, 39 (1982), 85–96.

[41] G. H. Golub and C. F. Van Loan, *Matrix Computations,* Third Edition, Johns Hopkins University Press, Baltimore and London, 1996.

[42] A. Graham, *Kronecker Products and Matrix Calculus with Applications,* Wiley, New York, 1981.

[43] M. Haviv and Y. Ritov, Bounds on the error of an approximate invariant subspace for non-self-adjoint matrices, *Numer. Math.*, 67 (1994), 491–500.

[44] P. Henrici, Bounds for iterates, inverses, spectral variation and fields of values of nonnormal matrices, *Numer. Math.*, 4 (1962), 24–39.

[45] D. J. Higham and N. J. Higham, Backward error and condition of structured linear systems, *SIAM J. Matrix Anal. Appl.*, 13 (1992), 162–175.

[46] D. J. Higham and N. J. Higham, Structured backward error and condition of generalized eigenvalue problems, *SIAM J. Matrix Anal. Appl.*, 20 (1999), 493–512.

[47] N. J. Higham, Computing the polar decomposition - with applications, *SIAM J. Sci. Stat. Comput.*, 7 (1986), 1160–1174.

[48] N. J. Higham, A survey of condition number estimation for triangular matrices, *SIAM Rev.*, 29 (1987), 575–596.

[49] N. J. Higham, Matrix nearness problems and applications. In *Applications of Matrix Theory,* M. J. C. Gover and S. Barnett, editors, Oxford University Press, Oxford, UK, 1989, pages 1–27.

[50] N. J. Higham, Computing error bounds for regression problems. In *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics* 112, P. J. Brown and W. A. Fuller , ed., American Mathematical Society, Providence, Rhode Island, 1990, 195-208.

[51] N. J. Higham, Perturbation theory and backward error for $AX - XB = C$, *BIT*, 33 (1993), 124–136.

[52] N. J. Higham, *Accuracy and Stability of Numerical Algorithms,* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

[53] M. W. Hirsch, *Differential Topology,* Springer-Verlag, New York, 1976.

[54] A. Horn, On the singular values of a product of completely continuous operators, *Proc. Nat. Acad. Sci. U. S. A.*, 36 (1950), 374–375.

[55] R. Horn and C. Johnson, *Matrix Analysis,* Cambridge University Press, New York 1985.

[56] R. Horn and C. Johnson, *Topics in Matrix Analysis,* Cambridge University Press, New York 1991.

[57] A. S. Householder, *The Theory of Matrices in Numerical Analysis,* Dover Publications, New York, 1964.

[58] B. Kågström and P. Poromaa, Distributed and shared memory block algorithms for the triangular Sylvester equation with sep^{-1} estimators, *SIAM J. Matrix Anal. Appl.,* 13 (1992), 90–101.

[59] B. Kågström and P. Poromaa, LAPACK-Style algorithms and software for solving the generalized Sylvester equation and estimating the separation between regular matrix pairs, *ACM Trans. Math. Software,* 22 (1996), 78–103.

[60] B. Kågström and P. Poromaa, Computing eigenspaces with specified eigenvalues of a regular matrix pair (A, B) and condition estimation: Theory, algorithms and software, *Numerical Algorithms,* 12 (1996), 369–407.

[61] W. M. Kahan, Inclusion theorems for clusters of eigenvalues of Hermitian matrices, Technical Report, Computer Science Department, University of Toronto, 1967.

[62] W. M. Kahan, B. N. Parlett, and E. Jiang, Residual bounds on approximate eigensystems of nonnormal matrices, *SIAM J. Numer. Anal.,* 19 (1982), 470-484.

[63] T. Kato, *Perturbation Theory for Linear Operators,* Springer-Verlag, New York, 1966.

[64] M. G. Kreǐn, The theory of self-adjoint extensions of semi-bounded Hermitian transformations and its applications, *Mat. Sb.,* 20(1947), 431–495; 21 (1947), 365–404 (in Russian).

[65] M. G. Kreǐn and M. A. Krasnoselsky, Fundamental theorems concerning the extension of Hermitian operators and some of their applications to the theory of orthogonal polynomials and the moment problem, *Uspekhi Mat. Nauk.,* 2, 3 (1947). In Russian.

[66] P. Lancaster, On eigenvalues of matrices dependent on a parameter, *Numer. Math.,* 6 (1964), 377–387.

[67] P. Lancaster and M. Tismenetsky, *The Theory of Matrices,* 2nd ed., Academic Press, New York, 1985.

[68] R. -C. Li, On eigenvalues of a Rayleigh quotient matrix, *Linear Algebra Appl.,* 169 (1992), 249–255.

[69] R. -C. Li, On eigenvalue variations of Rayleigh quotient matrix pencils of a definite pencil, *Linear Algebra Appl.,* 208/209 (1994), 471–483.

[70] V. B. Lidskii, Perturbation theory of non-conjugate operators, *U.S.S.R. Comput. Maths. Math. Phys.,* 1 (1965), 73–85.

[71] W. -W. Lin and J. -G. Sun, Perturbation analysis for the eigenproblem of periodic matrix pairs, *Linear Algebra Appl.,* 337 (2001), 157–187.

[72] X. -G. Liu, Differential expansion theory of matrix functions and its applications (in Chinese), Ph.D. Thesis, Computing Center, Academia Sinica, Beijing, 1990.

[73] X. -G. Liu and Y. -G. Xu, On the Rayleigh quotient matrix (in Chinese), *Math. Numer. Sinica,* 12 (1990), 208–213.

[74] A. G. J. MacFarlane and Y. S. Hung, Analytic properties of the singular values of a rational matrix, *Int. J. Control,* 37 (1983), 221–234.

[75] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications,* Academic Press, New York, 1979.

[76] R. Mathias, Quadratic residual bounds for the Hermitian eigenvalue problem, *SIAM J. Matrix Anal. Appl.,* 19 (1998), 541–550.

[77] C. D. Meyer and G. W. Stewart, Derivatives and perturbations of eigenvectors, *SIAM J. Numer. Math.,* 25 (1988), 679–691.

[78] L. Mirsky, Symmetric gauge functions and unitarily invariant norm, *Quart J. Math. Oxford,* 11 (1960), 50–59.

[79] 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 (1997), 793–817.

[80] W. Oettli and W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, *Numer. Math.,* 6 (1964), 405–409.

[81] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables,* Academic Press, New York, 1970.

[82] C. C. Paige, Eigenvalues of perturbed Hermitian matrices, *Linear Algebra Appl.,* 8 (1974), 1–10.

[83] B. N. Parlett, *The Symmetric Eigenvalue Problem,* Prentice-Hall, Englewood Cliffs, NJ, USA, 1980.

[84] S. Parrott, On a quotient norm and the Sz.-Nagy-Foias lifting theorem, *J. Funct. Anal.,* 30 (1978), 311–328.

[85] E. Polak and Y. Wardi, Nondifferentiable optimization algorithm for designing control systems having singular value inequalities, *Automatica,* 18 (1982), 267–283.

[86] F. Rellich, Störungstheorie der Spektralzerlegung, *Math. Ann.,* 113 (1937), 600–619.

[87] F. Rellich, *Perturbation Theory of Eigenvalue Problems,* Gordon & Breach, New York, 1969.

[88] J. R. Rice, A theory of condition, *J. SIAM Numer. Anal.,* 3 (1966), 287–310.

[89] J. L. Rigal and J. Gaches, On the compatibility of a given solution with the data of a linear system, *Journal of the Association for Computing Machinery,* 14 (1967), 543-548.

[90] Y. Saad, *Numerical Methods for Large Eigenvalue Problems: Theory and Algorithms,* John Wiley and Sons, New York, 1992.

[91] G. W. Stewart, Error and perturbation bounds for subspaces associated with certain eigenvalue problems, *SIAM Rev.,* 15 (1973), 727-764.

[92] G. W. Stewart, Gerschgorin theory for the generalized eigenvalue problem $Ax = \lambda Bx$, *Math. Comp.,* 29 (1975), 600–606.

[93] G. W. Stewart, On the perturbation of pseudo-inverses, projections and linear least squares problems, *SIAM Rev.,* 19 (1977), 634–662.

[94] G. W. Stewart, Research, development, and LINPACK, in *Mathematical Software III,* J. R. Rice, ed., Academic Press, New York, 1977, 1–14.

[95] G. W. Stewart, A second order perturbation expansion for small singular values, *Linear Algebra appl.,* 56 (1984), 231–235.

[96] G. W. Stewart, Two simple residual bounds for the eigenvalues of a Hermitian matrix, *SIAM J. Matrix Anal. Appl.,* 12 (1991), 205–208.

[97] G. W. Stewart and J. -G. Sun, *Matrix Perturbation Theory,* Academic Press, New York, 1990.

[98] G. W. Stewart and G. Zhang, Eigenvalues of graded matrices and the condition numbers of a multiple eigenvalue, *Numer. Math.,* 58 (1991), 703–712.

[99] J. -G. Sun, Perturbation analysis for the generalized eigenvalue problem and the generalized singular value problem. In B. Kågström and A. Ruhe, editors, *Matrix Pencils,* pages 221–244, Springer, New York, 1983.

[100] J. -G. Sun, On the perturbation of the eigenvalues of a normal matrix, *Math. Numer. Sinica,* 6 (1984), 334-336 (in Chinese).

[101] J. -G. Sun, Estimation of the separation of two matrices, *J. Comput. Math.,* 2(1984), 189–200; 3 (1985), 19-26.

[102] J. -G. Sun, Eigenvalues and eigenvectors of a matrix dependent on several parameters, *J. Comput. Math.,* 3 (1985), 351–364.

[103] J. -G. Sun, The stability analysis of the solutions of inverse eigenvalue problems, *J. Comput. Math.,* 4 (1986), 345–353.

[104] J. -G. Sun, *Matrix Perturbation Analysis,* Second Edition, Science Press, Beijing, 2001. (First Edition, Science Press, Beijing, 1987.) In Chinese.

[105] J. -G. Sun, A note on simple non-zero singular values, *J. Comput. Math.,* 6 (1988), 258–266.

[106] J. -G. Sun, Sensitivity analysis of zero singular values and multiple singular values, *J. Comput. Math.,* 6 (1988), 325–335.

[107] J. -G. Sun, A note on local behaviors of multiple eigenvalues, *SIAM J. Matrix Anal. Appl.,* 10 (1989), 533–541.

[108] J. -G. Sun, Multiple eigenvalue sensitivity analysis, *Linear Algebra and Appl.,* 137/138 (1990), 183–211.

[109] J. -G. Sun, Perturbation expansions for invariant subspaces, *Linear Algebra and Appl.,* 153 (1991), 85–97.

[110] J. -G. Sun, Perturbation bounds for the Cholesky and QR factorizations, *BIT,* 31 (1991), 341–352.

[111] J. -G. Sun, Eigenvalues of Rayleigh quotient matrices, *Numer. Math.,* 59 (1991), 603–614.

[112] J. -G. Sun, Rayleigh quotient and residual of a definite pair, *J. Comput. Math.,* 9 (1991), 247–255.

[113] J. -G. Sun, On the sensitivity of semisimple multiple eigenvalues, *J. Comput. Math.,* 10 (1992), 193–203.

[114] J. -G. Sun, On condition numbers of a nondefective multiple eigenvalue, *Numer. Math.,* 61 (1992), 265–275.

[115] J. -G. Sun, Backward perturbation analysis of certain characteristic subspaces, *Numer. Math.,* 65 (1993), 357–382.

[116] J. -G. Sun, A note on backward perturbations for the Hermitian eigenvalue problem, *BIT,* 35 (1995), 385–393.

[117] J. -G. Sun, On worst-case condition numbers of a nondefective multiple eigenvalue, *Numer. Math.,* 68 (1995), 373–382.

[118] J. -G. Sun, Residual bounds on approximate solutions for the unitary eigenproblem, *SIAM J. Matrix Anal. Appl.,* 17 (1996), 69–82.

[119] J. -G. Sun, Perturbation analysis of singular subspaces and deflating subspaces, *Numer. Math.,* 73 (1996), 235–263.

[120] J. -G. Sun, Backward errors for the unitary eigenproblem, Technical Report, UMINF 97.25, ISSN-0348-0542, Department of Computing Science, Umeå University, 1997.

[121] J. -G. Sun, Residual bounds of approximate solutions of the algebraic Riccati equation, *Numer. Math.,* 76 (1997), 249–263.

[122] J. -G. Sun, Condition number and backward error for the generalized singular value decomposition, *SIAM J. Matrix Anal. Appl.,* 22 (2000), 323–341.

[123] R. J. Vaccaro, A second-order perturbation expansion for the SVD, *SIAM J. Matrix Anal. Appl.,* 15 (1994), 661-671.

[124] J. M. Varah, On the separation of two matrices, *SIAM J. Numer. Anal.,* 16 (1979), 216–222.

[125] J. von Neumann, Some matrix-inequalities and metrization of matrix-space, *Bull. Inst. Math. Mécan. Univ. Kouybycheff Tomsk,* 1 (1935–1937), 286–300.

[126] J. Y. Wang and B. S. Garbow, A numerical method for solving inverse real symmetric eigenvalue problems, *SIAM J. Sci. Stat. Comput.,* 4 (1983), 45–51.

[127] P. -Å. Wedin, Perturbation bounds in connection with singular value decomposition, *BIT,* 12 (1972), 99—111.

[128] H. Weyl, Das asymtotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen, *Math. Ann.,* 71 (1912), 441–479.

[129] H. Wielandt, Inclusion theorems for eigenvalues. In *Simultaneous Linear Equations and the Determination of Eigenvalues,* Proceedings of a Symposium, Los Angels 1951. Nat. Bur. Standards Appl. Math. Ser. 29 (1953), 75–78. U.S. Government Printing Office, Washington, D.C.

[130] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, England 1965.

[131] J. H. Wilkinson, Note on matrices with a very ill-conditioned eigenproblem, *Numer. Math.,* 19 (1972), 176–178.

[132] H. K. Wimmer, Rellich's perturbation theorem on Hermitian matrices of holomorphic functions, *J. Math. Anal. appl.,* 114 (1986), 52–54.

[133] S. -F. Xu, Lower bound estimation for the separation of two matrices, *Linear Algebra appl.,* 262 (1997), 67–82.

[134] S. -F. Xu, *An Introduction to Inverse Algebraic Eigenvalue Problems,* Peking University Press and Vieweg Publishing, Hong Kong, 1998.

[135] T. Yamamoto, Error bounds for computed eigenvalues and eigenvectors, *Numer. Math.*, 34 (1980), 189–199.

[136] T. Yamamoto, Error bounds for computed eigenvalues and eigenvectors. II, *Numer. Math.*, 40 (1982), 201–206.

[137] E. Zeidler, *Nonlinear Functional Analysis and its Applications, I. Fixed-Point Theorems*, Springer-Verlag, New York, 1986.

Index