The Polynomial Eigenvalue Problem

Berhanu, Michael

2005

MIMS EPrint: **2006.358**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

THE POLYNOMIAL EIGENVALUE PROBLEM

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2005

Michael Berhanu

School of Mathematics

Contents

List of Tables

List of Figures

Abstract

In this thesis, we consider polynomial eigenvalue problems. We extend results on eigenvalue and eigenvector condition numbers of matrix polynomials to condition numbers with perturbations measured with a weighted Frobenius norm. We derive an explicit expression for the backward error of an approximate eigenpair of a matrix polynomial written in homogeneous form. We consider structured eigenvalue condition numbers for which perturbations have a certain structure such as symmetry, Hermitian or sparsity. We also obtain explicit and/or computable expressions for the structured backward error of an eigenpair.

We present a robust implementation of the HZ (or HR) algorithm for symmetric generalized eigenvalue problems. This algorithm has the advantage of preserving pseudosymmetric tridiagonal forms. It has been criticized for its numerical instability. We propose an implementation of the HZ algorithm that allows stability in most cases and comparable results with other classical algorithms for ill conditioned problems. The HZ algorithm is based on the HR factorization, an extension of the QR factorization in which the H factor is hyperbolic. This yields us to the sensitivity analysis of hyperbolic factorizations.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Statement

- The material in Chapter 4 is based on the technical report "Perturbation Bounds for Hyperbolic Matrix Factorizations", Numerical Analysis Report 469, Manchester Centre for Computational Mathematics, June 2005. This work has been submitted for publication in SIAM J. Matrix Anal. Appl.

- The material in Chapter 6 is based on the technical report "A Robust Implementation of the HZ Algorithm" (with Françoise Tisseur), Numerical Analysis Report, Manchester Centre for Computational Mathematics. In Preparation.

Acknowledgements

I am extremely grateful to my supervisor Françoise Tisseur for her help, guidance and for sharing with me her expertise.

I would like to express my gratitude to Nick Higham for his many helpful suggestions and constructive remarks.

Many thanks to my fellow students and friends Matthew Smith, Harikrishna Patel, Craig Lucas, Gareth Hargreaves, Anna Mills and Philip Davis for the enjoyable 3... years in Manchester.

$\epsilon\upsilon\chi\alpha\rho\iota\sigma\tau\omega\ \pi o\lambda\upsilon$ Maria Pampaka, Maria Mastorikou, Panagiotis Kallinikos ("Dr, elare"), *mucha gracias* to the spanish crew, Big Hands,...

Mariella Tsopela thank you for everything, $\phi i\lambda\alpha\kappa\iota\alpha$.

Thanks to my father Berhanu H/W who gave me in my childhood the thirst of knowledge. I am extremely grateful to my sisters Bethlam (Koki), Deborah (Lili), Myriam (Poly). Thanks Lili for your patience and help. Finally, a lot of thanks goes to my mother, Fiorenza Vitali, for her encouragement and unconditional love. I dedicate this thesis to her. *Merci beaucoup.*

Chapter 1

Introduction

We consider the matrix polynomial (or λ-matrix) of degree m

$$P(A, \lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \cdots + A_0, \tag{1.1}$$

where $A_k \in \mathbb{C}^{n \times n}$, $k = 0\colon m$. The polynomial eigenvalue problem (PEP) is to find an eigenvalue λ and corresponding nonzero eigenvector x satisfying

$$P(A, \lambda)x = 0.$$

The case $m = 1$ corresponds to the generalized eigenvalue problem (GEP)

$$Ax = \lambda Bx$$

and if $A_0 = I$ we have the standard eigenvalue problem (SEP)

$$Ax \;\; = \;\; \lambda x. \tag{1.2}$$

Another important case is the quadratic eigenvalue problem (QEP) with $m = 2$.

The importance of PEPs lies in the diverse roles they play in the solution of problems in science and engineering. We briefly outline some examples.

1.1 Applications of PEPs

QEPs and more generally PEPs appear in a variety of problems in a wide range of applications. There are numerous examples where PEPs arise naturally.

Some physical phenomena are modeled by a second order ordinary differential equation (ODE) with matrix coefficients

$$M\ddot{z} + D\dot{z} + Kz = f(t), \tag{1.3}$$

$$z(0) = a, \tag{1.4}$$

$$\dot{z}(0) = b. \tag{1.5}$$

The solutions of the homogeneous equation are of the form $e^{\lambda t}u$, with u a constant vector. This leads to the QEP

$$(\lambda^2 M + \lambda D + K)u = 0. \tag{1.6}$$

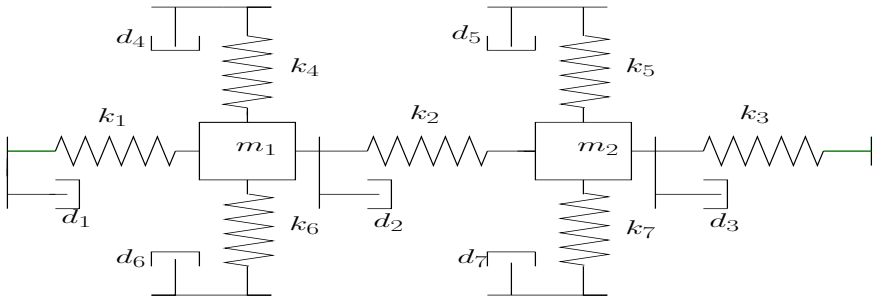

Figure 1.1: A 2 degree of freedom mass-spring damped system.

A well known example is the damped mass-spring system. In Figure 1.1, we consider the 2 degree of freedom mass-spring damped system. The dynamics of this system, under some assumptions, are governed by an ODE of the form (1.3)-(1.5). In this case $z = (x_1, y_1, x_2, y_2)$ denotes the coordinates of the masses m_1 and

m_2, $M = \mathrm{diag}(m_1, m_1, m_2, m_2)$ is the mass matrix, $D = \mathrm{diag}(d_1 + d_2, d_4 + d_6, d_2 + d_3, d_5 + d_7)$ is the damping matrix and $K = \mathrm{diag}(k_1 + k_2, k_4 + k_6, k_2 + k_3, k_5 + k_7)$ is the stiffness matrix with $d_i > 0$, $k_i > 0$ for $1 \leq i \leq 7$.

QEPs arise in structural mechanics, control theory, fluid mechanics and we refer to Tisseur and Meerbergen's survey [73] for more specific applications. Interesting practical examples of higher order PEPs are given in [52].

1.2 Notations

1.2.1 General Notations

$\mathbb{K}$ denotes the field $\mathbb{R}$ or $\mathbb{C}$.

- The colon notation: "$i = 1\colon n$" means the same as "$i = 1, 2, \ldots, n$".
- $\bar{\alpha}$ denotes the conjugate of the complex number α.
- $\mathbb{K}^{m\times n}$ denotes the set of $m \times n$ matrices with coefficients in $\mathbb{K}$.
- $\mathcal{M}_n(\mathbb{K})^m$ denotes the set of m-tuples of $n \times n$ matrices with coefficients in $\mathbb{K}$.
- For $x \in \mathbb{K}^n$, $x = (x_k)_{1\leq k\leq n} = (x_k)$, x_k denotes the kth component of x.
- e_k denotes the vector with the kth component equal to 1 and all the other entries are zero.
- For $A \in \mathbb{K}^{m\times n}$, $A = (\alpha_{ij})_{1\leq i\leq m,\ 1\leq j\leq n} = (\alpha_{ij})$, α_{ij} denotes the (i, j) element of A.
- We often use the tilde notation to denote a perturbed quantity and the hat notation to denote a computed quantity.

1.2.2 Matrix Notation and Special Matrices

Let $A \in \mathbb{K}^{m\times n}$, $A = (\alpha_{ij})$.

- A is a square matrix if $m = n$.
- $A^T \in \mathbb{K}^{n\times m}$ is the transpose of A and it is defined by $A^T = (\alpha_{ji})$.
- A is symmetric if $A^T = A$.
- A is J-symmetric if JA is symmetric for some $J \in \mathbb{R}^{n\times m}$.
- A is skewsymmetric if $A^T = -A$.
- $A^* \in \mathbb{K}^{n\times m}$ is the conjugate transpose of A and it is defined by $A^* = (\bar{\alpha}_{ji})$.
- A is Hermitian if $A^* = A$.
- A is skew-Hermitian if $A^* = -A$.
- A is diagonal if $\alpha_{ij} = 0$ for $i \neq j$.
- The identity matrix of order n, I_n or simply I , is the diagonal matrix that has all its diagonal entries equal to 1.
- A permutation matrix is a matrix obtained from the identity matrix by row or column permutation.
- $A \in \mathbb{K}^{m\times n}$ with $m \neq n$ is upper trapezoidal if $\alpha_{ij} = 0$ for $i > j$.
- A square matrix A is upper triangular if $\alpha_{ij} = 0$ for $i > j$ and lower triangular if $i < j$. If all the diagonal elements of A are equal to 1 then A is called a unit upper or lower triangular.
- A is an upper Hessenberg matrix if $\alpha_{ij} = 0$ for $i > j + 1$.

- A is a tridiagonal matrix if A and A^T are upper Hessenberg matrices.
- For a square matrix A, A^{-1} denotes its inverse. It is the unique matrix such that $A^{-1}A = A^{-1} = I$. A is also said to be nonsingular when A^{-1} exits. Otherwise A is singular.
- For $B = (b_{ij}) \in \mathbb{K}^{m\times n}$ the *Schur product* is defined by $A \circ B = (a_{ij}b_{ij})$.
- For $B = (b_{ij}) \in \mathbb{K}^{p\times q}$ the *Kronecker product* is defined by $A \otimes B = (a_{ij}B)$.

1.3 Mathematical Background

We recall in this Section some mathematical properties of norms, linear spaces and differentiable functions. A particular attention is given to the linear vector spaces $\mathbb{K}^n$ and $\mathbb{K}^{m\times n}$. In the rest of this chapter, $\mathcal{E}$ denotes a linear vector space over $\mathbb{K}$, $\mathbb{K}^n$ or $\mathbb{K}^{m\times n}$.

1.3.1 Linear Algebra

Let $V = \{v_1, \ldots, v_n\}$ where $v_k \in \mathcal{E}$ for $1 \leq k \leq n$. The linear subspace generated by V is defined by

$$\text{span}V = \left\{ \sum_{k=1}^{n} \alpha_k v_k,\ \alpha_k \in \mathbb{K} \right\}.$$

A linear combination is a vector of the type

$$\sum_{k=1}^{n} \alpha_k v_k,$$

where $(\alpha_1, \ldots, \alpha_2) \in \mathbb{K}^n$. The vectors in V are said to be linearly independent if

$$\sum_{k=1}^{n} \alpha_k v_k = 0 \ \Rightarrow\ \alpha_k = 0 \text{ for } k = 1\colon n.$$

The number of linearly independent vectors is the dimension of span$V$ in $\mathbb{K}$ and it is denoted by

$$\dim(V) = \dim_{\mathbb{K}}(V).$$

Let V_1 and V_2 be two linear subspaces of $\mathcal{E}$. If $V_1 \cap V_2 = \{0\}$ and $\mathcal{E} = V_1 + V_2$ then $\mathcal{E}$ is said to be the direct sum of V_1 and V_2 and the direct sum decomposition is denoted by

$$\mathcal{E} = V_1 \oplus V_2.$$

Let $A : \mathcal{E}_1 \to \mathcal{E}_2$ be a linear map or a matrix.

- The range of A is the linear subspace defined by

$$\text{range}(A) = \{y \in \mathcal{E}_2 : y = Ax,\ x \in \mathcal{E}_1\} = A(\mathcal{E}_1).$$

- The null space of A is the linear subspace defined by

$$\text{null}(A) = \{x \in \mathcal{E}_1 :\ Ax = 0\}.$$

- The rank of A is the dimension of range(A),

$$\text{rank}(A) = \dim(\text{range}(A)).$$

- With these notations, it follows that

$$\dim(\mathcal{E}_1) = \text{rank}(A) + \dim(\text{null}(A)).$$

- $A \in \mathbb{K}^{m\times n}$ is of full rank if $\text{rank}(A) = \min(m, n)$. If $\text{rank}(A) < \min(m, n)$ then A is rank deficient.

1.3.2 Normed Linear Vector Spaces

Definition 1.1 *Let $\mathcal{E}$ be a linear vector space. A norm is a map $\|\cdot\| : \mathcal{E} \longrightarrow \mathbb{R}$ satisfying the following properties:*

1. $\|x\| \geq 0$ *with equality if and only if* $x = 0$,

2. $\forall(\lambda, x) \in \mathbb{K} \times \mathcal{E},\ \|\lambda x\| = |\lambda|\|x\|$,

3. $\forall(x, y) \in \mathcal{E}^2,\ \|x + y\| \leq \|x\| + \|y\|$.

For $x \in \mathcal{E}$, $\mathcal{V}_x$ denotes an open neighborhood of x. The open ball of radius $\epsilon \geq 0$ centered at x is defined by

$$B(x, \epsilon) = \{y \in \mathcal{E}, \|y - x\| \leq \epsilon\}.$$

In this thesis, only $\mathcal{E} = \mathbb{K}^n$ and $\mathcal{E} = \mathbb{K}^{m\times n}$ are the spaces considered. Thus, all the norms are equivalent meaning that for any norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ on $\mathcal{E}$, there exists $\mu_1 > 0,\ \mu_2 > 0$ such that

$$\mu_1\|\cdot\|_\alpha \leq \|\cdot\|_\beta \leq \mu_2\|\cdot\|_\beta.$$

1.3.3 Scalar Product and Scalar Product Spaces

In this thesis, $\langle\cdot,\cdot\rangle$ denotes a bilinear form (respectively a sesquilinear form) over $\mathcal{E} \times \mathcal{E}$ if $\mathbb{K} = \mathbb{R}$ (respectively $\mathbb{K} = \mathbb{C}$). Let $M \in \mathbb{K}^{n\times n}$ be nonsingular. The form $\langle\cdot,\cdot\rangle_M$ is defined by $\langle x, y\rangle_M = \langle x, My\rangle = y^*M^*x$ for all $x, y \in \mathbb{K}^n$. In what follows, we assume that the form $\langle\cdot,\cdot\rangle_M$ is symmetric if $\mathbb{K} = \mathbb{R}$, that is

$$\langle x, y\rangle_M = \langle y, x\rangle_M$$

or Hermitian if $\mathbb{K} = \mathbb{C}$

$$\langle y, x\rangle_M = \overline{\langle x, y\rangle}_M.$$

Definition 1.2 *In this thesis, we say that the symmetric or Hermitian form* $\langle\cdot,\cdot\rangle_M$ *is a scalar product if* $\langle\cdot,\cdot\rangle_M$ *is positive definite, that is,*

$$\forall x \in \mathcal{E} \setminus \{0\},\ \langle x, x\rangle_M > 0. \tag{1.7}$$

Otherwise, we refer to $\langle\cdot,\cdot\rangle_M$ *as an indefinite scalar product.*

In the rest of this paragraph, we only consider definite positive scalar products. The Cauchy-Schwartz inequality

$$\forall (x,y) \in \mathcal{E}^2,\ |\langle x,y\rangle| \leq \sqrt{\langle x,x\rangle}\sqrt{\langle y,y\rangle}, \tag{1.8}$$

applies to any definite positive scalar product. Then, following Definition 1.1 and using (1.8), $x \mapsto \sqrt{\langle x,x\rangle}$ defines a norm over $\mathcal{E}$. This norm is known a the 2-norm and it is usually denoted by $\|\cdot\|_2$.

Definition 1.3 *For a given scalar product, matrices that preserve the scalar product are called orthogonal if $\mathbb{K} = \mathbb{R}$ or unitary if $\mathbb{K} = \mathbb{C}$. $\mathcal{O}_n$ (respectively $\mathcal{U}_n$) denotes the set of $n \times n$ orthogonal matrices (respectively the set of $m \times m$ unitary matrices). It follows immediately that*

$$\begin{aligned} Q^T Q &= I_n,\ Q \in \mathcal{O}_n, \\ Q^* Q &= I_n,\ Q \in \mathcal{U}_n. \end{aligned}$$

For $\mathcal{F} \subset \mathcal{E}$, $\mathcal{F}^\perp$ denotes the orthogonal complement of $\mathcal{F}$ and it is defined by

$$\mathcal{F}^\perp = \{x \in \mathcal{E} :\ \langle x,y\rangle = 0,\ \forall y \in \mathcal{F}\}.$$

If $\mathcal{F}$ is a linear subspace of $\mathcal{E}$ then we have the direct sum decomposition

$$\mathcal{E} = \mathcal{F} \oplus \mathcal{F}^\perp.$$

1.3.4 Matrices, Vectors and their Norms

$(x,y) \mapsto \langle x,y\rangle = y^* x$ is the usual scalar product over $\mathbb{K}^n$. The induced vector 2-norm is denoted by $\|\cdot\|_2$ and it is defined by

$$\|x\|_2 = \left(\sum_{k=1}^{n} |x_k|^2\right)^{\frac{1}{2}} = \sqrt{x^* x}.$$

Other useful norms over $\mathbb{K}^n$ are given by

$$\begin{aligned} \|x\|_1 &= \sum_{k=1}^{n} |x_k|, \\ \|x\|_\infty &= \max_{1\le k\le n} |x_k|. \end{aligned}$$

Let $A = (a_{ij}) \in \mathbb{K}^{m\times n}$. The subordinated matrix norm of A is defined by

$$\|A\|_{\alpha,\beta} = \sup_{x\neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\beta},$$

where $\|\cdot\|_\alpha$ is a norm over $\mathbb{K}^m$ and $\|\cdot\|_\beta$ is a norm over $\mathbb{K}^n$. It follows that

$$\begin{aligned} \|A\|_1 &= \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|, \\ \|A\|_2 &= \sqrt{\rho(A^*A)}, \\ \|A\|_\infty &= \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|, \end{aligned}$$

where for $X \in \mathbb{K}^{n\times n}$, the spectral radius $\rho(X)$ is

$$\rho(X) = \max\{|\lambda|,\ \det(X - \lambda I) = 0\}.$$

The matrix subordinated 2-norm is invariant under orthogonal or unitary transformations,

$$\|Q_1 X Q_2\|_2 = \|X\|_2,$$

for all $X \in \mathbb{K}^{m\times n}$ and orthogonal or unitary $Q_1,\ Q_2$.

The trace of a square matrix is the sum of its diagonal elements and for $X \in \mathbb{K}^{n\times n}$, $X = (x_{ij})$ it is denoted by

$$\mathrm{trace}(X) = \sum_{k=1}^{n} x_{kk}.$$

$(X, Y) \mapsto \mathrm{trace}(Y^*X)$ is the usual scalar product over $\mathbb{K}^{m\times n}$. The induced matrix norm is known as the Frobenius norm and it is defined by

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

The Frobenius norm is invariant under orthogonal or unitary transformations,

$$\|UXV\|_F = \|X\|_F,$$

for all $X \in \mathbb{K}^{m\times n}$, $U \in \mathcal{U}_m$ and $V \in \mathcal{U}_n$.

Definition 1.4 *Let $\mu = (\frac{1}{\mu_k})_{0\leq k\leq m}$, with $\mu_k > 0$. The μ-weighted Frobenius norm is induced by the inner-product over $\mathcal{M}_n(\mathbb{C})^{m+1}$,*

$$\langle A, B\rangle = \text{trace}\left(\sum_{k=0}^{m} \frac{1}{\mu_k} B_k^* A_k\right)$$

and it is denoted by $\|A\|_{F,\mu} = \sqrt{\langle A, A\rangle}$.

The μ-weighted 2-norm is defined by,

$$\|A\|_{2,\mu} = \left(\sum_{k=0}^{m} \left\|\frac{A_k}{\mu_k}\right\|_2^2\right)^{\frac{1}{2}}.$$

1.3.5 Differential Calculus

Let $f : \mathcal{E} \longrightarrow \mathcal{F}$, where $\mathcal{E}$, $\mathcal{F}$ are two normed vector spaces. f is differentiable or Fréchet differentiable at $x \in \mathcal{V}_x \subset \mathcal{E}$, where $\mathcal{V}_x$ is an open neighborhood of x if there exists a linear map $df(x) : \mathcal{E} \longrightarrow \mathcal{F}$, such that

$$\lim_{\|h\|\to 0} \frac{1}{\|h\|}(f(x+h) - f(x) - df(x)h) = 0.$$

In this thesis, we only consider the case where $\mathcal{E}$ has a finite dimension. Thus, if f is linear, then f is differentiable and $df = f$. All the vector spaces are vector spaces on $\mathbb{R}$ and thus all the functions are considered as functions of real variables and the differentiation is real. The following theorem is the well-known implicit function theorem [4], [63] that we are going to use several times in this thesis.

Theorem 1.1 *Let*

$$\begin{array}{rccc} f: & E\times F & \to & G \\ & (x,y) & \mapsto & f(x,y) \end{array}$$

be differentiable, where E, F and G are normed vector spaces. Assume that $f(x,y) = 0$ and that $\frac{\partial f}{\partial y}(x,y)$ is nonsingular for some $(x,y) \in E \times F$. Then, there exist a neighborhood of x, $\mathcal{V}_x$, a neighborhood of y, $\mathcal{V}_y$ and a differentiable function $\varphi : \mathcal{V}_x \to \mathcal{V}_y$ such that $y = \varphi(x)$ and for all $\widetilde{x} \in \mathcal{V}_x$, $f(\widetilde{x}, \varphi(\widetilde{x})) = 0$. Moreover,

$$d\varphi(x) = \left(\frac{\partial f}{\partial y}(x,y)\right)^{-1} \frac{\partial f}{\partial x}(x,y).$$

Definition 1.5 *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^p$. Assume that* $\mathrm{rank}(df(x)) = p$ *whenever $f(x) = 0$. Then, $f^{-1}(\{0\})$ is a $(n-p)$-dimensional manifold in $\mathbb{R}^n$.*

We now give a fundamental result from optimization, the Lagrange multipliers theorem [4].

Theorem 1.2 *Let $g : E \to \mathbb{R}$ be differentiable, where E is a normed vector spaces of finite dimension n. Let $\mathcal{S} \subset E$ be a differentiable manifold of dimension d defined by*

$$\mathcal{S} = \{y \in E,\ f_k(y) = 0,\ k = 1\!:\! n-d\}\,.$$

Assume that $x \in \mathcal{S}$ is an extremum of g on $\mathcal{S}$. Then, there exist $n-d$ scalars c_k, $k = 1\!:\! n-d$, such that

$$dg(x) = \sum_{k=1}^{n-d} c_k df_k(x).$$

We refer to [4] and [63] for a more detailed presentation of differential calculus and manifolds.

1.4 Special Matrix Subsets

$\triangle(\mathbb{K})$ denotes the set of upper triangular matrices in $\mathbb{K}^{n\times n}$ with a real diagonal. $\mathbf{Sym}(\mathbb{K})$ and $\mathbf{Skew}(\mathbb{K})$ are the linear subspaces of symmetric matrices and skew-symmetric matrices, respectively, with coefficients in $\mathbb{K}$. **Herm** and **SkewH**

are the linear subspaces of Hermitian matrices and skew-Hermitian matrices, respectively. dim denotes the dimension of a linear space in $\mathbb{R}$. We recall that

$$\begin{aligned} \dim \triangle(\mathbb{R}) &= \dim \mathbf{Sym}(\mathbb{R}) = \frac{n^2+n}{2}, && (1.9)\\ \dim \triangle(\mathbb{C}) &= \dim \mathbf{Herm} = \dim \mathbf{SkewH} = n^2, && (1.10)\\ \dim \mathbf{Skew}(\mathbb{R}) &= \frac{n^2-n}{2}, && (1.11)\\ \dim \mathbf{Sym}(\mathbb{C}) &= n^2+n, \quad \dim \mathbf{Skew}(\mathbb{C}) = n^2-n. && (1.12) \end{aligned}$$

Note that $\mathbf{SkewH} = i\mathbf{Herm}$. For $x \in \mathbb{K}^n$, $\mathrm{diag}(x)$ denotes the $n \times n$ diagonal matrix with diagonal x. For $X \in \mathbb{K}^{n\times n}$, we denote $\Pi_d(X)$ the diagonal part, $\Pi_u(X)$ the strictly upper triangular part and $\Pi_l(X)$, the strictly lower triangular part of X.

1.5 $(J, \widetilde{J})$-Orthogonal and $(J, \widetilde{J})$-Unitary Matrices

We denote by $\mathrm{diag}_n^k(\pm 1)$ the set of all $n \times n$ diagonal matrices with k diagonal elements equal to 1 and $n-k$ equal to -1. A matrix $J \in \mathrm{diag}_n^k(\pm 1)$ for some k is called a *signature* matrix. A matrix $H \in \mathbb{R}^{n\times n}$ is said to be $(J, \widetilde{J})$-*orthogonal* if $H^T J H = \widetilde{J}$, where $J,\ \widetilde{J} \in \mathrm{diag}_k^n(\pm 1)$. We denote by $\mathcal{O}_n(J, \widetilde{J})$ the set of $n \times n$ $(J, \widetilde{J})$-orthogonal matrices. If $J = \widetilde{J}$ then we say that H is J-orthogonal or pseudo-orthogonal and the set of J-orthogonal matrices is denoted by $\mathcal{O}_n(J)$. We say that a matrix is *hyperbolic* if it is $(J, \widetilde{J})$-orthogonal or pseudo-orthogonal with $J \neq \pm I$. We recall that if $J = \pm I$, then $\mathcal{O}_n(\pm I) = \mathcal{O}_n$ is the set of orthogonal matrices.

We extend the definition of $(J, \widetilde{J})$-orthogonal matrices to rectangular matrices in $\mathbb{R}^{m\times n}$, with $m \geq n$. $H \in \mathbb{R}^{m\times n}$ is $(J, \widetilde{J})$-orthogonal if $H^T J H = \widetilde{J}$ with

$J \in \text{diag}_m^k(\pm 1)$ and $J \in \text{diag}_n^q(\pm 1)$. We denote by $\mathcal{O}_{mn}(J, \widetilde{J})$ the set of $(J, \widetilde{J})$-orthogonal in $\mathbb{R}^{m\times n}$.

The definition of signature matrices can be extended and generalized to complex signature matrices. Let $\mathbb{U} = \{z \in \mathbb{C} : \ |z| = 1\}$ denote the unit circle in $\mathbb{C}$. We define the set of complex signature matrices as diagonal matrices such that each diagonal entry is in $\mathbb{U}$ and we denote the set of $n \times n$ complex signature matrices by $\text{diag}_n(\mathbb{U})$.

$(J, \widetilde{J})$-unitary matrices are the complex counterpart of $(J, \widetilde{J})$-orthogonal matrices and we say that a matrix $H \in \mathbb{K}^{n\times n}$ is $(J, \widetilde{J})$-unitary matrix if $H^*JH = \widetilde{J}$ where J and $\widetilde{J}$ are complex signature matrices. We denote by $\mathcal{U}_n(J, \widetilde{J})$ the set of $n\times n$ $(J, \widetilde{J})$-unitary matrices. A similar set is the set of complex $(J, \widetilde{J})$-orthogonal matrices that we denote by $\mathcal{O}_n(J, \widetilde{J}, \mathbb{C})$. We say that a matrix $H \in \mathbb{K}^{n\times n}$ is complex $(J, \widetilde{J})$-orthogonal if $H^TJH = \widetilde{J}$, where $J, \ \widetilde{J} \in \text{diag}_n(\mathbb{U})$. Similarly, we denote by $\mathcal{U}_{mn}(J, \widetilde{J})$ we denote the set of $m \times n$ $(J, \widetilde{J})$-unitary matrices and by $\mathcal{O}_{mn}(J, \widetilde{J}, \mathbb{C})$ the set of $m \times n$ complex $(J, \widetilde{J})$-orthogonal matrices.

We show that $\mathcal{O}_{mn}(J, \widetilde{J}), \mathcal{U}_{mn}(J, \widetilde{J})$ and $\mathcal{O}_{mn}(J, \widetilde{J}, \mathbb{C})$ can respectively be identified to $\mathbb{R}^d$, $\mathbb{R}^{n^2}$ and $\mathbb{R}^{2d}$, with $d = \frac{n^2-n}{2}$. We show that each of these sets are manifolds and we compute their dimension. Then, the introduction of local coordinate systems enable us to make the identification mentioned above.

Lemma 1.3 *$\mathcal{O}_n(J, \widetilde{J})$, $\mathcal{U}_n(J, \widetilde{J})$ and $\mathcal{O}_n(J, \widetilde{J}, \mathbb{C})$ are manifolds with respective dimension d, n^2 and $2d$ with $d = \frac{n^2-n}{2}$.*

Proof. Let $q_1 : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ and $q_2, q_3 : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ be defined by $q_1(X) = X^TJX - \widetilde{J}$, $q_2(X) = X^*JX - \widetilde{J}$ and $q_3(X) = X^TJX - \widetilde{J}$. We recall that $\mathcal{O}_n(J, \widetilde{J}) = q_1^{-1}(\{0\})$, $\mathcal{U}_n(J, \widetilde{J}) = q_2^{-1}(\{0\})$, and $\mathcal{O}_n(J, \widetilde{J}, \mathbb{C}) = q_3^{-1}(\{0\})$. For $1 \le k \le 3$, q_k is clearly differentiable. We have that

$$dq_1(H_1)\Delta H_1 \ = \ H_1^TJ\Delta H_1 + \Delta H_1^TJH_1,$$

$$
\begin{aligned}
dq_2(H_2)\Delta H_2 &= H_2^* J \Delta H_2 + \Delta H_2^* J H_2, \\
dq_3(H_3)\Delta H_3 &= H_3^T J \Delta H_3 + \Delta H_3^T J H_3.
\end{aligned}
$$

To compute the dimension of the three manifolds, we need to determine their tangent spaces that is the null space of each $dq_k(H_k)$, $k = 1\colon 3$, with H_k being in one of these manifolds. We have that

$$
\begin{aligned}
\text{null}(dq_1(H)) &= JH^{-T}\mathbf{Skew}(\mathbb{R}), \\
\text{null}(dq_2(H)) &= JH^{-*}\mathbf{SkewH}, \\
\text{null}(dq_3(H)) &= JH^{-T}\mathbf{Skew}(\mathbb{C}).
\end{aligned}
$$

Thus, following the dimensions given by (1.9)-(1.12), $\mathcal{O}_n(J, \widetilde{J})$ is a $\frac{n^2-n}{2}$ dimensional manifold, $\mathcal{U}_n(J, \widetilde{J})$ is a n^2 dimensional manifold and $\mathcal{O}_n(J, \widetilde{J}, \mathbb{C})$ is $n^2 - n$ a dimensional manifold. □

Let $X \in \mathcal{O}_{mn}(J, \widetilde{J})$, $Y \in \mathcal{U}_{mn}(J, \widetilde{J})$ and $Z \in \mathcal{O}_{mn}(J, \widetilde{J}, \mathbb{C})$. There exists differentiable one-to-one functions ϕ_k, $1 \leq k \leq 3$, open sets $\mathcal{V}_1 \subset \mathbb{R}^d$, $\mathcal{V}_2 \subset \mathbb{R}^{n^2}$, $\mathcal{V}_3 \subset \mathbb{R}^{2d}$, $\mathcal{V}_X \subset \mathbb{R}^{m\times n}$, $\mathcal{V}_Y \subset \mathbb{C}^{m\times n}$ and $\mathcal{V}_Z \subset \mathbb{C}^{m\times n}$ such that

$$
\begin{aligned}
\phi_1(\mathcal{V}_1) &= \mathcal{V}_X \cap \mathcal{O}_n(J, \widetilde{J}), && (1.13) \\
\phi_2(\mathcal{V}_2) &= \mathcal{V}_Y \cap \mathcal{U}_n(J, \widetilde{J}), && (1.14) \\
\phi_3(\mathcal{V}_3) &= \mathcal{V}_Z \cap \mathcal{O}_n(J, \widetilde{J}, \mathbb{C}). && (1.15)
\end{aligned}
$$

Moreover, the differential of these maps ϕ_k have full rank over the entire space where they are defined.

1.6 Matrix Operators Properties

For an operator or a linear map $\mathcal{T}$ defined on $\mathbb{K}^{n\times n}$, the 2-norm is defined by

$$
\|\mathcal{T}\|_2 = \sup_{\|X\|_F = 1} \|\mathcal{T}(X)\|_F.
$$

Some authors denote this norm by $\|\cdot\|_{F,F}$. The choice of this norm is justified by its differentiability properties and its computational simplicity. We now present some notations and we give some results that are needed throughout this thesis.

Theorem 1.4 *Let $A, B, X \in \mathbb{K}^{n\times n}$. We define the operators $\mathcal{T}_2 X = X \circ A$ and $\mathcal{T}_1 X = AXB$. Then,*

$$\|\mathcal{T}_2\|_2 = \max_{ij} |a_{ij}|, \tag{1.16}$$

$$\|\mathcal{T}_1\|_2 = \|(A \otimes B)\|_2 = \|A\|_2 \|B\|_2, \tag{1.17}$$

If A and B are nonsingular then

$$\min_{\|X\|_F=1} \|\mathcal{T}_1(X)\|_F = \|A^{-1}\|_2^{-1} \|B^{-1}\|_2^{-1}. \tag{1.18}$$

Proof. It is straightforward to show that the right hand side of (1.16) is an upper bound for $\|\mathcal{T}_2\|_2$. Let $|a_{pq}| = \max_{ij} |a_{ij}|$. Then, the bound is attained by $e_p e_q^T$.

Let $A = Q_1 S_1 Z_1^T$ and $B = Q_2 S_2 Z_2^T$ be the singular value decompositions of A and B. Then $(A \otimes B) = (Q_1 \otimes Q_2)(S_1 \otimes S_2)(Z_1^T \otimes Z_2^T)$ so that

$$\|(A \otimes B)\|_2 = \|(S_1 \otimes S_2)\|_2 = \|A\|_2 \|B\|_2$$

proving the second part of (1.17). We have

$$\begin{aligned} \|\mathcal{T}_1(X)\|_F &= \|(A \otimes B)\mathrm{vec}(X)\|_2, \\ \|\mathcal{T}_1\|_2 &= \|(A \otimes B)\|_2 = \|A\|_2 \|B\|_2. \end{aligned}$$

Similarly, for (1.18), we have

$$\begin{aligned} \min_{\|X\|_F=1} \|\mathcal{T}_1(X)\|_F &= \min_{\|X\|_F=1} \|(S_1 \otimes S_2)\mathrm{vec}(Z_2 X Z_1^T)\|_F, \\ &= \|A^{-1}\|_2^{-1} \|B^{-1}\|_2^{-1}. \end{aligned}$$

□

We now focus on particular matrix equations that arise in the following chapters. Let $A \in \mathbb{R}^{n\times n}$ be diagonalizable with the eigendecomposition

$$A = VDV^{-1}, \quad \text{with} \quad D = \operatorname{diag}(\lambda_k).$$

For $X \in \mathbf{Skew}(\mathbb{K})$, we consider the equation

$$AX \pm XA^T = Z_\pm,$$

where $Z_+ \in \mathbf{Skew}(\mathbb{K})$ and $Z_- \in \mathbf{Sym}(\mathbb{K})$. The two matrix operators that arise naturally are then defined on $\mathbf{Skew}(\mathbb{K})$ by

$$\mathcal{T}_\pm(A)X = AX \pm XA^T. \tag{1.19}$$

Let $\mathcal{F}$ and $\mathcal{G}$ be the linear subspaces of $\mathbf{Sym}(\mathbb{K})$ defined by

$$\begin{aligned} \mathcal{F} &= \{Y \in \mathbf{Sym}(\mathbb{K}):\ \Pi_d(V^{-1}YV^{-T}) = 0\}, \\ \mathcal{G} &= \{Y \in \mathbf{Sym}(\mathbb{K}):\ \Pi_d(Y) = 0\}. \end{aligned}$$

We define $M_\pm \in \mathbb{C}^{n\times n}$

$$M_\pm = \left(\frac{1}{\lambda_i \pm \lambda_j}\right)_{ij}. \tag{1.20}$$

Theorem 1.5 *Let $A \in \mathbb{R}^{n\times n}$ be diagonalizable, $A = VDV^{-1}$ with $D = \operatorname{diag}(\lambda_k)$ and let $\mathcal{T}_\pm(A)$ be the operator defined by* (1.19)*. Then,*

(i) $\mathcal{T}_+(A)$: $\mathbf{Skew}(\mathbb{K}) \longrightarrow \mathbf{Skew}(\mathbb{K})$ *is invertible if for all k_1, k_2, such that $1 \le k_1, k_2 \le n$ and $k_1 \neq k_2$, we have $\lambda_{k_1} + \lambda_{k_2} \neq 0$.*

(ii) $\mathcal{T}_-(A)$: $\mathbf{Skew}(\mathbb{K}) \longrightarrow \mathcal{F}$ *is invertible if the eigenvalues of A are distinct.*

Then, when $\mathcal{T}_\pm(A)^{-1}$ exists,

$$\mathcal{T}_\pm(A)^{-1}Z_\pm = V((V^{-1}Z_\pm V^{-T}) \circ M_\pm)V^T,$$

where $Z_+ \in \mathbf{Skew}(\mathbb{K})$, $Z_- \in \mathbf{Sym}(\mathbb{K})$ and $M_\pm$ is defined by (1.20)*.*

Proof. We consider the equation $\mathcal{T}_{\pm}(A)X = Z_{\pm}$. We have

$$VDV^{-1}X \pm XV^{-T}DV^T = Z_{\pm}, \quad D\widetilde{X} \pm \widetilde{X}D = \widetilde{Z}_{\pm},$$

where $\widetilde{X} = V^{-1}XV^{-T}$ and $\widetilde{Z}_{\pm} = V^{-1}Z_{\pm}V^{-T}$. Since $\widetilde{X}$ is complex skew-symmetric and D is diagonal we have $\Pi_d(D\widetilde{X} \pm \widetilde{X}D) = 0$. Also $\Pi_d(\widetilde{Z}_{\pm}) = 0$. Thus, if the eigenvalues have the properties required in each case then the solution exists and is unique. It is given by $X_{\pm} = V(\widetilde{Z}_{\pm} \circ M_{\pm})V^T$.

If $\mathbb{K} = \mathbb{R}$, we need to show now that $X_{\pm}$ is real. Without loss of generality, assume that

$$\begin{aligned} V &= [\, V_1 \quad V_2 \quad \overline{V}_2 \,], \quad V^{-T} = [\, U_1^T \quad U_2^T \quad \overline{U}_2^T \,] \quad \text{and} \\ D &= \operatorname{diag}(D_1, D_2, \overline{D}_2), \end{aligned}$$

where V_1, U_1 and D_1 are real and V_2, U_2 and D_2 are complex with a nontrivial imaginary part. Then, $V = \overline{V}P$ and $V^T = PV^*$, where

$$P = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}$$

is partitioned conformably to V. For $Y \in \mathbb{C}^{n\times n}$, $P\overline{Y}P = Y$ if and only if

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \overline{Y}_{12} \\ Y_{21} & Y_{22} & Y_{23} \\ \overline{Y}_{21} & \overline{Y}_{23} & \overline{Y}_{22} \end{bmatrix},$$

with Y_{11} real. Note that

$$\begin{aligned} P\overline{\widetilde{Z}_{\pm}}P &= \widetilde{Z}_{\pm} \ \text{ and } \ P\overline{M}_{\pm}P = M_{\pm}, \\ P\overline{\widetilde{Z}_{\pm} \circ M_{\pm}}P &= \widetilde{Z}_{\pm} \circ M_{\pm}. \end{aligned}$$

Hence, $\overline{X} = X$ and X is real. ∎

Theorem 1.6 *Let $A \in \mathbb{R}^{n\times n}$ be diagonalizable, $A = VDV^{-1}$ with $D = \mathrm{diag}(\lambda_k)$. Then,*

(i) $\widetilde{\mathcal{T}}_+(A)$: $\mathbf{Sym}(\mathbb{K}) \longrightarrow \mathbf{Sym}(\mathbb{K})$ *defined by* $\widetilde{\mathcal{T}}_+(A)X = AX + XA^T$ *is invertible if for all* k_1, k_2, *such that* $1 \leq k_1, k_2 \leq n$ *and* $k_1 \neq k_2$, *we have* $\lambda_{k_1} + \lambda_{k_2} \neq 0$.

(ii) $\widetilde{\mathcal{T}}_-(A)$: $\mathcal{G} \longrightarrow \mathbf{Skew}(\mathbb{K})$ *defined by* $\widetilde{\mathcal{T}}_-(A)X = AX - XA^T$ *is invertible if the eigenvalues of* A *are distinct.*

Then, when $\widetilde{\mathcal{T}}_\pm(A)^{-1}$ *exists,*

$$\widetilde{\mathcal{T}}_\pm(A)^{-1}Z_\pm = V((V^{-1}Z_\pm V^{-T}) \circ M_\pm)V^T,$$

where $Z_+ \in \mathbf{Skew}(\mathbb{K})$, $Z_- \in \mathbf{Sym}(\mathbb{K})$ *and* $M_\pm$ *is defined by* (1.20).

Proof. The proof is similar to Theorem 1.5. □

Applying the vec operator to $\mathcal{T}_\pm(A)^{-1}$ and $\widetilde{\mathcal{T}}_\pm(A)^{-1}$ in Theorems 1.5-1.6, we obtain

$$\|\mathcal{T}_\pm(A)^{-1}\|_2 = \|\widetilde{\mathcal{T}}_\pm(A)^{-1}\|_2 = \|(V \otimes V)\mathrm{diag}(\mathrm{vec}(M_\pm))(V^{-1} \otimes V^{-1})\|_2.$$

Note that if A is symmetric then V is orthogonal, thus

$$\begin{aligned} \|\widetilde{\mathcal{T}}_-(A)^{-1}\|_2 &= \|\mathcal{T}_-(A)^{-1}\|_2 = \|\mathrm{vec}(M_-)\|_\infty, \\ \|\widetilde{\mathcal{T}}_+(A)^{-1}\|_2 &= \|\mathcal{T}_+(A)^{-1}\|_2 = \frac{1}{2}\|A^{-1}\|_2. \end{aligned}$$

Furthermore, the adjoint operator of $\widetilde{\mathcal{T}}_-(A)$ is given by $\widetilde{\mathcal{T}}_-^T(A) = \mathcal{T}_-^T(A^T)$. $\mathcal{F}$ is the orthogonal complement of $\{X \in \mathbf{Sym}(\mathbb{R}),\ A^TX = XA\}$. This is a generalization of the orthogonal direct sum decomposition of $\mathbf{Sym}(\mathbb{R})$ given in [6].

1.7 Condition Number and Backward Error

We briefly describe the concept of condition numbers and backward errors. For that, we consider $f : \mathbb{R} \to \mathbb{R}$ to be twice differentiable and for $x \in \mathbb{R}$, we let $y = f(x)$.

The condition number is a measure of the sensitivity of the output y when the input x is subject to perturbation. Let Δx be a perturbation of x and $\tilde{y} = f(x + \Delta x)$. By Taylor's theorem, we have

$$\tilde{y} - y = f'(x)\Delta x + O(|\Delta x|^2).$$

Thus

$$|\tilde{y} - y| \leq |f'(x)||\Delta x| + O(|\Delta x|^2), \tag{1.21}$$

and if $f(x) \neq 0$, we have

$$\frac{|\tilde{y} - y|}{|y|} \leq \frac{|xf'(x)|}{|f(x)|}\frac{|\Delta x|}{|x|} + O(|\Delta x|^2). \tag{1.22}$$

The condition number or the idea of conditioning is to quantify the first order variation, that is, the coefficient of $|\Delta x|$ in (1.21) or the coefficient of $\frac{|\Delta x|}{|x|}$ in (1.22). The value $|f'(x)|$ is known as the absolute condition number and $\frac{|xf'(x)|}{|f(x)|}$ as the relative condition number. We expect $\tilde{y}$ to be close to y if $|f'(x)|$ is small, which we qualify as well-conditioned. If $|f'(x)|$ is relatively big then for a given Δx, small enough, we can have a large $|\tilde{y} - y|$, which is by analogy the ill-conditioned case.

Let $\hat{y}$ be an approximation of y. The aim of backward error analysis is to find Δx such that $\hat{y} = f(x + \Delta x)$. Δx might not be unique. Thus, we try to find Δx that has the smallest absolute value. We can define the backward error by,

$$\eta(\hat{y}) = \min\left\{\epsilon : \ \hat{y} = f(x + \Delta x), \ |\Delta x| \leq \epsilon\right\}.$$

The condition number and backward error can be used to bound the forward error. They are related by the following inequality [37], at the first order in the backward error

$$\text{forward error} \leq \text{condition number} \times \text{backward error}.$$

We conclude this paragraph by a necessary definition that will be used throughout this thesis.

Definition 1.6 *We define the condition number of a matrix with respect to inversion, generally called the condition number of the matrix by*

$$\kappa_\alpha(X) = \|X\|_\alpha \|X^{-1}\|_\alpha, \tag{1.23}$$

where $\|\cdot\|_\alpha$ *is a norm over* $\mathbb{K}^{n\times n}$.

1.8 The Polynomial Eigenvalue Problem

Let $P(A,\lambda)$ be an $n \times n$ matrix polynomial of degree m as in (1.1).

Definition 1.7 *We say that* $P(A,\lambda)$ *is regular if*

$$\det(P(A,\lambda)) \not\equiv 0.$$

We will assume throughout this thesis that $P(A,\lambda)$ is regular.

PEP is to find scalars λ and nonzero vectors x and y satisfying

$$P(A,\lambda)x = 0, \quad y^*P(A,\lambda) = 0.$$

λ is called an eigenvalue and, x and y are the corresponding right and left eigenvectors. Equivalently, the eigenvalues are the roots of the characteristic polynomial

$$\det(P(A,\lambda)) = 0.$$

Let d be the degree of the scalar polynomial $\det(P(A,\lambda))$. The d roots of $\det(P(A,\lambda))$ are called the finite eigenvalues. If $d < mn$, then we say that $P(A,\lambda)$ has $mn-d$ infinite eigenvalues. Let A_m be singular. We consider the PEP associated to $P(A,\mu)$, where $\mu = \frac{1}{\lambda}$. We see that the value $\mu = 0$ is an eigenvalue of the PEP and the eigenvectors are the vectors that generate null(A). The 0 eigenvalues of the PEP associated to $\lambda^m P(A, \frac{1}{\lambda})$ correspond to the infinite eigenvalues of the PEP associated to $P(A,\lambda)$.

We now give several definitions in order to characterize the eigenvalues.

Definition 1.8 *Let λ be an eigenvalue of $P(A,\lambda)$.*

The algebraic multiplicity of λ denoted by p is the multiplicity of λ as a root the characteristic polynomial $\det(P(A,\lambda))$.

The geometric multiplicity of λ denoted by q is the number of linearly independent eigenvectors spanning null$P(A,\lambda)$.

Let p be the algebraic multiplicity of λ and let q be the geometric multiplicity of λ. The eigenvalue λ is simple if $p = 1$. When $p > 1$, λ is a multiple eigenvalue. In the case where $p > 1$ and $q = p$, λ is semi-simple. Otherwise, λ is a defective eigenvalue.

Remark 1.7 *For the standard eigenvalue problem $Ax = \lambda x$, A is diagonalizable if its eigenvalues are simple or semi-simple. Otherwise A admits a non trivial Jordan form.*

While the polynomial eigenvalue problem is usually written as $P(A,\lambda)x = 0$, this representation has a disadvantage because it gives special emphasis to zero or infinite eigenvalues, which leads to difficulties in characterizing and computing condition numbers and backward errors. For example in [71] the condition number is not defined for zero eigenvalues. For the infinite eigenvalues, the condition number and the backward error can be obtained by computing the limit of the

appropriated quantity as $|\lambda|$ tends to infinity. With the homogeneous form of the PEP, finite and infinite eigenvalues are treated on the same footing. This facilitates the characterization and computation of condition numbers and backward errors. This alternative is discussed in the next section.

1.9 Homogeneous PEPs

Let $A = (A_0, A_1, \ldots, A_m) \in \mathcal{M}_n(\mathbb{C})^{m+1}$. We define the homogeneous matrix polynomial $P(A, \alpha, \beta)$ by

$$P(A, \alpha, \beta) = \sum_{k=0}^{m} \alpha^k \beta^{m-k} A_k, \tag{1.24}$$

that is, $P(A, \alpha, \beta)$ is homogeneous of degree m in $(\alpha, \beta) \in \mathbb{C}^2$.

We assume that $P(A, \alpha, \beta)$ is regular, that is $\det P(A, \alpha, \beta) \not\equiv 0$, for all $(\alpha, \beta) \neq (0, 0)$.

The homogeneous polynomial eigenvalue problem (PEP) is to find pairs of scalars $(\alpha, \beta) \neq (0, 0)$ and nonzero vectors $x, y \in \mathbb{C}^n$ satisfying

$$P(A, \alpha, \beta)x = 0, \quad y^* P(A, \alpha, \beta) = 0. \tag{1.25}$$

The vectors x, y are called right and left eigenvectors corresponding to the eigenvalue (α, β). Hence, an eigenvalue is now any line through the origin in $\mathbb{C}^2$ of solutions of $\det(P(A, \alpha, \beta)) = 0$.

For $\beta \neq 0$, we define $\lambda = \dfrac{\alpha}{\beta}$. Thus, we can link the non-homogeneous matrix polynomial $P(A, \lambda)$ to the homogeneous one by

$$P(A, \lambda) = \beta^m P(A, \alpha, \beta).$$

We see that solving the homogeneous PEP is equivalent to solving the non-homogeneous PEP.

For example, we have the well-known form

$$(\beta A - \alpha B)x = 0, \tag{1.26}$$

for the homogeneous generalized eigenvalue problem [68].

Chapter 2

Condition Numbers for Eigenvalues and Eigenvectors

Let $A = (A_0, A_1, \ldots, A_m) \in \mathcal{M}_n(\mathbb{K})^{m+1}$. We consider the homogeneous matrix polynomial $P(A, \alpha, \beta)$ defined by

$$P(A, \alpha, \beta) = \sum_{k=0}^{m} \alpha^k \beta^{m-k} A_k.$$

2.1 Introduction

The condition number of an eigenvalue reveals the sensitivity of the eigenvalue to perturbations in the data. Assume that $P(A, \alpha, \beta)$ is regular and that (α, β) is a simple eigenvalue of P. Let

$$\begin{aligned} \widetilde{A} &= (\widetilde{A}_0, \widetilde{A}_1, \ldots, \widetilde{A}_m), \\ &= (A_0 + \Delta A_0, A_1 + \Delta A_1, \ldots, A_m + \Delta A_m) \in \mathcal{M}_n(\mathbb{C})^{m+1} \end{aligned}$$

be a perturbation of A and let $(\widetilde{\alpha}, \widetilde{\beta})$ be the corresponding perturbation of (α, β). Two approaches can be taken to define condition numbers.

1. **Generalization of Stewart and Sun's approach for the GEP**

   We can use Rice's definition of condition numbers [61]. The condition number $c(\alpha, \beta)$ of a simple eigenvalue (α, β) can be defined by

   $$c(\alpha, \beta) = \lim_{\epsilon \to 0} \sup \left\{ \frac{d_c((\alpha, \beta), (\widetilde{\alpha}, \widetilde{\beta}))}{\epsilon}, \ \|E\| \leq \epsilon \right\},$$

   where d_c is the chordal distance given in Definition 2.3. It follows that in first order approximation the inequality

   $$d_c((\alpha, \beta), (\widetilde{\alpha}, \widetilde{\beta})) \leq c(\alpha, \beta)\|E\|$$

   holds. Stewart and Sun [68] and Sun [69] derive an expression for $c(\alpha, \beta)$ by bounding $d_c((\alpha, \beta), (\widetilde{\alpha}, \widetilde{\beta}))$.

2. **Dedieu and Tisseur's approach**

   Dedieu [25] and Dedieu and Tisseur [26] first define a map from a matrix $m + 1$-tuple to an eigenvalue. Then they define the condition operator of the differential map. The norm of this differential is the condition number.

Dedieu [25] showed for the GEP that approaches 1 and 2 are equivalent. The second approach has several advantages.

The differential calculus approach is described in more detail in the next section.

2.2 A Differential Calculus Approach

2.2.1 Preliminaries

In [25] and [26], the authors apply the implicit function theorem to the equation

$$f(A, x, \alpha, \beta) = 0,$$

where $f(A, x, \alpha, \beta) = P(A, \alpha, \beta)x$. The aim is to find a function g that maps $\tilde{A}$, in a neighborhood of A to the corresponding eigenpair $(\tilde{x}, \tilde{\alpha}, \tilde{\beta})$, in a neighborhood of (x, α, β). The implicit function theorem can not be applied as it is since the dimension of $f(A, x, \alpha, \beta) \in \mathbb{C}^n$ and $(x, \alpha, \beta) \in \mathbb{C}^{n+2}$ do not match. A way to overcome this problem is to introduce projective spaces. The use of these spaces for this problem arises naturally as we show below.

Let $\rho \in \mathbb{C} \setminus \{0\}$. We consider the PEP defined by (1.25). We see that $\rho(\alpha, \beta)$ is also an eigenvalue which is another representation of (α, β). Thus, it becomes natural to work with projective spaces.

2.2.2 Projective Spaces

Definition 2.1 *Let $\mathcal{R}$ be the equivalence relationship on $\mathbb{C}^k \setminus \{0\}$ defined by*

$$\forall (x, y) \in \mathbb{C}^k \times \mathbb{C}^k \quad x\mathcal{R}y \iff (\exists \rho \in \mathbb{C} \setminus \{0\} \quad y = \rho x)\,.$$

The quotient space $\mathbb{C}^k/\mathcal{R}$ is called the projective space and it is denoted by $\mathbb{P}(\mathbb{C}^k)$.

It follows immediately from Definition 2.1 that

$$\dim(\mathbb{P}(\mathbb{C}^k)) = k - 1.$$

Thus, we denote $\mathbb{P}(\mathbb{C}^k) = \mathbb{P}_{k-1}$. Note that $\mathbb{P}(\mathbb{C}^k)$ can also be identified to the quotient space of the unite sphere of $\mathbb{C}^k$

$$S = \{x \in \mathbb{C}^k : \|x\|_2^2 = 1\}$$

for the equivalence relationship $\widetilde{\mathcal{R}}$ defined by

$$\forall (x, y) \in \mathbb{C}^k \times \mathbb{C}^k \quad x\widetilde{\mathcal{R}}y \iff \left(y = e^{i\theta}x, \theta \in \mathbb{R}\right).$$

The quotient space associated to the eigenvalue (α, β) is the projective space $\mathbb{P}(\mathbb{C}^2) = \mathbb{P}_1$ of dimension 1. For the eigenvectors, it is more familiar to consider

projective spaces, since the eigenvectors span a linear subspace of $\mathbb{C}^n$. Also, projective spaces avoid the problem of choosing a normalization for the eigenvalues and eigenvectors (see [3, Sec. (4)] where several normalizations are discussed). Thus, we take

$$(x, \alpha, \beta) \in \mathbb{P}_{n-1} \times \mathbb{P}_1, \tag{2.1}$$

which has now dimension n. Let $T_x\mathbb{P}_{k-1}$ be the tangent space to $\mathbb{P}_{k-1}$. This tangent space is identified with

$$x^\perp = \{y \in \mathbb{C}^k : \langle y, x\rangle = 0\}.$$

The scalar product over $T_x\mathbb{P}_{k-1}$ and the induced norm are then given by

$$\begin{aligned} \langle y_1, y_2\rangle_{x^\perp} &= \frac{\langle y_1, y_2\rangle}{\langle x, x\rangle}, \\ \|y\|_{x^\perp} &= \sqrt{\langle y, y\rangle_{x^\perp}} \end{aligned} \tag{2.2}$$

and it is independent from the chosen representatives.

2.2.3 Condition Numbers

In this section, we compute the condition operators and the corresponding condition numbers of a simple eigenvalue and the associated eigenvector. We define a function g on $\mathcal{V}_A$ a neighborhood of $A \in \mathcal{M}_n(\mathbb{C})^{m+1}$ that maps to $\widetilde{A} \in \mathcal{V}_A$ to the corresponding eigenpair $(\widetilde{x}, \alpha, \beta)$, such that $P(\widetilde{A}, \widetilde{\alpha}, \widetilde{\beta})\widetilde{x} = 0$. The function g is not explicitly accessible but its differential can be computed. To characterize g we proceed as follow. As in [26], the main tool for this analysis is the implicit function theorem. Let

$$\begin{aligned} f: \ \mathcal{M}_n(\mathbb{C})^{m+1} \times \mathbb{P}_{n-1} \times \mathbb{P}_1 &\to \mathbb{C}^n, \\ (A, x, \alpha, \beta) &\mapsto P(A, \alpha, \beta)x, \end{aligned} \tag{2.3}$$

and let

$$V_P = \left\{(A, x, \alpha, \beta) \in \mathcal{M}_n(\mathbb{C})^{m+1} \times \mathbb{P}_{n-1} \times \mathbb{P}_1 : \quad P(A, \alpha, \beta)x = 0\right\}$$

be the set of polynomial eigenvalue problems. We define the following projections

$$\begin{aligned} &\Pi_1 \ : \ V_P \to \mathcal{M}_n(\mathbb{C})^{m+1}, && \Pi_2 \ : \ V_P \to \mathbb{P}_{n-1} \times \mathbb{P}_1, \\ &\Pi_1(A, x, \alpha, \beta) = A, && \Pi_2(A, x, \alpha, \beta) = (x, \alpha, \beta). \end{aligned}$$

Definition 2.2 *[26]* (A, x, α, β) *is defined as* ***a well-posed problem*** *when* Π_1 *is invertible. Otherwise, we refer to* (A, x, α, β) *as an ill-posed problem.*

We now focus on well-posed problems for which we can compute the condition operator and the condition number. In what follows, we see that a well-posed problem is equivalent to the eigenvalue (α, β) being simple. We define the vector v that is used throughout this section by

$$v = \left(\bar{\beta}\frac{\partial P(A, \alpha, \beta)}{\partial \alpha} - \bar{\alpha}\frac{\partial P(A, \alpha, \beta)}{\partial \beta}\right) x. \tag{2.4}$$

In the following theorem [26], we summarize the necessary results for the rest of this section.

Theorem 2.1 *Let* (α, β) *be a simple eigenvalue, let* x, y *be associated right and left eigenvectors and* v *be defined by* (2.4) *. Then,*

1. $v \notin \text{range}(P(A, \alpha, \beta))$,
2. $\Pi_{v^\perp} P(A, \alpha, \beta)_{|x^\perp}$ *is nonsingular,*
3. $y^* v \neq 0$,

where $\Pi_{v^\perp}$ *denotes the projection onto the orthogonal space to* v*.*

We write $d_2 f = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial \alpha} + \frac{\partial f}{\partial \beta}$ where f is defined by (2.3). The following theorem states a property of $d_2 f(A, x, \alpha, \beta)$ for a simple eigenvalue (α, β).

Theorem 2.2 *Let (α, β) be a simple eigenvalue of $P(A, \alpha, \beta)$ and let x and y be the corresponding right and left eigenvectors. Then,*

$$d_2 f(A, x, \alpha, \beta) : T_x \mathbb{P}_{n-1} \times T_{(\alpha,\beta)} \mathbb{P}_1 \to \mathbb{C}^n$$

is nonsingular.

Proof. Let $\Delta x \in T_x \mathbb{P}_{n-1}$ and $(\Delta\alpha, \Delta\beta) \in T_{(\alpha,\beta)} \mathbb{P}_1$. Then $\langle \Delta x, x \rangle = 0$ and $(\Delta\alpha, \Delta\beta) = \rho(\bar{\beta}, -\bar{\alpha})$, where $\rho \in \mathbb{C}$. Assume that

$$d_2 f(A, x, \alpha, \beta)(\Delta x, \Delta\alpha, \Delta\beta) = 0.$$

We have

$$\begin{aligned} d_2 f(Z_1) Z_2 &= P(A, \alpha, \beta)\Delta x + \Delta\alpha \frac{\partial P}{\partial \alpha}(A, \alpha, \beta)x + \Delta\beta \frac{\partial P}{\partial \beta}(A, \alpha, \beta)x, \\ &= P(A, \alpha, \beta)\Delta x + \rho v \ = \ 0, \end{aligned}$$

where $Z_1 = (A, x, \alpha, \beta)$ and $Z_2 = (\Delta x, \Delta\alpha, \Delta\beta)$, so that premultipication by y^* gives $\rho y^* v = 0$. By Theorem 2.1, $y^* v \neq 0$. Thus, $\rho = 0$ and therefore $(\Delta\alpha, \Delta\beta) = 0$. On the other hand $\Delta x \in x^\perp$ and $\Pi_{v^\perp} P(A, \alpha, \beta)_{|x^\perp}$ is nonsingular. Thus, $\Delta x = 0$. □

Note that when $d_2 f(A, x, \alpha, \beta)$ is nonsingular then the problem is well-posed, that is, $\Pi_1(A)$ is invertible. In this case, from the implicit function theorem, we know that there exists $\mathcal{V}_A$ a neighborhood of A, $\mathcal{V}_x \times \mathcal{V}_{(\alpha,\beta)}$ a neighborhood of $(x, (\alpha, \beta))$ and a differentiable map

$$g \ : \ \mathcal{V}_A \to \mathcal{V}_x \times \mathcal{V}_{(\alpha,\beta)}$$

such that for all $\tilde{A} \in \mathcal{V}_A$, we have

$$g(\tilde{A}) = (\tilde{x}, \widetilde{\alpha}, \widetilde{\beta}) \ , \ P(\tilde{A}, \widetilde{\alpha}, \widetilde{\beta})\tilde{x} = 0.$$

The differential at A, $dg(A)$ is then given by

$$\begin{aligned} dg(A) &= -(d_2 f(A,x,\alpha,\beta))^{-1} \frac{\partial f}{\partial A}(A,x,\alpha,\beta), \\ dg(A)\Delta A &= -(d_2 f(A,x,\alpha,\beta))^{-1} P(\Delta A,\alpha,\beta)x. \end{aligned} \tag{2.5}$$

We set $g = (g_1, g_2)$ such that for all $\tilde{A} \in \mathcal{V}_A$,

$$g_1(\tilde{A}) = \tilde{x},\ g_2(\tilde{A}) = (\widetilde{\alpha}, \widetilde{\beta}).$$

We can now define the condition operator for the eigenvector and for the eigenvalue to be

$$dg_1 \ \text{ and } \ dg_2.$$

We set $\Delta x = dg_1(A)\Delta A$ and $(\Delta\alpha, \Delta\beta) = dg_2(A)\Delta A$. We have that $\Delta x \in T_x\mathbb{P}_{n-1}$ and $(\Delta\alpha, \Delta\beta) \in T_{(\alpha,\beta)}\mathbb{P}_1$. Thus, $\Delta x \in x^\perp$ and $(\Delta\alpha, \Delta\beta) = \rho(\bar{\beta}, -\bar{\alpha})$. From (2.5), we get

$$P(A,\alpha,\beta)\Delta x + \rho v = -P(\Delta A,\alpha,\beta)x. \tag{2.6}$$

The condition numbers for the eigenvector x and the eigenvalue (α,β) are defined by

$$\begin{aligned} c_1(A,\alpha,\beta,x) &= \|dg_1(A)\|_{x^\perp} = \max_{\|\Delta A\|\leq 1} \frac{\|dg_1(A)\Delta A\|_2}{\|x\|_2}, \\ c_2(A,\alpha,\beta,x) &= \|dg_2(A)\|_{(\alpha,\beta)^\perp} = \max_{\|\Delta A\|\leq 1} \frac{\|dg_2(A)\Delta A\|_2}{\|(\alpha,\beta)\|_2}, \end{aligned}$$

where the norm on ΔA is arbitrary and the Hermitian structure of the projective spaces is defined by (2.2).

We now focus on computing the condition operator and condition number of the eigenvector and eigenvalue. We will take for the norm on $\mathcal{M}_n(\mathbb{K})^{m+1}$ the μ-weighted Frobenius norm of Definition 1.4. For a weight $\mu \in \mathbb{R}^{m+1}$, $\|.\|_{F,\mu}$ measures the perturbations of the coefficients of the matrix polynomial $P(A,\alpha,\beta)$

relative to the weights in $\mu = (\frac{1}{\mu_k})$. The absolute condition number is obtained by setting all the components of μ equal to 1 whereas the relative condition number is given by setting $\mu_k = \|A_k\|_F$ if $\|A_k\|_F \neq 0$, for $0 \leq k \leq m+1$. The case $\|A_k\|_F = 0$ is discussed at the end of this paragraph.

The following theorems give the condition operator and condition number of the eigenvector and eigenvalue. The method to obtain the condition operators is similar to [26] although here we apply the implicit function theorem directly. The condition numbers and the corresponding optimal perturbations (where the condition number is attained by $dg_k(A),\ k = 1, 2$) are computed with some differences from [26] since we use a weighted Frobenius norm. We define the scalar γ that is used in the following two theorems by

$$\gamma = \gamma(\alpha, \beta, \mu) = \left(\sum_{k=0}^{m} |\alpha|^{2k} |\beta|^{2(m-k)} \mu_k^2 \right)^{\frac{1}{2}}. \tag{2.7}$$

Theorem 2.3 *Let (α, β) be a simple eigenvalue of $P(A, \alpha, \beta)$, x and y be the corresponding left and right eigenvectors and v be as in* (2.4)*. The eigenvector condition operator is given by*

$$dg_1(A)\Delta A = -(\Pi_{v^\perp} P(A, \alpha, \beta)_{x^\perp})^{-1} \Pi_{v^\perp} P(\Delta A, \alpha, \beta)x. \tag{2.8}$$

The condition number of an eigenvector of $P(A, x, \alpha, \beta)$, with perturbations measured in the weighted Frobenius norm is given by

$$c_1(A, \alpha, \beta, x) = \left(\sum_{k=0}^{m} |\alpha|^{2k} |\beta|^{2(m-k)} \mu_k^2 \right)^{\frac{1}{2}} \|(\Pi_{v^\perp} P(A, \alpha, \beta)_{x^\perp})^{-1}\|_2. \tag{2.9}$$

Proof. Applying the orthogonal projection onto $v^\perp$ to (2.6), we obtain

$$\Pi_{v^\perp} P(A, \alpha, \beta)\widetilde{x} = -\Pi_{v^\perp} P(\Delta A, \alpha, \beta)x,$$

where $\widetilde{x} = dg_1(A)\Delta A$. We recall that $\widetilde{x} \in x^\perp$ and by Theorem 2.1, we know that $\Pi_{v^\perp} P(A, \alpha, \beta)_{|x^\perp}$ is nonsingular. Thus,

$$dg_1(A)\Delta A = -(\Pi_{v^\perp} P(A, \alpha, \beta)_{x^\perp})^{-1} \Pi_{v^\perp} P(\Delta A, \alpha, b)x.$$

Tacking norms and applying Cauchy-Schwarz inequality, we get

$$\begin{aligned}\frac{\|\widetilde{x}\|_2}{\|x\|_2} &\leq \|(\Pi_{v^\perp} P(A,\alpha,\beta)_{x^\perp})^{-1}\|_2 \sum_{k=0}^{m} |\alpha|^k |\beta|^{m-k} \|\Delta A_k\|_F \\ &\leq \gamma \|(\Pi_{v^\perp} P(A,\alpha,\beta)_{x^\perp})^{-1}\|_2 \|\Delta A\|_{F,\mu}, \end{aligned} \tag{2.10}$$

where γ is defined by (2.7). Let $u \in \mathbb{C}^n$ with $\|u\|_2 = 1$ be such that

$$\|(\Pi_{v^\perp} P(A,\alpha,\beta)_{x^\perp})^{-1} u\|_2 = \|(\Pi_{v^\perp} P(A,\alpha,\beta)_{x^\perp})^{-1}\|_2.$$

Note that $u \in v^\perp$. The inequality (2.10) is attained by

$$\Delta A_k = \frac{1}{\|x\|_2 \gamma} \bar{\alpha}^k \bar{\beta}^{m-k} u x^*, \ k = 0\colon m. \quad \square$$

Theorem 2.4 *Let (α,β) be a simple eigenvalue of $P(A,\alpha,\beta)$, x and y be the corresponding left and right eigenvectors and v be as in* (2.4)*. Then, the eigenvalue condition operator is given by*

$$dg_2(A)\Delta A = \frac{y^* P(\Delta A,\alpha,\beta) x}{y^* v}(-\bar{\beta},\bar{\alpha}). \tag{2.11}$$

The condition number of a simple eigenvalue (α,β) of $P(A,\alpha,\beta)$, with perturbations measured in the weighted Frobenius norm is given by

$$c_2(A,\alpha,\beta,x) = \frac{\|x\|_2 \|y\|_2}{|y^* v|} \left(\sum_{k=0}^{m} |\alpha|^{2k} |\beta|^{2(m-k)} \mu_k^2 \right)^{\frac{1}{2}}. \tag{2.12}$$

Proof. Let $(\widetilde{\alpha},\widetilde{\beta}) = dg_2(A)\Delta A$. Since $(\widetilde{\alpha},\widetilde{\beta}) \in T_{(\alpha,\beta)}\mathbb{P}_1$, then there exists a $\rho \in \mathbb{C}$ such that

$$(\widetilde{\alpha},\widetilde{\beta}) = \rho(\bar{\beta},-\bar{\alpha}),$$

and since (α,β) is an eigenvalue, we obtain from (2.6)

$$\rho y^* v = -y^* P(\Delta A,\alpha,\beta) x.$$

Thus, from Theorem 2.1, $y^*v \neq 0$ for a simple eigenvalue and

$$dg_2(A).\Delta A \;=\; \frac{y^*P(\Delta A,\alpha,\beta)x}{y^*v}(-\overline{\beta},\overline{\alpha}). \tag{2.13}$$

We have

$$c_2(A,\alpha,\beta,x) \;=\; \sup_{\|\Delta A\|_{F,\mu}\leq 1} \|(\widetilde{\alpha},\widetilde{\beta})\|_{(\alpha,\beta)^\perp} = \sup_{\|\Delta A\|_{F,\mu}\leq 1} \sqrt{\frac{|\widetilde{\alpha}|^2+|\widetilde{\beta}|^2}{|\alpha|^2+|\beta|^2}}. \tag{2.14}$$

Then we obtain

$$\begin{aligned}\frac{|\widetilde{\alpha}|^2+|\widetilde{\beta}|^2}{|\alpha|^2+|\beta|^2} &= \frac{|y^*P(\Delta A,\alpha,\beta)x|^2}{|y^*v|^2},\\ y^*P(\Delta A,\alpha,\beta)x &= \sum_{k=0}^{m}\alpha^k\beta^{m-k}y^*\Delta A_k x.\end{aligned}$$

Using the triangle inequality, we obtain

$$|y^*P(\Delta A,\alpha,\beta)x| \leq \sum_{k=0}^{m}\mu_k|\alpha|^k|\beta|^{m-k}|y^*(\frac{1}{\mu_k}\Delta A_k)x|,$$

and by applying the Cauchy-Schwarz inequality gives

$$|y^*P(\Delta A,\alpha,\beta)x| \leq \|x\|_2\|y\|_2 \left(\sum_{k=0}^{m}\mu_k|\alpha|^k|\beta|^{(m-k)}\|\frac{1}{\mu_k}\Delta A_k\|_F\right),$$

where $\|\cdot\|_F$ is the usual Frobenius norm. Applying once more the Cauchy-Schwarz inequality, we have

$$|y^*P(\Delta A,\alpha,\beta)x| \leq \gamma\|x\|_2\|y\|_2\|\Delta A\|_{F,\mu},$$

where γ is defined by (2.7). Hence

$$c_2(A,\alpha,\beta,x) \leq \frac{\|x\|_2\|y\|_2}{|y^*v|}\gamma.$$

By using the matrices

$$S = \left[\frac{y}{\|y\|_2} \;\;,\;\; 0\right] \;\text{ and }\; \Delta A_k = \frac{\overline{\alpha}^k\overline{\beta}^{m-k}\mu_k^2}{\gamma}S$$

with $\|\Delta A\|_{F,\mu} = 1$, we show that $c_2(A, \alpha, \beta, x)$ reaches $\dfrac{\|x\|_2\|y\|_2}{|y^*v|}\gamma$. Thus

$$c_2(A, \alpha, \beta, x) = \frac{\|x\|_2\|y\|_2}{|y^*v|}\left(\sum_{k=0}^{m} |\alpha|^{2k}|\beta^{2(m-k)}|\mu_k^2\right)^{\frac{1}{2}}. \quad \square$$

Note that the condition number $c_1(A, \alpha, \beta, x)$ and $c_2(A, \alpha, \beta, x)$ are well defined, since the right-hand side in (2.9) and (2.12) is independent of the choice of representatives of the eigenvector x and the eigenvalue (α, β).

The condition number given in Theorem 2.4 measures the absolute sensitivity of a simple eigenvalue if we choose $\mu_k = 1,\ k = 0\colon m$ or the relative sensitivity if $\mu_k = \frac{1}{\|A_k\|_F}$ with $A_k \neq 0,\ k = 0\colon m$. This means that all the coefficient matrices are subject to a perturbation. An interesting approach, physically meaningful, is to allow some of the matrices to not be perturbed. In order to compute the condition number, we see that $dg_2(A)$ is constant along the direction that are not perturbed. Thus, it is equivalent to allow the components of μ that correspond to the unperturbed directions to be zero in Theorem 2.4. We therefore define the weights by

$$\tilde{\mu}_k = \begin{cases} 1/\mu_k, & \text{if } \mu_k \neq 0, \\ 0, & \text{if } \mu_k = 0. \end{cases}$$

2.3 Perturbation Analysis

In this section, we investigate the first order variation of the perturbed eigenvalue $(\widetilde{\alpha}, \widetilde{\beta})$ by extending results in [25] for the GEP to the PEP. The following theorem enables us to work in a Hilbert space instead of a projective space.

Theorem 2.5 *Let (α, β) be a simple eigenvalue of $P(A, x, \alpha, \beta)$, normalized so that $\|(\alpha, \beta)\|_2 = 1$ and x be a right eigenvector of unit norm. Let ΔA be a*

perturbation of A and $(\widetilde{x}, (\widetilde{\alpha}, \widetilde{\beta})) \in \mathbb{P}(\mathbb{C}^2)$ be the perturbed eigenvalue. We choose the following representatives of x and $(\widetilde{\alpha}, \widetilde{\beta})$:

$$\begin{aligned} (\widetilde{\alpha}, \widetilde{\beta}) &= (\alpha, \beta) + (\alpha^\perp, \beta^\perp), \\ \widetilde{x} &= x + x^\perp, \end{aligned}$$

where

$$\begin{aligned} \left\langle (\alpha, \beta), (\alpha^\perp, \beta^\perp) \right\rangle &= 0, \\ \left\langle x, x^\perp \right\rangle &= 0. \end{aligned}$$

Then, we have

$$\begin{aligned} \widetilde{x} &= x + dg_1(A)\Delta A + O(\epsilon^2), \\ (\widetilde{\alpha}, \widetilde{\beta}) &= (\alpha, \beta) + dg_2(A)\Delta A + O(\epsilon^2), \end{aligned}$$

where $\epsilon = \|\Delta A\|_{F,\mu}$.

Proof. Let (e_1, e_2) be an orthonormal basis of $\mathbb{C}^2$ with $e_1 = (\alpha, \beta)$. Recall that $(\alpha, \beta) \neq (0, 0)$. We introduce the local chart:

$$V = \left\{ \xi_1 e_1 + \xi_2 e_2,\ (\xi_1, \xi_2) \in \mathbb{C}^2,\ \xi_1 \neq 0 \right\} \subset \mathbb{P}(\mathbb{C}^2),$$

$$\phi(\xi_1 e_1 + \xi_2 e_2) = \frac{\xi_2}{\xi_1} e_2 \in \mathbb{C}^2.$$

We have

$$\begin{aligned} g_2(A) &= (\alpha, \beta) \in \mathbb{P}(\mathbb{C}^2), \\ g_2(A + \Delta A) &= (\widetilde{\alpha}, \widetilde{\beta}) \in \mathbb{P}(\mathbb{C}^2), \\ \phi \circ g_2(A + \Delta A) &= (\alpha^\perp, \beta^\perp). \end{aligned}$$

On the other hand, we have

$$\phi \circ g_2(A + \Delta A) = \phi(g_2(A)) + d(\phi \circ g_2(A))\Delta A + O(\epsilon^2)$$

and

$$d(\phi \circ g_2(A)) = d\phi(\alpha, \beta) \circ dg_2(A)).$$

We obtain finally

$$d(\phi \circ g_2(A)) = dg_2(A)$$

since $dg_2(A)$ takes its values in $(\alpha, \beta)^{\perp}$ and $d\phi(\alpha, \beta)$ is the identity for unitary (α, β). Thus,

$$(\widetilde{\alpha}, \widetilde{\beta}) = (\alpha, \beta) + dg_2(A)\Delta A + O(\epsilon^2).$$

The perturbation expansion for the eigenvector is found in a similar way by considering the following local charts

$$\begin{aligned} \widetilde{V} &= \left\{ \sum_{k=1}^{n} \xi_k e_k,\ (\xi) \in \mathbb{C}^n,\ \xi_1 \neq 0 \right\} \subset \mathbb{P}(\mathbb{C}^n), \\ \widetilde{\phi}(\sum_{k=1}^{n} \xi_k e_k) &= (\frac{\xi_2}{\xi_1}, \ldots, \frac{\xi_n}{\xi_1}) \in \mathbb{C}^{n-1}, \end{aligned}$$

where $e_1 = x$ and the vectors $e_k,\ k = 1\colon n$ form an orthonormal basis of $\mathbb{C}^n$. □

Definition 2.3 *We consider the projective space $\mathbb{P}_{n-1}(\mathbb{C})$ with the usual scalar-product $\langle ., . \rangle$ over $\mathbb{C}^n$. The angle between $(u, v) \in \mathbb{C}^n \times \mathbb{C}^n$ is the Riemannian distance and it is defined by*

$$d_r(u, v) = \arccos\left(\frac{|\langle u, v \rangle|}{\|u\|_2 \|v\|_2} \right).$$

We define the chordal distance between $(u, v) \in \mathbb{C}^n \times \mathbb{C}^n$ by

$$\begin{aligned} d_c(u, v) &= \sin(d_r(u, v)), \\ &= \left(1 - \frac{|\langle u, v \rangle|^2}{\|u\|_2^2 \|v\|_2^2} \right)^{\frac{1}{2}}. \end{aligned}$$

For $n = 2$, $u = (\alpha, \beta)$ and $v = (\widetilde{\alpha}, \widetilde{\beta})$, the chordal distance becomes

$$d_c(u, v) = \frac{|\alpha\widetilde{\beta} - \widetilde{\alpha}\beta|}{\|(\alpha, \beta)\|_2 \|(\widetilde{\alpha}, \widetilde{\beta})\|_2}.$$

Corollary 2.6 *Let (α, β) be a simple eigenvalue of $P(A, \alpha, \beta)$ and x be the corresponding right eigenvector. For ΔA small enough, the perturbed polynomial $P(A + \Delta A, \alpha, \beta)$ has a simple eigenvalue $(\widetilde{\alpha}, \widetilde{\beta})$ with associated eigenvector $\tilde{x}$. Then, we have*

$$\begin{aligned} d_c(\widetilde{x}, x) &\leq c_1(A, \alpha, \beta, x)\epsilon + O(\epsilon^2), \\ d_c((\widetilde{\alpha}, \widetilde{\beta}), (\alpha, \beta)) &\leq c_2(A, \alpha, \beta, x)\epsilon + O(\epsilon^2), \end{aligned}$$

where $\epsilon = \|\Delta A\|$.

Proof. For any vectors $u, v \in \mathbb{C}^k$, we have the following identity

$$d_c(u, v) = \left(1 - \frac{|\langle u, v\rangle|^2}{\|u\|_2^2\|v\|_2^2}\right)^{\frac{1}{2}} = \left\| \frac{u}{\|u\|_2} - \frac{\langle u, v\rangle}{\|u\|_2\|v\|_2^2} v \right\|_2.$$

Thus, by applying Theorem 2.5 to

$$\frac{x}{\|x\|_2} \text{ and } \frac{\langle x, \tilde{x}\rangle}{\|\tilde{x}\|_2^2\|x\|_2}\tilde{x},$$

we obtain

$$\tilde{x} = x + dg_1(A)\Delta A + O(\epsilon^2).$$

Thus,

$$d_c(\tilde{x}, x) \leq c_1(A, \alpha, \beta, x)\epsilon + O(\epsilon^2).$$

The second inequality for the eigenvalue is obtained similarly. □

We know that for $0 \leq \theta < \frac{\pi}{2}$, we have

$$\sin(\theta) \leq \theta \leq \tan(\theta).$$

Applying this fact to the distances in Definition 2.3, we have

$$d_c \leq d_r \leq d_t,$$

where $d_t(u,v) = \tan(d_r(u,v))$. Thus, if we apply Theorem 2.5 to

$$\frac{(\alpha,\beta)}{\|(\alpha,\beta)\|_2} \quad \text{and} \quad \frac{\|(\alpha,\beta)\|_2(\widetilde{\alpha},\widetilde{\beta})}{\left\langle(\widetilde{\alpha},\widetilde{\beta}),(\alpha,\beta)\right\rangle}$$

and to

$$\frac{x}{\|x\|_2} \quad \text{and} \quad \frac{\|x\|_2\widetilde{x}}{\langle\widetilde{x},x\rangle},$$

we get the following inequalities

$$\begin{aligned} d_t(\tilde{x},x) &\leq c_1(A,\alpha,\beta,x)\epsilon + O(\epsilon^2), \\ d_t((\widetilde{\alpha},\widetilde{\beta}),(\alpha,\beta)) &\leq c_2(A,\alpha,\beta,x)\epsilon + O(\epsilon^2). \end{aligned}$$

Note that d_t is not a distance since it does not satisfy the triangular inequality [25].

2.4 Link to the Non-Homogeneous Form

Generally, matrix polynomials are considered in the non-homogeneous form. For $\beta \neq 0$, $\lambda = \frac{\alpha}{\beta}$, we have

$$P(A,\lambda) = P(A,\lambda,1). \tag{2.15}$$

Corollary 2.7 *For $\lambda = \frac{\alpha}{\beta}$ and $\tilde{\lambda} = \frac{\widetilde{\alpha}}{\widetilde{\beta}}$, we define the chordal distance by*

$$\chi(\tilde{\lambda},\lambda) = \frac{|\tilde{\lambda}-\lambda|}{\sqrt{1+|\tilde{\lambda}|^2}\sqrt{1+|\lambda|^2}}.$$

Then, we have $\chi(\tilde{\lambda},\lambda) \leq c_2(A,\alpha,\beta,x)\epsilon + O(\epsilon^2)$.

Proof. The result is obtained from Corollary 2.6:

$$d_c((\widetilde{\alpha},\widetilde{\beta}),(\alpha,\beta)) = \frac{|\widetilde{\alpha}\beta - \alpha\widetilde{\beta}|}{\|(\widetilde{\alpha},\widetilde{\beta})\|_2\|(\alpha,\beta)\|_2} = \frac{|\tilde{\lambda}-\lambda|}{\sqrt{1+|\tilde{\lambda}|^2}\sqrt{1+|\lambda|^2}}.$$

Thus, we have

$$\chi(\tilde{\lambda},\lambda) \leq c_2(A,\alpha,\beta,x)\epsilon + O(\epsilon^2). \square$$

2.5 Particular Case: the GEP

The generalized eigenvalue problem corresponds to the case of a matrix polynomial of degree $m = 1$ in (1.26). We consider the pair $(A, -B)$, where $A, B \in \mathbb{C}^{n\times n}$. We now focus on computing the condition number for the eigenvalue for the matrix polynomial

$$P((A, -B), \alpha, \beta) = \beta A - \alpha B.$$

From Theorem 2.4, the absolute eigenvalue condition number is given by

$$c_2((A, -B), \alpha, \beta, x) = \frac{(|\alpha|^2 + |\beta|^2)^{\frac{1}{2}}\|x\|_2\|y\|_2}{|y^*(\bar{\beta}B + \bar{\alpha}A)x|}, \tag{2.16}$$

where x and y are the right and left eigenvectors. We have

$$\begin{aligned}(\beta A - \alpha B)x &= 0,\\ \beta y^*Ax &= \alpha y^*Bx.\end{aligned}$$

Thus,

$$(\alpha, \beta) = \rho(y^*Ax, y^*Bx) \in \mathbb{P}_1,\ \rho \in \mathbb{C}.$$

By tacking the representative (y^*Ax, y^*Bx) for the eigenvalue (α, β), we obtain

$$c_2((A, -B), \alpha, \beta, x) = \frac{\|x\|_2\|y\|_2}{\sqrt{|\alpha|^2 + |\beta|^2}},$$

which is the condition number given by Stewart in [67, p. 140]. Now, for the standard eigenvalue problem, $B = I$ and $\beta \neq 0$ always. Letting $\lambda = \frac{\alpha}{\beta}$, (2.16) becomes

$$c_2((A, -I), \alpha, \beta, x) = \frac{\|x\|_2\|y\|_2}{|y^*x|}\frac{1}{\sqrt{1 + |\lambda|^2}}.$$

We recall that

$$\kappa(\lambda) = \frac{\|x\|_2\|y\|_2}{|y^*x|}$$

is the standard condition number of a simple eigenvalue for $Ax = \lambda x$ [78]. If A is normal, we can take $x = y$ so that

$$\kappa(\lambda) = 1,$$

and

$$c_2((A, -I), \alpha, \beta, x) \;=\; \frac{1}{\sqrt{1 + |\lambda|^2}} \;\leq\; 1.$$

Note that c_2 and κ have different interpretation: c_2 bounds the angle between the exact and perturbed eigenvalue whereas κ bounds the distance between the exact and perturbed eigenvalue.

2.6 Hermitian Structured Condition Numbers

We consider a Hermitian PEP $P(A, \alpha, \beta)$, which means that all the coefficient matrices are Hermitian. Let x be an eigenvector associated with (α, β). Then, $y = x$ is an eigenvector associated with $(\bar{\alpha}, \bar{\beta})$.

The Hermitian structured condition number for a simple eigenvalue (α, β) is defined by

$$c_{2,\mathbf{Herm}}(A, \alpha, \beta, x) \;=\; \max_{\Delta A \in \mathbf{Herm}^{m+1}, \|\Delta A\| \leq 1} \frac{\|dg_2(A)\Delta A\|_2}{\|(\alpha, \beta)\|_2}.$$

Clearly

$$c_{2,\mathbf{Herm}}(A, \alpha, \beta, x) \leq c_2(A, \alpha, \beta, x).$$

Let (α, β) be real. We see that in the proof of Theorem 2.4, the equality above is attained by the Hermitian perturbations

$$\Delta A_k = \frac{\alpha^k \beta^{m-k} \mu_k^2}{\gamma} \frac{x x^*}{\|x\|_2^2}, \qquad k = 0\colon m.$$

Thus, as for the standard eigenvalue problem, for real eigenvalues, we have

$$c_{2,\mathbf{Herm}}(A, \alpha, \beta, x) = c_2(A, \alpha, \beta, x).$$

2.7 Conclusion

This chapter focused on eigenvalue and eigenvector condition numbers of matrix polynomials. We generalized the work of Stewart and Sun, and Dedieu on the GEP to arbitrary degree matrix polynomial. The use of a weighted Frobenius norm allows flexibility on how the perturbations are measured. It enabled us first to define relative condition numbers. Then, by modifying the definition of weights, we showed that it also covers the case where of some the coefficient matrices are not perturbed (by setting to 0 the corresponding weights). In [69], this condition number is called the partial condition numbers since it corresponds to the norm of a partial differential.

Moreover, the results in this chapter and the results on backward errors in the next chapter contributed to the development of MATLAB's function `polyeig`. A pseudocode that computes the condition number (2.12) is given in Section 5.3.

Chapter 3

Backward Errors

3.1 Introduction

In backward error analysis, we consider that an approximate eigenpair $(\widehat{\alpha}, \widehat{\beta}, \widehat{x})$ of $P(A, \alpha, \beta)$ is the exact solution of a perturbed PEP $P(A + \Delta A, \alpha, \beta)$. Note that the perturbation ΔA may not be unique. We aim to characterize ΔA by focusing on perturbations that minimize the 2-norm or the Frobenius norm. If the backward error is in some sense small then the approximate solution is an exact solution of a nearby problem. The normwise backward error analysis is the study of perturbations that minimize a given norm. If we restrict the perturbations that minimize the norm to some subset of structured matrices, then the analysis is called structured normwise backward error analysis. The structures that we encounter in this chapter are symmetric and Hermitian.

In the first part of this chapter, we extend the results on backward error for nonhomogeneous PEPs [71] to homogeneous PEPs. The homogeneous form of P allows to treat on the same footing both finite and infinite eigenvalues. Then, we move on to structured backward errors.

3.2 Normwise Backward Error

The results on the backward error hold for the 2-norm or the Frobenius norm. Let $A = (\Delta A_k)_{0\le k\le m} \in \mathcal{M}_n(\mathbb{C})^{m+1}$ and let $(\widehat{\alpha}, \widehat{\beta})$ be an approximate eigenvalue of $P(A, \alpha, \beta)$ and let $\widehat{x}$, $\widehat{y}$ be the corresponding right and left eigenvectors. The vector $\mu = (\mu_k) \in \mathbb{R}^{m+1}$ contains the nonnegative weights μ_k that allows flexibility on how the perturbations are measured. We define

$$\begin{aligned}
\mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) &= \{\epsilon\colon P(A+\Delta A, \widehat{\alpha}, \widehat{\beta})\widehat{x} = 0, \|\Delta A_k\|_\delta \le \epsilon\mu_k, k = 0:m\},\\
\mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y}^*) &= \{\epsilon\colon \widehat{y}^*P(A+\Delta A, \widehat{\alpha}, \widehat{\beta}) = 0, \|\Delta A_k\|_\delta \le \epsilon\mu_k, k = 0:m\},\\
\mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*) &= \mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) \cap \mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y}^*). \qquad (3.1)
\end{aligned}$$

Definition 3.1 *The δ-norm backward error of $(\widehat{\alpha}, \widehat{\beta}, \widehat{x})$ is defined by*

$$\eta_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) = \min \mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}).$$

By analogy, the δ-norm backward error for the triplet $((\widehat{\alpha}, \widehat{\beta}), \widehat{x}, \widehat{y})$ is defined by

$$\eta_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*) = \min \mathcal{E}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*).$$

Definition 3.2 *For $z \in \mathbb{C}$, we define its sign by*

$$\operatorname{sign}(z) = \begin{cases} \frac{\bar{z}}{|z|} & if \quad z \neq 0,\\ 1 & if \quad z = 0. \end{cases}$$

Theorem 3.1 *An explicit expression for the 2-norm or the Frobenius norm backward error for the approximate eigenpair $((\widehat{\alpha}, \widehat{\beta}), \widehat{x})$ is given by*

$$\eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) = \eta_{F,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) = \frac{\|P(A, \widehat{\alpha}, \widehat{\beta})\widehat{x}\|_2}{\|\widehat{x}\|_2 \sum_{k=0}^m |\widehat{\alpha}|^k |\widehat{\beta}|^{m-k}\mu_k}. \qquad (3.2)$$

Proof. One can easily show that the right-hand side of (3.2) is a lower bound for η. This bound is attained by the following perturbations,

$$\Delta A_k = -\frac{1}{\gamma}\operatorname{sign}(\widehat{\alpha}^k\widehat{\beta}^{m-k})\mu_k\frac{P(A, \widehat{\alpha}, \widehat{\beta})\widehat{x}\widehat{x}^*}{\|\widehat{x}\|_2^2},$$

where $\gamma = \sum_{k=0}^{m} |\widehat{\alpha}|^k |\widehat{\beta}|^{m-k} \mu_k$. $\square$

Theorem 3.2 *The 2-norm backward error for the triplet $((\widehat{\alpha}, \widehat{\beta}), \widehat{x}, \widehat{y})$ is given by*

$$\eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*) = \max(\eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}), \eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y}^*)).$$

The Frobenius norm backward error for the triplet $((\widehat{\alpha}, \widehat{\beta}), \widehat{x}, \widehat{y})$ is given by

$$\eta_{F,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*) = \left(\frac{\|r\|_2^2}{\|\widehat{x}\|_2^2} + \frac{\|s\|_2^2}{\|\widehat{y}\|_2^2} - \frac{|s^* \widehat{x}|^2}{\|\widehat{x}\|_2^2 \|\widehat{y}\|_2^2} \right)^{\frac{1}{2}},$$

where $r = P(A, \widehat{\alpha}, \widehat{\beta})\widehat{x}$ and $s^* = y^* P(A, \widehat{\alpha}, \widehat{\beta})$.

Proof. Let $\epsilon \in \mathcal{E}_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*)$. As in Theorem 3.1, it can be shown that

$$\eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) \leq \epsilon, \quad \eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y}^*) \leq \epsilon.$$

Thus, we obtain that

$$\max(\eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}), \eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y}^*)) \leq \epsilon.$$

In order to show that this bound is attained, we use a result from [41]:

$$\min\{\|H\|_2 : \ H\widehat{x} = r, \ \widehat{y}^* H = s^*\} \quad = \quad \max\left\{ \frac{\|r\|_2}{\|\widehat{x}\|_2}, \frac{\|s\|_2}{\|\widehat{y}\|_2} \right\}. \tag{3.3}$$

Let H be the optimal matrix in (3.3) and for $0 \leq k \leq m$, let

$$\Delta A_k = -\frac{1}{\gamma} \text{sign}(\widehat{\alpha}^k \widehat{\beta}^{m-k}) \mu_k H,$$

where $\gamma = \sum_{k=0}^{m} |\widehat{\alpha}|^k |\widehat{\beta}|^{m-k} \mu_k$. We have that

$$P(A + \Delta A, \widehat{\alpha}, \widehat{\beta})\widehat{x} = 0, \quad \widehat{y}^* P(A + \Delta A, \widehat{\alpha}, \widehat{\beta}) = 0$$

and

$$\|A_k\|_2 = \mu_k \max(\eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}), \eta_{2,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y}^*)).$$

The backward error with the perturbations measured with the Frobenius norm is obtained by solving the optimization problem

$$\min\{\|H\|_F : \ H\widehat{x} = r, \ \widehat{y}^* H = s^*\}. \tag{3.4}$$

Let $\dfrac{\epsilon}{\sqrt{2}} \in \mathcal{E}_{F,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}, \widehat{y}^*)$. Then,

$$\eta_{F,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x})^2 + \eta_{F,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{y})^2 \leq \epsilon^2.$$

Then, the H that achieves (3.4) is given in [41] by

$$H = \frac{r\widehat{x}^*}{\|\widehat{x}\|_2^2} + \frac{\widehat{y}s^*}{\|\widehat{y}\|_2^2} - \frac{s^*\widehat{x}}{\|\widehat{x}\|_2^2\|\widehat{y}\|_2^2}\widehat{y}\widehat{x}^*.$$

with

$$\|H\|_F^2 = \frac{\|r\|_2^2}{\|\widehat{x}\|_2^2} + \frac{\|s\|_2^2}{\|\widehat{y}\|_2^2} - \frac{|s^*\widehat{x}|^2}{\|\widehat{x}\|_2^2\|\widehat{y}\|_2^2}. \qquad \square$$

We measured the perturbations individually in the definition of the backward error at the beginning of Section 3.2. In the previous chapter, when we computed the condition number, the perturbations are measured globally by the norm given in Definition 1.4. In order to be consistent, we need to compute the backward error using the same norm as for the condition number so that we can use the first order bound of the forward error. We define

$$\widetilde{\mathcal{E}}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) \ = \ \{\epsilon\colon P(A+\Delta A, \widehat{\alpha}, \widehat{\beta})\widehat{x} = 0, \|\Delta A\|_{\delta,\mu} \leq \epsilon\}.$$

Definition 3.3 *The δ-norm backward error of $(\widehat{\alpha}, \widehat{\beta}, \widehat{x})$ is then defined by*

$$\widetilde{\eta}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) = \min \widetilde{\mathcal{E}}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}).$$

We measure the perturbations with either the weighted 2-norm $\|\Delta A\|_{2,\mu}$ or the weighted Frobenius norm $\|\Delta A\|_{F,\mu}$.

Theorem 3.3 *The normwise backward error of an approximate eigenpair* $(\widehat{\alpha}, \widehat{\beta}, \widehat{x})$ *for the weighted* 2*-norm* $\|\cdot\|_{2,\mu}$ *or weighted Frobenius norm* $\|\cdot\|_{F,\mu}$ *is given by*

$$\widetilde{\eta}_{\delta,\mu}(\widehat{\alpha}, \widehat{\beta}, \widehat{x}) = \frac{\|r\|_2}{\gamma \|\widehat{x}\|_2}, \tag{3.5}$$

where $\delta = 2, F$, $r = P(A, \widehat{\alpha}, \widehat{\beta})x$ *and*

$$\gamma = \left(\sum_{k=0}^{m} |\widehat{\alpha}|^{2k} |\widehat{\beta}|^{2(m-k)} \mu_k^2 \right)^{\frac{1}{2}}.$$

Proof. We have

$$\begin{aligned} \frac{\|r\|_2}{\|\widehat{x}\|_2} &\leq \sum_{k=0}^{m} |\widehat{\alpha}|^k |\widehat{\beta}|^{m-k} \|A_k\|_\delta, \\ \frac{\|r\|_2}{\|\widehat{x}\|_2} &\leq \sum_{k=0}^{m} |\widehat{\alpha}|^k |\widehat{\beta}|^{m-k} \mu_k \| \frac{1}{\mu_k} A_k \|_\delta, \end{aligned}$$

where $\|\cdot\|$ in the inequalities above is 2-norm or the Frobenius norm. Using Cauchy-Schwarz inequality, these inequalities become

$$\frac{\|r\|_2}{\|\widehat{x}\|_2} \leq \gamma \|\Delta A\|_{\delta,\mu},$$

This bound is attained by the following perturbation

$$\Delta A_k = \frac{\|r\|_2}{\gamma^2 \|\widehat{x}\|_2^2} \mu_k^2 \bar{\widehat{\alpha}}^k \bar{\widehat{\beta}}^{m-k} r \widehat{x}^*. \quad \square$$

3.3 Normwise Structured Backward Error for the Symmetric PEP

In this section we consider structured backward errors for symmetric PEPs for which the coefficient matrices are symmetric or Hermitian. Our analysis is motivated by the development of structure preserving algorithms. It enables us to

check if an approximate eigenpair of a symmetric PEP is the exact eigenpair of a nearby symmetric PEP.

Let

$$\mathcal{E}_{S,\delta,\mu}(\widehat{\alpha},\widehat{\beta},\widehat{x}) = \left\{\epsilon\colon P(A+\Delta A,\widehat{\alpha},\widehat{\beta})\widehat{x}=0, \Delta A \in \mathbf{Sym}(\mathbb{R})^{m+1}, \|\Delta A\|_{\delta,\mu} \leq \epsilon\right\}.$$

Definition 3.4 *The normwise structured backward error for an approximate eigenpair $(\widehat{\alpha},\widehat{\beta},\widehat{x})$ is defined by*

$$\eta_{S,\delta,\mu}(\widehat{\alpha},\widehat{\beta},\widehat{x}) \;=\; \min \mathcal{E}_{S,\delta,\mu}(\widehat{\alpha},\widehat{\beta},\widehat{x}). \tag{3.6}$$

We measure the perturbations with either the weighted 2-norm $\|\Delta A\|_{2,\mu}$ or the weighted Frobenius norm $\|\Delta A\|_{F,\widetilde{\mu}}$ with $\widetilde{\mu} = \frac{\mu}{\sqrt{n}}$ (Definition 1.4).

Theorem 3.4 *The structured normwise backward error of a real eigenpair for the weighted 2-norm $\|\cdot\|_{2,\mu}$ or weighted Frobenius norm $\|\cdot\|_{F,\tilde{\mu}}$ with $\tilde{\mu} = \mu/\sqrt{n}$ is given by*

$$\eta_{S,\delta,\mu}(\widehat{\alpha},\widehat{\beta},\widehat{x}) = \frac{\|r\|_2}{\gamma\|\widehat{x}\|_2},$$

where

$$r = P(A,\widehat{\alpha},\widehat{\beta})\widehat{x} \;\text{ and }\; \gamma = \left(\sum_{k=0}^{m} |\widehat{\alpha}|^{2k}|\widehat{\beta}|^{2(m-k)}\mu_k^2\right)^{\frac{1}{2}}.$$

Proof. We have

$$\begin{aligned}
\frac{\|r\|_2}{\|\widehat{x}\|_2} &\leq \sum_{k=0}^{m} |\widehat{\alpha}|^k|\widehat{\beta}|^{m-k}\|A_k\|,\\
\frac{\|r\|_2}{\|\widehat{x}\|_2} &\leq \sum_{k=0}^{m} |\widehat{\alpha}|^k|\widehat{\beta}|^{m-k}\mu_k\|\frac{1}{\mu_k}A_k\|,
\end{aligned}$$

where $\|\cdot\|$ in the inequalities above is 2-norm or the Frobenius norm. Using Cauchy-Schwarz inequality, these inequalities become

$$\begin{aligned}
\frac{\|r\|_2}{\|\widehat{x}\|_2} &\leq \gamma\|\Delta A\|_{2,\mu},\\
\frac{\|r\|_2}{\|\widehat{x}\|_2} &\leq \gamma\|\Delta A\|_{F,\mu}.
\end{aligned}$$

Let S be a symmetric matrix such that $S\widehat{x} = r$. We can take $S = \frac{\|r\|_2}{\|\widehat{x}\|_2}H$, where H is Householder matrix if r and $\widehat{x}$ are linearly independent, otherwise $H = I$. Then, the optimal perturbations are given by

$$\Delta A_k = \frac{\|r\|_2}{\gamma^2\|\widehat{x}\|_2}\mu_k^2\widehat{\alpha}^k\widehat{\beta}^{m-k}H. \qquad \square$$

We see that the unstructured backward error (3.5) and the symmetric structured backward error in Theorem 3.4 are equal. Hence, imposing symmetric structures does not change the backward error.

In Chapter 6, we analyze the HZ algorithm that computes the eigenvalues of a real symmetric pair (A, B). Complication occur when (A, B) are real and the eigenpair is complex. In this case computing the symmetric structured backward error is an optimization problem that we solve in the next section.

3.4 Normwise Structured Backward Error for the Symmetric GEP

We consider the GEP in the nonhomogeneous form

$$Ax = \lambda Bx, \tag{3.7}$$

where A and B are $n \times n$ real symmetric matrices. We assume B nonsingular which justifies the use of the nonhomogeneous form.

3.4.1 Real Eigenpair

Suppose that the approximate eigenpair $(\widehat{\lambda}, \widehat{x})$ is real. Then, Theorem 3.4 with $\widehat{\lambda} = \widehat{\alpha}/\widehat{\beta}$ and $m = 1$ gives

$$\eta_{S,\delta}(\widehat{\lambda}, \widehat{x}) \;=\; \frac{1}{\sqrt{\mu_A^2 + |\widehat{\lambda}|^2 \mu_B^2}} \frac{\|(A - \widehat{\lambda}B)\widehat{x}\|_2}{\|\widehat{x}\|_2}, \tag{3.8}$$

This explicit expression for the backward error differs slightly from the one derived by D.J. Higham and N.J. Higham [32], where a different measure of the perturbations is used. If we restrict the perturbation to be real for a complex eigenvalue with a non trivial imaginary part, we face an optimization problem that is treated in detail in the next section.

3.4.2 Complex Eigenvalues

To compute $\eta_{S,\delta}$, we can use the Kronecker product approach described in [32] but the disadvantage of this technique is its computational cost. Our aim is to compute the structured backward error in $O(n)$ operations if the residual vector $r = (A - \widehat{\lambda}B)\widehat{x}$ is given or in $O(n^2)$ otherwise.

Let $(\widehat{\lambda}, \widehat{x})$ be an approximate complex eigenpair of the GEP (3.7). We know then that $(\bar{\widehat{\lambda}}, \bar{\widehat{x}})$ is also an approximate eigenpair of the GEP . Thus, we have the following system

$$(A + \Delta A)\widehat{x} \;=\; \widehat{\lambda}(B + \Delta B)\widehat{x}, \tag{3.9}$$

$$(A + \Delta A)\bar{\widehat{x}} \;=\; \bar{\widehat{\lambda}}(B + \Delta B)\bar{\widehat{x}}. \tag{3.10}$$

We write $\widehat{\lambda} = \tau + i\nu,\ \tau, \nu \in \mathbb{R}$ and $\widehat{x} = w + iz,\ w, z \in \mathbb{R}^n$. By adding first (3.9) to (3.10) and then by subtracting (3.9) to (3.10), we get the following equivalent system

$$(\Delta A - \tau \Delta B)w + \nu \Delta B z + r_1 \;=\; 0, \tag{3.11}$$

$$(\Delta A - \tau \Delta B)z - \nu \Delta B y + r_2 \;=\; 0, \tag{3.12}$$

where $r_1 = (A - \tau B)w + \nu Bz$ and $r_2 = (A - \tau B)z - \nu Bw$. We define

$$M(\Delta A, \Delta B) = \begin{bmatrix} \Delta A - \tau \Delta B & \nu \Delta B \\ -\nu \Delta B & \Delta A - \tau \Delta B \end{bmatrix}$$

and we rewrite (3.11)-(3.12) as

$$M(\Delta A, \Delta B)a + r = 0, \tag{3.13}$$

where

$$a = \begin{bmatrix} w \\ z \end{bmatrix} \quad \text{and} \quad r = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}.$$

We recall that $\nu \neq 0$ and that at least one of the components of z is non-zero. We define the map

$$\begin{aligned} g : \mathbf{Sym}(\mathbb{R}) \times \mathbf{Sym}(\mathbb{R}) &\rightarrow \mathbb{R}^{2n}, \\ (\Delta A, \Delta B) &\mapsto M(\Delta A, \Delta B)a + r. \end{aligned}$$

The perturbations are measured with the weighted Frobenius norm

$$\|(\Delta A, \Delta B)\|_{F,\mu} = \left(\frac{\|\Delta A\|_F^2}{n\mu_A} + \frac{\|\Delta B\|_F^2}{n\mu_B} \right)^{\frac{1}{2}}.$$

We rewrite the problem of computing the structured backward error as a constrained optimization problem. We define the objective function by

$$N(\Delta A, \Delta B) = \|(\Delta A, \Delta B)\|_{F,\tilde{\mu}}^2$$

and the feasible set

$$\Omega = \{(\Delta A, \Delta B),\ g(\Delta A, \Delta B) = 0\}.$$

Thus, the problem becomes to minimize the objective function N on Ω,

$$\min_{(\Delta A, \Delta B) \in \mathbf{Sym}(\mathbb{R})^2} N(\Delta A, \Delta B) \text{ subject to } g(\Delta A, \Delta B) = 0.$$

The Lagrange multipliers theorem 1.2 is the main tool to solve this optimization problem. Thus, we need to show that Ω is a differentiable manifold and then compute its dimension. The following lemma will help us to compute the dimension of Ω.

Lemma 3.5 *Let $\widehat{\lambda} = \tau + i\nu$ and $\widehat{x} = w + iz$ be an eigenpair of $(\widetilde{A}, \widetilde{B}) = (A + \Delta A, B + \Delta B)$. If the pencil $\widetilde{A} - \widehat{\lambda}\widetilde{B}$ is regular and $\nu \neq 0$ then w and z are linearly independent.*

Proof. Let $\nu \neq 0$ and let the pencil $(\widetilde{A} - \widehat{\lambda}\widetilde{B})$ be regular. Assume first that $w = 0$. Then, $(\widetilde{A} - \widehat{\lambda}\widetilde{B})\widehat{x} = 0$ implies

$$\widetilde{B}z = 0 \text{ and } \widetilde{A}z = 0.$$

Thus $z \in \text{null}\widetilde{A} \cap \text{null}\widetilde{B}$ and $(\widetilde{A} - \widehat{\lambda}\widetilde{B})$ is nonregular which contradicts the assumption. Since $\nu \neq 0$ then $w \neq 0$. Similarly, we show that $z = 0$ implies $(\widetilde{A} - \widehat{\lambda}\widetilde{B})$ is nonregular. Thus, $w \neq 0$ and $z \neq 0$.

Assume that $z = \xi w$, for some $\xi \in \mathbb{R} \setminus \{0\}$. Then,

$$\widetilde{A}w = \tau\widetilde{B}w + i\nu\widetilde{B}w.$$

Thus, $\nu = 0$ which contradicts the hypothesis. Hence, w and z are linearly independent. □

Theorem 3.6 *Ω is a $(n^2 - n)$-dimensional differentiable manifold, that is, the components of the gradient of g are made up $2n$ linearly independent functionals.*

Proof. Since g is linear, it is differentiable and

$$dg(\Delta A, \Delta B) = g - r.$$

Thus, applying the vec operator, $dg(\Delta A, \Delta B)$ becomes

$$dg(\Delta A, \Delta B)(E, F) = \widetilde{M}(y, z) \otimes I_n \begin{bmatrix} \text{vec}(E) \\ \text{vec}(F) \end{bmatrix},$$

where

$$\widetilde{M}(y, z) = \begin{bmatrix} y^T & (-\tau w + \nu z)^T \\ z^T & -(\tau w + \nu z)^T \end{bmatrix}.$$

By Lemma 3.5, w and z are linearly independent. Thus,

$$\operatorname{rank}(\widetilde{M}(w,z)\otimes I_n) = 2n.$$

Hence, Ω is a (n^2-n)-dimensional differentiable manifold by Definition 1.5. □

We recall that N and g are differentiable and we denote respectively their differentials by dN and dg. By the Lagrange multipliers theorem we know that if N has a minimizer $(\Delta A_*, \Delta B_*)$ on Ω then there exist $2n$ constants, $(c_i)_{1\leq i\leq 2n}$, such that

$$dN(\Delta A_*, \Delta B_*) = \sum_{i=1}^{2n} c_i dg_i(\Delta A_*, \Delta B_*), \tag{3.14}$$

where N reaches its local extremum. We define

$$u = -\tau w + \nu z \quad \text{and} \quad v = \tau z + \nu w.$$

We identify the coefficients in (3.14). We have

$$\Delta a_{ii} = \frac{\mu_A}{2}(c_i w_i + c_{n+i} z_i), \tag{3.15}$$

$$\Delta b_{ii} = \frac{\mu_B}{2}(c_i u_i - c_{n+i} v_i), \tag{3.16}$$

$$\Delta a_{ij} = \frac{\mu_A}{4}(c_i w_j + c_j w_i + c_{n+i} z_j + c_{n+j} z_i), \tag{3.17}$$

$$\Delta b_{ij} = \frac{\mu_B}{4}(c_i u_j + c_j u_i - c_{n+i} v_j - c_{n+j} v_i). \tag{3.18}$$

Since $(\Delta A, \Delta B) \in \Omega$, we have $g(\Delta A, \Delta B) = 0$. Also (3.15-3.18) are equivalent to

$$\Delta A = \frac{\mu_A}{4}(c_1 w^T + w c_1^T + c_2 z^T + z c_2^T), \tag{3.19}$$

$$\Delta B = \frac{\mu_B}{4}(c_1 u^T + u c_1^T - c_2 v^T - v c_2^T), \tag{3.20}$$

where $c_1 = c(1\colon n)$ and $c_2 = c(n+1\colon 2n)$. Then, using (3.19-3.20) in (3.11-3.12) and factorizing the Lagrange multipliers out gives

$$Tc = (S_0 \otimes I_n + S_1)c = 4r, \tag{3.21}$$

where $c = [\, c_1^T \quad c_2^T \,]^T$,

$$S_0 = \begin{bmatrix} \mu_A \|w\|_2^2 + \mu_B \|u\|_2^2 & \mu_A \langle w, z\rangle - \mu_B \langle u, v\rangle \\ \mu_A \langle w, z\rangle - \mu_B \langle u, v\rangle & \mu_A \|z\|_2^2 + \mu_B \|v\|_2^2 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

and

$$S_1 = \begin{bmatrix} \mu_A w w^T + \mu_B u u^T & \mu_A z w^T - \mu_B v u^T \\ \mu_A w z^T - \mu_B u v^T & \mu_A z z^T + \mu_B v v^T \end{bmatrix} \in \mathbb{R}^{2n\times 2n}.$$

Theorem 3.7 *The problem of minimizing N on Ω has a unique solution.*

Proof. Let $X = (X_1, X_2), Y = (Y_1, Y_2) \in \mathbf{Sym}(\mathbb{R})^2$ and let t be such that $0 \leq t \leq 1$. We have

$$\begin{aligned} \|tX_1 + (1-t)Y_1\|_F^2 &\leq t^2\|X_1\|_F^2 + (1-t)^2\|Y_1\|_F^2 + 2t(1-t)\|X_1\|_F\|Y_1\|_F \\ &\leq t\|X_1\|_F^2 + (1-t)\|Y_1\|_F^2. \end{aligned}$$

Similarly,

$$\|tX_2 + (1-t)Y_2\|_F^2 \leq t\|X_2\|_F^2 + (1-t)\|Y_2\|_F^2.$$

Thus,

$$N(tX + (1-t)Y) \leq tN(X) + (1-t)N(Y).$$

N is convex. Assume that $X, Y \in \Omega$. Then, by definition

$$g(X) = 0 \text{ and } g(Y) = 0.$$

Thus,

$$\begin{aligned} tg(X) + (1-t)g(Y) &= 0, \\ M(tX + (1-t)Y) + ((1-t)+t)r &= 0, \\ g(tX + (1-t)Y) &= 0. \end{aligned}$$

Hence, Ω and N are convex and $\lim_{+\infty} N = +\infty$. Thus the solution to the optimization problem exists [20].

Assume now that the optimization problem has several solutions. For each of these solutions, Equations (3.14)-(3.21) are valid. In particular, T in (3.21) is singular and the solutions are of the form $c = c_0 + \tilde{c}$ with $\tilde{c} \in \text{null}(T)$. Thus, for all $\xi \in \mathbb{R}$, $c = c_0 + \xi\tilde{c}$ is a solution. Let $\Delta A(\xi)$ and $\Delta B(\xi)$ be the corresponding optimal perturbations. Since,

$$\lim_{\xi \to \infty} N(\Delta A(\xi), \Delta B(\xi)) = +\infty,$$

the minimization problem cannot have a solution. Thus, $\text{null}(T)) = 0$ and T is nonsingular and the solution to the minimization problem is unique. □

In order to compute the structured backward error, we just need to solve (3.21). Now, if we know the values of $\langle c_k, w\rangle$ and $\langle c_k, z\rangle$, for $k = 1, 2$, we can obtain the optimal perturbations. Thus, we just need to apply successively w^T and z^T to (3.21). We obtain a 4×4 linear system

$$\widetilde{T}\widetilde{a} = \widetilde{r}, \tag{3.22}$$

where

$$\begin{aligned} \widetilde{a} &= [\,\langle c_1, w\rangle \quad \langle c_1, z\rangle \quad \langle c_2, w\rangle \quad \langle c_2, z\rangle\,]^T, \\ \widetilde{r} &= [\,\langle r_1, w\rangle \quad \langle r_1, z\rangle \quad \langle r_2, w\rangle \quad \langle r_2, z\rangle\,]^T. \end{aligned} \tag{3.23}$$

Note that $\widetilde{T}$ is nonsingular since T is nonsingular. Let

$$\begin{bmatrix} \widetilde{c}_1 \\ \\ \widetilde{c}_2 \end{bmatrix} = S_1 \begin{bmatrix} c_1 \\ \\ c_2 \end{bmatrix}.$$

$$\begin{aligned} \widetilde{c}_1 &= \langle c_1, w\rangle\,((\mu_A + \tau^2\mu_B)w - \mu_B\tau\nu z) + \langle c_1, z\rangle\,\mu_B(\nu^2 z - \tau\nu w) \\ &\quad + \langle c_2, w\rangle\,((\mu_A + \tau^2\mu_B)z + \mu_B\tau\nu w) - \langle c_2, z\rangle\,\mu_B(\nu^2 w + \tau\nu z), \\ \widetilde{c}_2 &= \langle c_1, w\rangle\,\mu_B(\tau\nu w - \nu^2 z) + \langle c_1, z\rangle\,((\mu_A + \tau^2\mu_B)w - \mu_B\tau\nu z) \\ &\quad + \langle c_2, w\rangle\,\mu_B(\nu^2 w + \tau\nu z) + \langle c_2, z\rangle\,((\mu_A + \tau^2\mu_B)z - \mu_B\tau\nu w). \end{aligned}$$

By apply successively w^T and z^T, $\widetilde{T} = (\widetilde{t}_{ij})$ is given by

$$\begin{aligned}
\widetilde{t}_{11} &= (2\mu_A + \tau^2\mu_B)\|w\|_2^2 + \mu_B(\|u\|_2^2 - \tau\nu\langle w, z\rangle),\\
\widetilde{t}_{12} &= \mu_B(\nu^2\langle w, z\rangle - \tau\nu\|w\|_2^2),\\
\widetilde{t}_{13} &= (2\mu_A + \tau^2\mu_B)\langle w, z\rangle + \mu_B(\tau\nu\|w\|_2^2 - \langle u, v\rangle),\\
\widetilde{t}_{14} &= -\mu_B(\nu^2\|w\|_2^2 + \tau\nu\langle w, z\rangle),\\
\widetilde{t}_{21} &= (\mu_A + \tau^2\mu_B)\langle w, z\rangle - \mu_B\tau\nu\|z\|_2^2,\\
\widetilde{t}_{22} &= \mu_B(\nu^2\|z\|_2^2 + \|u\|_2^2 - \tau\nu\langle w, z\rangle) + \mu_A\|w\|_2^2,\\
\widetilde{t}_{23} &= (\mu_A + \tau^2\mu_B)\|z\|_2^2 - \mu_B\tau\nu\langle w, z\rangle,\\
\widetilde{t}_{24} &= \mu_A\langle w, z\rangle - \mu_B(\tau\nu\|z\|_2^2 + \langle u, v\rangle + \nu^2\langle w, z\rangle),\\
\widetilde{t}_{31} &= \mu_A\langle w, z\rangle + \mu_B(\tau\nu\|w\|_2^2 - \nu^2\langle w, z\rangle - \langle u, v\rangle),\\
\widetilde{t}_{32} &= (\mu_A + \tau^2\mu_B)\|w\|_2^2 - \mu_B\tau\nu\langle w, z\rangle,\\
\widetilde{t}_{33} &= \mu_A\|z\|_2^2 + \mu_B(\|v\|_2^2 + \tau\nu\langle w, z\rangle + \nu^2\|w\|_2^2),\\
\widetilde{t}_{34} &= (\mu_A + \tau^2\mu_B)\langle w, z\rangle - \mu_B\tau\nu\|w\|_2^2,\\
\widetilde{t}_{41} &= (\mu_A + \tau\nu\mu_B)\langle w, z\rangle - \mu_B(\nu^2\tau\nu\|z\|_2^2 + \langle u, v\rangle),\\
\widetilde{t}_{42} &= (2\mu_A + \tau^2\mu_B)\langle w, z\rangle - \mu_B(\tau\nu\|z\|_2^2 + \langle u, v\rangle),\\
\widetilde{t}_{43} &= \mu_B(\tau\nu\|z\|_2^2 + \tau\nu\langle w, z\rangle),\\
\widetilde{t}_{44} &= (2\mu_A + \tau^2\mu_B)\|z\|_2^2 + \mu_B(\|v\|_2^2 + \tau\nu\langle w, z\rangle).
\end{aligned}$$

Now that (3.22) is solved, the values of $\langle c_k, w\rangle$ and $\langle c_k, z\rangle$ are known for $k = 1, 2$. Assume that $a, b \in \mathbb{R}^n$ with $a = (a_k), b = (b_k)$. Let $U(a, w, b, z) = aw^T + wa^T + bz^T + zb^T$. Then,

$$\begin{aligned}
\|U(a, w, b, z)\|_F^2 &= \sum_{i,j=1}^{n}(a_iw_j + w_ia_j + b_iz_j + z_ib_j)^2,\\
&= 2(\|a\|_2^2\|w\|_2^2 + \|b\|_2^2\|z\|_2^2 + \langle a, w\rangle^2 + \langle b, z\rangle^2\\
&\quad +2(\langle a, b\rangle\langle w, z\rangle + \langle a, z\rangle\langle w, b\rangle)). \qquad (3.24)
\end{aligned}$$

Applying formula (3.24) to $(\Delta A, \Delta B)$, the norm of the optimal perturbation are easily computed in $O(n)$ operation. Hence the structured backward error can be computed in $O(n)$ flops and if the optimal perturbations are required, they can be computed in $O(n^2)$ flops.

Algorithm 3.8 *Given an approximate eigenpair $(\widehat{\lambda}, \widehat{x})$, the residual vector $(A - \widehat{\lambda}B)\widehat{x}$ and the weights μ_A, μ_B, this algorithm computes the symmetric structured backward error in $O(n)$ flops.*

Set $\tau = \Re(\widehat{\lambda})$, $\nu = \Im(\widehat{\lambda})$. Set $u = -\tau w + \nu z$ and $v = \tau z + \nu w$.

Compute $\langle w, z \rangle, \langle u, v \rangle, \|w\|_2, \|z\|_2, \|u\|_2, \|v\|_2$.

Compute $r_1 = \Re((A - \widehat{\lambda}B)\widehat{x}), r_2 = \Im((A - \widehat{\lambda}B)\widehat{x})$ and set $r = [\, r_1 \quad r_2 \,]^T$.

Compute $\widetilde{r}$ in (3.23) and form $\widetilde{T}$.

Solve $\widetilde{T}\widetilde{c} = \widetilde{r}$.

Using (3.24), compute

$$\eta_{S,\delta}(\widehat{\lambda}, \widehat{x}) = \|(\Delta A, \Delta B)\|_{F,\mu} = \sqrt{\frac{\|\Delta A\|_F^2}{n\mu_A^2} + \frac{\|\Delta B\|_F^2}{n\mu_B^2}} \text{ with } \delta = (F, \tilde{\mu}).$$

If in practice the structured and normwise backward error are of the same order then the algorithm in this section would be only of a minor theoretical interest. Thus, in order to justify our work in this section we present a numerical example where the structured and unstructured normwise backward error have a large ratio. Our example is generated by the function `mdsmax` from N.J. Higham's Matrix Computation Toolbox [33]. Note that the size of the problem is small. The GEP was solved by the QZ algorithm (see Chapter 5) implemented in MATLAB

as `eig`. For $n = 5$, we obtained the symmetric pair (A, B) given by

$$A = \begin{bmatrix} 147 & -25.5 & 201.5 & 76 & -40.5 \\ -25.5 & 74 & -109.5 & 96 & 46.5 \\ 201.5 & -109.5 & -227 & -40 & -30.5 \\ 76 & 96 & -40 & 36 & 1 \\ -40.5 & 46.5 & -30.5 & 1 & -4 \end{bmatrix},$$

$$B = \begin{bmatrix} -211 & 146.5 & -9.5 & -12 & -4.5 \\ 146.5 & 57 & -96 & 127 & 3.5 \\ -9.5 & -96 & -218 & -43.5 & -50.5 \\ -12 & 127 & -43.5 & 41 & -35.5 \\ -4.5 & 3.5 & -50.5 & -35.5 & 159 \end{bmatrix}.$$

The pair (A, B) has three real eigenvalues and a complex conjugate pair. For the complex eigenpairs $(\widehat{\lambda}, \widehat{x})$ and $(\bar{\widehat{\lambda}}, \bar{\widehat{x}})$, we found for the unstructured backward error $\eta_{F,\delta}(\widehat{\lambda}, \widehat{x}) = 2.10^{-16}$ and for the structured backward error $\eta_{S,\delta}(\widehat{\lambda}, \widehat{x}) = 10^{-12}$ which gives

$$\frac{\eta_{S,\delta}}{\eta_{F,\delta}} \geq 10^3, \qquad \delta = (F, \tilde{\mu}).$$

This result is not surprising since QZ destroys any symmetry in the matrix pair. The HZ Algorithm (see Chapter 6) preserves the symmetry of the problem. On this example, the unstructured normwise backward error for the eigenpairs computed with HZ is the same as the one for the eigenpairs computed with QZ. But, for the structured backward error, there is a slight improvement, $\eta_{S,\delta}(\widehat{\lambda}, \widehat{x}) = 10^{-13}$, which gives a ratio

$$\frac{\eta_{S,\delta}}{\eta_{F,\tilde{\mu}}} \geq 10^2.$$

The results of this chapter are used in Sections 5.3 and 6.9.

Chapter 4

Matrix Factorizations and their Sensitivity

4.1 Introduction

In this chapter, we show how to introduce zeros in a vector or a matrix using $(J, \widetilde{J})$-orthogonal matrices defined in Section 1.5. In Paragraph 4.2.1, we start by recalling results on the so called unified rotations, then we describe generalized Householder reflectors. In the last part of Paragraph 4.2.1, we present zeroing strategies combined with a careful monitoring of the condition number of the hyperbolic transformations used. We also present an error analysis of the computation of hyperbolic rotations. The rest of this chapter focuses on matrix factorizations in which $(J, \widetilde{J})$-orthogonal factors are involved. We describe each factorization, we give the optimal first order perturbation bound, the condition number of the factorization and we present numerical experiments.

4.2 Zeroing with (J_1, J_2)-Orthogonal Matrices

4.2.1 Unified Rotations

Unified rotations include orthogonal and hyperbolic rotations. We present a brief summary; for a more detailed presentation see [11], [74].

Let $x = [x_1, x_2]^T$ and $J = \text{diag}(\sigma_1, \sigma_2)$. Unified rotations have the form

$$\begin{bmatrix} c & \frac{\sigma_1}{\sigma_2}s \\ -s & c \end{bmatrix} \in \mathbb{R}^{2\times 2}, \tag{4.1}$$

with $\sigma_1 c^2 + \sigma_2 s^2 = \pm 1$. The aim is to find a matrix H such that

$$Hx = \begin{bmatrix} \rho \\ 0 \end{bmatrix} \quad \text{and} \quad H^T J H \in \text{diag}_k^n(\pm 1),$$

when $\sigma_1|x_1| \neq \sigma_2|x_2|$. Unified rotations can be classified into three types. The first type is the well known Givens rotation, when $J = \pm I$. In this case, we have

$$c = \frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \ s = \frac{x_2}{\sqrt{x_1^2 + x_2^2}} \quad \text{and} \quad c^2 + s^2 = 1.$$

If $J \neq \pm I$ and $|x_1| > |x_2|$, we say that H is a hyperbolic rotation of type 1 and we have

$$c = \frac{x_1}{\sqrt{x_1^2 - x_2^2}}, \ s = \frac{x_2}{\sqrt{x_1^2 - x_2^2}} \quad \text{and} \quad c^2 - s^2 = 1.$$

Finally, when $J \neq \pm I$ and $|x_1| < |x_2|$, we say that H is a hyperbolic rotation of type 2 and we have

$$c = \frac{x_1}{\sqrt{x_2^2 - x_1^2}}, \ s = \frac{x_2}{\sqrt{x_2^2 - x_1^2}} \quad \text{and} \quad s^2 - c^2 = 1.$$

We recall that while orthogonal rotations are perfectly well conditioned, hyperbolic rotations satisfy

$$\kappa_2(H) = \frac{|c| + |s|}{||c| - |s||},$$

which means they can be arbitrarily ill conditioned.

4.2.2 Householder Reflectors

Let $J \in \text{diag}_n^q(\pm 1)$ and $u \in \mathbb{R}^n$ such that $\langle u, u \rangle_J \neq 0$ where $\langle u, u \rangle_J = \langle Ju, u \rangle$ and $(w, v) \mapsto \langle w, v \rangle$ denotes the usual inner product over $\mathbb{R}^n$. For $u \in \mathbb{R}^n$, a hyperbolic Householder reflector [60] has the form

$$H(u) = J - \frac{2}{\langle u, u \rangle_J} uu^T. \tag{4.2}$$

$H(u)$ is J-orthogonal and for any permutation P, $H(u)P$ is $(J, \widetilde{J})$-orthogonal with $\widetilde{J} = P^T JP$. The first purpose of this section is to solve the problem given: $x, y \in \mathbb{R}^n$ find $u \in \mathbb{R}^n$ such that

$$H(u)x = \alpha y, \tag{4.3}$$

for some $\alpha \in \mathbb{R} \backslash \{0\}$. In the second part of this section, we focus on the numerical stability of hyperbolic Householder reflector mainly by computing the condition number of these transformations.

Assume that (4.3) is satisfied. Then, since $H(u)$ preserves the indefinite norm, we have

$$\langle x, x \rangle_J = \langle H(u)x, H(u)x \rangle_J = \alpha^2 \langle y, y \rangle_J .$$

If $\langle x, x \rangle_J = 0$ then it implies that $\langle y, y \rangle_J = 0$ since $H(u)$ is nonsingular. In this case it is still possible to define $H(u)$ if $\langle x, y \rangle \neq 0$. But in most applications, $y = e_1$ which implies $\langle y, y \rangle_J \neq 0$ for all $J \in \text{diag}_q^n(\pm 1)$. Thus, if $y = e_1$ and $\langle x, x \rangle_J = 0$, we have to look for a permutation matrix P such that $\langle Px, Px \rangle_J \neq 0$. Finally, if both $\langle x, x \rangle_J \neq 0$ and $\langle y, y \rangle_J \neq 0$, we still need the sign of each quantities to agree in order to work with real matrices.

We are now able to give the following theorem that ensure the existence of hyperbolic Householder reflectors in some cases.

Theorem 4.1 *Let $J \in \mathrm{diag}_n^k(\pm 1)$ and let $x, y \in \mathbb{R}^n$ such that $\frac{\langle x,x\rangle_J}{\langle y,y\rangle_J} > 0$. Define $u = Jx - \alpha y \in \mathbb{R}^n$ with $\alpha = \pm\sqrt{\frac{\langle x,x\rangle_J}{\langle y,y\rangle_J}}$. Then, the hyperbolic Householder reflector $H(u)$ satisfies $H(u)x = \alpha y$.*

Proof. We have that

$$\langle u, u\rangle_J = 2(\langle x, x\rangle_J - \alpha \langle x, y\rangle) \quad \text{and} \quad \langle x, u\rangle = \langle x, x\rangle_J - \alpha \langle x, y\rangle .$$

Thus,

$$H(u)x = \left(1 - 2\frac{\langle x, u\rangle}{\langle u, u\rangle_J}\right) Jx + 2\alpha\frac{\langle x, u\rangle}{\langle u, u\rangle_J} y = \alpha y.$$

Finally, note that for any constant μ, $H(\mu u) = H(u)$ and if $H(u)x = \alpha y$ then u belongs to the linear subspace spanned by $Jx - \alpha y$. □

We are interested in computing the condition number of $H(u)$ for some $u \in \mathbb{R}^n$. Since $H(u)^T = H(u)$, we focus on the spectral properties of $H(u)$.

Theorem 4.2 *The eigenvalues of $H(u)$ are*

$$\lambda_{1,2} = -\frac{\|u\|_2^2}{\langle u, u\rangle_J} \pm \sqrt{\frac{\|u\|_2^4}{\langle u, u\rangle_J^2} - 1},$$

corresponding to the eigenvectors $v_{1,2} = \lambda_{1,2}u + Ju$ and $n-2$ eigenvalue equal ± 1 corresponding to the $n-2$ eigenvectors that lie in the $n-2$ orthogonal directions to $v_{1,2}$.

Proof. Let (λ, v) be an eigenpair of $H(u)$. If $\langle v, u\rangle = 0$ then λ is one of the diagonal element of J. Otherwise, we have that

$$H(u)v = Jv - 2\frac{\langle v, u\rangle}{\langle u, u\rangle_J} u,$$

which leads us to assume that $v = \alpha u + \beta Ju$ where $\alpha, \beta \in \mathbb{R}$. We have

$$H(u)v = \alpha Ju + \left(\beta - 2\frac{\langle v, u\rangle}{\langle u, u\rangle_J}\right) u = \lambda\alpha u + \lambda\beta Ju.$$

Since $J \neq \pm I$ (otherwise $H(u)$ will be the usual orthogonal Householder matrix), we have that u and Ju are linearly independent. Thus, we have to solve the system

$$\begin{aligned} \alpha &= \lambda\beta, & (4.4)\\ \beta - 2\frac{\langle \alpha u + \beta Ju, u\rangle}{\langle u, u\rangle_J} &= \lambda\alpha. & (4.5) \end{aligned}$$

From Equation (4.4) we have that $\alpha \neq 0$ and $\beta \neq 0$ since $\lambda \neq 0$. Thus, by substituting $\lambda = \frac{\alpha}{\beta}$ in (4.5), we get the quadratic equation

$$\lambda^2 + 2\frac{\|u\|_2^2}{\langle u, u\rangle_J} + 1 = 0.$$

Since $\frac{\|u\|_2^2}{\langle u,u\rangle_J} > 1$, the solution are real and they are given by

$$\lambda_{1,2} = -\frac{\|u\|_2^2}{\langle u, u\rangle_J} \pm \sqrt{\frac{\|u\|_2^4}{\langle u, u\rangle_J^2} - 1}.$$

Note that $\frac{1}{\lambda_1} = \lambda_2$. We get that v_1 is orthogonal to v_2 and any vector orthogonal to v_1 and v_2 is orthogonal to u which proves the theorem. □

The following corollary gives the condition number of a hyperbolic Householder matrix and shows that they can be arbitrarily ill conditioned as much as hyperbolic rotations.

Corollary 4.3 *The condition number of $H(u)$ for the 2-norm is given by*

$$\kappa_2(H(u)) = \left(\frac{\|u\|_2^2}{|\langle u, u\rangle_J|} + \sqrt{\frac{\|u\|_2^4}{\langle u, u\rangle_J^2} - 1}\right)^2.$$

4.2.3 Error Analysis

We start by presenting the standard model for floating point arithmetic. We assume the existence of the fl operator that satisfies

$$\begin{aligned} fl(x \diamond y) &= (x \diamond y)(1+\epsilon_1), \quad |\epsilon_1| \leq \mathrm{u},\\ fl(\sqrt{x}) &= \sqrt{x}(1+\epsilon_2), \qquad |\epsilon_2| \leq \mathrm{u}, \end{aligned}$$

where $\diamond$ denotes one of the algebraic operations $+, -, \times, /$ and u is the unit roundoff.

It is well know for the orthogonal case that the computed values of c and s in Givens rotation satisfies $fl(c) = c(1+\epsilon_c)$ and $fl(s) = c(1+\epsilon_s)$ with $|\epsilon_c| = O(\mathrm{u})$ and $|\epsilon_s| = O(\mathrm{u})$. For an orthogonal Householder reflector H, the computed matrix satisfies $\|fl(H) - H\|_2 \leq 10\mathrm{u}$. All these classical results and a more detailed presentation of a model for floating point numbers can be found in [31, 78]. In the hyperbolic case (with $J \neq \pm I$), several authors noticed that the way hyperbolic transformations are applied to a vector is crucial for stability and also the method of computing the c and s is of first importance to ensure a small relative error. We analyze this problem for a 2×2 hyperbolic rotation of type 1. The main problem in computing the values of c and s is how to compute the indefinite scalar product $\rho = x_1^2 - x_2^2$. First, we consider the two following ways of computing ρ:

$$\rho_1 = x_1^2 - x_2^2 \quad \text{and} \quad \rho_2 = (x_1 - x_2)(x_1 + x_2).$$

We have

$$\begin{aligned} fl(\rho_1) &= (x_1^2(1+\epsilon_1) - x_2^2(1+\epsilon_2))(1+\epsilon_3), \\ \frac{|fl(\rho_1) - \rho_1|}{|\rho_1|} &\leq \mathrm{u}(1 + \|x\|_2^2 \frac{(1+\mathrm{u})}{|\rho_1|}), \end{aligned}$$

where $|\epsilon_k| \leq \mathrm{u}$ for $k = 1{:}\,3$. We see that the relative error for ρ_1 is unbounded as x_1 becomes closer to x_2. For the second approach, we have

$$\begin{aligned} fl(\rho_2) &= \rho_2(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3), \\ \frac{|fl(\rho_2) - \rho_2|}{|\rho_2|} &\leq \mathrm{u}(3 + 3\mathrm{u} + \mathrm{u}^2), \end{aligned}$$

where $|\epsilon_k| \leq \mathrm{u}$ for $k = 1{:}\,3$. It is clear that the second method is numerically stable although it may suffer from overflow. Note that if $|x_1| > |x_2|$ and $t = \frac{x_2}{x_1}$ then the corresponding $\rho_3 = x_1(1-t)(1+t)$ is still unstable like the first method.

Another way of computing c and s is through the eigenvalues of H. For example, the eigenvalues of H in (4.1) (type 1 rotation) are given by

$$\lambda_{\pm} = c \pm s = \sqrt{\frac{x_1 \pm x_2}{x_1 \mp x_2}}.$$

Thus, we obtain

$$c = \frac{\lambda_+ + \lambda_-}{2} = \frac{\lambda_-^2 + 1}{2\lambda_-}, \tag{4.6}$$

$$s = \frac{\lambda_+ - \lambda_-}{2} = \frac{1 - \lambda_-^2}{2\lambda_-}. \tag{4.7}$$

We chose λ_- in the expression of c and s because in this case $|\lambda_-| \geq |\lambda_+|$ and thus the computation of λ_- is more stable than the computation of λ_+. The error analysis for λ_- gives

$$\begin{aligned} \frac{|fl(\lambda_-) - \lambda_-|}{|\lambda_-|} &= |\gamma_1 - 1|, \\ \gamma_1 &= (1+\epsilon_1)\sqrt{\frac{(1+\epsilon_2)(1+\epsilon_3)}{1+\epsilon_4}}, \end{aligned}$$

where $|\epsilon_k| \leq \mathrm{u}$ for $k = 1\colon 4$. We have that

$$\begin{aligned} \gamma_1 - 1 &\leq \mathrm{u}\frac{3+\mathrm{u}}{1-\mathrm{u}}, \qquad 1 - M \leq \mathrm{u}\frac{-3+\mathrm{u}}{1+\mathrm{u}}, \\ |\gamma_1 - 1| &\leq \mathrm{u}\max(\frac{3+\mathrm{u}}{1-\mathrm{u}}, \frac{-3+\mathrm{u}}{1+\mathrm{u}}), \\ |\gamma_1 - 1| &\leq \mathrm{u}\frac{3+\mathrm{u}}{1-\mathrm{u}} \leq 3\mathrm{u} + 8\mathrm{u}^2. \end{aligned}$$

Thus, $fl(\lambda_-) = (1+\alpha_1)\lambda_-$ where $|\alpha_1| \leq 3\mathrm{u} + 8\mathrm{u}^2$. Moreover, we have

$$\begin{aligned} fl(c) &= \gamma_2 \frac{1 + (1+\epsilon_1)(1+\alpha_1)^2\lambda_-^2}{2\lambda_-}, \\ \gamma_2 &= \frac{(1+\epsilon_2)(1+\epsilon_3)}{(1+\epsilon_4)(1+\alpha_1)}, \end{aligned}$$

where $|\epsilon_k| \leq \mathrm{u}$ for $k = 1\colon 4$. We have

$$\begin{aligned} |\gamma_2 - 1| &\leq 7\mathrm{u} + c_1\mathrm{u}^2, \quad \gamma_2 = 1 + \alpha_2, \quad |\alpha_2| \leq 7\mathrm{u} + c_1\mathrm{u}^2, \\ (1+\epsilon_1)(1+\alpha_1)^2 &= 1 + \alpha_3, \quad |\alpha_3| \leq 7\mathrm{u} + c_2\mathrm{u}^2, \end{aligned}$$

where c_1, c_2 are constants. Thus,

$$\frac{|fl(c)-c|}{|c|} \;\leq\; \alpha_2+\alpha_3+O(\mathrm{u}^2) \;\leq\; 14\mathrm{u}+O(\mathrm{u}^2).$$

Note that the computation of s with this method is unstable if λ_+ or λ_- are close to 1. But $|\lambda_\pm| \to 1$ if $\frac{x_1}{x_2} \to 0$ or $\frac{x_2}{x_1} \to 0$ and thus in this case the other methods mentioned at the beginning of the paragraph are stable.

We compare numerically the relative error of the computed c and s with the different methods. Let

$$\begin{aligned}
(\hat{c}_1, \hat{s}_1) &= (fl(\frac{x_1}{\sqrt{x_1^2-x_2^2}}), fl(\frac{x_2}{\sqrt{x_1^2-x_2^2}})),\\
(\hat{c}_2, \hat{s}_2) &= (fl(\frac{x_1}{\sqrt{(x_1-x_2)(x_1+x_2)}}), fl(\frac{x_2}{\sqrt{(x_1-x_2)(x_1+x_2)}})),\\
(\hat{c}_3, \hat{s}_3) &= (fl(\frac{\lambda_-^2+1}{2\lambda_-}), fl(\frac{1-\lambda_-^2}{2\lambda_-})).
\end{aligned}$$

The corresponding relative errors are denoted by

$$\begin{aligned}
Rc_k &= \frac{|\hat{c}_k-c|}{|c|} \quad \text{and} \quad Rs_k = \frac{|\hat{s}_k-s|}{|s|}, k = 1\colon 3,\\
R_k &= \max(Rc_k, Rs_k).
\end{aligned}$$

We compute the values of c and s for $x = [\,\xi \quad \xi-\delta\,]^T$ where $\xi \in \mathbb{R}$ and δ is small parameter. The exact values of c and s where computed in extended precision. Let $p = |\frac{\log(\delta)}{\log(10)}|$. The numerical results are displayed in Table 4.1. We see that the last to strategies are more stable numerically. They have acceptable residues, which confirms our analysis.

4.2.4 Zeroing Strategies

Let $x \in \mathbb{R}^n$ and $J \in \mathrm{diag}_n^k(\pm 1)$. Our aim in this paragraph is to discuss the different zeroing approaches. We start with rotations. We can apply $n-1$

Table 4.1: Relative errors for c and s.

p	$\xi = 1$			$\xi = 100$			$\xi = 10^4$		
	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3
1	0	0	0	5.10^{-14}	3.10^{-14}	3.10^{-14}	3.10^{-12}	2.10^{-12}	2.10^{-12}
2	0	0	0	5.10^{-13}	2.10^{-13}	2.10^{-13}	3.10^{-12}	6.10^{-17}	1.10^{-16}
3	6.10^{-16}	6.10^{-16}	5.10^{-16}	3.10^{-12}	2.10^{-12}	5.10^{-16}	2.10^{-10}	7.10^{-17}	8.10^{-17}
4	5.10^{-14}	6.10^{-14}	5.10^{-14}	2.10^{-12}	1.10^{-16}	2.10^{-12}	5.10^{-10}	6.10^{-17}	6.10^{-17}
5	2.10^{-12}	2.10^{-12}	2.10^{-12}	1.10^{-10}	3.10^{-17}	4.10^{-17}	1.10^{-9}	6.10^{-17}	1.10^{-16}
6	4.10^{-18}	4.10^{-18}	2.10^{-16}	2.10^{-9}	5.10^{-17}	9.10^{-17}	1.10^{-7}	1.10^{-16}	1.10^{-16}
7	8.10^{-17}	8.10^{-17}	2.10^{-16}	1.10^{-8}	5.10^{-17}	2.10^{-16}	1.10^{-6}	4.10^{-17}	8.10^{-17}
8	2.10^{-17}	2.10^{-17}	2.10^{-16}	2.10^{-7}	6.10^{-17}	9.10^{-17}	4.10^{-6}	9.10^{-17}	9.10^{-17}
9	7.10^{-17}	7.10^{-17}	9.10^{-17}	2.10^{-6}	7.10^{-17}	6.10^{-17}	6.10^{-5}	8.10^{-17}	8.10^{-17}

rotations H_k, with $1 \leq k \leq n-1$, such that

$$H = \prod_{k=1}^{n-1} H_k \quad \text{and} \quad Hx = \rho e_1.$$

Since hyperbolic rotations can be ill-conditioned, we need to monitor their condition number and minimize their number. For example, consider this case in $\mathbb{R}^3$:

$$x = [\, x_1 \quad x_2 \quad x_3 \,]^T, \quad J = \text{diag}(-1, -1, 1), \quad |x_1| < |x_2| < |x_3|.$$

If one chooses to annihilate x_3 first and then x_2 one needs two hyperbolic rotations. On the other hand, x_2 can be zeroed first with an orthogonal Givens rotation in the $(1, 2)$ plane and then a final hyperbolic rotation in the $(1, 3)$ plane is used to eliminate x_3. This strategy has two main advantages. First, it reduces the number of hyperbolic rotations used to at most 1. Secondly, it minimizes the risk of having two hyperbolic rotations acting in the same plane. This tends to reduce the growth of rounding errors and increases the chance that the largest condition number of the individual transformations H_k is of the same order of magnitude as the condition number of the overall transformation H.

Thus, the best strategy is to apply first all the orthogonal rotations and to apply last the hyperbolic rotations. With this strategy we only apply at most one hyperbolic rotations during the zeroing process. The following algorithm is the implementation of this strategy.

Algorithm 4.4 *Given* $x \in \mathbb{R}^n$ *and* $J = \mathrm{diag}(\sigma_k) \in \mathrm{diag}_n^k(\pm 1)$ *the following algorithm compute* $H \in \mathbb{R}^{n\times n}$ *such that* $Hx = \rho e_1$ *and* $H^T JH \in \mathrm{diag}_n^k(\pm 1)$.

Set $H = I$

Find I_+ the list of indices such that $\sigma_i = 1$ if $i \in I_+$

Find I_- the list of indices such that $\sigma_i = -1$ if $i \in I_-$

Let n_1, n_2 be the respective lengths of I_+ and I_-

Let i_1, $\tilde{i}_1$ be respectively the first elements of I_+ and I_-

for $k = 1 : n_1$

Apply a Givens rotation H_k in the (i_1, k) plane such that

$H_k \begin{bmatrix} x_{i_1} & x_k \end{bmatrix}^T = \rho_k e_1$

$H([i_1, k], :) = H_k H([i_1, k], :)$, $x([i_1, k]) = H_k x([i_1, k])$

end

for $k = 1 : n_2$

Apply a Givens rotation H_k in the $(\tilde{i}_1, k)$ plane such that

$H_k \begin{bmatrix} x_{\tilde{i}_1} & x_k \end{bmatrix}^T = \rho_k e_1$

$H([\tilde{i}_1, k], :) = H_k H([\tilde{i}_1, k], :)$, $x([\tilde{i}_1, k]) = H_k x([\tilde{i}_1, k])$

end

Set $k = (1, \max(i_1, \tilde{i}_1))$

Apply the hyperbolic rotation H_{n-1} in the $(1, k)$ plane such that

$H_{n-1} \begin{bmatrix} x_1 & x_k \end{bmatrix}^T = \rho_{n-1} e_1$

$H([1, k], :) = H_{n-1} H([1, k], :)$, $x([1, k]) = H_k x([1, k])$

In [9], the authors noticed that the way hyperbolic rotations are applied to a

vector is of first importance to maintain accuracy and stability. This method can be described as follows. Let $x = [\, x_1 \quad x_2 \,]^T$, let H be given by (4.1), a hyperbolic matrix of the first type and define $y = Hx$,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} cx_1 - sx_2 \\ -sx_1 + cx_2 \end{bmatrix}.$$

We have

$$\begin{bmatrix} x_1 \\ y_2 \end{bmatrix} = \widetilde{H} \begin{bmatrix} y_1 \\ x_2 \end{bmatrix}, \quad \widetilde{H} = \frac{1}{c} \begin{bmatrix} 1 & s \\ -s & 1 \end{bmatrix}.$$

$\widetilde{H}$ is a Givens rotation which suggests that the computation of y_2 is likely to be more stable than the computation of y_1. Now that we have the value of y_2 we can apply the same method for y_1. We have that

$$\begin{bmatrix} y_1 \\ x_2 \end{bmatrix} = \widehat{H} \begin{bmatrix} x_1 \\ y_2 \end{bmatrix}, \quad \widehat{H} = \frac{1}{c} \begin{bmatrix} 1 & -s \\ s & 1 \end{bmatrix}.$$

We recall that H and $\widetilde{H}$ are related by the exchange operator $\widetilde{H} = \mathrm{exc}(H)$ defined in [38]. The exchange operator maps hyperbolic matrices to orthogonal matrices and it satisfies $\mathrm{exc}(\mathrm{exc}) = \mathrm{exc}$.

We can also apply a Householder hyperbolic matrix described in Paragraph 4.2.2 to a vector to introduce zeros. In order to monitor the condition number of these transformations, we need the same type of strategy described in Algorithm 4.4. Let I_+ and I_- be defined as in Algorithm 4.4. Let $y = x(I_+) \in \mathbb{R}^{n_1}$ and $z = x(I_-) \in \mathbb{R}^{n_2}$. Let $G_1 = (I - 2uu^T) \in \mathbb{R}^{n_1 \times n_1}$ and $G_2 = (I - 2vv^T) \in \mathbb{R}^{n_2 \times n_2}$ be the orthogonal Householder matrices such that $G_1 y = \rho_1 e_1$ and $G_2 z = \rho_2 e_2$. We define $\tilde{u} \in \mathbb{R}^n$ and $\tilde{v} \in \mathbb{R}^n$ by

$$\tilde{u}_k = \begin{cases} u_{I_+(k)} & \text{if } k \in I_+, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \tilde{v}_k = \begin{cases} v_{I_-(k)} & \text{if } k \in I_-, \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Then, let $\widetilde{G}_1 = I - 2\tilde{u}\tilde{u}^T$ and $\widetilde{G}_2 = I - 2\tilde{v}\tilde{v}^T$. We have that

$$\widetilde{G}_1 \widetilde{G}_2 x = \widetilde{G}_2 \widetilde{G}_1 x = \tilde{\rho}_1 e_1 + \tilde{\rho}_2 e_k,$$

where $\tilde{\rho}_1, \tilde{\rho}_2 \in \mathbb{R}$ and $k = \max(i_1, \tilde{i}_1)$ where i_1 and $\tilde{i}_1$ are the first elements of I_+ and I_-. To finish the zeroing process, we just need to apply one 2×2 hyperbolic Householder reflector (or one 2×2 hyperbolic rotation) in the $(1, k)$ plane. The following algorithm describes the implementation of the zeroing strategy using Householder type transformations.

Algorithm 4.5 *Given $x \in \mathbb{R}^n$ and $J = \mathrm{diag}(\sigma_k) \in \mathrm{diag}_n^k(\pm 1)$ the following algorithm computes $H \in \mathbb{R}^{n \times n}$ such that $Hx = \rho e_1$ and $H^T J H \in \mathrm{diag}_n^k(\pm 1)$.*

Find I_+ the list of indices such that $\sigma_i = 1$ if $i \in I_+$
Find I_- the list of indices such that $\sigma_i = -1$ if $i \in I_-$
Let n_1, n_2 be the respective lengths of I_+ and I_-
Let i_1, $\tilde{i}_1$ be respectively the first elements of I_+ and I_-
Set $y = x(I_+)$, $z = x(I_-)$
Compute orthogonal Householder matrix $G_1 = I - 2uu^T$ such that $G_1 y = \rho e_1$
Compute orthogonal Householder matrix $G_2 = I - 2vv^T$ such that $G_2 z = \rho e_2$
Compute $\tilde{u}$ and $\tilde{v}$ using (4.8) and
the associated Householder matrices $\widetilde{G}_1$ and $\widetilde{G}_2$
Set $k = \max(i_1, \tilde{i}_1)$, $H = \widetilde{G}_1 \widetilde{G}_2$ and $x = Hx = \rho_1 e_1 + \rho_2 e_k$
Apply hyperbolic Householder transformation $\widetilde{G}$ in the $(1, k)$ plane.
Set $H = \widetilde{G} H$

One can easily see that using the first method seems a good choice to introduce zeros in a sparse matrix with unified rotations whereas the second approach with the Householder reflectors is a better choice to introduce zeros in a full matrix.

In the following sections, we investigate several matrix factorizations that have a hyperbolic or an orthogonal factor.

4.3 Introduction to Matrix Factorization

Matrix factorization is a common tool in different branches of mathematics. A general definition is given in [5]:

> A matrix factorization theorem is an assertion that a matrix A can be factorized into a product $A = A_1 A_2$, of two special matrices A_1, A_2. Some conditions may be necessary for such a decomposition to exist, and some further conditions may ensure the uniqueness of the factorization.

Throughout this chapter, we encounter matrix factorizations in which more than two matrices are involved. The aim of this work is to analyze the sensitivity of some matrix factorizations that involve at least one hyperbolic matrix and to give a first order perturbation bound for the factors. The optimal first order perturbation bound yields a condition number of the relevant matrix in the factorization, which measures its sensitivity to perturbations in the data. For $A, X, Y \in \mathbb{R}^{n\times n}$, let $\varphi(X, Y) = A$ be a factorization of A, where φ is a function describing the factorization. For instance for the QR factorization, $\varphi(X, Y) = XY$, where X is unitary and R upper triangular. The classical theory of condition numbers [61] employs the definitions

$$\begin{aligned}\kappa_X &= \limsup_{\epsilon\to 0}\left\{\epsilon^{-1}\|\Delta X\|,\ \varphi(X+\Delta X, Y+\Delta Y) = A+\Delta A,\ \|\Delta A\| \le \epsilon\right\},\\ \kappa_Y &= \limsup_{\epsilon\to 0}\left\{\epsilon^{-1}\|\Delta Y\|,\ \varphi(X+\Delta X, Y+\Delta Y) = A+\Delta A,\ \|\Delta A\| \le \epsilon\right\}.\end{aligned}$$

This definition has the advantage of being simple to present, although in most cases the necessary computations to bound the condition number or to show it is attained are far from being trivial. The method used in this thesis is quite different. Our aim is to define a function g in a neighborhood of A such that $\varphi(\widetilde{X}, \widetilde{Y}) = A + \Delta A$ with $(\widetilde{X}, \widetilde{Y}) = g(A + \Delta A)$. We define the condition number

as the norm of $\|dg(A)\|$, the differential of g at A. The main tool for this analysis is the implicit function theorem. Our method is described in detail in Section 4.4.

Several results that we cite later on are available concerning orthogonal matrix factorizations. In most cases, these results are only bounds and not the condition number. In the literature, condition numbers for hyperbolic matrix factorization have not been reported. In the rest of this chapter, we investigate some matrix factorizations and for each of them we compute the condition number. Our proof technique is not new. It was used in [5] and [24] to investigate perturbation bounds for several matrix factorizations.

The HR factorization is the generalization of the usual QR factorization when the orthogonal factor is allowed to be $(J, \widetilde{J})$-orthogonal. Perturbation bounds of the QR factorization are given for example in [5], [17], [24], [45] and [65]. We compute the condition number of the HR factorization and show that the classical perturbation bounds for the QR factorization are very weak. Our analysis (of the HR factorization) is closer to the ones presented in [5], [24] and [45]. For the singular value decomposition (SVD), it is well known that the condition number for the singular values is 1 (see for example [66]). Perturbation bounds for the singular vectors are also available in [64], [76]. In our case, we compute the condition number of the hyperbolic SVD (see Section 4.7), which is the generalization of the usual SVD. In several papers, the polar factorization have been analyzed (see for example [16], [34], [36], [44], [51]). Once more, we compute the condition number of the indefinite polar factorization and apply our results to the usual polar factorization, which allows us to give a short and easy computation of its condition number. We also refer to two surveys, [5] where perturbation bounds for several matrix factorizations are given and [36] where various conditioning problems are treated.

4.4 A General Method for Computing the Condition Number

Let $\mathbb{S}$ be a linear subspace of $\mathbb{K}^{n\times n}$, $X \in \mathbb{S}$ and let $\mathcal{V}_X \subset \mathbb{S}$ be an open neighborhood of X. $\mathbb{S}$ can be regarded as a set of matrices that have a particular structure such as symmetry, Hermitian or a sparse structure such as upper triangular. In this section, let $\mathcal{H}$ denote either $\mathcal{O}_{mn}(J, \widetilde{J})$, $\mathcal{O}_{mn}(J, \widetilde{J}, \mathbb{C})$ or $\mathcal{U}_{mn}(J, \widetilde{J})$. Let $\mathcal{E}$ be a linear subspace of $\mathbb{K}^{n\times n}$. A $(J, \widetilde{J})$-orthogonal or $(J, \widetilde{J})$-unitary factorization of a matrix $A \in \mathcal{E}$ can be described by a function

$$\varphi(X, Y) = A, \;\; X \in \mathcal{V}_X \;\; \text{and} \;\; Y \in \mathcal{H}$$

Our aim is to derive perturbation bounds for the X and Y factors when A is subject to some perturbation ΔA. The main tool for this analysis is the implicit function theorem. This technique was also used by Bhatia in [5]. This method is divided into three steps.

Step 1 Using (1.13), we define

$$\begin{aligned} f : \mathcal{V}_A \times \mathcal{V}_X \times \mathbb{K}^p &\rightarrow \mathcal{E}, \\ (\widetilde{A}, \widetilde{X}, \tilde{y}) &\mapsto \varphi(\widetilde{X}, \widetilde{Y}) - \widetilde{A}, \end{aligned}$$

where $\widetilde{Y} = \phi(\tilde{y})$ is defined according to $\mathcal{H}$ in Lemma 1.3 and p is the dimension of $\mathcal{H}$. Note that $f(A, X, y) = 0$, with $Y = \phi(y)$. Assume that f is differentiable. We denote the differential of f in the X and y direction by

$$df_2(A, X, y) = \frac{\partial f}{\partial X} + \frac{\partial f}{\partial y}.$$

For all the factorizations, $df_2(A, X, y)$ can be easily computed because φ is linear in X and at most quadratic in Y.

Step 2 In order to apply the implicit function theorem to f at (A, X, y), $df_2(A, X, y)$ has to be nonsingular. Thus, null$(df_2(A, X, y))$ needs to be computed, that is we need to solve the equation

$$df_2(A, X, y)(\Delta X, \Delta Y) = 0,\ \Delta X \in \mathcal{E},\ \Delta Y = d\phi(y)\Delta y,$$

with $\Delta y \in \mathbb{K}^p$. Using Section 1.5, ΔY is in the tangent space of $\mathcal{H}$. Assume that null$(df_2(A, X, y)) = \{0\}$. Then, by computing $d\varphi(X, Y)$, we have that

$$\{0\} = \text{range}\left(\frac{\partial \varphi}{\partial X}\right)_{|\mathbb{S}} \bigcap \text{range}\left(\frac{\partial \varphi}{\partial y}\right)_{|T(\mathcal{H})}.$$

Additionally, using (1.9)-(1.12) if $\dim \mathcal{E} = \dim \mathbb{S} + p$ then we have that $df_2(A, X, y)$ is invertible and the following splitting of $\mathcal{E}$ into a direct sum decomposition of the type

$$\mathcal{E} = \text{range}\left(\frac{\partial \varphi}{\partial X}\right)_{|\mathbb{S}} \bigoplus \text{range}\left(\frac{\partial \varphi}{\partial y}\right)_{|T(\mathcal{H})}, \tag{4.9}$$

holds, where $T(\mathcal{H}_n)$ is the tangent space of $\mathcal{H}_n$ at Y. The advantage of (4.9) is that it enable us to invert $df_2(A, X, y)$ by using the corresponding projector to the direct sum. Then, by the implicit function theorem, there exists a differentiable function $g = (g_X, g_Y)$ and an open neighborhood $\mathcal{V}_A$ of A satisfying

$$\begin{aligned} g : \mathcal{V}_A &\rightarrow \mathcal{V}_X \times \mathcal{V}_Y, \\ \widetilde{A} &\mapsto (g_X(\widetilde{A}), g_Y(\widetilde{A})), \end{aligned} \tag{4.10}$$

where $\mathcal{V}_X \times \mathcal{V}_Y$ is an open neighborhood of (X, Y). Moreover, g satisfies $g_X(A) = X,\ g_Y(A) = Y$ and

$$dg(A) = -(d_2 f(A, X, y))^{-1} \frac{\partial f}{\partial A}, \tag{4.11}$$

$$f(\widetilde{A}, g_X(\widetilde{A}), g_Y(\widetilde{A})) = 0, \quad \text{for all } \widetilde{A} \in \mathcal{V}_A, \tag{4.12}$$

that is $\widetilde{A} = \varphi(g_1(\widetilde{A}), g_1(\widetilde{A}))$ is the factorization of $\widetilde{A}$. Let $\Pi_{\mathbb{S}}$ and $\Pi_{T(\mathcal{H}_n)}$ denote the projectors corresponding to (4.9). We have that $\frac{\partial f}{\partial A} = -I$. Thus, (4.11) becomes

$$dg_X(A)\Delta A = \left(\frac{\partial \varphi}{\partial X}\right)^{-1} \Pi_{\mathbb{S}} \Delta A, \tag{4.13}$$

$$dg_Y(A)\Delta A = \left(\frac{\partial \varphi}{\partial Y}\right)^{-1} \Pi_{T(\mathcal{H})} \Delta A. \tag{4.14}$$

Step 3 The condition number of the factorization is given by the norm of the linear map $dg(A)$. In some cases, only a bound for the norm of $dg(A)$ will be given. Finally, for $\widetilde{A} \in \mathcal{V}_A$ and $\varphi(\widetilde{X}, \widetilde{Y}) = \widetilde{A}$, the first order perturbation bounds and expansion are obtained using Taylor's theorem

$$\|\widetilde{X} - X\|_F \leq \|dg_X(A)\|_2 \epsilon + O(\epsilon^2), \tag{4.15}$$

$$\|\widetilde{Y} - Y\|_F \leq \|dg_Y(A)\|_2 \epsilon + O(\epsilon^2), \tag{4.16}$$

where $\epsilon = \|\widetilde{A} - A\|_F$.

4.5 The HR Factorization

We say that $A \in \mathbb{R}^{n\times n}$ admits an HR factorization with respect to a signature matrix $J \in \operatorname{diag}_n^k(\pm 1)$ if

$$A = HR, \qquad R \in \triangle(\mathbb{R}),\ H \in \mathcal{O}_n(J, \widetilde{J}),$$

where $\widetilde{J} \in \operatorname{diag}_n^k(\pm 1)$. The next theorem from [14] shows that almost every matrix has an *HR factorization with respect to J*.

Theorem 4.6 *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and $J \in \operatorname{diag}_q^n(\pm 1)$. There exist $H, R \in \mathbb{R}^{n\times n}$ such that $H^T J H \in \operatorname{diag}_q^n(\pm 1)$, R is upper triangular and $A = HR$ if and only if all principal minors of $A^T J A$ are nonzero.*

Proof. The assumption that all principal minors of A^TJA are nonzero ensures that A^TJA has an LU factorization

$$A^TJA = LU,$$

with L a unit lower triangular matrix and U a nonsingular upper triangular matrix. Let $D = \mathrm{diag}(U)$ and $U = D\tilde{L}^T$, where $\tilde{L}$ is a unit lower triangular matrix. Since A^TJA is symmetric, we get that $L = \tilde{L}$. We define $\tilde{J} = \mathrm{sign}(D)$ and we have $A^TJA = L|D|^{\frac{1}{2}}\tilde{J}|D|^{\frac{1}{2}}L^T$. By Sylvester's inertia theorem $\tilde{J} \in \mathrm{diag}_q^n(\pm 1)$. We define $R = |D|^{\frac{1}{2}}L^T$ and $H = AR^{-1}$. We have

$$H^TJH = R^{-T}A^TJAR^{-1} = R^{-T}L|D|^{1/2}\widetilde{J}|D|^{1/2}L^TR^{-1} = \widetilde{J}.$$

Hence A can be factorized into $A = HR$ with H such that $H^TJH \in \mathrm{diag}_q^n(\pm 1)$ and R is upper triangular.

We suppose now that $A = HR$, where H is a $(J, \widetilde{J})$-orthogonal and R upper triangular. Then

$$A^TJA = R^TH^TJHR = R^T\widetilde{J}R.$$

Since A is nonsingular, R is nonsingular. Moreover, for $\widetilde{A} = A^TJA$,

$$\widetilde{A}(1\colon k, 1\colon k) = R^T(1\colon k, 1\colon k)\tilde{J}(1\colon k, 1\colon k)R(1\colon k, 1\colon k),\ k = 1\colon n,$$

which shows that all the leading principal submatrices of A^TJA are nonsingular. □

For nonsingular matrices the HR factorization is unique up to a signature matrix. We can make it unique by insisting that R has positive diagonal entries.

For $A \in \mathbb{R}^{m\times n}$, with $m > n$, the HR factorization with respect to $J \in \mathrm{diag}_q^m(\pm 1)$ is $A = HR$, where $H \in \mathbb{R}^{m\times m}$, $H^TJH \in \mathrm{diag}_q^m(\pm 1)$ and $R \in \mathbb{R}^{m\times n}$ is upper trapezoidal.We now give two theorems that we shall use later on. We need the following result for the implementation of the HZ algorithm (see Chapter 6) in the eventual case where a shift might be an eigenvalue.

Theorem 4.7 *Let $A \in \mathbb{R}^{m\times n}$ with $m > n$ having full rank and $J \in \mathrm{diag}_q^m(\pm 1)$. A has an HR factorization with respect to J if and only if all the principal minors of $A^T JA$ are nonzero.*

Proof. Assume that all the principal minors of $A^T JA$ are nonzero. Then, like in Theorem 4.6, $A^T JA$ can be factorized as

$$A^T JA = L|D|^{\frac{1}{2}}\tilde{J}_1|D|^{\frac{1}{2}}L^T,$$

where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular, $D \in \mathbb{R}^{n\times n}$ is nonsingular diagonal and $\tilde{J}_1 \in \mathrm{diag}_{q_1}^m(\pm 1)$ for some integer q_1. Let $\tilde{R} = |D|^{\frac{1}{2}}L^T$ and $R = [\tilde{R}^T 0]^T \in \mathbb{R}^{m\times n}$ and define $H_1 = A\tilde{R}^{-1} \in \mathbb{R}^{m\times n}$. We have that $H_1^T JH_1 = \tilde{J}_1$. Let $H_2 \in \mathbb{R}^{m\times(m-n)}$ such that $H = [H_1,\ H_2]$ is nonsingular. H_2 can be chosen such that its columns are $J-$orthogonal to the columns of H_1, that is, $H_1^T JH_2 = 0$. We now apply a Gram-Schmidt type process to the columns of $H_2 = [h_{n+1},\ \cdots,\ h_m]$ which is define by

$$\tilde{h}_i = h_i - \sum_{k=n+1}^{i-1} (h_k^T Jh_i)\tilde{h}_k,$$

for $n+1 \le i \le m$. Then, we set $H_2 = [\tilde{h}_{n+1}\ \cdots\ \tilde{h}_m]$ and we have

$$H^T JH = \begin{bmatrix} \tilde{J}_1 & 0 \\ 0 & \tilde{J}_2 \end{bmatrix},$$

where $\tilde{J}_2$ is diagonal. By Sylvester's law of inertia, none of the diagonal entries of $\tilde{J}_2$ can be zeros. Thus, we can normalize the columns of H_2 such that $\tilde{J}_2 \in \mathrm{diag}_{q_2}^m(\pm 1)$ with $q = q_1 + q_2$. Thus A has an HR factorization.

The converse is similar to the one in the proof of Theorem 4.6. □

Theorem 4.8 *Let $A \in \mathbb{R}^{n\times n}$, with $k = \mathrm{rank}(A) < n$ and assume that the first k columns of A are linearly independent. Write $A = [A_1,\ A_2]$, $A_1 \in \mathbb{R}^{n\times k}$, and assume that A_1 has an HR factorization with respect to J. Then, A has an HR factorization.*

Proof. We are given that $A_1 = H\tilde{R}$, where $\tilde{R}$ is upper trapezoidal and H is $(J_1, J_2)-$orthogonal for some $J_1, J_2 \in \text{diag}_q^n$. We have that range$(A_2) \subset$ range(A_1) and since A_1 has full rank, there exists a unique $P \in \mathbb{R}^{k\times(n-k)}$ such that

$$A_2 = A_1 P.$$

Thus, we define $R = [\tilde{R}\ ,\ \tilde{R}P]$ and we have $HR = H[\tilde{R}\ ,\ \tilde{R}P] = A$. □

Corollary 4.9 *Let $A \in \mathbb{R}^{n\times n}$, with $k_1 = \text{rank}(A) < n$, $J \in \text{diag}_q^n(\pm 1)$ and let k_2 be the rank of $A^T JA$. If $k_2 < k_1$, A does not have an HR factorization.*

Proof. The proof is a consequence of Theorem 4.8. □

The theorems and the corollary given above deal with HR factorization of real matrices. These results are needed in Chapter 6 for the implementation of the HZ algorithm. In the rest of this section, we focus on computing perturbation bounds for the HR factorization. The following theorem is a trivial extension of Theorem 4.6 to complex matrices.

Theorem 4.10 *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $J \in \text{diag}_n^k(\pm 1)$. There exist $H, R \in \mathbb{C}^{n\times n}$ such that $H^*JH \in \text{diag}_k^n(\pm 1)$, R is upper triangular and $A = HR$ if and only if all principal minors of A^*JA are nonzero.*

For rectangular matrix $A \in \mathbb{C}^{m\times n}$, the HR factorization with respect to a signature matrix $J \in \text{diag}_m^k(\pm 1)$ is defined by

$$A = HR, \qquad R \in \triangle(\mathbb{C}),\ H \in \mathcal{U}_{mn}(J, \widetilde{J}),$$

where $\widetilde{J} \in \text{diag}_n^q(\pm 1)$.

In the rest of this section, we use the following theorem to define the HR factorization for rectangular complex matrices. It enables us to investigate the perturbation of the HR factorization.

Theorem 4.11 *Let $A \in \mathbb{C}^{m\times n}$, with $m \geq n$, $\operatorname{rank}(A) = n$ and let $J \in \operatorname{diag}_m^k(\pm 1)$. A has an HR factorization with respect to J if and only if all principal minors of A^*JA are nonzero.*

Proof. Assume that all the principal minors of A^*JA are nonzero. Then, like in Theorem 4.6, A^*JA can be factorized as

$$A^*JA = L|D|^{\frac{1}{2}}\tilde{J}_1|D|^{\frac{1}{2}}L^*,$$

where $L \in \mathbb{C}^{n\times n}$ is unit lower triangular, $D \in \mathbb{R}^{n\times n}$ is nonsingular diagonal and $\tilde{J} \in \operatorname{diag}_n^q(\pm 1)$ for some integer q. Let $R = |D|^{\frac{1}{2}}L^*$ and define $H = AR^{-1} \in \mathbb{C}^{m\times n}$. We have that $H^*JH = \tilde{J}$. The converse is obtained like in Theorem 4.6. □

4.5.1 Perturbation of the HR Factorization

Now that the definition of the HR factorization is given, our aim is to derive perturbation bounds for the H factor and the R factor when A is subject to some perturbation ΔA. In this section, we first generalize the results on the perturbation bounds for the HR factorization of square matrices in [5] to complex rectangular matrices and then we extend the results concerning the QR factorizations in [17, 65, 45] to the HR factorization of complex rectangular matrices. We also compute the condition number of the HR factorization.

Let $\mathcal{V}_h \subset \mathbb{R}^p$ with $p = \frac{n^2-n}{2}$ and according to Section 4.4 $H = \phi(h)$. Following the general method developed in Section 4.4, we define

$$\begin{aligned} f : \mathbb{C}^{m\times n} \times \triangle(\mathbb{C}) \times \mathcal{V}_h &\rightarrow \mathbb{C}^{m\times n}, \\ (\widetilde{A}, \widetilde{R}, \tilde{h}) &\mapsto \widetilde{H}\widetilde{R} - \widetilde{A}, \end{aligned}$$

where from (1.13) $\widetilde{H} = \phi(\tilde{h})$. We have that $\varphi(H, R) = HR$. We get

$$d_2 f(A, h, R)(\Delta h, \Delta R) = \Delta H R + H\Delta R, \tag{4.17}$$

$$H^* J d_2 f(A,h,R)(\Delta h, \Delta R) R^{-1} \;=\; H^* J \Delta H + \widetilde{J} \Delta R R^{-1}, \qquad (4.18)$$

where $\Delta H = d\phi(h)\Delta h$. Note that $H^* J \Delta H \in \mathbf{SkewH}$ and $\widetilde{J}\Delta R R^{-1} \in \triangle(\mathbb{C})$. We define the two projectors Π_1 and Π_2 by

$$\begin{aligned} &\Pi_1 : \mathbb{C}^{n\times n} \to \triangle(\mathbb{C}), \qquad \Pi_1 = \Pi_d + \Pi_u + \Pi_l^*, \\ &\Pi_2 : \mathbb{C}^{n\times n} \to \mathbf{SkewH}, \quad \Pi_2 = \Pi_l - \Pi_l^*. \end{aligned}$$

Note that $X = (\Pi_1 + \Pi_2)X$ and $\mathrm{range}(\Pi_1) \cap \mathrm{range}(\Pi_2) = \emptyset$. Hence

$$\mathbb{C}^{n\times n} = \triangle(\mathbb{C}) \oplus \mathbf{SkewH}.$$

We have $\|\Pi_2(X)\|_F^2 = 2\|\Pi_l(X)\|_F^2$, thus, since Π_l is an orthogonal projection $\|\Pi_2\|_2 = \sqrt{2}$. It is straightforward to show that $\|\Pi_1\|_2 \le \sqrt{2}$. This bound is attained by $X = \frac{\sqrt{2}}{2}(e_i e_j^T + e_j e_i^T)$. Thus, from (4.17) and (4.17) and using (4.13) we get

$$dg_R(A)\Delta A \;=\; \Pi_1(H^* J \Delta A R^{-1}) R.$$

If $m = n$ then

$$dg_H(A)\Delta A \;=\; H\widetilde{J}\Pi_2(H^* J \Delta A R^{-1}).$$

If $m > n$, then there exits $G = [H\ H_0] \in \mathbb{C}^{m\times m}$ such that $G^* J G \in \mathrm{diag}_m^k(\pm 1)$ for some integer k. Note that G and k are obtain by a Gram-Schmidt type process. Thus,

$$\begin{aligned} dg_H(A)\Delta A &= JG^{-*} \begin{bmatrix} \Pi_2(H^* J \Delta A R^{-1}) \\ H_0^* J \Delta A R^{-1} \end{bmatrix}, \\ \|dg_H(A)\Delta A\|_F &= \|G\|_2 ((\|\Pi_2\|_2^2 \|H\|_2^2 + \|H_0\|_2^2)\|R^{-1}\|_2^2 \|\Delta A\|_F^2)^{\frac{1}{2}}, \\ \|dg_H(A)\|_2 &\le \sqrt{3}\kappa_2(G)\|R^{-1}\|_2. \end{aligned}$$

Finally, we obtain the bounds

$$\|dg_R(A)\|_2 \;\le\; \sqrt{2}\kappa_2(R)\|H\|_2, \qquad (4.19)$$

$$\|dg_H(A)\|_2 \;\le\; \sqrt{2}\kappa_2(G)\|R^{-1}\|_2. \qquad (4.20)$$

Following (4.15) and (4.16), we obtain the following theorem.

Theorem 4.12 *Let $A = HR$, $H \in \mathcal{U}_{mn}(J, \widetilde{J})$ be the HR factorization of A and for $\Delta A \in \mathbb{C}^{n\times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $A + \Delta A = \widetilde{H}\widetilde{R}$ be the HR factorization of $A + \Delta A$. Then*

$$\begin{aligned} \|\widetilde{R} - R\|_F &\leq \sqrt{2}\kappa_2(R)\|H\|_2\epsilon + O(\epsilon^2), & (4.21)\\ \|\widetilde{H} - H\|_F &\leq \sqrt{2}\kappa_2(H)\|R^{-1}\|_2\epsilon + O(\epsilon^2). & (4.22) \end{aligned}$$

Theorem 4.12 generalizes the result in [5] to complex rectangular matrices and also extends the results concerning the QR factorizations in [17, 65, 45] to the HR factorization of complex rectangular matrices. The bounds are similar to those obtained in [5, 45].

If we apply our result to the particular case of the QR factorization, then we get the well-known bounds

$$\begin{aligned} \|\widetilde{R} - R\|_2 &\leq \sqrt{2}\kappa_2(A)\epsilon + O(\epsilon^2), & (4.23)\\ \|\widetilde{H} - H\|_2 &\leq \sqrt{2}\|A^{-1}\|_2\epsilon + O(\epsilon^2). & (4.24) \end{aligned}$$

(4.19) and (4.20) give a bound on the condition number of the HR factorization. The exact condition number can be obtained by using a Kronecker product approach. Let M_1, M_2, C and $\widehat{r}$ be defined by

$$\begin{aligned} M_1 &= (I \otimes R^*\widetilde{J}) + (R^*\widetilde{J} \otimes I)C\mathbf{T},\\ M_2 &= (I \otimes A^*J) + (A^*J \otimes I)C\mathbf{T},\\ \mathrm{vec}(A^*) &= C\mathrm{vec}(A). \end{aligned}$$

From (4.18) or by differentiating $(A, R) \mapsto R^*\widetilde{J}R - A^*JA$, we get

$$R^*\widetilde{J}\widehat{R} + \widehat{R}^*\widetilde{J}R = A^*J\Delta A + \Delta A^*JA. \quad (4.25)$$

Applying the vec operator to (4.25), we obtain

$$\begin{aligned} M_1\widehat{r} &= M_2\text{vec}(\Delta A), \\ \widehat{r} &= (M_1)^{-1}_{|\triangle(\mathbb{C})}M_2\text{vec}(\Delta), \\ \|dg_R(A)\|_2 &= \|(M_1)^{-1}_{|\triangle(\mathbb{C})}M_2\|_2, \end{aligned} \tag{4.26}$$

where $(M_1)_{|\triangle(\mathbb{C})}$ is the restriction of M_1 to $\text{vec}(\triangle(\mathbb{C}))$. Combining (4.26) with the direct sum decomposition, we obtain

$$\begin{aligned} dg_H(A)\Delta A &= \Delta A R^{-1} - H(dg_R(A)\Delta A)R^{-1}, \\ \|dg_H(A)\|_2 &= \|R^{-T}\otimes I - (M_1)^{-1}_{|\triangle(\mathbb{C})}M_2\|_2, \end{aligned} \tag{4.27}$$

Using (4.26) and (4.27), we have the following theorem.

Theorem 4.13 *Let $A = HR$, $H \in \mathcal{U}_{mn}(J, \widetilde{J})$ be the HR factorization of A and for $\Delta A \in \mathbb{C}^{n\times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $A + \Delta A = \widetilde{H}\widetilde{R}$ be the HR factorization of $A + \Delta A$. Then, the sharpest perturbation bounds to first order are given by*

$$\|\widetilde{R} - R\|_F \leq \|(M_1)^{-1}_{|\triangle(\mathbb{C})}M_2\|_2\epsilon + O(\epsilon^2), \tag{4.28}$$

$$\|\widetilde{H} - H\|_F \leq \|R^{-T}\otimes I - (M_1)^{-1}_{|\triangle(\mathbb{C})}M_2\|_2\epsilon + O(\epsilon^2). \tag{4.29}$$

Theorem 4.12 is a generalization of the HR and QR factorization perturbation bounds that can be found in the literature. Although Theorem 4.12 and 4.13 are similar, the bounds in Theorem 4.13 are the best possible. In Table 4.3, we compare the bounds that are stated in these two theorems.

4.5.2 Numerical Experiments

The sensitivity of the HR factorization of A with respect to a signature matrix $J \neq \pm I$ is closely related to the minors of A^*JA. If one of the minors of A^*JA

vanishes or is close to zero, then R is ill conditioned which implies that H is also ill conditioned or does not exist (if R is singular). To illustrate this fact numerically, we construct a sequence of matrices A_ϵ such that their first column $a_\epsilon = A_\epsilon(:,1)$ is almost isotropic, that is, $a_\epsilon^T J a_\epsilon \to 0$ as $\epsilon \to 0$. We denote $\delta_\epsilon = \|A_{\epsilon_0} - A_\epsilon\|_F$ and $A_{\epsilon_0} = H_{\epsilon_0} R_{\epsilon_0}$. The results are in Table 4.2. In the second column of Table 4.2, the values of δ_ϵ are relatively small. We see that the values of $\|R_\epsilon - R_{\epsilon_0}\|_F$ in the third column and the values of $\|H_\epsilon - H_{\epsilon_0}\|_F$ in the fourth column do not depend on δ_ϵ. They depend instead on the values of $a_\epsilon^T J a_\epsilon$, in the sense that the bounds in the third and the sixth column get more accurate when $a_\epsilon^T J a_\epsilon$ increases and in the meantime the value δ_ϵ increases slowly. It confirms the fact that the sensitivity of the HR factorization depends on the minors of $A^* JA$. Note that the errors in R, in the third column (respectively H in the fifth column) are very close to the expected value in the fourth column (respectively the sixth column). This is due to the fact that we use the condition number. In the next numerical experiment, with the QR factorization, we see that if the bound is not sharp, then the expected values do not reflect the errors that are obtained.

Table 4.2: Perturbation bounds of the HR factorization.

$a_\epsilon^T J a_\epsilon$	δ_ϵ	$\|R_\epsilon - R\|_F$	$\|dg_R(A)\|_2 \delta_\epsilon$	$\|H_\epsilon - H\|_F$	$\|dg_H(A)\|_2 \delta_\epsilon$
$-7e-8$	$2e-15$	$1.77e-4$	$5.98e-4$	$6.3e-4$	$2.37e-4$
$-5e-6$	$2e-14$	$2.14e-6$	$2.97e-6$	$5.21e-7$	$2.02e-6$
$-2e-4$	$2e-13$	$1.34e-7$	$2e-7$	$3.78e-8$	$1.51e-7$
$-7e-3$	$2e-12$	$4.67e-8$	$6.61e-8$	$1.88e-8$	$1.36e-7$

For the QR factorization, we compare numerically (4.19) and (4.26). We consider the following 2×2 example

$$A_\epsilon = \begin{bmatrix} 1-\epsilon & 1 \\ 1 & 1+\epsilon \end{bmatrix}, \qquad \kappa_2(A_\epsilon) = \epsilon^{-2}(1+\sqrt{1+\epsilon^2})^2.$$

Let $Q_\epsilon R_\epsilon = A_\epsilon$ be the QR factorization of A_ϵ and let $\Delta A_\epsilon = A_0 - A_\epsilon$. We have that $\|\Delta A_\epsilon\|_F = |\epsilon|\sqrt{2}$. The numerical results are in Table 4.3. Note that the expected values in the second column, computed with our condition number, are just twice the error on the R factor. Note that $\|A_0\|_F = 2$. Thus, if we use relative errors our bounds are the same as the computed values. The expected values obtained with the usual bound are quite poor since the bound given by (4.23) is very poor. These results suggest that in this example the QR factorization of A_ϵ is a well conditioned problem independent of the condition number of the matrix that is factorized.

Table 4.3: Values of $\|dg_R(A)\|_2\|\Delta A_\epsilon\|_F$ and $\sqrt{2}\kappa_2(A_\epsilon)\|\Delta A_\epsilon\|_F$ as $\epsilon \to 0$.

ϵ	$\|R_\epsilon - R\|_F$	$\|dg_R(A)\|_2\epsilon$	$\sqrt{2}\kappa_2(A_\epsilon)\epsilon$
10^{-1}	$1.001e-1$	$2.107e-1$	$8e1$
10^{-2}	$1e-2$	$2.01e-2$	$8e2$
10^{-3}	$1e-3$	$2e-3$	$8e3$
10^{-4}	$1e-4$	$2e-4$	$8e4$
10^{-5}	$1e-5$	$2e-5$	$8e5$
10^{-6}	$1e-6$	$2e-6$	$8e6$

4.6 The Indefinite Polar Factorization

We say that $A \in \mathbb{R}^{n\times n}$ admits a polar factorization if $A = HS$ with H orthogonal and S symmetric definite positive. The indefinite polar factorization (IPF) is a generalization of the usual polar factorization, that is, we want to generalize the polar decomposition with H $(J, \widetilde{J})$-orthogonal.

$$A = HS,\ H^T J H = J,$$

The following theorem from [38] allows us to define this decomposition and it gives necessary conditions for the existence and uniqueness of the IPF.

Theorem 4.14 *If $A \in \mathbb{R}^{n\times n}$ and JA^TJA has no eigenvalues on the nonpositive real axis, then A has a unique IPF $A = HS$, where H is (J,J)-orthogonal and S is J-symmetric with eigenvalues in the open right half-plane.*

In this thesis, we define the IPF as in Theorem 4.14. Throughout this section, we assume that S is diagonalizable.

4.6.1 Perturbation of the IPF

We start by a preliminary result that will enable us to give the direct sum decomposition like in (4.9). We assume that A is nonsingular and that it admits the IPF $A = HS$, $H \in \mathcal{O}_n(J,J)$. Our aim is to derive perturbation bounds for the H factor and the factor S when A is subject to some perturbation ΔA. Using (1.13), we define

$$\begin{aligned} f : \mathbb{R}^{n\times n} \times \mathcal{V}_h \times J\mathbf{Sym}(\mathbb{R}) &\rightarrow \mathbb{R}^{n\times n}, \\ (\widetilde{A}, \widetilde{S}, \tilde{h}) &\mapsto \widetilde{H}\widetilde{S} - \widetilde{A}, \end{aligned}$$

where $\widetilde{H} = \phi(\tilde{h})$ and $H = \phi(h)$ and ϕ is defined by (1.13). Note that $f(A,h,S) = 0$. We define $d_2 f = \frac{\partial f}{\partial h} + \frac{\partial f}{\partial S}$. We have

$$\begin{aligned} d_2 f(A,h,S)(\Delta h, \Delta S) &= \Delta H S + H \Delta S, \\ H^T J d_2 f(A,h,S)(\Delta h, \Delta S) S^{-1} &= H^T J \Delta H + J \Delta S S^{-1}, \end{aligned}$$

where $\Delta H = d\phi(h)\Delta h$. Note that $H^T J \Delta H \in \mathbf{Skew}(\mathbb{R})$. In the following lemma, we establish the direct sum decomposition as in (4.9).

Lemma 4.15 *Let $J \in \mathrm{diag}_n^q(\pm 1)$ and let S be nonsingular, J-symmetric such that the eigenvalues of JS are positive. Then,*

$$\mathbb{R}^{n\times n} = \mathbf{Skew}(\mathbb{R}) \oplus \mathbf{Sym}(\mathbb{R})S^{-1}.$$

Furthermore, let Π_1 be the projector on $\mathbf{Skew}(\mathbb{R})$ and Π_2 be the projector on $\mathbf{Sym}(\mathbb{R})S^{-1}$. Then,

$$\Pi_1(Z) = \mathcal{T}_+(S^T)^{-1}(ZS - S^TZ), \tag{4.30}$$

$$\Pi_2(Z) = \widetilde{\mathcal{T}}_+(S^T)^{-1}(S^T(Z+Z^T)S)S^{-1}, \tag{4.31}$$

where $\mathcal{T}_+(S^T)$ and $\widetilde{\mathcal{T}}_+(S^T)$ are defined in Theorem 1.5.

Proof. Let $Z \in \mathbb{R}^{n\times n}$ and consider the equation $X + YS^{-1} = Z$ with $X \in \mathbf{Skew}(\mathbb{R})$ and $Y \in \mathbf{Sym}(\mathbb{R})$. We have that $-X + S^{-T}Y = Z^T$. Thus,

$$\begin{aligned} S^TX + XS &= ZS - S^TZ, \\ S^TY + YS &= S^T(Z^T + Z)S. \end{aligned}$$

We see then the solutions are given by

$$X = \mathcal{T}_+(S^T)^{-1}(ZS - S^TZ) \text{ and } Y = \widetilde{\mathcal{T}}_+(S^T)^{-1}(S^T(Z+Z^T)S)S^{-1}. \quad \square$$

To characterize g, we proceed as follows. We have $\frac{\partial f}{\partial A}(A,S,h) = -\Delta A$. We set $(\widehat{H},\widehat{S}) = (dg_H(A)\Delta A, dg_S(A)\Delta A)$. Thus,

$$\widehat{H}S + H\widehat{S} = \Delta A \text{ and } \widehat{H}^TJH + H^TJ\widehat{H} = 0.$$

Let $X = H^TJ\widehat{H} \in \mathbf{Skew}(\mathbb{R})$ and $\widetilde{\Delta A} = H^TJ\Delta A$. Thus,

$$S^TX + XS = \widetilde{\Delta A} - \widetilde{\Delta A}^T, \tag{4.32}$$

$$S^TJ\widehat{S} + \widehat{S}^TJS = S^T\widetilde{\Delta A} + \widetilde{\Delta A}^TS. \tag{4.33}$$

Thus, we obtain

$$\begin{aligned} dg_H(A)\Delta A &= HJ\widetilde{\mathcal{T}}_{S^T}^{-1}(H^TJ\Delta A - \Delta A^TJH), \\ dg_S(A)\Delta A &= J\mathcal{T}_{S^T}^{-1}(S^T)(A^TJ\Delta A + \Delta A^TJA). \end{aligned}$$

Let $S^T = VDV^{-1}$ be the eigendecomposition of S^T. We define

$$\begin{aligned} M_1 &= (V \otimes V^T)\mathrm{diag}(\mathrm{vec}(M))(V^{-1} \otimes V^{-T}), \\ M_2 &= -(H^TJ \otimes I)\mathbf{T} + I \otimes H^TJ, \\ \widetilde{M_2} &= (A^TJ \otimes I)\mathbf{T} + I \otimes A^TJ. \end{aligned}$$

Then, applying the vec operator, we obtain

$$\|dg_H(A)\|_2 = \|(I \otimes HJ)M_1M_2\|_2, \tag{4.34}$$

$$\|dg_S(A)\|_2 = \|M_1\widetilde{M_2}\|_2. \tag{4.35}$$

Using (4.34)-(4.35), we have the following theorem.

Theorem 4.16 *Let $A = HS$, $H \in \mathcal{O}_n(J)$ be the IPF of A and for $\Delta A \in \mathbb{R}^{n\times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $(A+\Delta A) = \widetilde{H}\widetilde{S}$ be the IPF of $A+\Delta A$. Then,*

$$\begin{aligned} \|\widetilde{S} - S\|_F &\leq \|M_1\widetilde{M_2}\|_2\epsilon + O(\epsilon^2), \\ \|\widetilde{H} - H\|_F &\leq \|M_1M_2\|_2\epsilon + O(\epsilon^2), \end{aligned}$$

where M_1, M_2 and $\widetilde{M_2}$ are the matrices involved in the differential of the implicit function in (4.34) *and* (4.35)

The above theorem gives the perturbation expansion of the IPF for a nonsingular A. If A is singular and 0 is at most a simple eigenvalue of A then it is possible to give the perturbation bounds of the factor S. We just need to apply the implicit function theorem to $(A, S) \mapsto S^TJS - A^TJA$. Also, from (4.34)-(4.35), we can

give bounds of the condition number that less expensive to compute that the exact condition numbers:

$$\begin{aligned} \|dg_H(A)\|_F &\leq 2m\kappa_2(V)^2\kappa_2(H), \\ \|dg_S(A)\|_F &\leq 2m\kappa_2(V)^2\|A\|_2, \end{aligned}$$

where $m = \max_{ij} |m_{ij}^+|$ and $M = (m_{ij}^+)$ is defined by (1.20).

4.6.2 The Polar Factorization

The polar factorization is the particular case that is obtained when $J = \pm I$. Thus, $A = QS$ is the polar factorization of A, with Q orthogonal and S symmetric. Note that if A is complex then the perturbation bounds remain the same for the unitary Q factor and the Hermitian factor S. In [5], a perturbation bound for the Hermitian factor that involves the 2-norm of A is given but in [34] and [35], the author found a constant bound $\sqrt{2}$. With our method, we obtain the condition number for the Hermitian factor and for the unitary factor in a simpler way than [16]. We proceed as follows.

Lemma 4.17 *Let the two matrix operators $\mathcal{T}_1$ and $\mathcal{T}_2$ be defined by $\mathcal{T}_1 X = (X - X^T) \circ M$ and $\mathcal{T}_2 X = (DX + X^T D) \circ M$ where M is defined in* (1.20) *and D real diagonal matrix with positive entries. Then*

$$\|\mathcal{T}_1\|_2 = \frac{2}{\lambda_{n-1} + \lambda_n}, \tag{4.36}$$

$$\|\mathcal{T}_2\|_2 = \sqrt{2}\frac{\sqrt{\lambda_n^2 + \lambda_1^2}}{\lambda_n + \lambda_1}, \tag{4.37}$$

where λ_{n-1} and λ_n are the two smallest diagonal entries of D and λ_1 the largest diagonal entry of D.

Proof. Let $X = (x_{ij}) \in \mathbb{R}^{n\times n}$ and assume that $Y = \mathcal{T}_1(X)$ with $Y = (y_{ij})$. We have

$$\begin{aligned}\|Y\|_F^2 &= \sum_{i,j=1}^{n} \frac{(x_{ij}-x_{ji})^2}{(\lambda_i+\lambda_j)^2} \leq 4\sum_{i,j=1}^{n} \frac{x_{ij}^2+x_{ji}^2}{(\lambda_i+\lambda_j)^2},\\ \|Y\|_F &\leq \frac{2}{\lambda_{n-1}+\lambda_n}\|X\|_F.\end{aligned}$$

The bound in (4.36) is attained by $E = \frac{1}{\sqrt{2}}(e_n e_{n-1}^T - e_{n-1}e_n^T)$ where e_k is the k-th column of the identity matrix.

We now focus on (4.37). Assume that $Y = \mathcal{T}_1(X)$ with $Y = (y_{ij})$. We have that $y_{ij} = \frac{1}{\lambda_i+\lambda_j}(\lambda_i x_{ij} + \lambda_j x_{ji})$, $y_{ii} = x_{ii}$. We define

$$\mu = \max_{i,j}\left(\frac{\lambda_i^2+\lambda_j^2}{(\lambda_i+\lambda_j)^2}\right).$$

and we have that $y_{ij}^2 \leq \mu(x_{ij}^2 + x_{ji}^2)$. Thus,

$$\begin{aligned}\|Y\|_F^2 &= \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1} 2y_{ij}^2 \leq 2\max_{i,j}\left(\frac{\lambda_i^2+\lambda_j^2}{(\lambda_i+\lambda_j)^2}\right)\|X\|_F^2,\\ \|\mathcal{T}_2\|_2 &\leq \sqrt{2\mu}.\end{aligned}$$

Let

$$E = \frac{1}{\sqrt{\lambda_p^2+\lambda_q^2}}\left(\lambda_p e_p e_q^T + \lambda_q e_q e_p^T\right)$$

with (p,q) the indices where μ is attained. Note that $\|E\|_F = 1$ and $\|\mathcal{T}_2(E)\|_F = 1$. Without loss of generality, assume that $\lambda_p \leq \lambda_q$ and define $t = \frac{\lambda_p}{\lambda_q}$, with $0 \leq t \leq 1$. We have that $\mu = \frac{1+t^2}{(1+t)^2}$. It is straightforward to see that $\tilde{\mu} : t \mapsto \frac{1+t^2}{(1+t)^2}$ is monotone and decreasing for $0 \leq t \leq 1$. Thus, $\tilde{\mu}$ attains its maximum for $t = 0$. Thus, $(p,q) = (n,1)$. □

Note that if A is nonsingular $\lambda_1 = \|A\|_2$ and $\lambda_n = 0$, thus $\|\mathcal{T}_2\|_2 = \sqrt{2}$. Otherwise if A is nonsingular $\lambda_n = \|A^{-1}\|_2^{-1}$ and we obtain

$$\|\mathcal{T}_2\|_2 = \sqrt{2}\frac{\sqrt{\|A^{-1}\|_2^{-2}+\|A\|_2^2}}{\|A^{-1}1\|_2^{-1}+\|A\|_2} = \sqrt{2}\frac{\sqrt{1+\kappa_2(A)^2}}{1+\kappa_2(A)}, \tag{4.38}$$

We consider (4.32)-(4.33), with H orthogonal, S symmetric and $S = V^T D V$ the eigendecomposition of S. Let $Z_1 = V\widetilde{\Delta A}V^T$ and $Z_2 = V^T\widetilde{\Delta A}V$. Then, (4.32)-(4.33) become

$$D\widetilde{X} + \widetilde{X}D = Z_1 - Z_1^T \text{ and } DY + YD = DZ_2 + Z_2^T D,$$

where $\widetilde{X} = VXV^T$ and $Y = V\hat{S}V^T$. Since $\|Z_1\|_F = \|Z_2\|_F = \|\Delta A\|_F$, applying Lemma 4.17 and using (4.38) , we obtain

$$\|dg_H(A)\|_2 = \frac{2}{\lambda_{n-1} + \lambda_n} \text{ and } \|dg_S(A)\|_2 = \sqrt{2}\frac{\sqrt{1 + \kappa_2(A)^2}}{1 + \kappa_2(A)}. \tag{4.39}$$

Note that $1 \leq \|dg_S(A)\|_2 \leq \sqrt{2}$. Both of these bounds are attained. If S is of the type $S = \lambda I$ or S is orthogonal, then $\|dg_S(A)\|_2 = 1$ and if A is singular then $\|dg_S(A)\|_2 = \sqrt{2}$. We have the following theorem.

Theorem 4.18 *Let $A = HS$, $H \in \mathcal{O}_n(I)$ be the polar factorization of A and for $\Delta A \in \mathbb{R}^{n\times n}$ such that $\epsilon = \|\Delta A\|_F$ is small enough, let $(A + \Delta A) = \widetilde{H}\widetilde{S}$ be the polar factorization of $A + \Delta A$. Then,*

$$\begin{aligned} \|\widetilde{H} - H\|_F &\leq \frac{2}{\lambda_{n-1} + \lambda_n}\epsilon + O(\epsilon^2), \\ \|\widetilde{S} - S\|_F &\leq \alpha\epsilon + O(\epsilon^2), \end{aligned}$$

where $\alpha = \sqrt{2}$ if A is singular or $\alpha = \sqrt{2}\dfrac{\sqrt{1 + \kappa_2(A)^2}}{1 + \kappa_2(A)}$ otherwise.

The bounds given in the above theorem are the sharpest possible to first order. Using the classical definition of condition number for the Hermitian factor, the same condition number as in (4.39) is obtained in [16]. Our method has the advantage of giving a shorter proof than [16] of several pages. Our method allows us also to compute explicitly the Fréchet derivative of the factors. In [51], the condition number in (4.39) for the orthogonal factor is given.

4.6.3 Numerical Experiments

To compute the indefinite polar factorization and the usual polar factorization, we used the iteration described in [38, Thm 5.2]. We recall that the iteration for the J-orthogonal factor is given by

$$H_0 \;=\; A, \qquad H_{n+1} = \frac{1}{2}(H_n + JH_n^{-T}J).$$

This iteration is guaranteed to converge if JA^TJA has no eigenvalue with a negative real part. We present two series of numerical tests. The first ones are quite standard, their purpose being to illustrate the perturbation bounds given in Theorem 4.16. We generated a matrix A_0 using the function `randn` of MATLAB. Then, we build a sequence of matrices A_ϵ that converges to A_0 as ϵ tends to zero. We denote $\delta_\epsilon = \|A_0 - A_\epsilon\|_F$ and $A_0 = H_0S_0$, $A_\epsilon = H_\epsilon S_\epsilon$ the indefinite polar factorization of A_0 and A_ϵ. J was obtained by

```
J =(-1).^randperm(n)
```

using MATLAB. We shifted all these matrices so that $JA_\epsilon^TJA_\epsilon$ has all its eigenvalues in the open right half-plane. The results are displayed in Table 4.4. We see that our perturbation bounds follow closely the computed values which confirms that in this case the bounds obtained by Theorem 4.16 are sharp.

We denote by c_H and c_S the bounds of the condition number of the hyperbolic and symmetric factors given by (4.36)-(4.36). Table 4.5 shows the first order perturbation bounds obtained by using c_H and c_S. The bounds obtained by using $c_s$and c_H in the first 4 rows in Table in 4.5 are accurate. In the last row, we see that the bound for the J-symmetric matrix is weak whereas the bound for the hyperbolic factor is more reliable. We conclude that the bounds c_S and c_H given by (4.36)-(4.36) should be used carefully when the norm of the perturbation is small.

Table 4.4: Perturbation bounds of the indefinite polar factorization.

δ_ϵ	$\|S_\epsilon - S_0\|_F$	$\|dg_{S_0}(A_0)\|_2\delta_\epsilon$	$\|H_\epsilon - H_0\|_F$	$\|dg_{H_0}(A_0)\|_2\delta_\epsilon$
$1e-15$	$1e-15$	$1e-14$	$1e-15$	$2e-15$
$1e-9$	$1e-9$	$2e-9$	$2e-8$	$5e-8$
$1e-5$	$3e-5$	$9e-5$	$2e-5$	$6e-5$
$1e-3$	$7e-3$	$1.6e-2$	$5e-3$	$2e-2$
$1e-2$	$1e-2$	$2.3-2$	$2e-2$	$3.4e-2$

Table 4.5: Perturbation bounds of the IPF using bounds for the condition numbers c_H and c_S.

δ_ϵ	$c_S\delta_\epsilon$	$c_H\delta A_\epsilon$
$1e-15$	$3.7e-13$	$7.5e-14$
$1e-9$	$3.7e-7$	$7.5e-8$
$1e-5$	$3.7e-3$	$7.5e-4$
$1e-3$	$3.7e-1$	$7.5e-2$
$1e-2$	3.7	$7.5e-1$

The aim of the second numerical experiment series is to give an example where the bounds given by (4.36)-(4.36)) are very poor approximations of the exact condition numbers. The test matrices are Hilbert matrices, built in MATLAB and they can be called by the function `hilb`. The (i,j) element of a Hilbert matrix is given by $1/(i+j-1)$. These Hilbert matrices are symmetric and very ill conditioned. The signature matrix $J \in \mathrm{diag}_n^k(\pm 1)$ is given by

$$J = \mathrm{diag}(-I_{\lfloor n/2 \rfloor}, I_{\lfloor n/2 \rfloor}).$$

The logarithm of the condition number $\log_{10}(\|dg_S(A)\|_2)$ for the J-symmetric factor is represented by $\star$ and by $+$ for $\log_{10}(\|dg_H(A)\|_2)$, the logarithm of the condition number of the hyperbolic factor. The logarithm of the bound denoted by c_S in (4.36) is represented by $\square$ and by $\circ$ the bound c_H in (4.36). We see in Figure 4.1, in all the test matrices the exact condition number is very small

compare to c_S, the biggest ratio being of order 10^{18}. For the hyperbolic factor, the difference is less, the biggest ratio being of order 10^4.

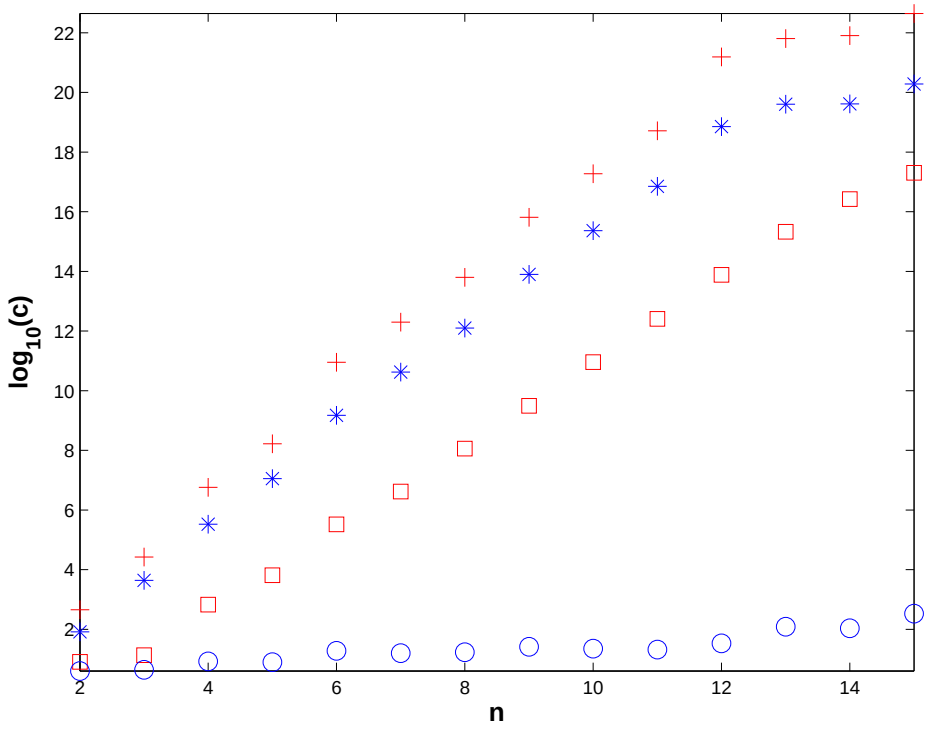

Figure 4.1: Condition number and perturbation bounds of the IPF of Hilbert matrices with $\log_{10}(\|dg_S(A)\|_2)$ ($\bigcirc$), $\log_{10}(\|dg_H(A)\|_2)$ ($\square$), $\log_{10}(c_S)$ ($*$) and $\log_{10}(c_H)$ ($+$).

4.7 The Hyperbolic Singular Value Decomposition

Let $A \in \mathbb{R}^{m\times n}$ with $m \geq n$. We say that A admits a hyperbolic singular value decomposition (HSVD) if

$$A = QDH^T$$

with D diagonal, Q orthogonal and $H \in \mathcal{O}_n(J, \widetilde{J})$. The hyperbolic singular value decomposition (HSVD) and the indefinite least square problem were analyzed in [10], [12] and [55]. $\mathcal{E}_D$ denotes the set of real diagonal matrices. The following theorem establishes the existence of the HSVD. The theorem and the proof are similar to those in [12, Sec. 2].

Theorem 4.19 *Let $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ be a full rank matrix, $J \in \operatorname{diag}_n^k(\pm 1)$ and assume that $\operatorname{rank}(AJA^T) = n$. Then, there exists a positive nonsingular diagonal matrix $D \in \mathbb{R}^{m\times n}$, Q orthogonal, $\widetilde{J} \in \operatorname{diag}_n^k(\pm 1)$ and $H \in \mathcal{U}_n(J, \widetilde{J})$ such that*

$$A = QDH^T.$$

Proof. Let $AJA^T = QSQ^T$ be an eigendecomposition. Assume that AJA^T is nonsingular. We define $D = |S|^{\frac{1}{2}}$ and $\widetilde{J} = \operatorname{sign}(S)$. Let

$$H = A^T Q \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix}.$$

We have

$$H^T J H = \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix}^T Q^T A J A^T Q \begin{bmatrix} D^{-1} \\ 0 \end{bmatrix} = \widetilde{J}. \quad \square$$

In the definition of the HSVD, we see that only the n first columns of Q are necessary to define the decomposition. Thus, in the rest of this section, we consider that the HSVD of $A \in \mathbb{R}^{m\times n}$ is given by

$$A = QDH^T,\ Q \in \mathcal{O}_{mn}(I),\ H \in \mathcal{O}_n(J, \widetilde{J}),\ J, \widetilde{J} \in \operatorname{diag}_n^k(\pm 1).$$

4.7.1 Perturbation of the HSVD

The linear subspace of $n \times n$ real diagonal matrices is identified with $\mathbb{R}^n$ and it is denoted by $\mathcal{E}_D$. Let

$$\begin{aligned} f : \mathbb{R}^{m\times n} \times \mathcal{E}_D \times \mathcal{V}_q \times \mathcal{V}_h &\rightarrow \mathbb{R}^{n\times n}, \\ (\widetilde{A}, \widetilde{D}, \widetilde{q}, \tilde{h}) &\mapsto \widetilde{Q}\widetilde{D}\widetilde{H}^T - \widetilde{A}, \end{aligned}$$

with $\widetilde{Q} = \phi_1(\widetilde{q})$ and $\widetilde{H} = \phi_2(\tilde{h})$, where ϕ_1 and ϕ_2 are defined by (1.13). Note that $f(A, D, q, h) = 0$. We define $d_2 f = \frac{\partial f}{\partial D} + \frac{\partial f}{\partial q} + \frac{\partial f}{\partial h}$, $\Delta Q = d\phi_1(q)\Delta q$ and $\Delta H = d\phi_2(h)\Delta h$. We have

$$\begin{aligned} d_2 f(A, D, q, h)(\Delta A, \Delta Q, \Delta H) &= \Delta Q D H^T + Q D \Delta H^T + Q \Delta D H^T, \\ Q^T d_2 f(A, D, q, h)(\Delta A, \Delta Q, \Delta H) H^{-1} &= Q^T \Delta Q D + D \Delta H^T H^{-T} + \Delta D, \end{aligned}$$

with $Q^T \Delta Q$ and $\Delta H^T H^{-T} \widetilde{J}$ skew-symmetric. The following lemma establishes the direct sum decomposition (4.9).

Lemma 4.20 *Let $D = \mathrm{diag}(\lambda_k)$. If the diagonal elements of $\widetilde{J}D^2$ are distinct, then we have the following direct sum decomposition*

$$\mathbb{R}^{n\times n} = \mathcal{E}_D \oplus \mathbf{Skew}(\mathbb{R})D \oplus D\mathbf{Skew}(\mathbb{R})\widetilde{J}.$$

The corresponding projector Π_1 on $\mathcal{E}_D$ is just Π_d whereas for all $Z \in \mathbb{R}^{n\times n}$ the projector on $\mathbf{SkewH}D$ is $\Pi_2 D$ and the projector on $D\mathbf{SkewH}\widetilde{J}$ is $D\Pi_3\widetilde{J}$ where

$$\Pi_2(Z) = (\widetilde{J}ZD + DZ^T\widetilde{J}) \circ \Lambda, \quad \Pi_3(Z) = (DZ + Z^T D) \circ \Lambda, \tag{4.40}$$

$$\Lambda = (\mu_{ij}), \quad \mu_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \dfrac{1}{\widetilde{\sigma}_j\lambda_i^2 - \widetilde{\sigma}_i\lambda_j^2} & \text{otherwise,} \end{cases} \tag{4.41}$$

where $D = \mathrm{diag}(\lambda_i)$ and $\widetilde{J} = \mathrm{diag}(\tilde{\sigma}_i)$ Moreover, the norms of the operators, $\|\Pi_2\|_2$ and $\|\Pi_3\|_2$ are given by

$$\|\Pi_2\|_2 = \|\Pi_3\|_2 = \sqrt{2}\max_{i\neq j} \frac{\sqrt{\lambda_i^2 + \lambda_j^2}}{\left|\widetilde{\sigma}_i\lambda_j^2 - \widetilde{\sigma}_j\lambda_i^2\right|}. \tag{4.42}$$

Proof. Let $Z \in \mathbb{R}^{m\times n}$, $Z = (z_{ij})$ and assume that $\Delta D + XD + DY\widetilde{J} = Z$ where $\Delta D \in \mathcal{E}_D$ and $X, Y \in \mathbf{Skew}(\mathbb{R})$. Since X and Y are skew symmetric, we have $\Pi_d(XD + DY) = 0$. Thus,

$$\Delta D = \Pi_d(Z).$$

By computing the elements of $(\Pi_l + \Pi_u)(XD - DY\widetilde{J})$, we get $\frac{n^2-n}{2}$ 2×2 linear systems

$$E_{ij} \begin{bmatrix} x_{ij} & y_{ij} \end{bmatrix}^T = \begin{bmatrix} z_{ij} & z_{ji} \end{bmatrix}^T,$$

where E_{ij} is defined by

$$E_{ij} = \begin{bmatrix} \lambda_j & \widetilde{\sigma}_i\lambda_i \\ -\lambda_i & -\widetilde{\sigma}_j\lambda_j \end{bmatrix}.$$

We have $\det E_{ij} = \widetilde{\sigma}_i\widetilde{\sigma}_j(\widetilde{\sigma}_j\lambda_j^2 - \widetilde{\sigma}_j\lambda_i^2) \neq 0$. Hence, for $i \neq j$, we obtain the $\frac{n^2-n}{2}$ solutions

$$x_{ij} = -\frac{\sigma_i\lambda_j z_{ij} + \widetilde{\sigma}_j\lambda_i z_{ji}}{\widetilde{\sigma}_j\lambda_i^2 - \widetilde{\sigma}_i\lambda_j^2}, \tag{4.43}$$

$$y_{ij} = \frac{\lambda_i z_{ij} + \lambda_j z_{ji}}{\widetilde{\sigma}_j\lambda_i^2 - \widetilde{\sigma}_i\lambda_j^2}. \tag{4.44}$$

With (4.43) and (4.44), we obtain

$$X = -(\widetilde{J}ZD + DZ^T\widetilde{J}) \circ \Lambda, \tag{4.45}$$

$$Y = (DZ + Z^TD) \circ \Lambda, \tag{4.46}$$

where $\Lambda = (\mu_{ij})$ is given by (4.41). Finally, we obtain $\Pi_1 = \Pi_d$, $\Pi_2(Z) = X$ and $\Pi_3(Z) = Y$.

Using the Cauchy-Schwarz inequality, we have that

$$\begin{aligned} x_{ij}^2 &\leq \frac{\lambda_i^2 + \lambda_j^2}{\widetilde{\sigma}_i\lambda_j^2 - \widetilde{\sigma}_j\lambda_i^2}(z_{ij}^2 + z_{ji}^2), \\ \|X\|_F^2 &\leq 2\max_{ij, i\neq j} \frac{\lambda_i^2 + \lambda_j^2}{\left|\widetilde{\sigma}_i\lambda_j^2 - \widetilde{\sigma}_j\lambda_i^2\right|}\|Z\|_F^2. \end{aligned}$$

The bound (4.33) is attained by

$$E = \frac{\sigma_p \lambda_q e_{pq} + \sigma_q \lambda_p e_{qp}}{\sqrt{\lambda_p^2 + \lambda_q^2}},$$

with $\|E\|_F = 1$ and where (p, q) are the indices where

$$\max_{ij, i \neq j} \frac{\sqrt{\lambda_i^2 + \lambda_j^2}}{\left| \widetilde{\sigma}_i \lambda_j^2 - \widetilde{\sigma}_j \lambda_i^2 \right|}$$

is attained. Similarly, using the same method, we show the second part of (4.33) and that the bound is attained by

$$\widetilde{E} = \frac{\lambda_p e_{pq} + \lambda_q e_{qp}}{\sqrt{\lambda_p^2 + \lambda_q^2}},$$

with $\|\widetilde{E}\|_F = 1$. □

To characterize g, we proceed as in (4.11) and (4.13-4.14). We have $\frac{\partial f}{\partial A} \Delta A = -\Delta A$. We set $(\widehat{D}, \widehat{Q}, \widehat{H}) = (dg_D(A)\Delta A, dg_Q(A)\Delta A, dg_H(A)\Delta A)$. We obtain the linear system

$$\begin{aligned}
\widehat{Q} D H^T + Q \widehat{D} H^T + Q D \widehat{H}^T &= \Delta A, \qquad (4.47)\\
Q^T \widehat{Q} D + D \widehat{H}^T J H \widetilde{J} + \widehat{D} &= Q^T \Delta A J H \widetilde{J},\\
Q^T \widehat{Q} + \widehat{Q}^T Q &= 0,\\
\widehat{H}^T J H + H^T J \widehat{H} &= 0.
\end{aligned}$$

Thus, by Lemma 4.20,

$$dg_D(A)\Delta A = \Pi_d(Q^T \Delta A J H \widetilde{J}), \qquad (4.48)$$

$$dg_H(A)\Delta A = \widetilde{J} H^T J \Pi_3(Q^T \Delta A J H \widetilde{J}). \qquad (4.49)$$

If $m = n$, then

$$dg_Q(A)\Delta A = Q \Pi_2(Q^T \Delta A J H \widetilde{J}). \qquad (4.50)$$

If $m > n$, then we know that there exist $G = [Q, Q_0] \in \mathbb{R}^{n\times n}$ such that $G^T G = I$. G is obtained as in Section 4.5 by the classical Gram-Schmidt process. Using (4.47), we have

$$dg_Q(A)\Delta A = G \begin{bmatrix} \Pi_2(Q^T \Delta A J H \widetilde{J}) \\ Q_0^T \Delta A H^{-T} D^{-1} \end{bmatrix}. \tag{4.51}$$

Let $\tilde{h}_k$ denote the k-th column $H\widetilde{J}$. We have

$$\begin{aligned} \|dg_D(A)\Delta A\|_2 &= \sup_{\|\Delta A\|_F=1} \|\Pi_d(Q^T \Delta A J H \widetilde{J})\|_F, \\ &= \sup_{\|\Delta A\|_F=1} \|\Pi_d(\Delta A H \widetilde{J})\|_F, \\ &= \|W\|_2, \end{aligned}$$

where $W \in \mathbb{R}^{n\times n^2}$ has its k-th row defined by $\tilde{h}_k^T \otimes e_k^T$. Thus,

$$\|dg_D(A)\Delta A\|_2 = \|W\|_2 = \max_k \|\tilde{h}_k\|_2 = \max_k \|H(k,:)\|_2. \tag{4.52}$$

We define

$$M_1 = \widetilde{J} H^T J \otimes Q^T, \tag{4.53}$$

$$M_2 = D \otimes \widetilde{J} + (\widetilde{J} \otimes D)\mathbf{T}, \quad \widetilde{M_2} = I \otimes D + (D \otimes I)\mathbf{T}. \tag{4.54}$$

Applying the vec operator to (4.49) and taking norms, we obtain

$$\|dg_H(A)\|_2 = \|(I \otimes H^T J)\mathrm{diag}(\mathrm{vec}(\Lambda))\widetilde{M_2} M_1\|_2. \tag{4.55}$$

Similarly, for Q factor, we obtain from (4.51)

$$\|dg_Q(A)\|_2 = \begin{cases} \|\mathrm{diag}(\mathrm{vec}(\Lambda))M_2 M_1\|_2, & \text{if } m = n, \\ \\ \|\mathrm{diag}(\mathrm{vec}(\Lambda))M_2 M_1 + (D^{-1}H^{-1}) \otimes I_n\|_2, & \text{if } m > n. \end{cases} \tag{4.56}$$

We are now able to give the first order expansion of the three factors of the HSVD.

Theorem 4.21 *Let $A = QDH^T$, $H \in \mathcal{O}_n(J, \widetilde{J})$ be the HSVD of A and for $\Delta A \in \mathbb{R}^{n\times n}$ such that $\epsilon = \|\Delta A\|_F$ is small, let $(A + \Delta A) = \widetilde{Q}\widetilde{D}\widetilde{H}^T$ be the HSVD of $A + \Delta A$. Then, using (4.56-4.55) and (4.52-4.61)*

$$\begin{aligned} \|D - \widetilde{D}\|_F &\leq \max_k \|H(k,:)\|_2 \epsilon + O(\epsilon^2), && (4.57)\\ \|Q - \widetilde{Q}\|_F &\leq \|dg_Q(A)\|_2 \epsilon + O(\epsilon^2), && (4.58)\\ \|H - \widetilde{H}\|_F &\leq \|dg_H(A)\|_2 \epsilon + O(\epsilon^2), && (4.59) \end{aligned}$$

where $\|dg_Q(A)$ and $\|dg_H(A)\|_2$ are given by (4.55) *and* (4.56)*. These bounds are the sharpest possible to first order.*

Using (4.50) and (4.49), note that the condition number of the HSVD can be bounded by

$$\begin{aligned} \|dg_Q(A)\|_2 &\leq \begin{cases} \dfrac{2}{m}\|D\|_2\|H\|_2, & \text{if } m = n, \\ \left(\dfrac{4}{m^2}\|D\|_2^2\|H\|_2^2 + \|H^{-T}D^{-1}\|_2^2\right)^{\frac{1}{2}}, & \text{if } m > n, \end{cases} && (4.60)\\ \|dg_H(A)\|_2 &\leq \frac{2}{m}\|D\|_2\kappa_2(H), && (4.61) \end{aligned}$$

where $m = \min_{ij} |\widetilde{\sigma}_i\lambda_i^2 - \widetilde{\sigma}_j\lambda_j^2| = \|\mathrm{diag}(\mathrm{vec}(\Lambda))\|_2$. These bounds are less sharp than (4.56) and (4.55) but they are easily computable. We also can give better bounds than (4.60)-(4.61), using (4.42),

$$\begin{aligned} \|dg_Q(A)\|_2 &\leq \begin{cases} \alpha\|H\|_2, & \text{if } m = n, \\ \left(\alpha^2\|H\|_2^2 + \|H\widetilde{J}D^{-1}\|_2^2\right)^{\frac{1}{2}}, & \text{if } m > n, \end{cases} && (4.62)\\ \|dg_H(A)\|_2 &\leq \alpha\kappa_2(H), && (4.63) \end{aligned}$$

where $\alpha = \|\Pi_2\|_2 = \|\Pi_3\|_2$ is defined in (4.42).

For the usual SVD, H is orthogonal. We get the well-known result (see for example [66]) for the singular values

$$\|dg_D(A)\|_2 \;=\; 1,\ \|D-\widetilde{D}\|_F \le \epsilon + O(\epsilon^2).$$

The condition numbers of Q and H can be easily computed since H is also orthogonal

$$\|dg_Q(A)\|_2 \;=\; \begin{cases} \alpha, & \text{if } m=n, \\ \left(\alpha^2+\dfrac{1}{\lambda_n^2}\right)^{\frac{1}{2}}, & \text{if } m>n, \end{cases} \tag{4.64}$$

$$\|dg_H(A)\|_2 \;=\; \alpha, \tag{4.65}$$

where $\alpha = \|\Pi_2\|_2 = \|\Pi_3\|_2$ is defined in (4.42) and λ_n is the smallest singular value of A. In [42] and [64], a bound for the singular vectors is proposed. This bound is obtained by applying the fact that H is orthogonal in (4.60) and (4.61).

4.7.2 Numerical Experiments

We consider a 3×3 example with

$$D_0 = \mathrm{diag}(10, 9.9, 1) \ \text{ and } A_0 = U_0 D_0 V_0^T,$$

where (U_0, V_0) is a randomly generated orthogonal-hyperbolic matrix pair and the signature matrices $(J, \widetilde{J})$ such that $V_0^T J V = \widetilde{J}$ are defined by

$$\begin{aligned} J &= \mathrm{diag}(-1,-1,1), \\ \widetilde{J} &= \mathrm{diag}(-1,1,-1). \end{aligned}$$

We construct a sequence of matrices $A_\epsilon = U_0 D_\epsilon V_0^T$ with $D_\epsilon = D_0 + \epsilon e_2 e_2^T$ where e_2 denotes the second column of the identity. The results are in Table 4.6 for the

singular values and in Table 4.7 for the orthogonal and hyperbolic factor, with $A_\epsilon = U_\epsilon D_\epsilon V_\epsilon^T$ be the HSVD of A_ϵ, $\delta_\epsilon = \|A_0 - A_\epsilon\|_F$. We see that the expected bound for the hyperbolic singular values are very close to the computed values. It is due to the fact that the bound on the hyperbolic singular value depends only on the norm of the hyperbolic factor

$$V_\epsilon = A_\epsilon^T U_\epsilon D_\epsilon^{-1},$$

with $\kappa_2(D_\epsilon) = 10$. The orthogonal and hyperbolic factors are more sensitive to the fact that one of the hyperbolic singular values is becoming double which does not appear easily in Theorem 4.21. But, we see in the expressions of dg_Q and dg_H in (4.49) and (4.50)–(4.51) that the sensitivity of the orthogonal and hyperbolic factors depend on Π_2 and Π_3 in (4.40). Moreover, the norms of these projectors (4.42) vary proportionally to the inverse of $\min_{ij} ||\lambda_j| - |\lambda_i||$ which explains the numerical test in Table 4.7. The bounds in the last row of Table 4.7 (column 3 and 5) are quite poor. The first explanation is the fact that the value of $\delta_\epsilon = 10^{-5}$ is big, the corresponding perturbation is not in the required neighborhood $\mathcal{V}_A$ (see Section 4.4) in order to apply the implicit function theorem. Consequently, the result on the condition number and the perturbation expansion in Theorem 4.21 are not valid. Another fact that we need to keep in mind is that the perturbation expansion given in Theorem 4.21 gives a bound for the predicted result but it does not guarantee any accuracy of these bounds.

In Figure 4.2, we plot the logarithms of the exact condition number of the orthogonal and hyperbolic factor, the bounds given by (4.60), (4.61), (4.62) and (4.63), against the value of ϵ. The exact condition number for the orthogonal factor $\|dg_Q(A)\|_2$ and the condition number for hyperbolic factor $\|dg_H(A)\|_2$ are represented by $\circ$ and $\square$. We denote by $c_{Q,1}$ and $c_{H,1}$ the bounds of the condition numbers given by (4.60), (4.61) and we denote by $c_{Q,2}$ and $c_{H,2}$ the bounds defined

by (4.62) and (4.63). In Figure 4.2, the symbols + and $\lhd$ represent the logarithm of the bounds given by $c_{Q,1}$ and $c_{H,1}$ whereas the symbols $\star$ and $\rhd$ are for the logarithms of $c_{Q,2}$ and $c_{H,2}$. These values are labeled by $\log_{10}(c)$ on the y-axes. We see that the condition number and the bounds are of the same order and seem to be the same asymptotically.

Table 4.6: Perturbation bounds for the singular values from HSVD.

δ_ϵ	$\|D_0 - D_\epsilon\|_F$	$\|dg_D(A_0)\|_2\delta_\epsilon$
10^{-13}	9.10^{-14}	10^{-13}
10^{-10}	2.10^{-10}	3.10^{-10}
10^{-6}	9.10^{-7}	1.10^{-6}
3.10^{-5}	3.10^{-5}	3.10^{-5}

Table 4.7: Perturbation bounds for the orthogonal and hyperbolic factors.

δ_ϵ	$\|Q_\epsilon - Q_0\|_F$	$\|dg_Q(A_0)\|_2\delta_\epsilon$	$\|H_\epsilon - H_0\|_F$	$\|dg_H(A_0)\|_2\delta_\epsilon$
10^{-13}	10^{-11}	10^{-9}	10^{-11}	$1.4.10^{-9}$
2.10^{-10}	10^{-12}	10^{-6}	10^{-12}	10^{-6}
10^{-6}	4.10^{-12}	10^{-1}	4.10^{-12}	1.10^{-1}
10^{-5}	10^{-12}	4	10^{-12}	4

The behaviour of the usual SVD and HSVD can be quite different and unexpected. For $n = 2$, if the two singular values are close then the condition number of the singular vector is large since the condition number for the orthogonal factors, (4.64)–(4.65) is unbounded. In the hyperbolic case, with $J = \mathrm{diag}(1, -1)$, the condition number for the orthogonal factor and the hyperbolic factor is uniformly bounded on any subset of $\mathbb{R}^{2\times 2} \setminus \{0\}$ at a positive distance of the zero matrix.

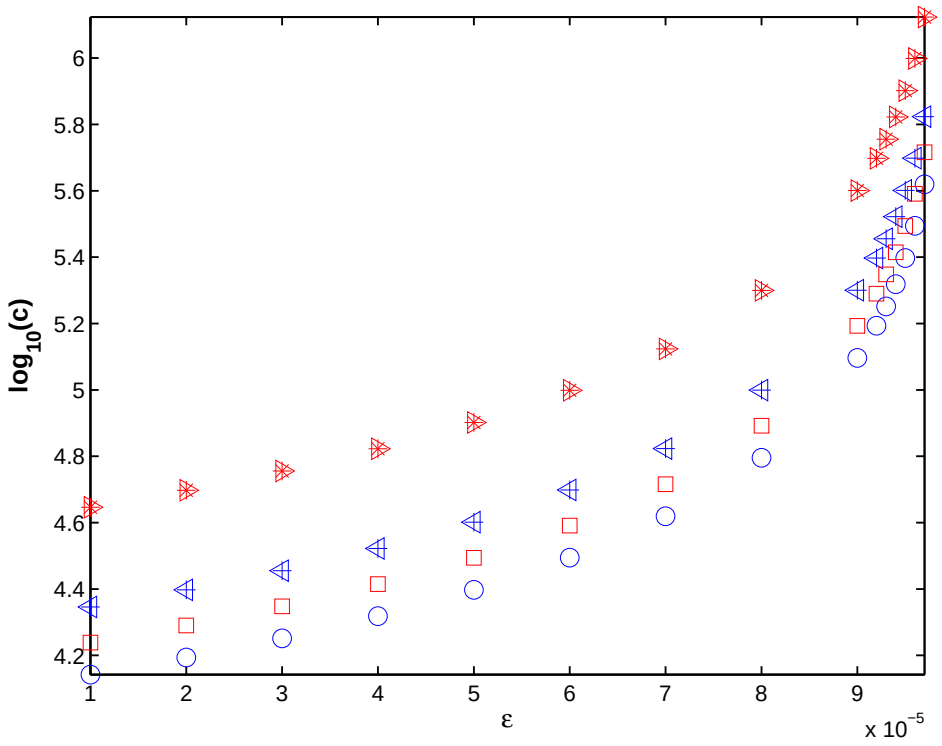

Figure 4.2: Comparison between the condition number and its bounds with $\log_{10}(\|dg_Q(A)\|_2)$ ($\bigcirc$), $\log_{10}(\|dg_H(A)\|_2)$ ($\square$), $\log_{10}(c_{Q,1})$ ($+$), $\log_{10}(c_{H,1})$ ($\triangleleft$), $\log_{10}(c_{Q,2})$ ($*$) and $\log_{10}(c_{H,2})$ ($\triangleright$).

4.8 Sensitivity of Hyperbolic Eigendecompositions

In this section, we consider a pair $(S, \widetilde{J})$ with $S \in \mathbb{R}^{n\times n}$ symmetric and $\widetilde{J} \in \mathrm{diag}_n^k(\pm 1)$. Throughout this section, we assume that the eigenvalues of $\widetilde{J}S$ are simple. We say that S is diagonalizable with respect to $\widetilde{J}$ if there exists $H \in \mathcal{O}_{mn}(J, \widetilde{J}, \mathbb{C})$ with $J \in \mathrm{diag}_n(\mathbb{U})$ such that

$$S = H^T D H, \tag{4.66}$$

$$\widetilde{J} = H^T J H. \tag{4.67}$$

Note that

$$\widetilde{J}^{-1}S = H^{-1}(J^{-1}D)H,$$

that is, $\widetilde{J}^{-1}S$ and $J^{-1}D$ are similar. If $\widetilde{J} = \pm I$, then this process is the usual diagonalization of a symmetric matrix, with H orthogonal and the existence of this decomposition is well known and understood. For $\widetilde{J} \neq \pm I$, we need to justify the existence of (4.66)–(4.67). When the eigenvalues of $\widetilde{J}^{-1}S$ are simple, Lemma 2 in [70] tells us that there exits a nonsingular matrix $X \in \mathbb{C}^{n\times n}$ such that

$$X^*(S, \widetilde{J})X = (\widetilde{D}, \Sigma),$$

where $\widetilde{D}, \Sigma$ are block diagonal matrices with 1×1 blocks corresponding to real eigenvalues and 2×2 blocks corresponding to complex conjugate eigenvalues. The k-th 2×2 diagonal block $(\widetilde{D}_k, \Sigma_k)$ of $(\widetilde{D}, \Sigma)$ has the following form

$$\widetilde{D}_k = \begin{bmatrix} 0 & \widetilde{d}_k \\ \overline{\widetilde{d}_k} & 0 \end{bmatrix}, \qquad \Sigma_k = \begin{bmatrix} 0 & \beta_k \\ \overline{\beta_k} & 0 \end{bmatrix}.$$

Let $(x_k, \overline{x_k})$ be the $k-$th and $(k+1)$-th columns of X. We have

$$x_k^*(S, \widetilde{J})x_k = (0, 0), \quad x_k^*(S, \widetilde{J})\overline{x_k} = (\widetilde{d}_k, \beta_k),$$

or equivalently by taking conjugate,

$$x_k^T(S,\widetilde{J})\overline{x_k} \;=\; (0,0), \quad x_k^T(S,\widetilde{J})x_k \;=\; (\overline{\widetilde{d}_k},\overline{\beta_k}).$$

Hence, $X^T\widetilde{J}X = \mathrm{diag}(\overline{\beta_k})$. Let $J \in \mathrm{diag}_n(\mathbb{U})$ be such that $\Sigma = |\Sigma|J$ and define

$$H = |\Sigma|^{-1/2}X^{-1} \qquad \text{and} \qquad D = \overline{\mathrm{diag}(\widetilde{d}_k)}|\Sigma|.$$

Then,

$$H^TDH = X^{-T}|\Sigma|^{-1/2}D|\Sigma|^{-1/2}X^{-1} = X^{-T}\overline{\mathrm{diag}(\widetilde{d}_k)}X^{-1} = S.$$

Also $H^TJH = \widetilde{J}$. Therefore, H satisfies (4.66)–(4.67).

Note that the the derivation is still valid if the eigenvalues of $\widetilde{J}^{-1}S$ are semi-simple. In Chapter 6, we present a method, the HZ algorithm, to compute $Y \in \mathbb{R}^{n\times n}$ such that $(S,\widetilde{J}) = Y^T(D_0,J_0)Y$, with $D_0 \in \mathbb{R}^{n\times n}$ block diagonal (with 1×1 and 2×2 blocks) and $J_0 \in \mathrm{diag}_n^k(\pm 1)$. The decomposition defined by (4.66)–(4.67) is obtained by splitting the 2×2 blocks with normalized eigenvector matrix V such that $V^TJ_0V \in \mathrm{diag}_n(\mathbb{U})$.

4.8.1 Perturbation Analysis of the Diagonalization by Hyperbolic Matrices

The sensitivity of the eigendecomposition of the standard eigenvalue problem is analyzed in [2] where no structure in the perturbations is taken into account. Their analysis depends on the pseudospectra of the matrix and the distance to the nearest defective matrix. The HZ algorithm described in Chapter 6 computes the eigendecomposition (4.66)–(4.67). For this reason, we concentrate in this section on decomposition of this type only. We also analyze some cases where supplementary structures are imposed to S in (4.66).

We consider the decomposition (4.66) and (4.67). Let n_1 be the number of real eigenvalues of $(S, \widetilde{J})$ and $2n_2$ the number of complex eigenvalues of S. Without loss of generality, we can assume that

$$D \;=\; \mathrm{diag}(D_1, D_2, \overline{D}_2), \quad J = \mathrm{diag}(J_1, J_2, \overline{J}_2), \quad H^T = [\, V_1 \quad V_2 \quad \overline{V}_2 \,],$$

where $D_1 \in \mathbb{R}^{n_1 \times n_1}$, $D_2 \in \mathbb{C}^{n_2 \times n_2}$, $J_1 \in \mathrm{diag}^k_{n_1}(\pm 1)$, $J_2 \in \mathrm{diag}_{n_2}(\mathbb{U})$, $V_1 \in \mathbb{R}^{n \times n_1}$ and $V_2 \in \mathbb{C}^{n \times n_2}$. D_1 is the diagonal matrix that contains the real eigenvalues of $(S, \widetilde{J})$ that corresponds to the eigenvectors in $\widetilde{J}V_1J_1$ whereas D_2 is diagonal and contains the complex eigenvalues of $(S, \widetilde{J})$ that corresponds to the eigenvectors in $\widetilde{J}V_2J_2$. Let $\mathcal{A} \subset \mathbb{C}^{n \times n}$ be the linear space of matrices that are partitioned as in (4.68), that is,

$$\mathcal{A} = \left\{ A \in \mathbb{C}^{n \times n} : \; A^T = [\, A_1 \quad A_2 \quad \overline{A_2} \,], \; A_1 \in \mathbb{R}^{n \times n_1}, A_2 \in \mathbb{C}^{n \times n_2} \right\}.$$

We define

$$\begin{aligned} \mathcal{E}_H &= \mathcal{O}_n(J, \widetilde{J}, \mathbb{C}) \cap \mathcal{A}, \\ \mathcal{E}_D &= \mathrm{diag}(\mathbb{C}^n) \cap \mathcal{A}. \end{aligned} \tag{4.68}$$

We recall that $\mathcal{E}_H$ is an $\frac{n^2-n}{2}$ dimensional differentiable manifold. Moreover, the function ϕ defined in Section 1.5 has the following property

$$H^T J d\phi(h) \in \mathbf{Skew}(\mathbb{C}).$$

Lemma 4.22 *If the eigenvalues of $\widetilde{J}^{-1}S$ are simple, then*

$$\mathbf{Sym}(\mathbb{R}) = \mathrm{range}(\mathcal{T}_-(S\widetilde{J})) \oplus H^T \mathcal{E}_D H,$$

where $\mathcal{T}_-$ is defined in Theorem 1.5 *and $\mathcal{E}_D$ is defined in* (4.68)*. Moreover, the projector $\Pi_{\mathcal{E}_D}$ on $\mathcal{E}_D$ is given by $\Pi_{\mathcal{E}_D}(X) = \Pi_d(H^{-T} X H^{-1})$.*

Proof. The proof is a consequence of Theorem 1.5. □

We define

$$\begin{aligned} f : \mathbf{Sym}(\mathbb{R}) \times \mathcal{E}_D \times \mathcal{E}_H &\rightarrow \mathbf{Sym}(\mathbb{R}), \\ (\widetilde{S}, \widetilde{D}, \tilde{h}) &\mapsto \widetilde{H}^T \widetilde{D} \widetilde{H} - \widetilde{S}, \end{aligned}$$

where $\widetilde{H} = \phi(\tilde{h})$. Note that $f(A, D, h) = 0$, with $H = \phi(h)$.

Lemma 4.23 *$d_2 f(S, D, h)$ is nonsingular if the eigenvalues of $\widetilde{J}^{-1} S$ are distinct.*

Proof. Let $\Delta D \in \mathcal{E}_D$ and $\Delta h \in \mathbb{R}^{\frac{n^2-n}{2}}$ be such that $d_2 f(S, D, h)(\Delta D, \Delta h) = 0$. Set $\Delta H = d\phi(h)\Delta h$ and $X = H^T J \Delta H \in \mathbf{Skew}(\mathbb{C})$. Thus,

$$\begin{aligned} d_2 f(S, D, h)(\Delta D, \Delta h) = 0 &\Leftrightarrow \Delta H^T D H + H^T D \Delta H + H^T \Delta D H = 0, \\ &\Leftrightarrow \Delta H^T J H S \widetilde{J} + S \widetilde{J} H^T J \Delta H + H^T \Delta D H = 0, \\ &\Leftrightarrow S \widetilde{J} X - X \widetilde{J} S + H^T \Delta D H = 0. \end{aligned}$$

Since DJ is diagonal, Lemma 4.22 implies $\Delta D = 0$ and $\Delta H = X = 0$. □

When the eigenvalues of $\widetilde{J} S$ are simple, $d_2 f(S, D, h)$ is nonsingular and the implicit function theorem shows there exists a differentiable function $g = (g_D, g_H)$ and an open neighborhood $\mathcal{V}_S$ of S satisfying

$$\begin{aligned} g : \mathcal{V}_S &\rightarrow \mathcal{V}_D \times \mathcal{V}_H, \\ \widetilde{S} &\mapsto (g_D(\widetilde{S}), g_H(\widetilde{S})), \end{aligned}$$

where $\mathcal{V}_D \times \mathcal{V}_H$ is an open neighborhood of (D, H) with $H = \phi(h)$. Moreover, g satisfies $g_D(S) = D,\ g_H(S) = H$ and

$$\forall \widetilde{S} \in \mathcal{V}_S, \qquad f(\widetilde{S}, g_D(\widetilde{S}), g_H(\widetilde{S})) = (0, 0),$$

that is, $\widetilde{S} = g_H(\widetilde{S})^T g_D(\widetilde{D}) g_H(\widetilde{S}) = \widetilde{H}^T \widetilde{D} \widetilde{H}$ is the diagonalization of $\widetilde{S}$ with respect to $\widetilde{J}$. To characterize g, we proceed as follows. Let $d_1 f(S, D, h) =$

$\frac{\partial f}{\partial S}(S, D, h)$. The differential of g at A is then given by

$$dg(S)\Delta S = -(d_2 f(S, D, h))^{-1} d_1 f(S, D, h)\Delta A.$$

We have $d_1 f(S, D, h)\Delta S = -\Delta S$. We set $(\widehat{D}, \widehat{H}) = (dg_D(S)\Delta S, dg_H(S)\Delta S)$ and thus

$$d_2 f(S, D, h)(\widehat{D}, \widehat{H}) = \Delta S.$$

Hence,

$$\widehat{H}^T DH + H^T D\widehat{H} + H^T \widehat{D} H \;\; = \;\; \Delta S. \tag{4.69}$$

Let $X = H^T J \widehat{H} \in \mathbf{Skew}(\mathbb{C})$ and $Y = H^{-T} X H^{-1}$. Then, (4.69) becomes

$$\Delta S \widetilde{J} \;\; = \;\; S\widetilde{J}X - X\widetilde{J}S + \widehat{D}, \tag{4.70}$$

$$H^{-T}\Delta S H^{-1} \;\; = \;\; DJY - YDJ + \widehat{D}. \tag{4.71}$$

Using Theorem 1.5, we obtain the condition operators

$$dg_D(S)\Delta S \;\; = \;\; \widehat{D} = \Pi_d(H^{-T}\Delta S H^{-1}), \tag{4.72}$$

$$dg_H(S)\Delta S \;\; = \;\; \widehat{H} = JH^{-T}\mathcal{T}_-^{-1}(S\widetilde{J})(\Delta S). \tag{4.73}$$

One can show that

$$\|dg_D(S)\|_2 \;\; = \;\; \|W\|_2,$$

where $W \in \mathbb{C}^{n \times n^2}$ has its k-th row defined by $u_k \otimes u_k$ with $u_k = \sigma_k H(k, :)\widetilde{J}$. Note that $\|W\|_2 \leq \|H\|_2^2$. Thus, we obtain the following theorem that gives the first order perturbation bound of D and H.

Theorem 4.24 *Let $\widetilde{S} = S + \Delta S$ be such that $\epsilon = \|\Delta S\|_F$ is small enough. Let $\widetilde{S} = \widetilde{H}^T \widetilde{D} \widetilde{H}$ be the diagonalization of $\widetilde{S}$ with respect to $\widetilde{J}$, where $\widetilde{H} \in \mathcal{O}_n(J, \widetilde{J}, \mathbb{C})$. Then,*

$$\|D - \widetilde{D}\|_F \;\; \leq \;\; \|W\|_2^2 \epsilon + O(\epsilon^2), \tag{4.74}$$

$$\|H - \widetilde{H}\|_F \;\; \leq \;\; \gamma\epsilon + O(\epsilon^2), \tag{4.75}$$

where $\gamma = \|(I \otimes JH^{-T})(H^T \otimes H^T \mathrm{diag}(\mathrm{vec}(M_-))H^{-T} \otimes H^{-T}\|_2$ *and* M_- *is given by* (1.20).

The bounds in the above theorem are the sharpest possible to first order. Since $\|W\|_2 \leq \|H\|_2^2$, the condition numbers of all the eigenvalues of $(S, \widetilde{J})$ are bounded by $\kappa_2(H) = \|H\|_2^2$. The expression for γ in Theorem 4.24 is long and complicated. Thus, we give a bound for γ that is easy and less expensive to compute. The first bound can be obtain by tacking norms

$$\gamma \leq \frac{\kappa_2(H)^2 \|H\|_2}{\min\limits_{i \neq j} |\lambda_i \sigma_i - \lambda_j \sigma_j|}.$$

But, using (4.71), we can obtain a better bound. Let $Y = (y_{ij})$ and let $Z = (z_{ij}) = H^{-T} \Delta S H^{-1} - \widehat{D}$. We have

$$\begin{aligned} |y_{ij}|^2 &\leq \frac{|z_{ij}|^2}{\min\limits_{i \neq j} |\lambda_i \sigma_i - \lambda_j \sigma_j|^2}, \\ \|Y\|_F &\leq \sqrt{2} \frac{\kappa_2(H) \|\Delta S\|_F}{\min\limits_{i \neq j} |\lambda_i \sigma_i - \lambda_j \sigma_j|}. \end{aligned}$$

Using Theorem 1.4, we get

$$\begin{aligned} \|\widehat{H}\|_F &\leq \sqrt{2} \frac{\|H\|_2 \kappa_2(H) \|\Delta S\|_F}{\min\limits_{i \neq j} |\lambda_i \sigma_i - \lambda_j \sigma_j|}, \\ \|dg_H(S)\|_2 &\leq \sqrt{2} \frac{\|H\|_2 \kappa_2(H)}{\min\limits_{i \neq j} |\lambda_i \sigma_i - \lambda_j \sigma_j|}. \end{aligned}$$

Similar results can be found in [18, Thm 4.2.1] for the standard eigenvalue problem.

We apply our results to the standard symmetric eigenvalue problem. We consider (4.66) and (4.67) with $\widetilde{J} = \pm I$. Then, H is orthogonal. Equation (4.72) and (4.73) become

$$\|dg_D(S)\|_2 = 1, \tag{4.76}$$

$$\|dg_H(S)\|_2 \;=\; \frac{\sqrt{2}}{\min\limits_{i,j,i\neq j}|\lambda_i-\lambda_j|}.$$

A consequence of (4.76) is that $g_D(S+B(0,\epsilon))\subset D+B(0,\epsilon)$ for ϵ sufficiently small.

4.8.2 Condition Number Theorems

The matrix factorization that we analyzed in this section is the eigendecomposition of a symmetric-diagonal pair when the eigenvectors are normalized such that they form a $(J,\widetilde{J})-$orthogonal matrix. In this paragraph, we use the result on the sensitivity of the eigendecomposition (4.66)–(4.67) to deduce structured eigenvalue and eigenvector condition numbers. We specialize to the case of a simple eigenvalue of $(S,\widetilde{J})$ and we compute real structured condition numbers.

Consider the decomposition (4.66)-(4.67). Let $\lambda=\lambda_k$ be an eigenvalue of $(S,\widetilde{J})$, the $k-$th diagonal element of D. The corresponding eigenvector is given by

$$x=H^{-1}e_k=\sigma_k\widetilde{J}H^Te_k=\sigma_k\widetilde{J}H(k,:)^T.$$

From (4.73), we have that the condition number for the eigenvector is given by

$$\begin{aligned}
c(x) &= \|e_k^T dg_H(S)e_k\|_2 \\
&= \sup_{\Delta S\in\mathbf{Sym}(\mathbb{R}),\|\Delta S\|_F=1}\|Y(k,:)H\|_2, \\
&= \sup_{\Delta S\in\mathbf{Sym}(\mathbb{R}),\|\Delta S\|_F=1}\|x^T\Delta SH^{-1}\mathrm{diag}(M_-(k,:))H\|_2,
\end{aligned}$$

where Y satisfies (4.71) and M_- is defined by (1.20). Let P and δ be defined by

$$\begin{aligned}
P &= H^{-1}\mathrm{diag}(M_-(k,:))H, \\
\delta &= \max(\mathrm{diag}(M_-(k,:))) = \max_{j,j\neq k}\frac{1}{|\sigma_k\lambda_k-\sigma_j\lambda_j|}.
\end{aligned}$$

Thus, we have that

$$\delta\leq c(x)\leq\|x\|_2\|P\|_2\leq\delta\|x\|_2\kappa_2(H).$$

Note that the above inequalities hold for the unstructured eigenvector condition number for the problem $Ax = \lambda x$ [18]. The condition number for the eigenvalue is given by

$$\begin{aligned} c(\lambda) &= \|e_k^T dg_D(S) e_k\|_2 = \sup_{\Delta S \in \mathbf{Sym}(\mathbb{R}), \|\Delta S\|_F = 1} |H(:,k)^T \Delta S H(:,k)|, \\ c(\lambda) &= \sup_{\Delta S \in \mathbf{Sym}(\mathbb{R}), \|\Delta S\|_F = 1} |x^T \Delta S x|, \end{aligned} \tag{4.77}$$

where $dg_D(S)$ is given by (4.72).

Theorem 4.25 *Let λ be a simple eigenvalue of the symmetric diagonal pair $(S, \widetilde{J})$ and let x be the corresponding right eigenvector normalized so that $|x^T \widetilde{J} x| = 1$. Then, the structured condition number $c(\lambda)$ in (4.77) is given by*

$$c(\lambda) = \|x\|_2^2.$$

Proof. We have that

$$|x^T \Delta S x| = |(x^T \otimes x^T) \mathrm{vec}(\Delta S)|.$$

Let $\Delta s \in \mathbb{R}^{\frac{n^2+n}{2}}$ such that $\mathrm{vec}(\Delta S) = T_{\mathbf{Sym}}(\mathbb{R}) \Delta s$ where $T_{\mathbf{Sym}}(\mathbb{R}) \in \mathbb{R}^{n^2 \times \frac{n^2+n}{2}}$ is an isometric mapping. We have that

$$c(\lambda) = \|(x^T \otimes x^T) T_{\mathbf{Sym}}(\mathbb{R})\|_2$$

and

$$\begin{aligned} \mathrm{vec}(\Delta S) &= T_{\mathbf{Sym}}(\mathbb{R}) \Delta s, \\ &= \sum_{k=1}^{n} \Delta s_{ii} (e_k \otimes e_k) + \frac{1}{\sqrt{2}} \sum_{i=1}^{n-1} \sum_{j=2}^{n} \Delta s_{ij} (e_i \otimes e_j + e_j \otimes e_i). \end{aligned}$$

Thus,

$$\begin{aligned} (x^T \otimes x^T) T_{\mathbf{Sym}}(\mathbb{R}) &= \sum_{k=1}^{n} x_k^2 (e_k \otimes e_k) + \frac{1}{\sqrt{2}} \sum_{i=1}^{n-1} \sum_{j=2}^{n} x_i x_j (e_i \otimes e_j + e_j \otimes e_i), \\ \|(x^T \otimes x^T) T_{\mathbf{Sym}}(\mathbb{R})\|_2^2 &= \sum_{k=1}^{n} |x_k|^4 + \sum_{i=1}^{n-1} \sum_{j=2}^{n} |x_i x_j|^2. \end{aligned}$$

Hence, $c(\lambda) = \|(x^T \otimes x^T) T_{\mathbf{Sym}}(\mathbb{R})\|_2 = \|x\|_2^2$. □

In [43], Karow, Kressner and Tisseur use a different normalization for x, $\|x\|_2 = 1$ and they obtain a different expression for the structured condition number: $c(\lambda) = 1/|x^T J x|$.

In [15], the authors analyze the condition number of complex eigenvalues of real matrices under real perturbations. They prove inequality

$$\frac{1}{\sqrt{2}}\kappa(\lambda) \leq \kappa_{\mathbb{R}}(\lambda) \leq \kappa(\lambda),$$

where $\kappa_{\mathbb{R}}(\lambda)$ is the condition number computed for which the perturbations are forced to be real and $\kappa(\lambda)$ the usual condition number. That is, the real structured condition number is within a small factor of the standard (complex) condition number.

We now focus on eigenvalue condition numbers where supplementary structures are imposed to the perturbations. Let $\mathbb{S} \subset \mathbf{Sym}(\mathbb{R})$ be a class of sparse symmetric matrices. We recall that $\mathbb{S}$ is a linear subspace of $\mathbf{Sym}(\mathbb{R})$. Let $\Pi_{\mathbb{S}}$ be the orthogonal projection on $\mathbb{S}$ and let $m = \dim(\mathbb{S})$. We define $T_{\mathbb{S}} \in \mathbb{R}^{n^2 \times m}$, the injection from $\mathbb{R}^m$ on $\mathbb{R}^{n^2}$ such that

$$\mathrm{vec}(Z) = T_{\mathbb{S}} z \qquad \text{and} \qquad \|Z\|_F = \|z\|_2,$$

for all $Z \in \mathbb{S}$ and $z \in \mathbb{R}^m$. The structured condition number is defined by

$$c(\lambda, \mathbb{S}) = \sup_{\Delta S \in \mathbb{S}, \|\Delta S\|_F = 1} |x^T \Delta S x|. \tag{4.78}$$

For $Z = (z_{ij}) \in \mathbb{S}$ such that $\|\mathrm{vec}(Z)\|_2 = \|T_{\mathbb{S}} z\|_2$, we know that

$$\begin{aligned}
Z &= \sum_{k \in K_1} z_{kk} e_k e_k^T + \frac{1}{\sqrt{2}} \sum_{(i,j) \in K_2} z_{ij} (e_i e_j^T + e_j e_i^T), \\
\mathrm{vec}(Z) &= \sum_{k \in K_1} z_{kk} (e_k \otimes e_k) + \frac{1}{\sqrt{2}} \sum_{(i,j) \in K_2} z_{ij} (e_i \otimes e_j + e_j \otimes e_i), \qquad (4.79) \\
\mathrm{vec}(Z) &= \sum_{k \in K_1} T_{\mathbb{S}}(:,k) z + \sum_{k \in K_2} T_{\mathbb{S}}(:,k) z, \qquad (4.80)
\end{aligned}$$

with $\mathrm{vec}(Z) = T_{\mathbb{S}} z$ and the number of the elements in K_1 and K_2 being m.

Theorem 4.26 *Let λ be a simple eigenvalue of the symmetric diagonal pair $(S, \widetilde{J})$ and let x be the corresponding right eigenvector normalized so that $|x^T \widetilde{J} x| = 1$. Then, for a symmetric sparse structure $\mathbb{S}$, the structured condition number $c(\lambda, \mathbb{S})$ in (4.78) is given by*

$$c(\lambda, \mathbb{S}) = \|\Pi_{\mathbb{S}}(xx^T)\|_F,$$

where $\Pi_{\mathbb{S}}$ is the projector on $\mathbb{S}$.

Proof. We apply to (4.78) the result in (4.79)-(4.80) which gives

$$\begin{aligned}
c(\lambda, \mathbb{S}) &= \|(x^T \otimes x^T) T_{\mathbb{S}}\|_2, \\
&= \|(x^T \otimes x^T)(\sum_{k \in K_1} T_{\mathbb{S}}(:, k) + \sum_{k \in K_2} T_{\mathbb{S}}(:, k))\|_2, \\
&= \|(x^T \otimes x^T) \left(\sum_{k \in K_1} (e_k \otimes e_k) + \frac{1}{\sqrt{2}} \sum_{(i,j) \in K_2} (e_i \otimes e_j + e_j \otimes e_i) \right) \|_2, \\
&= \left(\sum_{k \in K_1} |x_k|^4 + 2 \sum_{(i,j) \in K_2} |x_i x_j|^2 \right)^{\frac{1}{\sqrt{2}}}.
\end{aligned}$$

Thus, from the last equation above, we obtain $c(\lambda, \mathbb{S}) = \|\Pi_{\mathbb{S}}(xx^T)\|_F$. □

From Theorem 4.26, we can easily deduce the condition number when $\mathbb{S}$ is the class of symmetric tridiagonal matrices. In this case the condition number is given by the Frobenius norm of the tridiagonal part of xx^T. Similar results are in [54], where the authors analyze the structured eigenvalue condition numbers with sparsity structure for the standard eigenvalue problem.

We now give an algorithm that allows us to compute structured condition numbers for any symmetric linear structure that has $m \leq \frac{n^2+n}{2}$ degree of freedom. Note that if $m = \frac{n^2+n}{2}$ then the computed condition number is the one given by Theorem 4.25. We represent a symmetric linear structure $\mathbb{S}$ with m matrices

$M_k \in \mathbb{R}^{k\times 2}$, with $k = 1{:}\, m$. The columns of M_k are the indices of the elements of the matrix $S \in \mathbb{S}$ that are equal, that is, for all (i_1, j_1) and (i_2, j_2), entries of M_k, $S(i_1, j_1) = S(i_2, j_2)$. Using MATLAB's notations, we obtain that the elements of $S(M_k(:, 1), M_k(:, 2))$ are equal.

Algorithm 4.27 *Given an eigenvector x of a symmetric-diagonal pair $(S, \widetilde{J})$ such that $|x^T Jx| = 1$, m matrices M_k defining the structure $\mathbb{S}$, this algorithm computes the structured condition number $C(\lambda, \mathcal{S})$.*

Set $C(\lambda_k, \mathbb{S}) = 0$, $t = 0$

For $k = 1 : m$

Get p the number of rows of M_k

Set $t = 0$

For q=1:p

$i = M_k(q, 1)$, $j = M_k(q, 2)$

$t = t + x_i x_j$

end

$C(\lambda_k, \mathbb{S}) = C(\lambda_k, \mathbb{S}) + |t|^2$

end

$C(\lambda_k, \mathbb{S}) = \sqrt{C(\lambda_k, \mathbb{S})}$

Chapter 5

Numerical Solutions of PEPs

5.1 Introduction

The standard approach for the numerical solution of GEPs is to reduce the matrices involved to some simpler form that reveals the eigenvalues, for instance the generalized Schur form for a pair (A, B). Unfortunately, these canonical forms do not generalize to λ-matrices of degree greater than one. This is a major complication for the numerical solution of PEPs.

There are two major divisions in numerical methods for solving PEPs.

1. The first division is into methods that tackle the problem in its original form and those that linearize it into a GEP of larger dimension and then apply GEP techniques.

2. The second division is into methods for dense, small to medium size problems and iterative methods for large scale and sparse problems.

3. Particular cases that can be directly transformed into GEPs or can be solved by alternative ways.

In the first part of this chapter, we focus on QEPs with a rank one damping matrix and we describe a method to solve it. Then, we concentrate on methods for general PEPs based on linearizations. We also present a method to access the quality of computed eigenpairs that is based on the concepts of condition number and backward error presented in Chapters 2 and 3 and we give some numerical experiments. Finally, we give an overview of three methods for solving symmetric GEPs.

5.2 QEPs with a Rank one Damping Matrix

5.2.1 Preliminaries

We consider the quadratic eigenvalue problem

$$(\lambda^2 M - \lambda\sigma uu^T + K)x = 0, \tag{5.1}$$

where M, K are $n \times n$ nonsingular symmetric matrices, $\sigma = \pm 1$ and $u \in \mathbb{R}^n$. We assume that the pair (M, K) is diagonalizable, in the sense that there exists nonsingular matrices (Q, Z) such that

$$Q(M, K)Z = (D, \widetilde{D}),$$

where D and $\widetilde{D}$ are diagonal matrices, possibly complex. We transform the problem (5.1) into an equivalent problem by multiplying by Q and Z and we get

$$(\lambda^2 D - \lambda\sigma v\widetilde{v}^* + \widetilde{D})y = 0, \tag{5.2}$$

where $v = Qu$, $\widetilde{v} = Z^*u$ and $y = Zx$. From (5.2), we have

$$(\lambda^2 D + \widetilde{D})y = \lambda\sigma \langle y, \widetilde{v}\rangle v. \tag{5.3}$$

We know that if the k^{th} component of v is zero then $Z^{-1}e_k$ is an eigenvector of the QEP associated to the eigenvalue λ solution of $\lambda^2 d_k + \tilde{d}_k = 0$.

We assume for the rest of this section that $\langle y, \tilde{v} \rangle \neq 0$, λ^2 is not an eigenvalue of $(D, \widetilde{D})$. Then, we know that $(\lambda^2 D + \widetilde{D})$ is nonsingular. Thus, Equation (5.3) becomes

$$\begin{aligned} y &= \lambda\sigma \langle y, \tilde{v} \rangle (\lambda^2 D + \widetilde{D})^{-1} v, \\ \langle y, \tilde{v} \rangle &= \lambda\sigma \langle y, \widetilde{v} \rangle \left\langle (\lambda^2 D + \widetilde{D})^{-1} v, \widetilde{v} \right\rangle, \\ \lambda\sigma \sum_{k=1}^{n} \frac{v_k \overline{\widetilde{v}}_k}{\lambda^2 d_k + \tilde{d}_k} &= 1, \end{aligned} \tag{5.4}$$

where $D = \mathrm{diag}(d_k)$, $\widetilde{D} = \mathrm{diag}(\tilde{d}_k)$, $v = (u_k)$ and $\widetilde{v} = (\widetilde{v}_k)$. Equation (5.4) is similar to the *secular equation* in the divide and conquer algorithm for the symmetric eigenvalue problem. In the symmetric case, the secular equation has several nice properties. The roots are all real and they satisfy an interlacing property with the poles. We refer to [46] and [48] for efficient solutions to the secular equation. In our case, the zeros of (5.4) are usually complex and there is no apparent link between the roots and the poles. In what follows, we first solve a particular case of (5.4) with M positive definite and K negative definite. In this case, all the eigenvalues are real. We then describe a method to solve (5.4) in the general case.

In the rest of this section, we derive global bounds for the eigenvalues of the QEP define by (5.2). These bounds allow us to localize the roots of (5.4).

Theorem 5.1 *Let (λ, x) be an eigenpair of $(\lambda^2 A + \lambda B + C)x = 0$ with $\|x\|_2 = 1$. We assume that A and C are nonsingular. Then, there exist two positive constants r_1 and r_2 such that*

$$r_1 \leq |\lambda| \leq r_2, \tag{5.5}$$

with

$$(r_1, r_2) = \begin{cases} (\delta_-, \widetilde{\delta}_+) & \text{if } \widetilde{\delta}_- \leq \delta_-, \\ (\widetilde{\delta}_-, \delta_+) & \text{otherwise,} \end{cases} \tag{5.6}$$

where

$$\delta_{\pm} = \frac{1}{2\|A\|_2}\left(-\|B\|_2 \pm \sqrt{\|B\|_2^2 + 4\frac{\|A\|_2}{\|C^{-1}\|_2}}\right), \tag{5.7}$$

$$\widetilde{\delta}_{\pm} = \frac{\|A^{-1}\|_2}{2}\left(\|B\|_2 \pm \sqrt{\|B\|_2^2 + 4\frac{\|C\|_2}{\|A^{-1}\|_2}}\right). \tag{5.8}$$

Proof. On one hand, if the quadratic matrix polynomial has a zero eigenvalue, then $r_1 = 0$, otherwise we know that there exits $r_1 > 0$ such that $|\lambda| > r_1$. On the other hand, M is nonsingular, thus there exits always an upper bound r_2 such that $|\lambda| < r_2$. Our purpose is to evaluate these bounds for the eigenvalues.

Since (λ, x) is an eigenpair of the QEP, we have $(\lambda^2 A + C)x = -\lambda B x$, which implies

$$\big||\lambda^2|\|Ax\|_2 - \|Cx\|_2\big| \leq |\lambda|\|B\|_2.$$

Thus, we obtain

$$-|\lambda|\|B\|_2 + \|Cx\|_2 \leq |\lambda|^2\|Ax\|_2 \leq |\lambda|\|B\|_2 + \|Cx\|_2.$$

By using (1.18) in Theorem 1.4, we get the following inequalities

$$|\lambda^2|\|A\|_2 + \|B\|_2|\lambda| - \frac{1}{\|C^{-1}\|_2} \geq 0, \tag{5.9}$$

$$\frac{|\lambda|^2}{\|A^{-1}\|_2} - \|B\|_2|\lambda| - \|C\|_2 \leq 0. \tag{5.10}$$

By solving each of these inequalities (5.9) and (5.10), we obtain the values the values of $\delta_{\pm}$ and $\widetilde{\delta}_{\pm}$ defined in (5.7)–(5.8). The solution of these inequalities are given by $\mathbb{R} \setminus (\delta_-, \delta_+)$ for (5.9) and by $[\widetilde{\delta}_-, \widetilde{\delta}_+]$ for (5.10). We know that the intersection of these sets is not void since the bounds for the eigenvalues exist. Thus, by tacking the intersection of the solution sets, we obtain the solution

of the system defined by the inequalities (5.9)–(5.9) and the bounds r_1 and r_2 defined by (5.6). □

A more general version of this theorem is given in [40, Lemma 3.1], with a different proof.

5.2.2 Real Eigenvalues with $M > 0,\ K \leq 0$

We consider the case where M is definite positive, K is semi-negative definite and the eigenvalues of (M, K) are distinct. We recall that in this case the eigenvalues of (M, K) are real. Let $M = G_1^T D_1 G$ be the eigendecomposition of M, with G_1 orthogonal. Define $\widetilde{K} = D_1^{-1/2} G_1 K G_1^T D_1^{-1/2}$ and let $\widetilde{K} = G_2^T \widetilde{D} G_2$ be its eigendecomposition, with G_2 orthogonal. We define $Q = G_2 D_1^{-1/2} G_1$ and $Z = Q^{-1}$. Thus, by left and right multiplication by Q and Z, the QEP (5.2) becomes with $D = I$,

$$(\lambda^2 I + \lambda v v^T + \widetilde{D})y = 0.$$

The associated secular equation is given by

$$f(\lambda) = 0,$$

where f is defined on $\mathbb{R}$ by

$$f(\lambda) = \lambda\sigma \sum_{k=1}^{n} \frac{v_k^2}{\lambda^2 + \widetilde{d}_k} - 1. \tag{5.11}$$

Note that by Sylvester's theorem, $D^{-1/2} G_1 K G_1^T D^{-1/2}$ is semi-definite negative. Thus, f has $2n$ poles given by $\pm\sqrt{-\widetilde{d}_k}$. On each interval $(\sqrt{-d_k}\ \sqrt{-d_{k+1}})$, the derivative of f has a constant sign equal to $-\sigma$. f is monotone in each interval $(\sqrt{-d_k}\ \sqrt{-d_{k+1}})$. Thus, we know that f has a zeros in each of these intervals.

We first apply the bisection method in each interval to approximate the zeros of f. The approximate solutions can then be used as starting points for Newton's

method. Without loss of generality, we choose $\sigma = 1$ which implies that f is decreasing. Let $z_0 \in (\sqrt{-d_k}\ \sqrt{-d_{k+1}})$ be a zeros of f for some k and let ϵ be a given parameter to locate the zeros of f within the interval $(z_0 - \epsilon, z_0 + \epsilon)$. The iterations of the bisection method are then given by

$$\begin{aligned} a_0 &= \sqrt{-d_k}, \quad b_0 = \sqrt{-d_{k+1}}, \\ (a_{n+1}, b_{n+1}) &= \begin{cases} (a_n, \frac{a_n+b_n}{2}) & \text{if } \ f(\frac{a_n+b_n}{2}) \geq 0, \\ (\frac{a_n+b_n}{2}, b_n) & \text{otherwise.} \end{cases} \end{aligned}$$

We perform the above iterations until $b_n - a_n < \epsilon$. For $\epsilon = 10^{-p}$, one can easily see that the number of bisection iterations that are required is given by $n \in \mathbb{N}$ such that

$$n > \frac{\log(\sqrt{-d_{k+1}} - \sqrt{-d_k})}{\log(2)} + p\frac{\log(10)}{\log(2)}.$$

A global number of bisection iterations can be given using the bounds obtained in Theorem 5.1,

$$n = \lfloor \frac{\log(r_2 - r_1) + p\log(10)}{\log(2)} \rfloor + 1.$$

Newton's method in this case is the classical method for finding the zeros of a nonlinear scalar function. The Newton iteration is defined by

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}.$$

In the next section, we analyze the general case where the eigenvalues can be complex.

5.2.3 General Case

We describe a basic method to solve (5.4) that consists of transforming (5.4) into a root finding problem of a scalar polynomial. We consider the secular equation (5.4) and we define the scalar polynomials

$$q(z) = \prod_{k=1}^{n}(d_k z^2 + \widetilde{d}_k), \quad \widetilde{q}_j(z) = \prod_{k=1, k\neq j}^{n}(d_k z^2 + \widetilde{d}_k),\ j = 1\colon n.$$

By multiplying (5.4) by q, we obtain an equivalent equation

$$q(z) - \sigma z \sum_{k=1}^{n} v_k \overline{\widetilde{v}}_k \widetilde{q}_k(z) = 0. \tag{5.12}$$

We define the scalar polynomial p by

$$p(z) = q(z) - \sigma\lambda \sum_{k=1}^{n} v_k \overline{\widetilde{v}}_k \widetilde{q}_k.$$

Our aim is to obtain the coefficients of p in its developed form. Let

$$p(z) = \sum_{k=0}^{2n} a_k z^k.$$

p is a $2n$ degree polynomial and thus we need $2n+1$ distinct values of p in order to determine a_k for $k = 0\colon 2n$. Then, the coefficients of p are obtained by solving a linear system $Ax = b$, with $A \in \mathbb{C}^{(2n+1)\times(2n+1)}$, $b = (b_k) \in \mathbb{R}^{2n+1}$ is a vector containing the $2n+1$ distinct values of p and $x = (a_k)$ is the vector containing the coefficients of p. The entries of b are of the form $b_k = p(z_k)$, where $z_k \in \mathbb{C}$ and $k = 1\colon 2n+1$. Our aim is to minimize the condition number of A so that the solution x is accurate. We chose

$$z_q = \exp\left(i\frac{2q\pi}{2n+1}\right), \qquad q = 0\colon 2n.$$

We have

$$p(z_q) = \sum_{k=0}^{2n} a_k \exp\left(i\frac{2kq\pi}{2n+1}\right).$$

Thus, we obtain the matrix $A = (\alpha_{kq})$ with

$$\alpha_{q+1,k+1} = \exp\left(i\frac{2kq\pi}{2n+1}\right), \qquad k, q = 0\colon 2n.$$

We see that $(2n+1)^{-1/2}A$ is unitary which is the best possible choice to solve the system $Ax = b$ by dividing each side by $\sqrt{2n+1}$. Once we have the coefficients

of p in its developed form, any method that finds the root of a polynomial can be used. We refer to [8] where the author uses the Erhlich-Aberth method (see Section 5.5.1) and Rouché's theorem to approximate the zeros of a polynomial. We believe more efficient methods can be derived to solve the secular equation (5.4).

5.3 Solving PEPs Through Linearization

5.3.1 Different Linearisations

In order to solve numerically PEPs of order higher than one, we transform them into GEPs of larger size (mn). This is the same idea as transforming an ordinary differential equation (ODE) of order higher than one into an ODE of order one. These transformations are known as linearizations, and they are not unique for a given PEP.

Definition 5.1 *[30] We say that the pair $(\mathcal{A}, \mathcal{B})$ is a linearization of the $n \times n$ matrix polynomial $P(A, \lambda)$ of degree m, if there exist two matrices $E(\lambda)$ and $F(\lambda)$, of size mn, with a constant non-zero determinant such that*

$$E(\lambda)(\mathcal{A} - \lambda\mathcal{B})F(\lambda) = \begin{pmatrix} P(A,\lambda) & 0 \\ 0 & I_{n(m-1)} \end{pmatrix}.$$

After the linearization of the PEP, we end up with the new problem

$$(\beta\mathcal{A} - \alpha\mathcal{B})z_r = 0, \quad z_l^*(\beta\mathcal{A} - \alpha\mathcal{B}) = 0, \tag{5.13}$$

where (α, β) are the eigenvalues of the PEP, and we still need to recover x and y, the eigenvectors of the PEP from z_r and z_l.

In the following subsections, we focus on numerical methods and algorithms for GEPs and then for PEPs, analyzing different types of linearizations. We first state a well known theorem: *the generalized Schur decomposition.*

Theorem 5.2 *Let $(A, B) \in \mathcal{M}_n(\mathbb{C})^2$. There exist two unitary matrices Q and Z such that $Q^*AZ = T$ and $Q^*BZ = S$ are triangular. For $i = 1\colon n$, we denote α_{ii} and β_{ii}, the diagonal elements of T and S. If there exists an i such that $\alpha_{ii} = \beta_{ii} = 0$, then the spectrum of the matrix polynomial defined by the pair (A, B) is the entire complex plane $\mathbb{C}$, which corresponds to the non-regular case. Otherwise the eigenvalues are given by $(\alpha_{ii}, \beta_{ii})$, for $i = 1\colon n$.*

Proof. We refer to [18] and [68] for two different proofs . □

The standard way of solving the GEPs is via the QZ algorithm of C.B. Moler and G.W. Stewart [53]. For a given matrix pair, this algorithm computes the generalized Schur decomposition from which the eigenvalues are recovered as in the above theorem.

We now focus on some linearizations.

5.3.2 Companion Linearization

Let

$$\mathcal{A} = \begin{pmatrix} A_0 & 0 & \cdots & 0 \\ 0 & I_n & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & I_n \end{pmatrix},$$

$$\mathcal{B} = \begin{pmatrix} -A_1 & -A_2 & \cdots & -A_m \\ I_n & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & \cdots & I_n & 0 \end{pmatrix}.$$

The pencil $\mathcal{A}-\lambda\mathcal{B}$ is a linearization of $P(A, \lambda)$ called the companion linearization. A left and right eigenvector of $P(A, \lambda)$, say y and x can be recovered from a left

and right eigenvector of $\mathcal{A} - \lambda\mathcal{B}$. We first assume that $\alpha \neq 0$ and $\beta \neq 0$. We define $\lambda = \frac{\alpha}{\beta}$. For a vector $z \in \mathbb{C}^{mn}$, we write $z^k = z(nk+1\colon n(1+k))$, for $0 \leq k \leq m-1$. Let z_l and z_r be the left and right eigenvectors satisfying (5.13).

We have

$$((\beta\mathcal{A} - \alpha\mathcal{B})z_r)^0 = \beta A_0 z_r^1 + \alpha \sum_{k=1}^{m} A_k z_r^k$$

and for $1 \leq k \leq m-1$,

$$((\beta\mathcal{A} - \alpha\mathcal{B})z_r)^k = \beta z_r^{k+1} - \alpha z_r^k.$$

By induction on $2 \leq k \leq m-1$, when (5.13) is satisfied, we show that

$$\begin{aligned} z_r^k &= \lambda^{k-1} z_r^1, \\ \sum_{k=0}^{m} \alpha^k \beta^{m-k} A_k z_r^k &= 0. \end{aligned}$$

Thus, we can choose $x = z_r^k$, for some $0 \leq k \leq m-1$. We choose the z_r^k for which the backward error is minimal.

For the left eigenvector y, we can show by induction on m that $y = z_l^0$, that is, we can recover y by reading off the n leading components of z_l.

Let's assume now that $\beta = 0$. Then, we have for $1 \leq k \leq m-1$, $z_r^k = 0$ and $x = z_r^m$, with $x \in \text{null}(A_m)$. In the case $\alpha = 0$, we have for $2 \leq k \leq m$, $z_r^k = 0$ and $x = z_r^1$, with $x \in \text{null}(A_0)$.

More generally, we can show [50] that x is an eigenvector of P with a finite eigenvalue λ if and only if

$$\begin{bmatrix} 1 \\ \lambda \\ \vdots \\ \lambda^{m-2} \\ \lambda^{m-1} \end{bmatrix} \otimes x$$

is an eigenvector of $\mathcal{A} - \lambda\mathcal{B}$ with eigenvalue λ. If λ is an infinite eigenvalue then x is a right eigenvector of P if and only if $e_1 \otimes x$ is a right eigenvector of $\mathcal{A} - \lambda\mathcal{B}$ [50].

5.3.3 Symmetric Linearization

Let

$$\mathcal{A} = \begin{pmatrix} 0 & \cdots & \cdots & 0 & A_0 \\ \vdots & & & A_0 & A_1 \\ \vdots & & & A_1 & \vdots \\ 0 & A_0 & & & A_{m-2} \\ A_0 & A_1 & \cdots & A_{m-2} & A_{m-1} \end{pmatrix},$$

$$\mathcal{B} = \begin{pmatrix} 0 & \cdots & 0 & A_0 & 0 \\ \vdots & 0 & A_0 & A_1 & \vdots \\ 0 & & & \vdots & \vdots \\ A_0 & A_1 & \cdots & A_{m-2} & 0 \\ 0 & \cdots & \cdots & 0 & -A_m \end{pmatrix}.$$

If A_m is nonsingular, one can show that $\mathcal{A} - \lambda\mathcal{B}$ is a linearization [30]. We see that in the case where all the matrices A_k are symmetric, the matrices $\mathcal{A}$ and $\mathcal{B}$ are symmetric. The main difference (with the companion linearization) is that the left eigenvector has the same structure as the right one, which is $z_l^k = \lambda^{k-1} z_l^1$ for $2 \leq k \leq m$. For the zeros and infinite eigenvalues, the eigenvector is obtained using the same analysis that was done for the companion linearization.

5.3.4 Influence of the Linearization

Amongst the two linearizations defined in Sections 5.3.2 and 5.3.3, the companion linearization is the most used in practice. Several open questions remain in the

numerical solution of PEPs. It is clear that the linearization should have an influence on the computed eigenpairs, it is a fact that we illustrate with a QEP example in the second part of this paragraph. The first question is how to choose the best linearization so that the eigenpairs are computed the most accurately. The second question is how to describe all possible linearizations of a matrix polynomial.

Let $P(A, \lambda)$ be a matrix polynomial and $\mathcal{A} - \lambda\mathcal{B}$ be one of its linearizations. The condition number of λ as an eigenvalue of the PEP should be less than the condition number of λ as an eigenvalue of the GEP obtained after the linearization process. There is no proof of this result. The heuristic argument is that the class of possible perturbations for the PEP is smaller than the class of possible perturbations for the corresponding GEP. An interesting question is how to find the "best" linearization, that is, a linearization for which the condition number of λ as an eigenvalue of the GEP is minimal.

In [71], Tisseur analyzed three linearizations of QEPs and found bounds for the corresponding condition numbers based on the norms of the coefficient matrices and the modulus of the eigenvalue. More recently, a wide class of linearizations is described in [50]. The conditioning of these linearizations is analyzed in [39]. When the problem is not too badly scaled, two particular linearizations are shown to be almost optimal: they are about as well conditioned as the original polynomial. Balancing or scaling matrices is common for SEPs and GEPs, [31], [67]. In [39] and [71], the authors analyze the effect of scaling a PEP on the condition number.

We consider a quadratic eigenvalue problem $(\lambda^2 M + \lambda D + K)x = 0$ where $M, D, K \in \mathbb{R}^{n\times n}$ are symmetric, with $n = 200$. In this example, $D = e_s e_s^T$ with $s = 100$ and M and K are positive definite. The problem is stable, all the eigenvalues are in the left half plane. This problem is described in more detail in

Section 7.8.2.

Figures 5.1 and 5.2 illustrate the sensitivity of the linearization process. We plotted in the complex plane the eigenvalues of a QEP that we first solved by the companion linearization in Figure 5.1 and then by the symmetric linearization in Figure 5.2. After the linearization, we used the QZ algorithm. We see in Figure 5.1 that some eigenvalues computed using the companion linearization have a positive real part whereas those computed with the symmetric linearization have a negative or 0 real part. Thus, in this case the symmetric linearization performs better than the companion linearization. This fact can be explained by the fact that the symmetric linearization preserves the symmetry structure whereas the companion linearization destroys it. In Section 7.8.2, using the symmetric linearization, we compare the eigenvalues obtained by a symmetry structure preserving algorithm (the HZ algorithm, see Chapter 6) and the QZ algorithm.

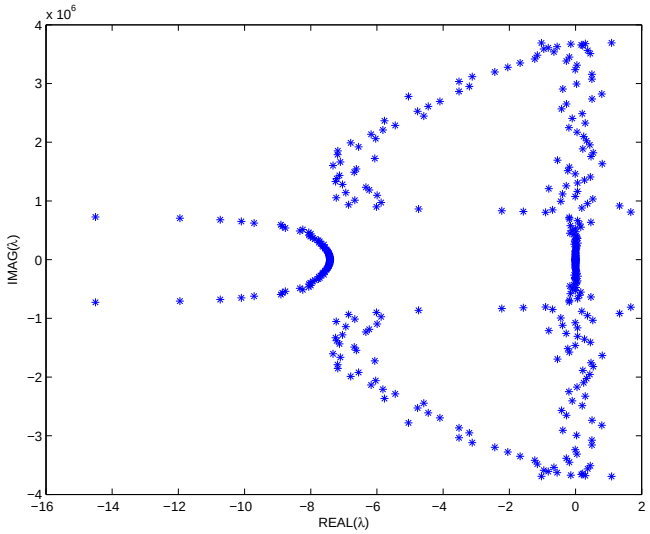

Figure 5.1: Spectrum computed with the companion linearization.

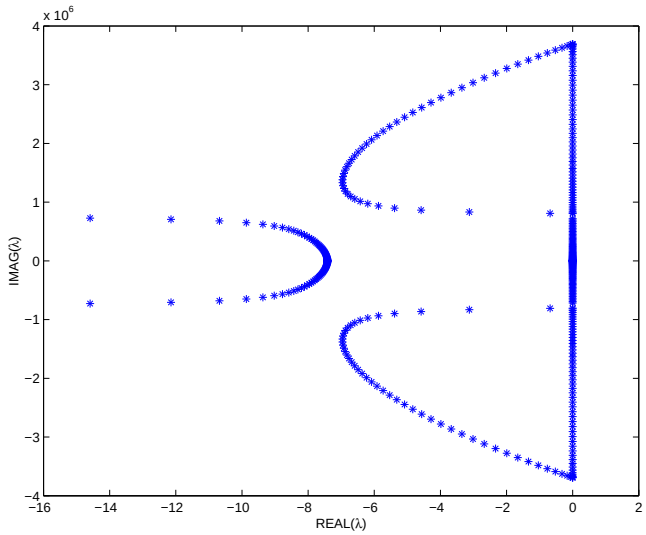

Figure 5.2: Spectrum computed with the symmetric linearization.

5.3.5 Pseudocode

The following pseudocode solves the PEP by a companion linearization, then gives the condition number $c_2(A, \alpha, \beta)$.

Form $(\mathcal{A}, \mathcal{B})$ such that $\beta\mathcal{A} - \alpha\mathcal{B}$ is a companion linearization of $P(\alpha, \beta)$.

Compute the generalized Schur decomposition

$S = Q^*AZ$, $T = Q^*BZ$.

For $k = 1 : nm$

$$\lambda_k = \begin{cases} s_{kk}/t_{kk} \text{ if } t_{kk} \neq 0, \\ \text{inf} \ \text{ if } t_{kk} = 0. \end{cases}$$

Solve $u_l^*(t_{kk}S - s_{kk}T) = 0$, $z_l = Q^*u_l$.

$y_k = z_l(1 : n)$.

Solve $(t_{kk}S - s_{kk}T)u_r = 0$, $z_r = Z^*u_r$.

if $s_{kk} = 0$,

$x_k = z_r(1 : n)$,

end if.

if $t_{kk} = 0$,

$x_k = z_r((m-1)n + 1 : nm)$,

end if.

% For the finite non zero eigenvalues, we take the right eigenvector

% of the PEP from the big eigenvector of the GEP. We choose the right

% eigenvector that has the minimal residual.

if ($s_{kk} \neq 0$ and $t_{kk} \neq 0$),

$r_1 = \|P(s_{kk}, t_{kk})z(1\colon n)\|$

For $i = 1 : m - 1$

$r_{i+1} = \|P(s_{kk}, t_{kk})z(in + 1\colon (i + 1)n)\|$

$$x_k = \begin{cases} z_r((i-1)n + 1 : in) \text{ if } \|r_i\| \leq \|r_{i+1}\|, \\ z(in + 1 : (i + 1)n) \text{ otherwise.} \end{cases}$$

end.

end if.

% Compute $c_k = c_2(A, s_{kk}, t_{kk})$.

Set $c_k = 0$.

For $k = 0 : m$

$c_k = c_k + |\alpha|^{2k}|\beta|^{2(m-k)}$

end

$c_k = \|x\|_2\|y\|_2 c_k$.

$v = \bar{\beta}\beta^{m-1}A_1 x$

For $k = 2 : m$

$v = v + k\bar{\beta}\alpha^{k-1}\beta^{m-k}A_k x$

end.

For $k = 0 : m - 1$

$v = v + (m - k)\bar{\alpha}\alpha^k\beta^{m-k-1}A_k x$

end.

$c_k = \frac{c_k}{|y^* v|}$

end.

5.4 Numerical Examples with `condpolyeig`

5.4.1 Lack of Numerical Tools

In Table 5.1, we give numerical tools for eigenvalue problems available in different software. In the second row, `polyeig` is a MATLAB function that solves PEPs. We see that `polyeig` is the only routine available for PEPs and that there are no other routines (routines for the SEP) such as `psa` (computation of pseudospectra), `condeig` (condition number). There is a lack of numerical tools for solving PEPs and analyzing their sensitivity.

Table 5.1: List of eigentools.

Problem	MATLAB built-in	MATLAB other	Scilab
$A - \lambda I$	`eig(A)`, `eigs(A)`, `condeig(A)`, `schur(A)`	`psa(A)` (Wright), `fv(A)`, `gersh(A)`,	`spec(A)`, `bdiag(A)`, `htrianr(A)`
$\lambda^2 A + \lambda B + C$	`polyeig(C,B,A)`		

In an earlier version of MATLAB (version 6.1.0.450 (R12.1)), `polyeig` failed to return the right eigenvector corresponding to infinite eigenvalues. Here is an example of a MATLAB output:

```
>> A=rand(3);B=rand(3);C=rand(3);
>> C(:,3)=0;
>> [X E]=polyeig(A,B,C)
```

```
Warning: Divide by zero.
> In /opt/matlab6.1/toolbox/matlab/matfun/polyeig.m at line 76
X =
  Columns 1 through 4
NaN    -0.4726 - 0.0561i   0.6889 + 0.2037i  -0.0728 - 0.0984i
NaN     0.4687 + 0.0556i  -0.5042 - 0.1490i -0.3683 - 0.4977i
NaN     0.7369 + 0.0875i  -0.4369 - 0.1292i  -0.4614 - 0.6235i
  Columns 5 through 6
  -0.4799 + 0.2784i   0.4033 + 0.3811i
   0.1084 + 0.4092i  -0.1993 + 0.3735i
   0.4340 - 0.5697i  -0.2918 - 0.6540i
E =
      Inf
  18.4943 + 0.0000i
   2.3520 + 0.0000i
  -1.3149 - 0.0000i
  -0.1048 - 0.6647i
  -0.1048 + 0.6647i
```

This problem is fixed in the new version 7.0.1 of MATLAB by analyzing separately the right eigenvectors corresponding to zero or infinite eigenvalues.

In the next section, we present a MATLAB routine, `condpolyeig`, in our quest of filling this lack of numerical tools for the PEP.

5.4.2 `condpolyeig`

Let $(A_0, \cdots, A_m) \in \mathcal{M}_n(\mathbb{C})^{m+1}$, $\mu \in \mathbb{C}^{m+1}$. Once we have solved the corresponding PEP, we would like to know how sensitive the eigenvalues are to perturbations

in the data. For this reason, we present a MATLAB routine, `condpolyeig`, that computes the condition number of a simple eigenvalue.

The call

```
>>[X,Y,E,s]=condpolyeig(A0,....,Am)
```

in the MATLAB prompt, returns the right and left eigenvectors in the $n \times mn$ matrices `X` and `Y`. The eigenvalues, $\frac{\alpha}{\beta}$ are in the mn vector `E` and the corresponding condition numbers in `s`. If the number of output is one then `condpolyeig` returns, by default the condition numbers. These condition numbers are computed by using the formula (2.4) where the default vector of weights μ is given by

$$\mu_j = \begin{cases} \frac{1}{\|E_j\|_F}, & E_j \neq 0, \\ 1, & E_j = 0. \end{cases}$$

The call

```
>>[X,Y,E,s]=condpolyeig(A0,....,Am,mu)
```

allows to compute the condition number using the weight defined in `mu`.

In order to solve a PEP of degree m, for $n \times n$ matrices, we use a companion linearisation, described in Section 5.3.2. Then we use the `qz` function of MATLAB that returns the right and left eigenvectors and the eigenvalues. We select the left and right eigenvectors as explained in Section 5.3.2. Finally, we compute the condition number.

5.4.3 Numerical Examples

We consider the quadratic eigenvalue problem with

$$P(A_\theta, \alpha, \beta) = \begin{bmatrix} \alpha^2 - 3\alpha\beta + 2\beta^2 & -\alpha^2 + \alpha\beta & -\alpha^2 + 9\beta^2 \\ 0 & \alpha^2 - \alpha\beta(1+\theta) & 0 \\ 0 & 0 & \alpha\beta - 3\beta^2 \end{bmatrix},$$

where θ is a positive parameter. Since $P(A_\theta, \alpha, \beta)$ is upper triangular, the exact eigenvalues are readily available. They are given in Table 5.2.

Table 5.2: Eigenvalues of $P(A_\theta, \alpha, b)$.

k	1	2	3	4	5	6
(α_k, β_k)	$(0,1)$	$(1,1)$	$(1+\theta,1)$	$(2,1)$	$(3,1)$	$(1,0)$
$\lambda = \frac{\alpha_k}{\beta_k}$	0	1	$1+\theta$	2	3	∞

Our aim is to analyze the behaviour of the eigenvalues, when the parameter θ tends to -1. The condition number for the infinite eigenvalue is 1. The eigenvalue 0 becomes double when $\theta = -1$. Tables 5.3 and 5.4 give $\chi(\tilde{\lambda}, \lambda)$ (see Corollary 2.7), where $\tilde{\lambda}$ is computed by `condpolyeig`, $\lambda = \lambda_\theta$ the exact eigenvalue of $P(A_\theta, \alpha, \beta)$, the eigenvalue condition number $c_2(A_\theta, \alpha, \beta, x)$ in (2.12) and the backward error $\eta(\hat{\alpha}, \hat{\beta}, \hat{x})$ in (3.5). Recall that from Corollary 2.7, for small $\epsilon = \|\Delta A\|_F$,

$$\chi(\tilde{\lambda}, \lambda) \leq c_2(A_\theta, \alpha, \beta, x)\epsilon + O(\epsilon^2),$$

where $\chi(\tilde{\lambda}, \lambda)$ is the chordal distance between $\tilde{\lambda}$ and λ. We see that

$$c_2(A_\theta, \alpha, \beta, x)\eta(\hat{\alpha}, \hat{\beta}, \hat{x})$$

is in most cases of the same order as $\chi(\tilde{\lambda}, \lambda)$. This is an illustration of the formula:

$$\text{forward error} \leq \text{condition number} \times \text{backward error}.$$

Table 5.3: Condition number and backward error for $\lambda = 0$.

θ	$\chi(\tilde{\lambda}, 0)$	$c_2(A_\theta, \alpha, \beta, x)$	$\eta(\hat{\alpha}, \hat{\beta}, \hat{x})$	$c_2(A_\theta, \alpha, \beta, x)\eta(\hat{\alpha}, \hat{\beta}, \hat{x})$
$-1+10^{-4}$	8.10^{-14}	1.10^{4}	7.10^{-18}	7.10^{-14}
$-1+10^{-6}$	4.10^{-12}	1.10^{6}	4.10^{-18}	3.10^{-12}
$-1+10^{-8}$	1.10^{-9}	9.10^{7}	1.10^{-17}	1.10^{-9}
$-1+10^{-10}$	9.10^{-10}	1.10^{9}	7.10^{-19}	9.10^{-10}

Table 5.4: Condition number and backward error for $\lambda = 1 + \theta$.

θ	$\chi(\tilde{\lambda}, \lambda)$	$c_2(A_\theta, \alpha, \beta, x)$	$\eta(\hat{\alpha}, \hat{\beta}, \hat{x})$	$c_2(A_\theta, \alpha, \beta, x)\eta(\hat{\alpha}, \hat{\beta}, \hat{x})$
$-1 + 10^{-4}$	7.10^{-14}	9.10^{11}	7.10^{-18}	7.10^{-6}
$-1 + 10^{-6}$	4.10^{-12}	2.10^{11}	4.10^{-18}	1.10^{-6}
$-1 + 10^{-8}$	1.10^{-9}	7.10^{8}	1.10^{-17}	1.10^{-8}
$-1 + 10^{-10}$	9.10^{-10}	1.10^{9}	7.10^{-19}	8.10^{-10}
$-1 + 10^{-16}$	2.10^{-9}	6.10^{8}	7.10^{-18}	2.10^{-9}

5.5 An Overview of Algorithms for Symmetric GEPs

We consider the generalized eigenvalue problem $Ax = \lambda Bx$, where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$ are symmetric. For such problem, the eigenvalues λ can be real or complex but they come in pairs $(\lambda, \bar{\lambda})$. Symmetric GEPs arise as intermediate steps in a variety of eigenvalue problems. For example, the quadratic eigenvalue problem $(\lambda^2 M + \lambda D + K)x = 0$ with symmetric coefficient matrices is frequently encountered in structural mechanics [73]. The standard way of dealing with this problem in practice is to reformulate it as a generalized eigenvalue problem (GEP) $Ax = \lambda Bx$ of twice the dimension. We recall that this process is called linearization as the GEP is linear in λ. Symmetry in the problem is maintained with an appropriate choice of linearization. For example, we can take

$$A = \begin{bmatrix} 0 & K \\ K & D \end{bmatrix}, \quad B = \begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix}, \quad x = \begin{bmatrix} u \\ \lambda u \end{bmatrix}.$$

The resulting A and B are symmetric but not definite, and in general the pair (A, B) is indefinite. Thus the Cholesky-QR algorithm [22], [31], or the symmetric Lanczos algorithm in the sparse case, cannot be applied. Usually, the symmetric indefinite GEP is solved by applying general GEP techniques that destroy any symmetry in the problem; for example the QZ algorithm or an Arnoldi process if A and B are large and sparse. Though symmetric indefinite GEPs do not

have any special spectral properties there are some advantages to preserving the symmetry, such as reductions in storage and the computational cost. When $A - \lambda B$ is of small to medium size, it can be reduced to a symmetric tridiagonal-diagonal form $T - \lambda S$ using one of the procedures described by Tisseur [74]. For large and sparse matrices the pseudo-Lanczos algorithm of Parlett and Chen [56] applied to $A - \lambda B$ yields a projected problem of the form $Ax = \lambda Bx$.

Tridiagonal-diagonal pencils arise also when solving nonsymmetric eigenvalue problems $Ax = \lambda x$. In the dense case, A can be reduced to nonsymmetric tridiagonal form $\tilde{T}$ [29], and in the sparse case, the nonsymmetric Lanczos algorithm produces an nonsymmetric tridiagonal matrix $\widetilde{T}$. Assuming that $\widetilde{T}$ has no zero subdiagonal and superdiagonal entries, one can easily construct two nonsingular diagonal matrices D_1, D_2 such that $T - \lambda J = D_1(\widetilde{T} - \lambda I)D_2$ with T symmetric tridiagonal and J diagonal with diagonal entries ± 1.

We are interested in robust and efficient algorithms that compute all the eigenvalues and eigenvectors of $T - \lambda S$ while preserving the structure of the problem. When applied to the nonsymmetric tridiagonal matrix $S^{-1}T$, the QR algorithm [31] does not preserve the tridiagonal structure: the matrix $S^{-1}T$ is considered as a Hessenberg matrix and the upper part of $S^{-1}T$ is filled in during the iterations. Therefore the QR algorithm requires some extra storage. Two alternatives applicable to $S^{-1}T$ are the LR algorithm [62] for nonsymmetric tridiagonal matrices and the HR algorithm [13], [14]. Both algorithms preserve the tridiagonal form of $S^{-1}T$ but may be unstable as they use non-orthogonal transformations. Another alternative is the Erhlich-Aberth method.

5.5.1 The Erhlich-Aberth Method

We start by describing a method to compute roots of a polynomial. Let p be a scalar polynomial of degree n and $z = (z_j) \in \mathbb{C}^n$. The Erhlich-Aberth iterations given in [1] and [28], are defined by

$$z_j^{(k+1)} = z_j^{(k)} - \frac{\frac{P(z_j^{(k)})}{P'(z_j^{(k)})}}{1 - \frac{P(z_j^{(k)})}{P'(z_j^{(k)})} \sum_{q=1, q\neq j}^{n} \frac{1}{z_j^{(k)} - z_q^{(k)}}}, \ j = 1\colon n. \tag{5.14}$$

These iterations converge locally to the n roots of p allowing to approximate simultaneously all the roots. A detailed presentation and implementation is given in [8], where the starting approximations are obtained by Rouché's Theorem.

D. Bini and F. Tisseur proposed a method to compute the eigenvalues of a symmetric GEP in [7] based on the Erhlich-Aberth iterations. An efficient and robust implementation of these iterations depends on the set of starting values $z^{(0)}$ and how the Newton correction $p(\lambda)/p'(\lambda)$ are computed. Let T be symmetric tridiagonal and J be a signature matrix. The Erhlich-Aberth method is then applied to $p(\lambda) = \det(T - \lambda J)$. The Newton correction is given by

$$\frac{p(\lambda)}{p'(\lambda)} = -\frac{1}{\operatorname{trace}(T - \lambda J)^{-1}}.$$

They propose a robust method to compute the Newton's correction based on the QR factorization of $T - \lambda J$. The initial approximations for (5.14) are obtained using a divide and conquer strategy.

5.5.2 LR Algorithm

The LR algorithm [62] is an iterative process to compute the eigenvalues of a matrix based on an LU factorization. The LR iterations are given by

$$T_0 \ = \ T,$$

$$\begin{aligned} T_k &= L_k U_k \text{ (LU factorization)}, \\ T_{k+1} &= U_k L_k. \end{aligned}$$

If T is tridiagonal, the LR iterations preserve the tridiagonal form. Note that a successful implementation needs a pivoted LU factorization [78]. A detailed analysis of the LR algorithm, its implementation and the first step that consists of reducing a general matrix to a tridiagonal form is presented in [57] and [23].

5.5.3 HR Algorithm

The HR algorithm of Brebner and Grad [13] and Bunse-Gerstner [14] is an iterative procedure that begins with the pseudosymmetric matrix $\widetilde{T}_0 = J^{-1}T$. It produces a sequence of similar pseudosymmetric matrices $\widetilde{T}_k = J_k T_k$ obtained from an HR factorization with respect to J_{k-1},

$$p_k(\widetilde{T}_{k-1}) = H_k R_k, \qquad H_k^T J_{k-1} H_k = J_k, \tag{5.15}$$

where p_k is a polynomial. It sets

$$\widetilde{T}_k = H_k^{-1} \widetilde{T}_{k-1} H_k. \tag{5.16}$$

Note that $\widetilde{T}_k = (J_k H_k^T J_{k-1}) \widetilde{T}_{k-1} H_k$, which implies that if $\widetilde{T}_{k-1}$ is pseudosymmetric then $\widetilde{T}_k$ is pseudosymmetric. Also, $\widetilde{T}_k = R_k \widetilde{T}_{k-1} R_k^{-1}$ (with R_k upper triangular) which implies that if $\widetilde{T}_{k-1}$ is tridiagonal then $\widetilde{T}_k$ is tridiagonal. Hence the HR iterations preserve pseudosymmetric tridiagonal forms.

This algorithm is analyzed in the following chapter.

Chapter 6

The HZ Algorithm

6.1 Introduction

Our aim is to derive an efficient and robust implementation of the HZ algorithm [13], [14]. We consider the symmetric pair (A, B). We assume that the pencil $A - \lambda B$ is *regular*, that is, $\det(A - \lambda B) \not\equiv 0$ and that B is nonsingular. For nonregular pencils or singular B we refer to Lucas [49] who shows that the pencil can be deflated and reduced to a regular pencil $A - \lambda B$, where B is nonsingular.

The method we consider consists of three main steps.

6.1.1 Symmetric–Diagonal Reduction

We present briefly two methods that reduce the symmetric pair (A, B) to a symmetric-diagonal pair $(A, B) = M^T(C, J)M$.

We can use the eigendecomposition of $B = Q^T DQ$, where Q is orthogonal and D diagonal. Then, $J = \text{sign}(D)$, $C = |D|^{-1/2}QAQ^T|D|^{-1/2}$ and $M = Q^T|D|^{-1/2}$.

Since B is indefinite we can also use a block LDL^T factorization [37, Ch.11]

$$P^T BP = LDL^T, \tag{6.1}$$

where P is a permutation matrix, L is unit lower triangular and D is block diagonal with 1×1 or 2×2 blocks on its diagonal. Let

$$D = X|\Lambda|^{1/2}\widetilde{J}|\Lambda|^{1/2}X^T, \quad \widetilde{J} \in \mathrm{diag}_q^n(\pm 1), \tag{6.2}$$

be the eigendecomposition of D, where X is orthogonal and Λ is the diagonal matrix of eigenvalues of D. The pair $(C, \widetilde{J})$ with

$$C = M^T A M, \quad M = P L^{-T} X |\Lambda|^{-1/2} \tag{6.3}$$

is congruent to (A, B) and is in symmetric-diagonal form. This reduction is not as stable as the one based on the eigendecomposition of B since it uses non-orthogonal transformations. We refer to [74] for an analysis of its numerical stability. It is however a lot less expensive than computing the whole eigendecomposition of B.

6.1.2 Tridiagonal–Diagonal Reduction

The symmetric matrix C in (6.3) can be tridiagonalized using a sequence of congruence transformations $Q_1 Q_2 \cdots Q_{n-2} = Q$ that preserve the diagonal form of the second matrix $\widetilde{J}$,

$$Q^T C Q = T, \quad Q^T \widetilde{J} Q = J, \quad J \in \mathrm{diag}_q^n(\pm 1).$$

For the Q_i, Tisseur [74] suggests to use a product of two Householder reflectors followed by a hyperbolic rotation. We refer to [74] for the details of the implementation.

6.1.3 HR or HZ Iterations

The HR algorithm [13], [14] is an iterative process that begins with the pseudosymmetric matrix $\widetilde{T}_0 = J^{-1}T$. It produces a sequence of similar matrices $\widetilde{T}_k$,

$k \geq 1$ obtained, when it exists, from an HR factorization with respect to J_{k-1},

$$p_k(\widetilde{T}_{k-1}) = H_k R_k, \qquad H_k^T J_{k-1} H_k = J_k, \tag{6.4}$$

where p_k is a polynomial. It sets

$$\widetilde{T}_k = H_k^{-1} \widetilde{T}_{k-1} H_k. \tag{6.5}$$

If $\widetilde{T}_{k-1} := J_{k-1} T_{k-1}$ is pseudosymmetric ($T_{k-1} = T_{k-1}^T$) then

$$\widetilde{T}_k = (J_k H_k^T J_{k-1}) \widetilde{T}_{k-1} H_k = J_k (H_k^T T_{k-1} H_k) := J_k T_k \tag{6.6}$$

is pseudosymmetric. Also, $\widetilde{T}_k = R_k \widetilde{T}_{k-1} R_k^{-1}$ with R_k upper triangular so that if $\widetilde{T}_{k-1}$ is tridiagonal then $\widetilde{T}_k$ is tridiagonal. Hence the HR iteration (6.4)–(6.5) preserves pseudosymmetric tridiagonal forms.

Using (6.6), the kth HR step (6.4)–(6.5) can be rewritten as

$$\begin{aligned} p_k(\widetilde{T}_{k-1}) &= H_k R_k, \\ J_k &= H_k^T J_{k-1} H_k, \\ T_k &= H_k^T T_{k-1} H_k. \end{aligned} \tag{6.7}$$

We will refer to (6.7) as the *kth HZ step* of the HZ algorithm. In an analogous way to the QZ algorithm, the "Z" in HZ is explained by the fact that the iteration (6.7) acts on the symmetric tridiagonal–diagonal pair (T, J) rather than on the single pseudosymmetric tridiagonal matrix $\widetilde{T}$.

The HR algorithm belongs to the broader class of GR algorithms [75] and convergence results on GR algorithms apply [75, Theorem 3.2]. One can show that if the cumulative transforming matrices H_k are uniformly bounded then the sequence T_k converges to block diagonal form with 1×1 and 2×2 blocks on the diagonal, thus exposing the eigenvalues of the pseudosymmetric tridiagonal matrix $\widetilde{T}_0 = J^{-1} T$. Note that the HR factorization may not always exist. This

may prevent the convergence of the HZ iterations. In practice we can modify the polynomial of p so that $p(T) = HR$ exists but there is no guarantee of convergence.

The HZ algorithm and its practical implementation are studied in more detail in the next sections.

6.2 Preliminaries

So far, we have enough conditions in order to analyze the existence of the HR factorization. We now present two theorems that enable us to classify difficulties that we might face during the execution of the HZ algorithm. The following theorem describes a property of the spectrum of the matrix pair (T, J).

Theorem 6.1 *Let $T \in \mathbb{R}^{n\times n}$ be tridiagonal symmetric and unreduced. Then each eigenvalue of the pair (T, J) has geometric multiplicity* 1.

Proof. Let $\lambda \in \mathbb{C}$ be an eigenvalue and x a corresponding eigenvector. Write $J = \mathrm{diag}(\sigma_i)$. From $(T - \lambda J)x = 0$ we have

$$\begin{aligned} (t_{11} - \lambda\sigma_1)x_1 + t_{12}x_2 &= 0, \\ t_{i,i-1}x_{i-1} + (t_{ii} - \lambda\sigma_i)x_i + t_{i,i+1}x_{i+1} &= 0,\ i = 2\colon n-1, \\ t_{n,n-1}x_{n-1} + (t_{nn} - \lambda\sigma_n)x_n &= 0. \end{aligned}$$

Since T is unreduced, we have that $t_{i,i+1} \neq 0$ for $1 \leq i \leq n-2$. Thus, by induction we can express each component of x as a multiple of x_1 in a unique way. It follows that the eigenspace corresponding to λ has dimension 1, that is λ has geometric multiplicity 1. □

Theorem 6.2 *Let $A \in \mathbb{R}^{n\times n}$ be pseudosymmetric for some $J \in \mathrm{diag}_q^n(\pm 1)$. If λ is a defective multiple eigenvalue of A then $p(A) = (A - \lambda I)(A - \bar{\lambda} I)$ does not have an HR factorization with respect to J.*

Proof. Let λ be an eigenvalue of A. We have by pseudosymmetry

$$p(A)^T J p(A) = J p(A)^2.$$

If λ is defective then

$$\mathrm{rank}(p(A)^T J p(A)) < \mathrm{rank}(p(A)).$$

Thus, by Corollary 4.9 $p(A)$ does not have an HR factorization with respect to J. □

Combining Theorem 6.1 and 6.2 we have that for an unreduced tridiagonal symmetric pair (T, J), the HR factorization of $p(JT) = (JT - \lambda I)(JT - \bar{\lambda} I)$ does not exist if the shift λ is a defective eigenvalue. Hence, we may expect difficulties for matrix pair with nontrivial Jordan blocks.

6.3 Practical Implementation of One HZ Step

We consider the symmetric pair (T, J) where T is unreduced tridiagonal and $J \in \mathrm{diag}_n^k(\pm 1)$ a signature matrix. We recall that if T is not unreduced then

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix},$$

and the spectrum of T is the union of the spectrum T_1 and T_2. The problem can be split into smaller unreduced problems. From now on, we assume that T is unreduced and $J \in \mathrm{diag}_n^q(\pm 1)$.

We consider one single HZ step on an unreduced T:

$$p(JT) = HR, \tag{6.8}$$

$$\widehat{T} = H^T T H, \quad \widehat{J} = H^T J H. \tag{6.9}$$

The degree of the polynomial p is called the multiplicity of the step. If p has degree 1, it is a single step. In our implementation, p is chosen to be the quadratic

$$p(z) = (z - \omega_1)(z - \omega_2), \tag{6.10}$$

where $\omega_1, \omega_2 \in \mathbb{C}$ are called shifts. We discuss later on the choice of these shifts.

The HZ step can be carried out either explicitly or implicitly without forming the matrix $p(JT)$ and its HR factorization [75]. We adopt the implicit form as it involves fewer operations. For that, we need to build a $(J, \widehat{J})$-orthogonal matrix $\widetilde{H}$ whose first column is the same as that of H in (6.8) and such that the matrix $\widetilde{T} = \widetilde{H}^T T \widetilde{H}$ is symmetric tridiagonal. Then $\widetilde{T}$ and $\widehat{T}$ are essentially the same [31], [67], [74].

Since T is in tridiagonal form, the first column of $p(JT)$ has a simple form. If $x = p(JT)e_1$ then $x = [x_1, x_2, x_3, 0, \ldots, 0]^T$, where

$$\begin{aligned} x_1 &= t_{11}^2 - t\sigma_1 t_{11} + \sigma_1\sigma_2 t_{12} t_{21} + d, \\ x_2 &= \sigma_2 t_{21}(\sigma_1 t_{11} + \sigma_2 T_{22} - t), \\ x_3 &= \sigma_2\sigma_3 t_{21} t_{32}, \end{aligned}$$

with $T = (t_{ij})_{1 \le i,j \le n}$, $J = \mathrm{diag}(\sigma_i)$, $t = \omega_1 + \omega_2$ and $d = \omega_1\omega_2$. The aim is to construct $\widetilde{H}$ such that $\widetilde{H}x$ is a multiple of e_1, and $\widetilde{T} = \widetilde{H}^{-1} T \widetilde{H}^{-T}$ is symmetric tridiagonal.

We first describe the matrix tools needed to construct $\widetilde{H}$.

6.4 Implementing the Bulge Chasing

First we determine a $(J, J_0)-$orthogonal matrix H_0 such that $H_0 x$ is a multiple of e_1 and compute $T_0 = H_0^T T H_0$. This creates a bulge at the top left corner of

T_0 as illustrated by the generic 8×8 matrix T_0

$$T_0 = \begin{bmatrix} \times & \times & \times & \times & & & & \\ \times & \times & \times & \times & & & & \\ \times & \times & \times & \times & & & & \\ \times & \times & \times & \times & \times & & & \\ & & & \times & \times & \times & & \\ & & & & \times & \times & \times & \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times \end{bmatrix}.$$

To recover the tridiagonal form the bulge is chased to the bottom right by applying carefully chosen unified rotations.

At the first stage of this process two unified rotations are used to introduce zeros in positions $(4,1)$ and $(3,1)$. At stage p, $1 \leq p \leq n-4$, we get the pair

$$(T_p, J_p) = H_p^T (T_{p-1}, J_{p-1}) H_p,$$

where H_p is $(J_{p-1}, J_p)-$orthogonal. H_p is the product of H_{p-1} and two unified rotations that introduce zeros in position $(p+3, p)$ and $(p+2, p)$. We write H_p as

$$H_p = H_{p-1} G_{q,p+3} G_{p+1,p+3},$$

where $q \in \{p+1, p+2\}$ and $G_{q,p+3}$ is a unified rotation in the $(q, p+3)$ plane. After $n-4$ steps, the 4×4 bulge is at the bottom right corner of T_{n-4}. We just need to apply the zeroing process once more to obtain the tridiagonal-diagonal pair $(T_{n-3}, J_{n-4}) = (\widetilde{T}, \widetilde{J})$ defined by (6.8) and (6.9). We have

$$\begin{aligned} H &= \Pi_{i=0}^{i=n-3} H_i, \\ \widehat{J} &= H^T J H, \\ \widehat{T} &= H^T T H. \end{aligned} \tag{6.11}$$

Hyperbolic rotations are not orthogonal and therefore may be numerically unstable. We aim to use as few of them as possible. This goal can be achieved if we use at most one hyperbolic rotation per step. At step $p \leq n-3$ of the tridiagonalization process, we need to introduce zeros in positions $(p+3, p)$ and $(p+2, p)$, we need to compute a 3×3 matrix $\widetilde{H}_p$ such that $\widetilde{H}_p^T T(j+p+2 \colon j+p+3, j) = \rho e_1$. We apply first all possible orthogonal rotations and if necessary we finish the zeroing by applying a hyperbolic rotation. With this zeroing strategy, we apply at most $n-2$ hyperbolic rotations during the bulge chasing and tridiagonalization process. Moreover, this zeroing strategy is also described in Algorithm 4.4 for vectors in $\mathbb{R}^n$ in Section 4.2.4.

Bojanczyk, Brent and Van Dooren [9] noticed that how hyperbolic rotations are applied to a vector is crucial to the stability of the computation. In our implementation we use the mixed application of hyperbolic rotations as in Section 4.2.4 (see [9], [11] for a detailed description). Tisseur [74] showed that the residual $\|H^T T H - \widehat{T}\|/\|T\|/\|H\|^2$ can be much smaller when one applies the hyperbolic rotations in a mixed way rather than in a direct way.

In [74], Tisseur analyzes the tridiagonalization of a symmetric matrix with respect to a signature matrix. The bulge chasing process is a tridiagonalization. If at a step j of the tridiagonalization process (Algorithm 6.3), a hyperbolic matrix does not exist or its condition number is too large, then we can apply a random unified rotation on the first two rows and column. We then have nonzero entries at the positions $(3, 1)$ and $(1, 3)$ and we can restart the tridiagonalization and chase the bulge with $j-2$ unified rotations until we reach the new j^{th} column [74].

If at the k-th HZ iteration, the first column of the shifted matrix $x = p(JT)e_1$ is isotropic, that is $\langle x, x \rangle_J = 0$ or the condition number of the hyperbolic matrix such that $Qx = \rho e_1$ is too large, then one can apply a random shift and hope that

this technique prevents a breakdown. Assume now that x_1, the first column of the shifted matrix at the first HZ iteration is not isotropic and that the bulge chasing process is accomplished successfully. Then, if the following first columns x_k of the shifted matrix at the k-th HZ iteration ($k \geq 2$) are not isotropic, then all the corresponding bulge chasing processes are accomplished successfully. The reason is that 2×2 hyperbolic matrices preserve the modulus of the indefinite scalar product. Thus, if $\langle x_k, x_k \rangle_J \neq 0$ for all $k \geq 2$ and the first bulge chasing process is successful, then there will be no major breakdown during the HZ iterations.

Moreover, during the bulge chasing process, we need to check for deflations that may occur before the process is completed. This is particularly important if the last rotation at a given step k is hyperbolic. Let $x = \begin{bmatrix} t_{k+1,k} & t_{q,k} \end{bmatrix}^T$ with $q = k+2$ or $q = k+3$ and $\widetilde{J} = \pm\text{diag}(1, -1)$. During our numerical tests, we noticed that $\|x\|_2$ can be very small. If $|x_1| \approx |x_2|$, the hyperbolic transformation mapping x to $\pm(x^T Jx)e_1$ has a large condition number, therefore affecting the numerical stability of the process. Hence it makes sense to deflate before applying the hyperbolic transformation rather than after. In our implementation, we chose to set x to 0 if $\|x\|_2 \leq \epsilon \|T\|_2$.

6.5 Pseudocodes

The following algorithm describes this process and the computation of $\widetilde{H}_p$ (See Section 6.3). The shifts are analyzed in Section 6.6 where we compare several shifting strategies. Thus, in the next algorithm, we assume that we have the shifts and the first column of the shifted matrix. We now give two algorithms in order to describe the HZ implementation. We start by the implementation of a single HZ step and we carry on with the HZ algorithm for a tridiagonal-diagonal pair.

Algorithm 6.3 *Given an unreduced symmetric tridiagonal $n \times n$ matrix T and a signature matrix J, this algorithm applies an implicit double HZ step to the pair (T, J) with shifts $\omega_1,\ \omega_2$. It returns a tridiagonal diagonal pair $(\widetilde{T}, \widetilde{J})$ and a matrix H such that $(\widetilde{T}, \widetilde{J}) = H^T(T, J)H$.*

Set $H = I$

Compute $s = \omega_1 + \omega_2$ and $t = \omega_1\omega_2$ (see Section 6.6).

Compute $x = p(JT)e_1 = ((JT)^2 - tJT + dI)e_1$.

Apply Algorithm 4.4 to x and obtain Q such that $Q^T x = \rho e_1$

Compute $Q^T T(1\colon 4, 1\colon 4)Q$ and update $H(:, 1\colon 3) = HQ$.

$j = 1$

While $p \leq n - 3$

  Set $x = [\, t_{j+1,j} \quad t_{j+2,j} \quad t_{j+3,j} \,]^T$

  Apply Algorithm 4.4 to x and $J(j + 1 : j + 3)$

  and obtain Q such that $Q^T x = \rho e_1$.

  Compute $Q^T T(j + 1\colon j + 3, j + 1\colon j + 3)Q$.

  Update $H(2\colon j + 3, j + 1 : j + 3) = H(2\colon j + 3, j + 1 : j + 3)Q$.

  $j = j + 1$

end

Construct unified rotation $G(n - 1, n)$ to zero out $T(n - 1, n)$.

Update $T = G_{n-1,n}^T T(n - 1 : n, n - 1 : n)G_{n-1,n}$.

Update $H(:, n - 1 : n - 2) = H(:, n - 1 : n - 2)G_{n-1,n}$.

Algorithm 6.4 *Given an unreduced symmetric tridiagonal $n \times n$ T_1, a signature matrix J_1 and a tolerance ϵ, this algorithm computes a block diagonal matrix T_2 with 1×1 and 2×2 blocks, a signature matrix J_2 and H such that $T_2 = H^T T_1 H$ and $J_2 = H^T J_1 H$.*

Set $H = I$, $T_2 = T_1$, $J_2 = J_1$ and $p = 1$

while $n > 2$

Set $k = \min(n, 4)$, $q = k + p - 1$

Apply a double implicit step (Algorithm 6.3) on $(T_2(p:q, p:q), J_2(p:q))$

Update $H(:, p:n) = H(:, p:n)Q$ where Q is return by Algorithm 6.3

for $i = p:n$

if $|t_{i,i-1}| \leq \epsilon(|t_{i-1,i-1}| + |t_{ii}|)$

Set $t_{i,i-1} = 0$, $t_{i-1,i} = 0$

end

end

% Get smallest n such that $t_{n,n-1}$ or $t_{n-1,n-2}$ is non-zero

while or $(t_{n,n-1} = 0, t_{n-1,n-2} = 0)$

if $T_{n,n-1} = 0$

$n = n - 1$

elseif $t_{n-1,n-2} = 0$

$n = n - 2$

end

if $n \leq 2$

break

end

end

Get biggest p such that $T(p:n, p:n)$ is unreduced

end

In Algorithm 6.3, the tolerance ϵ is usually u the unit roundoff and the deflation is numerically allowed since rounding errors of order $u\|T\|$ are present during the computations [31].

6.6 Shifting Strategies

The shifts ω_1, ω_2 in Section 6.3 (Equation (6.10)) are based on the eigenvalues $\lambda_1,\ \lambda_2$ of the bottom right corner 2×2 subpencil

$$T(n-1\colon n, n-1\colon n) - \lambda J(n-1\colon n, n-1\colon n).$$

The *Francis* shift consists of taking $\omega_1 = \lambda_1$ and $\omega_2 = \lambda_2$ [31]. The *Wilkinson* shifts correspond to taking $\omega_1 = \omega_2 = \widetilde{\lambda}$, where $\widetilde{\lambda}$ is the nearest eigenvalue to $\sigma_n t_{nn}$. We consider three shifting strategies:

1. Francis shifting strategy, where Francis shifts are used exclusively,

2. "mix 1" shifting strategy, where Francis shifts are used when

   $$J(n-1\colon n, n-1\colon n) = \pm\mathrm{diag}(1,-1)$$

   and Wilkinson shifts are used when

   $$J(n-1:n, n-1:n) = \pm I,$$

3. "mix 2" shifting strategy, described below, that is based on the eigenvalues of the bottom right corner 3×3 subpencil.

The first two shifting strategies (Francis and "mix 1") are commonly used in eigenvalue algorithms. We give some justifications for our third shifting strategy choice "mix 2". "mix 2" uses a double shift even though the shifts are the eigenvalues of the bottom right corner 3×3 subpencil. The reason is that once the iteration is converging the eigenvalues of the bottom right corner 3×3 subpencil are better approximations to the matrix eigenvalues than eigenvalues of the bottom right corner 2×2 subpencil. Furthermore, it allows us to have a heuristic criteria to "guess" if the next eigenvalue that appears in the Hessenberg

form after deflation is complex or real. At the end of this section, we present numerical experiments that show the number of HZ iterations is less using the "mix 2" shifting strategy than using the other two shifting strategies.

In the rest of this section, we describe first the shifting strategy "mix 2" then we compare numerically the three shifting strategies.

Let

$$\begin{aligned} q(z) &= \sigma_{n-2}\sigma_{n-1}\sigma_n \det(T(n-2\colon n, n-2\colon n) - zJ(n-2\colon n, n-2\colon n)), \\ &= \sigma_{n-2}\sigma_{n-1}\sigma_n(z^3 + a_2 z^2 + a_1 z + a_0), \end{aligned}$$

where

$$\begin{aligned} a_2 &= -\mathrm{trace}(J(n-2\colon n, n-2\colon n)T(n-2\colon n, n-2\colon n)), \\ a_0 &= -\det(J(n-2\colon n, n-2\colon n)T(n-2\colon n, n-2\colon n)), \\ a_1 &= \sigma_{n-2}\sigma_n t_{nn} t_{n-2,n-2} + \sigma_{n-1}\sigma_n t_{nn} t_{n-1,n-1} + \sigma_{n-2}\sigma_{n-1} t_{n-2,n-2} t_{n-1,n-1} \\ &\quad -\sigma_{n-1}\sigma_n t_{n,n-1}^2 - \sigma_{n-2}\sigma_{n-1} t_{n-1,n-2}^2. \end{aligned}$$

We know that q has at least one real root. We have that

$$q'(z) = \sigma_{n-2}\sigma_{n-1}\sigma_n(3z^2 + 2a_2 z + a_1)$$

and that $q' \geq 0$ or $q' \leq 0$ if the determinant $\Delta = 4(a_2^2 - 3a_1) \leq 0$. In this case q has two complex conjugate roots. Now, if $\Delta > 0$, let ξ_k, $k = 1, 2$ with $\xi_1 \leq \xi_2$ be the two distinct real roots of q'. If $q(\xi_k) = 0$ then, q has two real distinct roots (one simple and one double). If $q(\xi_1)q(\xi_2) < 0$ then q has three distinct real roots. Otherwise, if $q(\xi_1) < 0$ or $q(\xi_2) > 0$, then q has one real root and two conjugate complex roots. The roots of a scalar cubic polynomial can be obtain explicitly by using Vieta's substitution and Cardan's formula. A practical method to obtain the roots of q is to apply the HZ algorithm on $(\widetilde{T}, \widetilde{J})$, where

$$\widetilde{T} = T(n-2\colon n, n-2\colon n), \quad \widetilde{J} = J(n-2\colon n, n-2\colon n),$$

by using the first two shifting strategies. Since the size (3×3) of the problem is small, there is no difference in practice between the Francis and "mix 2" shifting strategies. They need to perform one iteration on the 3×3 problem for a deflation to occur. We obtain $Q \in \mathbb{R}^{3\times 3}$ such that $\widehat{T} = Q^T \widetilde{T} Q$ is block diagonal and $Q^T \widetilde{J} Q \in \text{diag}_3^k(\pm 1)$. From $\widehat{T}$, we obtain easily the eigenvalues μ_1, μ_2 and μ_3 of $(\widetilde{T}, \widetilde{J})$. The shifting strategy "mix 2" consists of applying a double Wilkinson shift if all μ_k are real, the shift being μ_j the nearest eigenvalue to $\sigma_n t_{nn}$. If q has complex conjugate roots then $\widetilde{T}$ has one of the following forms:

$$\begin{bmatrix} \times & & \\ & \times & \times \\ & \times & \times \end{bmatrix}, \tag{6.12}$$

$$\begin{bmatrix} \times & \times & \\ \times & \times & \\ & & \times \end{bmatrix}. \tag{6.13}$$

If $\widetilde{T}$ has the structure in (6.12), then we apply a double Francis shift with complex shifts $\mu_k, \overline{\mu_k}$, otherwise if $\widetilde{T}$ has the structure in (6.13), we use a double Wilkinson shift with real shift μ_k, the only real root of q. Algorithm 6.5 describes the shifting strategy "mix 2".

Algorithm 6.5 *Given $T = (t_{ij})$, an unreduced tridiagonal symmetric $n \times n$ matrix, a signature matrix $J = \text{diag}(\sigma_k)$ and a tolerance parameter ϵ, this algorithm chooses a shift for (T, J) described by the method "mix 2".*

Set $\widetilde{T} = T(n-2:n, n-2:n)$, $\quad \widetilde{J} = J(n-2:n, n-2:n)$.
Compute $\widehat{T} = Q^T \widetilde{T} Q$, block diagonal with $\widehat{J} = Q^T \widetilde{J} Q \in \text{diag}_3^k(\pm 1)$.
Compute the eigenvalues μ_k, $k = 1, 2, 3$ of $(\widehat{T}, \widehat{J})$.
if $\sigma_n = \sigma_{n-1} = \sigma_{n-2}$
    Apply a double Wilkinson shift with μ such that

$|\sigma_n t_{nn} - \mu| = \min_{k=1,2,3} |\sigma_n t_{nn} - \mu_k|$

else

if $\Im(\mu_k) = 0,\ k = 1, 2, 3$

if $|\widehat{t}_{32}| \leq \epsilon(|\widehat{t}_{22}| + |\widehat{t}_{33}|)$

Apply a double Wilkinson shift with μ such that

$|\sigma_n t_{nn} - \mu| = \min_{k=1,2,3} |\sigma_n t_{nn} - \mu_k|$

else apply a double Francis shift with μ and $\tilde{\mu}$ such that

$|\sigma_n t_{nn} - \mu| = \min_{k=1,2,3} |\sigma_n t_{nn} - \mu_k|$

$|\sigma_n t_{nn} - \tilde{\mu}| = \min_{k=1,2,3} |\sigma_n t_{n-1,n-1} - \mu_k|$

end

else if $|\widehat{t}_{32}| \leq \epsilon(|\widehat{t}_{22}| + |\widehat{t}_{33}|)$

Apply a double Wilkinson shift with μ such that

$|\sigma_n t_{nn} - \mu| = \min_{k=1,2,3} |\sigma_n t_{nn} - \mu_k|$

else Apply a double Francis shift with $\mu \in \mathbb{C} \setminus \mathbb{R}$ and $\bar{\mu}$

the eigenvalues of $(\widehat{T}, \widehat{J})$.

end

end

end

We used random symmetric tridiagonal matrices generated with MATLAB's `randn`. J was obtained by MATLAB random permutation generator `randperm`:

```
J = diag((-1).^randperm(n)).
```

For each value of n, we used 100 tridiagonal-diagonal pairs. The total number of iterations on average and the number of iterations per eigenvalue on average are shown in Table 6.1 and Table 6.2, respectively. We see that the Francis shifting strategy and the "mix 1" shifting strategy are equivalent in the sense

that they require more or less the same amount of iterations. The Francis shifting strategy seems to require less iterations than the mixed Francis-Wilkinson shifting strategy. The last shifting strategy "mix 2" performs better than the two other methods. It is due to the fact that the eigenvalues of the bottom right corner 3×3 subpencil are better approximations to the eigenvalues of (T, J) than the eigenvalues of the bottom right corner 2×2 subpencil. The disadvantage of the "mix 2" shifting strategy is that it is slightly more expensive to implement and it fails if the bottom right corner 3×3 subpencil has a non-trivial Jordan block. In Chapter 7, we present numerical experiments with different type of matrices in order to see the behavior of the shifting strategies in the case of ill conditioned problems.

6.7 Flops Count and Storage

The first step of the HZ algorithm described in Subsection 6.1.1 requires $n^3/3$ flops for the LDL^T decomposition of B and there is an additional cost of n^3 flops to update A. If instead, we use the QR algorithm to diagonalize B, we need approximately $(2/3 + 5)n^3$ flops and an additional $2n^3$ flops to update A.

The second step, that is, the reduction to a tridiagonal-diagonal pair requires approximately $(1/3)n^3 + (1/2)n^2$ flops using Tisseur's algorithm [74]. Finally, the HZ algorithm on a tridiagonal-diagonal pair involves in average $10(n^2 + n) + 5n^2$ operations for the bulge chasing and an additional $10n$ flops to compute the shifts with Algorithm 6.5.

In Table 6.3, we compare the number of floating point operations in the HZ and QZ algorithm. The first step in the QZ algorithm is the Hessenberg-triangular reduction. Then, we apply the QZ algorithm to compute the real generalized Schur decomposition. This algorithm is presented in detail in [53] and in [31],

Table 6.1: Average number of iterations for each shifting strategy.

n	Francis	mix 1	mix 2
10	14.375	13.575	11.7
50	75.55	75.25	64
100	152.225	153.575	128.3
150	228.6	231.225	193.72
200	306.75	308.55	259.65
300	462.525	467.375	389.62
400	614.5	623.05	519.72

Table 6.2: Average number of iterations per eigenvalue for each shifting strategy.

n	Francis	mix 1	mix 2
10	1.44	1.36	1.17
50	1.51	1.5	1.28
100	1.52	1.53	1.28
150	1.52	1.54	1.29
200	1.53	1.54	1.3
300	1.54	1.56	1.3
400	1.54	1.56	1.3

where all the flops counts are available. In total, the QZ algorithm requires $33n^3$ flops whereas the HZ algorithm requires $16.4n^3$ if we use an LDL^T factorization or $23n^3$ if we apply a symmetric QR algorithm.

We now focus on storage. We need n^2+n size vector to store the pair (A, B) and a vector size n^2 for the hyperbolic matrix. During the HZ iteration, we need $2(2n-1)+2n$ size vector to store the two tridiagonal-diagonal pairs and an additional 5×5 workspace. In comparison, the QZ algorithm requires n^2+n size vector to store the pair (A, B) and a vector size $2n^2$ for the two orthogonal matrices Q and Z. An additional n^2+n is required to store the Hessenberg-triangular pair. The storage saved with the HZ algorithm is of order $2n^2$.

Table 6.3: Comparison of the number of floating point operations in the HZ and QZ algorithms.

Step	QZ	HZ	
1	$9n^3$	QR	LDL^T
		$8n^3$	$(4/3)n^3$
2		$(1/3)n^3$	
3	$26n^2$	$15n^2$	

6.8 Eigenvectors

We consider the matrices T_2 and J_2 returned by Algorithm 6.4. T_2 is block diagonal. We have to consider two cases depending on the size of the blocks. For a 1×1 block in the i^{th} position, the corresponding eigenvector is just e_i.

For a 2×2 block, with real or complex eigenvalues we need to solve the equation

$$T_2(i\colon i+1, i\colon i+1)y = \lambda_i J(i\colon i+1, i\colon i+1)y. \tag{6.14}$$

The matrix $T_2(i\colon i+1, i\colon i+1) - \lambda_i J(i\colon i+1)$ has rank 1. Thus, (6.14) has a linear subspace of dimension 1. We rewrite (6.14) as $Ax = 0$ with $A = (a_{ij})_{1\le i,j\le 2}$. We get two expressions for the same linear subspace which is spanned by

$$\begin{aligned} x_1 &= [\,-a_{12} \quad a_{11}\,]^T, \\ x_2 &= [\,-a_{22} \quad a_{21}\,]^T, \\ x_2 &= \rho x_1. \end{aligned} \tag{6.15}$$

Although x_1 and x_2 are linearly dependent in theory, in finite arithmetic one of them can be computed more accurately than the other one. In the following paragraphs, we present the chosen method for computing the eigenvector.

For real eigenvalues (6.14) gives us two expressions of y for each eigenvalue.

We can choose y that minimizes the residual

$$\|(T_2(i\colon i+1, i\colon i+1) - \lambda_i J(i\colon i+1, i\colon i+1))y\|_2$$

for a normalized eigenvector $\|y\|_2 = 1$. Then, the eigenvector x corresponding to λ_i has zero entries except for the row i, $i+1$ and we have $x(i\colon i+1) = y$.

Complex eigenvalues come in pairs $(\lambda_i, \bar{\lambda_i})$ as the corresponding eigenvectors. We solve equation (6.14) and we get two expressions for y and two for $\bar{y}$. We can now choose y that minimizes

$$\|(T_2(i\colon i+1, i\colon i+1) - \lambda_i J(i\colon i+1, i\colon i+1))y\|_2 \text{ or } \|(T_2(i\colon i+1, i\colon i+1) - \bar{\lambda_i} J(i\colon i+1, i\colon i+1))\bar{y}\|_2.$$

Finally, to obtain the eigenvectors of the original T_1 we just need to multiply x by the accumulated transformations H.

6.9 Iterative Refinement

6.9.1 Newton's Method

The iterative refinement is done by Newton's method. In [72], Tisseur studied Newton's method in floating point arithmetic and showed how to apply iterative refinement to the GEP. We apply the Newton method to the function $f : \mathbb{K}^n \times \mathbb{K} \to \mathbb{K}^{n+1}$ defined by

$$f\left(\begin{bmatrix} x \\ \lambda \end{bmatrix}\right) = \begin{bmatrix} (A - \lambda B)x \\ \mu e_s^T x - \mu \end{bmatrix},$$

where $\mathbb{K}$ denotes $\mathbb{R}$ or the complex field $\mathbb{C}$, $\mu > 0$ and some $1 \leq s \leq n$. Then, Newton's method for the GEP is finding the zeros of f. The Jacobian matrix of f is given by

$$G\left(\begin{bmatrix} x \\ \lambda \end{bmatrix}\right) = \begin{bmatrix} A - \lambda B & -Bx \\ \mu e_s^T & 0 \end{bmatrix}.$$

Given a starting guess of the eigenpair (x_0, λ_0), the Newton iterations are defined by

$$G\left(\begin{bmatrix} x_p \\ \lambda_p \end{bmatrix}\right)\begin{bmatrix} \Delta x_{p+1} \\ \Delta\lambda_{p+1} \end{bmatrix} = -f\left(\begin{bmatrix} x_p \\ \lambda_p \end{bmatrix}\right), \tag{6.16}$$

where $\Delta x_{p+1} = x_{p+1} - x_p$ and $\Delta\lambda_{p+1} = \lambda_{p+1} - \lambda_p$.

We assume that all the convergence conditions for the Newton iterations are satisfied. Those conditions are informally that the Jacobian matrix (in the Newton iteration) is not too ill conditioned, the linear system solver is not too unstable and the starting pair (λ_0, x_0) for the iteration is a good enough approximation. Then, we have the following result on the backward error in the $\infty-$norm of a refined eigenpair $(\tilde{\lambda}, \tilde{x})$ with residuals computed in fixed precision [72, Corollary 3.5]

$$\eta(\tilde{\lambda}, \tilde{x}) \leq \gamma_n + u(3 + |\lambda|)\max\left(\frac{\|A\|_\infty}{\|B\|_\infty}, \frac{\|B\|_\infty}{\|A\|_\infty}\right), \tag{6.17}$$

where u is the unit roundoff, $\gamma_n = \frac{cnu}{(1-cnu)}$ and c is a small integer . We can expect this backward error to be small enough (of order $\tilde{c}nu$, with $\tilde{c}$ a small constant) if

$$|\lambda|\max\left(\frac{\|A\|_\infty}{\|B\|_\infty}, \frac{\|B\|_\infty}{\|A\|_\infty}\right) \leq 1.$$

If $|\lambda| > 1$, we can consider the reciprocal matrix pencil $B - \lambda^{'}A$, with $\lambda^{'} = \frac{1}{\lambda}$. If the GEP is not well balanced, that is

$$\frac{\|B\|_\infty}{\|A\|_\infty} \gg 1 \text{ or } \frac{\|B\|_\infty}{\|A\|_\infty} \ll 1,$$

we can consider the equivalent pencil

$$\gamma A - (\gamma\lambda)B,$$

with $\gamma = \frac{\|B\|_\infty}{\|A\|_\infty}$. Thus, in each case, we can change the GEP in order to obtain a small backward error for the refined eigenpair.

We now consider Newton's method and its implementation.

6.9.2 Implementation

The direct implementation of this iteration is too expensive since it requires $O(n^3)$ flops per iteration [72]. By using Tisseur's reduction to tridiagonal-diagonal form [74], we obtain the equivalent pencil

$$(T, J) = H_1^T(A, B)H_1. \tag{6.18}$$

Then, from the HZ algorithm 6.3 and 6.4, we know that there exists a matrix H_2 such that $H_2^T T H_2 = D$, $H_2^T J H_2 = \widetilde{J}$, with T block diagonal and $\widetilde{J} \in \text{diag}_k^n(\pm 1)$ and we set $H = H_1 H_2$.

In the rest of this paragraph, we describe a generalization of the method used in [72] to complex eigenpairs. By using the eigendecomposition computed by the HZ algorithm, the cost of the implementation can be reduced to $O(n^2)$ flops per iteration. After manipulating Newton's equation (6.16) and applying the same ideas as [72], we obtain the iteration

$$H^T M_p \delta_{p+1} = -H^T(A - \lambda_p B)x_p, \tag{6.19}$$

where

$$M_p = (A - \lambda_p B) - ((A - \lambda_p B)e_s + Bx)e_s^T \quad \text{and} \quad \delta_{p+1} = \Delta x_{p+1} + (\Delta\lambda_{p+1} - 1)e_s.$$

We define $w_{p+1} = H^{-1}\delta_{p+1}$, $r_p = (A - \lambda_p B)x_p$, $v_p = H^T(A - \lambda_p B)e_s + H^T B x_p$ and $d = H^T e_s$. Then, Equation (6.19) becomes

$$((D - \lambda_p \widetilde{J}) - v_p d^T)w_{p+1} = -H^T r_p. \tag{6.20}$$

To solve (6.20) we proceed as follows. We use Givens rotations to compute the orthogonal matrix E_p such that

$$E_p^T v_p = \pm ||v_p||_2 e_1.$$

The matrix $E_p^T((D-\lambda_p\widetilde{J})-v_pd^T)$ is upper Hessenberg with an extra subdiagonal and its QR factorization requires only $O(n^2)$ flops. Hence w_{p+1} in (6.20) can be obtained in $O(n^2)$ flops.

Algorithm 6.6 *Given a tolerance ϵ, a symmetric pair (A,B), H, a block diagonal D and a signature matrix $\widetilde{J}$ such that $H^TAH=D$ and $H^TBH=\widetilde{J}$ and an approximate eigenpair (λ,x) with $\|x\|_\infty=x_s=1$, this algorithm applies iterative refinement to λ and x.*

Repeat until convergence

While $\eta(\lambda,x)>\epsilon$

$v=H^T(A-\lambda B)e_s+H^TBx$

$g=H^Te_s$

Compute orthogonal matrix E such that

$E^Tv=\pm||v||_2e_1$

Compute orthogonal matrix F such that

$R=F^TE^T((D-\lambda\widetilde{J})-vd^T)$ is upper triangular

Solve $Rw=-F^TE^TH^T(A-\lambda B)x$

$\delta=Hw$

$\lambda=\lambda+\delta^Te_s$

$x=x+\delta-(\delta^Te_s)e_s$

end

The implementation of Newton's method described in the first part of this section requires $O(n^2)$ operation per iterations. Any method that necessitates a matrix-vector multiplication will requires $O(n^2)$ operations per iteration. We are now going to present a method that is based on a modified version of the Sherman-Morrison-Woodbury formula. Our first motivation is to reduce the cost

to at most one n^2 flops operation and to apply Newton's method in real arithmetic for complex eigenpair. The second aim of this method is to improve the numerical stability without increasing the cost of computation. The tridiagonal-diagonal pair (T, J) (6.18) is the most compact form that can be obtained in a finite number of steps. The HZ iterations use hyperbolic rotations that add instability to the reduction. Hence, the accumulated errors are larger when we use the eigendecomposition within the Newton iteration rather than the tridiagonal-diagonal form (T, J).

The iterations with the tridiagonal-diagonal pair (T, J) in (6.18) are given by

$$H_1^T M_p \delta_{p+1} = -H_1^T (A - \lambda_p B) x_p, \tag{6.21}$$

where

$$M_p = (A - \lambda_p B) - ((A - \lambda_p B)e_s + Bx)e_s^T \quad \text{and} \quad \delta_{p+1} = \Delta x_{p+1} + (\Delta\lambda_{p+1} - 1)e_s.$$

We define $w_{p+1} = H_1^{-1}\delta_{p+1}$, $r_p = (A - \lambda_p B)x_p$, $v_p = H_1^T(A - \lambda_p B)e_s + H_1^T B x_p$ and $d = H_1^T e_s$. Then, (6.21) becomes

$$((T - \lambda_p J) - v_p d^T)w_{p+1} = -H_1^T r_p. \tag{6.22}$$

Refining real eigenpairs: To solve (6.22) when λ_p is real, we proceed as follows. Let $QR = (T - \lambda_p J)$ be the QR factorization of the tridiagonal matrix $(T - \lambda_p J)$. This factorization can be done in $O(n)$ operations. Premultiplying (6.22) by Q^T gives

$$(R - vd^T)w_{p+1} = -Q^T H_1^T r_p, \; v = Q^T v_p. \tag{6.23}$$

Note that since λ_p approaches an eigenvalue of (T, J), R is nearly singular so we cannot use the Sherman-Morrison-Woodbury formula as it is. Let $\widetilde{R} = (R - vd^T) + vd^T + uu^T$ for some $u \in \mathbb{R}^n$ such that $\widetilde{R}$ is nonsingular. Then,

$$R - vd^T = \widetilde{R} - \begin{bmatrix} v & u \end{bmatrix} \begin{bmatrix} d^T \\ u^T \end{bmatrix}$$

and (6.23) becomes

$$\left(\widetilde{R} - \begin{bmatrix} v & u \end{bmatrix} \begin{bmatrix} d^T \\ u^T \end{bmatrix}\right) w_{p+1} = -Q^T H_1^T r_p.$$

Since $\widetilde{R}$ is nonsingular, we can use the Sherman-Morrison-Woodbury formula. This gives

$$w_{p+1} = \left(I_n + \widetilde{R}^{-1} \begin{bmatrix} v & u \end{bmatrix} C^{-1} \begin{bmatrix} d^T \\ u^T \end{bmatrix}\right) \widetilde{R}^{-1} b, \tag{6.24}$$

where

$$C = I_2 - \begin{bmatrix} d^T \\ u^T \end{bmatrix} \widetilde{R}^{-1} \begin{bmatrix} v & u \end{bmatrix} \quad \text{and} \quad b = -Q^T H_1^T r_p.$$

Note that for the choice $u = e_s$ with s such that

$$|r_{ss}| = \min_{1 \le k \le n} |r_{kk}|,$$

$\widetilde{R}$ is upper triangular with only 3 superdiagonals and any calculation of the form $\widetilde{R}^{-1} z$ cost $O(n)$ operations.

Refining complex eigenpair: We now consider the complex case. Let (λ, x) be a complex eigenpair, $\lambda = \alpha + i\beta$ for some $(\alpha, \beta) \in \mathbb{R}^2$ with $\beta \neq 0$ and $x = y + iz$ for some $(y, z) \in \mathbb{R}^{n \times 2}$ with $z \neq 0$. Separating real and imaginary parts in (6.22) yields

$$M_p \widetilde{w}_{p+1} = -\widetilde{r}_p, \tag{6.25}$$

where $M_p = \widetilde{M}_p - \widetilde{v}_p \widetilde{d}^T$ and

$$\widetilde{M}_p = \begin{bmatrix} T - \alpha_p J & \beta_p J \\ -\beta_p J & T - \alpha_p J \end{bmatrix}, \quad \widetilde{v}_p = \begin{bmatrix} \Re(v_p) & \Im(v_p) \\ -\Im(v_p) & \Re(v_p) \end{bmatrix},$$

$$\widetilde{w}_{p+1} = \begin{bmatrix} H^{-1}\Re(w_{p+1}) \\ H^{-1}\Re(w_{p+1}) \end{bmatrix}, \quad \widetilde{r}_p = \begin{bmatrix} H^{-1}\Re(r_p) \\ H^{-1}\Re(w_p) \end{bmatrix},$$

$$\widetilde{d} = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix}.$$

The first approach is similar to the real case. Let $\widetilde{M}_p = QR$ be the QR factorization of the sparse matrix $\widetilde{M}_p$. This factorization can be done in $O(n^2)$ operations. Premultiplying (6.25) by Q^T gives

$$(R - v\widetilde{d}^T)w_{p+1} = -Q^T\widetilde{r}_p, \qquad v = Q^T\widetilde{v}_p. \tag{6.26}$$

Note that since λ_p approaches an eigenvalue of (T, J), R is nearly singular. It is approaching a matrix of rank $2(n-1)$. Once more, we cannot use the Sherman-Morrison-Woodbury formula as it is. Let $\widetilde{R} = (R - v\widetilde{d}^T) + v\widetilde{d}^T + u_1u_1^T + u_2u_2^T$ for some $u_1, u_2 \in \mathbb{R}^{2n}$ such that $\widetilde{R}$ is nonsingular. Then,

$$R - v\widetilde{d}^T = \widetilde{R} - [\, v \quad u_1 \quad u_2 \,] \begin{bmatrix} \widetilde{d}^T \\ u_1^T \\ u_2^T \end{bmatrix}$$

and (6.26) becomes

$$\left(\widetilde{R} - [\, v \quad u_1 \quad u_2 \,] \begin{bmatrix} \widetilde{d}^T \\ u_1^T \\ u_2^T \end{bmatrix} \right) w_{p+1} = -Q^T\widetilde{r}_p. \tag{6.27}$$

Since $\widetilde{R}$ is nonsingular, we can use the Sherman-Morrison-Woodbury formula. This gives

$$w_{p+1} = \left(I_{2n} + \widetilde{R}^{-1} [\, v \quad u_1 \quad u_2 \,] C^{-1} \begin{bmatrix} \widetilde{d}^T \\ u_1^T \\ u_2^T \end{bmatrix} \right) \widetilde{R}^{-1}Q^T\widetilde{r}_p,$$

where

$$C = I_4 - \begin{bmatrix} \widetilde{d}^T \\ u_1^T \\ u_2^T \end{bmatrix} \widetilde{R}^{-1} [\, v \quad u_1 \quad u_2 \,].$$

Note that for the choices $u_1 = e_{s_1},\ u_2 = e_{s_2}$ with $s_1,\ s_2$ such that

$$\begin{aligned} |r_{s_1s_1}| &= \min_{1\le k\le n} |r_{kk}|, \\ |r_{s_2s_2}| &= \min_{1\le k\le n, k\ne s_1} |r_{kk}|, \end{aligned}$$

$\widetilde{R}$ is band upper triangular.

Throughout the Newton iterations, we have computed several QR factorizations of tridiagonal matrices. For a band matrix that has p subdiagonals and p superdiagonals the R factor of the QR factorization can be obtained in $O(n)$ operations if $p \ll n$. R is upper triangular with $2p$ superdiagonals. The Q factor can be obtained in $O(n^2)$ operations. But, we recall that Q is not explicitly required in order to solve the linear system within Newton's iterations. Further details on the QR factorization of a tridiagonal matrix can be found in [7].

Algorithm 6.7 *Given a tolerance ϵ, a symmetric pair (A, B), H, a tridiagonal T and a signature matrix J such that $H^T AH = T$ and $H^T BH = J$ and an approximate real eigenpair (λ, x) with $\|x\|_\infty = x_s = 1$, this algorithm applies iterative refinement to λ and x.*

Repeat until convergence

While $\eta(\lambda, x) > \epsilon$

$r = (A - \lambda B)x$

$v = H^T(A - \lambda B)e_s + H^T Bx$, $d = H^T e_s$

Compute the QR factorization of $T - \lambda J = QR$

Compute s such that $|r_{ss}| = \min_{1 \le k \le n} |r_{kk}|$ and set $\widetilde{R} = R + e_s e_s^T$

Apply Sherman-Morrison-Woodbury formula (6.24) to solve

$$(\widetilde{R} - \begin{bmatrix} v & e_s \end{bmatrix} \begin{bmatrix} d^T \\ e_s^T \end{bmatrix}) w = -Q^T H^T r$$

$\delta = Hw$

$\lambda = \lambda + \delta^T e_s$

$x = x + \delta - (\delta^T e_s)e_s$

end

Algorithm 6.8 *Given a tolerance ϵ, a symmetric pair (A, B), H, a tridiagonal T and a signature matrix J such that $H^T AH = T$ and $H^T BH = J$ and an approximate complex eigenpair (λ, x) with $\|x\|_\infty = x_s = 1$, this algorithm applies iterative refinement to λ and x.*

Repeat until convergence

While $\eta(\lambda, x) > \epsilon$

% Compute the residues

$r_1 = (A - \Re(\lambda)B)y + \Im(\lambda)Bz$, $r_2 = (A - \Re(\lambda)B)z - \Im(\lambda)By$

Compute the QR factorization of

$$M = \begin{bmatrix} T - \Re(\lambda)J & \Im(\lambda)J \\ -\Im(\lambda)J & T - \Re(\lambda)J \end{bmatrix} = QR$$

Compute s_1 such that $|r_{s_1 s_1}| = \min_{1 \le k \le n} |r_{kk}|$

Compute s_2 such that $|r_{s_2 s_2}| = \min_{1 \le k \le n, k \ne s_1} |r_{kk}|$

Set $\widetilde{R} = R + e_{s_1} e_{s_1}^T + e_{s_2} e_{s_2}^T$

%Compute the rank 2 updates

$v_1 = H_1^T((A - \Re(\lambda)B)e_s + B\Re(x))$, $v_2 = H_1^T(\Im(\lambda)Be_s - B\Im(x))$,

$v_3 = H_1^T(-\Im(\lambda)Be_s + B\Im(x))$, $v_4 = H_1^T((A - \Re(\lambda)B)e_s + B\Re(x))$

$$v = \begin{bmatrix} v_1 & v_2 \\ v_3 & v_4 \end{bmatrix}, \widetilde{d} = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix}, d = H^T e_s$$

$u_1 = [\, v \quad e_{s_1} \quad e_{s_2} \,]$, $u_2 = [\, \widetilde{d} \quad e_{s_1} \quad e_{s_2} \,]$

Apply Sherman-Morrison-Woodbury formula to solve

$$(\widetilde{R} - u_1 u_2^T)w = -Q^T \begin{bmatrix} H_1^T r_1 \\ H_1^T r_2 \end{bmatrix}$$

$\delta_1 = Hw(1\colon n)$, $\delta_2 = Hw(n+1\colon 2n)$

$\Re(\lambda) = \Re(\lambda) + \delta_1^T e_s$, $\Im(\lambda) = \Im(\lambda) + \delta_2^T e_s$

$\Re(x) = \Re(x) + \delta_1 - (\delta_1^T e_s)e_s$, $\Im(x) = \Im(x) + \delta_2 - (\delta_2^T e_s)e_s$

end

Algorithm 6.8 uses only real arithmetic to apply iterative refinement to a complex λ and x. But, in this case the QR factorization is more expensive than in Algorithm 6.7. Note that Algorithm 6.7 can be used for complex eigenpairs. Its advantage will be that the only operation that requires n^2 flops is one matrix-vector multiplication. For a tridiagonal-diagonal pair, eigenvectors computed in Section 6.8 might not be accurate. Thus, we can use the eigenvectors in Section 6.8 as starting approximations for the inverse iteration or the Newton iteration.

Chapter 7

Numerical Experiments with HZ and Comparisons

7.1 The HZ Algorithm

In Chapter 6, we described the HZ algorithm for symmetric GEPs. The first step is a symmetric-diagonal reduction. Then, by applying Tisseur's tridiagonalization process, the pair is reduced to a tridiagonal-diagonal pair on which we perform the HZ iterations.

The HZ algorithm can also be used to solve the standard real unsymmetric eigenvalue problem (A, I). In this case, the first step is to reduce the pair (A, I) to tridiagonal-diagonal pair (T, I). Details on this reduction were analyzed in [23], [78] and more recently in [29]. Then, the tridiagonal matrix can be transformed into an equivalent symmetric-diagonal pair $(\widetilde{T}, J)$. For the reduction of a general matrix to tridiagonal form, there is a an implementation in Fortran 77 [27] and the codes are available on the web at http://www.netlib.org/toms/710. Finally, the pair (T, I) can be transformed into an equivalent symmetric-diagonal pair

$(\widetilde{T}, J)$ as follows:

$$\sigma_1 = 1, \tag{7.1}$$

$$a_k = t_{k,k-1}t_{k-1,k},\ \tilde{t}_{k,k-1} = \sqrt{|a_k|},\ \sigma_k = \text{sign}(a_k),\ k = 2\colon n. \tag{7.2}$$

In the following sections, we present various numerical experiments. Unless otherwise stated, we use the shifting strategy "mix 2" (see Algorithm 6.5, Section 6.6) with the HZ algorithm. In the tables and figures, HZ stands for our implementation of the HZ algorithm, LR for our implementation of the LR algorithm, EA for Tisseur's implementation of the Erhlich-Aberth method and QR and QZ for MATLAB's built in implementations of the QR and QZ algorithms. The HZ and LR algorithms and the Erhlich-Aberth method are also implemented in MATLAB.

7.2 Standard Numerical Experiment

We consider test matrices generated by MATLAB's function `randn` as follows:

```
a = randn(n,1);
b = randn(n-1,1);
T = diag(a)+diag(b,1)+diag(b,-1);
J = diag((-1).^randperm(n));
```

T is a $n \times n$ symmetric tridiagonal matrix and J is a signature matrix. We generate 100 test matrices for each size $n = 100,\ 200,\ 300$ and $n = 400$. The number of HZ iterations, in each case is on average 1.3 iterations per eigenvalue and there are on average $(3/4)n$ Newton iterations. For $n = 100$, 95 eigenpairs have a backward error of order 10^{-16}, 85, 65 and 65, respectively for $n = 200$, $n = 300$ and $n = 400$, before applying the iterative refinement. These results are not surprising

since these eigenvalue problems are well conditioned: the unstructured normwise condition numbers are all between 5 and 18. In all these experiments, we find that the following ratio between the structured and unstructured normwise condition numbers

$$t = \frac{\kappa(\lambda)}{C(\lambda, \mathbb{S})}$$

satisfies $1.5 \leq t \leq 1.75$, where $\kappa(\lambda)$ is the usual Wilkinson condition number and $C(\lambda, \mathbb{S})$ is given in Corollary 4.26. Here, $\mathbb{S}$ is the class of symmetric tridiagonal matrices. The results on iterative refinement are summarized in Table 7.1. η_1 denotes the normwise backward error on average for each size, before we apply iterative refinement. η_2 denotes the normwise backward error on average, obtained after iterative refinement. On average, 80% of the eigenpairs require iterative refinement. The second and third column of Table 7.1 show that the ratio η_2/η_1 is of order 10^{-6} except for $n = 400$ where it is of order 10^{-8}. The last column shows that applying iterative refinement to an approximate eigenpair can reduce the backward error to a quantity close to machine precision.

Table 7.1: Numerical results for randomly generated tridiagonal-diagonal pairs.

n	Number of Newton Iterations	η_1	η_2
100	77	10^{-10}	1.6×10^{-16}
200	158	8×10^{-8}	9×10^{-16}
300	237	5×10^{-9}	3×10^{-15}
400	320	10^{-7}	8×10^{-15}

Similarly, we present standard tests for symmetric GEPs $Ax = \lambda Bx$ that are randomly generated by

```
A = randn(n); A=A+A';
B = randn(n); B=B+B';
```

In Table 7.2, we see that the number of Newton iterations is about two iterations per eigenvalue. The average of the largest backward errors is given in column two before iterative refinement (η_1) and in column three after refinement (η_2). We see that iterative refinement improves the backward error. The ratio η_2/η_1 is between 10^{-4} and 10^{-6}. Note that the matrix that reduces the symmetric pair into a tridiagonal-diagonal pair has a large condition number. For this reason, the backward error in column three is only of order 10^{-12} for $n = 400$.

Table 7.2: Numerical results with randomly generated symmetric pairs.

n	Number of Newton Iterations	η_1	η_2	$\kappa_2(H)$
100	184	3^{-10}	2×10^{-14}	10^3
200	411	10^{-9}	6×10^{-14}	1.8×10^4
300	649	8×10^{-8}	5×10^{-13}	3×10^4
400	879	10^{-6}	1×10^{-12}	10^5

7.3 Symmetric GEPs and Iterative Refinement

We first consider an example taken from the Harwell-Boeing Collection available from http://math.nist.gov/MatrixMarket. The matrix A is '`LUND_A`' and the matrix B is '`LUND_B`'. A and B are both indefinite. The size of the problem is $n = 147$. We have

$$\kappa_2(A) = 2.2 \times 10^8 \qquad \text{and} \qquad \kappa_2(B) = 7.4 \times 10^3.$$

The eigenvalues of A are in the region $80 \leq |\lambda_A| \leq 2.3 \times 10^8$ and those of B are in the region $0.2 \leq |\lambda_B| \leq 7.4 \times 10^3$. The eigenvalues of (A, B) are real and they are in the interval

$$200 \leq |\lambda| \leq 1.4 \times 10^6.$$

The matrix that reduces the pair (A, B) to a tridiagonal-diagonal pair has a condition number of order 150. The HZ algorithm performed 140 iterations and 154 Newton iterations were required. The largest backward error is of order 10^{-6} before and 10^{-15} after iterative refinement is applied.

In Figure 7.1, we plot in logarithmic scale the unstructured normwise backward error against the modulus of the eigenvalues. The dashed line is the value of γ_n in the expression of the bound of the backward error (6.17). We see that the iterative refinement reduces the backward error and that the bound (6.17) is satisfied.

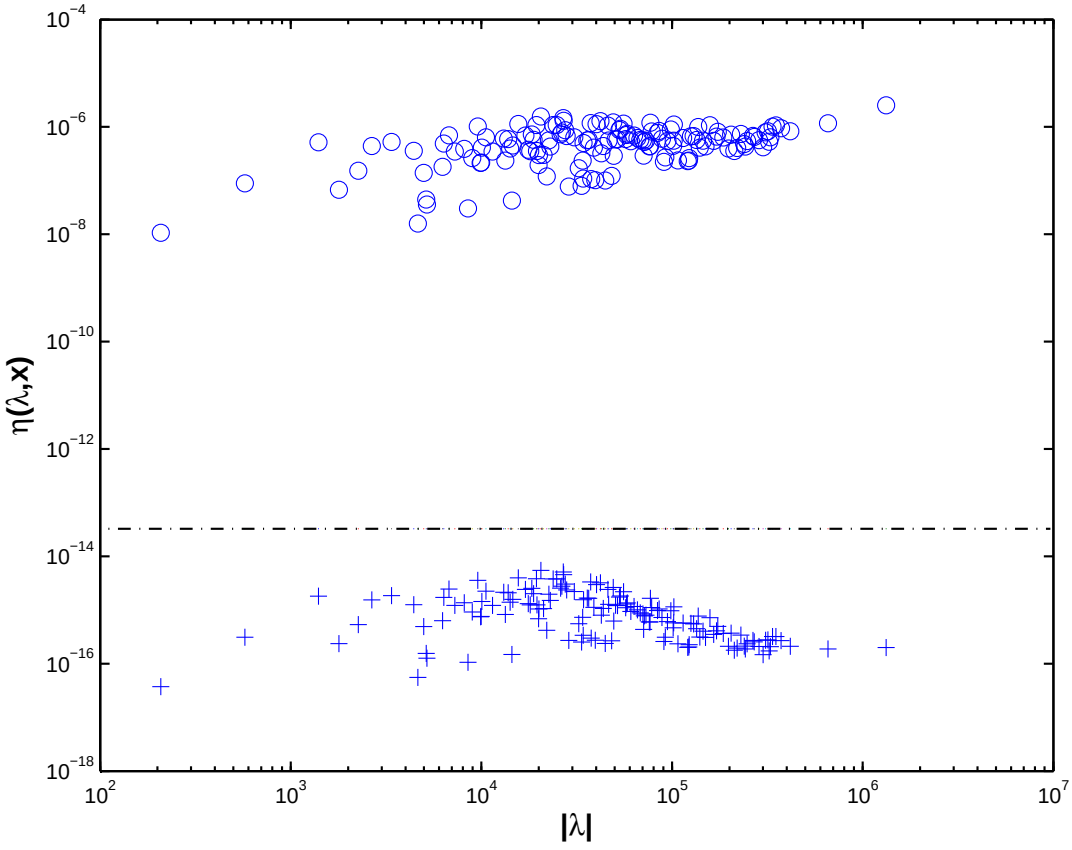

Figure 7.1: Normwise unstructured backward errors before ($\circ$) and after ($+$) iterative refinement.

7.4 HZ on Tridiagonal-Diagonal Pairs

The following examples can be found in [7]. These test matrices are not symmetric thus we use the process in (7.1)–(7.2) to obtain tridiagonal-diagonal pairs. In factored form, these tridiagonal matrices are given by

$$T = D^{-1}\text{tridiag}(1, \alpha, 1),\ D = \text{diag}(\delta),\ \alpha, \delta \in \mathbb{R}^n.$$

$$\begin{aligned}
&\text{Test 1}: && a_k = k(-1)^{\lfloor k/8 \rfloor},\ \delta_k = (-1)^k/k,\ k = 1{:}\,n,\\
&\text{Test 2}: && a_k = 10(-1)^{\lfloor k/8 \rfloor},\ \delta_k = (-1)^{\lfloor k \rfloor},\ k = 1{:}\,n,\\
&\text{Test 3}: && a_k = k,\ \delta_k = n - k + 1,\ k = 1{:}\,n,\\
&\text{Test 4}: && a_k = (-1)^k,\ \delta_k = (-1)^{\lfloor k \rfloor}20,\ k = 1{:}\,n,\\
&\text{Test 5}: && a_k = 10^{5(-1)^k}(-1)^{\lfloor k/4 \rfloor},\ \delta_k = (-1)^{\lfloor k/3 \rfloor},\ k = 1{:}\,n,\\
&\text{Test 6}: && a_k = 2,\ \delta_k = 1,\ k = 1{:}\,n,\\
&\text{Test 7}: && a_k = \frac{1}{k} + \frac{1}{n-k+1},\ \delta_k = \frac{1}{k}(-1)^{\lfloor k/9 \rfloor},\ k = 1{:}\,n,\\
&\text{Test 8}: && a_k = k^{\lfloor k/5 \rfloor \lfloor k/13 \rfloor},\ \delta_k = (n-k+1)^2(-1)^{\lfloor k/11 \rfloor},\ k = 1{:}\,n,\\
&\text{Test 9}: && a_k = 1,\ k = 1{:}\,n,\ \delta_k = 1 \text{ if} k < n/2,\ \delta_k = -1 \text{ if } k \geq n/2,\\
&\text{Test 9}: && a_k \text{ and } \delta_k \text{ are uniformly distributed in } [-0.5, 0.5].
\end{aligned} \tag{7.3}$$

The eigenvalues of these test matrices have a variety of distribution as shown in Figures 7.2, 7.3 and 7.4. We denote by λ_k the k-th eigenvalue computed in extended precision and by $\hat{\lambda}_k$ its approximation computed with either our implementation of the HZ algorithm, the LR algorithm or by Tisseur's implementation of the Erhlich-Aberth method. We compute the relative error for the test matrices 1–10 with $n = 100$ (Table 7.4) and $n = 150$ (Table 7.5).

The largest eigenvalue condition number for these test matrices are shown in Table 7.3. They vary between 2 (test 4) and 10^{10} (test 5) for $n = 100$ and they

are slightly larger for $n = 150$. Table 7.4 shows that the relative error on the computed eigenvalues increases with the condition number. The approximations obtained with the Erhlich-Aberth method are relatively accurate whereas the ones returned by the LR algorithm have poor accuracy. The HZ algorithm has an intermediate accuracy but for the test matrix 5 with $n = 150$ it fells to return an acceptable approximation. The backward error with the HZ algorithm is of order 10^{-16} except for the test 5 for which it is of order 10^{-11}. These good results on the backward error are not enough to ensure a small relative error.

Table 7.3: Largest eigenvalue condition number for test matrices 1–10 with $n = 100$ and $n = 150$

Test	1	2	3	4	5	6	7	8	9	10
$\max_k(C(\lambda_k)),\ n = 100$	$3e4$	239	$4e4$	2	$1.7e10$	$4e3$	$6e2$	$4e6$	$2e2$	637
$\max_k(C(\lambda_k)),\ n = 150$	$6e4$	$6e2$	$9e4$	2	$4e10$	$9e3$	$7e2$	$1e7$	$4e2$	$5e3$

An HZ iteration requires approximately $80n$ operations per iteration whereas the EA iteration in [7] necessitates approximately $57n$ operations per iteration. Thus, an HZ iteration requires 1.33 times more operations than an EA iteration. In Table 7.6, we compare the number of iterations between the Erhlich-Aberth method and the HZ algorithm. For $n = 150$, the ratio between the number of iterations between the Erhlich-Aberth method and the HZ algorithm lies between 1.4 and 14. The large number of iterations in Table 7.6 for the Erhlich-Aberth method is due to the quality of the starting approximations of the eigenvalues. We illustrate this fact in the next numerical experiment by changing the starting approximations of the eigenvalues for the EA method. Another disadvantage of the Erhlich-Aberth method is the fact that it uses complex arithmetic and as a result, it does not preserve the symmetry of the spectrum. We illustrate this fact in Section 7.5.

Table 7.4: Largest relative error of the computed eigenvalues for test matrices 1–10 with $n = 100$.

Test	1	2	3	4	5	6	7	8	9	10
HZ	6.8e-14	3e-14	5e-15	4e-15	1e-5	9e-13	1e-14	1e-14	3e-15	1e-14
LR	2.9e-13	4e-11	1e-14	7.4e-9	5e7	9e-10	1e-5	2e-10	6e-7	1e-8
EA	5.8e-16	2e-16	5e-16	1.9e-16	1e-10	4e-14	6e-16	6e-16	2e-15	2e-14

Table 7.5: Largest relative error of the computed eigenvalues for test matrices 1–10 with $n = 150$.

Test	1	2	3	4	5	6	7	8	9	10
HZ	6e-13	3e-12	3e-14	2e-15	1.3	6e-13	2e-12	2e-13	2e-14	2e-11
LR	2.9e-12	9e-9	3e-15	5e-9	2e2	2e-10	2e-7	8e-8	2e-8	2e-6
EA	3e-16	2e-14	2e-16	1e-16	2e-7	1e-13	7e-16	3e-16	2e-15	3e-16

The eigenvalues computed with HZ can be used as starting approximations to the Erhlich-Aberth iteration. This can be viewed as an iterative refinement of the eigenvalues only. The Erhlich-Aberth iteration fails to converge for test 5 and $n = 100, 150$: in this case, the eigenvalues computed with the HZ algorithm are poor approximations of the exact eigenvalues, which explains the non-convergence of the Erhlich-Aberth iterations. For the test matrices 1 to 4 and 7 to 9, we obtain a relative error of order 10^{-16} with at most two Erhlich-Aberth iterations per eigenvalue. For the test matrix 6, the relative error is of order 10^{-13}, with a single Erhlich-Aberth iteration per eigenvalue. This represents a reduction of 85% in the total number of iterations compared to the case if we had used the Erhlich-Aberth method only.

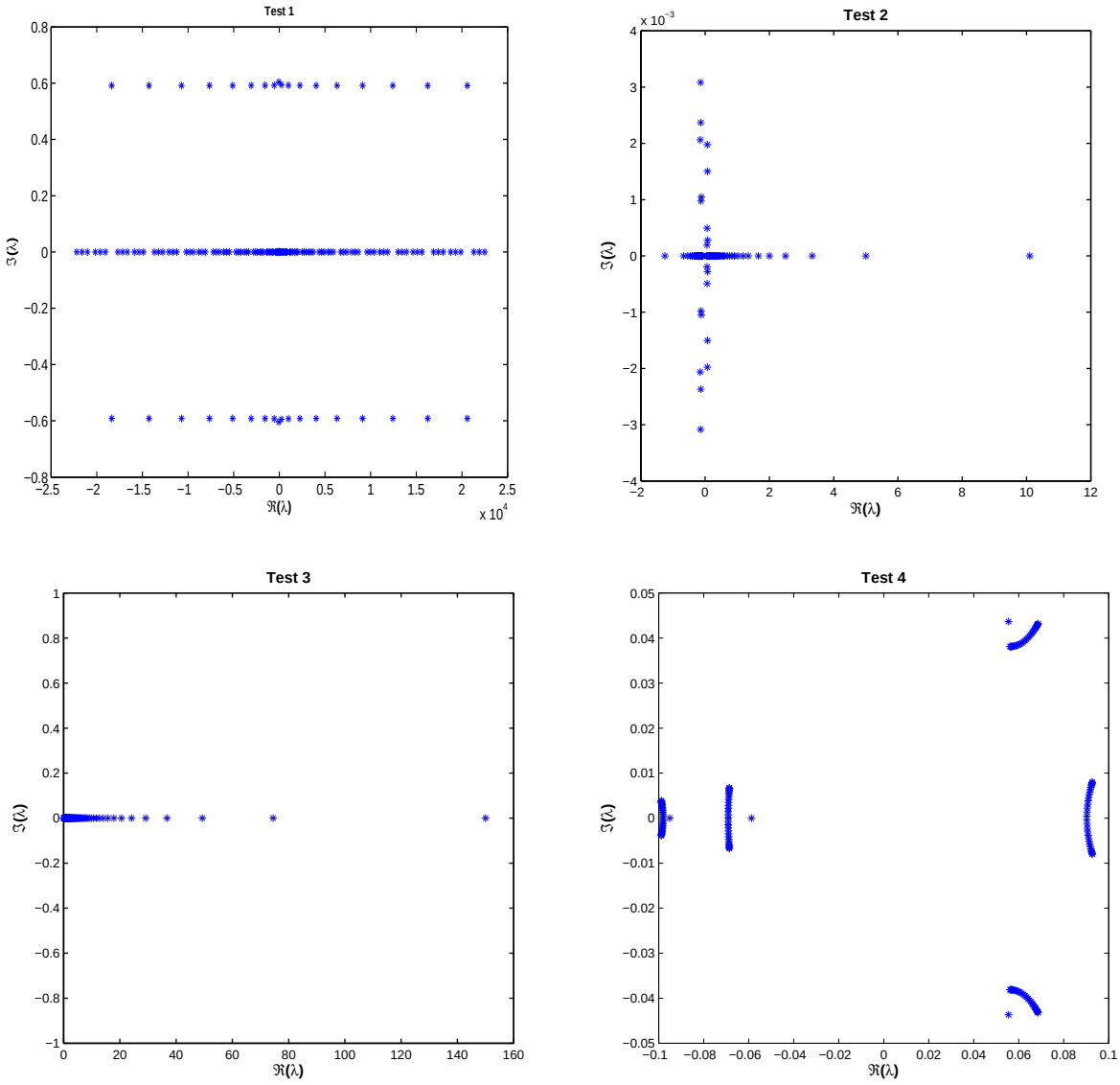

Figure 7.2: The eigenvalues of tests 1 to 4 in the complex plan for $n = 150$.

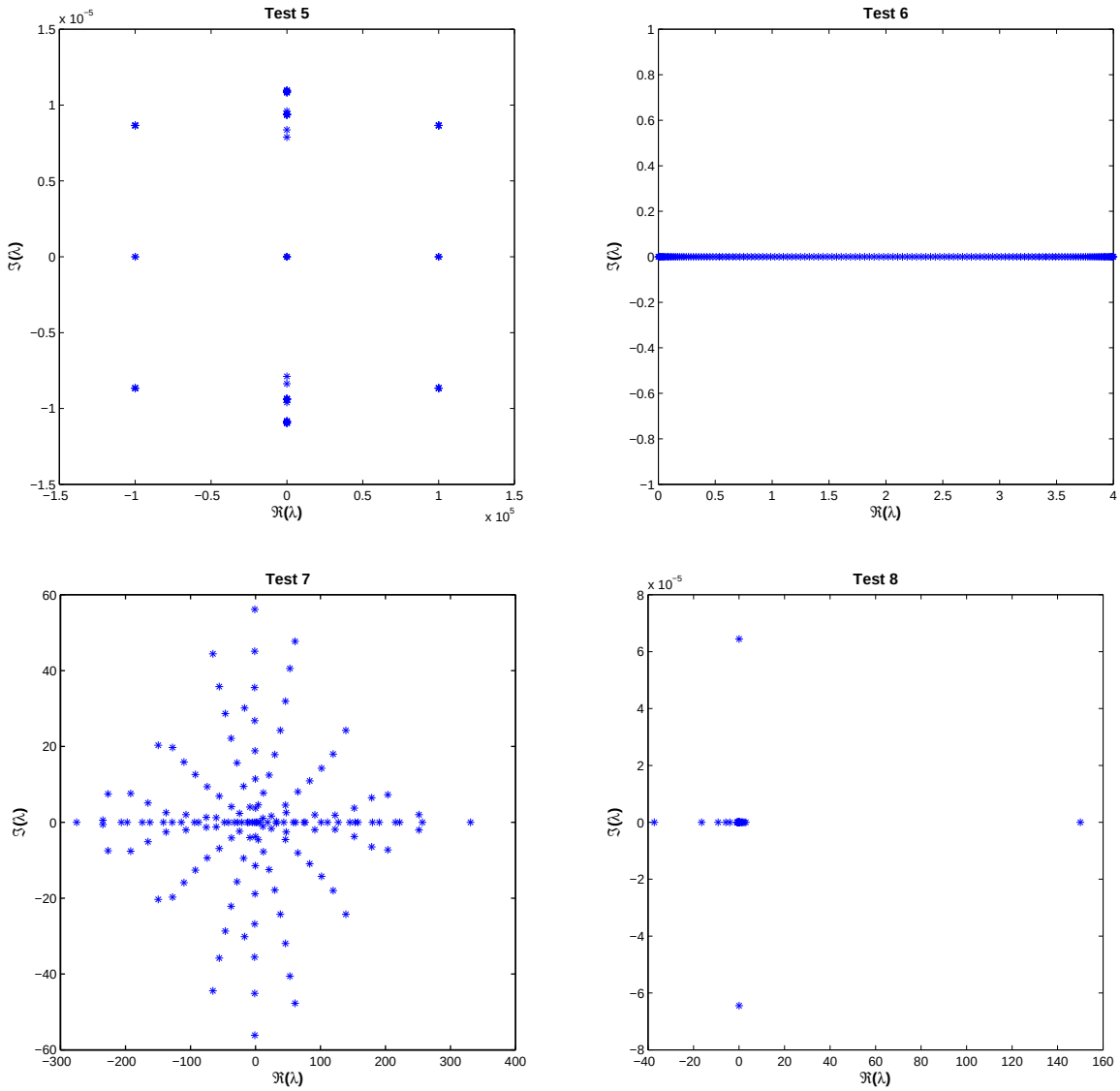

Figure 7.3: The eigenvalues of tests 5 to 8 in the complex plan for $n = 150$.

Table 7.6: Number of HZ iterations and Erhlich-Aberth iterations, $n = 150$.

Test	1	2	3	4	5	6	7	8	9	10
HZ	190	232	148	303	664	307	268	170	289	278
EA	540	526	419	972	954	2062	765	355	4082	835

Table 7.7: Normwise backward errors for test matrices 1-10 with $n = 150$.

Test	1	2	3	4	5	6	7	8	9	10
$\max_i(\eta_1(\lambda_i, x_i))$	1.9e-12	2e-13	8e-17	2e-7	2e-4	9e-16	2e-11	1e-15	1.9e-11	3e-11
$\max_i(\eta_2(\lambda_i, x_i))$	1.8e-16	3e-15		8e-17	1e-11	2e-16	4e-17	1e-16	1.7e-16	1.6e-16
k	65	70	0	86	337	76	92	4	120	95

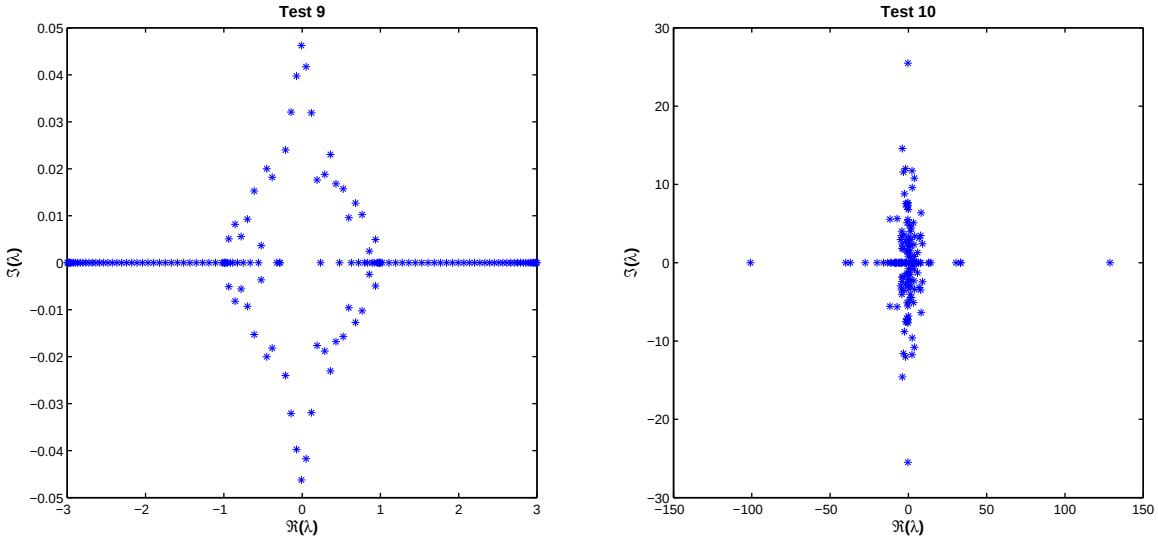

Figure 7.4: The eigenvalues of tests 9 and 10 in the complex plan for $n = 150$.

7.5 Bessel Matrices

Nonsymmetric tridiagonal Bessel matrices associated with the generalized Bessel polynomials [59] are defined by $T_a = \text{tridiag}(\beta, \alpha, \gamma)$ with

$$\begin{aligned}
\alpha_1 &= -\frac{2}{a},\ \gamma_1 = -\alpha_1,\ \beta_1 = \frac{\alpha_1}{a+1}, \\
\alpha_k &= -2\frac{a-2}{(2k+a-2)(2k+a-4)}, \quad k = 2\colon n, \\
\beta_k &= -\frac{2k}{(2k+a-1)(2k+a-2)}, \quad k = 2\colon n-1, \\
\gamma_k &= 2\frac{k+a-2}{(2k+a-2)(2k\ +a-3)}, \quad k = 2\colon n-1.
\end{aligned}$$

We carry out two experiments. In the first one, we take $n = 18$ $a = -8.5$. In this case, the condition numbers of the eigenvalues lie between 10^8 and 9×10^{12}. In the second experiment, we take $n = 60$ and $a = 12$ and in this case the condition numbers of the eigenvalue are between 4.4×10^3 and 5.9×10^{15}. The HZ algorithm performed 18 and 83 iterations to compute the eigenvalues for $n = 18$ and $n = 60$, respectively. In both cases, the largest backward error obtained with the HZ algorithm for an eigenpair is of order 10^{-16}. For an exact eigenvalue λ_0 and a corresponding approximation λ_1, we denote the relative error by $\epsilon(\lambda_0, \lambda_1) = |\lambda_0 - \lambda_1|/|\lambda_0|$. Figure 7.5 shows that the relative error decreases as the real part of the eigenvalues increases. It shows that relative to conditionning all the algorithms provide a good approximation of the eigenvalues with a real part greater than -0.06 as shown in Figure 7.6. The Bessel matrix with $n = 18$ and $a = -8.5$ is the only example for which the HZ algorithm provides the best approximations. In this case, the Erhlich-Aberth method performed 112 iterations, which is six times more than the HZ algorithm. Thus, there might be less error accumulated with the HZ algorithm. For the Bessel matrix with $n = 18$ and $a = -8.5$, we see in Figure 7.6 that the Erhlich-Aberth method does not preserve the symmetry of the spectrum.

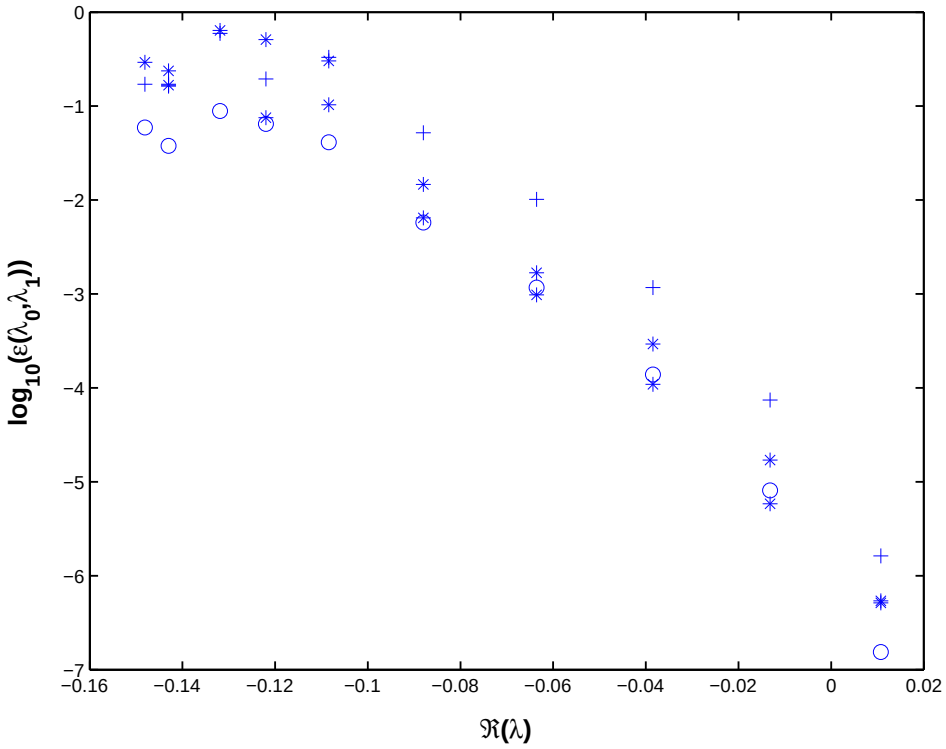

Figure 7.5: Relative errors of the eigenvalues of the Bessel matrix with $n = 18$, $a = -8.5$ computed with HZ ($\circ$), EA ($*$) and with QR ($+$).

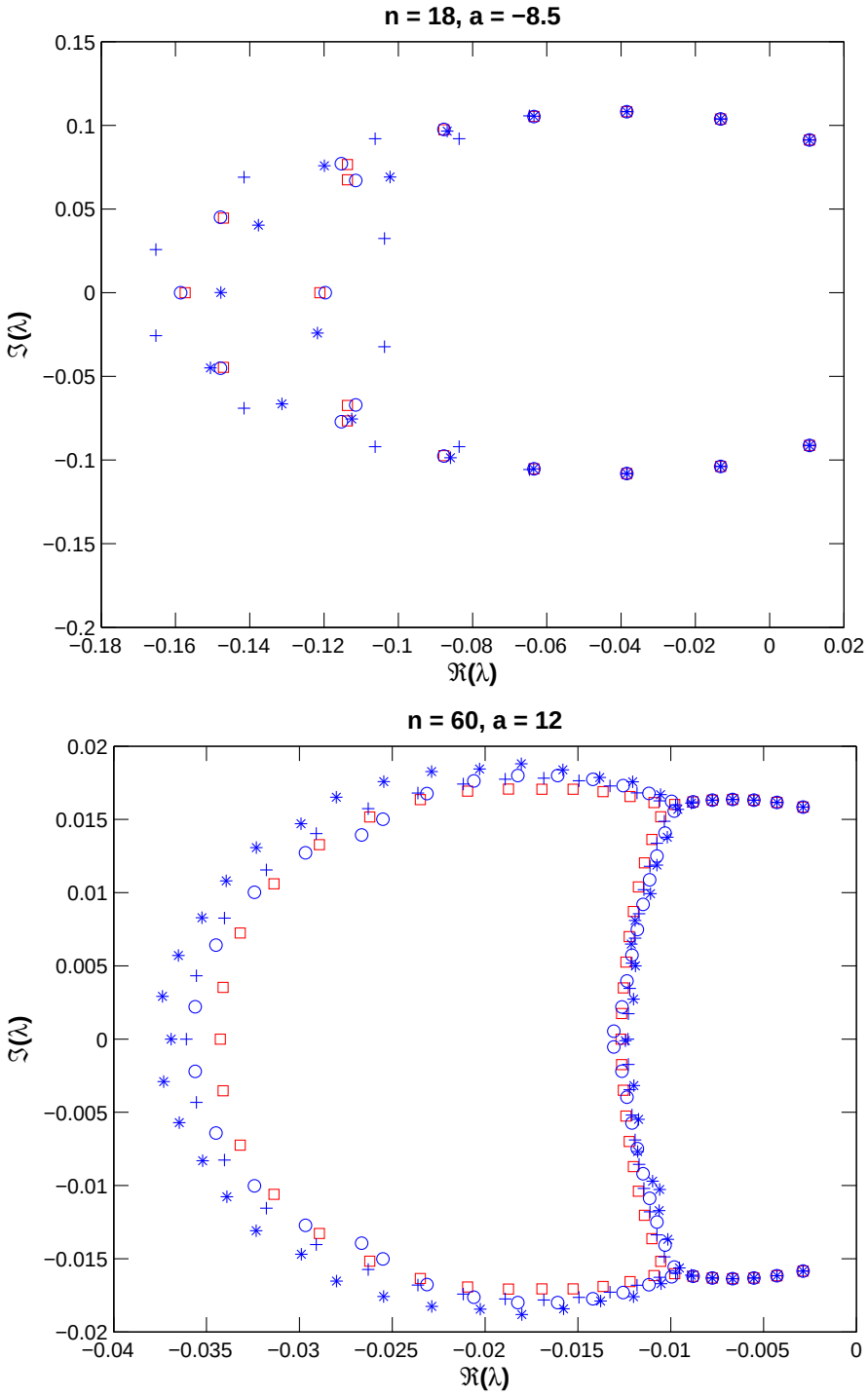

Figure 7.6: Eigenvalues of Bessel matrices computed in extended precision (□) and with HZ (∘), EA (∗) and with QR (+).

7.6 Lui Matrices

In [58], for a given $A \in \mathbb{C}^{n\times n}$, the author describes a process to build a pair (T, J) with T symmetric tridiagonal and $J \in \mathrm{diag}_n^k(\pm 1)$ such that the pencils (A, I) and (T, J) are equivalent. This process is based on the following theorem [58, Thm. 5.6].

Theorem 7.1 *Let $A \in \mathbb{C}^{n\times n}$, $p, q \in \mathbb{R}^n$, $H_n^{(k)}(A, p, q) = (h_{ij}^{(k)}) \in \mathbb{R}^{n\times n}$ where $h_{ij}^{(k)} = p^T A^{k+i+j-2} q$, $i, j = 1\colon n$. If $H_n^{(0)}$ is nonsingular and permits triangular decomposition $H_n^{(0)} = L_n J_n L_n^T$ then the following pencils are equivalent*

$$(H_n^{(1)}, H_n^{(0)}),\ (A, I),\ (T_n, J_n).$$

Here T_n is the unreduced symmetric tridiagonal matrix $T_n = L_n^{-1} H_n^{(1)} L_n^{-T}$.

In [47], Liu applies this method to an $n \times n$ Jordan block A which has a unique eigenvalue 0. The equivalent tridiagonal-diagonal pair (T, J) with

$$\begin{aligned} T &= \mathrm{tridiag}(b, a, b), \quad a \in \mathbb{R}^n,\ b \in \mathbb{R}^{n-1}, \\ J &= \mathrm{diag}(\sigma), \quad \sigma \in \mathbb{R}^n, \end{aligned}$$

is given below for $n = 5, 14$ and 28.

$$\begin{aligned} n = 5: \quad & a_k = 0,\ k = 1\colon n,\ b = [\,-1 \quad \sqrt{2} \quad -1/\sqrt{2} \quad -1/\sqrt{2}\,], \\ & \sigma = [\,1 \quad -1 \quad -1 \quad 1 \quad -1\,], \end{aligned}$$

$$n = 14: \quad a_k = \begin{cases} 1 & \text{if } k = 7, 8, \\ 0 & \text{otherwise}, \end{cases}$$

$$b_k = 1,\ k = 1\colon n-1,$$

$$\sigma = [\,1 \quad -1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1 \quad 1 \quad -1 \quad -1 \quad 1 \quad 1 \quad 1 \quad -1\,],$$

$$n = 28: \qquad a_k = \begin{cases} 1 & \text{if } k = 7, 8, 14, 14, 21, 22 \\ 0 & \text{otherwise,} \end{cases}$$

$$b_k = 1, \ k = 1{:}\, n-1,$$

$$\begin{aligned} \sigma(1:7) &= [\,1 \quad -1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1\,], \\ \sigma(8:14) &= [\,1 \quad -1 \quad -1 \quad 1 \quad 1 \quad 1 \quad -1\,], \\ \sigma(15:21) &= [\,1 \quad -1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1\,], \\ \sigma(22:28) &= [\,1 \quad -1 \quad -1 \quad 1 \quad 1 \quad 1 \quad -1\,]. \end{aligned}$$

For the Liu matrices, the HZ algorithm with the shifting strategies described in Algorithm 6.5 fails to converge. The first column x of the shifted matrix is either e_k, with $k = 1, 2, 3$ or $x^T J x = 0$. For the same reasons, the HZ algorithm does not converge with the shifting strategy that consists of Francis's shifts only. Thus our first series of experiments with the Liu matrices are with the shifting strategy that consists of using a double Wilkinson shift if $J(n) = J(n-1)$ and a double Francis shift otherwise. For $n = 5$, the spectrum is plotted in Figure 7.7. The figure on the right is a zoom of the center of the figure on the left. In Figure 7.7 (on the right), we see that the HZ Algorithm returns poor results for 4 eigenvalues. But surprisingly, it finds one zero eigenvalue (of order 10^{-17}, on the left). For $n = 14$ and 28, in Figure 7.8, the HZ algorithm returns approximations that are similar to the Erhlich-Aberth method and MATLAB's implementation of the QR algorithm `eig`.

In our second series of tests with Liu's matrices 14 and 28, we modified Algorithm 6.5 by adding a random shift. This random shift is used when the first

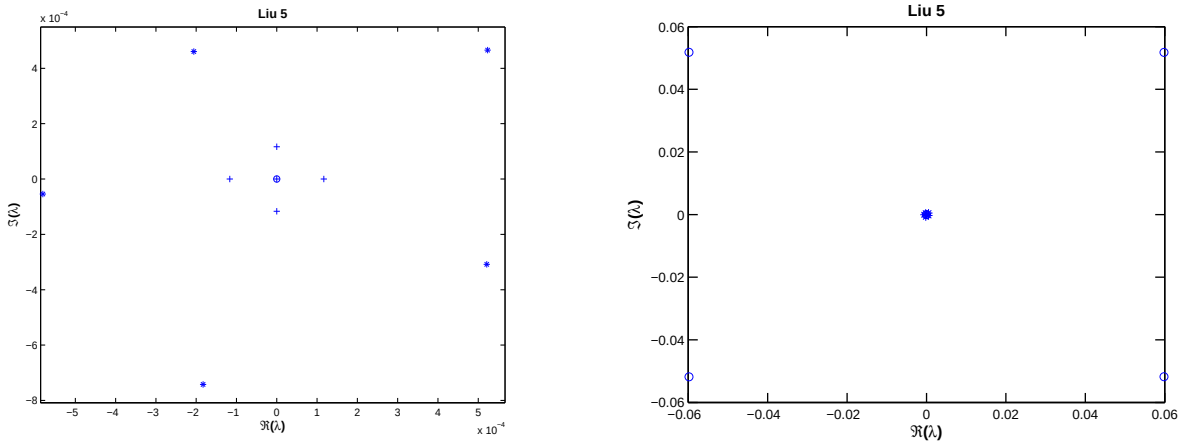

Figure 7.7: The eigenvalues of Liu's matrix 5 computed with HZ ($\circ$), EA ($*$) and QR ($+$).

column of the shifted matrix is isotropic or when it is e_k, $k = 1, 2, 3$. With this shifting strategy, the Algorithm fails to return 4 eigenvalues for $n = 14$ and 6 for $n = 28$ with any reasonable accuracy. For the other eigenvalues, in Figure 7.9, the HZ algorithm returns better approximations than the Erhlich-Aberth method or even `eig`. For $n = 28$, in Figure 7.9, the HZ algorithm returns better approximations for 10 eigenvalues.

In all these experiments with Liu's matrices, we see that the shifting strategy influences strongly the approximations. One interesting question and a practical problem to solve in the future is how to choose the optimal shifting strategy for each matrix or at each HZ iteration. Even though Algorithm 6.5 gives good approximations for the eigenvalues with a low number of iterations in most cases, it does not perform well on Liu's matrices.

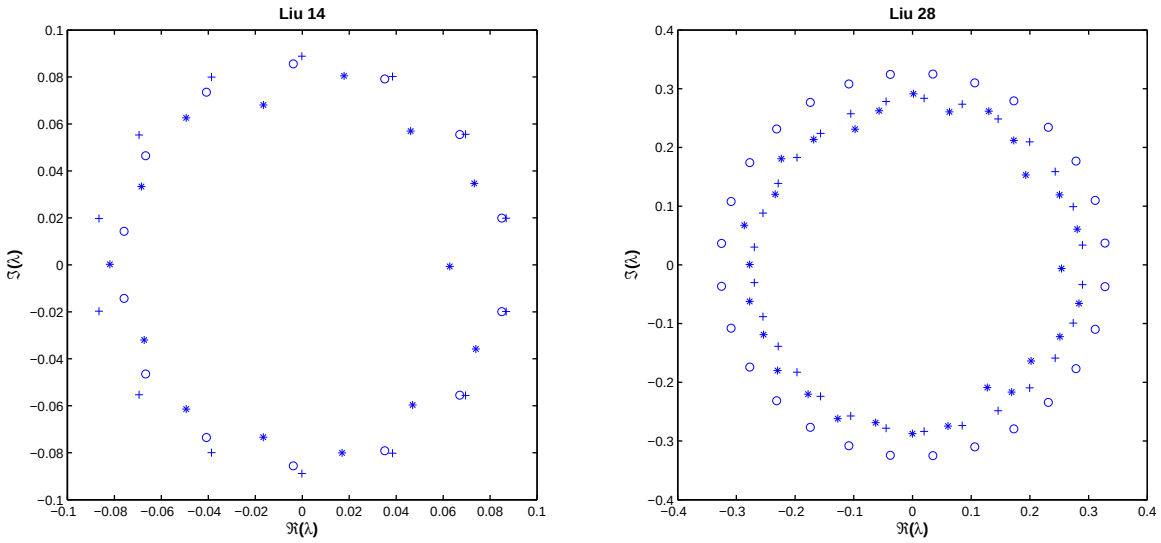

Figure 7.8: The eigenvalues of Liu's matrices 14 and 28 computed with HZ ($\circ$) using shifting strategy "mix 1", EA ($*$) and QR ($+$).

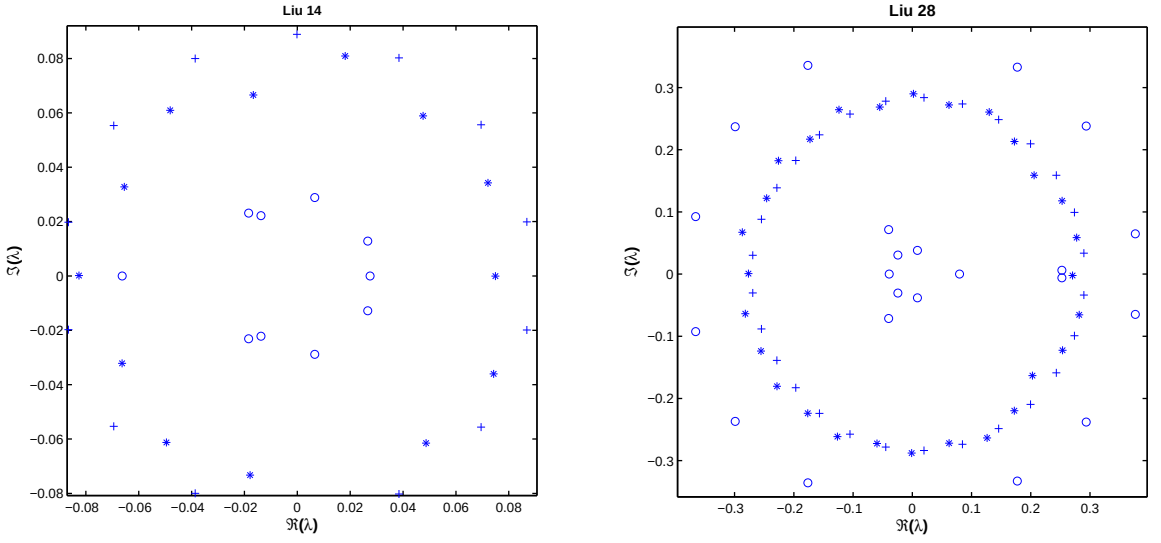

Figure 7.9: The eigenvalues of Liu's matrices 14 and 28 computed with HZ ($\circ$) using shifting strategy "mix 2" and random shifts, EA ($*$) and QR ($+$).

7.7 Clement Matrices

Clement's matrices are nonsymmetric tridiagonal and they were generated for test purposes [21]. A famous example of a Clement matrix T also analyzed in [7], is defined by $T = \text{tridiag}(\beta, 0, \gamma)$ with $\beta_j = \gamma_{n-j}$, $\gamma_j = j$, $k = 1\!:\!n-1$. Its eigenvalues are $\pm(n-1), \pm(n-3), \ldots, \pm 1$ for n even and $\pm(n-1), \pm(n-3), \ldots, 0$ for n odd. We see in Figure 7.10 that the eigenvalue condition numbers are large in the middle of the spectrum and small at its ends.

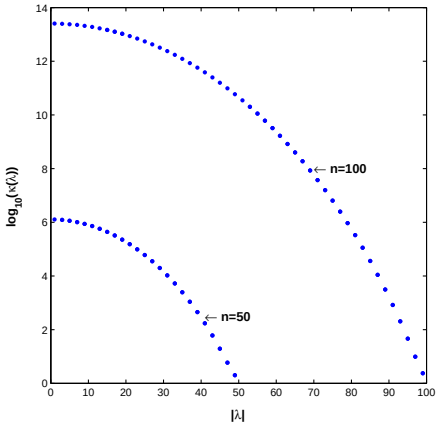

Figure 7.10: Eigenvalue condition numbers for the Clement matrix for $n = 50$ and 100.

Since the sign of the product $\beta_j\gamma_j$, $j = 1\!:\!n-1$ is constant, the process described in (7.1)–(7.2), yields a standard symmetric eigenvalue problem. In Figure 7.11, we plot in the complex plane the eigenvalues of the problem for $n = 200$ and $n = 300$, computed with QR. We see that they are a very poor approximation of the exact eigenvalues. Note that applying the process (7.1)–(7.2) to T yields a symmetric eigenvalue problem and in this case, the symmetric QR algorithm produces good approximations. In this case, we do not present the results with the HZ algorithm since it is equivalent to a symmetric QR.

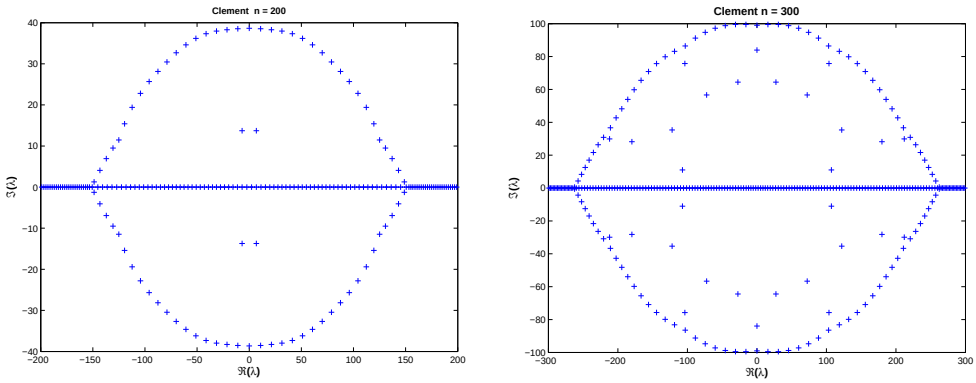

Figure 7.11: Eigenvalues of the Clement matrix with $n = 200$ and $n = 300$ computed with MATLAB's function `eig`.

We modify the definition of Clement matrices in order to test the HZ algorithm. For $n = 50$ and $n = 100$, we take

$$\begin{aligned} T &= \text{tridiag}(\beta, 0, \gamma), \\ \gamma_j &= (-1)^j j, \qquad k = 1\!:\! n-1, \\ \beta_j &= \gamma_{n-j}, \qquad k = 1\!:\! n-1. \end{aligned}$$

For these matrices, we have that $\text{sign}(\beta_j\gamma_j) = -\text{sign}(\beta_{j+1}\gamma_{j+1})$. Thus, the eigenvalue problem $(T - \lambda J)$, where the pair (T, J) is obtained by the process described in (7.1)–(7.2) is not a standard symmetric eigenvalue problem since $J \neq \pm I$. The spectrum of these matrices is plotted in Figures 7.12 and 7.13.

For $n = 50$, the condition numbers of the eigenvalues lie between 29 and 3×10^6 and for $n = 100$, they lie between 4×10^6 and 2×10^{21}. The condition number decreases as the modulus of the eigenvalue increases at the same rate as in Figure 7.10. The HZ algorithm performed 216 and 328 iterations for $n = 50$ and $n = 100$, respectively, while the Erhlich-Aberth method needed 352 and 758 for $n = 50$ and $n = 100$, respectively.

In Table 7.8, we compute the largest relative error between the eigenvalues

Table 7.8: Largest relative error of the computed eigenvalues of the modified Clement matrices with $n = 50$ and $n = 100$.

n	50	100
HZ	4e-15	1.8e-14
QR	4e-11	6e-4
EA	1.3e-16	2.8e-16

computed in extended precision and the approximations obtained by the HZ algorithm, QR or EA. Those returned by the Erhlich-Aberth method have the smallest relative error. We see that in this example, the HZ algorithm returns better approximations than `eig` (QR algorithm). The eigenvalues with the smallest modulus have the largest relative error. This is due to the fact that the condition number is big in the middle of the spectrum and small at its ends.

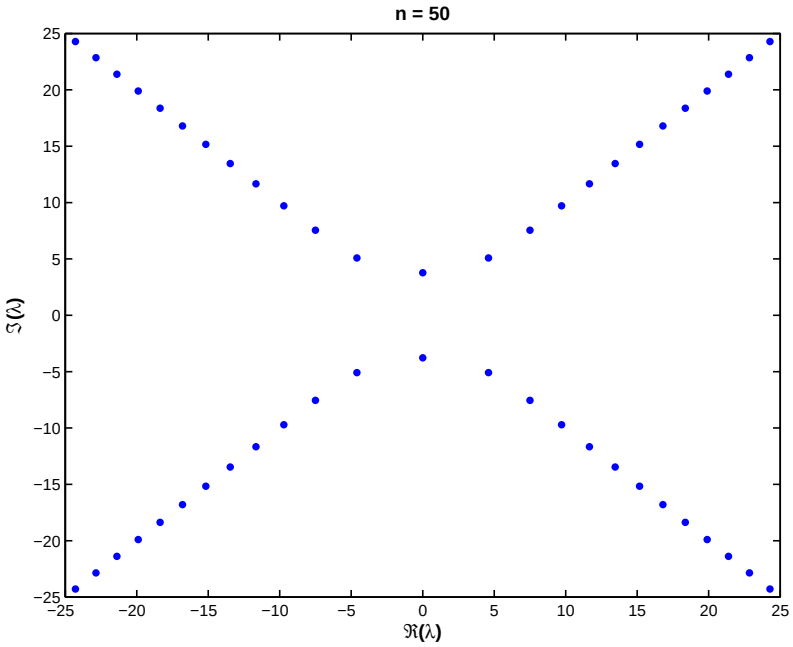

Figure 7.12: The eigenvalues of the modified Clement matrices for $n = 50$.

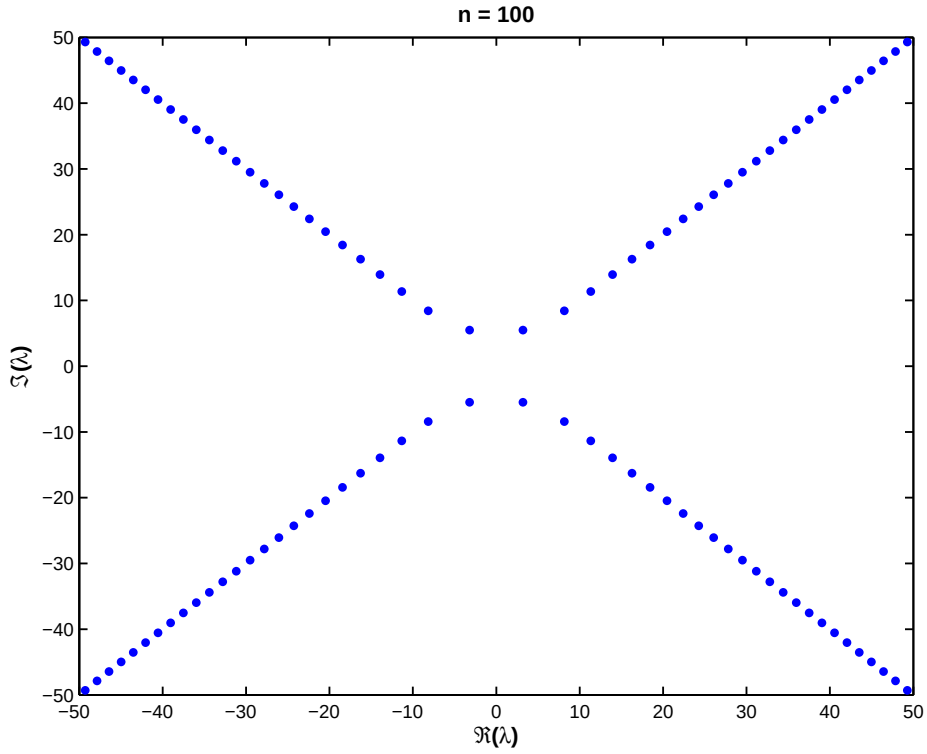

Figure 7.13: The eigenvalues of the modified Clement matrices for $n = 100$.

7.8 Symmetric QEPs

We now perform numerical experiments on symmetric QEPs. The QEP is first linearized using a symmetric linearization and then the eigenpairs are computed using either the QZ algorithm (that destroys symmetry) or the HZ algorithm that exploits the symmetry. The relative backward error is computed using (3.2).

7.8.1 Wave Equation

This example was presented in [40]. The equation of a free vibrating string clamped at both ends in a spatially inhomogeneous environment is given by

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + \epsilon a(x)\dfrac{\partial u}{\partial t} = \Delta u, \ \epsilon > 0, \ x \in (0, \pi), \\ u(t,0) = u(t,\pi) = 0, \\ u(0,x) = u_0(x). \end{cases}$$

We search for solutions in the form

$$u(x,t) = \sum_{k=1}^{n} q_k(t)\sin(kx)$$

and by applying the Galerkin method, we obtain the second order differential equation

$$M\ddot{q} + \epsilon C\dot{q} + Kq = 0,$$

where $q = [\, q_1, \quad \ldots, \quad q_n \,]$, $M = (\pi/2)I$, $K = (\pi/2)\mathrm{diag}(j^2)$ and

$$C = (c_{kj}), \quad c_{kj} = \int_0^{\pi} a(x)\sin(kx)\sin(jx)dx.$$

We take $a(x) = x^2(\pi - x)^2 - \delta$, $\delta = 2.7$ and $\epsilon = 0.1$. The quadratic matrix polynomial of interest is then defined by

$$Q(\lambda) = \lambda^2 M + \lambda\epsilon C + K.$$

Its eigenvalues are plotted in Figure 7.14.

We compute the eigenpairs of Q for $n = 50, 100$ and $n = 200$. In Table 7.9, we compare the QEP normwise backward errors for eigenpairs computed with the HZ or the QZ algorithm. We see that the eigenpairs computed with the HZ algorithm have smaller backward errors than those computed with QZ.

For $n = 200$, we plot in Figure 7.15 the modulus of the eigenvalues against the logarithm of the backward errors. On this example the eigenvalues computed

Table 7.9: Largest normwise QEP backward error.

n	50	100	200
HZ	5e-14	9e-13	4e-11
QZ	1.9e-12	2e-11	1.5e-10

with the HZ algorithm, marked with $\circ$ have a smaller backward error than the ones computed with the QZ algorithm (+). This also illustrates the fact that QZ is not necessarily backward stable for the solution of QEPs. The condition number of the matrix H that reduces the pair (A, B) obtained from a symmetric linearization to a tridiagonal-diagonal pair is relatively large, 5.2×10^2 for $n = 50$, 6×10^3 for $n = 100$ and 8.6×10^4 for $n = 200$. It appears that this ill conditioned matrix does not have a high influence on the backward error of the approximate eigenpairs of the QEP.

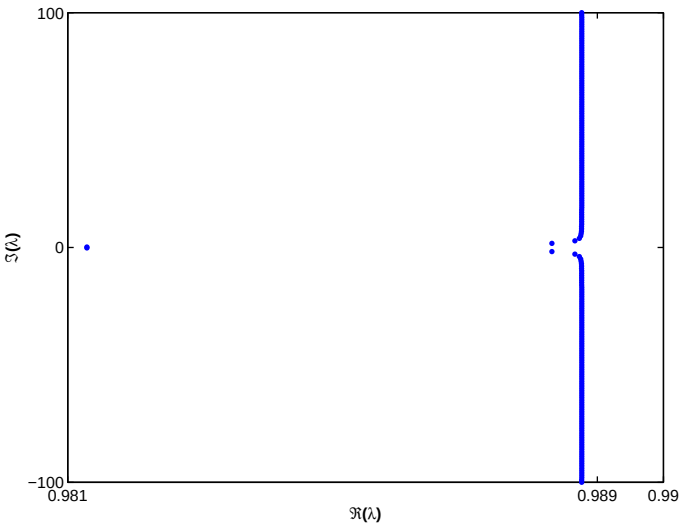

Figure 7.14: Eigenvalues of the wave equation for $n = 200$.

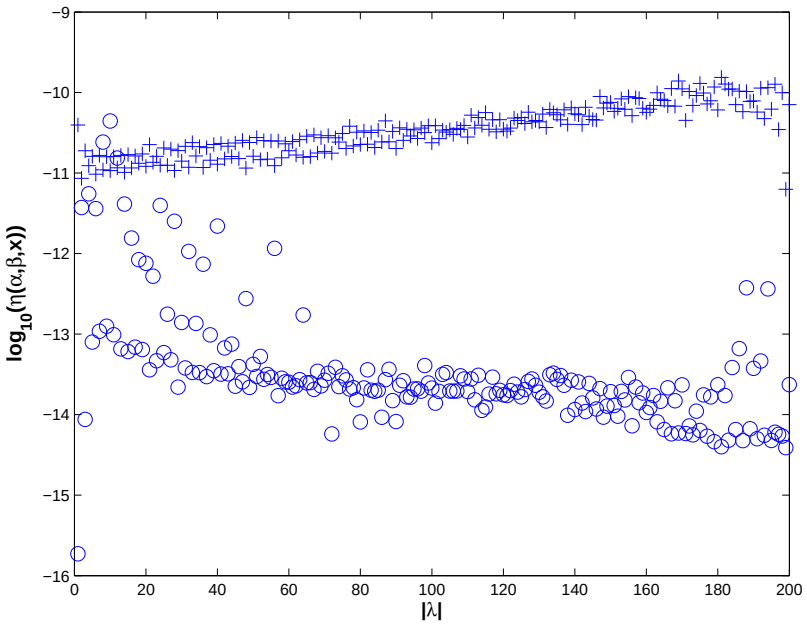

Figure 7.15: Backward errors of the approximate eigenpairs (with $\lambda = \alpha/\beta$) of the wave problem computed with HZ ($\circ$) and QZ ($+$) with $n = 200$.

7.8.2 Simply Supported Beam

The model of a simply supported beam can be described by [77]

$$\begin{cases} EI\dfrac{\partial^4 u}{\partial x^4} + \rho a\dfrac{\partial^2 u}{\partial t^2} + \delta_{(x-x_p)}\dfrac{\partial u}{\partial t} = 0,\ 0 < x < L,\ t > 0, \\ u(0,t) = u(L,t) = 0, \\ \frac{\partial^2 u}{\partial x^2}(0,t) = \frac{\partial^2 u}{\partial x^2}(L,t), \end{cases}$$

where $\delta_{(x-x_p)}$ is the Dirac measure centered at x_p and

$$E = 7 \times 10^{10},\ I = 6.25 \times 10^{-9},\ L = 1,\ \rho a L = 0.675.$$

Using the Galerkin method as in the previous example, we obtain the quadratic matrix polynomial

$$Q(\lambda) = \lambda^2 M + \lambda D + K,$$

where M and K are symmetric and $D = e_k e_k^T$. We took $n = 200$ and $x_p = L/2$. In this case $k = 100$. The spectrum is plotted in Figure 7.16.

The QEP backward errors for the approximate eigenpairs lie between 10^{-7} and 10^{-8} with QZ and between 8×10^{-9} and 3×10^{-18} with HZ. In Figure 7.17, we see that the backward errors obtained from QZ is almost constant and large whereas the backward errors from the HZ algorithm decreases exponentially with the modulus of the eigenvalues. We see that on this example, the HZ algorithm is more backward stable for solving the QEP than QZ. Note that only 367 HZ iterations are performed which is less than one iteration per eigenvalue. Thus, in this case the HZ algorithm is highly competitive with QZ to solve QEPs.

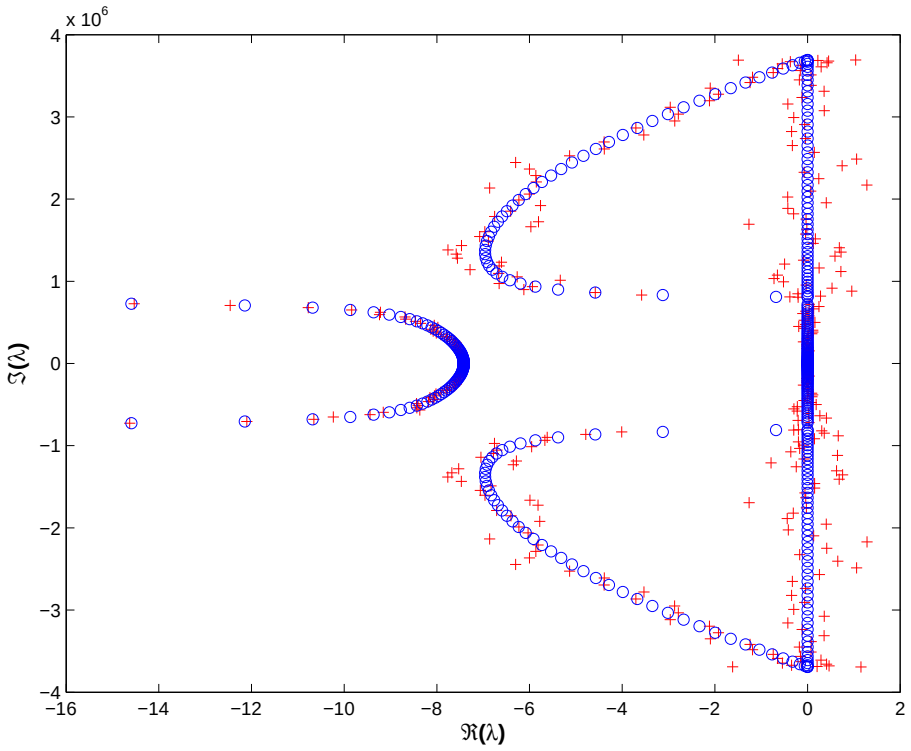

Figure 7.16: Eigenvalues of the beam problem with $n = 200$ computed with HZ ($\circ$) and QZ ($+$).

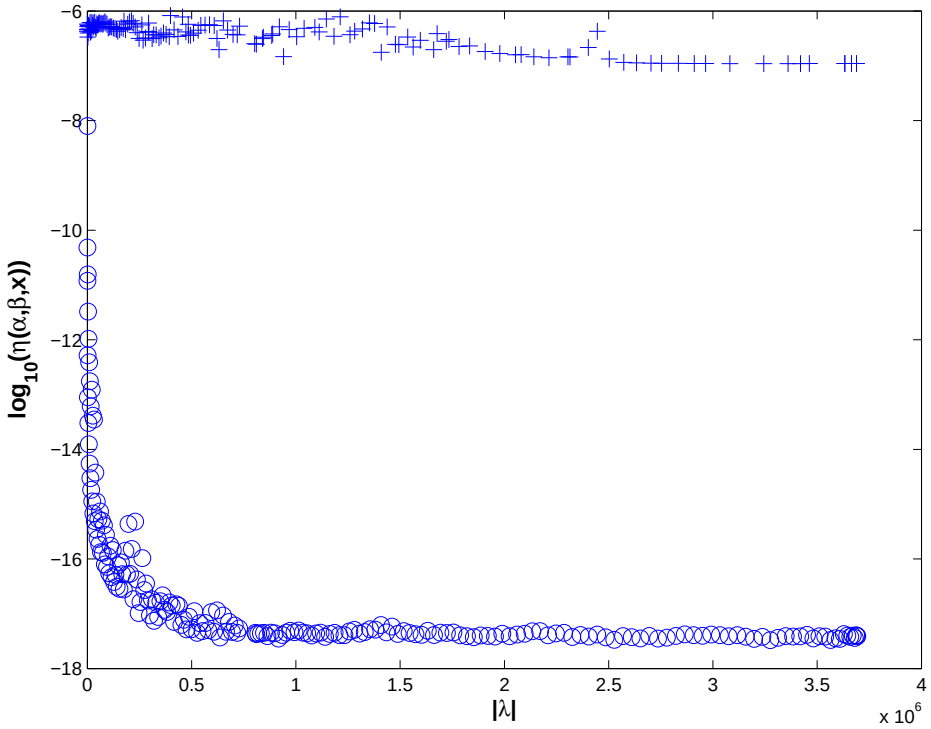

Figure 7.17: Backward errors of the approximate eigenpairs (with $\lambda = \alpha/\beta$) of the beam problem computed with HZ ($\circ$) and QZ ($+$) with $n=200$.

Chapter 8

Conclusion

8.1 Summary

We gave the condition number for an eigenpair of a PEP in its homogeneous form, where the perturbations are measured with weighted Frobenius norms. It has the advantage of defining relative condition numbers and partial condition numbers where we assume that some coefficient matrices are not perturbed. Using this approach, the condition number for $\lambda = 0$ and $\lambda = \infty$ is defined. We also computed the backward error of an approximate eigenpair of a homogeneous PEP. This work contributed to the improvement of `polyeig` in MATLAB (version 7.1) which allows this routine to return condition numbers in the case of an infinite or zero eigenvalue. We also computed structured condition numbers and backward errors. We gave a method that computes the structured backward error of an approximate eigenpair of a symmetric GEP in $O(n^2)$ operations where the $O(n^2)$ operations comes from the computation of the residual vector.

We proved that the set of hyperbolic matrices or generally the set of $(J, \widetilde{J})$-unitary and the set of $(J, \widetilde{J})$-orthogonal matrices are differentiable manifolds. It allows us to define local coordinates which permits the application of the implicit

function theorem to analyze the perturbation expansion of matrix factorizations. Since the HZ algorithm is based on the HR factorization, we started by analyzing its perturbation bounds for each of its factors and we continued by giving a detailed analysis of perturbation bounds for several matrix factorizations: the indefinite polar factorization, the hyperbolic singular value decomposition and the diagonalization of a symmetric pair with respect to a signature matrix. In each case, we computed the condition number of the factorization. For the last factorization, we considered the eigenvalue problem $(S - \lambda\widetilde{J})x = 0$ with S symmetric, $\widetilde{J} \in \mathrm{diag}(\pm 1)$ and the associated factorization $(S, \widetilde{J}) = H^T(D, J)H$ where H is $(J, \widetilde{J}, \mathbb{C})$-orthogonal, $J \in \mathrm{diag}_n(\mathbb{U})$ and D diagonal. We gave explicit computable expressions for structured eigenvalue condition numbers and described an algorithm to compute them when the structure is linear.

We presented an implementation of the HZ algorithm with several improvements. The problems with single shifts are:

1. Need to use complex arithmetic to solve a real problem.

2. The HR factorization does not exist when the shift is a complex eigenvalue by an argument similar to that Theorem 6.2. This may prevent the convergence of the algorithm.

We have seen that an implementation with a double shift allows to define the matrices of the next step for almost every unreduced pseudosymmetric tridiagonal starting matrix. We also analyzed a shifting strategy that reduces in most cases the number of iterations to 1.3 on average per eigenvalue. Moreover, the HZ algorithm preserves the pseudosymmetric form and all the computations are done in real arithmetic. It has a low operation cost. As we have seen in the numerical examples, it returns a very good approximation of an eigenpair for well conditioned problems and it returns comparable results to other classical

algorithms when the problem is ill conditioned.

8.2 Future Projects and Improvements

In Chapter 6, we presented three shifting strategies for the HZ algorithm. The shifting strategy influences the speed of the converges. We also have seen that the HZ algorithm may fail to converge with one shifting strategy and converge with another one. We see that there is a crucial need in obtaining an optimal shifting strategy for the HZ algorithm.

The second improvement to be made in the HZ algorithm would be a stable implementation of a Newton's method for the iterative refinement that solves directly the problem of diagonalizing a symmetric matrix with respect of a signature matrix. This idea can be explained as follows. We have computed the condition number of an eigenpair by giving an explicit expression of the condition operator. This condition operator can be used in the Newton method with the notation in Section 4.8:

$$\begin{aligned} dg_{H_n}(H_n^T S H_n)\Delta H_{n+1} &= J H_n^{-T} \mathcal{T}_-^{-1}(H_n^T D_n H_n)(R_n), \\ dg_{D_n}(H_n^T S_n H_n)\Delta D_{n+1} &= \Pi_d(H_n^{-T} R_n H^{-1}), \end{aligned}$$

where $R_n = S - H_n^T D_n H_n$, $\Delta H_{n+1} = H_{n+1} - H_n$, $\Delta D_{n+1} = D_{n+1} - D_n$ and dg_{H_n} and dg_{D_n} are given by (4.72)–(4.73). Theoretically, at each step $H_n \in \mathcal{O}(J, \widetilde{J}, \mathbb{C})$ but in practice due to rounding off errors the matrices H_n are not necessarily on the manifold. Thus, the problem becomes finding an implementation that guarantees that these matrices are on a nearby manifold of the type $H^T J H = \widetilde{J} + O(\epsilon)$ with ϵ small. For the orthogonal case, several authors suggested to apply a QR factorization to the orthogonal factor at each iteration (see for example [19]) which allows the matrices to be numerically orthogonal (which implies that ϵ is

relatively small). This occurs because the QR factorization is a more stable process than the HR factorization.

There have been recent improvements on linearizations of matrix polynomials where a whole class of linearizations were described. There were also recent improvements on the conditioning of linearizations of matrix polynomials. There are two unsolved problems that persist. We need to characterize all the linearizations of matrix polynomials and we have to find an algorithm that will allows us to choose the appropriate linearization to a given polynomial eigenvalue problem.

Bibliography

[1] Oliver Aberth. Iteration methods for finding all zeros of a polynomial simultaneously. *Math. Comp.*, 27:339–344, 1973.

[2] R. Alam and S. Bora. On sensitivity of eigenvalues and eigendecompositions of matrices. *Linear Algebra Appl.*, 396:273–301, 2005.

[3] A. L. Andrew, K. E. Chu, and P. Lancaster. Derivatives of eigenvalues and eigenvectors of matrix functions. *SIAM J. Matrix Anal. Appl.*, 14(4):903–926, 1993.

[4] A. Avez. *Calcul différentiel*. Collection Maîtrise de Mathématiques Pures. [Collection of Pure Mathematics for the Master's Degree]. Masson, Paris, 1983.

[5] Rajendra Bhatia. Matrix factorizations and their perturbations. *Linear Algebra Appl.*, 197/198:245–276, 1994.

[6] Rajendra Bhatia and Kalyan B. Sinha. Derivations, derivatives and chain rules. *Linear Algebra Appl.*, 302/303:231–244, 1999. Special issue dedicated to Hans Schneider (Madison, WI, 1998).

[7] Dario A. Bini, Luca Gemignani, and Françoise Tisseur. The Ehrlich-Aberth method for the nonsymmetric tridiagonal eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 27(1):153–175, 2005.

[8] Dario Andrea Bini. Numerical computation of polynomial zeros by means of Aberth's method. *Numer. Algorithms*, 13(3-4):179–200 (1997), 1996.

[9] A. W. Bojańczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog. A note on downdating the Cholesky factorization. *SIAM J. Sci. Statist. Comput.*, 8(3):210–221, 1987.

[10] Adam Bojanczyk, Nicholas J. Higham, and Harikrishna Patel. The equality constrained indefinite least squares problem: theory and algorithms. *BIT*, 43(3):505–517, 2003.

[11] Adam Bojanczyk, Sanzheng Qiao, and Allan O. Steinhardt. Unifying unitary and hyperbolic transformations. *Linear Algebra and Appl.*, 316(1-3):183–197, 2000.

[12] Adam W. Bojańczyk, Ruth Onn, and Allan O. Steinhardt. Existence of the hyperbolic singular value decomposition. *Linear Algebra Appl.*, 185:21–30, 1993.

[13] M. A. Brebner and J. Grad. Eigenvalues of $Ax = \lambda Bx$ for real symmetric matrices A and B computed by reduction to a pseudosymmetric form and the HR process. *Linear Algebra and Appl.*, 43:99–118, 1982.

[14] A. Bunse-Gerstner. An analysis of the HR algorithm for computing the eigenvalues of a matrix. *Linear Algebra and Appl.*, 35:155–173, 1981.

[15] R. Byers and D. Kressner. On the condition of a complex eigenvalue under real perturbations. *BIT*, 44(2):209–214, 2004.

[16] F. Chaitin-Chatelin and S. Gratton. On the condition numbers associated with the polar factorization of a matrix. *Numer. Linear Algebra Appl.*, 7(5):337–354, 2000.

[17] Xiao-Wen Chang, Christopher C. Paige, and G. W. Stewart. Perturbation analyses for the QR factorization. *SIAM J. Matrix Anal. Appl.*, 18(3):775–791, 1997.

[18] Françoise Chatelin. *Valeurs Propres de Matrices*. Masson, Paris, France, 1988.

[19] Moody T. Chu and Kenneth R. Driessel. The projected gradient method for least squares matrix approximations with spectral constraints. *SIAM J. Numer. Anal.*, 27(4):1050–1060, 1990.

[20] Philippe G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1982.

[21] Paul A. Clement. A class of triple-diagonal matrices for test purposes. *SIAM Rev.*, 1:50–52, 1959.

[22] Philip I. Davies, Nicholas J. Higham, and Françoise Tisseur. Analysis of the Cholesky method with iterative refinement for solving the symmetric definite generalized eigenproblem. *SIAM J. Matrix Anal. Appl.*, 23(2):472–493, 2001.

[23] A. Dax and S. Kaniel. The ELR method for computing the eigenvalues of a general matrix. *SIAM J. Numer. Anal.*, 18(4):597–605, 1981.

[24] Jean-Pierre Dedieu. Approximate solutions of numerical problems, condition number analysis and condition number theorem. In *The mathematics of numerical analysis (Park City, UT, 1995)*, volume 32 of *Lectures in Appl. Math.*, pages 263–283. Amer. Math. Soc., Providence, RI, 1996.

[25] Jean-Pierre Dedieu. Condition operators, condition numbers, and condition number theorem for the generalized eigenvalue problem. *Linear Algebra and Appl.*, 263:1–24, 1997.

[26] Jean-Pierre Dedieu and Françoise Tisseur. Perturbation theory for homogeneous polynomial eigenvalue problems. *Linear Algebra Appl.*, 358:71–94, 2003. Special issue on accurate solution of eigenvalue problems (Hagen, 2000).

[27] J. J. Dongarra, G. A. Geist, and C. H. Romine. Algorithm 710: FORTRAN subroutines for computing the eigenvalues and eigenvectors of a general matrix by reduction to general tridiagonal form. *ACM Trans. Math. Software*, 18(4):392–400, 1992.

[28] L.W. Ehrlich. A modified Newton method for polynomials. *Commun. ACM*, 10:107–108, 1967.

[29] George A. Geist. Reduction of a general matrix to tridiagonal form. *SIAM J. Matrix Anal. Appl.*, 12(2):362–373, 1991.

[30] I. Gohberg, Peter Lancaster, and Leiba Rodman. *Matrix Polynomials*. Academic Press, New York, 1982.

[31] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[32] Desmond J. Higham and Nicholas J. Higham. Structured backward error and condition of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl*, 1998.

[33] Nicholas J. Higham. The matrix computation toolbox. `http://www.ma.man.ac.uk/~higham/mctoolbox`.

[34] Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM J. Sci. Statist. Comput.*, 7(4):1160–1174, 1986.

[35] Nicholas J. Higham. The matrix sign decomposition and its relation to the polar decomposition. *Linear Algebra and Appl.*, 212/213:3–20, 1994.

[36] Nicholas J. Higham. A survey of componentwise perturbation theory in numerical linear algebra. In *Mathematics of Computation 1943–1993: a half-century of computational mathematics (Vancouver, BC, 1993)*, volume 48 of *Proc. Sympos. Appl. Math.*, pages 49–77. Amer. Math. Soc., Providence, RI, 1994.

[37] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[38] Nicholas J. Higham. J-orthogonal matrices: properties and generation. *SIAM Rev.*, 45(3):504–519 (electronic), 2003.

[39] Nicholas J. Higham, D. Steven Mackey, and Tisseur Françoise. The conditioning of linearizations of matrix polynomials. Numerical Analysis Report No. 415, Manchester Centre for Computational Mathematics, Manchester, England, 2005.

[40] Nicholas J. Higham and Françoise Tisseur. Bounds for eigenvalues of matrix polynomials. *Linear Algebra Appl.*, 358:5–22, 2003.

[41] W. Kahan, B. N. Parlett, and E. Jiang. Residual bounds on approximate eigensystems of nonnormal matrices. *SIAM J. Numer. Anal.*, 19(3):470–484, 1982.

[42] David W. Kammler. A perturbation analysis of the intrinsic conditioning of an approximate null vector computed with a SVD. *J. Comput. Appl. Math.*, 9(3):201–204, 1983.

[43] Michael Karow, Daniel Kressner, and Françoise Tisseur. Structured eigenvalue condition numbers. Numerical Analysis Report No. 467, Manchester Centre for Computational Mathematics, Manchester, England, April 2005.

[44] Charles Kenney and Alan J. Laub. Polar decomposition and matrix sign function condition estimates. *SIAM J. Sci. Statist. Comput.*, 12(3):488–504, 1991.

[45] A. Largillier. Bounds for relative errors of complex matrix factorizations. *Appl. Math. Lett.*, 9(6):79–84, 1996.

[46] Ren-Cang Li. Solving secular equations stably and efficiently. Numerical Analysis Report No. 89, November 1994. LAPACK Working Note 152.

[47] Zhishun A. Liu. *On the extended HR algorithm.* Pam-564, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, USA, august 1992.

[48] Oren E. Livne and Achi Brandt. N roots of the secular equation in $O(N)$ operations. *SIAM J. Matrix Anal. Appl.*, 24(2):439–453 (electronic), 2002.

[49] Craig Lucas. *Algorithms for Cholesky and QR Factorizations, and the Semidefinite Generalized Eigenvalue Problem.* PhD thesis, School of Mathematics, The University of Manchester, Manchester, UK, 2004.

[50] D. Steven Mackey, Niloufer Mackey, Christian Mehl, and Volker Mehrmann. Vector spaces of linearizations for matrix polynomials. Preprint, DFG Research Center, Technische Universitt, Berlin, Germany, 2005.

[51] Roy Mathias. Perturbation bounds for the polar decomposition. *SIAM J. Matrix Anal. Appl.*, 14(2):588–597, 1993.

[52] Volker Mehrmann and David Watkins. Polynomial eigenvalue problems with Hamiltonian structure. *Electron. Trans. Numer. Anal.*, 13:106–118 (electronic), 2002.

[53] C. B. Moler and G. W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.*, 10(2):241–256, 1973.

[54] Silvia Noschese and L Pasquini. Eigenvalue condition numbers:zero-structured vrsus traditional. Preprint, Mathematics Departement, University of Rome, La Sapienza, Italy, 2004.

[55] Ruth Onn, Steinhardt Allan O, and Adam Bojanczyk. The hyperbolic singular value decomposition and applications. *Applied mathematics and computing, Trans. 8th Army Conf., Ithaca/NY (USA) 1990, ARO Rep. 91-1, 93-108*, 1991.

[56] B. N. Parlett and H. C. Chen. Use of indefinite pencils for computing damped natural modes. *Linear Algebra and Appl.*, 140:53–88, 1990.

[57] Beresford Parlett. The development and use of methods of LR type. *SIAM Rev.*, 6:275–295, 1964.

[58] Beresford N. Parlett. Reduction to tridiagonal form and minimal realizations. *SIAM J. Matrix Anal. Appl.*, 13(2):567–593, 1992.

[59] L. Pasquini. Accurate computation of the zeros of the generalized Bessel polynomials. *Numer. Math.*, 86(3):507–538, 2000.

[60] Charles M. Rader and Allan O. Steinhardt. Hyperbolic Householder transformations. *IEEE Trans. Acoust. Speech Signal Process.*, 34:1589–1602, 1986.

[61] John R. Rice. A theory of condition. *SIAM J. Numer. Anal.*, 3(2):287–310, 1966.

[62] Heinz Rutishauser. Solution of eigenvalue problems with the LR-transformation. In *Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues*, number 49 in Applied Mathematics Series, pages 47–81. National Bureau of Standards, United States Department of Commerce, Washington, D. C., 1958.

[63] Michael Spivak. *Calculus on manifolds. A modern approach to classical theorems of advanced calculus.* W. A. Benjamin, Inc., New York-Amsterdam, 1965.

[64] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.*, 15:727–764, 1973.

[65] G. W. Stewart. Perturbation bounds for the QR factorization of a matrix. *SIAM J. Numer. Anal.*, 14(3):509–518, 1977.

[66] G. W. Stewart. A note on the perturbation of singular values. *Linear Algebra Appl.*, 28:213–216, 1979.

[67] G. W. Stewart. *Matrix Algorithms. Volume II: Eigensystems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[68] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory.* Academic Press, London, 1990.

[69] Ji-guang Sun. Stability and accuracy: Perturbation analysis of algebraic eigenproblems. Report UMINF 98-07, Department of Computing Science, University of Umeå, Sweden, August 1998.

[70] R. C. Thompson. The characteristic polynomial of a principal subpencil of a Hermitian matrix pencil. *Linear Algebra and Appl.*, 14(2):135–177, 1976.

[71] Françoise Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra and Appl.*, 309:339–361, 2000.

[72] Françoise Tisseur. Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 22(4):1038–1057, 2001.

[73] Françoise Tisseur and Karl Meerbergen. The quadratic eigenvalue problem. *SIAM Review*, 43(2):235–286, 2001.

[74] Françoise Tisseur. Tridiagonal-diagonal reduction of symmetric indefinite pairs. *SIAM J. Matrix Anal. Appl.*, 26(1):215–232 (electronic), 2004.

[75] David Watkins and Ludwig Elsner. Theory of decomposition and bulge-chasing algorithms for the generalized eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 15(3):943–967, 1994.

[76] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1):99–111, 1972.

[77] Nils Wegner. Simply supported beam. Private Communication.

[78] J. H. Wilkinson. *The algebraic eigenvalue problem*. Clarendon Press, Oxford, 1965.