ABSTRACT

REHMAN, RIZWANA. Numerical Computation of the Characteristic Polynomial of a Complex Matrix. (Under the direction of Ilse C. F. Ipsen.)

In this dissertation we present algorithms, and sensitivity and stability analyses for the numerical computation of characteristic polynomials of complex matrices. In Quantum Physics, for instance, characteristic polynomials are required to calculate thermodynamic properties of systems of fermions.

The general consensus seems to be that numerical methods for computing characteristic polynomials are numerically inaccurate and unstable. However, in order to judge the numerical accuracy of a method, one needs to investigate the sensitivity of the coefficients of the characteristic polynomial to perturbations in the matrix. We derive forward error bounds for the coefficients of the characteristic polynomial of an $n \times n$ complex matrix. These bounds consist of elementary symmetric functions of singular values. Furthermore, we investigate the numerical stability of two methods for the computation of characteristic polynomials. The first method determines the coefficients of the characteristic polynomial of a matrix from its eigenvalues. The second method requires a preliminary reduction of a complex matrix A to its Hessenberg form H. The characteristic polynomial of H is obtained from successive computations of characteristic polynomials of leading principal submatrices of H. Our numerical experiments suggest that the second method is more accurate than the determination of the characteristic polynomial from eigenvalues.

Numerical Computation of the Characteristic Polynomial of a Complex Matrix

by
Rizwana Rehman

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2010

APPROVED BY:

Dr. Ilse C. F. Ipsen Chair of Advisory Committee	Dr. Dean J. Lee Co-Chair of Advisory Committee
Dr. Stephen L. Campbell	Dr. Carl D. Meyer

DEDICATION

I dedicate this thesis to my beloved husband, Zahid, who helped me through times when nobody else could have. You have been with me every step of the way. Thank you for all the unconditional love and support that you have always given me.

BIOGRAPHY

Rizwana Rehman was born on Thursday, February 11, 1971 in the town of Quetta, Pakistan. At age seven, she was already an accomplished leader of her class with great academic results. Her special interest in math started in grade four, when she began helping her younger sister, Salma, who was struggling with math homework. She took her education very seriously, and finished first in class from kindergarten through high school without exception.

After finishing high school, during summer vacations, she chose to teach math to children in the Islamia Girls School. There she saw many little girls with tremendous potential to excel in education, forced to abandon school, because of tribal traditions where education in women was considered immoral. This situation made her very sad and she decided to put more emphasis on her studies. She acquired her B.S. (Honors) degree in Applied Mathematics in 1993 claiming first in class in Mathematics and Science Faculty. She was also awarded the Quaid-E-Azam Scholarship, which is given to the nation's top students.

At age 22 she got married and then went on to finish her Masters from University of Balochistan. Once again, taking first in class in Mathematics and Science Faculty.

Rizwana had two children and decided to be a homemaker for four years until she decided to pursue her aim to educate young girls by becoming a lecturer in the Government Girls College in Quetta. In May 2000 she was offered a job as a lecturer at the University of Balochistan, Pakistan.

In December 2000, she migrated to the United States to increase her opportunities for success. One year later, she joined Wake Technical College as a math instructor. In 2003 she decided to persue her PhD, and was accepted into North Carolina State University. As a graduate student she taught math on a wide range of scales, varying from one-on-one sessions, to class sizes in excess of 150. She was awarded both the Nicholas J. Rose Scholarship and the Lowell S. Winton Scholarship in 2009.

ACKNOWLEDGEMENTS

"Go down deep enough into anything and you will find mathematics"
Dean Schlicter

It is with heartfelt thanks that I acknowledge the persons that have contributed to my completion of this dissertation. First, I want to thank my academic advisor and dissertation chair, IIse Ipsen, who has been a significant contributor in this dissertation. Her ability to understand complex issues and her insights have strengthened my understanding of this dissertation significantly. I am grateful to her, for investing numerous hours of her precious time in my educational development. She has always been there for me when I needed her support and guidance. Her constructive criticisms have no doubt greatly improved the quality of this dissertation and it has been an honor to work with her. Special thanks to Dr. Dean Lee for giving me the opportunity to work on such a challenging topic and helping me whenever I needed him.

This achievement would not be possible without the support of Dr. Stephen Campbell. I still remember vividly the first day I visited NC State University and met with him. I was very nervous and unsure if I would be accepted as a graduate student. Dr Campbell gave me enough time to answer all my questions and encouraged me to apply to graduate school and continue my dream to become a Dr. Rehman. I admit that without Dr. Campbell's support I may have not achieved this goal.

I thank all the wonderful professors of the department of mathematics at NC State University who enhanced my knowledge and helped me to better understand complex mathematical concepts. In particular, I want to thank Dr Carl Meyer for teaching me Applied Linear Algebra. Due to his incredible teaching style, I fell in love with this discipline of mathematics. Thanks to Dr. Ernie Stitzinger for his moral support in the beginning years of my studies when I needed it most. I am

grateful to Dr. John Griggs for providing me the opportunity to teach hundreds of wonderful students. He was always available to give me his skillful advice regarding difficult classroom situations. I am also very thankful to the teachers at University of Balochistan, Pakistan, who provided me with their best efforts to prepare me for higher education. I would like to show my gratitude to Ms. Lala Rukh for making me a better person.

It is my pleasure to thank the NCSU math department for providing financial support during the past years through teaching assistantships and summer employment. Also, I am thankful for the Lowell S. Winston and Nicholas J. Rose Scholarship awards.

I would like to show my gratitude to Seyma Bennett-Shabbir, Di Bucklad, Carolyn Gunton, Nicole Newkirk Dahlke, and Denise Seabrooks during my time at NC State. I could not have survived at NC State without their help. I am indebted to the members of the Graduate Student Numerical Analysis Seminar for providing me with excellent feedback and ideas to make me a better presenter.

I cannot adequately express thanks to my husband, Zahid, for his constant support and overwhelming belief in me. Especially, keeping our three kids occupied during agonizing days of completing this thesis. I want to thank my children Mahnoor, Hamza, and Nida for supporting me in the days when I needed the energy to continue working on this dissertation. Throughout the years, I have received unwavering support from my parents and my siblings. They always are there to cheer me on. I want to pay my gratitude to my grandmother who took care of me, and my sisters during the prolonged sickness of my mother.

Finally, I am indebted to my dear friends Rebacca Kalhorn and Rebecca Wills who helped in proofreading my dissertation. I would not have completed this dissertation without their unconditional support. I am also grateful to Sukhbir Dhillon for helping me in writing the matlab codes for my research. Thanks for the friendship of Kelly Dickson, Monique Taylor, April Alston and Ivan Andjelkovic.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

Chapter 1 Introduction 1

Chapter 2 The Characteristic Polynomial in Quantum Physics 3

Chapter 3 Perturbation Bounds for the Characteristic Polynomial . 7
- 3.1 Introduction 7
- 3.2 Determinants 8
  - 3.2.1 Expansions 9
  - 3.2.2 Absolute Perturbation Bounds 13
  - 3.2.3 Relative Perturbation Bounds 19
  - 3.2.4 Local Sensitivity 20
- 3.3 Characteristic Polynomial 22
  - 3.3.1 General Matrices 24
  - 3.3.2 Normal Matrices 26
  - 3.3.3 Normwise Bounds 29

Chapter 4 Stability of Existing Numerical Methods 31
- 4.1 Leverrier's Method 32
- 4.2 Krylov's Method 34
- 4.3 Danilewski's Method 36
- 4.4 Hyman's Method 37

Chapter 5 The Computation of Characteristic Polynomials from Eigenvalues 39
- 5.1 Introduction 39
- 5.2 Elementary Symmetric Functions 40
  - 5.2.1 Absolute Perturbation Bounds 41
  - 5.2.2 Relative Perturbation Bounds 44
  - 5.2.3 The Summation Algorithm 51
  - 5.2.4 Forward Stability of the Summation Algorithm for Real Numbers 52
  - 5.2.5 Forward Stability of the Summation Algorithm for Complex Numbers 59
- 5.3 Bounds for Characteristic Polynomials from Eigenvalue Perturbations 63

5.3.1 Absolute Perturbation Bounds 63
5.3.2 Relative Perturbation Bounds 68
5.3.3 The Summation Algorithm Applied to Computed Eigenvalues 74
5.4 Numerical Tests 77

Chapter 6 La Budde's Method 83
6.1 Introduction 83
6.2 The Sturm Sequence Method 85
6.2.1 Running Error Bounds 87
6.3 La Budde's Method 91
6.3.1 Running Error Bounds for Real Matrices 93
6.3.2 Running Error Bounds for Complex Matrices 97
6.4 Combined Error Bounds 100
6.5 Numerical Tests 103
6.6 Comparison of La Budde's Method with the Eigenvalue Method . . . 105

Chapter 7 Conclusion and Future Research 107
7.1 Our Contributions 107
7.2 Conclusion 108
7.3 Future Research 109

Bibliography 110

Appendices 115

Appendix A Matlab File: sturm.m 116

Appendix B Matlab File: labudde.m 119

LIST OF TABLES

Table 6.1 Running error bounds of $p(\lambda)$ of Hansen's matrix 104

Table 6.2 Running error bounds of $p(\lambda)$ of the matrix R 104

LIST OF FIGURES

Figure 5.1 Results for $p(\lambda)$ of Forsythe matrix of order 200 78

Chapter 1

Introduction

The characteristic polynomial of an $n \times n$ complex matrix A is defined as

$$p(\lambda) \equiv \det(\lambda I - A) \equiv \lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n,$$

where, in particular $c_n = (-1)^n \det(A)$ and $c_1 = -\text{trace}(A)$. The roots of the characteristic polynomial $p(\lambda)$ of A are eigenvalues of A.

In this dissertation we investigate the numerical computation of the characteristic polynomial of a complex matrix. The coefficients of $p(\lambda)$ of a complex matrix are of central importance in a Quantum Physics application. Characteristic polynomials also have applications in Physics [27], Chemistry [35] and Engineering [36].

This dissertation is organized as follows: In chapter 2 we describe the physical application of the characteristic polynomial. The coefficients of $p(\lambda)$ are needed to determine thermodynamic properties of fermionic systems. To assess the numerical accuracy of our computed coefficients, we derive absolute perturbation bounds for the coefficients of $p(\lambda)$ in chapter 3. The bounds are expressed in terms of elementary symmetric functions of singular values. The results in chapter 3 were published in July 2008 in the SIAM Journal of Matrix Analysis and Applications [34]. In chapter 4 we analyze the numerical stability of some well known methods for the computation

of $p(\lambda)$.

Chapter 5 focuses on the computation of the characteristic polynomial of A from its eigenvalues. To obtain forward error bounds of coefficients of $p(\lambda)$ from this method we initially consider elementary symmetric functions $s_k(\lambda)$ of complex numbers $\lambda_1, \ldots, \lambda_n$. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then elementary symmetric functions of the eigenvalues of A equal the coefficients of $p(\lambda)$ of A, up to a sign. That is, $c_k = (-1)^k s_k(\lambda)$. Therefore, the perturbation bounds of elementary symmetric functions lead to condition numbers of coefficients, given the coefficients are computed from the eigenvalues of A. Furthermore, we investigate the numerical stability of the Summation Algorithm that determines elementary symmetric functions of $\lambda_1, \ldots, \lambda_n$, and we show that the Summation Algorithm is numerically forward stable. We also perform tests on various matrices by implementing MATLAB's "poly" function. The "poly" function first computes the eigenvalues of the given matrix and then determines $p(\lambda)$ from computed eigenvalues by using the Summation Algorithm. The results of our experiments suggest that the perturbation bounds of coefficients are accurate, when the cofficients are computed from the eigenvalues of A.

In chapter 6 we examine the numerical stability of a relatively unknown method due to La Budde [19], that computes the characteristic polynomial of A from its Hessenberg form. For our application in Quantum Physics which requires some initial coefficients of $p(\lambda)$ of A, we modify La Budde's method to produce the required coefficients. Our experiments suggest that La Budde's method yields more accurate results than the computation of $p(\lambda)$ from eigenvalues of A as implemented by MATLAB's "poly" function. In chapter 7 we discuss our future research.

Chapter 2

The Characteristic Polynomial in Quantum Physics

Our research concerns the computation of characteristic polynomials of complex matrices for an application in Quantum Physics. The characteristic polynomial of an $n \times n$ matrix A is defined as

$$p(\lambda) \equiv \det(\lambda I - A) = \lambda^n + c_1\lambda^{n-1} + \cdots + c_n.$$

The physical application consists of calculating thermodynamic properties of systems of interacting fermions, which is required, for instance, to understand the structure and evolution of neutron stars [39]. Thermodynamic properties of fermionic systems such as average energy, total energy, entropy and heat capacity can be derived from partition function Z. We give a brief description of fermions and partition functions.

Noninteracting Fermions

Fermions, named after the Italian physicist Enrico Fermi, are the basic building blocks of matter. They include electrons, protons and neutrons.

Fermions are elementary particles with odd half-integer spin. They are indistinguishable and have an antisymmetric wave function. Fermions are "antisocial": According to Pauli's exclusion principle, formulated in 1925 by the Austrian physicist Wolfgang Pauli, no two fermions in an atom can exist in the same quantum state, e.g. have the same position, energy state, spin, etc.

A quantum state is formally represented by a vector in an abstract linear space, and each physical observable such as energy is associated with a self-adjoint linear operator. The possible values from a physical measurement are the real eigenvalues of this operator. Of special interest to us is the single particle Hamiltonian operator H, whose eigenvalues E_1, E_2, ..., E_k represent the possible energies of a single fermion. We can regard H as an $n \times n$ matrix.

Since two identical fermions cannot occupy the same quantum state, the full set of possible energies for two identical noninteracting fermions are all possible sums of two eigenvalues, $E_{j_1} + E_{j_2}$, for $1 \leq j_1 < j_2 \leq n$. Similarly the full set of possible energies for k identical noninteracting fermions are all possible sums of k eigenvalues,

$$E_{j_1} + E_{j_2} + \cdots + E_{j_k}, \qquad 1 \leq j_1 < j_2 \cdots < j_k \leq n.$$

Partition Function

An object of considerable interest in studying systems of fermions is the partition function Z. As we described earlier, it is used to calculate the average energy of the system, its heat capacity, and many other thermal quantities. The partition function Z is given by

$$Z \equiv \text{trace}\left[\exp\left(-\beta H\right)\right].$$

Here $\beta = \frac{1}{KT}$, where K is Boltzmann's constant and T is the temperature in degrees Kelvin. Alternatively we can express Z in terms of the eigenvalues E_i of the

Hamiltonian H,

$$Z = \sum_i e^{-\beta E_i}.$$

As previously stated, for a system of k noninteracting fermions the energies are given by the single particle energy sums $E_{j_1} + E_{j_2} + \cdots + E_{j_k}$. They lead to the partition function Z_k corresponding to a system of k noninteracting fermions,

$$Z_k = \sum_{1 \le j_1 < j_2 \cdots < j_k \le n} \exp\left[-\beta\left(E_{j_1} + E_{j_2} + \cdots + E_{j_k}\right)\right].$$

Let us define the $n \times n$ matrix $A \equiv \exp(-\beta H)$ whose eigenvalues are $\lambda_j \equiv e^{-\beta E_j}$. We can represent Z_k as sums of products of eigenvalues (elementary symmetric functions)

$$Z_k = \sum_{1 \le j_1 < j_2 \cdots < j_k \le n} \lambda_{j_1}\lambda_{j_2}\cdots\lambda_{j_k}, \qquad 1 \le k \le n.$$

In particular, for the systems with the smallest and largest number of fermions, we have

$$Z_1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n = \mathrm{trace}(A), \qquad Z_n = \lambda_1 \cdots \lambda_n = \det(A),$$

and for a system of two fermions,

$$Z_2 = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \cdots + \lambda_1\lambda_n + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \cdots + \lambda_2\lambda_n + \cdots + \lambda_{n-1}\lambda_n.$$

The elementary symmetric functions Z_k equal the coefficients of the characteristic polynomial $p(\lambda) = \det(\lambda I - A)$, up to a sign. That is, $c_k = (-1)^k Z_k$. Therefore, calculating partition functions for a system of noninteracting fermions is esentially the same as computing the coefficients of the characteristic polynomial of the $n \times n$ matrix A. The index k associated with the coefficient c_k corresponds to the number of fermions.

Interacting Fermions

We have mentioned how the coefficients of the characteristic polynomial come up in Quantum Statistics. Unfortunately, the system of noninteracting fermions we just described is rather trivial. The problems of real interest are interacting quantum systems.

In many situations [39] we can use an integral identity called a Hubbard-Stratonovich transformation [33, 46] to write the interacting partition function as an integral over noninteracting partition functions, $\int Ds\, Z_N(s)$. Here, the variable s is actually a function of space and time. If we discretize space and time as a lattice of points, then Ds is the product of $ds(x, y, z, t)$ for each point (x, y, z, t). This is sometimes called a functional integral measure and the integral is a functional integral. We can again define an $n \times n$ matrix $A(s)$ with eigenvalues $\lambda_i(s)$ and partition function

$$Z_k(s) = \sum_{1 \le j_1 < j_2 \cdots < j_k \le n} \lambda_{j_1}(s)\lambda_{j_2}(s) \cdots \lambda_{j_k}(s).$$

Numerical Considerations

The matrices $A(s)$ are produced by a code over which we have no control. They are dense and currently of order $n \le 2000$, but in the future n may become larger. The matrices have no discernible structure, and the eigenvalues can be complex with a wide range of magnitudes. Characteristic polynomials have to be computed for various matrices $A(s)$. However, matrices at different points s appear to have no obvious common properties that could be exploited. No accuracy is specified for the computation, but an estimate should be available for the absolute accuracy of the coefficients.

Chapter 3

Perturbation Bounds for the Characteristic Polynomial

3.1 Introduction

Many methods for computing the characteristic polynomial $p(\lambda)$ of a complex matrix A were developed in the first half of the twentieth century as a precursor to an eigenvalue computation. However, we found very few bounds for the numerical sensitivity of the coefficients of the characteristic polynomial. To devise reliable numerical methods and to judge their accuracy we need to know the *numerical conditioning* of the coefficients. That is, if the matrix A is perturbed by E, then how do the coefficients of the characteristic polynomial of $A + E$ compare to those of $p(\lambda)$? Conditioning reflects sensitivity in *exact* arithmetic, with no reference to any algorithm. To this end, we derive perturbation bounds for absolute normwise perturbations. The basis for all bounds is an expansion of the determinant of a perturbed diagonal matrix.

Overview

Section 3.2 deals with determinants. We first derive expansions for determinants (section 3.2.1), and from them absolute perturbation bounds in terms of elementary symmetric functions of singular values (section 3.2.2), as well as relative bounds for determinants (section 3.2.3), and local sensitivity results (section 3.2.4). Section 3.3 deals with coefficients c_k of the characteristic polynomial. We derive absolute perturbation bounds for general matrices (section 3.3.1) and normal matrices (section 3.3.2), as well as normwise bounds (section 3.3.3).

Notation

The matrix A is a $n \times n$ complex matrix with singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, and eigenvalues λ_i, labelled so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$. The two-norm is $\|A\|_2 = \sigma_1$, and A^* is the conjugate transpose of A. The matrix $I = \text{diag}\begin{pmatrix}1 & \dots & 1\end{pmatrix}$ is the identity matrix, with columns e_i, $i \geq 1$. We denote by A_i the principal submatrix of order $n-1$ that is obtained by removing row and column i of A, and by $A_{i_1 \dots i_k}$ the principal submatrix of order $n-k$, obtained by removing rows and columns $i_1 \dots i_k$.

3.2 Determinants

We derive expansions and perturbation bounds for determinants. We start with expansions for determinants of perturbed matrices (section 3.2.1), and from them derive absolute perturbation bounds in terms of elementary symmetric functions of singular values (section 3.2.2), as well as relative bounds for determinants (section 3.2.3), and local sensitivity results (section 3.2.4).

3.2.1 Expansions

We derive expansions for determinants of perturbed matrices in several steps, by considering perturbations that have only a single nonzero diagonal element (Lemma 3.1), perturbations of diagonal matrices (Theorem 3.3), and at last perturbations of general matrices (Corollary 3.4).

Lemma 3.1. Let A be a $n \times n$ complex matrix, α a scalar, and A_i the principal submatrix of order $n-1$ obtained by deleting row and column i of A.

If $B = A + \alpha e_i e_i^*$, then $\det(B) = \det(A) + \alpha \det(A_i)$, $1 \leq i \leq n$.

Proof. This follows from a cofactor expansion [38, Theorem 2.3.1] along row i or column i of B. □

The above expansion can be used to expand the determinant of a perturbed diagonal matrix. Before deriving this expansion, we motivate its expression on matrices of order 2 and 3.

Example 3.2. If

$$D = \begin{pmatrix} \delta_1 & \\ & \delta_2 \end{pmatrix}, \qquad F = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix},$$

then $\det(D+F) = \det(D) + \det(F) + S_1$, where $S_1 \equiv \delta_1 f_{22} + \delta_2 f_{11}$.

If

$$D = \begin{pmatrix} \delta_1 & & \\ & \delta_2 & \\ & & \delta_3 \end{pmatrix}, \qquad F = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix},$$

then $\det(D+F) = \det(D) + \det(F) + S_1 + S_2$, where

$$S_1 \equiv \delta_1 \det \begin{pmatrix} f_{22} & f_{23} \\ f_{32} & f_{33} \end{pmatrix} + \delta_2 \det \begin{pmatrix} f_{11} & f_{13} \\ f_{31} & f_{33} \end{pmatrix} + \delta_3 \det \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix},$$

and $S_2 \equiv \delta_1\delta_2 f_{33} + \delta_1\delta_3 f_{22} + \delta_2\delta_3 f_{11}$.

These examples illustrate that the expansion of $\det(D+F)$ can be written as a sum, where each term consists of a product of k diagonal elements of D and the determinant of the "complementary" submatrix of order $n-k$ of F.

To derive expansions for diagonal matrices of any order, we denote by $F_{i_1 \ldots i_k}$ the principal submatrix of order $n-k$ obtained by deleting rows and columns $i_1 \ldots i_k$ of the $n \times n$ matrix F.

Theorem 3.3 (expansion for diagonal matrices)**.** Let D and F be $n \times n$ complex matrices. If $D = \operatorname{diag}\begin{pmatrix}\delta_1 & \ldots & \delta_n\end{pmatrix}$, then

$$\det(D+F) = \det(D) + \det(F) + S_1 + \cdots + S_{n-1},$$

where

$$S_k \equiv \sum_{1 \le i_1 < \cdots < i_k \le n} \delta_{i_1} \cdots \delta_{i_k} \det(F_{i_1 \ldots i_k}), \qquad 1 \le k \le n-1.$$

In particular, if $\delta_1 = \cdots = \delta_j = 0$ for some $1 \le j \le n-1$, then

$$\det(D+F) = \det(F) + S_1 + \cdots + S_{n-j},$$

where

$$S_k = \sum_{j+1 \le i_1 < \cdots < i_k \le n} \delta_{i_1} \cdots \delta_{i_k} \det(F_{i_1 \ldots i_k}), \qquad 1 \le k \le n-j.$$

Proof. The proof is by induction over the matrix order n, and Example 3.2 represents the induction basis. Assuming the statement is true for matrices of order $n-1$, we show that it is also true for matrices of order n. Let

$$D^{(j)} \equiv \operatorname{diag}\begin{pmatrix}0 & \ldots & 0 & \delta_{j+1} & \ldots & \delta_n\end{pmatrix}$$

be a diagonal matrix of order n with j leading zeros.

Applying Lemma 3.1 to $A \equiv D^{(1)} + F$ and $B \equiv A + \delta_1 e_1 e_1^*$ gives

$$\det(D + F) = \delta_1 \det(D_1 + F_1) + \det(D^{(1)} + F).$$

We repeat this process on the second summand $\det(D^{(1)} + F)$ to remove the diagonal elements δ_j one by one; $j \geq 2$. To this end, we apply Lemma 3.1 to $A \equiv D^{(j)} + F$ and $B \equiv A + \delta_j e_j e_j^*$, and denote by $(D^{(j)})_j$ the matrix of order $n-1$ obtained by removing row and column j from $D^{(j)}$. This gives

$$\det(D^{(1)} + F) = \sum_{j=2}^{n-1} \delta_j \det\left((D^{(j)})_j + F_j\right) + \delta_n \det(F_n) + \det(F).$$

Putting the above expression into the expansion for $\det(D + F)$ yields

$$\det(D + F) = \det(F) + \delta_1 \det(D_1 + F_1) + \sum_{j=2}^{n-1} \delta_j \det\left((D^{(j)})_j + F_j\right) + \delta_n \det(F_n).$$

Since $D_1 + F_1$ and $(D^{(j)})_j + F_j$ are matrices of order $n-1$, we can apply the induction hypothesis. To take advantage of the fact that the $j-1$ top diagonal elements of $(D^{(j)})_j$ are zero, we define the following sums for matrices of order $n-1$,

$$S_k^{(j)} \equiv \sum_{j+1 \leq i_1 < \cdots < i_k \leq n} \delta_{i_1} \cdots \delta_{i_k} \det(F_{j i_1 \ldots i_k}), \qquad 1 \leq j \leq n-1, \quad 1 \leq k \leq n-j,$$

where $F_{j i_1 \ldots i_k}$ is the matrix of order $n-k-1$ obtained by removing rows and columns $j, i_1, \ldots, i_k$ of F. The induction hypothesis yields

$$\begin{aligned}
\det(D_1 + F_1) &= \det(D_1) + \det(F_1) + S_1^{(1)} + \cdots + S_{n-2}^{(1)}, \\
\det\left((D^{(j)})_j + F_j\right) &= \det(F_j) + S_1^{(j)} + \cdots + S_{n-j}^{(j)}, \qquad 2 \leq j \leq n-2, \\
\det\left((D^{(n-1)})_{n-1} + F_{n-1}\right) &= \det(F_{n-1}) + S_1^{(n-1)}.
\end{aligned}$$

Now substitute the above expansions into the expression for $\det(D+F)$ and use the fact that $\delta_1 \det(D_1) = \det(D)$, $\sum_{i=1}^{n} \delta_i \det(F_i) = S_1$, and

$$\sum_{i=1}^{n-j} \delta_i S_j^{(i)} = S_{j+1}, \qquad 1 \le j \le n-2.$$

When the leading j diagonal elements of D are zero, then at most $n-j$ of the S_k are nonzero, and within each S_k one needs to account only for the nonzero summands. We now extend Theorem 3.3 to general matrices, by transforming them to diagonal form via the SVD. Let $A = U\Sigma V^*$ be a SVD of A, where $\Sigma = \operatorname{diag}\begin{pmatrix}\sigma_1 & \dots & \sigma_n\end{pmatrix}$ with $\sigma_1 \ge \cdots \ge \sigma_n \ge 0$, and U and V are unitary.

Corollary 3.4 (expansion for general matrices)**.** Let A and E be $n \times n$ complex matrices, and $F \equiv U^*EV$. Then

$$\det(A+E) = \det(A) + \det(E) + S_1 + \cdots + S_{n-1},$$

where

$$S_k \equiv \det(UV^*) \sum_{1 \le i_1 < \cdots < i_k \le n} \sigma_{i_1} \cdots \sigma_{i_k} \det(F_{i_1 \dots i_k}), \qquad 1 \le k \le n-1.$$

If $\operatorname{rank}(A) = r$ for some $1 \le r \le n-1$, then

$$\det(A+E) = \det(E) + S_1 + \cdots + S_r,$$

where

$$S_k = \det(UV^*) \sum_{1 \le i_1 < \cdots < i_k \le r} \sigma_{i_1} \cdots \sigma_{i_k} \det(F_{i_1 \dots i_k}), \qquad 1 \le k \le r.$$

Proof. The SVD of A implies $A + E = U(\Sigma + F)V^*$, and Theorem 3.3 implies

$$\det(\Sigma + F) = \det(\Sigma) + \det(F) + \hat{S}_1 + \cdots + \hat{S}_{n-1},$$

where

$$\hat{S}_k \equiv \sum_{1 \le i_1 < \cdots < i_k \le n} \sigma_{i_1} \cdots \sigma_{i_k} \det(F_{i_1,\ldots,i_k}), \qquad 1 \le k \le n-1.$$

With $S_k \equiv \det(UV^*)\hat{S}_k$ we obtain $\det(A + E) = \det(A) + \det(E) + S_1 + \cdots + S_{n-1}$.

Now suppose $\operatorname{rank}(A) = r \le n - 1$. Then $n - r$ singular values are zero, so that all products of $r + 1$ or more singular values are zero. In particular, $\det(A) = 0$. If $\operatorname{rank}(A) = r < n - 1$, then $S_{r+1} = \cdots = S_{n-1} = 0$. Moreover, the terms $S_1, \ldots, S_r$ contain only the nonzero singular values $\sigma_1, \ldots, \sigma_r$. □

Corollary 3.4 shows that the number of summands in the expansion decreases with the rank of the matrix.

3.2.2 Absolute Perturbation Bounds

We derive absolute perturbation bounds for determinants in terms of elementary symmetric functions of singular values. These bounds give rise to absolute first-order condition numbers. We also derive simpler, but weaker normwise bounds.

To bound the perturbations we need the following inequalities.

Lemma 3.5 (Hadamard's inequality)**.** If B is a $n \times n$ complex matrix, then

$$|\det(B)| \le \prod_{i=1}^{n} \|Be_i\|_2 \le \|B\|_2^n.$$

Proof. The first inequality is Hadamard's inequality [28, Corollary 7.8.2]. □

The bounds also contain elementary symmetric functions, which are defined as follows [28, Definition 1.2.9].

Definition 3.1 (elementary symmetric functions of singular values)**.** Let A be a $n \times n$ matrix with singular values $\sigma_1 \geq \ldots \geq \sigma_n$. The expressions

$$s_0 \equiv 1, \qquad s_k \equiv \sum_{1 \leq i_1 < \cdots < i_k \leq n} \sigma_{i_1} \cdots \sigma_{i_k}, \qquad 1 \leq k \leq n,$$

are the kth elementary symmetric functions of the singular values of A.

Now we are ready to derive the first perturbation bound for determinants of general matrices.

Corollary 3.6 (general matrices)**.** Let A and E be $n \times n$ complex matrices. Then

$$|\det(A) - \det(A+E)| \leq \sum_{i=1}^{n} s_{n-i} \|E\|_2^i.$$

If $\operatorname{rank}(A) = r$ for some $1 \leq r \leq n-1$, then

$$|\det(A+E)| \leq \|E\|_2^{n-r} \sum_{i=0}^{r} s_{r-i} \|E\|_2^i,$$

where the s_j are elementary symmetric functions in the r largest singular values of A, $1 \leq j \leq r$.

The bounds hold with equality for $E = \epsilon UV^*$ with $\epsilon > 0$, where $A = U\Sigma V^*$ is a SVD of A.

Proof. Corollary 3.4 implies $|\det(A) - \det(A+E)| \leq |\det(E)| + |S_1| + \cdots + |S_{n-1}|$. To bound $|S_k|$ use the fact that $|\det(UV^*)| = 1$ and $\sigma_i \geq 0$ to obtain

$$\begin{aligned} |S_k| &\leq \max_{1 \leq i_1 < \cdots < i_k \leq n} |\det(F_{i_1 \ldots i_k})| \sum_{1 \leq i_1 < \cdots < i_k \leq n} \sigma_{i_1} \cdots \sigma_{i_k} \\ &= \max_{1 \leq i_1 < \cdots < i_k \leq n} |\det(F_{i_1 \ldots i_k})| s_k. \end{aligned}$$

Lemma 3.5 implies $|\det(E)| \leq \|E\|_2^n$, and $|\det(F_{i_1 \ldots i_k})| \leq \|F\|_2^{n-k} = \|E\|_2^{n-k}$. Hence $|S_k| \leq s_k \|E\|_2^{n-k}$, $1 \leq k \leq n-1$.

Now suppose $\operatorname{rank}(A) = r$. Then Corollary 3.4 implies

$$|\det(A+E)| \leq |\det(E)| + |S_1| + \cdots + |S_r| \leq \|E\|_2^{n-r} \sum_{i=0}^{r} s_{r-i} \|E\|_2^i,$$

where the terms s_{r-i} contain only nonzero singular values.

If $E = \epsilon UV^*$, then $F = \epsilon I$ and $\det(F_{i_1 \ldots i_k}) = \epsilon^{n-k} = \|E\|_2^{n-k}$, so that $S_k = |S_k| = \|E\|_2^{n-k} s_k$. □

Corollary 3.6 bounds the absolute error in $\det(A+E)$ by elementary symmetric functions of singular values and powers of $\|E\|_2$. Although the bounds for nonsingular and rank-r matrices look different, because the sums start at different indices, they are consistent. If $\operatorname{rank}(A) \leq n-k$ for some $k \geq 1$, then $|\det(A+E)|$ is bounded by a multiple of $\|E\|_2^k$. Hence if $\|E\|_2 < 1$ then determinants of rank-deficient matrices tend to be better conditioned in the absolute sense.

Remark 3.7 (Hermitian positive-definite matrices)**.** In the special case when A is Hermitian positive-definite, singular values are equal to eigenvalues, so that we can write the elementary symmetric functions in terms of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. Hence in Corollary 3.6

$$s_k = \sum_{1 \leq i_1 < \cdots < i_k \leq n} \lambda_{i_1} \cdots \lambda_{i_k}, \qquad 1 \leq k \leq n-1.$$

Note that $A+E$ does not have to be Hermitian positive-definite, because no restrictions are placed on E.

Remark 3.8 (first-order absolute condition numbers)**.** Let A be a $n \times n$ complex matrix with $\operatorname{rank}(A) \geq n-1$ and $\|E\|_2 < 1$. Corollary 3.6 implies the first-order

bound

$$|\det(A) - \det(A+E)| \leq s_{n-1}\|E\|_2 + \mathcal{O}\left(\|E\|_2^2\right),$$

where $s_{n-1} \leq n\sigma_1 \ldots \sigma_{n-1}$. Hence we can view s_{n-1} or $n\sigma_1 \ldots \sigma_{n-1}$ as first-order condition numbers for absolute perturbations in A.

Example 3.9. The perturbation of a diagonally scaled Jordan block below illustrates that the first-order bound in Remark 3.8 can hold with equality. Let

$$A = \begin{pmatrix} 0 & \alpha_1 & 0 & \ldots & 0 \\ \vdots & 0 & \alpha_2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & 0 & \alpha_{n-1} \\ 0 & 0 & \ldots & \ldots & 0 \end{pmatrix}, \qquad E = \epsilon e_n e_1^*,$$

where $|\epsilon| \leq 1$ and $\alpha_i > 0$, $1 \leq i \leq n-1$. Then $|\det(A+E) - \det(A)| = \alpha_1 \ldots \alpha_{n-1}\epsilon$. Since the singular values of A are 0 and $\alpha_i > 0$, $1 \leq i \leq n-1$, we obtain

$|\det(A+E) - \det(A)| = s_{n-1}\|E\|_2$.

Replacing the singular values in Corollary 3.6 by powers of $\|A\|_2$ gives the simpler, but weaker bounds below.

Corollary 3.10 (normwise bounds)**.** Let A and E be $n \times n$ complex matrices. Then

$$\begin{aligned} |\det(A+E) - \det(A)| &\leq \sum_{i=1}^{n} \binom{n}{i} \|A\|_2^{n-i}\|E\|_2^i \\ &= (\|A\|_2 + \|E\|_2)^n - \|A\|_2^n. \end{aligned}$$

If $\operatorname{rank}(A) = r$ for some $1 \leq r \leq n-1$, then

$$\begin{aligned} |\det(A+E)| &\leq \|E\|_2^{n-r} \sum_{i=0}^{r} \binom{r}{i} \|A\|_2^{r-i}\|E\|_2^i \\ &= \|E\|_2^{n-r}(\|A\|_2 + \|E\|_2)^r. \end{aligned}$$

Proof. This follows from Corollary 3.6 and $s_{n-i} \leq \binom{n}{n-i}\|A\|_2^{n-i} = \binom{n}{i}\|A\|_2^{n-i}$, $1 \leq i \leq n-1$. □

A bound similar to the one in Corollary 3.10 was already derived in [5, section 20], [6, Problem I.6.11], [14, Theorem 4.7] for any p-norm, by taking Fréchet derivatives of wedge products. Below we give a basic proof from first principles for the two-norm.

Theorem 3.11 (section 20 in [5], problem I.6.11 in [6], Theorem 4.7 in [14])**.** Let A and E be $n \times n$ complex matrices. Then

$$|\det(A+E) - \det(A)| \leq n\|E\|_2 \max\{\|A\|_2, \|A+E\|_2\}^{n-1}.$$

Proof. We first show the statement for a diagonal matrix. That is, if $D = \operatorname{diag}\begin{pmatrix}\delta_1 & \dots & \delta_n\end{pmatrix}$ is diagonal, then

$$\det(D+F) = \det(D) + z, \qquad \text{where} \quad |z| \leq n\|F\|_2 \max\{\|D\|_2, \|D+F\|_2\}^{n-1}.$$

The proof is by induction. For $n = 2$

$$D = \begin{pmatrix}\delta_1 & \\ & \delta_2\end{pmatrix}, \qquad F = \begin{pmatrix}f_{11} & f_{12} \\ f_{21} & f_{22}\end{pmatrix},$$

and

$$z \equiv \det(D+F) - \det(D) = \delta_1 f_{22} + \det\begin{pmatrix}f_{11} & f_{12} \\ f_{21} & \delta_2 + f_{22}\end{pmatrix}.$$

Lemma 3.5 implies

$$\begin{aligned} |z| &\leq \|F\|_2\|D\|_2 + \left\|\begin{pmatrix}f_{11} \\ f_{21}\end{pmatrix}\right\|_2 \left\|\begin{pmatrix}f_{12} \\ \delta_2 + f_{22}\end{pmatrix}\right\|_2 \leq \|F\|_2\|D\|_2 + \|F\|_2\|D+F\|_2 \\ &\leq 2\|F\|_2 \max\{\|D\|_2, \|D+F\|_2\}. \end{aligned}$$

This completes the induction basis. Assuming the statement is true for matrices of

order $n-1$, we show that it is also true for matrices of order n. As in the proof of Theorem 3.3, let $D^{(1)} \equiv \text{diag}\begin{pmatrix} 0 & \delta_2 & \dots & \delta_n \end{pmatrix}$ be the matrix obtained from D by replacing δ_1 with 0, and apply Lemma 3.1 to conclude

$$\det(D+F) = \delta_1 \det(D_1 + F_1) + \det(D^{(1)} + F).$$

Since $D_1 + F_1$ is a matrix of order $n-1$, the induction hypothesis applies and gives $\det(D_1 + F_1) = \det(D_1) + z_1$, where

$$\begin{aligned} |z_1| &\leq (n-1)\|F_1\|_2 \max\{\|D_1\|_2, \|D_1+F_1\|_2\}^{n-2} \\ &\leq (n-1)\|F\|_2 \max\{\|D\|_2, \|D+F\|_2\}^{n-2}. \end{aligned}$$

Substitute the above expression into the expansion for $\det(D+F)$ to obtain

$$z \equiv \det(D+F) - \det(D) = \delta_1 z_1 + \det(D^{(1)} + F),$$

where $|\delta_1 z_1| \leq (n-1)\|F\|_2 \max\{\|D\|_2, \|D+F\|_2\}^{n-1}$. Applying Lemma 3.5 to $\det(D^{(1)} + F)$ yields

$$\det(D^{(1)} + F) \leq \|Fe_1\|_2 \prod_{i=2}^{n} \|(D+F)e_i\|_2 \leq \|F\|_2 \|D+F\|_2^{n-1}.$$

Therefore we have proved the theorem for diagonal matrices D.

To prove the theorem for general matrices A, let $A = U\Sigma V^*$ be a SVD of A. Then $\det(A+E) = \det(UV^*)\det(\Sigma + F)$, where $F \equiv U^*EV$. Since Σ is diagonal, $\det(\Sigma + F) = \det(\Sigma) + z$, where

$$|z| \leq n\|F\|_2 \max\{\|\Sigma\|_2, \|\Sigma+F\|_2\}^{n-1} = n\|E\|_2 \max\{\|A\|_2, \|A+E\|_2\}^{n-1}.$$

Hence $\det(A+E) - \det(A) = \det(UV^*)z$, and the result follows from $|\det(UV^*)| = 1$. □

3.2.3 Relative Perturbation Bounds

We derive expansions for relative perturbations of determinants, as well as relative perturbation bounds that improve existing bounds.

Theorem 3.12 (expansion)**.** Let A and E be $n \times n$ complex matrices. If A is nonsingular, then

$$\frac{\det(A+E) - \det(A)}{\det(A)} = \det(A^{-1}E) + S_1 + \cdots + S_{n-1},$$

where

$$S_k \equiv \sum_{1 \le i_1 < \cdots < i_k \le n} \det((A^{-1}E)_{i_1 \ldots i_k}), \qquad 1 \le k \le n-1.$$

Proof. Write $\det(A+E) = \det(A)\det(I + A^{-1}E)$ and apply Theorem 3.3 to $\det(I + A^{-1}E)$. □

Corollary 3.13 (relative perturbation bound)**.** Let A and E be $n \times n$ complex matrices. If A is nonsingular, then

$$\frac{|\det(A+E) - \det(A)|}{|\det(A)|} \le \left(\kappa \frac{\|E\|_2}{\|A\|_2} + 1\right)^n - 1,$$

where $\kappa \equiv \|A\|_2 \|A^{-1}\|_2$.

Proof. Apply Corollary 3.6 to

$$\frac{|\det(A+E) - \det(A)|}{|\det(A)|} = |\det(I + A^{-1}E) - \det(I)|,$$

and bound $\|A^{-1}E\|_2 \le \kappa \|E\|_2 / \|A\|_2$. □

Remark 3.14. Corollary 3.13 is more general and tighter than the following bound from [20, (1.6)], [26, Problem 14.15]:

$$\frac{|\det(A+E)-\det(A)|}{|\det(A)|} \leq \frac{n\kappa\|E\|_2/\|A\|_2}{1-n\kappa\|E\|_2/\|A\|_2},$$

which holds only for $n\kappa\|E\|_2/\|A\|_2 < 1$. This is true because of the following. With $q \equiv \|A^{-1}\|_2\|E\|_2 = \kappa\|E\|_2/\|A\|_2$ we can write the first term in the bound of Corollary 3.13 as

$$(q+1)^n = \sum_{i=0}^{n}\binom{n}{i}q^i \leq \sum_{i=0}^{n} n^i q^i \leq \sum_{i=0}^{\infty}(nq)^i.$$

If $nq < 1$, then $\sum_{i=0}^{\infty}(nq)^i = \frac{1}{1-nq}$, so that

$$(q+1)^n - 1 \leq \frac{1}{1-nq} - 1 = \frac{nq}{1-nq}.$$

This implies for the bound in Corollary 3.13

$$\left(\kappa\frac{\|E\|_2}{\|A\|_2}+1\right)^n - 1 \leq \frac{n\kappa||E\|_2/\|A\|_2}{1-n\kappa\|E\|_2/\|A\|_2},$$

where the last expression is the bound in [20, inequality (1.6)], [26, Problem 14.15].

3.2.4 Local Sensitivity

We derive a local condition number for determinants from directional derivatives. The directional derivative for $\det(A)$ in the direction E is $\frac{d^k}{dx^k}\det(A+xE)$.

Although we derive the expressions below from the expansion in Theorem 3.3, we could have also used the expression for derivatives of $A(x)$ in [29, equation (6.5.9)].

Theorem 3.15. Let A and E be $n \times n$ complex matrices, $F \equiv U^*EV$, and x a real scalar. Then

$$\det(A+xE) = \sum_{i=1}^{n} S_{n-i}x^i + \det(A),$$

where

$$S_0 \equiv \det(E), \qquad S_k \equiv \det(UV^*) \sum_{1 \le i_1 < \cdots < i_k \le n} \sigma_{i_1} \cdots \sigma_{i_k} \det(F_{i_1 \ldots i_k}), \quad 1 \le k \le n-1,$$

and

$$\frac{d^k}{dx^k} \det(A + xE)|_{x=0} = k! S_{n-k}, \qquad 1 \le k \le n.$$

Proof. If $D = \operatorname{diag}\begin{pmatrix} \delta_1 & \ldots & \delta_n \end{pmatrix}$ is a diagonal matrix, then Theorem 3.3 implies $\det(D+xF) = \det(xF) + \tilde{S}_1 + \cdots + \tilde{S}_{n-1} + \det(D)$, where $\det(xF) = x^n \det(F) = x^n S_0$ and

$$\tilde{S}_k = \sum_{1 \le i_1 < \cdots < i_k \le n} \delta_{i_1} \cdots \delta_{i_k} \det(xF_{i_1 \ldots i_k}) = x^{n-k} S_k.$$

To derive the expansion for a general matrix, use the SVD as in Corollary 3.4. □

The first derivative gives the local condition number of the determinant with regard to small perturbations.

Corollary 3.16 (local condition number)**.** Let A and E be $n \times n$ complex matrices, and x a real scalar. Then

$$\left| \frac{d}{dx} \det(A + xE)|_{x=0} \right| \le s_{n-1} \|E\|_2, \qquad \text{where} \quad s_{n-1} \le n\sigma_1 \ldots \sigma_{n-1}.$$

Proof. Theorem 3.15 implies for the first derivative

$$\frac{d}{dx} \det(A + xE)|_{x=0} = \det(UV^*) \sum_{1 \le i_1 < \cdots < i_{n-1} \le n} \sigma_{i_1} \cdots \sigma_{i_{n-1}} \det(F_{i_1 \ldots i_{n-1}}),$$

where $F_{i_1 \ldots i_{n-1}}$ is a diagonal element of F. Lemma 3.5 implies $|\det(F_{i_1 \ldots i_{n-1}})| \le \|F\|_2 = \|E\|_2$. □

Corollary 3.16 shows that the sensitivity of $\det(A)$ to small perturbations in any direction E is determined by s_{n-1} or $n\sigma_1 \ldots \sigma_{n-1}$. A comparison with Remark 3.8

shows that the local condition number for $\det(A)$ is identical to the first-order condition number.

3.3 Characteristic Polynomial

Based on the determinant results in section 3.2, we derive absolute perturbation bounds for the coefficients of the characteristic polynomial for general matrices (section 3.3.1) and normal matrices (section 3.3.2), as well as simpler, but weaker normwise bounds (section 3.3.3).

Applying Theorem 3.3 to the characteristic polynomial

$$\det(\lambda I - A) = \lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n$$

of the $n \times n$ matrix A gives the well-known expressions [28, Theorem 1.2.12]

$$c_{n-k} = (-1)^{n-k} \sum_{1 \le i_1 < \ldots < i_k \le n} \det(A_{i_1 \ldots i_k}), \quad 0 \le k \le n-1,$$

where $A_{i_1 \ldots i_k}$ is the principal submatrix of order $n-k$ obtained by deleting rows and columns $i_1 \ldots i_k$ of A. The characteristic polynomial of the perturbed matrix $A + E$ is

$$\det(\lambda I - (A+E)) = \lambda^n + \tilde{c}_1\lambda^{n-1} + \cdots + \tilde{c}_{n-1}\lambda + \tilde{c}_n,$$

where $\tilde{c}_n = (-1)^n \det(A+E)$ and

$$\tilde{c}_{n-k} = (-1)^{n-k} \sum_{1 \le i_1 < \ldots < i_k \le n} \det(A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}), \qquad 1 \le k \le n-1.$$

The example on next page illustrates that products of singular values play an important role in the conditioning of the coefficients c_k.

Example 3.17 (companion matrices)**.** The $n \times n$ matrix

$$A = \begin{pmatrix} \alpha_1 & \alpha_2 & \dots & \dots & \alpha_n \\ \eta & 0 & \dots & \dots & 0 \\ 0 & \eta & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \eta & 0 \end{pmatrix}, \qquad \eta > 0,$$

is a multiple of a companion matrix, and let $E = e_1 \begin{pmatrix} \epsilon & \dots & \epsilon \end{pmatrix}$ with $\epsilon > 0$. The respective coefficients of the characteristic polynomials of A and $A+E$ are [26, section 28.6]

$$c_i = \alpha_i \eta^{i-1}, \qquad \tilde{c}_i = (\alpha_i + \epsilon)\eta^{i-1}, \qquad 1 \le i \le n.$$

Then $|\tilde{c}_i - c_i| = \epsilon\eta^{i-1}$, $1 \le i \le n$. The singular values of A are [26, section 28.6]

$$\sigma_1^2 = \frac{1}{2}\left(\alpha + \sqrt{\alpha^2 - 4|\alpha_n|^2}\right), \qquad \sigma_n^2 = \frac{1}{2}\left(\alpha - \sqrt{\alpha^2 - 4|\alpha_n|^2}\right),$$

where $\alpha \equiv 1 + |\alpha_1|^2 + \dots + |\alpha_n|^2$, and $\sigma_i = \eta$, $2 \le i \le n-1$. Therefore the conditioning of the coefficients c_k is determined by products of singular values.

The products of singular values in our perturbation bounds are expressed in terms of elementary symmetric functions of only the largest singular values of A.

Definition 3.2 (elementary symmetric functions in the largest singular values)**.** Let A be a $n \times n$ matrix with singular values $\sigma_1 \ge \dots \ge \sigma_n$. Denote by

$$s_0^{(k)} \equiv 1, \qquad s_j^{(k)} \equiv \sum_{1 \le i_1 < \dots < i_j \le k} \sigma_{i_1} \dots \sigma_{i_j}, \qquad 1 \le j \le k, \quad 1 \le k \le n,$$

where $s_j^{(n)} = s_j$. The expression $s_j^{(k)}$ is the jth elementary symmetric function in the k largest singular values of A.

3.3.1 General Matrices

We use the determinant expansion in Corollary 3.4 to derive perturbation bounds for coefficients c_k of general matrices.

Theorem 3.18 (general matrices)**.** Let A and E be $n \times n$ complex matrices. Then

$$|\tilde{c}_k - c_k| \leq \binom{n}{k} \sum_{i=1}^{k} s_{k-i}^{(k)} \|E\|_2^i, \qquad 1 \leq k \leq n.$$

If $\operatorname{rank}(A) = r$ for some $1 \leq r \leq n-1$, then

$$|\tilde{c}_k - c_k| \leq \binom{n}{k} \|E\|_2^{k-r} \sum_{i=0}^{r} s_{r-i}^{(k)} \|E\|_2^i, \qquad r+1 \leq k \leq n.$$

Proof. In the perturbed coefficient

$$\tilde{c}_{n-k} = (-1)^{n-k} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \det(A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}),$$

the matrices $A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}$ are of order $n-k$. Fix the indices $i_1, \ldots, i_k$; set $B \equiv A_{i_1 \ldots i_k}$ and $F \equiv E_{i_1 \ldots i_k}$; and let $\mu_1 \geq \ldots \geq \mu_{n-k}$ be the singular values of B. Corollary 3.4 implies $\det(B+F) = \det(B) + \det(F) + S_1 + \cdots + S_{n-k-1}$, where

$$S_j = \sum_{1 \leq i_1 < \ldots < i_j \leq n-k} \mu_{i_1} \ldots \mu_{i_j} \det(F_{i_1 \ldots i_j}), \qquad 1 \leq j \leq n-k-1.$$

Since B is a submatrix of A, the singular values interlace [28, Theorem 7.3.9], so that $\sigma_j \geq \mu_j$, $1 \leq j \leq n-k$. With Lemma 3.5 we obtain $|S_j| \leq s_j^{(n-k)} \|E\|_2^{n-k-j}$. Hence $|S_1| + \cdots + |S_{n-k-1}| \leq \sum_{i=1}^{n-k} s_{n-k-i}^{(n-k)} \|E\|_2^i$. Summing up the terms associated with all $\binom{n}{k}$ submatrices $A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}$ gives the desired bound for $|\tilde{c}_{n-k} - c_{n-k}|$.

Now suppose $\operatorname{rank}(A) = r \leq n-1$. Since r singular values are nonzero, the

elementary symmetric functions $s_j^{(k)}$ in the k largest singular values remain unchanged for $k \leq r$.

Since $n-r$ singular values are equal to zero, all products of $r+1$ or more singular values are zero. Hence for $k \geq r+1$ we have $s_j^{(k)} = 0$ whenever $j \geq r+1$, so that

$$\sum_{i=1}^{k} s_{k-i}^{(k)} \|E\|_2^i = \|E\|_2^{k-r} \sum_{i=0}^{r} s_{r-i}^{(k)} \|E\|_2^i.$$

Moreover, for $j \leq r$ the $s_j^{(k)}$ are functions of the r largest singular values only, so that $s_j^{(k)} = s_j^{(r)}$. Therefore $\sum_{i=0}^{k} s_{k-i}^{(k)} \|E\|_2^i = \|E\|_2^{k-r} \sum_{i=0}^{r} s_{r-i}^{(r)} \|E\|_2^i$, giving the desired bound for $|\tilde{c}_k - c_k|$ when $k \geq r+1$. □

For the two extreme coefficients, Theorem 3.18 produces the expected bounds: In the case of $c_n = (-1)^n \det(A)$, the bound coincides with the determinant bound in Corollary 3.6, while for $c_1 = -\mathrm{trace}(A)$ we obtain $|\tilde{c}_1 - c_1| \leq n\|E\|_2$. Theorem 3.18 shows that the conditioning of c_k with regard to absolute perturbations is determined by the binomial term $\binom{n}{k}$ and the elementary symmetric functions in the k largest singular values. The binomial coefficient is largest for c_k with $k \approx n/2$, because $\binom{n}{n-k} = \binom{n}{k}$, and $\binom{n}{k}$ is monotonically increasing for $k < n/2$. In particular, if n is even, then for $k = n/2$ we have $k\binom{n}{k} \geq k\left(\frac{n}{k}\right)^k = n2^{n/2-1}$.

If $\mathrm{rank}(A) = r \leq n-2$, then the bounds for the coefficients $c_{r+1}, \ldots, c_n$ contain higher powers of $\|E\|_2$. Hence if $\|E\|_2 < 1$, then the coefficients $c_{r+1}, \ldots, c_n$ of rank-deficient matrices tend to be better conditioned in the absolute sense.

Remark 3.19 (first-order absolute condition numbers for general matrices)**.** Theorem 3.18 implies for $\|E\|_2 < 1$ the first-order bound

$$|\tilde{c}_k - c_k| \leq \binom{n}{k} s_{k-1}^{(k)} \|E\|_2 + \mathcal{O}(\|E\|_2^2), \qquad 1 \leq k \leq n,$$

where $s_{k-1}^{(k)} \leq k\sigma_1 \ldots \sigma_{k-1}$. Hence we can view $\binom{n}{k} s_{k-1}^{(k)}$ or $\binom{n}{k} k\sigma_1 \ldots \sigma_{k-1}$ as first-order condition numbers for absolute perturbations in the coefficient c_k.

3.3.2 Normal Matrices

We show that for normal matrices, the conditioning of the coefficients improves because the binomial term is smaller, and the elementary symmetric functions depend on all singular values, not just the largest ones. Note that all statements for normal matrices apply in particular to Hermitian matrices.

Theorem 3.20 (normal matrices)**.** If the $n \times n$ matrix A is normal, then

$$|\tilde{c}_k - c_k| \leq \sum_{i=1}^{k} \binom{n-k+i}{i} s_{k-i} \|E\|_2^i, \qquad 1 \leq k \leq n.$$

The bound holds with equality if $E = \epsilon I$ with $\epsilon > 0$.

Proof. Since A is normal, it has an eigenvalue decomposition $A = V\Lambda V^*$, where $\Lambda = \operatorname{diag}\begin{pmatrix}\lambda_1 & \cdots & \lambda_n\end{pmatrix}$ is complex diagonal, $|\lambda_1| \geq \ldots \geq |\lambda_n|$, and V is unitary. Set $D \equiv \lambda I - \Lambda$ and $F \equiv -V^*EV$, so that $\det(\lambda I - (A+E)) = \det(D+F)$. Theorem 3.3 implies $\det(D+F) = \det(D) + \det(F) + S_1 + \cdots + S_{n-1}$. Substituting $\det(D) = \lambda^n + \sum_{k=1}^{n} c_k \lambda^{n-k}$ and $\det(D+F) = \lambda^n + \sum_{k=1}^{n} \tilde{c}_k \lambda^{n-k}$ in the above expansion gives

$$\sum_{k=1}^{n} (\tilde{c}_k - c_k)\lambda^{n-k} = \det(F) + S_1 + \cdots + S_{n-1}.$$

Thus $\tilde{c}_k - c_k$ is equal to the coefficient of λ^{n-k} on the right-hand side, i.e., in $\det(F) + S_1 + \cdots + S_{n-1}$. Since

$$S_{n-j} \equiv \sum_{1 \leq i_1 < \cdots < i_{n-j} \leq n} (\lambda - \lambda_{i_1}) \cdots (\lambda - \lambda_{i_{n-j}}) \det(F_{i_1 \ldots i_{n-j}}), \qquad 1 \leq j \leq n-1,$$

has as highest power λ^{n-j}, the term λ^{n-k} can occur only in $S_{n-k}, \ldots, S_{n-1}$. This means $\tilde{c}_k - c_k$ is the sum of the coefficients of λ^{n-k} in $S_{n-1}, \ldots, S_{n-k}$. To bound the coefficient of λ^{n-k} in S_{n-j} in particular, we first bound all coefficients in S_{n-j}.

Observe that S_{n-j} is a sum of $\binom{n}{n-j}$ products $(\lambda - \lambda_{i_1}) \cdots (\lambda - \lambda_{i_{n-j}})$. For fixed $i_1, \ldots, i_{n-j}$ we can write the product as

$$(\lambda - \lambda_{i_1}) \cdots (\lambda - \lambda_{i_{n-j}}) = \lambda^{n-j} + \gamma_1 \lambda^{n-j-1} + \cdots + \gamma_{n-j-1}\lambda + \gamma_{n-j}.$$

The coefficient γ_l is a sum of $\binom{n-j}{l}$ products $\lambda_{j_1} \ldots \lambda_{j_l}$. Hence S_{n-j} contains $\binom{n}{n-j}\binom{n-j}{l}$ such products. Therefore we can bound $|S_{n-j}|$ by a sum of $\binom{n}{n-j}\binom{n-j}{l}$ products $|\lambda_{j_1}| \ldots |\lambda_{j_l}|$. Since A is normal $|\lambda_i| = \sigma_i$, so that these products are also summands of the elementary symmetric function s_l. The sum s_l contains $\binom{n}{l}$ such summands. Therefore the number of occurrences of s_l in the bound for $|S_{n-j}|$ is $\binom{n}{n-j}\binom{n-j}{l}/\binom{n}{l} = \binom{n-l}{j}$.

Now we are ready to return to the coefficient of λ^{n-k} in particular; it is γ_{k-j}. Applying the above counting argument with $l = k-j$ shows that the coefficient of λ^{n-k} in S_{n-j} is bounded by $\binom{n-k+j}{j} s_{k-j} |\det(F_{i_1 \ldots i_{n-j}})|$. Lemma 3.5 implies $|\det(F_{i_1 \ldots i_{n-j}})| \leq \|F\|_2^j = \|E\|_2^j$. Summing up the contributions from all S_{n-j}, $1 \leq j \leq k$, gives the desired result.

If $E = \epsilon I$, then $F = \epsilon I$ and $\det(F_{i_1 \ldots i_k}) = \epsilon^{n-k} = \|E\|_2^{n-k}$. □

Remark 3.21 (first-order absolute condition numbers for normal matrices)**.** If A is normal and $\|E\|_2 < 1$, then Theorem 3.20 implies the first-order bound

$$|\tilde{c}_k - c_k| \leq (n-k+1) s_{k-1} \|E\|_2 + \mathcal{O}(\|E\|_2^2), \qquad 1 \leq k \leq n,$$

where $s_{k-1} \leq k|\lambda_1 \ldots \lambda_{k-1}|$. Hence we can view $(n-k+1)s_{k-1}$ or $(n-k+1)k|\lambda_1 \ldots \lambda_{k-1}|$ as first-order condition numbers for absolute perturbations in the coefficient c_k.

For Hermitian positive-define matrices, the bound in Theorem 3.20 can be expressed in terms of the coefficients c_k.

Corollary 3.22 (Hermitian positive-definite matrices)**.** If the $n \times n$ matrix A is Hermitian positive-definite, then

$$|\tilde{c}_k - c_k| \leq \sum_{i=1}^{k} \binom{n-k+i}{i} |c_{k-i}| \|E\|_2^i, \qquad 1 \leq k \leq n.$$

Proof. The coefficients c_k are also elementary symmetric functions in the eigenvalues [28, section 1.2], and the eigenvalue of a Hermitian positive-definite is equal to the singular values. Thus $c_k = (-1)^k s_k$, and the result follows from Theorem 3.20. □

To first order, the conditioning of coefficient c_k is determined by the magnitude of the preceding coefficient, $|c_{k-1}|$. As in Corollary 3.7, the matrix $A+E$ in Corollary 3.22 does not have to be Hermitian positive-definite, because E can be arbitrary. Below we illustrate that one cannot use the expression in Corollary 3.22 for indefinite matrices; that is, positive-definiteness of A is crucial for the expression in Corollary 3.22.

Example 3.23. Corollary 3.22 is not valid for indefinite Hermitian matrices and in particular matrices with zero trace.

To see this, let

$$A = \begin{pmatrix} \alpha & \\ & -\alpha \end{pmatrix}, \qquad \tilde{A} = \begin{pmatrix} \alpha - \epsilon & \\ & -\alpha + \epsilon \end{pmatrix},$$

where $\alpha > 0$ and $\epsilon > 0$. The characteristic polynomials are

$$\det(\lambda I - A) = \lambda^2 - \alpha^2, \qquad (\lambda I - (A + E)) = \lambda^2 - (\alpha - \epsilon)^2,$$

so that $\tilde{c}_2 - c_2 = 2\alpha\epsilon - \epsilon^2$. However, $|\tilde{c}_2 - c_2|$ cannot be bounded in terms of c_1, as required by Corollary 3.22, because $c_1 = 0$.

3.3.3 Normwise Bounds

Replacing the singular value products by powers of $\|A\|_2$ gives the following simpler, but weaker bounds.

Corollary 3.24 (normwise bounds)**.** Let A and E be $n \times n$ complex matrices. Then

$$\begin{aligned} |\tilde{c}_k - c_k| &\leq k\binom{n}{k}\sum_{i=1}^{k}\binom{k}{i}\|A\|_2^{k-i}\|E\|_2^i, \\ &= \binom{n}{k}\left((\|A\|_2 + \|E\|_2)^k - \|A\|^k\right), \qquad 1 \leq k \leq n. \end{aligned}$$

If $\operatorname{rank}(A) = r$ for some $1 \leq r \leq n-1$, then

$$\begin{aligned} |\tilde{c}_k - c_k| &\leq k\binom{n}{k}\|E\|_2^{k-r}\sum_{i=1}^{r}\binom{k}{i}\|A\|_2^{r-i}\|E\|_2^i, \\ &= \binom{n}{k}\|E\|_2^{k-r}\left((\|A\|_2 + \|E\|_2)^r - \|A\|^r\right), \qquad r+1 \leq k \leq n. \end{aligned}$$

Proof. This follows from Theorem 3.18 and

$$s_{k-i}^{(k)} \leq \binom{k}{k-i}\|A\|_2^{k-i} = \binom{k}{i}\|A\|_2^{k-i}, \qquad 1 \leq i \leq k-1.$$

A similar bound was was already derived in [5, section 20] and [6, Problem I.6.11] for any p-norm, by taking Fréchet derivatives of wedge products. Below we give a basic proof from first principles for the two-norm.

Theorem 3.25 (section 20 in [5], problem I.6.11 in [6])**.** Let A and E be $n \times n$ complex matrices. Then

$$|\tilde{c}_k - c_k| \leq k\binom{n}{k}\|E\|_2 \max\{\|A\|_2, \|A+E\|_2\}^{k-1}, \qquad 1 \leq k \leq n.$$

Proof. As in the proof of Theorem 3.18, we use

$$\tilde{c}_{n-k} = (-1)^{n-k} \sum_{1 \le i_1 < \ldots < i_k \le n} \det(A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}).$$

This gives for the absolute error

$$|\tilde{c}_{n-k} - c_{n-k}| \le \sum_{1 \le i_1 < \ldots < i_k \le n} |\det(A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}) - \det(A_{i_1 \ldots i_k})|.$$

Theorem 3.11 implies that $|\det(A_{i_1 \ldots i_k} + E_{i_1 \ldots i_k}) - \det(A_{i_1 \ldots i_k})|$ is bounded by

$$(n-k)\|E_{i_1 \ldots i_k}\|_2 \max\{\|A_{i_1 \ldots i_k}\|_2, \|(A+E)_{i_1 \ldots i_k}\|_2\}^{n-k-1}.$$

Bounding the principal submatrices by the norms of the respective matrices and recognizing that the sum contains $\binom{n}{n-k}$ summands yields the desired bound. □

Chapter 4

Stability of Existing Numerical Methods

In the first half of the twentieth century the characteristic polynomial $p(\lambda)$ was often computed as a precursor to computing eigenvalues. In the second half of the twentieth century; however, Wilkinson and others demonstrated that computing eigenvalues as roots of characteristic polynomials is numerically unstable [49]. As a consequence, characteristic polynomials and methods for computing them fell out of favor with the numerical algebra community. They can be found in old books by Faddeeva [12], Gantmacher [16], and Householder [31]. Wilkinson was the first to analyze the numerical stability of many methods in detail [48, §3.14, §6.20, §6.52, §7.6, §7.18, §7.19]. Yet, he did not take into account the conditioning of the coefficients of characteristic polynomials. We give a brief description of popular methods along with their shortcomings.

4.1 Leverrier's Method

The first practical method for computing the characteristic polynomial of an $n \times n$ complex matrix A was developed by Leverrier in 1840. It is based on Newton's identities [41, page 504]

$$c_1 = -\text{trace}(A), \qquad c_k = -\frac{1}{k}\text{trace}(A^k + c_1 A^{k-1} + \ldots + c_{k-1}A), \quad 2 \leq k \leq n.$$

This can be expressed recursively as

$$c_k = -\frac{1}{k}\text{trace}(AB_{k-1}), \quad \text{where} \quad B_1 = A + c_1 I, \quad B_k = AB_{k-1} + c_k I.$$

Leverrier's method has been discovered and modified many times [30, 31]. Due to its generality Leverrier's method still continues to attract attention for many applications [4, 7]. However, it is not considered practical for the computation of the characteristic polynomial of a matrix A [12, §3.28], [41, page 504], [48, §7.19]. The first prohibitive feature is the potential numerical instability of the method and the second is huge operation count proportional to n^4. Regarding Leverrier's method, Wilkinson said [48, §7.19],

> "We find that it is common for severe cancellation to take place when the c_i are computed, as can be verified by estimating the orders of magnitudes of the various contributions to c_i."

Wilkinson described two factors responsible for the inaccurate result for c_k. One is errors in traces of powers of the matrix A and the other is errors in previously computed coefficients $c_1, \ldots, c_{k-1}$. To judge the numerical accuracy of Leverrier's method, in light of our perturbation results, we computed characteristic polynomials of many test matrices with well conditioned coefficients. Our results were unsatisfactory in most cases and they appear to testify to the accuracy of Wilkinson's analysis. We found that even for well conditioned coefficients of the characteristic polynomial

of A, the method seems to give inaccurate results. We observed that, in particular, if the coefficients c_k are very small or very large in comparison to the traces of powers of A, then we should not expect satisfactory results from Leverrier's method. We present two examples below to make our point. The tests are performed on MATLAB 7.6(R2008a) with machine precision $u \approx 1.1 \times 10^{-16}$.

Example 4.1. Consider the matrix A of order n with elements $a_{ij} = 1$, $1 \leq i, j \leq n$. The characteristic polynomial of A is

$$p(\lambda) = \lambda^n - n\lambda^{n-1}.$$

The coefficients $c_2, \ldots, c_n$ are zero. A has one non zero singular value $\sigma_1 = n$ and $\sigma_2 = \ldots = \sigma_n = 0$. Our perturbation results in Theorem 3.18 show that the characteristic polynomial of A is well conditioned since all but one elementary symmetric function of singular values are zero. We computed the coefficients of the characteristic polynomial of A for $n = 40$ from Leverrier's method and found that the computed c_{22} through c_{40} are in the range of 10^{18} to 10^{47}. The computed coefficients are much larger than our perturbation bounds. This test illustrates that Leverrier's method is numerically unstable.

In the second test we present the example of Wilkinson [48, §7.19].

Example 4.2 (§7.9 in [48])**.** Consider a diagonal matrix A where $a_{ii} = 2^{1-i}$, $1 \leq i \leq 20$. The results in Corollary 3.22 indicate that the coefficients of the characteristic polynomial of A are well conditioned in the absolute and relative sense. However, Leverrier's method fails to produce even a single correct digit of c_{12} through c_{20}. In fact some of the last coefficients are computed with wrong signs. This example again illustrates our observation about the instability of Leverrier's method.

To improve the accuracy of Leverrier's method one could think of implementing the method with higher machine precision, but the above example shows that in some

situations very high precision would be needed to get accurate results from Leverrier's algorithm. This makes the method almost impractical due to huge operation count.

4.2 Krylov's Method

In 1931 Krylov presented a method that implicitly tries to reduce A to a companion matrix C whose first row contains the coefficients of $p(\lambda)$. The matrix C is given below.

$$\begin{pmatrix} -c_1 & -c_2 & -c_3 & \cdots & -c_{n-1} & -c_n \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Explicitly, the method constructs a matrix K from what we now call Krylov vectors: $v, Av, A^2v, \ldots$, where $v \neq 0$ is an arbitrary vector. Let $m \geq 1$ be the smallest index for which the vectors $v, Av, \ldots, A^{m-1}v$ are linearly independent, but the inclusion of one more vector $A^m v$ makes the vectors linearly dependent. Then the linear system

$$Kx + A^m v = 0, \qquad \text{where} \qquad K = (v \quad Av \quad \ldots \quad A^{m-1}v)$$

has a unique solution x. Krylov's method solves the linear system $Kx = -A^m v$ for x. m is known as the grade of the initial vector v. In the fortunate case when $m = n$, v has a grade of n, the solution x contains the coefficients of $p(\lambda)$, and $x_i = c_{n-i+1}, 1 \leq i \leq n$. In this case, if C is the companion matrix of $p(\lambda)$, then direct multiplication shows that

$$K^{-1}AK = C.$$

If $m < n$, then x contains only a divisor of the minimum polynomial of A, which in turn is a divisor of $p(\lambda)$.

We can still continue the process by applying Krylov's method to matrices of smaller dimensions [41, Example 7.11.3]. In the end, the characteristic polynomial is recovered as a product of its divisors. Even though Kylov's method is very general, it has many shortcomings. We do not know in advance the grade of the initial vector v; therefore, we may end up with a divisor of the characteristic polynomial of A. Also K is usually a dense matrix even when A is sparse. Due to these reasons, a tremendous amount of work may be involved in the computation of the characteristic polynomial [41, page 650]. If A is derogatory, i.e. the eigenvalues of A have geometric multiplicity 2 or larger, then every starting vector v is of grade less than n, and Krylov's method does not produce the characteristic polynomial of A.

If the matrix A is non derogatory, then it is similar to its companion matrix. Therefore, almost every starting vector should give the characteristic polynomial. Still it is possible to start with a vector v of grade $m < n$, and Krylov's method fails to produce $p(\lambda)$ even for a non derogatory matrix A [23, Example 4.2].

To analyze Krylov's method, Wilkinson considered diagonal matrices with elements λ_i, $1 \leq i \leq n$, where $|\lambda_1| > |\lambda_2| \cdots > |\lambda_n|$ [48, §6.20 to §6.25, §7.6]. He describes that in most situations, Krylov matrix K is very ill conditioned. As m grows, the later Krylov vectors tend to be almost linearly dependent. The conditioning of K depends upon the initial vector v also. If any component v_i, $1 \leq i \leq n$, of v is small, K may be ill conditioned. If we make a favorable choice of selecting the initial vector v with components $v_i = 1$, $1 \leq i \leq n$, then the matrix K is a Vandermonde matrix. The i^{th} row of K is given by $e_i^T K = [1 \quad \lambda_i \quad \lambda_i^2 \quad \ldots \quad \lambda_i^{n-1}]$. Wilkinson discusses the conditioning of K with respect to many eigenvalue distributions. He concludes that for very harmless looking distributions of eigenvalues, K might be very ill conditioned and Krylov's method may fail drastically. In particular, if there is a considerable variation in the sizes of eigenvalues, then any quest for an initial vector v which provides a well conditioned K is doomed to failure. Wilkinson also concludes that Krylov's method only gives good results if the eigenvalues are "well distributed"

in the complex plane. His final assessment of Krylov's method is that it has severe limitations as a general purpose method.

4.3 Danilewski's Method

Danilewski's method [23, §3.1] reduces an $n \times n$ matrix A to its companion form by $n-1$ similarity transformations. The similarity transformations resemble those in a Gauss Jordan decomposition. Each transformation produces zeros in a particular row with one particular entry (pivot) being 1. At the end of process, we obtain the companion matrix C as described in Section 4.2. If A is derogatory, then it cannot be similar to any companion matrix because a companion matrix is always non derogatory. However, in this case A decomposes in such a way that we can remove a factor of the characteristic polynomial and apply Danilewski's method to a matrix of smaller dimension [23, §3.2]. At the end of the process, the characteristic polynomial $p(\lambda)$ of A is recovered as a product of its factors. Danilewski's method is attractive because of its generality and efficiency.

Householder proved that Danilewski's method, as well as many other methods, such as Weber-Voetter's, Bryan's and Samuelson's are particular implementations of Krylov's method [31, §6]. The problem with Krylov's method, as well as Danilewski's method is that they try to compute, either implicitly or explicitly, a similarity transformation to a companion matrix. However, such a transformation only exists if A is non derogatory. Hammarling showed that even for a non derogatory matrix A, in finite precision arithmetic Danilewski's method can produce very inaccurate results [23, Exercise 3.1]. We face numerical instability in Danilewski's method when the pivot element is small. The emergence of a small pivot element should indicate that the matrix A is close to being derogatory. Nevertheless, in practice a small pivot may emerge due to rounding errors also. Like Gauss Jordan we can interchange columns to bring a larger element in magnitude to the pivot position. But our choice of pivoting

is limited because we want to preserve the structure of the matrix C. It is therefore not clear that remedies like those proposed for Danilewski's method in [24], [32, pg 36] and [47] would be fruitful to make the method numerically reliable.

4.4 Hyman's Method

Hyman's method requires a preliminary reduction of A to its Hessenberg form H. His method evaluates $p(\lambda) = \det(\lambda I - H)$ at a specific value of λ [48, §7.11]. Misra, Quintana and Van Dooren [42] proposed Hyman's method for the computation of the characteristic polynomial of a real Hessenberg matrix H. Their basic idea can be described as follows. Let B be an $n \times n$ real matrix, and partition

$$B = \begin{array}{c} \\ 1 \\ n-1 \end{array} \begin{array}{c} \begin{array}{cc} n-1 & 1 \end{array} \\ \begin{pmatrix} b_1^T & b_{12} \\ B_2 & b_2 \end{pmatrix} \end{array}.$$

If B_2 is nonsingular then $\det(B) = (-1)^{n-1} \det(B_2)(b_{12} - b_1^T B_2^{-1} b_2)$. Specifically, if $B = \lambda I - H$, where H is an unreduced upper Hessenberg matrix then B_2 is nonsingular and upper triangular, so that $\det(B_2) = (-1)^{n-1} h_{21} \dots h_{n,n-1}$ is just the product of the subdiagonal elements. Thus

$$p(\lambda) = h_{21} \dots h_{n,n-1}(b_{12} - b_1^T B_2^{-1} b_2).$$

The quantity $B_2^{-1} b_2$ can be computed as the solution of a triangular system. However, b_1, B_2, and b_2 are functions of λ. Recovering the coefficients of λ^i requires the solutions of n upper triangular systems. A structured backward error bound for the characteristic polynomial of H under certain conditions has been derived in [42], and iterative refinement is suggested for improving backward accuracy. However, it is not clear that this will help in general. The numerical stability of Hyman's method

depends on the condition number with respect to inversion of the triangular matrix B_2. Since the diagonal elements of B_2 are $h_{21}, \ldots, h_{n,n-1}$, B_2 can be ill conditioned with respect to inversion if H has small subdiagonal elements.

Chapter 5

The Computation of Characteristic Polynomials from Eigenvalues

5.1 Introduction

In chapter 4 we discussed methods for computing coefficients c_k of the characteristic polynomial $p(\lambda)$ of an $n \times n$ complex matrix A. We concluded that these methods are numerically unstable or limited in their application. In this chapter we consider the computation of the characteristic polynomial $p(\lambda)$ of matrix A from its eigenvalues $\lambda_1, \ldots, \lambda_n$. The coefficients c_k of the characteristic polynomial $p(\lambda)$ are related to elementary symmetric functions $s_k(\lambda)$ of the eigenvalues.

Lemma 5.1 (page 494 in [41])**.** Define elementary symmetric functions of complex numbers numbers $\lambda_1, \ldots, \lambda_n$ as follows.

$$s_0(\lambda) = 1, \qquad s_k(\lambda) = \sum_{1 \le i_1 < \ldots < i_k \le n} \lambda_{i_1} \cdots \lambda_{i_k}, \quad 1 \le k \le n.$$

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A then

$$c_k = (-1)^k s_k(\lambda), \qquad 1 \leq k \leq n.$$

Moreover $s_k(\lambda)$ is the sum of all $k \times k$ principal minors of A.

Computation of the characteristic polynomial of matrix A from its eigenvalues has many advantages over the algorithms described in chapter 4. This method is very simple and we can easily implement and analyze it. Many numerically stable algorithms such as the QR algorithm for computing eigenvalues have been developed [21, §7.3, §7.5]. In the past twenty years many new algorithms for certain classes of matrices have emerged which compute the eigenvalues with high relative accuracy [9, 3, 10, 44]. High relative accuracy of eigenvalues in turn can lead to a more accurate computation of the characteristic polynomials.

Overview

Section 5.2 deals with elementary symmetric functions. The perturbation bounds of elementary symmetric functions lead to forward error bounds for the coefficients c_k in Section 5.3. In Section 5.4 we present numerical tests to check the accuracy of perturbation bounds and the method of computing the coefficients of the characteristic polynomials of matrices from their eigenvalues.

5.2 Elementary Symmetric Functions

We first derive absolute and relative perturbation bounds for elementary symmetric functions $s_k(\lambda)$, $1 \leq k \leq n$, in terms of perturbations in λ_i. These bounds relate the changes in λ_i, $1 \leq i \leq n$, to changes in $s_k(\lambda)$.

5.2.1 Absolute Perturbation Bounds

We consider absolute perturbations $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$ with $\tilde{\lambda}_i = \lambda_i + \epsilon_i$, $1 \leq i \leq n$. The corresponding elementary symmetric functions are

$$\begin{aligned} s_0(\tilde{\lambda}) = 1, \quad s_k(\tilde{\lambda}) &= \sum_{1 \leq i_1 < \ldots < i_k \leq n} \tilde{\lambda}_{i_1} \cdots \tilde{\lambda}_{i_k} \\ &= \sum_{1 \leq i_1 < \ldots < i_k \leq n} (\lambda_{i_1} + \epsilon_{i_1}) \cdots (\lambda_{i_k} + \epsilon_{i_k}), \qquad 1 \leq k \leq n. \end{aligned}$$

To bound the absolute error $|s_k(\tilde{\lambda}) - s_k(\lambda)|$ we construct diagonal matrices A and $A+E$ consisting of λ_i and $\lambda_i + \epsilon_i$. We apply bounds derived in chapter 3 for matrices to get perturbation bounds for $s_k(\lambda)$ in terms of perturbations in the matrix elements. We use Theorem 3.20 to derive the result below, where the absolute value applies component wise, i.e. $|\lambda| \equiv \big(|\lambda_1|, \ldots, |\lambda_n|\big)$. Theorem 3.20 states:

Let A and $A + E$ be $n \times n$ complex matrices with respective characteristic polynomials

$$\begin{aligned} \det(zI - A) &= z^n + c_1 z^{n-1} + \cdots + c_{n-1} z + c_n \\ \det(zI - (A + E)) &= z^n + \tilde{c}_1 z^{n-1} + \cdots + \tilde{c}_{n-1} z + \tilde{c}_n. \end{aligned}$$

If A is normal (or Hermitian), then

$$|\tilde{c}_k - c_k| \leq \sum_{i=1}^{k} \binom{n-k+i}{i} s_{k-i}\left(|\lambda|\right) \|E\|_2^i, \qquad 1 \leq k \leq n.$$

Theorem 3.20 helps us to bound the absolute error $|s_k(\tilde{\lambda}) - s_k(\lambda)|$ in terms of elementary symmetric functions of the absolute values $|\tilde{\lambda}_1|, \ldots |\tilde{\lambda}_n|$ as follows.

Theorem 5.2. If $\tilde{\lambda}_i = \lambda_i + \epsilon_i$, $1 \leq i \leq n$, then

$$|s_k(\tilde{\lambda}) - s_k(\lambda)| \leq \sum_{i=1}^{k} \binom{n-k+i}{i} s_{k-i}(|\lambda|)\, \epsilon_{abs}^i, \qquad 1 \leq k \leq n,$$

where $\epsilon_{abs} \equiv \max_{1 \leq i \leq n} |\epsilon_i|$.

Proof. Apply Theorem 3.20 to diagonal matrices $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $E = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_n)$. Lemma 5.1 implies $|s_k(\tilde{\lambda}) - s_k(\lambda)| = |\tilde{c}_k - c_k|$. Since E is diagonal we have $\|E\|_2 = \epsilon_{abs}$. □

Theorem 5.2 bounds the absolute error in $s_k(\tilde{\lambda})$ in terms of the "preceding" elementary functions $s_1, \ldots, s_{k-1}$ of the absolute values of $\lambda_i, 1 \leq i \leq n$. In particular, Theorem 5.2 implies that $s_1(\lambda) = \lambda_1 + \cdots + \lambda_n$ is well-conditioned in the absolute sense because

$$|s_1(\tilde{\lambda}) - s_1(\lambda)| \leq n\epsilon_{abs}.$$

Furthermore, if $\epsilon_{abs} < 1$, then $s_n(\lambda) = \lambda_1 \cdots \lambda_n$ satisfies the first order bound

$$|s_n(\tilde{\lambda}) - s_n(\lambda)| \leq s_{n-1}(|\lambda|)\, \epsilon_{abs} + \mathcal{O}(\epsilon_{abs}^2).$$

A similar result for s_n is derived in [15, Lemma 3] by means of an inequality due to Mitrinović [43, pg 315]. The first order bounds for the other elementary symmetric functions are

$$|s_k(\tilde{\lambda}) - s_k(\lambda)| \leq (n-k+1)\, s_{k-1}(|\lambda|)\, \epsilon_{abs} + \mathcal{O}(\epsilon_{abs}^2), \qquad 2 \leq k \leq n-1.$$

If $\lambda_i \geq 0$, then we can bound the absolute error in $s_k(\tilde{\lambda})$ in terms of the preceding functions $s_j(\lambda)$, $1 \leq j \leq k-1$.

Corollary 5.3. If $\tilde{\lambda}_i = \lambda_i + \epsilon_i$, $1 \leq i \leq n$, and $\lambda_i \geq 0$, then

$$|s_k(\tilde{\lambda}) - s_k(\lambda)| \leq \sum_{i=1}^{k} \binom{n-k+i}{i} s_{k-i}(\lambda)\, \epsilon_{abs}^{i}, \qquad 1 \leq k \leq n.$$

To compute error bounds of characteristic polynomials of matrices with zero eigenvalues later in this chapter, we derive the perturbation bounds for elementary symmetric functions when some λ_i are zero.

Corollary 5.4. Suppose $\lambda_1 = \ldots = \lambda_{n-r} = 0$ for some $1 \leq r \leq n-1$, then,

$$|\tilde{s}_k - s_k| \leq \epsilon_{abs}^{k-r} \sum_{i=0}^{r} \binom{n-r+i}{k-r+i} s_{r-i}(|\lambda|) \epsilon_{abs}^{i}, \qquad r+1 \leq k \leq n.$$

Proof. If r values of λ_i are non zero, then all products of $r+1$ or more λ_i are zero. This implies that $s_{r+1}(|\lambda|), \ldots, s_n(|\lambda|)$ are zero because each kth elementary symmetric function of $|\lambda_i|$ is a sum of products of k input values, and when $k > r$, then these products become zero. The surviving terms in the bound of Theorem 5.2 contain only $s_1, \ldots, s_r$.

$$\begin{aligned} |\tilde{s}_k - s_k| \;\leq\; & \binom{n-r}{k-r} s_r(|\lambda|)\epsilon_{abs}^{k-r} + \binom{n-r+1}{k-r+1} s_{r-1}(|\lambda|)\epsilon_{abs}^{k-r+1} + \\ & \binom{n-r+2}{k-r+2} s_{r-2}(|\lambda|)\epsilon_{abs}^{k-r+2} + \ldots + \binom{n}{k}\epsilon_{abs}^{k} \\ =\; & \epsilon_{abs}^{k-r} \sum_{i=0}^{r} \binom{n-r+i}{k-r+i} s_{r-i}(|\lambda|)\, \epsilon_{abs}^{i}, \qquad r+1 \leq k \leq n. \end{aligned}$$

□

Corollary 5.4 shows that when some input values λ_i are zero, then the elementary symmetric functions become better conditioned. The conditioning improves with more zero input values λ_i. We also mention that for $k \leq r$, the $s_k(|\lambda|)$ are functions of only non zero r input values $\lambda_i, 1 \leq i \leq r$, and, hence, contain fewer terms. Therefore,

the absolute bounds for the elementary symmetric functions $s_k(\lambda)$, $1 \leq k \leq r$, also improve and become tighter.

5.2.2 Relative Perturbation Bounds

We consider relative perturbations $\hat{\lambda}_i = \lambda_i(1+\epsilon_i)$, $1 \leq i \leq n$. We express λ_i and $\hat{\lambda}_i$ as elements of diagonal matrices A and $A+E$, respectively, and apply our results from chapter 3 to get bounds for elementary symmetric functions.

We first derive a bound for $s_n(\lambda) = \lambda_1 \cdots \lambda_n$ from Theorem 3.12 and then use this bound in turn to derive bounds for the other elementary symmetric functions. Theorem 3.12 states:

Let A and E be $n \times n$ complex matrices. If A is nonsingular then,

$$\frac{\det(A+E) - \det(A)}{\det(A)} = \det(A^{-1}E) + S_1 + \cdots + S_{n-1},$$

where

$$S_k \equiv \sum_{1 \leq i_1 < \cdots < i_k \leq n} \det((A^{-1}E)_{i_1 \ldots i_k}), \qquad 1 \leq k \leq n-1.$$

Here $((A^{-1}E)_{i_1 \ldots i_k})$ is the principal submatrix of order $n-k$ obtained by removing rows and columns $i_1, \ldots, i_k$ from $A^{-1}E$.

We express the relative error in $s_n(\hat{\lambda})$ in terms of elementary symmetric functions of $\epsilon = \big(\epsilon_1, \ldots, \epsilon_n\big)$ below.

Theorem 5.5. If $\hat{\lambda}_i = \lambda_i(1+\epsilon_i)$, $1 \leq i \leq n$, then

$$s_n(\hat{\lambda}) - s_n(\lambda) = s_n(\lambda) \sum_{i=1}^{n} s_i(\epsilon).$$

Proof. If $\lambda_i = 0$ for some i then $s_n(\lambda) = \lambda_1 \ldots \lambda_n = 0$. Moreover $\hat{\lambda}_i = \lambda_i(1+\epsilon_i) = 0$ so that $s_n(\hat{\lambda}) = 0$ and the desired result holds.

Now assume that $\lambda_i \neq 0$, $1 \leq i \leq n$. Define $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ so that $\det(A) = s_n(\lambda)$. Also define $E = \mathrm{diag}(\lambda_1\epsilon_1, \ldots, \lambda_n\epsilon_n)$ so that $A^{-1}E = D = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_n)$ and $\det(D) = s_n(\epsilon)$. Applying Theorem 3.12 to A and $A + E$ gives

$$s_n(\hat{\lambda}) - s_n(\lambda) = s_n(\lambda)\left(s_n(\epsilon) + S_1 + \cdots + S_{n-1}\right),$$

where

$$S_k \equiv \sum_{1 \leq i_1 < \cdots < i_k \leq n} \det(D_{i_1 \ldots i_k}), \qquad 1 \leq k \leq n-1.$$

Lemma 5.1 implies $S_k = s_{n-k}(\epsilon)$, $1 \leq k \leq n-1$. Hence $\sum_{k=1}^{n-1} S_k = \sum_{k=1}^{n-1} s_k(\epsilon)$. □

The expression in Theorem 5.5 implies the following relative error bound for $s_n(\hat{\lambda})$.

Corollary 5.6. If $\hat{\lambda}_i = \lambda_i(1 + \lambda_i)$, $1 \leq i \leq n$, then

$$|s_n(\hat{\lambda}) - s_n(\lambda)| \leq |s_n(\lambda)| \sum_{i=1}^{n} \binom{n}{i} \epsilon_{rel}^i = |s_n(\lambda)| \left[(1 + \epsilon_{rel})^n - 1\right].$$

where $\epsilon_{rel} = \max_{1 \leq i \leq n} |\epsilon_i|$.

Proof. In the right-hand side expression of Theorem 5.5 each $s_i(\epsilon)$ is a sum of $\binom{n}{i}$ terms, where each term is a product of i factors ϵ_j, $1 \leq i, j \leq n$. Therefore, $|s_i(\epsilon)| \leq \binom{n}{i}\epsilon_{rel}^i$. □

For $\epsilon_{rel} < 1$, Corollary 5.6 implies the first-order relative bound

$$|s_n(\hat{\lambda}) - s_n(\lambda)| \leq |s_n(\lambda)|\, n\epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2).$$

This means if n is sufficiently small, then $s_n(\lambda)$ is well conditioned with respect to relative perturbations in λ.

In order to explain the expansion for any elementary symmetric function, we present an example.

Example 5.7. For $n = 4$ we determine $s_3(\hat{\lambda}) - s_3(\lambda)$.

With $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_4)$, Lemma 5.1 implies that $s_3(\lambda)$ is a sum of 3×3 principal minors of A. That is,

$$s_3(\lambda) = \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_3\lambda_4 + \lambda_2\lambda_3\lambda_4.$$

Applying the expansion in Theorem 5.5 to each of these four products gives

$$\begin{aligned} s_3(\hat{\lambda}) - s_3(\lambda) \;=\; & \lambda_1\lambda_2\lambda_3\ [(\epsilon_1+\epsilon_2+\epsilon_3)+(\epsilon_1\epsilon_2+\epsilon_1\epsilon_3+\epsilon_2\epsilon_3)+\epsilon_1\epsilon_2\epsilon_3]+ \\ & \lambda_1\lambda_2\lambda_4\ [(\epsilon_1+\epsilon_2+\epsilon_4)+(\epsilon_1\epsilon_2+\epsilon_1\epsilon_4+\epsilon_2\epsilon_4)+\epsilon_1\epsilon_2\epsilon_4]+ \\ & \lambda_1\lambda_3\lambda_4\ [(\epsilon_1+\epsilon_3+\epsilon_4)+(\epsilon_1\epsilon_3+\epsilon_1\epsilon_4+\epsilon_3\epsilon_4)+\epsilon_1\epsilon_3\epsilon_4]+ \\ & \lambda_2\lambda_3\lambda_4\ [(\epsilon_2+\epsilon_3+\epsilon_4)+(\epsilon_2\epsilon_3+\epsilon_2\epsilon_4+\epsilon_3\epsilon_4)+\epsilon_2\epsilon_3\epsilon_4]. \end{aligned}$$

The ϵ terms on the right-hand side are elementary symmetric functions of three elements of ϵ. For instance, the right-hand side of

$$\hat{\lambda}_1\hat{\lambda}_2\hat{\lambda}_4 - \lambda_1\lambda_2\lambda_4 = \lambda_1\lambda_2\lambda_4\ [(\epsilon_1+\epsilon_2+\epsilon_4)+(\epsilon_1\epsilon_2+\epsilon_1\epsilon_4+\epsilon_2\epsilon_4)+\epsilon_1\epsilon_2\epsilon_4]$$

contains elementary symmetric functions of ϵ_1, ϵ_2, ϵ_4. Denote them by $s_j^{(124)}(\epsilon)$. That is,

$$s_1^{(124)}(\epsilon) = \epsilon_1+\epsilon_2+\epsilon_4, \quad s_2^{(124)}(\epsilon) = \epsilon_1\epsilon_2+\epsilon_1\epsilon_4++\epsilon_2\epsilon_4, \quad s_3^{(124)}(\epsilon) = \epsilon_1\epsilon_2\epsilon_4.$$

Then, we can write

$$\hat{\lambda}_1\hat{\lambda}_2\hat{\lambda}_4 - \lambda_1\lambda_2\lambda_4 = \lambda_1\lambda_2\lambda_4\ \left[s_1^{(124)}(\epsilon)+s_2^{(124)}(\epsilon)+s_3^{(124)}(\epsilon)\right].$$

If we do this for all four products $s_3(\hat{\lambda}) - s_3(\lambda)$, we obtain

$$\begin{aligned} s_3(\hat{\lambda}) - s_3(\lambda) &= \lambda_1\lambda_2\lambda_3 \left[s_1^{(123)}(\epsilon) + s_2^{(123)}(\epsilon) + s_3^{(123)}(\epsilon)\right] + \\ &\quad \lambda_1\lambda_2\lambda_4 \left[s_1^{(124)}(\epsilon) + s_2^{(124)}(\epsilon) + s_3^{(124)}(\epsilon)\right] + \\ &\quad \lambda_1\lambda_3\lambda_4 \left[s_1^{(134)}(\epsilon) + s_2^{(134)}(\epsilon) + s_3^{(134)}(\epsilon)\right] + \\ &\quad \lambda_2\lambda_3\lambda_4 \left[s_1^{(234)}(\epsilon) + s_2^{(234)}(\epsilon) + s_3^{(234)}(\epsilon)\right]. \end{aligned}$$

In order to extend this example to any n, we introduce notation for elementary symmetric functions of subsets of elements. For n complex numbers $\lambda_1, \ldots, \lambda_n$, and $1 \le i_1 < \ldots < i_k \le n$, we denote the jth elementary function of $\lambda_{i_1}, \ldots, \lambda_{i_k}$ by $s_j^{(i_1 \ldots i_k)}(\lambda)$, $1 \le j \le k$. In particular, $s_j^{(1\ldots n)}(\lambda) = s_j(\lambda)$, $1 \le j \le n$. Now we are ready to extend the expansion in Theorem 5.5 to other elementary symmetric functions.

Theorem 5.8. If $\hat{\lambda}_i = \lambda_i(1 + \epsilon_i)$, $1 \le i \le n$, then

$$s_k(\hat{\lambda}) - s_k(\lambda) = \sum_{1 \le i_1 < \ldots < i_k \le n} \lambda_{i_1} \ldots \lambda_{i_k} \sum_{j=1}^{k} s_j^{(i_1 \ldots i_k)}(\epsilon), \qquad 1 \le k \le n.$$

Proof. According to Lemma 5.1, each $s_k(\hat{\lambda})$ is a sum of $\binom{n}{k}$ principal minors of order k. If $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, then such a principal minor is of the form $\hat{\lambda}_{i_1} \cdots \hat{\lambda}_{i_k}$. Applying Theorem 5.5 to $\hat{\lambda}_{i_1} \cdots \hat{\lambda}_{i_k}$ yields

$$\hat{\lambda}_{i_1} \cdots \hat{\lambda}_{i_k} - \lambda_{i_1} \cdots \lambda_{i_k} = \lambda_{i_1} \cdots \lambda_{i_k} \sum_{j=1}^{k} s_j^{(i_1 \ldots i_k)}(\epsilon).$$

Summing up these expansions for all principal minors gives the desired result. □

The connection to Theorem 5.5 may be even clearer if we use the fact that

$\lambda_{i_1} \dots \lambda_{i_k} = s_k^{(i_1 \dots i_k)}(\lambda)$ and we express Theorem 5.8 as

$$s_k(\hat{\lambda}) - s_k(\lambda) = \sum_{1 \le i_1 < \dots < i_k \le n} s_k^{(i_1 \dots i_k)}(\lambda) \sum_{j=1}^{k} s_j^{(i_1 \dots i_k)}(\epsilon), \qquad 1 \le k \le n.$$

Theorem 5.8 implies the following perturbation bound.

Corollary 5.9. If $\hat{\lambda}_i = \lambda_i(1 + \epsilon_i)$, $1 \le k \le n$, then

$$|s_k(\hat{\lambda}) - s_k(\lambda)| \le s_k(|\lambda|) \sum_{j=1}^{k} \binom{k}{j} \epsilon_{rel}^j = s_k(|\lambda|) \left((1 + \epsilon_{rel})^k - 1 \right),$$

$\epsilon_{rel} = \max_{1 \le i \le n} |\epsilon_i|$.

Proof. Applying the triangle inequality to the expression in Theorem 5.8 gives

$$|s_k(\hat{\lambda}) - s_k(\lambda)| \le \sum_{1 \le i_1 < \dots < i_k \le n} |\lambda_{i_1}| \dots |\lambda_{i_k}| \sum_{j=1}^{k} |s_j^{(i_1 \dots i_k)}(\epsilon)|, \qquad 1 \le k \le n.$$

The elementary symmetric functions can be bounded by $|s_j^{(i_1 \dots i_k)}(\epsilon)| \le \binom{k}{j} \epsilon_{rel}^j$. Summing up all the bounds yields

$$\begin{aligned} \sum_{1 \le i_1 < \dots < i_k \le n} |\lambda_{i_1}| \dots |\lambda_{i_k}| \sum_{j=1}^{k} |s_j^{(i_1 \dots i_k)}(\epsilon)| &\le \sum_{1 \le i_1 < \dots < i_k \le n} |\lambda_{i_1}| \dots |\lambda_{i_k}| \sum_{j=1}^{k} \binom{k}{j} \epsilon_{rel}^j \\ &= s_k(|\lambda|) \sum_{j=1}^{k} \binom{k}{j} \epsilon_{rel}^j. \end{aligned}$$

□

Remark 5.10 (Relative Error Bound)**.** Corollary 5.9 implies the following relative error bound for $s_k(\lambda) \neq 0$:

$$\frac{|s_k(\hat{\lambda}) - s_k(\lambda)|}{|s_k(\lambda)|} \le \frac{s_k(|\lambda|)}{|s_k(\lambda)|} \left((1 + \epsilon_{rel})^k - 1 \right).$$

The bound suggests that $s_k(\lambda)$ is sensitive to relative perturbations in λ if $s_k(|\lambda|) \gg |s_k(\lambda)|$. If $\epsilon_{rel} < 1$, then we obtain to first order

$$\frac{|s_k(\hat{\lambda}) - s_k(\lambda)|}{|s_k(\lambda)|} \leq k \, \frac{s_k(|\lambda|)}{|s_k(\lambda)|} \epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2).$$

Thus, $ks_k(|\lambda|)/|s_k(\lambda)|$ can be interpreted as a first-order relative condition number for $s_k(\lambda)$.

When all λ_i are positive, $s_k(\lambda)$ is insensitive to relative perturbations in λ provided k is sufficiently small.

Corollary 5.11. If $\hat{\lambda}_i = \lambda_i(1 + \epsilon_i)$, and $\lambda_i > 0$, $1 \leq i \leq n$, then

$$\begin{aligned} \frac{|s_k(\hat{\lambda}) - s_k(\lambda)|}{s_k(\lambda)} &\leq (1 + \epsilon_{rel})^k - 1, \qquad 1 \leq k \leq n, \\ &= k\epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2). \end{aligned}$$

Corollary 5.11 implies that elementary symmetric functions are well-conditioned with respect to relative perturbations, if the input values are positive.

We can derive another weaker bound from Corollary 5.9.

Corollary 5.12. If $\hat{\lambda}_i = \lambda_i(1 + \epsilon_i)$, $1 \leq i \leq n$, and $n\epsilon_{rel} < 1$, then

$$\frac{|s_k(\hat{\lambda}) - s_k(\lambda)|}{|s_k(\lambda)|} \leq \frac{s_k(|\lambda|)}{|s_k(\lambda)|} \, \frac{k\epsilon_{rel}}{1 - k\epsilon_{rel}}, \qquad 1 \leq k \leq n.$$

Proof. Write

$$(1 + \epsilon_{rel})^k = \sum_{j=0}^{k} \binom{k}{j} \epsilon_{rel}^j \leq \sum_{j=0}^{k} k^j \epsilon_{rel}^j \leq \sum_{j=0}^{\infty} (k\epsilon_{rel})^j.$$

From $k\epsilon_{rel} \leq n\epsilon_{rel} < 1$ follows $\sum_{j=0}^{\infty}(k\epsilon_{rel})^j = \frac{1}{1-k\epsilon_{rel}}$, so that

$$(1+\epsilon_{rel})^k - 1 \leq \frac{1}{1-k\epsilon_{rel}} - 1 = \frac{k\epsilon_{rel}}{1-k\epsilon_{rel}}.$$

□

If $\lambda_i > 0$, then Corollary 5.12 reproduces [8, Proposition 7.1].

$$\frac{|s_k(\hat{\lambda}) - s_k(\lambda)|}{s_k(\lambda)} \leq \frac{k\epsilon_{rel}}{1-k\epsilon_{rel}}, \qquad 1 \leq k \leq n.$$

Computing Elementary Symmetric Functions

We derived perturbation bounds for elementary symmetric functions $s_k(\lambda)$ of complex numbers $\lambda_1, \ldots, \lambda_n$. We now consider the computation of elementary symmetric functions $s_k(\lambda)$. There are $\binom{n}{k}$ summands in each $s_k(\lambda)$ for a given set of complex numbers $\lambda_1, \ldots, \lambda_n$; therefore, straightforward computation of $s_k(\lambda)$ is very expensive. We need an efficient and numerically stable algorithm for the computation of elementary symmetric functions. A paper by Baker and Harwell presents a collection of algorithms [2]. These algorithms include the Difference Algorithm, the Summation Algorithm and the Grouping Property Algorithm. Baker and Harwell mention that these algorithms have not been studied for numerical stability. We show that the Summation Algorithm is forward stable. If $\lambda_i > 0$, then there are no subtractions in the computation of the elementary symmetric functions. Subtractions of almost equal numbers may cause inaccurate results when computations are carried out in floating point arithmetic. Therefore, for $\lambda_i > 0$ the computed elementary symmetric functions from the Summation Algorithm are accurate. We describe the algorithm in detail below.

5.2.3 The Summation Algorithm

In this section we adopt the following notation. We denote the kth elementary symmetric function of i numbers $\lambda_1, \ldots, \lambda_i$ by $s_k^{(i)}$. In particular $s_k^{(n)} = s_k(\lambda)$. This notation will help in understanding the Summation Algorithm better. We can compute the elementary symmetric functions recursively [11], see also [13, pp 250, eqn. 14.3.11]. We present the Summation Algorithm below and describe the recursion by an example.

Algorithm 1 Summation Algorithm

Input: $\lambda_1, \cdots, \lambda_n$

Output: Elementary symmetric functions $s_k(\lambda)$

Set $s_0^{(l)} = 1, \quad 1 \leq l \leq n-1$
Set $s_k^{(l)} = 0$ for $k > l$
Set $s_1^{(1)} = \lambda_1$
for $i = 2 : n$ **do**
    for $k = 1 : i$ **do**
        $s_k^{(i)} = s_k^{(i-1)} + \lambda_i \, s_{k-1}^{(i-1)}$
    end for
end for
{At the end of recursion: $s_k^{(n)} = s_k(\lambda)$}

Example 5.13. Consider the computation of elementary symmetric functions for four complex numbers $\lambda_1, \ldots, \lambda_4$ by the Summation Algorithm. The recursion begins by establishing an elementary symmetric function of order 1. For i=1, we have $s_1^{(1)} = \lambda_1$. For $i = 2$, we obtain

$$\begin{aligned} s_1^{(2)} &= s_1^{(1)} + \lambda_2 \, s_0^{(1)} = \lambda_1 + \lambda_2, \\ s_2^{(2)} &= s_2^{(1)} + \lambda_2 \, s_1^{(1)} = \lambda_2 \lambda_1. \end{aligned}$$

Similarly, for $i = 3$, we have

$$\begin{aligned} s_1^{(3)} &= s_1^{(2)} + \lambda_3 s_0^{(2)} = \lambda_1 + \lambda_2 + \lambda_3, \\ s_2^{(3)} &= s_2^{(2)} + \lambda_3 s_1^{(2)} = \lambda_2\lambda_1 + \lambda_3(\lambda_1 + \lambda_2), \\ s_3^{(3)} &= s_3^{(2)} + \lambda_3 s_2^{(2)} = \lambda_3\lambda_2\lambda_1. \end{aligned}$$

In the final step of the recursion $i = 4$, so that $s_k^{(4)} = s_k(\lambda)$, $1 \le k \le 4$. We get all elementary symmetric functions as follows:

$$\begin{aligned} s_1(\lambda) &= s_1^{(4)} = s_1^{(3)} + \lambda_4 s_0^{(3)} = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4, \\ s_2(\lambda) &= s_2^{(4)} = s_2^{(3)} + \lambda_4 s_1^{(3)} = \lambda_2\lambda_1 + \lambda_3(\lambda_1 + \lambda_2) + \lambda_4(\lambda_1 + \lambda_2 + \lambda_3), \\ s_3(\lambda) &= s_3^{(4)} = s_3^{(3)} + \lambda_4 s_2^{(3)} = \lambda_3\lambda_2\lambda_1 + \lambda_4\left[\lambda_2\lambda_1 + \lambda_3(\lambda_1 + \lambda_2)\right], \\ s_4(\lambda) &= s_4^{(4)} = s_4^{(3)} + \lambda_4 s_3^{(3)} = \lambda_4\lambda_3\lambda_2\lambda_1. \end{aligned}$$

Note that $s_n(\lambda)$ is computed as a product $\lambda_1 \ldots \lambda_n$. The Summation Algorithm requires $\frac{n(n-1)}{2}$ multiplications and $\frac{n(n-1)}{2} + (n+1)$ additions [2].

We present a forward rounding error analysis of the Summation Algorithm below.

5.2.4 Forward Stability of the Summation Algorithm for Real Numbers

Suppose $\lambda_1, \ldots, \lambda_n$ are real numbers. To carry out rounding error analysis of the Summation Algorithm we use the following models of basic floating point operations [25, §2.2].

In the standard floating point model for real floating point numbers x and y, assuming no underflow or overflow, we have

$$\mathrm{fl}(x \; op \; y) = (x \; op \; y)(1 + \delta), \quad |\delta| \le u, \quad \mathrm{op} = +, -, \times, /, \tag{5.1}$$

where u is the unit roundoff. The modified standard floating point model is given as follows:

$$\mathrm{fl}(x\ op\ y) = \frac{x\ op\ y}{1+\epsilon}, \quad |\epsilon| \le u. \tag{5.2}$$

The following relations are required for our error analysis.

Lemma 5.14 (§3.1, §3.4 in [25]). 1. If $|\delta_i| \le u$ and $\rho_i = \pm 1$ for $1 \le i \le n$ and $nu < 1$, then

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1+\theta_n,$$

where

$$|\theta_n| \le \frac{nu}{1-nu} =: \gamma_n.$$

2. For positive integers j and l,

$$(1+\theta_j)(1+\theta_l) = (1+\theta_{j+l}).$$

Whenever we write γ_n, there is an implicit assumption that $nu < 1$.

5.2.4.1 Worst Case Error Bounds

The Lemma below leads to forward error bounds for $s_k(\lambda)$, $1 \le k \le n-1$.

Lemma 5.15. Denote the computed elementary symmetric functions from the Summation Algorithm by $\hat{s}_k(\lambda)$, $1 \le k \le n-1$. If $\hat{s}_k(\lambda)$ are computed as

$$\hat{s}_k^{(i)} = \mathrm{fl}\left[\hat{s}_k^{(i-1)} + \mathrm{fl}\left[\lambda_i \hat{s}_{k-1}^{(i-1)}\right]\right], \quad 2 \le i \le n.$$

Then, for $1 \le k \le n-1$,

$$\begin{aligned}\hat{s}_k(\lambda) &= \hat{s}_k^{(n)} = \mathrm{fl}\left[\hat{s}_k^{(n-1)} + \mathrm{fl}\left[\lambda_n \hat{s}_{k-1}^{(n-1)}\right]\right] \\ &= \sum_{1\le i_1<\ldots<i_k\le n} \lambda_{i_1}\cdots\lambda_{i_k}(1+\theta_t^{(i_1\ldots i_k)}),\end{aligned}$$

where t is a value that satisfies $1 \leq t \leq 2n$.

Proof. The poof of Lemma 5.15 is by induction on the number of data $\lambda_1, \ldots, \lambda_i$. For $i = 1$, no floating point operation occurs. For $i = 2$, using the standard floating point model (5.1),

$$\hat{s}_1^{(2)} = (\lambda_1 + \lambda_2)(1 + \delta_1) = (\lambda_1 + \lambda_2)(1 + \theta_1),$$

where we used relation 1 of Lemma 5.14 to replace $(1 + \delta_1)$ with $(1 + \theta_1)$. Similarly,

$$\hat{s}_2^{(2)} = \lambda_1\lambda_2(1 + \delta_2) = \lambda_1\lambda_2(1 + \theta_1^{(12)}).$$

Hence, the statement of Lemma 5.15 is true for $i = 2$. We suppose that the statement of Lemma 5.15 is true for $n - 1$ input values λ_i and prove that the statement is also true for n input values. Using the Summation Algorithm under the standard floating point model 5.1, the computed $\hat{s}_k^{(n)}$ is

$$\hat{s}_k^{(n)} = \hat{s}_k^{(n-1)}(1 + \delta_3) + \lambda_n\hat{s}_{k-1}^{(n-1)}(1 + \delta_3)(1 + \delta_4) = \hat{s}_k^{(n-1)}(1 + \theta_1) + \lambda_n\hat{s}_{k-1}^{(n-1)}(1 + \theta_2).$$

From the induction hypothesis on $\hat{s}_k^{(n-1)}$,

$$\hat{s}_k^{(n-1)} = \sum_{1 \leq i_1 < \ldots < i_k \leq n-1} \lambda_{i_1} \cdots \lambda_{i_k}(1 + \theta_l^{(i_1 \ldots i_k)}), \quad 1 \leq l \leq 2n - 2.$$

Similarly, from the induction hypothesis on $\hat{s}_{k-1}^{(n-1)}$,

$$\hat{s}_{k-1}^{(n-1)} = \sum_{1 \leq i_1 < \ldots < i_{k-1} \leq n-1} \lambda_{i_1} \cdots \lambda_{i_{k-1}}(1 + \theta_m^{(i_1 \ldots i_{k-1})}), \quad 1 \leq m \leq 2n - 2.$$

Now we write

$$\begin{aligned}\hat{s}_k^{(n)} &= \sum_{1\le i_1<\ldots<i_k\le n-1} \lambda_{i_1}\cdots\lambda_{i_k}(1+\theta_l^{(i_1\ldots i_k)})(1+\theta_1) + \\ &\quad \lambda_n \sum_{1\le i_1<\ldots<i_{k-1}\le n-1} \lambda_{i_1}\ldots\lambda_{i_{k-1}}(1+\theta_m^{(i_1\ldots i_{k-1})})(1+\theta_2).\end{aligned}$$

We use relation 2 of Lemma 5.14 on both summands of the above equation and obtain

$$\hat{s}_k^{(n)} = \sum_{1\le i_1<\ldots<i_k\le n} \lambda_{i_1}\cdots\lambda_{i_k}(1+\theta_t^{(i_1\ldots i_k)}), \quad 1\le t\le 2n.$$

□

Lemma 5.15 leads to the following forward error in $\hat{s}_k(\lambda)$.

Theorem 5.16. Denote the computed elementary symmetric functions from the Summation Algorithm by $\hat{s}_k(\lambda)$, $1\le k\le n-1$. If $\hat{s}_k(\lambda)$ are computed as

$$\hat{s}_k(\lambda) = \mathrm{fl}\left[\hat{s}_k^{(n-1)} + fl\left[\lambda_n \hat{s}_{k-1}^{(n-1)}\right]\right],$$

and $2nu<1$, then

$$|\hat{s}_k(\lambda) - s_k(\lambda)| \;\le\; \gamma_{2n}s_k(|\lambda|) \le \frac{2nu}{1-2nu}s_k(|\lambda|),$$

and

$$|\hat{s}_n(\lambda) - s_n(\lambda)| \;\le\; \gamma_{n-1}|s_n(\lambda)| \le \frac{(n-1)u}{1-(n-1)u}|s_n(\lambda)|.$$

Proof. For $1\le k\le n-1$, Lemma 5.15 implies

$$\hat{s}_k(\lambda) = \sum_{1\le i_1<\ldots<i_k\le n} \lambda_{i_1}\cdots\lambda_{i_k}(1+\theta_t^{(i_1\ldots i_k)}), \quad 1\le t\le 2n. \tag{5.3}$$

Expanding the right hand side of equation (5.3) gives

$$\hat{s}_k(\lambda) = s_k(\lambda) + \sum_{1 \le i_1 < \ldots < i_k \le n} \theta_t^{(i_1 \ldots i_k)} \lambda_{i_1} \cdots \lambda_{i_k}, \quad \text{where} \quad 1 \le t \le 2n.$$

We apply the triangle inequality on every term of the difference $|\hat{s}_k(\lambda) - s_k(\lambda)|$ to obtain

$$|\hat{s}_k(\lambda) - s_k(\lambda)| \le \sum_{1 \le i_1 < \ldots < i_k \le n} |\theta_t^{(i_1 \ldots i_k)}||\lambda_{i_1} \cdots \lambda_{i_k}|, \quad 1 \le t \le 2n.$$

We use relation 1 of Lemma 5.14 to replace every $|\theta_t^{(i_1 \ldots i_k)}|$ by γ_{2n} to get the desired result. The conclusion for $\hat{s}_n(\lambda)$ follows from the fact that the Summation Algorithm computes $s_n(\lambda)$ as a product $\lambda_1 \cdots \lambda_n$. □

Remark 5.17. A comparison of Theorem 5.16 with Corollary 5.12 shows that the Summation Algorithm is forward stable. The error bounds of the computed elementary symmetric functions $\hat{s}_k(\lambda)$ from the Summation Algorithm are small multiples of the condition numbers of $s_k(\lambda)$, $1 \le k \le n$.

Theorem 5.16 also shows that the Summation Algorithm computes $s_n(\lambda)$ and elementary symmetric functions of positive values $\lambda_1, \ldots, \lambda_n$ with high relative accuracy.

Corollary 5.18. If the Summation Algorithm is used to compute the elementary symmetric functions of $\lambda_i > 0$, $1 \le i \le n$, then

$$\frac{|\hat{s}_k(\lambda) - s_k(\lambda)|}{s_k(\lambda)} \le \gamma_{2n} \le \frac{2nu}{1 - 2nu}, \quad 1 \le k \le n.$$

5.2.4.2 Running Error Analysis

The error bounds of the Summation Algorithm in Theorem 5.16 are worst case bounds that do not depend on actual rounding errors committed during the computations. These bounds do not take into account the intermediate quantities in which cancellation can occur. We derive sharper running error bounds for the Summation

Algorithm. The underlying idea is to compute the error bounds from the computed values of elementary symmetric functions at every step of the recursion. In this way, we can take advantage of cancellation that might occur in intermediate quantities. There are, of course, rounding errors in the computation of the running error bounds, but their effect is negligible [25, §3.3].

In our running error analysis, we use both standard and modified standard floating point models. We compute $\hat{s}_k^{(i)}$ as

$$\hat{s}_k^{(i)} = \mathrm{fl}\left[\hat{s}_k^{(i-1)} + \mathrm{fl}\left[\lambda_i \hat{s}_{k-1}^{(i-1)}\right]\right].$$

In the beginning of the recursion, when $i = 1$, we have $\hat{s}_1^{(1)} = s_1^{(1)} = \lambda_1$. For $2 \le i \le n$, we write

$$\hat{s}_k^{(i)} = s_k^{(i)} + e_k^{(i)}.$$

Here $s_k^{(i)}$ is the exact kth elementary symmetric function of i input values λ_i, and $e_k^{(i)}$ is the error in $\hat{s}_k^{(i)}$. We first give the running error bounds for $\hat{s}_1(\lambda)$.

Theorem 5.19. The following recursion produces the running error bound for $\hat{s}_1(\lambda)$,

$$|e_1^{(i)}| \le |e_1^{(i-1)}| + u|\hat{s}_1^{(i)}|, \quad 2 \le i \le n.$$

Proof. Using the modified floating point model (5.2), we write

$$\hat{s}_1^{(i)} = \frac{\hat{s}_1^{(i-1)} + \lambda_i}{1 + \epsilon^{(i)}}, \quad |\epsilon^{(i)}| \le u, \quad 2 \le i \le n.$$

This implies

$$(1 + \epsilon^{(i)})\hat{s}_1^{(i)} = \hat{s}_1^{(i-1)} + \lambda_i.$$

Distributing

$$\hat{s}_1^{(i)} + \epsilon^{(i)}\hat{s}_1^{(i)} = \hat{s}_1^{(i-1)} + \lambda_i.$$

We write $\hat{s}_1^{(i)}$ and $\hat{s}_1^{(i-1)}$ in terms of their errors, and simplify to get

$$e_1^{(i)} = e_1^{(i-1)} - \epsilon^{(i)}\hat{s}_1^{(i)}.$$

We use the triangle inequality to obtain

$$|e_1^{(i)} \leq |e_1^{(i-1)}| + u|\hat{s}_1^{i)}|.$$

□

We now give running error bounds for $\hat{s}_k(\lambda)$, $2 \leq k \leq n-1$.

Theorem 5.20. The error in $\hat{s}_k^{(k)}$, $2 \leq k \leq n-1$, can be calculated from the following recursion:

$$|e_k^{(k)}| \leq |\lambda_k e_{k-1}^{(k-1)}| + u|\hat{s}_k^{(k)}|,$$

and for $k < i \leq n$, the error in $\hat{s}_k^{(i)}$ is given by the following recursion:

$$|e_k^{(i)}| \leq |e_k^{(i-1)}| + |\lambda_i e_{k-1}^{(i-1)}| + u\left(|\lambda_i \hat{s}_{k-1}^{(i-1)}| + |\hat{s}_k^{(i)}|\right).$$

Proof. Note that for $k > i$, $s_k^{(i)} = 0$. Therefore, we start accumulating errors in $\hat{s}_k^{(i)}$ at step $i = k$. At step $i = k$, we compute

$$\hat{s}_k^{(k)} = \mathrm{fl}\left[\lambda_k \hat{s}_{k-1}^{(k-1)}\right].$$

We then use the modified floating point model (5.2) to write

$$(1+\epsilon)\hat{s}_k^{(k)} = \lambda_k \hat{s}_{k-1}^{(k-1)}, \quad \text{where} \quad |\epsilon| \leq u.$$

We write $\hat{s}_k^{(k)}$ and $\hat{s}_{k-1}^{(k-1)}$ in terms of their errors and simplify to obtain

$$e_k^{(k)} = \lambda_k e_{k-1}^{(k-1)} - \epsilon \hat{s}_k^{(k)}.$$

Applying the triangle inequality gives the running error bound for $\hat{s}_k^{(k)}$.

For $k < i \leq n$, we use both standard and modified floating point models to get

$$(1 + \epsilon^{(i)})\hat{s}_k^{(i)} = \hat{s}_k^{(i-1)} + \lambda_i \hat{s}_{k-1}^{(i-1)}(1 + \delta^{(i)}),$$

where $|\epsilon^{(i)}|, |\delta^{(i)}| \leq u$. Writing $\hat{s}_k^{(i)}$ and $\hat{s}_{k-1}^{(i-1)}$ in terms of their errors and simplifying produces the following error in $\hat{s}_k^{(i)}$:

$$e_k^{(i)} = e_k^{(i-1)} + \lambda_i e_{k-1}^{(i-1)} + \delta^{(i)} \lambda_i \hat{s}_{k-1}^{(i-1)} - \epsilon^{(i)} \hat{s}_k^{(i)}.$$

We use the triangle inequality to get the forward error bound. □

5.2.5 Forward Stability of the Summation Algorithm for Complex Numbers

For simplicity, we presented the error analysis for elementary symmetric functions of real values first. The error analysis of the Summation Algorithm is still valid for complex input values. However, the constants in the forward error bounds for complex values increase modestly in comparison to real values. The standard model for addition and multiplication of complex numbers x and y in the absence of underflow or overflow implies [25, §3.6],

$$\mathrm{fl}(x \pm y) = (x \pm y)(1 + \delta), \quad |\delta| \leq u, \tag{5.4}$$

and

$$\mathrm{fl}(xy) = xy(1 + \delta), \quad |\delta| \leq \gamma_3. \tag{5.5}$$

Similarly, under the modified model,

$$\mathrm{fl}(x \pm y) = \frac{(x + y)}{1 + \epsilon}, \quad |\epsilon| \leq u, \tag{5.6}$$

and

$$\mathrm{fl}(xy) = \frac{xy}{1+\eta}, \quad |\eta| \leq \gamma_3. \tag{5.7}$$

We also use the following relation in our error analysis [25, Lemma 3.3].

$$\gamma_k + \gamma_j + \gamma_k\gamma_j \leq \gamma_{k+j}. \tag{5.8}$$

Worst Case Bounds

We derive worst case error bounds for the Summation Algorithm applied to complex numbers. For our error analysis, we use the following Lemma.

Lemma 5.21. The following bound holds for a product $\lambda_1 \ldots \lambda_j$ of floating point complex numbers.

$$\mathrm{fl}(\lambda_1 \ldots \lambda_j) = \lambda_1 \ldots \lambda_j(1+\alpha), \quad \text{where} \quad |\alpha| \leq \gamma_{3(j-1)}.$$

Proof. The proof is by induction on j. For $j = 1$, the statement is trivial.

For $j = 2$, the statement follows from (5.5). We assume that the statement is true for $j = n-1$ and prove it for $j = n$.

$$\mathrm{fl}\Big[\lambda_1 \ldots \lambda_n\Big] = \mathrm{fl}\Big[\lambda_1 \ldots \lambda_{n-1}\Big]\lambda_n(1+\delta), \quad \text{where} \quad |\delta| \leq \gamma_3.$$

Applying the induction hypothesis to $\mathrm{fl}\Big[\lambda_1 \ldots \lambda_{n-1}\Big]$, we have

$$\mathrm{fl}[\lambda_1 \ldots \lambda_n] = (\lambda_1 \ldots \lambda_{n-1})\lambda_n(1+\alpha_1)(1+\delta), \quad \text{where} \quad |\alpha_1| \leq \gamma_{3(n-2)}.$$

Let us define

$$1 + \alpha = (1+\alpha_1)(1+\delta) = 1 + \alpha_1 + \delta + \alpha_1\delta.$$

This implies

$$|\alpha| \leq |\alpha_1| + |\delta| + |\alpha_1\delta| \leq \gamma_{3(n-2)} + \gamma_3 + \gamma_3\gamma_{3(n-2)}.$$

From (5.8), we have $|\alpha| \leq \gamma_{3(n-1)}$.

□

Forward error bounds for elementary symmetric functions $\hat{s}_k(\lambda)$, $1 \leq k \leq n$, computed from the Summation Algorithm are derived below.

Theorem 5.22. If the elementary symmetric functions $s_k(\lambda)$ of complex numbers $\lambda_1, \ldots, \lambda_n$ are computed from the Summation Algorithm as follows,

$$\hat{s}_k^{(n)} = \mathrm{fl}\left[\hat{s}_k^{(n-1)} + \mathrm{fl}\left[\lambda_n \hat{s}_{k-1}^{(n-1)}\right]\right],$$

then the forward error in $\hat{s}_k(\lambda)$, $1 \leq k \leq n-1$, is bounded by

$$|\hat{s}_k(\lambda) - s_k(\lambda)| \leq \gamma_{2(n+k-1)} s_k(|\lambda|), \quad 1 \leq k \leq n-1.$$

For $s_n(\lambda)$ we have

$$|\hat{s}_n(\lambda) - s_n(\lambda)| \leq \gamma_{3(n-1)} |s_n(\lambda)|.$$

Proof. Rounding errors in the computation of $s_k(\lambda)$ arise from multiplication and addition of complex numbers. Each term in $\hat{s}_k(\lambda)$ is a product of k complex numbers. Therefore,

$$\hat{s}_k(\lambda) = \sum_{1 \leq i_1 < \ldots < i_k \leq n} \lambda_{i_1} \cdots \lambda_{i_k} (1 + \theta_l^{(i_1 \ldots i_k)})(1 + \alpha_1^{(i_1 \ldots i_k)}),$$

where from the standard model (5.4) for addition of complex numbers, $|\theta_l^{(i_1 \ldots i_k)}| \leq \gamma_{2n-k+1}$, and from multiplication of k complex numbers using Lemma 5.21, $|\alpha_1^{(i_1 \ldots i_k)}| \leq \gamma_{3(k-1)}$. Let us write

$$(1 + \theta_l^{(i_1 \ldots i_k)})(1 + \alpha_1^{(i_1 \ldots i_k)}) = (1 + \alpha^{(i_1 \ldots i_k)}),$$

where

$$\alpha^{(i_1\ldots i_k)} = \theta_l^{(i_1\ldots i_k)} + \alpha_1^{(i_1\ldots i_k)} + \theta_l^{(i_1\ldots i_k)}\alpha_1^{(i_1\ldots i_k)}.$$

Inequality (5.8) implies

$$|\alpha^{(i_1\ldots i_k)}| \leq |\theta_l^{(i_1\ldots i_k)}| + |\alpha_1^{(i_1\ldots i_k)}| + |\theta_l^{(i_1\ldots i_k)}\alpha_l^{(i_1\ldots i_k)}| \leq \gamma_{2(n+k-1)}.$$

Then,

$$\hat{s}_k(\lambda) = s_k(\lambda) + \sum_{1\leq i_1<\ldots<i_k\leq n} \alpha^{(i_1,\ldots,i_k)}\lambda_{i_1}\cdots\lambda_{i_k}.$$

Applying the triangle inequality to the difference $|\hat{s}_k(\lambda) - s_k(\lambda)|$ provides the forward error bound for $\hat{s}_k(\lambda)$, $1 \leq k \leq n-1$. The bound for $s_n(\lambda)$ follows from Lemma 5.21. □

Running Error Bounds

Running error bounds for the Summation Algorithm can also be derived by using standard and modified models for complex numbers.

Theorem 5.23. If the Summation Algorithm is used to compute elementary symmetric functions of complex numbers $\lambda_1, \ldots, \lambda_n$, then the error in $\hat{s}_1(\lambda)$ is given by the following recursion:

$$|e_1^{(i)}| \leq |e_1^{(i-1)}| + u|\hat{s}_1^{(i)}|, \quad 2 \leq i \leq n.$$

Proof. The proof is similar to Theorem 5.19 and follows from applying the modified model for complex numbers. □

Running error bounds for $s_k(\lambda)$, $2 \leq k \leq n$, are given as follows.

Theorem 5.24. If the Summation Algorithm is used to compute elementary symmetric functions of complex numbers $\lambda_1, \ldots, \lambda_n$, then the error in $\hat{s}_k^{(k)}$, $2 \leq k \leq n$, is

given by

$$|e_k^{(k)}| \leq |\lambda_k e_{k-1}^{(k-1)}| + \gamma_3 |\hat{s}_k^{(k)}|.$$

For $k < i \leq n$, the error in $\hat{s}_k^{(i)}$ is bounded by

$$|e_k^{(i)}| \leq |e_k^{(i-1)}| + |\lambda_i e_{k-1}^{(i-1)}| + \gamma_3 |\lambda_i \hat{s}_{k-1}^{(i-1)}| + u|\hat{s}_k^{(k)}|.$$

Proof. The proof is similar to that of Theorem 5.20 and follows by applying standard and modified models for complex numbers. □

5.3 Bounds for Characteristic Polynomials from Eigenvalue Perturbations

The method of computing coefficients of the characteristic polynomial of a matrix A from its eigenvalues consists of two steps. The eigenvalues of A are computed in the first step. In the second step, the coefficients of the characteristic polynomial of A are determined from the computed eigenvalues. We derive perturbation bounds for coefficients c_k, $1 \leq k \leq n$, of the characteristic polynomial $p(\lambda)$ of A in terms of perturbations in eigenvalues. These bounds apply when the characteristic polynomial of A is determined from its computed eigenvalues. These bounds are derived from perturbation bounds of elementary symmetric functions presented in the previous section.

5.3.1 Absolute Perturbation Bounds

We first state absolute forward error bounds for coefficients c_k, $1 \leq k \leq n$, in terms of absolute errors in eigenvalues of A. In this section we denote the eigenvalues of $A + E$ by $\tilde{\lambda}_i$ and the eigenvalues of A by λ_i.

Corollary 5.25. If $|\tilde{\lambda}_i - \lambda_i| = \epsilon_i$ and $\max_i |\epsilon_i| = \epsilon_{abs} < 1$, then

$$|\tilde{c}_k - c_k| \leq (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs} + \mathcal{O}(\epsilon_{abs}^2), \quad 1 \leq k \leq n.$$

Proof. The proof follows from Theorem 5.2 and

$$|\tilde{c}_k - c_k| = |\tilde{s}_k(\lambda) - s_k(\lambda)|,$$

which is a result of Lemma 5.1. □

The above error bound suggests that if the maximum absolute error ϵ_{abs} in the eigenvalues of A is less than 1, and n is not too large, then $s_{k-1}(|\lambda|)$ is the first order condition number for c_k with respect to absolute perturbations in eigenvalues of A. If $\lambda_i > 0$, then the error in the kth coefficient c_k can be expressed in terms of the preceding coefficient c_{k-1}.

Corollary 5.26. If $\lambda_i > 0$ and $\epsilon_{abs} < 1$, then

$$|\tilde{c}_k - c_k| \leq (n-k+1)|c_{k-1}|\epsilon_{abs} + \mathcal{O}(\epsilon_{abs}^2), \quad 1 \leq k \leq n.$$

Proof. The proof follows from Corollary 5.25 and Lemma 5.1. □

We apply the results of Corollary 5.25 and Corollary 5.26 to general matrices. We substitute the value of ϵ_{abs} to derive absolute forward error bounds for coefficients of characteristic polynomials when the coefficients are determined from computed eigenvalues.

General Matrices

For a general matrix A we have the following bound for ϵ_{abs}.

Theorem 5.27 (Corollary 2.2 in [45])**.** Let A and $A+E$ be $n \times n$ complex matrices. Let $Q^{-1}AQ = diag(J_1, \ldots, J_l)$ be the Jordan form of A and m the order of the largest

Jordan block. Define $\rho = \max\left\{\|\sqrt{n}Q^{-1}EQ\|_2, \|\sqrt{n}Q^{-1}EQ\|_2^{1/m}\right\}$. There exists a permutation τ of $\{1, 2, \ldots, n\}$ such that

$$\epsilon_{abs} = \max_{1\leq i\leq n} |\tilde{\lambda}_{\tau(i)} - \lambda_i| \leq \sqrt{n}(1+\sqrt{n-l})\rho.$$

The above bound suggests that for a small n the magnitude of largest absolute error ϵ_{abs} depends upon the quantity $\rho = \max\left\{\|\sqrt{n}Q^{-1}EQ\|_2, \|\sqrt{n}Q^{-1}EQ\|_2^{1/m}\right\}$. In general, there do not exist conditions under which $\epsilon_{abs} < 1$. However, if we assume that $\rho < \frac{1}{\sqrt{n}(1+\sqrt{n-l})}$, then we get the following bound for the coefficients of the characteristic polynomial of A.

Theorem 5.28. Under the assumptions of Theorem 5.27 with $\rho < \frac{1}{\sqrt{n}(1+\sqrt{n-l})}$,

$$|\tilde{c}_k - c_k| \leq \sqrt{n}(n-k+1)(1+\sqrt{n-l})s_{k-1}(|\lambda|)\rho + \mathcal{O}(\rho)^2, \quad 1 \leq k \leq n.$$

Proof. This follows from Corollary 5.25 and Theorem 5.27. □

In the following corollaries we assume that $\rho < \frac{1}{\sqrt{n}(1+\sqrt{n-l})}$.

Corollary 5.29. If all eigenvalues of A are positive, then

$$|\tilde{c}_k - c_k| \leq \sqrt{n}(1+\sqrt{n-l})(n-k+1)|c_{k-1}|\rho + \mathcal{O}(\rho^2), 1 \leq k \leq n.$$

Proof. This follows from Corollary 5.26 and Theorem 5.27. □

So, the error in the kth coefficient c_k of A with all positive eigenvalues depends upon the magnitude of the preceding coefficient c_{k-1}.

The following Corollary shows that coefficients of the characteristic polynomial of a matrix with some zero eigenvalues are better conditioned.

Corollary 5.30. If $\lambda_1 = \ldots = \lambda_{n-r} = 0$ for some $1 \leq r \leq n-1$, then

$$|\tilde{c}_k - c_k| \leq \left(\sqrt{n}(1+\sqrt{n-l})\right)^{k-r} \binom{n-r}{k-r} s_r(|\lambda|)\rho^{k-r} + \mathcal{O}(\rho^{k-r+1}), \quad r+1 \leq k \leq n.$$

Proof. The proof follows from Corollary 5.4 and Lemma 5.1. □

Diagonalizable Matrices. For a diagonalizable matrix A in the bound of Theorem 5.27, $m = 1$ and $n = l$. We get the following bound for ϵ_{abs}.

Theorem 5.31 (Theorem 2.4 in [45])**.** Let A be a diagnolizable matrix and $A + E$ be a complex matrix. Define $\rho = \sqrt{n}\|Q^{-1}EQ\|_2$. There exists a permutation τ of $\{1, 2, \ldots, n\}$ such that

$$\epsilon_{abs} = \max_{1 \leq i \leq n} |\tilde{\lambda}_{\tau(i)} - \lambda_i| \leq \sqrt{n}\rho.$$

Theorem 5.31 implies the following bound for coefficients of the characteristic polynomial of a diagonalizable matrix A.

Theorem 5.32. Let A be a diagonalizable matrix with $\rho < \frac{1}{\sqrt{n}}$. If coefficients of the characteristic polynomial of A are computed from eigenvalues of A, then

$$|\tilde{c}_k - c_k| \leq \sqrt{n}(n-k+1)(n-k+1)s_{k-1}(|\lambda|)\rho + \mathcal{O}(\rho^2), \quad 1 \leq k \leq n.$$

Proof. The proof follows from Corollary 5.25 and Theorem 5.31. □

When matrix A has all positive eigenvalues, then we can express the error in the kth coefficient in terms of the preceding coefficient.

Corollary 5.33. Let A be a diagonalizable matrix A with $\lambda_i > 0$. If $\rho < \frac{1}{\sqrt{n}}$, then

$$|\tilde{c}_k - c_k| \leq \sqrt{n}(n-k+1)|c_{k-1}|\rho + \mathcal{O}(\rho^2), \quad 1 \leq k \leq n.$$

Proof. This follows from Corollary 5.26 and Theorem 5.31. □

Normal Matrices

For a normal matrix A, we get the following bound for ϵ_{abs}.

Theorem 5.34 (Corollary 2.4 in [40])**.** Let A be a normal matrix and $\tilde{A} = A + E$ a general complex matrix. Let U be a unitary matrix such that

$$U^*(A+E)U = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_l \end{pmatrix}, \quad 1 \le l \le n.$$

A_j are upper triangular matrices. There exists a permutation τ of $\{1, 2, \ldots, n\}$ such that

$$\epsilon_{abs} = \max_{1 \le i \le n} |\tilde{\lambda}_{\tau(i)} - \lambda_i| \le \sqrt{n(n-l+1)}\|E\|_2.$$

The above bound suggests that if the dimension of matrix A is not too large and $A + E$ is close to a normal matrix (i.e. if $l \approx n$ and if matrices A_j are close to being normal), then eigenvalues of A are well conditioned in absolute sense. In particular, if $l = n$, then $A + E$ is a normal matrix. In the worst case, we know nothing about $A + E$ and $l = 1$. Theorem 5.34 yields the following bound for the kth coefficient c_k of a normal matrix A.

Theorem 5.35. If $A + E$ is non normal and $\sqrt{n(n-l+1)}\|E\|_2 < 1$, then

$$|\tilde{c}_k - c_k| \le \sqrt{n(n-l+1)}(n-k+1)s_{k-1}(|\lambda|)\|E\|_2 + \mathcal{O}(\|E\|_2^2).$$

If $A + E$ is normal, then

$$|\tilde{c}_k - c_k| \le \sqrt{n}(n-k+1)s_{k-1}(|\lambda|)\|E\|_2 + \mathcal{O}(\|E\|_2^2).$$

Proof. The proof follows from Corollary 5.25 and Theorem 5.34. □

For a Hermitian matrix A, Weyl's Theorem gives the following improved bound for ϵ_{abs}.

Theorem 5.36 (page 551 in [41])**.** If A, E, and $A+E$ are Hermitian matrices, then

$$\max_i |\tilde{\lambda}_i - \lambda_i| = \epsilon_{abs} \leq \|E\|_2.$$

The above result provides the following bound for coefficients of the characteristic polynomial of a Hermitian matrix A.

Theorem 5.37. If A, E and $A+E$ are Hermitian matrices, then

$$|\tilde{c}_k - c_k| \leq (n-k+1)s_{k-1}(|\lambda|)\|E\|_2 + \mathcal{O}(\|E\|_2^2).$$

Proof. The proof follows from Corollary 5.25 and Theorem 5.36. □

5.3.2 Relative Perturbation Bounds

We derive perturbation bounds for the coefficients of the characteristic polynomial of A in terms of relative errors in eigenvalues. These bounds apply when the characteristic polynomial is determined from computed eigenvalues of A, and are derived from our results of section 5.2.2 for elementary symmetric functions. We denote the eigenvalues of A and $A+E$ by λ_i and $\hat{\lambda}_i$, respectively, where $\hat{\lambda}_i = \lambda_i(1+\epsilon_i)$. The kth coefficient of the characteristic polynomial of $A+E$ is denoted by $\hat{c}_k$. We define $\epsilon_{rel} = \max_{1\leq i\leq n} |\epsilon_i|$.

Corollary 5.38. If $\max_i |\hat{\lambda}_i - \lambda_i| = \epsilon_{rel} < 1$, then for non zero $c_k, 1 \leq k \leq n$,

$$\frac{|\hat{c}_k - c_k|}{|c_k|} \leq k\frac{s_k(|\lambda|)}{|c_k|}\epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2), \quad 1 \leq k \leq n.$$

Proof. The proof follows from Corollary 5.9 and Lemma 5.1. □

The above bound suggests that if $\epsilon_{rel} < 1$, then the first oder relative condition number for non zero kth coefficient c_k with respect to relative perturbations is $k\frac{s_k(|\lambda|)}{|c_k|}$.

Corollary 5.39. If $\lambda_i > 0, 1 \leq i \leq n$, and if $\epsilon_{rel} < 1$, then

$$\frac{|\hat{c}_k - c_k|}{|c_k|} \leq k\epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2), \quad 1 \leq k \leq n.$$

Proof. The proof follows from Corollary 5.11 and Lemma 5.1. □

We derive perturbation bounds for coefficients of characteristic polynomials of certain classes of matrices by substituting value of ϵ_{rel} in Corollary 5.38 and Corollary 5.39.

Non Singular Normal Matrices

We can obtain relative bounds for coefficients of the characteristic polynomial of a non singular matrix based on the following theorem by Wen Li and Weiwei Sun.

Theorem 5.40 (Corollary 3.3 in [40])**.** Let A be a non singular normal matrix. Let U be a unitary matrix such that

$$U^*(A+E)U = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_l \end{pmatrix}, \quad 1 \leq l \leq n,$$

where A_j are upper triangular matrices. Then, there exists a permutation τ of $\{1, 2, \ldots, n\}$ such that

$$\epsilon_{rel} = \max_{1 \leq i \leq n} \frac{|\hat{\lambda}_{\tau_i} - \lambda_i|}{|\lambda_i|} \leq \sqrt{n(n-l+1)}\|A^{-1}\|_2 \, \|E\|_2.$$

The above bound suggests that if $A + E$ is close to a normal matrix (i.e. if $l \approx n$ and if A_j are close to being normal, and if $\|A^{-1}\|_2 \, \|E\|_2 < 1$), then there is a

permutation τ of $\{1, 2, \ldots, n\}$ under which the eigenvalues of normal matrix A are well conditioned in a relative sense.

In particular, when $l = n$, $A + E$ is normal. In the worst case $l = 1$, so that

$$\max_{1 \le i \le n} \frac{|\hat{\lambda}_{\tau_i} - \lambda_i|}{|\lambda_i|} \le n \|A^{-1}\|_2 \, \|E\|_2.$$

We substitute the value of ϵ_{rel} from the above theorem in the Corollary 5.38, and obtain the forward error bound for the kth coefficient of the characteristic polynomial of normal matrix A.

Theorem 5.41. If $\sqrt{n(n-l+1)}\|A^{-1}\|_2\|E\|_2 < 1$, then under the assumptions of Theorem 5.40

$$\frac{|\hat{c}_k - c_k|}{|c_k|} \le k\sqrt{n(n-l+1)}\frac{s_k(|\lambda|)}{|c_k|}\|A^{-1}\|_2 \, \|E\|_2 + \mathcal{O}(\|A^{-1}\|_2\|E\|_2)^2, \quad 1 \le k \le n.$$

Proof. This follows from Corollary 5.38 and Theorem 5.40. □

For a unitary matrix A, the conditioning of the characteristic polynomial improves.

Corollary 5.42. Under the assumptions of Theorem 5.40 for unitary matrices A and $A + E$, if $\sqrt{n}\|E\|_2 < 1$, then

$$\frac{|\hat{c}_k - c_k|}{|c_k|} \le \sqrt{n}k\frac{s_k(|\lambda|)}{|c_k|} \, \|E\|_2 + \mathcal{O}(\|E\|_2)^2, \quad 1 \le k \le n.$$

Proof. In the bound of Theorem 5.41, substitute $l = n$ and $\|A^{-1}\|_2 = 1$. □

Real Symmetric Matrices

We present forward error bounds for coefficients of the characteristic polynomial of a real symmetric matrix A. These bounds are different from the previous bounds. We assume that the eigenvalues of A are computed from the Symmetric Rank Revealing Decomposition RRD of A. The Symmetric RRD of A with $rank(A) = r$

is a factorization $A = XDX^T$ where X is a matrix of order $n \times r$, D is a diagonal $r \times r$ non singular matrix and X has full column rank and is well conditioned [10]. The following perturbation bounds show that small relative perturbations in the elements of D and small normwise relative perturbations in X cause only small relative perturbations in the eigenvalues of A.

Theorem 5.43 (Theorem 2.1 in [10])**.** Let $A{=}XDX^T$ and $\hat{A} = \hat{X}\hat{D}\hat{X}^T$ be *RRDs* of the real symmetric $n \times n$ matrices A and $\hat{A}$ with

$$\frac{\|\hat{X} - X\|_2}{\|X\|_2} \leq \beta, \quad \frac{|\hat{D}_{ii} - D_{ii}|}{|D_{ii}|} \leq \beta \quad \text{for all} \quad i,$$

where $0 \leq \beta < 1$. Define $\kappa_2(X) = \|X\|_2\|X^{-1}\|_2$. If $\eta = \beta(2+\beta)\kappa_2(X) < 1$, then

$$|\hat{\lambda}_i - \lambda_i| \leq (2\eta + \eta^2)|\lambda_i|, \qquad 1 \leq i \leq n.$$

This bound implies that for well conditioned X, the eigenvalues of A are well conditioned with regard to relative perturbations in the Symmetric *RRD* of A.

We obtain the following result if the k^{th} coefficient c_k of the characteristic polynomial of A is determined from eigenvalues that have been computed from an *RRD* of A.

Theorem 5.44. Under the assumptions of Theorem 5.43, if $\eta < 1$, then

$$\frac{|\hat{c_k} - c_k|}{|c_k|} \leq 2k\,\frac{s_k(|\lambda|)}{|c_k|}\eta + \mathcal{O}(\eta^2), \quad 1 \leq k \leq n.$$

When the matrix A is symmetric positive definite, we have the following interesting bound.

Corollary 5.45. Under the assumptions of Theorem 5.43 with symmetric positive definite matrix A

$$\frac{|\hat{c_k} - c_k|}{|c_k|} \leq 2k\,\eta + \mathcal{O}(\eta^2), \quad 1 \leq k \leq n.$$

Proof. This follows from Corollary 5.39 and Theorem 5.43. □

The above bound implies that the coefficients of the characteristic polynomial of a symmetric positive definite matrix A are well conditioned, if we determine the coefficients from eigenvalues of A and the eigenvalues have been computed from the Symmetric RRD.

Accurate eigenvalues from the Symmetric $RRDs$ of many classes of symmetric matrices can be obtained. Some include scaled diagonally dominant matrices [3], diagonally scaled well conditioned positive definite matrices [9], Cauchy matrices, diagonally scaled Cauchy matrices, Vandermonde matrices, totally non negative matrices [10], total signed compound matrices and diagonally scaled totally unimodular matrices [44].

Non singular TN matrices

Matrices with all non negative minors are called totally nonnegative TN. All the eigenvalues of a non singular TN matrix are positive [37]. Like real symmetric matrices, we compute eigenvalues of a non singular TN matrix from its factorization. The following theorem by Gasca and Peña establishes necessary and sufficient conditions for a TN matrix.

Theorem 5.46 (Theorem 4.2 in [17])**.** A real $n \times n$ nonsingular matrix A is TN if and only if it can be uniquely factored as

$$A = L^{(1)} \cdots L^{(n-1)} \, D \, U^{(n-1)} \cdots U^{(1)},$$

where D is a diagonal matrix with positive diagonal elements. $L^{(j)}$ and $U^{(j)}$ are lower and upper unit bidiagonal matrices with non negative off–diagonal elements.

The following perturbation bounds show that small relative perturbations in the

elements of $L^{(j)}$, $U^{(j)}$ and D cause only small relative perturbations in the eigenvalues of A.

Theorem 5.47 (Corollary 7.3 in [37])**.** Let A and $\hat{A}$ be TN matrices. If

$$\hat{A} = \hat{L}^{(1)} \dots \hat{L}^{(n-1)} \, \hat{D} \, \hat{U}^{(n-1)} \cdots \hat{U}^{(1)},$$

where for $1 \leq j \leq n-1$, and $\delta \leq \frac{1}{2n^2}$,

$$|\hat{L}^{(j)}_{i+1,i} - L^{(j)}_{i+1,i}| \leq \delta |L^{(j)}_{i+1,i}|,$$

$$|\hat{U}^{(j)}_{i-1,i} - U^{(j)}_{i-1,i}| \leq \delta |U^{(j)}_{i-1,i}|,$$

$$|\hat{D}_{ii} - D_{ii}| \leq \delta |D_{ii}|.$$

then

$$|\hat{\lambda}_i - \lambda_i| \leq \frac{2n^2\delta}{1 - 2n^2\delta} \, \lambda_i, \quad 1 \leq i \leq n.$$

The above result implies that the first order condition number of each eigenvalue of A with respect to component wise relative perturbations in factors of A is $2n^2$. We get the following relative perturbation bound for the kth coefficient of the characteristic polynomial of a TN matrix A, if it has been determined from the eigenvalues of A, and the eigenvalues have been computed from the factors $L^{(j)}$, $U^{(j)}$ and D.

Theorem 5.48. For a nonsingular TN matrix, under the assumptions of Theorem 5.47,

$$\frac{|\hat{c}_k - c_k|}{|c_k|} \leq 2kn^2\delta + \mathcal{O}(\delta^2), \quad 1 \leq k \leq n.$$

Proof. The proof follows from Corollary 5.39 and Theorem 5.47. □

For moderate n, the above result shows that the coefficients of the characteristic polynomial of a TN matrix A are well conditioned with respect to relative perturbations in the eigenvalues of A provided the eigenvalues are determined from factors

of A. A numerically stable algorithm for computing eigenvalues of a TN matrix has been presented in [37].

5.3.3 The Summation Algorithm Applied to Computed Eigenvalues

We present the forward error bounds for the coefficients of $p(\lambda)$ of A from using the Summation Algorithm on computed eigenvalues. These bounds provide the overall error from both steps of computing the characteristic polynomial from eigenvalues.

Corollary 5.49. Suppose that the exact and computed eigenvalues of a matrix A are real numbers λ_i and $\tilde{\lambda}_i$,respectively, where $\tilde{\lambda}_i = \lambda_i + \epsilon_i$, $1 \leq i \leq n$. Suppose that $\max |\epsilon_i| = \epsilon_{abs} < 1$. Denote the computed coefficients of the characteristic polynomial of A from the Summation Algorithm by $\hat{c}_k$, $1 \leq k \leq n$. The forward error in $\hat{c}_k$ is given by

$$\begin{aligned} |\hat{c}_k - c_k| \quad \leq \quad & (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs} + \\ & \gamma_{2n}\Big(s_k(|\lambda|) + (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs}\Big) + \mathcal{O}(\epsilon_{abs}^2) \end{aligned}$$

Proof. From the triangle inequality we can write

$$|\hat{c}_k - c_k| \leq |\hat{c}_k - \tilde{c}_k| + |\tilde{c}_k - c_k|, \quad 1 \leq k \leq n. \tag{5.9}$$

Here $\tilde{c}_k$ are the exact coefficients of the characteristic polynomial of a matrix with eigenvalues $\tilde{\lambda}_i$. Corollary 5.25 implies

$$|\tilde{c}_k - c_k| \leq (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs} + \mathcal{O}(\epsilon_{abs}^2) \tag{5.10}$$

Lemma 5.1 and Theorem 5.16 imply

$$|\hat{c}_k - \tilde{c}_k| \leq \gamma_{2n} s_k(|\tilde{\lambda}|), \quad 1 \leq k \leq n. \tag{5.11}$$

The bound for $s_k(|\tilde{\lambda}|)$ follows from Theorem 5.2.

$$s_k(|\tilde{\lambda}|) \leq s_k(|\lambda|) + (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs} + \mathcal{O}(\epsilon_{abs}^2), \quad 1 \leq k \leq n. \tag{5.12}$$

Substituting the bounds of $|\tilde{c}_k - c_k|$, $|\hat{c}_k - \tilde{c}_k|$ and $|s_k(|\tilde{\lambda}|)$ from (5.10), (5.11) and (5.12) in the error bound of (5.9) yields the result. □

The above bound shows that two factors $s_{k-1}(|\lambda|)$ and $s_k(|\lambda|)$ may be responsible for inaccurate results of $\hat{c}_k$ when eigenvalues are known with absolute accuracy. When the eigenvalues have been computed with some estimate of relative accuracy and we use the Summation Algorithm to determine coefficients of $p(\lambda)$ from computed eigenvalues, we get the following error bounds.

Corollary 5.50. Suppose that the exact and computed eigenvalues of a matrix A are real numbers λ_i and $\hat{\lambda}_i$ respectively, where $\hat{\lambda}_i = \lambda_i(1+\epsilon_i)$, $1 \leq i \leq n$. Suppose that $\max|\epsilon_i| = \epsilon_{rel} < 1$. Denote the computed coefficients from the Summation Algorithm by $\bar{c}_k$, $1 \leq k \leq n$. The forward error in $\bar{c}_k$ is given by

$$|\bar{c}_k - c_k| \leq s_k(|\lambda|)\big(2k\epsilon_{rel} + \gamma_{2n}\big) + \mathcal{O}(\epsilon_{rel}^2), \quad 1 \leq k \leq n.$$

Proof. The proof is similar to that of Corollary 5.49. From the triangle inequality we write

$$|\bar{c}_k - c_k| \leq |\bar{c}_k - \hat{c}_k| + |\hat{c}_k - c_k|, \quad 1 \leq k \leq n. \tag{5.13}$$

Here $\hat{c}_k$ are the exact coefficients of a matrix with eigenvalues $\hat{\lambda}_i$, $1 \leq i \leq n$. Corollary 5.38 implies

$$|\hat{c}_k - c_k| \leq k s_k(|\lambda|)\epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2), \quad 1 \leq k \leq n. \tag{5.14}$$

Theorem 5.16 and Lemma 5.1 imply

$$|\bar{c}_k - \hat{c}_k| \leq \gamma_{2n} s_k(|\hat{\lambda}|) \tag{5.15}$$

From Corollary 5.9 we write the bound for $s_k(|\hat{\lambda}|)$

$$s_k(|\hat{\lambda}|) \leq s_k(|\lambda|) + k s_k(|\lambda|)\epsilon_{rel} + \mathcal{O}(\epsilon_{rel}^2) \tag{5.16}$$

We substitute the bounds of $s_k(|\hat{\lambda}|)$, $|\bar{c}_k - \hat{c}_k|$ and $|\hat{c}_k - c_k|$ from (5.16), (5.15) and (5.14) in (5.13). We also use the fact that $\gamma_{2n} < 1$ to simplify and obtain the desired bound. □

The bound of Corollary 5.50 shows that when the eigenvalues of a matrix A are known with some estimate of relative accuracy $\epsilon_{rel} < 1$ and the Summation Algorithm is applied to determine the coefficients of the characteristic polynomial then $s_k(|\lambda|)$, $1 \leq k \leq n$, is the first order condition number of error in computed coefficients.

We consider the case when some or all computed eigenvalues of A are complex, where eigenvalues of A may be real or complex numbers. Because more rounding errors are committed in determining the coefficients from computed complex eigenvalues, therefore, the constant γ_{2n} in the error bounds of Corollary 5.49 increases modestly to $\gamma_{2(n+k-1)}$.

Corollary 5.51. Suppose that the computed eigenvalues of A are complex numbers $\tilde{\lambda}_i$, where $\tilde{\lambda}_i = \lambda_i + \epsilon_i$, $1 \leq i \leq n$. Suppose that $\max|\epsilon_i| = \epsilon_{abs} < 1$. Denote the computed coefficients from the Summation Algorithm by $\hat{c}_k$, $1 \leq k \leq n$. Then the forward error in c_k is given by

$$\begin{aligned} |\hat{c}_k - c_k| \quad \leq \quad & (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs} + \\ & \gamma_{2(n+k-1)}\left[s_k(|\lambda|) + (n-k+1)s_{k-1}(|\lambda|)\epsilon_{abs}\right] + \mathcal{O}(\epsilon_{abs}^2) \end{aligned}$$

Proof. The proof is similar to that of Corollary 5.49 and follows by applying Theorem 5.22. □

5.4 Numerical Tests

We verify the accuracy of our derived bounds for computing the coefficients of the characteristic polynomial of a matrix from its eigenvalues on various test matrices. In our experiments we use MATLAB's "poly" function. This MATLAB's function yields the coefficients of the characteristic polynomial of a given matrix. The "poly" function first computes eigenvalues of the given matrix by using "eig" function, then determines the coefficients from the Summation Algorithm [1]. The machine precision u of MATLAB is approximately 1.1×10^{-16}. We compute the exact eigenvalues of test matrices by MATLAB's symbolic toolbox and compare them with eigenvalues computed from MATLAB's "eig" function. This allows us to approximate the absolute error (ϵ_{abs}) and the relative error (ϵ_{rel}) in eigenvalues of test matrices. The computed coefficients of characteristic polynomials of test matrices from "poly" function are denoted by $\hat{c}_k$.

Test 1: Forsythe Matrix

Conside the following Forsythe matrix [22, Example 5.22].

$$F = \begin{pmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & 1 \\ \eta & 0 & \dots & 0 \end{pmatrix}.$$

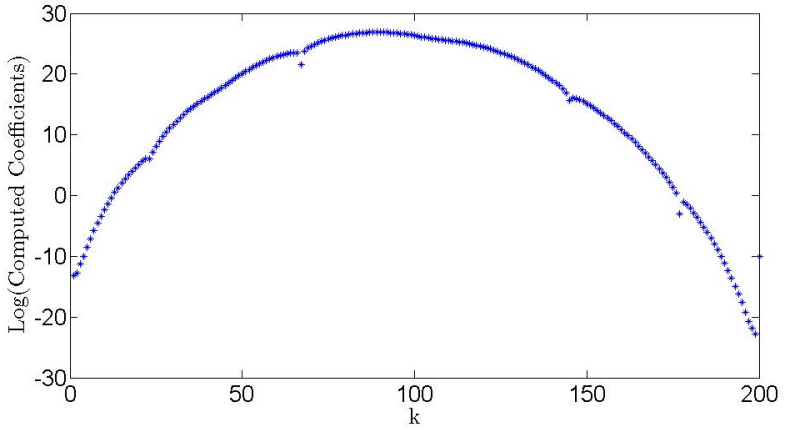

Figure 5.1: Results for $p(\lambda)$ of Forsythe matrix of order 200

The characteristic polynomial of F is $p(\lambda) = \lambda^n - \eta$, and the eigenvalues are given by

$$\lambda_k = \sqrt[n]{|\eta|}\, exp\frac{2k\pi i}{n}, \qquad 1 \leq k \leq n.$$

All the coefficients of the characteristic polynomial of F are zero beside c_n, and the eigenvalues of F are complex numbers. We compute $p(\lambda)$ of the Forsythe matrix of order 200 with $\eta = 10^{-10}$. We find that some of the computed coefficients are gigantic in magnitude. Figure 5.1 shows the Log of the computed coefficients. The first order condition number of the coefficient c_k is $(n-k+1)s_{k-1}(|\lambda|)$, as Corollary 5.25 shows. The first order condition numbers of coefficients c_{60} through c_{120} are in the range 10^{38} to 10^{53}. The perturbation bounds suggest that these coefficients might have large absolute errors. The magnitudes of the computed coefficients $\hat{c}_k$ confirm our analysis.

Test 2: Hansen's Matrix

Characteristic polynomials of symmetric positive definite matrices are well conditioned, if eigenvalues are computed with high relative accuracy, as Corollary 5.39

shows. The first order relative condition number of the coefficient c_k is $k\epsilon_{rel}$. We consider the following symmetric positive definite matrix A of order n used by Hansen to check the accuracy of Danilewiski's method [24, Section: Experimental Results]).

$$A = \begin{pmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$$

The coefficients of the characteristic polynomial of A are given by the following formula.

$$c_{n-m+1} = (-1)^{n-m+1} \binom{n+m-1}{n-m+1} \qquad 1 \leq m \leq n.$$

In particular, the trace of A is $2n-1$ and the determinant of A is 1. We compute $p(\lambda)$ of the matrix A of order 100. We determine $\epsilon_{rel} \approx 1.70 \times 10^{-13}$ and

$$\max_k \frac{|\hat{c}_k - c_k|}{|c_k|} \approx 10^{-13}, \quad 1 \leq k \leq n.$$

Every coefficient of $p(\lambda)$ of A is computed to at least 13 digits of accuracy. The error bounds correctly predict the results.

Test 3: A matrix generated from the inverse of Hansen's Matrix

The inverse of Hansen's matrix A presented in Test 3 can be given explicitly [24, Section: Experimental Results]. Let us denote the inverse of A by B. Then,

$$b_{ij} = \min(n-i+1, n-j+1), \quad 1 \leq i, j \leq n.$$

If we denote the kth coefficient of the characteristic polynomial of A by $c_k(A)$ and the kth coefficient of its inverse B by $c_k(B)$, then $c_k(A) = c_{n-k}(B)$, $1 \leq k \leq n-1$. In particular, $\det(B) = 1$. Hence, for the coefficients of the characteristic polynomial of B, we obtain the same numbers as for the coefficients of A but in the reverse order. The matrix B is symmetric positive definite. We observe that MATLAB generally computes eigenvalues of symmetric positive definite matrices with high relative accuracy. In order to make a more interesting example, we consider matrix R of order 100, where $R = PBP^{-1}$, and, P is a random matrix. R and B have the same characteristic polynomials. The first order relative condition number of the coefficient c_k of $p(\lambda)$ of R is $k\epsilon_{rel}$. We find that $\epsilon_{rel} \approx 10^{-9}$ and

$$\max_k \frac{|\hat{c}_k - c_k|}{|c_k|} \approx 10^{-9}, \quad 1 \leq k \leq n.$$

The perturbation bounds correctly predict the accuracy of the coefficients of $p(\lambda)$ of R.

Test 4: Tridiagonal Matrix

In this test we compute the coefficients of the characteristic polynomial of a tridiagonal matrix T of order 100 with the following entries:

$$T = \begin{pmatrix} 0 & -1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & -1 \\ & & 1 & 0 \end{pmatrix}.$$

We compute the exact coefficients of the characteristic polynomial of T by MATLAB's symbolic toolbox and find that every odd coefficient c_k of the characteristic polynomial of matrix A is zero. In particular $c_1 = 0$ and $c_n = 1$. The non zero coefficients of $p(\lambda)$

of T are computed accurately to at least 13 digits by MATLAB's "poly" function. However, most of the zero coefficients are inaccurate. In particular, the computed values of c_{37} to c_{77} are in the range of 10^0 to 10^3. The first oder absolute condition number of the coefficient c_k with respect to absolute perturbations in eigenvalues is $(n-k+1)s_{k-1}(|\lambda|)$. The first order condition numbers for c_{37} through c_{77} are in the range of 10^{26} to 10^{30}. Even though, the perturbation bounds are pessimistic, they still qualitatively provide accurate information about the coefficients.

Test 5: Frank Matrix

We consider the Frank matrix of order 20 from MATLAB's gallery of test matrices. The Frank matrix is an upper Hessenberg matrix with determinant 1. The coefficients of $p(\lambda)$ of the Frank matrix appear in pairs in the sense that $c_k = c_{n-k}$, $1 \le k \le n-1$. The eigenvalues of the Frank matrix are positive and occur in reciprocal pairs. We determine the exact coefficients by MATLAB's symbolic toolbox and compare them with the computed coefficients by using the "poly" function. We observe that the last four computed coefficients have very large absolute errors. In particular, the exact values of c_{19} and c_{20} are -210 and 1, respectively, and the computed coefficients are of order 10^6. Lemma 5.1 and Theorem 5.2 suggest the following error in the computed coefficient $\hat{c}_k$.

$$|\hat{c}_k - c_k| \le \sum_{i=1}^{k} \binom{n-k+i}{i} |c_{k-i}| \epsilon_{abs}^i, \qquad 1 \le k \le n.$$

The error in $\hat{c}_k$ may be large if the magnitudes of the preceding coefficients are large. We compute $\epsilon_{abs} \approx 0.34$. The approximated first order condition numbers for c_{19} and c_{20} are 10^4 and 100, respectively. The first order condition numbers are much smaller than the actual errors in the coefficients. This example shows that a small first order condition number does not guarantee a small error in $\hat{c}_k$. These huge errors in $\hat{c}_{19}$ and $\hat{c}_{20}$ are due to the fact that some of the preceding coefficients are large in magnitude.

For example, the magnitudes of coefficients from c_{14} to c_{18} are in the range of 10^4 to 10^9. The perturbation bounds correctly predict the errors in $\hat{c}_{19}$ and $\hat{c}_{20}$.

Chapter 6

La Budde's Method

6.1 Introduction

We present a little known method for computing the characteristic polynomial of a complex matrix A. The method was first introduced in 1956 by Wallace Givens at the third High Speed Computer Conference at Louisiana State University [19]. According to Givens, the method was brought to his attention by his coder Donald La Budde. Finding no earlier reference to this method, we credit its development to La Budde and, thus, name it " La Budde's method". The method begins with a preliminary reduction of A to an upper Hessenberg form H by orthogonal similarity transformations. The coefficients of the characteristic polynomial of H are determined by successively computing characteristic polynomials of leading principal submatrices of H. Because H and A are similar, they have the same characteristic polynomials. If A is symmetric, then H is a symmetric tridiagonal matrix, and La Budde's method simplifies to what we call the Sturm sequence method used by Givens to compute eigenvalues of a symmetric tridiagonal matrix T [18]. Regarding La Budde's method Givens said:

> "Since no division occurs in this second stage of the computation and the detailed examination of the first stage for the symmetric case was success-

> ful in guaranteeing its accuracy there, one may hope that the proposed method of getting the characteristic equation will often yield accurate results. It is, however, probable that cancellations of large numbers will sometimes occur in the floating point additions and will thus lead to excessive errors."

Wilkinson also preferred La Budde's method over the method of reducing a Hessenberg matrix to Frobenius form for obtaining the characteristic polynomial. He stated: [48, §6.57]

> " We have described the determination of the Frobenius form in terms of similarity transformations for the sake of consistency and in order to demonstrate it's relation to Danilewski's method. However, it is more straightforward to think in terms of a direct derivation of the characteristic polynomials of H. This polynomial may be obtained by recurrence relations in which we determine successively the characteristic polynomials of each of the leading principal submatrices. No special difficulties arise if some of the subdiagonal entries of H are small or even zero."

La Budde's method has attractive features for computing the characteristic polynomial of A. In the first stage, we reduce A to H by Householder's orthogonal similarity transformations which are unconditionally numerically stable [48, §6.6]. We derived perturbation bounds for coefficients of the characteristic polynomial of A in chapter 3. These bounds are in terms of elementary symmetric functions of singular values σ_i, $1 \leq i \leq n$, of A. Singular values of A are invariant under orthogonal similarity transformations, and, therefore, A and H have the same singular values. Hence, in reducing A to H, the condition numbers of coefficients of the characteristic polynomial of A remain unchanged. In the second stage of La Budde's method, we compute the characteristic polynomial of H by successive computations of characteristic polynomials of leading principal submatrices of H. We will show how to derive running error bounds for coefficients computed in the second stage. The errors from both stages of La Budde's method can be combined to obtain the forward errors in coefficients of the characteristic polynomial of A. We will also show that if a single

coefficient c_k, $1 \leq k \leq n$, of the characteristic polynomial of A is required, then La Budde's method can be modified to produce the desired coefficient. There are $\frac{5}{3}n^3$ floating point operations in reducing A to H by Householder's method [21, page 223] and $\frac{1}{6}n^3$ floating point operations in the second stage of La Budde's method [48, page 411]. This implies that the computation of the characteristic polynomial is efficient.

Overview

We present the Sturm sequence method for the computation of the characteristic polynomial $p(\lambda)$ of a real symmetric tridiagonal matrix T in Section 6.2. Furthermore, we derive running error bounds for coefficients of $p(\lambda)$. In Section 6.3, we describe La Budde's method for computing the characteristic polynomial of a Hessenberg matrix H and include its running error analysis. In Section 6.4, we present combined error bounds of La Budde's method for a real matrix A, i.e. the error in reducing A to H and the error in determining the characteristic polynomial of the computed Hessenberg matrix. Numerical tests are given is Section 6.5. In Section 6.6 we compare La Budde's method with MATLAB's "poly" function for the computation of $p(\lambda)$ of a given matrix from its eigenvalues.

6.2 The Sturm Sequence Method

Let a real $n \times n$ symmetric tridiagonal matrix T be defined as

$$T = \begin{pmatrix} a_1 & b_2 & & & \\ b_2 & a_2 & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & b_n \\ & & & b_n & a_n \end{pmatrix}.$$

Let us denote the characteristic polynomial of a principal submatrix T_i of order i by $p_i(\lambda)$, $1 \leq i \leq n$, where $p_i(\lambda) = \det(\lambda I - T_i)$. The recursion for computing coefficients of $p(\lambda)$ of T is given as follows [21, §8.5]:

Algorithm 2 Sturm sequence method for $p(\lambda)$ of a tridiagonal matrix T

Input: $n \times n$ real symmetric tridiagonal matrix T
Output: The characteristic polynomial $p(\lambda)$ of T
  Set $p_0(\lambda) = 1$
  Set $p_1(\lambda) = \lambda - a_1$
  for $i = 2 : n$ **do**
    $p_i(\lambda) = (\lambda - a_i)p_{i-1}(\lambda) - b_i^2 p_{i-2}(\lambda)$
  end for

The correctness of the Sturm sequence method follows from a simple determinantal expansion [21, §8.5]. In the process of computing the characteristic polynomial $p(\lambda)$ of T, the Sturm sequence method computes characteristic polynomials of all leading principal submatrices of T. Let us denote the coefficients of $p_i(\lambda)$ by $c_k^{(i)}$, $1 \leq k \leq i$. Then,

$$p(\lambda) = p_n(\lambda), \qquad \text{and} \quad c_k = c_k^{(n)}, \quad 1 \leq k \leq n.$$

Writing $p_i(\lambda)$, $p_{i-1}(\lambda)$ and $p_{i-2}(\lambda)$ in terms of their coefficients and equating like powers of λ on both sides of the recursion of the Sturm sequence method, we observe that the computation of $c_k^{(i)}$ requires $c_k^{(i-1)}$, $c_{k-1}^{(i-1)}$ and $c_{k-2}^{(i-2)}$, $2 \leq i \leq n$. This allows us to write the modified Sturm sequence method which computes a coefficient c_k of the characteristic polynomial of T. In addition $c_1, \ldots, c_{k-1}$ are also computed.

The inner loop of the modified Sturm sequence method computes $c_k^{(i-1)}$, $c_{k-1}^{(i-1)}$ and $c_{k-2}^{(i-2)}$, which are required in the computation of $c_k^{(i)}$. The outer loop computes $c_1^{(i)}, \ldots, c_k^{(i)}$ for $2 \leq i \leq n$. At the end of the recursion, we obtain $c_1, \ldots, c_k$.

Algorithm 3 Modified Sturm sequence method for a single coefficient c_k

Input: A real $n \times n$ symmetric tridiagonal matrix T, index k

Output: A coefficient c_k of $p(\lambda)$ of T

Set $c_0^{(l)} = 1$ for $0 \leq l \leq n-1$
Set $c_s^{(l)} = 0$ for $s > l$ or $s < 0$
Set $c_1^{(1)} = -a_1$
for $i = 2 : n$ **do**
    for $j = 1 : k$ **do**
        $c_j^{(i)} = c_j^{(i-1)} - a_i c_{j-1}^{(i-1)} - b_i^2 c_{j-2}^{(i-2)}$
    end for
end for
{At step n: $c_k^{(n)} = c_k$}

6.2.1 Running Error Bounds

To derive running error bounds of the Sturm sequence method we use the standard and modified floating point models described in chapter 5 (section 5.2.4). We denote the computed coefficients of the characteristic polynomial $p_i(\lambda)$ of order i by $\hat{c}_k^{(i)}$. We compute $\hat{c}_k^{(i)}$ as

$$\hat{c}_k^{(i)} = \mathrm{fl}\left[\mathrm{fl}\left[\hat{c}_k^{(i-1)} - \mathrm{fl}\left[a_i \hat{c}_{k-1}^{(i-1)}\right]\right] - \mathrm{fl}\left[b_i^2 \hat{c}_{k-2}^{(i-2)}\right]\right].$$

Also,

$$\hat{c}_k^{(i)} = c_k^{(i)} + e_k^{(i)}, \quad 2 \leq i \leq n, \quad 1 \leq k \leq n.$$

Here $c_k^{(i)}$ is the exact coefficient and $e_k^{(i)}$ is the error in $\hat{c}_k^{(i)}$. In the beginning of the recursion $\hat{c}_1^{(1)} = c_1^{(1)} = -a_1$ and $e_1^{(1)} = 0$. We state the running error bound for $\hat{c}_1$ and then present the error bound for $\hat{c}_2$.

Theorem 6.1. The error in $\hat{c}_1$ is bounded as follows:

$$|e_1^{(i)}| \leq |e_1^{(i-1)}| + u|\hat{c}_1^{(i)}|, \quad 2 \leq i \leq n.$$

Proof. The Sturm sequence method computes c_1 as a recursive sum of diagonal entries of T. At step i, we write

$$\hat{c}_1^{(i)} = \mathrm{fl}\Big[\hat{c}_1^{(i-1)} - a_i\Big].$$

Using the modified standard model (5.2), we get

$$(1+\epsilon^{(i)})\hat{c}_1^{(i)} = \hat{c}_1^{(i-1)} - a_i, \quad \text{where} \quad |\epsilon^{(i)}| \leq u.$$

Writing the computed coefficients $\hat{c}_1^{(i)}$ and $\hat{c}_1^{(i-1)}$ in terms of their errors and simplifying the expression, we get

$$e_1^{(i)} = e_1^{(i-1)} - \epsilon^{(i)}\hat{c}_1^{(i)}.$$

This implies

$$|e_1^{(i)}| \leq |e_1^{(i-1)}| + u|\hat{c}_1^{(i)}|, \quad 2 \leq i \leq n.$$

□

Theorem 6.2. The error in $\hat{c}_2^{(2)}$ is given by the following inequality.

$$|e_2^{(2)}| \leq u\big(|a_2 a_1| + |b_2^2| + |\hat{c}_2^{(2)}|\big).$$

For $2 < i \leq n$,

$$|e_2^{(i)}| \leq |e_2^{(i-1)}| + |a_i e_1^{(i-1)}| + u\big(|\hat{c}_2^{(i-1)}| + |b_i^2| + |\hat{c}_2^{(i)}|\big) + \gamma_2|a_i\hat{c}_1^{(i-1)}|.$$

Proof. $\hat{c}_2^{(2)}$ is computed as

$$\hat{c}_2^{(2)} = \mathrm{fl}\Big[\mathrm{fl}\Big[a_1 a_2\Big] - \mathrm{fl}\Big[b_2^2\Big]\Big].$$

Using the standard floating point model (5.1), we get

$$\hat{c}_2^{(2)} = \mathrm{fl}\Big[a_2a_1(1+\delta) - b_2^2(1+\eta)\Big], \quad |\delta|, |\eta| \leq u.$$

Now, we use the standard modified model (5.2) to obtain

$$(1+\epsilon)\hat{c}_2^{(2)} = a_2a_1(1+\delta) - b_2^2(1+\eta), \quad |\epsilon| \leq u.$$

Expressing $\hat{c}_2^{(2)}$ in terms of its error and simplifying produces the result for the error in $\hat{c}_2^{(2)}$.

When $2 < i \leq n$,

$$\begin{aligned}
\hat{c}_2^{(i)} &= \mathrm{fl}\Big[\mathrm{fl}\Big[\hat{c}_2^{(i-1)} - \mathrm{fl}\Big[a_i\hat{c}_1^{(i-1)}\Big]\Big] - \mathrm{fl}\Big[b_i^2\Big]\Big] \\
&= \mathrm{fl}\Big[\hat{c}_2^{(i-1)}(1+\delta^{(i)}) - a_i\hat{c}_1^{(i-1)}(1+\theta_2^{(i)}) - b_i^2(1+\eta^{(i)})\Big],
\end{aligned}$$

where $|\delta^{(i)}|$, $|\eta^{(i)}| \leq u$ and $|\theta_2^{(i)}| \leq \gamma_2$. We use the modified model (5.2) to write

$$(1+\epsilon^{(i)})\hat{c}_2^{(i)} = \hat{c}_2^{(i-1)}(1+\delta^{(i)}) - a_i\hat{c}_1^{(i-1)}(1+\theta_2^{(i)}) - b_i^2(1+\eta^{(i)}),$$

where $|\epsilon^{(i)}| \leq u$. We express $\hat{c}_2^{(i-1)}$ and $\hat{c}_1^{(i-1)}$ in terms of their errors and simplify to get

$$e_2^{(i)} = e_2^{(i-1)} + \delta^{(i)}\hat{c}_2^{(i-1)} - a_ie_1^{(i-1)} - \theta_2^{(i)}a_i\hat{c}_1^{(i-1)} - b_i^2\eta^{(i)} - \epsilon^{(i)}\hat{c}_2^{(i)}.$$

Applying the triangle inequality gives the forward error in $\hat{c}_2^{(i)}$. □

For $3 \leq k \leq n$, the running forward error bounds are given below.

Theorem 6.3. For $3 \leq k \leq n-1$, the error in $\hat{c}_k^{(k)}$ is bounded by the following:

$$|e_k^{(k)}| \leq |a_ke_{k-1}^{(k-1)}| + |b_k^2e_{k-2}^{(k-2)}| + \gamma_2|b_k^2\hat{c}_{k-2}^{(k-2)}| + u\big(|a_k\hat{c}_{k-1}^{(k-1)}| + |\hat{c}_k^{(k)}|\big).$$

For $k < i \leq n$, we have

$$\begin{aligned} |e_k^{(i)}| \;\leq\; & |e_k^{(i-1)}| + |a_i e_{k-1}^{(i-1)}| + |b_i^2 e_{k-2}^{(i-2)}| + \\ & u\big(|\hat{c}_k^{(i-1)}| + |\hat{c}_k^{(i)}|\big) + \gamma_2\big(|a_i \hat{c}_{k-1}^{(i-1)}| + |b_i^2 \hat{c}_{k-2}^{(i-2)}|\big). \end{aligned}$$

Proof. Note that for $k > i$, $c_k^{(i)} = 0$. Therefore, we start accumulating errors at step k. Also, $c_k^{(k-1)} = 0$. We compute $\hat{c}_k^{(k)}$ as

$$\hat{c}_k^{(k)} = -\mathrm{fl}\Big[\mathrm{fl}\Big[a_k \hat{c}_{k-1}^{(k-1)}\Big] + \mathrm{fl}\Big[b_k^2 \hat{c}_{k-2}^{(k-2)}\Big]\Big].$$

Applying the standard floating model (5.1), we get

$$\hat{c}_k^{(k)} = -\mathrm{fl}\Big[a_k \hat{c}_{k-1}^{(k-1)}(1+\delta) + b_k^2 \hat{c}_{k-2}^{(k-2)}(1+\theta_2)\Big],$$

where $|\delta| \leq u$ and $|\theta_2| \leq \gamma_2$. From the modified model (5.2), we obtain

$$(1+\epsilon)\hat{c}_k^{(k)} = -a_k \hat{c}_{k-1}^{(k-1)}(1+\delta) - b_k^2 \hat{c}_{k-2}^{(k-2)}(1+\theta_2), \quad |\epsilon| \leq u.$$

Now, the proof is similar to that of previous proofs. We write $\hat{c}_k^{(k)}$, $\hat{c}_{k-1}^{(k-1)}$ and $\hat{c}_{k-2}^{(k-2)}$ in terms of their errors and simplify to get

$$e_k^{(k)} = -a_k e_{k-1}^{(k-1)} - b_k^2 e_{k-2}^{(k-2)} - \theta_2 b_k^2 \hat{c}_{k-2}^{(k-2)} - \delta a_k \hat{c}_{k-1}^{(k-1)} - \epsilon \hat{c}_k^{(k)}.$$

The forward error bounds follow from the triangle inequality. When $k < i \leq n$, we write

$$\hat{c}_k^{(i)} = \mathrm{fl}\Big[\hat{c}_k^{(i-1)} - \mathrm{fl}\Big[a_i \hat{c}_{k-1}^{(i-1)} + b_i^2 \hat{c}_{k-2}^{(i-2)}\Big]\Big].$$

This implies

$$\hat{c}_k^{(i)} = \mathrm{fl}\Big[\hat{c}_k^{(i-1)}(1+\delta^{(i)}) - a_i \hat{c}_{k-1}^{(i-1)}(1+\theta_2^{(i)}) - b_i^2 \hat{c}_{k-2}^{(i-2)}(1+\hat{\theta}_2^{(i)})\Big],$$

where $|\delta^{(i)}| \leq u$ and $|\theta_2^{(i)}|,|\hat{\theta}_2^{(i)}| \leq \gamma_2$. We apply the modified standard model (5.2) and obtain

$$(1+\epsilon^{(i)})\hat{c}_k^{(i)} = \hat{c}_k^{(i-1)}(1+\delta^{(i)}) - a_i\hat{c}_{k-1}^{(i-1)}(1+\theta_2^{(i)}) - b_i^2\hat{c}_{k-2}^{(i-2)}(1+\hat{\theta}_2^{(i)}), \quad |\epsilon^{(i)}| \leq u.$$

Writing the computed coefficients in terms of their errors produces the following result.

$$\begin{aligned} e_k^{(i)} &= e_k^{(i-1)} + \delta^{(i)}\hat{c}_k^{(i-1)} - \theta_2^{(i)}a_i\hat{c}_{k-1}^{(i-1)} - a_ie_{k-1}^{(i-1)} - \\ &\quad b_i^2e_{k-2}^{(i-2)} - \hat{\theta}_2^{(i)}b_i^2\hat{c}_{k-2}^{(i-2)} - \epsilon^{(i)}\hat{c}_k^{(i)}. \end{aligned}$$

We apply the triangle inequality to bound the forward error in $\hat{c}_k^{(i)}$, $3 \leq k \leq n$. □

Remark 6.4. A straight forward analysis shows that if T is a diagonal matrix, then the recursion of the Sturm sequence simplifies to the Summation Algorithm. We presented the Summation Algorithm along with its error analysis in chapter 5.

6.3 La Budde's Method

In La Budde's method, the coefficients of the characteristic polynomial of an upper Hessenberg matrix H are computed by successively determining the characteristic polynomials of each of the leading principal submatrices H_i, $1 \leq i \leq n$, of H[19]. Let us denote characteristic polynomials of leading principal submatrices H_i of order i by $p_i(\lambda)$, where $p_i(\lambda) = \det(\lambda I - H_i)$, $1 \leq i \leq n$. The characteristic polynomial of H is computed from Algorithm 4.

Algorithm 4 with slight modifications can be used to compute the characteristic polynomial of a lower Hessenberg matrix H as well. The recurrence relations of La Budde's method for computing the characteristic polynomial of H can also be unraveled like those for a tridiagonal matrix T. By writing $p_i(\lambda), \ldots, p_1(\lambda)$ in terms

Algorithm 4 La Budde's method for $p(\lambda)$ of an upper Hessenberg matrix H

Input: An $n \times n$ upper Hessenberg matrix H
Output: The characteristic polynomial $p(\lambda)$ of H
Set $p_0(\lambda) = 1$
Set $p_1(\lambda) = \lambda - h_{11}$
for $i = 2 : n$ **do**
$p_i(\lambda) = (\lambda - h_{ii})p_{i-1}(\lambda) - \sum_{m=1}^{i-1} h_{i-m,i}\, h_{i,i-1} \cdots h_{i-m+1,i-m}\, p_{i-m-1}(\lambda)$
end for

of their coefficients $c_j^{(i)}$, $1 \leq j \leq i$, $2 \leq i \leq n$, and equating like powers of λ on both sides of the recursion of La Budde's method, we find out how each coefficient is computed. In order to compute a coefficient $c_k^{(i)}$, La Budde's method requires the computation of $c_k^{(i-1)}, c_{k-1}^{(i-1)}, \ldots, c_1^{(i-k+1)}$, $2 \leq i \leq n$. This provides us with the following algorithm that can be used to compute a coefficient c_k. In addition, we also get $c_1, \ldots, c_{k-1}$.

Algorithm 5 La Budde's method for a single coefficient c_k of $p(\lambda)$ of H

Input: An $n \times n$ upper Hessenberg matrix H, index k
Output: A coefficient c_k of $p(\lambda)$ of H
Set $c_0^{(l)} = 1$ for $0 \leq l \leq n-1$
Set $c_s^{(l)} = 0$ for $s > l$ or $s < 0$
Set $c_1^{(1)} = -h_{11}$
for $i = 2 : n$ **do**
for $j = 1 : k$ **do**
$c_j^{(i)} = c_j^{(i-1)} - h_{ii} c_{j-1}^{(i-1)} - \sum_{m=1}^{j-1} h_{i-m,i}\, h_{i,i-1} \cdots h_{i-m+1,i-m}\, c_{j-m-1}^{(i-m-1)}$
end for
end for
{At step n: $c_k^{(n)} = c_k$}

For computation of $c_k^{(i)}$, the inner loop of Algorithm 5 determines $c_k^{(i-1)}$, $c_k^{(i-1)}$, $c_{k-1}^{(i-1)}, \ldots, c_1^{(i-k+1)}$ and the outer loop computes $c_1^{(i)}, \ldots, c_k^{(i)}$. Therefore, at step $i = n$, we obtain $c_1, \ldots, c_k$.

6.3.1 Running Error Bounds for Real Matrices

We present running error bounds for the coefficients of $p(\lambda)$ of a real Hessenberg matrix H. To make the error analysis easier, we explain how $c_1, \ldots, c_4$ are computed. The coefficient c_1 is computed as a recursive sum of diagonal entries of H.

$$c_1^{(i)} = c_1^{(i-1)} - h_{ii}, \quad 2 \leq i \leq n.$$

The recursion for c_2 simplifies to

$$c_2^{(i)} = c_2^{(i-1)} - h_{ii}c_1^{(i-1)} - h_{i-1,i}h_{i,i-1}, \quad 2 \leq i \leq n.$$

Similarly, c_3 is computed as

$$c_3^{(i)} = c_3^{(i-1)} - h_{ii}c_2^{(i-1)} - h_{i-1,i}h_{i,i-1}c_1^{(i-2)} - h_{i-2,i}h_{i,i-1}h_{i-1,i-2}, \quad 3 \leq i \leq n.$$

Finally, the recursion formula for c_4 is given as follows:

$$\begin{aligned} c_4^{(i)} &= c_4^{(i-1)} - h_{ii}c_3^{(i-1)} - h_{i-1,i}h_{i,i-1}c_2^{(i-2)} - h_{i-2,i}h_{i,i-1}h_{i-1,i-2}c_1^{(i-3)} \\ &\quad -h_{i-3,i}h_{i,i-1}h_{i-1,i-2}h_{i-2,i-3}, \quad 4 \leq i \leq n. \end{aligned}$$

The running error bounds for $\hat{c}_1$ and $\hat{c}_2$ are analogous to those of Theorems 6.1 and 6.2. We state them below.

Theorem 6.5. The following recursion produces the error bound for $\hat{c}_1$ of $p(\lambda)$ of an $n \times n$ upper Hessenberg matrix H.

$$|e_1^{(i)}| \leq |e_1^{(i-1)}| + u|\hat{c}_1^{(i)}|, \quad 2 \leq i \leq n.$$

Proof. The proof is similar to that of Theorem 6.1. □

Here is the error bound for $\hat{c}_2$.

Theorem 6.6. For $i = 2$, the error in $\hat{c}_2^{(2)}$ is computed as follows:

$$|e_2^{(2)}| \leq u\left(|h_{11}h_{22}| + |h_{12}h_{21}| + |\hat{c}_2^{(2)}|\right),$$

and for $2 < i \leq n$,

$$|e_2^{(i)}| \leq |e_2^{(i-1)}| + |h_{ii}e_1^{(i-1)}| + u\left(|\hat{c}_2^{(i-1)}| + |h_{i-1,i}h_{i,i-1}| + |\hat{c}_2^{(i)}|\right) + \gamma_2|h_{ii}\hat{c}_1^{(i-1)}|.$$

Proof. The proof is similar to that of Theorem 6.2. □

Now, we give the proof for the error bound in any $\hat{c}_k$, $3 \leq k \leq n$.

Theorem 6.7. The error in $\hat{c}_k^{(k)}$, $3 \leq k \leq n$ is bounded by the following:

$$\begin{aligned}|e_k^{(k)}| \;\leq\; & |h_{kk}e_{k-1}^{(k-1)}| + \sum_{m=1}^{k-2}|h_{k-m,k}h_{k,k-1}h_{k-m+1,k-m}e_{k-m-1}^{(k-m-1)}| + \\ & u\big(|\hat{c}_k^{(k)}| + |h_{kk}\hat{c}_{k-1}^{(k-1)}|\big) + \gamma_k\big(|h_{k-1,k}h_{k,k-1}\hat{c}_{k-2}^{(k-2)}| + |h_{1k}h_{k,k-1}\cdots h_{21}|\big) + \\ & \gamma_{k+1}\big(\sum_{m=2}^{k-2}|h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}|\big).\end{aligned}$$

The error bound in $\hat{c}_k^{(i)}$, $3 < i \leq n$, is given by

$$\begin{aligned}|e_k^{(i)}| \;\leq\; & |e_k^{(i-1)}| + |h_{ii}e_{k-1}^{(i-1)}| + \sum_{m=1}^{k-2}|h_{i-m,i}h_{i,i-1}\cdots h_{i-m+1,i-m}e_{k-m-1}^{(i-m-1)}| + \\ & \gamma_k\big(|h_{i-1,i}h_{i,i-1}\hat{c}_{k-2}^{(i-2)}| + |h_{i-k+1,i}h_{i,i-1}\cdots h_{i-k+2,i-k+1}|\big) + u\big(|\hat{c}_k^{(i)}| + |\hat{c}_k^{(i-1)}|\big) \\ & +\gamma_2|h_{ii}\hat{c}_{k-1}^{(i-1)}| + \gamma_{k+1}\big(\sum_{m=2}^{k-2}|h_{i-m,i}h_{i,i-1}\cdots h_{i-m+1,i-m}\hat{c}_{k-m-1}^{(i-m-1)}|\big).\end{aligned}$$

Proof. We compute $\hat{c}_k^{(k)}$ as

$$\hat{c}_k^{(k)} = -\mathrm{fl}\left[\mathrm{fl}\left[h_{kk}\hat{c}_{k-1}^{(k-1)}\right] + \mathrm{fl}\left[\sum_{m=1}^{k-1}h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}\right]\right].$$

There are $k-1$ terms in the sum and the initial $k-2$ terms consist of the product of $m+2$ floating point numbers. The last term in the sum is a product of k numbers. Using the standard floating point model (5.1),

$$\begin{aligned} \mathrm{fl}\left[h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}\right] &= h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)} \\ &\quad (1+\theta_{m+1}), \end{aligned}$$

where $|\theta_{m+1}| \leq \gamma_{m+1}$. The last term is a product of k floating point numbers; therefore,

$$\mathrm{fl}\left[h_{1k}h_{k,k-1}\cdots h_{21}\right] = h_{1k}h_{k,k-1}\cdots h_{21}(1+\theta_{k-1}),$$

where $|\theta_{k-1}| \leq \gamma_{k-1}$. Adding terms from left to right in the sum and using the standard floating point model (5.1), we obtain

$$\begin{aligned} \mathrm{fl}\left[\sum_{m=1}^{k-1} h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}\right] &= h_{k-1,k}h_{k,k-1}\hat{c}_{k-2}^{(k-2)}(1+\theta_k) + \\ &\quad h_{1k}h_{k,k-1}\cdots h_{21}(1+\hat{\theta}_k) + \\ &\quad \sum_{m=2}^{k-2} h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m} \\ &\quad \hat{c}_{k-m-1}^{(k-m-1)}(1+\theta_{k+1}^{(m)}), \end{aligned}$$

where $|\theta_{k+1}^{(m)}| \leq \gamma_{k+1}$ and $|\theta_k|,|\hat{\theta}_k| \leq \gamma_k$. Also,

$$\mathrm{fl}\left[h_{kk}\hat{c}_{k-1}^{(k-1)}\right] = h_{kk}\hat{c}_{k-1}^{(k-1)}(1+\delta),$$

where $|\delta| \leq u$. Using the modified standard model (5.2) to add $\mathrm{fl}\left[h_{kk}\hat{c}_{k-1}^{(k-1)}\right]$ and

$\mathrm{fl}\left[\sum_{m=1}^{k-1} h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}\right]$, we get

$$\begin{aligned}(1+\epsilon)\hat{c}_k^{(k)} &= -\big(h_{kk}\hat{c}_{k-1}^{(k-1)}(1+\delta) + h_{k-1,k}h_{k,k-1}\hat{c}_{k-2}^{(k-2)}(1+\theta_k) + \\ &\quad h_{1k}h_{k,k-1}\cdots h_{21}(1+\hat{\theta}_k) + \\ &\quad \sum_{m=2}^{k-2} h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}(1+\theta_{k+1}^{(m)})\big),\end{aligned}$$

where $|\epsilon| \leq u$. Writing the computed coefficients in terms of their errors and simplifying produces

$$\begin{aligned}e_k^{(k)} &= -\Bigg(h_{kk}e_{k-1}^{(k-1)} + \sum_{m=1}^{k-2} h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}e_{k-m-1}^{(k-m-1)} + \\ &\quad h_{kk}\hat{c}_{k-1}^{(k-1)}\delta + h_{k-1,k}h_{k,k-1}\hat{c}_{k-2}^{(k-2)}\theta_k + h_{1k}h_{k,k-1}\cdots h_{21}\hat{\theta}_k + \epsilon\hat{c}_k^{(k)} + \\ &\quad \sum_{m=2}^{k-2} h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}\theta_{k+1}^{(m)}\Bigg).\end{aligned}$$

Applying the triangle inequality yields the forward error bound for $\hat{c}_k^{(k)}$.

When $k < i \leq n$, then we have another term $\hat{c}_k^{(i-1)}$ in the computation of $\hat{c}_k^{(i)}$. We compute $\hat{c}_k^{(i)}$ as follows:

$$\hat{c}_k^{(i)} = \mathrm{fl}\left[\mathrm{fl}\left[\hat{c}_k^{(i-1)} - \mathrm{fl}\left[h_{ii}\hat{c}_{k-1}^{(i-1)}\right]\right] - \mathrm{fl}\left[\sum_{m=1}^{k-1} h_{i-m,i}h_{i,i-1}\cdots h_{i-m+1,i-m}\hat{c}_{k-m-1}^{(i-m-1)}\right]\right].$$

Due to inclusion of this term, in standard floating point arithmetic,

$$\mathrm{fl}\left[\hat{c}_k^{(i-1)} - \mathrm{fl}\left[h_{ii}\hat{c}_{k-1}^{(i-1)}\right]\right] = \hat{c}_k^{(i-1)}(1+\delta^{(i)}) - h_{ii}\hat{c}_{k-1}^{(i-1)}(1+\theta_2^{(i)}),$$

where $|\delta^{(i)}| \leq u$ and $|\theta_2^{(i)}| \leq \gamma_2$. The rest of the proof is similar to the case $i = k$. □

Remark 6.8 (Potential instability of La Budde's method)**.** The running error bounds

reflect the potential instability of La Budde's method. The coefficient $c_k^{(i)}$ of characteristic polynomial of H_i at step i, $2 \leq i \leq n$ is computed from the preceding coefficients $c_k^{(i-1)}, c_{k-1}^{(i-1)}, \ldots, c_1^{(i-k+1)}$. La Budde's method can produce inaccurate results for $c_k^{(i)}$, if the magnitudes of preceding coefficients are very large in comparison to $c_k^{(i)}$ so that catastrophic cancellation may occur in the computation of $c_k^{(i)}$. This means that the error in the computed coefficient $\hat{c}_k$ of $p(\lambda)$ of H might be large when the preceding coefficients of characteristic polynomials of leading principal submatrices of H are larger than c_k.

6.3.2 Running Error Bounds for Complex Matrices

We derive running error bounds of La Budde's method for coefficients c_k of $p(\lambda)$ of a complex matrix by applying standard and modified models for complex numbers. The proofs of running error bounds for coefficients of $p(\lambda)$ of a complex matrix are similar to those of a real matrix. The multiplication of j real numbers results in the error multiplier γ_{j-1}, and the product of j complex numbers gives $\gamma_{3(j-1)}$; therefore, the error multipliers in the running error bounds of the coefficients of $p(\lambda)$ of a complex matrix increase more than those of a real matrix. The error bound for $\hat{c}_1$ of $p(\lambda)$ of A is similar to the error bound for a real matrix and follows from applying the modified model (5.6) for addition of complex numbers. Below, we present the error bound for $\hat{c}_2$.

Theorem 6.9. The error in $\hat{c}_2^{(2)}$ is bounded by the following:

$$|e_2^{(2)}| \leq \gamma_3\big(|h_{11}h_{22}| + |h_{12}h_{21}|\big) + u|\hat{c}_2^{(2)}|.$$

The error in $\hat{c}_2^{(i)}$, $3 \leq i \leq n$, is given by the following recursion:

$$\begin{aligned} |e_2^{(i)}| \;\leq\; & |e_2^{(i-1)}| + |h_{ii}e_1^{(i-1)}| + \gamma_4|h_{ii}\hat{c}_1^{(i-1)}| + \\ & u\left(|\hat{c}_2^{(i-1)}| + \hat{c}_2^{(i)}|\right) + \gamma_3|h_{i-1,i}h_{i,i-1}|. \end{aligned}$$

Proof. $\hat{c}_2^{(2)}$ is computed as

$$\hat{c}_2^{(2)} = \mathrm{fl}\Big[\mathrm{fl}\Big[h_{11}h_{22}\Big] - \mathrm{fl}\Big[h_{12}h_{21}\Big]\Big].$$

Applying the standard model (5.5) of multiplication of complex numbers,

$$\mathrm{fl}\Big[h_{11}h_{22}\Big] = h_{11}h_{22}(1+\alpha_1), \quad \text{where} \quad |\alpha_1| \leq \gamma_3.$$

Similarly,

$$\mathrm{fl}\Big[h_{12}h_{21}\Big] = h_{12}h_{21}(1+\hat{\alpha}_1), \quad \text{where} \quad |\hat{\alpha}_1| \leq \gamma_3.$$

We add $\mathrm{fl}\Big[h_{11}h_{22}\Big]$ and $\mathrm{fl}\Big[h_{12}h_{21}\Big]$ by using the modified model (5.6). We Write $\hat{c}_2^{(2)}$ in terms of its error, simplify, and apply the triangle inequality to get the bound.

For $3 \leq i \leq n$, we compute $\hat{c}_2^{(i)}$ as

$$\hat{c}_2^{(i)} = \mathrm{fl}\Big[\mathrm{fl}\Big[\hat{c}_2^{(i-1)} - \mathrm{fl}\Big[h_{ii}\hat{c}_1^{(i-1)}\Big]\Big] - \mathrm{fl}\Big[h_{i-1,i}h_{i,i-1}\Big]\Big].$$

After applying standard models (5.4) and (5.5) of addition and multiplication of complex numbers, we get

$$\hat{c}_2^{(i)} = \mathrm{fl}\Big[\hat{c}_2^{(i-1)}(1+\theta_1^{(i)}) - h_{ii}\hat{c}_1^{(i-1)}(1+\theta_1^{(i)})(1+\alpha_1^{(i)}) - h_{i-1,i}h_{i,i-1}(1+\hat{\alpha}_1^{(i)})\Big],$$

where $|\theta_1^{(i)}| \leq \gamma_1$ and $|\alpha_1^{(i)}|, |\hat{\alpha}_1^{(i)}| \leq \gamma_3$. Let us write

$$h_{ii}\hat{c}_1^{(i-1)}(1+\theta_1^{(i)})(1+\alpha_1^{(i)}) = h_{ii}\hat{c}_1^{(i-1)}(1+\alpha^{(i)}),$$

where

$$\alpha^{(i)} = \alpha_1^{(i)} + \theta_1^{(i)} + \alpha_1^{(i)}\theta_1^{(i)}; \quad \text{and using (5.8) yields} \quad |\alpha^{(i)}| \leq \gamma_4.$$

Then,

$$\hat{c}_2^{(i)} = \mathrm{fl}\Big[\hat{c}_2^{(i-1)}(1+\theta_1^{(i)}) - h_{ii}\hat{c}_1^{(i-1)}(1+\alpha^{(i)}) - h_{i-1,i}h_{i,i-1}(1+\hat{\alpha}_1^{(i)})\Big].$$

Now, we apply the modified model (5.6) to add $\hat{c}_2^{(i-1)}(1+\theta_1^{(i)}) - h_{ii}\hat{c}_1^{(i-1)}(1+\alpha^{(i)})$ and $h_{i-1,i}h_{i,i-1}(1+\hat{\alpha}_1^{(i)})$. The rest of the proof is similar to that of Theorem 6.2. □

We present the running error bounds for $\hat{c}_k$, $3 \le k \le n$, below.

Theorem 6.10. The error bound for $\hat{c}_k^{(k)}$ is given as

$$\begin{aligned}
|e_k^{(k)}| \;\le\; & |h_{kk}e_{k-1}^{(k-1)}| + \sum_{m=1}^{k-2} |h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}e_{k-m-1}^{(k-m-1)}| + u|\hat{c}_k^{(k)}| + \\
& \sum_{m=2}^{k-2} |h_{k-m,k}h_{k,k-1}\cdots h_{k-m+1,k-m}\hat{c}_{k-m-1}^{(k-m-1)}|\gamma_{2m+k+3} + \\
& \gamma_3|h_{kk}\hat{c}_{k-1}^{(k-1)}| + \gamma_{3k-2}|h_{1k}h_{k,k-1}\cdots h_{21}| + \gamma_{k+4}|h_{k-1,1}h_{h,k-1}\hat{c}_{k-2}^{(k-2)}|.
\end{aligned}$$

The error bound for $\hat{c}_k^{(i)}$, $3 < i \le n$ is given by

$$\begin{aligned}
|e_k^{(i)}| \;\le\; & |e_k^{(i-1)}| + |h_{ii}e_{k-1}^{(i-1)}| + \sum_{m=1}^{k-2} |h_{i-m,i}h_{i,i-1}\cdots h_{i-m+1,i-m}e_{k-m-1}^{(i-m-1)}| + \\
& u\big(|\hat{c}_k^{(i-1)}| + |\hat{c}_k^{(i)}|\big) + \sum_{m=2}^{k-2} |h_{i-m,i}h_{i,i-1}\cdots h_{i-m+1,i-m}\hat{c}_{k-m-1}^{(i-m-1)}|\gamma_{2m+k+3} + \\
& \gamma_{3k-2}|h_{i-k+1,i}h_{i,i-1}\cdots h_{i-k+2,i-k+1}| + \gamma_{k+4}|h_{i-1,1}h_{i,i-1}\hat{c}_{k-2}^{(i-2)}| + \\
& \gamma_4|h_{ii}\hat{c}_{k-1}^{(i-1)}|.
\end{aligned}$$

Proof. The proof is very similar to that of Theorem 6.7 and follows from applying standard and modified models of complex floating point numbers. In addition, we use (5.8) as in Theorem 6.9. □

6.4 Combined Error Bounds

In this section we present the combined error bounds of La Budde's method for real matrices. We first reduce the matrix A by using the Householder's transformations to its Hessenberg form H, and then determine coefficients of the characteristic polynomial of the computed Hessenberg matrix. To estimate the errors in coefficients of $p(\lambda)$ of A from both stages, we first investigate how much error is produced in the coefficients by reducing A to H. We use the following result.

Lemma 6.11 (page 351 in [48])**.** Let $\tilde{H}$ be the upper Hessenberg matrix computed in floating point arithmetic by applying Householder similarity transformations to the $n \times n$ real matrix A then $\tilde{H} = Q^T(A + G)Q$, where $Q^TQ = I$ and

$$\|G\|_F \leq \nu\, n^2 u \|A\|_F,$$

where ν is a small constant of order uity, u is the unit roundoff and $\|.\|_F$ denotes the Frobenius norm.

The above Lemma implies the following result.

Lemma 6.12. Let $\tilde{H} = H + E$ be the upper Hessenberg matrix computed in floating point arithmetic by applying Householder similarity transformations to the $n \times n$ real matrix A then,

$$\|E\|_2 \leq \nu n^2 u \|A\|_F,$$

where ν is a small constant of order unity, u is the unit roundoff and $\|.\|_F$ denotes the Frobenius norm.

Proof. From Lemma 6.11,

$$\tilde{H} = H + Q^T G Q.$$

This implies

$$\|E\|_2 \leq \|Q^T G Q\|_2 \leq \|Q\|_2 \, \|G\|_2 \, \|Q^T\|_2.$$

Because $\|Q\|_2 = \|Q^T\|_2 = 1$ and $\|G\|_2 \leq \|G\|_F$, we get by applying Lemma 6.11,

$$\|E\|_2 \leq n^2 u \|A\|_F.$$

□

We can use our perturbation results from chapter 3 to estimate how much error is introduced in the coefficients of $p(\lambda)$ of A when A is reduced to its Hessenberg form.

Lemma 6.13. Let $\tilde{H} = H + E$ be the upper Hessenberg matrix computed in floating point arithmetic by applying Householder similarity transformations to the $n \times n$ real matrix A. If $\tilde{c}_k$ are the coefficients of the characteristic polynomial of $\tilde{H}$ and $\|A\|_F < \frac{1}{\nu n^2 u}$, then

$$|\tilde{c}_k - c_k| \leq \nu n^2 \binom{n}{k} s_{k-1}^{(k)} \|A\|_F u + \mathcal{O}(u^2), \quad 1 \leq k \leq n,$$

where $s_{k-1}^{(k)}$ is the $(k-1)st$ elementary symmetric function in the k largest singular values of A and ν is a small constant of order unity.

Proof. The characteristic polynomials and singular values of A and H are the same; therefore, using Remark 3.19, we write

$$|\tilde{c}_k - c_k| \leq \binom{n}{k} s_{k-1}^{(k)} \|E\|_2 + \mathcal{O}(\|E\|_2^2).$$

Substituting the bound on $\|E\|_2$ from Lemma 6.12 in the above equation we obtain the result. □

We bound the error in the coefficients when La Budde's method is applied to the computed Hessenberg form $\tilde{H}$.

Theorem 6.14. Let $\tilde{H}$ be the upper Hessenberg matrix computed in floating point arithmetic by applying Householder similarity transformations to the $n \times n$ real matrix

A, and suppose that Algorithm 5 is used to determine the coefficients of $p(\lambda)$ of $\tilde{H}$. Suppose that $\hat{c}_k$ are the computed coefficients of $\tilde{H}$ from La Budde's method and $\|A\|_F < \frac{1}{\nu n^2 u}$.
For $1 \leq k \leq n$,

$$|c_k - \hat{c}_k| \leq \nu n^2 \binom{n}{k} s_{k-1}^{(k)} \|A\|_F u + |e_k^{(n)}| + \mathcal{O}(u^2),$$

where $|e_k^{(n)}|$ is the estimated running error bound of $\hat{c}_k$ and $s_{k-1}^{(k)}$ is the $(k-1)st$ elementary symmetric function in the k largest singular values of A. ν is a small constant of order unity.

Proof. From the triangle inequality, we write

$$|\hat{c}_k - c_k| \leq |\hat{c}_k - \tilde{c}_k| + |\tilde{c}_k - c_k|, \tag{6.1}$$

where $\tilde{c}_k$, $1 \leq k \leq n$ are the exact coefficients of the characteristic polynomial of $\tilde{H}$. From Lemma 6.13,

$$|\tilde{c}_k - c_k| \leq n^2 \binom{n}{k} s_{k-1}^{(k)} \|A\|_F u + \mathcal{O}(u^2), \quad 1 \leq k \leq n.$$

The error $|\hat{c}_k - \tilde{c}_k|$ is given by the running error bound $|e_k^{(n)}|$. Substituting the error bounds for $|\tilde{c}_k - c_k|$ and $|\hat{c}_k - \tilde{c}_k|$ in (6.1) yields the result. □

The above theorem shows that the error in $\hat{c}_k$ can be large due to two reasons: the first order condition number $\nu n^2 \binom{n}{k} s_{k-1}^{(k)} \|A\|_F$ and the error introduced by La Budde's method. As we discussed earlier, the error in La Budde's method can be large when $|c_k|$ is very small compared to preceding coefficients $|c_1|, \ldots, |c_{k-1}|$.

If A is normal, then H is also normal, and the coefficients of the characteristic polynomials of H are better conditioned. The combined error bound for the coefficients is given as follows.

Theorem 6.15. Under the assumptions of Theorem 6.14, if A is a normal matrix, then

$$|c_k - \hat{c}_k| \leq \nu n^2 \, (n-k+1) \, s_{k-1} \|A\|_F u + e_k + \mathcal{O}(u^2), \quad 1 \leq i \leq n,$$

where s_{k-1} is the $(k-1)st$ elementary symmetric function of singular values of A.

Proof. The proof is similar to that of Theorem 6.14 and follows from Remark 3.21. □

6.5 Numerical Tests

We implement La Budde's method on the test matrices presented in section 5.4. Our code is written in MATLAB 7.6(R2008a). As discussed in chapter 5 (section 5.4), we either know characteristic polynomials of these test matrices explicitly or we use MATLAB's symbolic toolbox to determine the exact coefficients of characteristic polynomials.

Test 1: Forsythe Matrix

We compute the characteristic polynomial of the Forsythe matrix of order 200. The characteristic polynomial of the Forsythe matrix is $p(\lambda) = \lambda^{200} - 10^{-10}$. We find that all the coefficients computed by La Budde's method are exact. Except for $\hat{c}_{200}$, the running error bounds for all coefficients are zero, and the relative running error bound for $\hat{c}_{200}$ is approximately 10^{-14}.

Test 2: Hansen's Matrix

We compute the characteristic polynomial of Hansen's matrix A of order 100 with the Sturm sequence method. We compare the relative errors in the coefficients, and observe that the computed coefficients are correct to at least 15 digits. The results of running error bounds vary over a wide range. We present the relative error results of the running error bounds in the following table.

Table 6.1: Running error bounds of $p(\lambda)$ of Hansen's Matrix

Range of Coefficients	Running Error Bounds
(c_1, c_{30})	$(10^{-15}, 10^{-14})$
(c_{31}, c_{59})	$(10^{-13}, 10^{-10})$
(c_{60}, c_{84})	$(10^{-9}, 10^{-1})$
(c_{85}, c_{100})	$(10, 10^{22})$

The running error bounds of the first thirty coefficients of $p(\lambda)$ of A are correct. As k grows, the running error bounds become more pessimistic. These bounds are determined from the computed values of coefficients of characteristic polynomials of principal submatrices of A. Because these intermediate quantities are large; therefore, the running error bounds are inaccurate.

Test 3: A matrix generated from the inverse of Hansen's Matrix

We compute $p(\lambda)$ of the matrix R presented in Test 3 of section 5.4 with La Budde's method. We compare the relative errors of the coefficients. The computed coefficients are correct to at least 9 digits. The running error bounds provide accurate information about the initial coefficients. The results for relative error bounds are summarized in the table below.

Table 6.2: Running error bounds of $p(\lambda)$ of the matrix R

Range of Coefficients	Running Error Bounds
(c_1, c_{30})	$(10^{-14}, 10^{-9})$
(c_{31}, c_{96})	$(10^{-9}, 10^{-1})$
(c_{97}, c_{100})	$(1.26, 2.05)$

Test 4: Tridiagonal Matrix

We compute the coefficients of $p(\lambda)$ of the tridiagonal matrix T presented in chapter 5. All the zero coefficients are computed exactly by La Budde's method. We calculate the relative errors in the non-zero coefficients, and find that all non zero coefficients are correct to at least 15 digits. Our running error bounds correctly predict the zero coefficients. The relative running error bounds for non zero coefficients indicate that the computed coefficients are correct to at least 14 digits.

Test 5: Frank Matrix

For the Frank matrix of order $n = 20$, La Budde's method produces the exact characteristic polynomial. However, as n grows, this method yields poor results for later coefficients. It is due to the fact that these coefficients are ill conditioned. The first order condition number of c_k is $\binom{n}{k} s_{k-1}^{(k)}$, where $s_{k-1}^{(k)}$ is the $(k-1)st$ elementary symmetric function in the k largest singular values of the Frank matrix. We observe that these quantities are enormous for a Frank matrix of large order. As an example, the computed value of the determinant (the exact value of the determinant is 1) of the Frank matrix of order 50 is approximately 10^{46}, whereas its first order condition number is approximately 10^{63}.

6.6 Comparison of La Budde's Method with the Eigenvalue Method

We compare the results of the tests presented in sections 5.4 and 6.5. In section 5.4 we computed the characteristic polynomials with MATLAB's "poly" command. In section 6.5 we computed the same values with La Budde's method. In these tests, as well as others not presented here, we find that La Budde's method is at least as accurate as MATLAB's poly command. In fact, in some cases, namely: Forsythe

Matrix, Frank Matrix of order 20, and tridiagonal matrix, La Budde's method is significantly more accurate. We also observe that for the symmetric positive definite matrices we considered, La Budde's method produces 2 or 3 more significant digits than MATLAB's poly command. The error bounds for both methods are accurate for most test matrices; however, as we have shown they can be pessimistic. These tests, as well as the fact that La Budde's method does not depend on the computation of eigenvalues for determining the characteristic polynomial supports the conclusion that La Budde's method is more accurate than MATLAB's poly command.

Chapter 7

Conclusion and Future Research

7.1 Our Contributions

The aim of this work was to investigate the numerical computation of the characteristic polynomial of a complex matrix A. In Quantum Physics, for instance, characteristic polynomials are required to calculate thermodynamic properties of systems of fermions. Characteristic polynomials attracted mathematicians in the middle of the twentieth century for determining the eigenvalues of A. Later work on numerical methods for computing characteristic polynomials seems to have stopped.

In our early research we found that little was known about the sensitivity of $p(\lambda)$ to perturbations in the matrix, i.e. if the matrix A is perturbed by E, then how do the coefficients of the characteristic polynomial of $A + E$ compare to those of $p(\lambda)$? As a first step toward the solution of our problem, we derived perturbation bounds for the coefficients of $p(\lambda)$ which we present in chapter 3. These perturbation bounds consist of elementary symmetric functions of singular values and suggest that coefficients of characteristic polynomials of normal matrices are better conditioned with regard to absolute perturbations than those of general matrices. We also improved relative and absolute perturbation bounds for determinants.

Once the conditioning of the coefficients is known, we can analyze numerical methods for computing characteristic polynomials. We presented the analysis of some general known methods for the computation of characteristic polynomials in chapter 4. Furthermore, we investigated the computation of $p(\lambda)$ of A from its eigenvalues in chapter 5. This method consists of two steps: first we compute the eigenvalues of A and then we determine the coefficients from the computed eigenvalues. We derived bounds that show the sensitivity of the coefficients of $p(\lambda)$ to changes in eigenvalues of a complex matrix A, when the eigenvalues of A are used to determine the coefficients. To determine the coefficients of $p(\lambda)$ from the computed eigenvalues of A, we investigated the numerical stability of the Summation Algorithm. We showed that the Summation Algorithm is forward stable. In addition, we used MATLAB's "poly" function to check the accuracy of the perturbation bounds for many test matrices. The "poly" function determines the characteristic polynomial of a given matrix from its eigenvalues. We found that the perturbation bounds accurately predict the conditioning of coefficients of the characteristic polynomial.

In chapter 6 we examined the numerical stability of La Budde's method for the computation of $p(\lambda)$. In this method, first we reduce the matrix A to its Hessenberg form H, and then the coefficients of $p(\lambda)$ of H are determined by successively computing the characteristic polynomials of leading principal submatrices of H. For our application in quantum physics, we modified La Budde's method to compute an individual coefficient c_k. We derived running error bounds for c_k, $1 \leq k \leq n$, which provide an estimate of the forward errors. We tested the accuracy of La Budde's method for characteristic polynomials of test matrices presented in chapter 5.

7.2 Conclusion

Our numerical tests suggest that La Budde's method gives more accurate results than the MATLAB's poly command, which computes the characteristic polynomial

of a matrix from its eigenvalues. We also observe in our experiments that La Budde's method is practically stable, i.e. for well conditioned coefficients, this method produces accurate results. In addition, the fact that La Budde's method does not depend on the computation of eigenvalues for determining the characteristic polynomial suggests that this method is superior to the computation of the characteristic polynomial of a matrix from its eigenvalues.

7.3 Future Research

This research was motivated by an application of characteristic polynomials in quantum physics. The matrices $A(s)$ in this application are dense and have no obvious structure. They are currently of order $n \leq 2000$. The computation of $p(\lambda)$ has to be started from scratch for every $A(s)$, due to the lack of exploitable relations among the matrices. La Budde's method has produced accurate results for test matrices of smaller order ($n \leq 100$). The coefficients of characteristic polynomials of the matrices $A(s)$ can become very large as the matrix dimension n grows, and double precision may not be enough for their computation. To address this problem we need a multiple precision package. We have spoken with David Bailey of Berkeley National Laboratory who has developed software for higher precision computations. He suggested to run the code that produces the matrices $A(s)$ with his QD package (double-double and quad-double precision)[1]. To further investigate the accuracy of La Budde's method we will compute the coefficients of the characteristic polynomials of the matrices $A(s)$ from their eigenvalues by using QD package and compare the results. We will also investigate classes of structured perturbations to which the coefficients are less sensitive.

[1] http://crd.lbl.gov/~dhbailey/mpdist/

Bibliography

[1] *http://www.mathworks.com.*

[2] F. B. Baker and M. R. Harwell, *Computing elementary symmetric functions and their derivatives: A didactic*, Applied Psychological Measurement, 20 (1996), pp. 169–192.

[3] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Num Anal., 27 (1990), pp. 776–791.

[4] S. Barnett, *Leverrier's algorithm: A new proof and extensions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 551–556.

[5] R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman Scientific & Technical, New York, 1987.

[6] ——, *Matrix Analysis*, Springer-Verlag New York, 1997.

[7] M. D. Bingham, *A new method for obtaining the inverse matrix*, J. Amer. Statist. Assoc., 36 (1941), pp. 530–534.

[8] J. Demmel and P. Koev, *Accuarate and efficient evaluation of Schur and Jack functions*, Math. Comp., 75 (2005), pp. 223–239.

[9] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[10] F. M. Dopico and P. Koev, *Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices*, SIAM J. Matrix Anal.Appl., 28 (2006), pp. 1126–1156.

[11] A. Eisinberg and G. Fedele, *A property of the elementary symmetric functions*, Calcolo Springer-Verlag, 42 (2005), pp. 31–36.

[12] V. N. Faddeeva, *Computational Methods of Linear Algebra*, Dover, New York, 1959.

[13] G. H. Fischer, *Introduction to the theory of Psychological tests*, Bern: Huber, first ed., 1974.

[14] S. Friedland, *Variation of tensor powers and permanents*, Linear Multilinear Algebra, C (1982), pp. 81–98.

[15] A. Galántai and C. J. Hegedüs, *Perturbation bounds for polynomials*, Numer. Math., 109 (2008), pp. 77–100.

[16] F. R. Gantmacher, *The Theory of Matrices, vol. I*, AMS Chelsea Publishing, Providence Rhode Island, 1998.

[17] M. Gasca and J. M. Peña, *On factorizations of totally positive matrices*, in Total positivity and its applications, vol. 359, Kluwer Academic Publishers Group, 1996, pp. 109–130.

[18] W. B. Givens, *Numerical computation of the characteristic values of a real symmetric matrix*, tech. report, Oak Ridge National Labortary, 1953.

[19] ——, *The characteristic value-vector problem*, J. Assoc. Comput. Mach., 4 (1957), pp. 298–307.

[20] S. K. Godunov, A. G. Antonov, O. P. Kiriljuk, and V. I. Kostin, *Guaranteed Accuracy in Numerical Linear Algebra*, Kluwer Academic Publishers, Dordrecht, 1993.

[21] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, 1983.

[22] R. T. Gregory and D. L.Karney, *A Collection of matrices for Testing Computational Algorithms*, Wiley, Interscience, first ed., 1969.

[23] S. J. Hammarling, *Latent Roots and Latent Vectors*, The University of Toronto Press, 1970.

[24] E. R. Hansen, *On the Danilewski method*, J. Assoc. Comput. Mach., 10 (1963), pp. 102–109.

[25] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, 1995.

[26] ——, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2nd ed., 2002.

[27] Z. Hong-Hao, Y. Wen-Bin, and L. Xue-Song, *Trace formulae of characteristic polynomial and Cayley Hamilton's theorem, and applications to chiral perturbation theory and general relativity*, Commun. Theor. Phys., 49, pp. 801–808.

[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[29] ——, *Topics in Matrix Analysis*, Cambridge University Press, London, 1991.

[30] P. Horst, *A method for determining the coefficients of the characteristic equation*, Ann. Math. Statistics, 6 (1935), pp. 83–84.

[31] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1964.

[32] A. S. Householder and F. L. Bauer, *On certain methods for expanding the characteristic polynomial*, Numer. Math., 1 (1959), pp. 29–37.

[33] J. Hubbard, *Calculations of partition functions*, Phys. Rev. Lett., 3 (1959), pp. 77–78.

[34] I. C. F. Ipsen and R. Rehman, *Perturbation bounds for determinants and characteristic polynomials*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 762–776.

[35] k. balasubramanian, *Characteristic polynomials of spirographs*, Journal of Mathematical Chemistry, 3 (1989), pp. 147–159.

[36] T. Kailath, *Linear Systems*, Englewood Cliffs, Prentice-Hall NJ, 1980.

[37] P. Koev, *Accurate eigenvalues and SVDs of totally nonnegative matrices*, SIAM J. Matrix Anal. Appl, 27 (2005), pp. 1–23.

[38] P. Lancaster and M. Tismenetsky, *The theory of matrices*, Computer Science and Applied Mathematics, Academic Press Inc., Orlando, FL, second ed., 1985.

[39] D. J. Lee and T. Schaefer, *Neutron matter on the lattice with pionless effective field theory*, Phys. Rev., C72 (2005), p. 024006.

[40] W. Li and W. Sun, *The perturbation bounds for eigenvalues of normal matrices*, Numer. Linear Algebra Appl., 12 (2005), pp. 89–94.

[41] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.

[42] P. Misra, S. Quintana, and P. V. Dooren, *Numerically reliable computation of characteristic polynomials*, in Proceedings of the American Control Conference, vol. 6, IEEE, 1995, pp. 4025–4029.

[43] D. S. Mitrinović, *Analytic Inequalities*, Springer, Heildelberg, 1970.

[44] M. J. Peláez and J. Moro, *Accurate factorization and eigenvalue algorithms for symmetric DSTU and TSC matrices*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1173–1198.

[45] Y. Z. Song, *A note on the variation of the spectrum of an arbitrary matrix*, Linear Algebra Appl., 342 (2002), pp. 41–46.

[46] R. L. Stratonovich, *On a method of calculating quantum distribution functions*, Soviet Phys. Dokley, 2 (1958), pp. 416–419.

[47] J. W. Wang and C. T. Chen, *On the computation of the characteristic polynomial of a matrix*, IEEE Trans. Automat. Control, 27 (1982), pp. 449–451.

[48] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

[49] ——, *The perfidious polynomial*, in Studies in Numerical Analysis, vol. 24, Math. Assoc. America, 1984, pp. 1–28.

Appendices

Appendix A

Matlab File: sturm.m

```
function[char]=sturm(A)
% The following code computes the coefficients of the
characteristic polynomial of a real symmetric matrix by using
modified Sturm Sequence Method (Algorithm 3 of chapter 6).

% Same code can be used for a non symmetric tri-diagonal matrix.

% INPUT:  A = a symmetric or a non symmetric tri-diagonal   matrix

% OUTPUT: char = a row vector containing coefficients of p(λ) of A


% Step 1: Use Householder's reduction to bring A to its tri-
  diagonal form by using hess function of MATLAB.

T=hess(A);
 % Step 2: Compute coefficients of p(λ) of T.

[n,n] = size(T);
% if k=n, then all coefficients are obtained. If some initial low
order coefficients are required, change k

k=n;
% c matrix stores coefficients of characteristic polynomials of
  principal submatrices of T

c = zeros(n,k);

c(1,1) = -T(1,1);

alpha=diag(T); % alpha stores diagonal entries of T

% subdiagonal entries of T are stored in gamma

gamma=zeros(n,1);

for i = 2:n
    gamma(i) = T(i,i-1);
end
%superdiagonal entries of T are stored in beta
beta=zeros(n,1);
for i = 2:n
    beta(i) = T(i-1,i);
end
```

```
%Compute the coefficients of the characteristic polynomial

for m=2:n
    for j=1:k
        if(m>=j)
            if(j==1)
                c(m,j) = c(m-1,j)-alpha(m);
            elseif(j==2)
                c(m,j) = c(m-1,j)-alpha(m)*c(m-1,j-1)-
                         beta(m)*gamma(m);
            else
                c(m,j) = c(m-1,j)-alpha(m)*c(m-1,j-1)-
                         beta(m)*gamma(m)*c(m-2,j-2);
            end
        end
    end
end

char=c(end, :);
```

Appendix B

Matlab File: labudde.m

```
function[char]=labudde(A)

%The following code computes the coefficients of the
characteristic polynomial of a complex matrix A by
using Labudde's Method. It uses Algorithm 5 of chapter
6
%INPUT: a complex matrix
%OUTPUT: coefficients of characteristic polynomial
%Step 1: Use Matlab's hess function to reduce A to
hessenberg form.

H=hess(A);
[n,n] = size(H);

%k=n produces all coefficients, if some initial low
order coefficients are required, k can be changed.
k=n;

%c matrix  stores coefficients of characteristic
polynomials of principal submatrices of H
c = zeros(n,k);

c(1,1) = -H(1,1);

% gamma stores subdiagonsal entries of H

gamma=zeros(n,1);
for s = 2:n
    gamma(s) = H(s,s-1);
end

for m=2:n
    for j=1:k
        if(j<=m)
            if(j==1)
                c(m,j) = c(m-1,j)-H(m,m);
            else
                Prod = gamma(m)*ones(j-1,1);
                Sum = 0;
                if(j>2)
```

```
for s=1:j-2
      Prod(s+1)= Prod(s)*gamma(m-s);
      Sum = Sum+(H(m-s,m)*Prod(s)*c(m-s-1,j-s-1));
         end
       Sum = Sum+(H(m-j+1,m)*Prod(j-1));
                End

                  if(j==2)
                   Sum = H(m-j+1,m)*Prod(j-1);
                 end
             c(m,j) = c(m-1,j)-H(m,m)*c(m-1,j-1)-Sum;
            end
        end
    end
end

char=c(end,:);
```