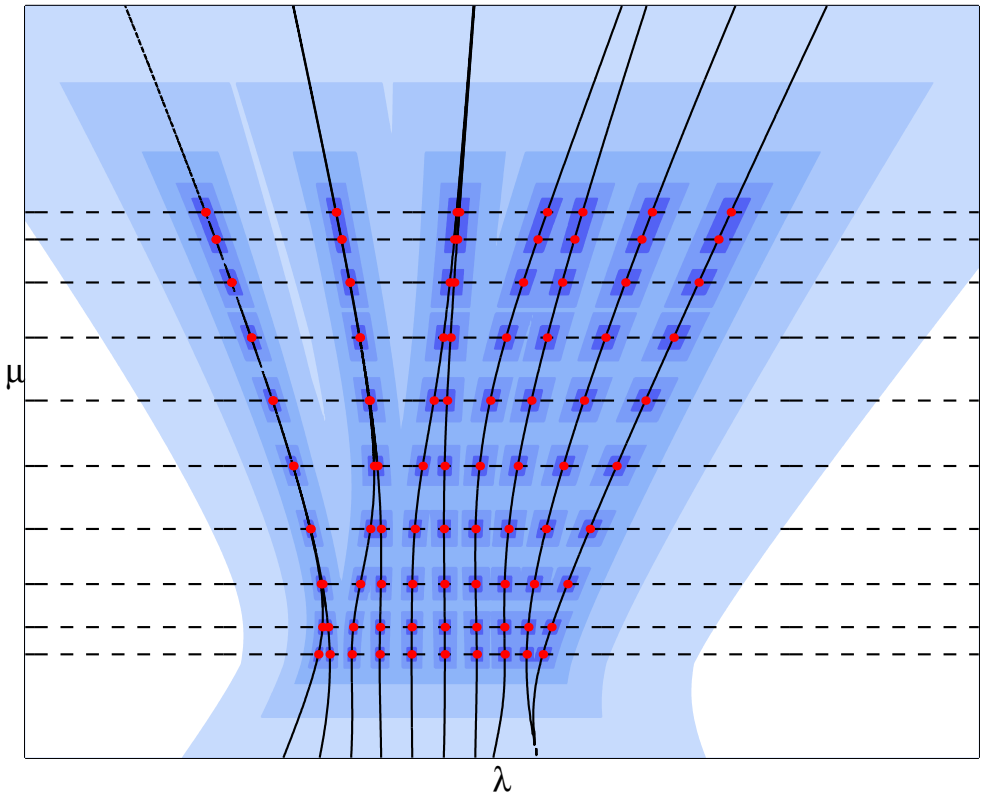
Subspace Methods for Eigenvalue Problems



Michiel E. Hochstenbach

Subspace Methods for Eigenvalue Problems

Deelruimte Methoden voor Eigenwaarde Problemen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de Rector Magnificus, Prof. dr. W. H. Gispen, ingevolge het besluit van het College van Promoties in het openbaar te verdedigen op woensdag 28 mei 2003 des middags te 12.45 uur

door

Michiel Erik Hochstenbach

geboren op 24 maart 1973 te Gouda

Promotor: Prof. dr. H.A. van der Vorst
Faculteit der Wiskunde en Informatica
Universiteit Utrecht

Mathematics Subject Classification: 65F15, 65F50, 65F35, 15A18, 15A69, 93E24.

Hochstenbach, Michiel Erik
Subspace Methods for Eigenvalue Problems
Proefschrift Universiteit Utrecht – Met een samenvatting in het Nederlands.

ISBN 90-393-3353-X

Aan Ineke

Contents

List of Algorithms

Chapter 1

Introduction

1.1 Background

This thesis treats a number of aspects of subspace methods for various eigenvalue problems. In this section we discuss the origin of eigenvalue problems, while the next section gives an introduction to the mathematical side. For more information see the references in Section 1.7.

Vibrations and their corresponding eigenvalues (or frequencies) arise in science, engineering, and daily life. Matrix eigenvalue problems come from a large number of areas, such as

- chemistry (chemical reactions, energy levels of a molecule),
- mechanics (design of earthquake resistant buildings)
- dynamical systems (stability, bifurcation analysis of systems depending on a parameter),
- Markov chains (stationary distribution of random processes),
- magneto-hydrodynamics,
- oceanography,
- economics,
- signal and image processing,
- control theory,
- pattern recognition,
- and statistics.

Eigenvalues and eigenvectors give valuable information about the behavior and properties of a matrix; therefore it may not be surprising that eigenvalue problems have been the subject of study for over one and a half century, partly before the current matrix notation became standard. Depending on the application, one is interested in one or more eigenvalues at the end of the spectrum, or rather in eigenvalues in the interior of the spectrum, or in the number of eigenvalues in an interval.

Methods for eigenvalue problems are often subdivided into two categories. The first category, the *direct methods* such as the QR-method and the divide-and-conquer method, aims to (accurately) find all eigenvalues of relatively small (say order 10^3) matrices.

Although these approaches work in an iterative way, they are called direct because they are (almost) guaranteed to converge in a fixed number of steps. These methods are efficient, and the underlying mathematics is quite well understood.

Many applications however, for instance in chemistry, give rise to eigenvalue problems where the size of the matrix easily exceeds one million. These problems often come from discretized partial differential equations; typically only a small portion of the eigenvalues is needed. Moreover, the matrices are often *sparse*, this means that the matrix contains relatively many entries which are zero. Therefore, one can compute a matrix-vector product economically, that is, quickly, also for large matrices. For these matrices, the direct approaches are often intractable, because they consume too much computer time and/or memory, even on modern (and future) computers. Because of all these reasons, *iterative methods*, and in particular the important subclass of *subspace methods*, are often the ones of choice for large sparse matrices. In a subspace method, the matrix is projected onto a low-dimensional subspace; the projected matrix is then solved by direct methods. In this way, we get approximate eigenpairs from a low-dimensional subspace.

For large sparse problems, there is often no such a thing as "*the* best method". The method of choice may depend upon certain properties of the matrix (structure, size), the data of interest (what, to which accuracy), the available operations (transpose of the matrix, preconditioner), and the machine architecture. In this thesis, we hope to give a contribution to the interesting and active field of subspace methods for eigenvalue problems. We study various eigenvalue problems, namely

- the (standard) eigenvalue problem,
- the generalized eigenvalue problem,
- the singular value problem,
- the polynomial eigenvalue problem,
- and the multiparameter eigenvalue problem.

The standard and generalized eigenproblem are the most common ones, originating from numerous applications. The singular value problem plays an important role in applications such as signal and image processing, control theory, pattern recognition, statistics, and search engines for the internet. But it also has a central position in the numerical linear algebra itself, for instance for the least squares problem, the numerical rank of a matrix, angles between subspaces, the sensitivity (condition) of the solution of linear systems, the pseudospectrum, and the (Euclidean) norm of a matrix.

The polynomial eigenvalue problem arises in the study of the vibrations of a mechanical system caused by an external force (the effects of the wind on a bridge), in the simulation of electronic circuits, and in fluid mechanics.

An example of the origin of the multiparameter eigenvalue problem is the mathematical physics when the method of separation of variables is used to solve boundary value problems.

An overview of the contributions of this thesis will follow in Section 1.5.

1.2 Various eigenvalue problems

We now briefly describe the mathematical formulation of the different types of eigenvalue problems that are studied. For more information we refer to [5] and other references collected in Section 1.7. See the appendix and Section 1.6 for notations that are used.

1.2.1 The standard eigenvalue problem

The *(standard) eigenvalue problem* (EP) is to find nontrivial solutions (i.e., $\lambda \in \mathbb{C}$, $x \in \mathbb{C}^n \backslash \{0\}$) to

$$Ax = \lambda x, \tag{1.2.1}$$

were A is a complex $n \times n$ matrix. Here λ is an *eigenvalue*, x is a *(right) eigenvector*, and (λ, x) is called an *eigenpair*. A vector $y \in \mathbb{C}^n \backslash \{0\}$ satisfying $y^*A = \lambda y^*$ is a *left eigenvector*. When A is *normal* ($A^*A = AA^*$), there exists an orthonormal basis of eigenvectors. In this case, a right eigenvector is also a left eigenvector corresponding to the same eigenvalue. When A is *Hermitian* ($A = A^*$), then, in addition, the *spectrum*, the set of all eigenvalues, is a subset of $\mathbb{R}$. *Real symmetric* matrices ($A \in \mathbb{R}^{n \times n}$, $A = A^T$) have a real spectrum and an orthonormal basis of real eigenvectors. A matrix is *complex symmetric* if $A \in \mathbb{C}^{n \times n}$ and $A = A^T$. Although these matrices are in general not normal, they have the property that if x is an eigenvector corresponding to a simple eigenvalue, then $\overline{x}$ is the corresponding left eigenvector.

1.2.2 The singular value problem

Although the *singular value problem* (SVP) does not contain the word "eigenvalue", it is closely related to the (Hermitian) eigenproblem. Given the $m \times n$ matrix A, the singular value problem is to find a *singular triple* (σ, u, v), where $\sigma \geq 0$, and $u \in \mathbb{C}^m \backslash \{0\}$ and $v \in \mathbb{C}^n \backslash \{0\}$ satisfy

$$\begin{aligned} Av &= \sigma u, \\ A^*u &= \sigma v. \end{aligned}$$

Here, σ is called a *singular value*, u a *left singular vector*, and v a *right singular vector*. The singular value problem gives rise to two different equivalent Hermitian eigenvalue problems. First, the nonzero eigenvalues of the $n \times n$ matrix A^*A or the $m \times m$ matrix AA^* are the squares of the nonzero singular values of A. Their eigenvectors are the right and left singular vectors of A, respectively. Second, the eigenvalues of the *augmented matrix*

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}$$

are plus and minus the singular values of A, and we can extract the left and right singular vectors from the eigenvectors by taking the first and second part (see Section 3.2). The singular value problem is subject of study in Chapters 3 and 4.

1.2.3 The generalized eigenvalue problem

The *generalized eigenvalue problem* (GEP)

$$Ax = \lambda Bx$$

is a generalization of the EP. As for the EP, λ is an eigenvalue, $x \neq 0$ is a (right) eigenvector, and a vector $y \neq 0$ satisfying $y^*A = \lambda y^*B$ is a left eigenvector. When B is nonsingular, this problem can be transformed to a standard eigenvalue problem by left multiplying by the inverse of B, although computationally, this is often unattractive to do. Therefore, and also because a singular B leads to new properties unique to the generalized eigenvalue problem, this problem fully deserves its own treatment. The GEP will be the subject of Section 2.7.1.

1.2.4 The polynomial eigenvalue problem

The *polynomial eigenvalue problem* (PEP)

$$\left(\lambda^l A_l + \lambda^{l-1} A_{l-1} + \cdots + \lambda A_1 + A_0\right) x = 0$$

is a generalization of the EP and the GEP. The concepts of eigenvalue and left and right eigenvector are defined similarly as for these problems. In particular, $l = 2$ gives the *quadratic eigenvalue problem* (QEP), and $l = 1$ yields the GEP. Aspects of the polynomial eigenvalue problem are discussed in Chapter 8 and in Section 2.7.3.

1.2.5 The multiparameter eigenvalue problem

The *multiparameter eigenvalue problem* (MEP) is another generalization of the EP and the GEP. Here the problem is to find a k-tuple values $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k) \in \mathbb{C}^k$ and nonzero vectors $x_i \in \mathbb{C}^{n_i}$ for $i = 1, \ldots, k$ such that

$$\left(V_{i0} - \sum_{j=1}^{k} \lambda_j V_{ij}\right) x_i = 0, \quad i = 1, \ldots, k,$$

where the V_{ij} are $n_i \times n_i$ matrices over $\mathbb{C}$. When $k = 1$, this yields the GEP.

The k-tuple $\boldsymbol{\lambda} \in \mathbb{C}^k$ is called an *eigenvalue* and the tensor product $\boldsymbol{x} = x_1 \otimes x_2 \otimes \cdots \otimes x_k$ is the corresponding *(right) eigenvector*. A left eigenvector can be defined similarly. The multiparameter problem, and in particular the two-parameter eigenvalue problem, is the subject of Chapters 5, 6, and 7.

1.2.6 Relations

Relations between the various eigenvalue problems are summarized in Tables 1.1 and 1.2 ("$\uparrow$" stands for "can lead to", and "$\subset$" means "is a special case of"):

TABLE 1.1: Acronyms for the various eigenvalue problems

acronym	meaning
SEP	symmetric eigenvalue problem
SVP	singular value problem
HEP	Hermitian eigenvalue problem
NEP	normal eigenvalue problem
CSEP	complex symmetric eigenvalue problem
EP	standard eigenvalue problem
GEP	generalized eigenvalue problem
QEP	quadratic eigenvalue problem
PEP	polynomial eigenvalue problem
MEP	multiparameter eigenvalue problem

TABLE 1.2: Relations between the various eigenvalue problems

$$\begin{array}{ccccccccccc} & & & & & & & & \subset & \text{QEP} & \subset \text{ PEP} \\ \text{SEP} & \subset & \text{HEP} & \subset & \text{NEP} & \subset & \text{EP} & \subset & \text{GEP} & & \\ & & \uparrow & & & & \cup & & & \subset & \text{MEP} \\ & & \text{SVP} & & & & \text{CSEP} & & & & \end{array}$$

1.3 Subspace methods

In this section we consider the standard eigenvalue problem (1.2.1). As mentioned in Section 1.1, subspace methods may be used for the numerical solution of eigenvalue problems. The idea of subspace methods is to compute accurate eigenpairs from low-dimensional subspaces. This approach reduces computational time and memory usage and thus enables us to tackle larger problems that are too expensive for methods that work in the entire space.

A subspace method to find an eigenpair works as follows. We start with a given search subspace from which approximations to eigenpairs are computed *(extraction)*. In the extraction we usually have to solve a smaller eigenvalue problem of the same type as the original one. After each step we expand the subspace by a new direction *(expansion)*. In some methods, but not all, the expansion depends upon the selected approximation. The idea is that, as the search subspace grows, the eigenpair approximations will converge to an eigenpair of the original problem. In order to keep computation costs low, we usually do not expand the search space to the whole space. If the process does not converge in a certain number of iterations, then the method is *restarted* with a few selected approximations as the basis of a new search space. If one or more eigenpairs have already been found, and we want to find other pairs, we can use *deflation* techniques to avoid finding the same pair again.

In the following subsections we discuss various aspects of subspace methods. For an overview of subspace methods see [5] and other references in Section 1.7.

1.3.1 Subspace extraction

Let $\mathcal{U}$ be a k-dimensional search space (a subspace of $\mathbb{C}^n$ or $\mathbb{R}^n$), where one should think of the typical situation $k \ll n$. Subspace extraction deals with the question: having the search space $\mathcal{U}$, how do we get approximate eigenpairs (θ, u), so $\theta \approx \lambda$ and $u \approx x$, with $u \in \mathcal{U}$? Given θ and u, the *residual* is defined by $r = Au - \theta u$. When (θ, u) is an exact eigenpair, the residual is zero. To find approximate eigenpairs, a common approach is to impose a *Galerkin condition* on the residual, that is, to require that the residual is orthogonal to a certain test space. In the *Ritz–Galerkin* approach, the test space is equal to the search space, while in *Petrov–Galerkin* variants it is not. The approximations that arise in this way are called *Ritz/Petrov values* and *Ritz/Petrov vectors* (respectively *Ritz/Petrov pairs*). In this context, it should be noted that some authors only use the terms Ritz value and Ritz vector when A is Hermitian. In this case, the Ritz–Galerkin approach is also called the *Rayleigh–Ritz* method. Sometimes this name is also used for non-Hermitian matrices.

Especially for interior eigenvalues, the performance of Rayleigh–Ritz can be disappointing, in the sense that the resulting approximations are of poor quality (see, e.g., [82, p. 282]). One option is to compute a *refined Ritz vector* after the Rayleigh–Ritz process (see, e.g., [43] and [82, p. 289]). Another alternative that may lead to better approximate eigenpairs is the *harmonic Rayleigh–Ritz* procedure (see, e.g., [82, p. 292]). In this thesis we will consider generalizations of standard Rayleigh–Ritz, harmonic Rayleigh–Ritz, and refined Ritz vectors for various eigenvalue problems.

1.3.2 Subspace expansion

Some subspace methods perform a simple action for the subspace expansion: they (implicitly) multiply a vector repeatedly by the matrix A. For Hermitian A, this gives the Lanczos method [51], for non-Hermitian A this leads to Arnoldi [2]. Both of these methods construct a *Krylov subspace* of dimension k generated by A and a starting vector u:

$$K_k(A, u) = \text{span}\{u, Au, \ldots, A^{k-1}u\}.$$

Other methods use the residual r to expand the search space. The Davidson method [19] preconditions this residual. Jacobi–Davidson (JD) [75] expands the search space by the (approximate) solution of the so-called *correction equation*, see Section 2.2. The fact that in these methods all iterates are stored to build up a search space is referred to as *subspace acceleration*. When, as in Jacobi–Davidson, the expansion can be seen as a Newton step, one also speaks of an accelerated Newton method. Often one does not solve the Newton equation to full precision, in this case the term *accelerated inexact Newton* is used. In this thesis, we will generalize the Jacobi–Davidson method to other eigenvalue problems.

1.3.3 Two-sided subspace methods

Characteristic for (ordinary or one-sided) subspace methods is that the test space coincides with the search space, or some transformation of the search space. Two-sided

subspace methods, on the other hand, build up a search space and a test space independently of each other. Two-sided Lanczos [51] uses multiplication by A and by A^* for this goal. In Chapter 2 we present a two-sided Jacobi–Davidson method. A pro of two-sided methods is that the subspace extraction may yield better approximations to the eigenpair; sometimes the resulting projected system is also easier to solve. Some of the cons are that two-sided methods are often more expensive per step and that they may suffer from a breakdown. The oblique projections may also cause problems with stability.

1.3.4 Asymptotic convergence

In exact arithmetic and if no breakdown occurs, all subspace methods will (trivially) converge in a finite number of steps in the absence of restarts. In [90, p. 652], the following is stated about the convergence of Ritz values to eigenvalues:

> "Strictly mathematically speaking it is not very meaningful, of course, to speak of convergence and convergence behavior of Ritz values, in view of the finiteness of the set of Ritz values. However, as is well known, in many practical situations one or more extremal eigenvalues are approximated by the corresponding Ritz values to a sufficient degree of accuracy long before their degree reaches the dimension of the matrix, and in this stage of the process those Ritz values display a behavior which is very reminiscent of that of a converging infinite sequence close to its limit. It is this that we have in mind when speaking of convergence and convergence behavior."

One can make similar statements about the convergence of approximate eigenvectors and the subspace method itself. Thus, by the "*asymptotic convergence*" of subspace methods, we mean the convergence behavior of these methods in a situation where we have a (very) good approximation to an eigenpair, rather than the situation where the dimension of the subspace goes to infinity.

1.4 Various issues

1.4.1 Modified Gram–Schmidt

In a search space method, it is often of practical importance to have an orthonormal basis of a subspace at one's disposal. A common tool to obtain such a basis is the Gram–Schmidt algorithm. It is well known (see for example [31, pp. 231–232]) that classical Gram–Schmidt may lose orthogonality in finite precision arithmetic. Modified Gram–Schmidt, which rearranges the calculations, does a better job, but still can be insufficient in the case of an almost dependent set of vectors. Moreover, the method is not parallelizable. Repeating the (modified) Gram–Schmidt method once gives good numerical properties ("twice is enough", see, for instance, the discussion in [7, Section 2.4.5]).

In this thesis, we will denote any numerically stable form of Gram–Schmidt, (such as repeated (modified) Gram–Schmidt) by the acronym MGS because most readers will be more familiar with this than with RGS or RMGS.

1.4.2 Perturbation theory

When we have approximated an eigenpair, we may be interested in the error. While it is often expensive (or impossible) to compute or bound the *(forward) error*, the *backward error* may be readily available. This is a measure of the perturbation of the matrix that is necessary such that the computed eigenpair is an exact eigenpair of the perturbed matrix. The *condition number* of an eigenvalue (or eigenvector) gives information about the sensitivity of that eigenvalue (or eigenvector) for perturbations in the matrix. The *pseudospectrum* of a matrix gives a graphical oversight of the sensitivity of more (or all) eigenvalues at the same time. The following relation often holds approximately:

$$\text{forward error} \lesssim \text{condition number} \cdot \text{backward error}.$$

Chapter 7 is dedicated to perturbation theory of the multiparameter eigenvalue problem; see Section 1.7 for references on the subject of perturbation theory and pseudospectra.

1.4.3 Numerical experiments

Most numerical experiments are carried out in MATLAB 5 on a SUN workstation. We use some typical MATLAB notation in the thesis, such as $\mathrm{diag}(1:n)$ for the diagonal matrix constructed from scalars $1, \ldots, n$ and $[\,]$ for the empty matrix; see also the appendix on notations. When we used MATLAB's function `rand(m,n)` to create a $m \times n$ matrix with random entries (chosen from a uniform distribution on the interval (0,1)), we first put the "seed" to zero by the command "`rand('seed',0)`" so that our results are reproducible. Additionally, we used MAPLE 5 for Section 8.4.

As mentioned in Section 1.1, most of the methods developed in this thesis are designed for large sparse matrices. We would like to remark that partly due to limitations of MATLAB, the size of the matrices in the numerical experiments does not exceed $\mathcal{O}(10^3)$. Although we realize that some practical matters have to be taken care of for (much) larger matrices, we do not expect major obstacles in an implementation. Moreover, although preconditioning is an important (or even crucial) subject, we do not pay special attention to the choice of a preconditioner in this thesis. In most cases, we take an (inexact) LU decomposition, often based on a target.

MATLAB codes of all methods are available from the author on request.

1.5 Overview

Part of this thesis is formed by four chapters that consider Jacobi–Davidson type methods for various eigenvalues problems:

- for the (nonnormal) standard, complex symmetric, generalized, and polynomial eigenvalue problem in Chapter 2;
- for the singular value problem in Chapter 3 (with Chapter 4 as a continuation);
- and for the multiparameter eigenvalue problem (especially the case of two parameters) in Chapters 5 en 6.

To begin with, we study two Jacobi–Davidson type methods for *nonnormal matrices*, called *two-sided* and *alternating Jacobi–Davidson* in Chapter 2. For these matrices, the right and left eigenvectors are generally not identical, as is the case for normal matrices. This motivates the presence of two search spaces, one for the right, and one for the left eigenvector. The search space for the left vector is the test space for the right vector and vice versa. The correction equation, that serves for the expansion of the search spaces, contains oblique projections, instead of the orthogonal projections that are characteristic for the standard Jacobi–Davidson method. These methods can be applied to the standard, the complex symmetric, the generalized, and the polynomial eigenvalue problem.

Chapter 3 introduces a Jacobi–Davidson type method for the *singular value problem*. As in Chapter 2 we have two search spaces, this time one for the right, and one for the left singular vector. This gives rise to a method with cubic convergence when the correction equation is solved exactly. In practice, this equation will often be solved inexactly, resulting in linear convergence. The method can be seen as an accelerated inexact Newton process and as an accelerated inexact Rayleigh quotient iteration. In Chapter 4, special attention is given to the approximation of the *smallest* and *interior singular values*. For these values, the standard Galerkin subspace extraction is no longer satisfactory. Just as for the standard eigenvalue problem, harmonic and refined approaches are more promising. We also discuss applications of the methods to the least squares problem and the approximation of a matrix by means of a truncated singular value decomposition.

Chapters 5 and 6 treat a Jacobi–Davidson type method for the *multiparameter eigenvalue problem*, in particular for the case of two parameters. In Chapter 5 we consider the so-called *right definite* multiparameter eigenvalue problem. In the case of two parameters, we have again two search spaces, one for each of the components of the decomposable tensor. The extraction of the search space is done by a generalization of the Rayleigh–Ritz method, that ensures monotonic convergence to the extreme eigenvalues. For the subspace expansion, we present two different correction equations: one with orthogonal one-dimensional projections which neglect second-order terms, and one with two-dimensional oblique projections that only disregards third-order terms. Because standard deflation techniques are not applicable for this problem, we use a selection criterion for the Ritz values when we are interested in more eigenpairs.

In Chapter 6, we study the wider class of the *nonsingular* multiparameter eigenvalue problems. This is a challenging problem, where we need many techniques to attack it. For instance, we choose here for a two-sided approach (different test and search spaces), comparable to Chapter 2.

In Chapter 7, we examine numerical important aspects of the multiparameter problem: *backward error* and *condition* of eigenvalues and eigenvectors. These concepts give an indication how good a certain obtained approximation is, and how sensitive the eigenvalues and eigenvectors are for perturbations in the problem. Also, the *pseudospectrum* for the multiparameter problem is introduced. This may give an impression of the sensitivities of a couple or all eigenvalues.

For the standard eigenvalue problem, the extraction of Ritz pairs from a search space is well studied. For the *polynomial eigenvalue problem* the situation is less clear. In Chapter 8, we consider approximations to an eigenvalue that can be obtained from a

certain search space. The emphasis is on the quadratic eigenvalue problem and one-dimensional search spaces. Three new methods are given, based on a Galerkin or minimum residual approach. The methods are compared using perturbation results and backward errors, and then generalized to general polynomial problems and extraction from more-dimensional search spaces.

In Chapter 9, we develop *probabilistic bounds* for the extreme eigenvalues of a Hermitian matrix with the Lanczos method. These bounds are obtained with Lanczos, Ritz and Chebyshev polynomials. Because we assume that the starting vector contains a sufficient component of the desired eigendirection, we thus get bounds that are correct with a certain (large) probability. The bounds may be used as a stopping criterion. As a second application of the techniques, we get an estimation for the number of steps of the Lanczos method that are (still) necessary to get an extreme eigenvalue with a prescribed tolerance.

Chapters 2 through 9 have appeared as separate papers. For this thesis, they have all been edited to some extent, varying from small editorial changes, to enlargement by extra subsections. Some notations have been changed to ensure uniformity. Chapter 2 (without Section 2.7.2) is based on [38]:

> M. E. Hochstenbach, G. L. G. Sleijpen, *Two-sided and alternating Jacobi–Davidson,* Lin. Alg. Appl. 358(1-3), pp. 145–172, 2003, reprinted with permission from Elsevier

while Section 2.7.2 is a summary of [1]:

> P. Arbenz, M. E. Hochstenbach, *A Jacobi–Davidson method for complex symmetric matrices,* Preprint 1255, Dept. of Math., Utrecht University, September 2002.

Chapters 3 and 4 are essentially [33]:

> M. E. Hochstenbach, *A Jacobi–Davidson type SVD method,* SIAM J. on Sci. Comp. 23(2), pp. 606–628, 2001. Second place student paper competition 6th Copper Mountain Conference 2000

and the following-up paper [34]:

> M. E. Hochstenbach, *Harmonic and refined extraction methods for the singular value problem, with applications in least squares problems,* Preprint 1263, Dept. of Math., Utrecht University, December 2002. Winner travel award student/new PhD paper competition 6th International Symposium on Iterative Methods in Scientific Computing,

but the chapters have been integrated and extended (Sections 3.6.2 and 3.7.9 are new). Chapter 5 is based on [36]:

> M. E. Hochstenbach, B. Plestenjak, *A Jacobi–Davidson type method for a right definite two-parameter eigenvalue problem,* SIAM J. on Matrix Anal. Appl. 24(2), pp. 392–410, 2002,

but Sections 5.4.3, 5.4.4, and 5.7 are new. Chapter 6 is the following-up paper [35]:

M. E. Hochstenbach, T. Košir, B. Plestenjak, *A Jacobi–Davidson type method for the two-parameter eigenvalue problem,* Preprint 1262, Dept. of Math., Utrecht University, November 2002.

Chapter 7 is [37]:

M. E. Hochstenbach, B. Plestenjak, *Backward error, condition and pseudospectra for the multiparameter eigenvalue problem,* Preprint 1225, Dept. of Math., Utrecht University, February 2002.

Chapter 8 is [39]:

M. E. Hochstenbach, H. A. van der Vorst, *Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem,* Preprint 1212, Dept. of Math., Utrecht University, November 2001. Accepted for publication in SIAM J. on Sci. Comp. Winner student/new PhD paper competition 7th Copper Mountain Conference 2002.

Chapter 9 has appeared as [95]

J. L. M. van Dorsselaer, M. E. Hochstenbach, H. A. van der Vorst, *Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method,* SIAM J. on Matrix Anal. Appl. 22(3), pp. 837–852, 2000.

Table 1.3 gives an overview of the contents of some of the chapters with respect to the two main aspects of subspace methods: extraction and expansion.

Table 1.3: Contents of some of the chapters, with respect to subspace extraction and expansion.

problem	extraction	expansion
EP	Ch. 2	Ch. 2
SVP	Ch. 3, 4	Ch. 3
CSEP	Sec. 2.7.2	Sec. 2.7.2
GEP	Sec. 2.7.1	Sec. 2.7.1
QEP	Ch. 8	Sec. 2.7.3
PEP	Sec. 8.3.1	Sec. 2.7.3
MEP	Ch. 5, 6	Ch. 5, 6

1.6 Notations

In this thesis we use the standard conventions in numerical linear algebra, sometimes called the *Householder notation.* Capital Roman letters denote matrices or operators. Vectors are indicated by lowercase Roman letters, and lowercase Greek letters stand for scalars. A script letter (e.g., $\mathcal{U}$) denotes a subspace, where it is a custom that the corresponding capital Roman letter (e.g., U) stands for a matrix of which the columns form a (often, but not always, orthonormal) basis for that subspace. We call such a matrix a *search matrix.* Letters in boldface denote a tuple of corresponding items, for instance, $\mathbf{A}$ stands for a tuple of matrices, and $\boldsymbol{\alpha}$ for a tuple of scalars.

There are, however, some exceptions on the general rules above, for example m and n are standardly used for the size of a matrix. Sometimes the usual notations can also be slightly overloaded. For example, u denotes an approximation to a *right eigenvector* (Chapter 2), but also an approximation to a *left singular vector* (Chapters 3 and 4). Likewise, v is used for an approximation to a *left eigenvector*, but also for an approximation to a *right singular vector*. The same notational overload concerns x (*right eigenvector* as well as *left singular vector*), and y (*left eigenvector* and *right singular vector*). An approximation to an eigenvalue λ is often indicated by θ or ρ (to stress the fact that it is a Rayleigh quotient).

While many notations are summarized in the table in an appendix, we like to highlight some specific ones. In this thesis $\|\cdot\|$ (without subscript) always stands for the Euclidean norm $\|\cdot\|_2$. By $\mathrm{span}(A)$, the space spanned by the columns of A is meant. Be warned: while $\lambda_j(A)$ is used for the jth *smallest* eigenvalue of a Hermitian A, $\sigma_j(A)$ denotes the jth *largest* singular value of A. The letter κ is used for both the condition number of an eigenvalue $\kappa(\lambda)$ and that of a matrix $\kappa(A) = \|A\| \cdot \|A^{-1}\|$. Besides for a matrix, the letter B is used in Chapter 7 for Euler's beta function $B(\alpha, \beta)$. The letter e is used in e_j, the jth canonical vector, and for "error vectors".

In addition to Table 1.1, Table 1.4 lists acronyms that are used in the thesis.

TABLE 1.4: Acronyms used in the thesis.

Acronym	meaning
Bi-MGS	MGS for bi-orthogonal bases
BiCG	bi-conjugate gradients
COCG	complex orthogonal conjugate gradients
CSYM	solver for a complex symmetric linear system
GMRES	generalized minimum residual method
MINRES	minimum residual method
JD	Jacobi–Davidson
JDCS	Jacobi–Davidson for the complex symmetric eigenvalue problem
JDSVD	Jacobi–Davidson for the singular value problem
MGS	numerically stable form of Gram–Schmidt
MGS-CS	MGS for a complex orthogonal basis
MV	matrix-vector product
QMR	quasi-minimal residual method
RQI	Rayleigh quotient iteration
SVD	singular value decomposition

1.7 Literature

Valuable sources for eigenvalue problems in general are (in reverse chronological order) Van der Vorst [91], Stewart [82], Bai et. al. [5], Saad [69], Stewart & Sun [83], Horn & Johnson [40, 41], Golub & van Loan [31], Parlett [61], Varga [98], and Wilkinson [101]. Some of the papers and books that are cited throughout the thesis, are especially relevant for topics as in Table 1.5 (per topic in chronological order).

TABLE 1.5: Some references divided per topic.

subject	references
origin of eigenvalue problems	[69]
review on eigenvalue problems	[92, 5, 30]
symmetric eigenvalue problem	[61]
Lanczos	[51, 61, 68, 62, 18]
two-sided Lanczos	[51]
Arnoldi	[2]
Davidson	[19]
Jacobi–Davidson	[75, 73, 76, 78, 81, 23, 77, 89]
Lanczos bidiagonalization	[28, 29, 17, 71]
complex symmetric eigenvalue problem	[18]
quadratic eigenvalue problem	[86]
polynomial eigenvalue problem	[50, 73, 84, 85]
multiparameter eigenvalue problem	[3, 4, 8, 9, 10, 11, 13, 79, 100, 21, 6, 47, 64, 65]
singular value problem	[29, 17, 41, 96, 97, 63]
perturbation theory	[46, 83]
pseudospectrum	[87, 88, 85]
Rayleigh quotient iteration	[55, 56, 50, 60, 61]
harmonic Rayleigh–Ritz	[54, 58, 82, 74]
refined Ritz vectors	[54, 43, 82]
accelerated inexact Newton	[22]
least squares problem	[59, 7]
Chebyshev polynomials	[67]

Chapter 2

Two-sided and alternating Jacobi–Davidson

Abstract. We discuss two variants of a two-sided Jacobi–Davidson method, which have asymptotically cubic convergence for nonnormal matrices, and aim to find both right and left eigenvectors. These methods can be seen as Jacobi–Davidson analogues of Ostrowski's two-sided Rayleigh quotient iteration. Some relations between (exact and inexact) two-sided Jacobi–Davidson and (exact and inexact) two-sided Rayleigh quotient iteration are given, together with convergence rates.

Furthermore, we introduce an alternating Jacobi–Davidson process that can be seen as the Jacobi–Davidson analogue of Parlett's alternating Rayleigh quotient iteration. The methods are extended to the generalized, complex symmetric, and polynomial eigenproblem. Advantages of the methods are illustrated by numerical examples.

Key words: Jacobi–Davidson, Rayleigh quotient iteration, Ostrowski's two-sided Rayleigh quotient iteration, Parlett's alternating Rayleigh quotient iteration, two-sided Lanczos, correction equation, nonnormal matrix, accelerated inexact Newton, rate of convergence, generalized eigenproblem, complex symmetric eigenproblem, polynomial eigenproblem.

AMS subject classification: 65F15, 65F50.

2.1 Introduction

We are interested in the computation of one or more eigenvalues and the corresponding left and right eigenvectors of the (possibly nonnormal) matrix A. It is well known that when Rayleigh quotient iteration (RQI) converges to a simple eigenvalue of a normal matrix, the asymptotic convergence rate is cubic (see, for example, [60, p. 683] and [61, p. 77]). For a nonnormal eigenvalue of a nonnormal matrix, RQI has locally quadratic convergence at best [60, p. 688].

*Based on joint work with Gerard L. G. Sleijpen and Peter Arbenz, see Section 1.5.

Ostrowski's two-sided RQI [55] works with the *two-sided (or generalized) Rayleigh quotient*

$$\theta(u, v) := \frac{v^* A u}{v^* u},$$

where u and v are approximate right and left eigenvectors. It can be shown that when two-sided RQI converges to a simple eigenvalue, the local convergence is cubic (see [60, p. 689] and Section 2.3). In (two-sided) RQI, one has to solve linear systems of the form $(A - \theta I)\widetilde{u} = u$. For large sparse matrices, these computations, and therefore (two-sided) RQI as a whole, may be less attractive.

Jacobi–Davidson (JD) [75] is an efficient method to compute a few eigenvalues and corresponding (right) eigenvectors of A. The essence of JD is its correction equation, where the shifted operator $A - \theta I$ is restricted to the subspace orthogonal to the current approximation to an eigenvector. When we solve this equation exactly, then JD can be considered as accelerated RQI (see [75] and Section 2.3).

Because of this, it is of interest to investigate JD analogues of two-sided RQI, leading to an acceleration of two-sided RQI. The idea of a two-sided JD is already, though somewhat hidden, present in [73], in particular in Remark 3.5 and Section 5.1.3. We will see that two-sided JD has two search spaces, one for the right and one for the left eigenvector. When the correction equations are solved exactly, the method has locally cubic convergence.

In practice, it is neither necessary nor advisable to solve the correction equation in the JD method accurately. Instead, we may solve it only approximately, for instance to a certain precision. This principle can also be applied to the two-sided processes, leading to inexact two-sided JD and inexact two-sided RQI. At the price of slower convergence, the methods thus become computationally more attractive. An attempt to merge the two search spaces of two-sided JD gives rise to alternating JD, which can be viewed as an acceleration of Parlett's alternating RQI [60].

This chapter has been organized as follows. Section 2.2 introduces some notations and definitions and gives a presentation of JD. In Section 2.3 we review Ostrowski's two-sided RQI, and in Section 2.4 we consider two flavors of two-sided JD. Inexact variants of these two-sided methods, and some relations between them, as well as convergence rates, can be found in Section 2.5. Section 2.6 proposes alternating JD, and Section 2.7 extends the two-sided methods to the complex symmetric, generalized, and polynomial eigenvalue problem. In Section 2.8 we discuss various aspects of the methods. Numerical experiments are presented in Section 2.9, and a discussion and some conclusions can be found in Section 2.10.

2.2 Jacobi–Davidson and Rayleigh quotient iteration

Let us first introduce some notations. Throughout this chapter, λ denotes a simple eigenvalue of the $n \times n$ matrix A, $n > 1$, with x and y as its normalized right and left eigenvectors. The (finite) condition of λ is equal to $\kappa(\lambda) := |y^* x|^{-1}$. Approximations to the eigentriple are indicated by θ for the eigenvalue and u, v for the right and left

eigenvectors. We assume that θ is not equal to an eigenvalue of A, which is equivalent to the assumption that $A-\theta I$ is invertible. To avoid confusion, we remark that, throughout this chapter, the word "right" is used as the opposite of "left" (e.g., right eigenvector versus left eigenvector), and does not have the meaning of "correct".

Since $x \not\perp y$, $(A-\lambda I)|_{y^\perp} : y^\perp \to y^\perp$ is invertible; in particular it has a finite condition number, denoted by $\kappa((A-\lambda I)|_{y^\perp})$. Later in this chapter, we use the following definition.

Definition 2.2.1 (cf. [83, p. 145]) We define the effective condition number of a nonzero matrix C as

$$\kappa_e(C) := \|C\| \cdot \|C^+\| = \sigma_{\max}(C) \,/\, \min_{\sigma_j \neq 0} \sigma_j(C),$$

where C^+ is the pseudoinverse of C, and the $\sigma(C)$s are the singular values of C. ⊘

Next, we give a presentation of "standard" JD, such that two-sided JD will follow as a natural generalization for nonnormal matrices in Section 2.4. The JD method [75] consists of two ingredients. The first part, the well-known *Rayleigh–Ritz* approach, deals with the question: having a k-dimensional search space $\mathcal{U}$ (where one should think of the typical situation $k \ll n$), how do we get an approximate eigenpair (θ, u), where $u \in \mathcal{U}$? Let the columns of U form an orthonormal basis for $\mathcal{U}$, and define the *residual* r by

$$r := Au - \theta u.$$

Imposing the *Ritz–Galerkin condition* on the residual

$$r = Au - \theta u \perp \mathcal{U}, \tag{2.2.1}$$

and writing $u = Uc$ (where c is a k-dimensional vector), we find that (θ, c) should be a solution of the low-dimensional projected eigenproblem

$$U^*AUc = \theta c,$$

so a *Ritz pair* $(\theta, u) = (\theta, Uc)$ is a backtransformed eigenpair of the *projected matrix* U^*AU. In particular, if (θ, u) is a Ritz pair, we have

$$\theta = \theta(u) := \frac{u^*Au}{u^*u} \quad \text{and} \quad r \perp u,$$

that is, θ is the *Rayleigh quotient* of u and the corresponding residual is orthogonal to u.

The second ingredient of JD gives an answer to the question: having an approximate eigenpair (θ, u) to (λ, x), how do we expand the search space $\mathcal{U}$ to get an even better approximation? For this, JD looks for an orthogonal correction $s \perp u$ such that

$$A(u+s) = \lambda(u+s),$$

i.e., such that $u+s$ is a multiple of the eigenvector x. This equation can be rewritten to obtain

$$(A-\theta I)s = -r + (\lambda-\theta)u + (\lambda-\theta)s. \tag{2.2.2}$$

During the process, λ, and hence also the last two terms on the right-hand side, are unknown. We neglect the term $(\lambda - \theta)s$, this may be seen as "throwing away second order terms" (both $\lambda - \theta$ and s will be asymptotically small). This suggests that JD is in fact a Newton method, which is true indeed [76]. (When we choose θ to be the two-sided Rayleigh quotient, then $\lambda - \theta$ is second order, so $(\lambda - \theta)s$ is a third order term, and we may even expect cubic convergence, see Sections 2.3 and 2.4.)

Then we are interested in the projection of (2.2.2) (without the third term on the right-hand side) that maps u (and so the second term on the right-hand side) to 0 and keeps r fixed. Because $r \perp u$, this projection is $I - uu^*$, the orthogonal projection onto the orthogonal complement of u. The result of neglecting the third term of (2.2.2) and projecting the equation is

$$(I - uu^*)(A - \theta I)s = -r.$$

Using

$$(I - uu^*)s = s,$$

we derive the JD correction equation:

$$(I - uu^*)(A - \theta I)(I - uu^*)s = -r \qquad \text{where } s \perp u, \tag{2.2.3}$$

from which we see that the operator $A - \theta I$ is restricted to the orthogonal complement of u. In practice, (2.2.3) is often solved only *approximately* (or *inexactly*), for example by an iterative method, e.g. a few steps of (preconditioned) GMRES. The approximate solution is used to expand the search space $\mathcal{U}$, this is called *subspace acceleration*. JD can therefore be viewed as an *accelerated inexact Newton method* for the eigenvalue problem [76].

However, when we solve (2.2.3) exactly, then we find (see [75])

$$s = -(A - \theta I)^{-1} r + \alpha (A - \theta I)^{-1} u = -u + \alpha (A - \theta I)^{-1} u,$$

where $\alpha = (u^* (A - \theta I)^{-1} u)^{-1}$ is such that $s \perp u$. JD uses s to expand the search space $\mathcal{U}$. Since already $u \in \mathcal{U}$, we get the same subspace expansion using $\widetilde{s} = (A - \theta I)^{-1} u$. Here we recognize a step of RQI, and we conclude that *exact* JD (i.e., JD where we solve the correction equation exactly) can also be seen as accelerated RQI.

In RQI, when the approximations (θ_k, u_k) converge, they converge asymptotically cubically for normal matrices:

Theorem 2.2.2 (Constant of cubic convergence of RQI.) *If A is normal and $u_k \to x$ as $k \to \infty$, then*

$$\lim_{k \to \infty} \|u_{k+1} - x\| \,/\, \|u_k - x\|^3 \leq 1.$$

Proof: See [60, p. 683]. □

The underlying reason for the cubic convergence is the following property of the Rayleigh quotient for normal matrices [60, p. 681]:

$$\theta(u) = \frac{u^* A u}{u^* u} \text{ is stationary} \iff u \text{ is an eigenvector of } A. \tag{2.2.4}$$

(Recall that stationary means that all directional derivatives are zero.) We have already seen that exact JD can be considered as accelerated RQI. Because JD uses subspace acceleration, it will trivially converge in a finite number of steps. Yet, in view of Section 1.3.4, we can speak of the asymptotic convergence of JD. When we neglect the effect of the subspace acceleration on the asymptotic convergence, JD "inherits" the asymptotic convergence of RQI. This explains the expression that "Jacobi–Davidson has asymptotically cubic convergence for normal matrices."

2.3 Two-sided Rayleigh quotient iteration

If A is nonnormal, property (2.2.4) is lost for nonnormal eigenvalues. This implies that RQI (and therefore also exact JD) converges asymptotically at best (only) quadratically to a nonnormal eigenpair (λ, x) [60, p. 688]. But instead of (2.2.4), we have the following property for the *two-sided Rayleigh quotient* $\theta(u, v)$ [60, p. 688]:

$$\theta(u,v) := \frac{v^*Au}{v^*u} \text{ is stationary} \iff \begin{array}{l} u \text{ and } v \text{ are right and left eigenvector} \\ \text{of } A \text{ with eigenvalue } \theta \text{ and } v^*u \neq 0. \end{array} \tag{2.3.1}$$

Because of this property, one may expect cubic convergence for simple eigenvalues of nonnormal matrices when we approximate the left and the right eigenvector simultaneously. For this reason Ostrowski proposes a *two-sided Rayleigh quotient iteration* [55]. In every step of this method, we solve the two equations

$$(A - \theta_k I)u_{k+1} = u_k \quad \text{and} \quad (A - \theta_k I)^* v_{k+1} = v_k, \tag{2.3.2}$$

for u_{k+1} and v_{k+1}, respectively, where $\theta_k = \theta(u_k, v_k)$. This leads to Algorithm 2.3.1.

Input: initial vectors u_1 and v_1 with unit norm, such that $v_1^*u_1 \neq 0$
Output: an eigentriple of A (or failure)

for $k = 1, 2, \ldots$
1. Compute $\theta_k := \theta_k(u_k, v_k) = \frac{v_k^* A u_k}{v_k^* u_k}$
2. If $A - \theta_k I$ is singular, solve $(A - \theta_k I)x = 0$ and $(A - \theta_k I)^* y = 0$ and stop
3. Solve $(A - \theta_k I)u_{k+1} = u_k$ and normalize u_{k+1}
4. Solve $(A - \theta_k I)^* v_{k+1} = v_k$ and normalize v_{k+1}
5. If $v_{k+1}^* u_{k+1} = 0$ then method fails

ALGORITHM 2.3.1: Ostrowski's two-sided Rayleigh quotient iteration [55]

In [60, p. 689] it is shown that when this two-sided RQI converges to a simple eigenvalue, it has locally cubic convergence. However, the following theorem states that the speed of the cubic convergence might be significantly slower in the nonnormal case. Note that by writing

$$u_k = \left(\frac{xy^*}{y^*x}\right) u_k + \left(I - \frac{xy^*}{y^*x}\right) u_k \quad \text{and} \quad v_k = \left(\frac{yx^*}{x^*y}\right) v_k + \left(I - \frac{yx^*}{x^*y}\right) v_k,$$

we see that u_k and v_k can be written in the form

$$u_k = \alpha_k(x + \delta_k d_k) \quad \text{and} \quad v_k = \beta_k(y + \varepsilon_k e_k), \tag{2.3.3}$$

where $\delta_k, \varepsilon_k \geq 0$, $d_k \perp y$, $e_k \perp x$, and u_k, v_k, x, y, d_k, and e_k all have unit norm.

Theorem 2.3.1 (Locally cubic convergence of two-sided RQI.) *Suppose that u_k and v_k converge to x and y, respectively, as $k \to \infty$. Then $\theta_k \to \lambda$, and*

$$\delta_{k+1} \leq \gamma \delta_k^2 \varepsilon_k + h.o.t. \quad \text{and} \quad \varepsilon_{k+1} \leq \gamma \delta_k \varepsilon_k^2 + h.o.t.$$

Here

$$\gamma := \kappa(\lambda)\kappa((A - \lambda I)|_{y^\perp}),$$

and h.o.t. stands for "higher order terms in δ_k and ε_k" (i.e., in the statement above h.o.t. stands for terms of order $\mathcal{O}(\delta_k^i \varepsilon_k^j)$, where $i + j > 3$).

Proof: This is a slight extension of a result in [60, p. 689], where Parlett shows that (in our notation) there exist nonzero $\alpha_{k+1}, \beta_{k+1}$ such that

$$\begin{aligned} u_{k+1} &= \alpha_{k+1}(x + \delta_k(\lambda - \theta_k)\,(A - \theta_k I)^{-1}\, d_k), \\ v_{k+1} &= \beta_{k+1}(y + \varepsilon_k(\lambda - \theta_k)^*(A - \theta_k I)^{-*} e_k), \end{aligned}$$

where

$$\theta_k - \lambda = \delta_k \varepsilon_k \frac{e_k^*(A - \lambda I)d_k}{y^* x + \delta_k \varepsilon_k e_k^* d_k}.$$

Hence

$$|\lambda - \theta_k| = \delta_k \varepsilon_k \kappa(\lambda) |e_k^*(A - \lambda I)d_k| + \text{h.o.t.} \tag{2.3.4}$$

Since $(A - \lambda I)^{-1}$ exists on $y^\perp$, and $(A - \lambda I)^{-*}$ exists on $x^\perp$, we have

$$\begin{aligned} \|\,(A - \theta_k I)^{-1}\, d_k\| &\leq \|\left((A - \lambda I)|_{y^\perp}\right)^{-1}\| + \text{h.o.t.}, \\ \|(A - \theta_k I)^{-*} e_k\| &\leq \|((A - \lambda I)|_{x^\perp})^{-*}\| + \text{h.o.t.} \end{aligned}$$

We can conclude that

$$\begin{aligned} \delta_{k+1} &\leq \delta_k^2 \varepsilon_k \kappa(\lambda)\kappa((A - \lambda I)|_{y^\perp}) + \text{h.o.t.}, \\ \varepsilon_{k+1} &\leq \delta_k \varepsilon_k^2 \kappa(\lambda)\kappa((A - \lambda I)^*|_{x^\perp}) + \text{h.o.t.} \end{aligned}$$

The proof is completed by the observation

$$\begin{aligned} \kappa((A - \lambda I)|_{y^\perp}) &= \kappa_e\left(\left(I - \frac{xy^*}{y^*x}\right)(A - \lambda I)\left(I - \frac{xy^*}{y^*x}\right)\right) \\ &= \kappa_e\left(\left(I - \frac{yx^*}{x^*y}\right)(A - \lambda I)^*\left(I - \frac{yx^*}{x^*y}\right)\right) \\ &= \kappa((A - \lambda I)^*|_{x^\perp}). \end{aligned}$$

□

Comparing Theorems 2.2.2 and 2.3.1, one may observe two differences. First, Theorem 2.2.2 can also be expressed in terms of the angle $\angle(u_k, x)$ (see [61, Theorem 4.7.1]), but in the nonnormal case this is not obvious. Second, because of the possibly large constant of Theorem 2.3.1, the cubic convergence may have less significance in practice.

2.4 Two-sided Jacobi–Davidson

Inspired by two-sided RQI, we design a *two-sided JD method*. We work with two search spaces, $\mathcal{U}$ for the right and $\mathcal{V}$ for the left eigenvector. Suppose that we have k-dimensional search spaces $\mathcal{U}$ and $\mathcal{V}$, and approximations $u \in \mathcal{U}$ and $v \in \mathcal{V}$ to the right and left eigenvectors, $u \not\perp v$. We now would like to take

$$\theta = \theta(u, v) = \frac{v^*Au}{v^*u}$$

as approximation to the eigenvalue. Note that this holds if and only if $(A - \theta I)u \perp v$ and $(A - \theta I)^*v \perp u$. This suggests the imposition of *Petrov–Galerkin conditions* on the *right residual* r_u and *left residual* r_v to determine approximate eigenvectors u and v:

$$r_u := (A - \theta I)u \perp \mathcal{V} \quad \text{and} \quad r_v := (A - \theta I)^*v \perp \mathcal{U}.$$

Now write $u = Uc$ and $v = Vd$, where the columns of U and V form bases for $\mathcal{U}$ and $\mathcal{V}$ (not necessarily orthogonal, see Sections 2.4.1 and 2.4.2), and c and d are k-dimensional vectors. We see that the desired c and d are the right and left eigenvectors corresponding to the eigenvalue θ of the projected (generalized) eigensystem

$$V^*AUc = \theta\, V^*Uc \quad \text{and} \quad U^*AVd = \overline{\theta}\, U^*Vd. \tag{2.4.1}$$

To expand the search spaces $\mathcal{U}$ and $\mathcal{V}$, the two-sided JD method looks for corrections s and t (not necessarily orthogonal, see Sections 2.4.1 and 2.4.2) such that

$$A(u + s) = \lambda(u + s) \quad \text{and} \quad A^*(v + t) = \overline{\lambda}(v + t).$$

For the *right correction equation* this means (cf. (2.2.2))

$$(A - \theta I)s = -r_u + (\lambda - \theta)u + (\lambda - \theta)s. \tag{2.4.2}$$

As in the previous section, we consider the projection of this equation that maps u to 0 and fixes r_u. In this situation $r_u \perp v$, so the sought (oblique) projector is given by $P = I - \frac{uv^*}{v^*u}$. P is an approximation to the spectral projector, just as $I - uu^*$ is in the normal case (see Section 2.2). When we neglect $(\lambda - \theta)s$, which is now of third order (see (2.3.4)), and project (2.4.2), this yields

$$\left(I - \frac{uv^*}{v^*u}\right)(A - \theta I)s = -r_u. \tag{2.4.3}$$

In a similar way we get for the *left correction equation*

$$\left(I - \frac{vu^*}{u^*v}\right)(A - \theta I)^*t = -r_v. \tag{2.4.4}$$

We now discuss two variants of the two-sided JD approach: one where the columns of U and V are bi-orthogonal, and one where both U and V have orthogonal columns.

2.4.1 The columns of the search spaces bi-orthogonal

For the first variant of two-sided JD, we want the columns of U and V to be bi-orthogonal, that is, V^*U should be a diagonal matrix. This is a natural idea, because the right eigenvector corresponding to a particular eigenvalue is orthogonal to the left eigenvector corresponding to a different eigenvalue. This choice has the advantage that the projected eigenproblem (2.4.1) is easily transformed into a standard eigenproblem. Since in this variant we look for bi-orthogonal corrections $s \perp v$ and $t \perp u$, the correction equations (2.4.3) and (2.4.4) can be written as

$$\begin{aligned} \left(I - \frac{uv^*}{v^*u}\right)(A - \theta I)\left(I - \frac{uv^*}{v^*u}\right)s &= -r_u \qquad (s \perp v), \\ \left(I - \frac{vu^*}{u^*v}\right)(A - \theta I)^*\left(I - \frac{vu^*}{u^*v}\right)t &= -r_v \qquad (t \perp u). \end{aligned}$$

The operator in the first equation is the conjugate transpose of the operator in the second equation, so these equations may be solved simultaneously by bi-conjugate gradients (BiCG). Note that BiCG tries to solve two equations; but often only one approximate solution is used, the other solution solves a shadow equation and has no practical interest. In this situation we do use both approximate solutions from BiCG; r_v takes the role of the shadow residual. Of course, we can also deal with the correction equations separately; for instance we may try to solve each of them by a few steps of (preconditioned) GMRES, see the numerical experiments. The resulting algorithm for the computation of the eigenvalue with the largest magnitude is shown in Algorithm 2.4.2.

If one is interested in other eigenvalues, one should change the choice in Step 4 of Algorithm 2.4.2 accordingly (possibly using refined or harmonic Ritz vectors). Also remember that $V_k^*U_k$ is a diagonal matrix. In Step 2 of the algorithm, Bi-MGS stands for (repeated) bi-modified Gram–Schmidt, used to make the columns of U_k and V_k bi-orthogonal in a numerically stable way.

Note that if the algorithm terminates, we have in general found only one eigenvector, say the right eigenvector, to the prescribed tolerance. Often we will also have a good approximation to the left eigenvector (see also the numerical experiments), but this is not necessarily the case. In any case, it is not sensible to continue with the algorithm, for we would then perform superfluous calculations for one of the eigenvectors. If we want to have both eigenvectors accurately, then, at the end of Algorithm 2.4.2, it suffices to (reasonably accurately) solve $t \perp v$ from the system

$$(I - vv^*)(A - \theta I)^*(I - vv^*)\,t = -r_v, \tag{2.4.5}$$

where v is the (often good) approximate left eigenvector from Algorithm 2.4.2. Solving one such system will in general be enough, since θ is a very good approximation to λ. Instead of (2.4.5), we may also solve a correction equation with oblique projections, but experiments suggest that (2.4.5) uses less computational effort.

2.4.2 The columns of both search spaces orthogonal

Another obvious idea is to keep the columns of both U and V orthogonal. Because in this variant we look for updates $s \perp u$, $t \perp v$, the two correction equations now take the

Input: a device to compute Ax and A^*x for arbitrary x,
starting vectors u_1 and v_1 ($v_1^*u_1 \neq 0$), and a tolerance ε
Output: an approximation (θ, u, v) to an eigentriple of A
satisfying $\min\{\|(A-\theta I)u\|, \|(A-\theta I)^*v\|\} \leq \varepsilon$

1. $s = u_1,\ t = v_1,\ U_0 = [\,],\ V_0 = [\,]$
   for $k = 1, 2, \ldots$
2. $(U_k, V_k) = \text{Bi-MGS}\ (U_{k-1}, V_{k-1}, s, t)$
3. Compute kth column of $W_k = AU_k$
   Compute kth row and column of $H_k = V_k^* W_k$
4. Compute eigentriples (θ, c, d) of the matrix $(V_k^*U_k)^{-1}(V_k^*AU_k)$
   and select one (e.g., θ with largest magnitude)
5. $u = U_k c/\|U_k c\|,\ v = V_k d/\|V_k d\|,\ (\theta = \frac{v^*Au}{v^*u})$
6. $r_u = (A-\theta I)u = W_k c/\|U_k c\| - \theta u$
   $r_v = (A-\theta I)^*v$
7. Stop if $\min\{\|r_u\|, \|r_v\|\} \leq \varepsilon$ (and compute second vector at will)
8. Solve (approximately) $s \perp v$, $t \perp u$ from
   $$\left(I - \tfrac{uv^*}{v^*u}\right)(A-\theta I)\left(I - \tfrac{uv^*}{v^*u}\right)s = -r_u$$
   $$\left(I - \tfrac{vu^*}{u^*v}\right)(A-\theta I)^*\left(I - \tfrac{vu^*}{u^*v}\right)t = -r_v$$

ALGORITHM 2.4.2: Bi-orthogonal two-sided Jacobi–Davidson

form

$$\begin{cases} \left(I - \frac{uv^*}{v^*u}\right)(A-\theta I)\,(I-uu^*)\,s &= -r_u \qquad (s \perp u), \\ \left(I - \frac{vu^*}{u^*v}\right)(A-\theta I)^*\,(I-vv^*)\,t &= -r_v \qquad (t \perp v). \end{cases} \tag{2.4.6}$$

This leads to Algorithm 2.4.3 for the computation of the eigenvalue with the largest magnitude. In Step 2 of the algorithm, MGS stands for modified Gram–Schmidt, used to make the columns of U_k and V_k orthogonal. A problem in this variant is that the operator in the first equation in (2.4.6) maps $u^\perp$ onto $v^\perp$, while the operator in the second equation maps $v^\perp$ onto $u^\perp$. As also observed in [73, Section 3.3], it is unnatural to repeat such an operator, so it seems unattractive to solve the equations in (2.4.6) by a Krylov solver.

As has been noted in [73], we can fix this by working with a preconditioner M for $A - \theta I$. We know [73] that the inverse of the projected preconditioner

$$\left(I - \frac{uv^*}{v^*u}\right) M\,(I-uu^*) : u^\perp \to v^\perp$$

is given by

$$\left(I - \frac{M^{-1}uu^*}{u^*M^{-1}u}\right) M^{-1} \left(I - \frac{uv^*}{v^*u}\right) : v^\perp \to u^\perp.$$

For the first equation, this operator maps $v^\perp$ back to $u^\perp$, while

$$\left(I - \frac{M^{-*}vv^*}{v^*M^{-*}v}\right) M^{-*} \left(I - \frac{vu^*}{u^*v}\right)$$

for the second equation maps $u^\perp$ back to $v^\perp$. Let us study the simplest case, $M = I$, for a moment. Then we get

$$\begin{cases} (I - uu^*)(A - \theta I)(I - uu^*)\, s &= -(I - uu^*)\, r_u \\ (I - vv^*)(A - \theta I)^*(I - vv^*)\, t &= -(I - vv^*)\, r_v. \end{cases} \tag{2.4.7}$$

We recognize these equations as the correction equations of standard JD (2.2.3) applied to A and A^*; so for this special case we get a version of two-sided JD where the correction equations are of the same form as in standard JD. Of course, preconditioning may also be useful to speed up the convergence of the inner iteration.

Input: a device to compute Ax and A^*x for arbitrary x,
starting vectors u_1 and v_1 ($v_1^* u_1 \neq 0$), and a tolerance ε
Output: an approximation (θ, u, v) to an eigentriple of A
satisfying $\min\{\|(A-\theta I)u\|, \|(A-\theta I)^* v\|\} \leq \varepsilon$

1. $s = u_1, t = v_1,\ U_0 = [\,],\ V_0 = [\,]$
   for $k = 1, 2, \ldots$
2. $U_k = \text{MGS}\ (U_{k-1}, s)$
   $V_k = \text{MGS}\ (V_{k-1}, t)$
3. Compute kth column of $W_k = AU_k$
   Compute kth row and column of $H = V_k^* W_k$
4. Compute eigentriples (θ, c, d) of the pencil $(V_k^* AU_k,\ V_k^* U_k)$
   and select one (e.g., θ with largest magnitude)
5. $u = U_k c,\ v = V_k d,\ (\theta = \frac{v^* Au}{v^* u})$
6. $r_u = (A - \theta I)u = W_k c - \theta u$
   $r_v = (A - \theta I)^* v$
7. Stop if $\min\{\|r_u\|, \|r_v\|\} \leq \varepsilon$ (and compute second vector at will)
8. Solve (approximately) $s \perp u,\ t \perp v$ from
   $$\left(I - \tfrac{uv^*}{v^* u}\right)(A - \theta I)(I - uu^*)\, s = -r_u$$
   $$\left(I - \tfrac{vu^*}{u^* v}\right)(A - \theta I)^*(I - vv^*)\, t = -r_v$$

ALGORITHM 2.4.3: Orthogonal two-sided Jacobi–Davidson

The following theorem states that exact two-sided JD, like two-sided RQI, has locally cubic convergence.

Theorem 2.4.1 *If the two correction equations* (2.4.3) *and* (2.4.4) *are solved exactly, both the bi-orthogonal and the orthogonal variant of the two-sided JD process converge asymptotically cubically to an eigenvalue, if that eigenvalue is simple.*

Proof: The solution to (2.4.3) is

$$s = -u + \zeta\,(A - \theta I)^{-1} u.$$

In the bi-orthogonal variant $s \perp v$, so $\zeta = v^* u / (v^*\,(A - \theta I)^{-1} u)$. In the orthogonal variant $s \perp u$, then $\zeta = (u^*\,(A - \theta I)^{-1} u)^{-1}$. We recognize the updated vector $u + s$ as a multiple of the one from two-sided RQI. For the left correction equation we have a similar expression. Now apply Theorem 2.3.1. □

2.5 Inexact two-sided RQI and Jacobi–Davidson

In Section 2.1 we have already mentioned that JD and RQI are in practice often very expensive when we solve the linear systems, occuring in the methods ((2.3.2), respectively (2.4.3) and (2.4.4)), accurately. In this section, we therefore consider inexact variants. In Sections 2.5.1 and 2.5.2, we investigate two-sided RQI and two-sided JD when the linear systems are solved to a certain precision (minimal residual approach). In Section 2.5.3, a relation between two-sided RQI and two-sided JD is established, when the linear systems are solved by a number of BiCG-steps (bi-orthogonal residual approach).

2.5.1 Inexact two-sided RQI

In [80] and [89], the authors study inexact RQI for Hermitian matrices. They show that the asymptotic convergence rate under certain assumptions is quadratic. Here we give a generalization for inexact two-sided RQI.

Consider the situation where we solve the two equations (2.3.2) of the two-sided RQI method inexactly, by which we mean that we are contented with u_{k+1}, v_{k+1} satisfying

$$\|(A-\theta I)u_{k+1} - u_k\| \leq \xi_1 < 1 \quad \text{and} \quad \|(A-\theta I)^* v_{k+1} - v_k\| \leq \xi_2 < 1. \tag{2.5.1}$$

Note that if we have nonsingular preconditioners $M_1 \approx A - \theta I$ and $M_2 \approx (A-\theta I)^*$, such that

$$\|(A-\theta I)M_1^{-1} - I\| \leq \xi_1 \quad \text{and} \quad \|(A-\theta I)^* M_2^{-1} - I\| \leq \xi_2,$$

then only one action with each preconditioner (that is, take $u_{k+1} := M_1^{-1}u_k$ and $v_{k+1} := M_2^{-1}v_k$) is enough to satisfy (2.5.1). However, since $A - \theta I$ is almost singular if $\theta \approx \lambda$, it is not a realistic assumption to have such preconditioners at our disposal.

To study the convergence rate of inexact two-sided RQI, the following lemma is useful.

Lemma 2.5.1 *Let $v = \beta(y + \varepsilon e)$, $e \perp x$ (cf. (2.3.3)). The following statements are true:*

(a) $\left\| I - \frac{xy^*}{y^*x} \right\| = \left\| \frac{xy^*}{y^*x} \right\| = \kappa(\lambda)$;

(b) $\left\| (I - \frac{xy^*}{y^*x})|_{v^\perp} \right\| \leq \sqrt{1 + \varepsilon^2 \kappa(\lambda)^2}$;

(c) $\left\| \frac{xy^*}{y^*x}|_{v^\perp} \right\| \leq \varepsilon\kappa(\lambda)$.

Proof: Define $Q := I - \frac{xy^*}{y^*x}$. By examining the eigenpairs of Q^*Q we see that all singular values of Q restricted to the space $\mathrm{span}\{x, y\}^\perp$ are equal to one. Likewise, the singular values of $I - Q$ restricted to the space $y^\perp$ are zero. Therefore, one may check that, up to a normalizing constant, $\mathrm{argmax}_{z \neq 0} \frac{\|Qz\|}{\|z\|} = y - (x^*y)x$, and $\mathrm{argmax}_{z \neq 0} \frac{\|(I-Q)z\|}{\|z\|} = y$, both with maximum $|y^*x|^{-1}$. This proves (a). (In fact, this is a special case of the result that for all projections $0 \neq P \neq I$, we have $\|P\| = \|I - P\|$ [45].) Using $y = \beta^{-1}v - \varepsilon e$, we get that $I - \frac{xy^*}{y^*x} = I + \varepsilon \frac{xe^*}{y^*x}$ on the subspace $v^\perp$. Similar to the proof of (a), it is only of interest to consider the singular values of this operator on the subspace $\mathrm{span}\{x, e\}^\perp$. Now $\|Qe\|^2 = \|e + \varepsilon (y^*x)^{-1} x\|^2 = 1 + \varepsilon^2\kappa(\lambda)^2$, because $e \perp x$. Because in general $e \not\perp v$, (b) follows. Finally, (c) can be proved by noting that $\frac{xy^*}{y^*x} = -\varepsilon \frac{xe^*}{y^*x}$ on $v^\perp$. □

From the proof of part (a), we see that, when $\kappa(\lambda)$ is very large, $\operatorname{argmax}(I - \frac{xy^*}{y^*x}) \approx \operatorname{argmax}(\frac{xy^*}{y^*x})$. This implies that when the right eigenvector x and the corresponding left eigenvector y are nearly orthogonal, the decomposition $v = \eta_1 x + \eta_2 d, d \perp y$, has (almost equally) large η_1 and η_2 components. We are now ready to state the following result. As in Theorem 2.2.2, h.o.t. stands for "higher order terms in δ_k and ε_k".

Theorem 2.5.2 (Locally quadratic convergence of inexact two-sided RQI, generalization of [80, Corollary 4.3] and [89, Proposition 2.2].) *Suppose that* $\max\{\xi_1, \xi_2\} \cdot \kappa(\lambda) < 1$. *For one step of inexact two-sided RQI, where the equations are solved inexactly according to* (2.5.1), *we have (using the notation in* (2.3.3)*)*

$$\delta_{k+1} \leq \gamma_1 \delta_k \varepsilon_k + h.o.t. \quad \textit{and} \quad \varepsilon_{k+1} \leq \gamma_2 \delta_k \varepsilon_k + h.o.t.$$

Here

$$\gamma_i = \kappa(\lambda)\kappa((A - \lambda I)|_{y^\perp}) \frac{\xi_i \kappa(\lambda)}{1 - \xi_i \kappa(\lambda)} \qquad (i = 1, 2).$$

Proof: From the first equation of (2.5.1) we know that there exists a $\widetilde{\xi}$, $0 \leq \widetilde{\xi} \leq \xi_1$, and a unit vector f such that

$$(A - \theta_k I)u_{k+1} = u_k + \widetilde{\xi} f.$$

Decomposing f in an x-component and a component orthogonal to y, we get using Lemma 2.5.1(a) that

$$(A - \theta_k I)u_{k+1} = \widetilde{\alpha} x + \widetilde{\delta}\widetilde{d},$$

where $\widetilde{d} \perp y$, $|\widetilde{\alpha}| \geq |\alpha_k| - \xi_1 \kappa(\lambda)$, and $\widetilde{\delta} \leq |\alpha_k|\delta_k + \xi_1 \kappa(\lambda)$. Moreover, we have the estimates

$$|\alpha_k| = 1 + \text{h.o.t.}, \quad \text{and} \quad |\beta_k| = 1 + \text{h.o.t.} \tag{2.5.2}$$

The value of γ_1 now follows, analogous to the proof of Theorem 2.3.1, from bounding $|\widetilde{\delta}/\widetilde{\alpha}|$, and γ_2 is derived in a similar manner. □

When we have preconditioners at our disposal, we may also try to solve the (left) preconditioned equations to a certain precision, e.g.

$$\|M_1^{-1}((A - \theta I)u_{k+1} - u_k)\| \leq \xi_1 < 1,$$

Just as for Theorem 2.5.2, one can prove that this yields locally quadratic convergence, now with constants

$$\gamma_i = \kappa(\lambda)\kappa((A - \lambda I)|_{y^\perp}) \frac{\xi_i \kappa(\lambda)\|M_i\|}{1 - \xi_i \kappa(\lambda)\|M_i\|} \qquad (i = 1, 2).$$

As $\theta \approx \lambda$, the condition number of the matrix $A - \theta_k I$ increases. Therefore, it may get more and more expensive to solve (2.5.1) to a certain tolerance. This provides a motivation to study inexact two-sided JD.

2.5.2 Inexact two-sided Jacobi–Davidson

In [89], the author studies inexact JD for Hermitian matrices. He shows that the asymptotic convergence rate is linear under certain assumptions. Here we give a generalization for two-sided JD.

Consider the situation where we solve the two equations (2.4.3) and (2.4.4) of the two-sided JD method inexactly, by which we mean that we are satisfied with $\widetilde{s} \perp v$ and $\widetilde{t} \perp u$ (bi-orthogonal variant) or $\widetilde{s} \perp u$ and $\widetilde{t} \perp v$ (orthogonal variant) where

$$\left\| \left(I - \frac{uv^*}{v^*u} \right) (A - \theta I)\widetilde{s} + r_u \right\| \leq \xi_1 \|r_u\| \tag{2.5.3}$$

and

$$\left\| \left(I - \frac{vu^*}{u^*v} \right) (A - \theta I)^*\widetilde{t} + r_v \right\| \leq \xi_2 \|r_v\|, \tag{2.5.4}$$

for some $0 < \xi_1, \xi_2 < 1$. The next theorem states that the resulting local convergence is linear.

Theorem 2.5.3 (Locally linear convergence of inexact two-sided JD, generalization of [89, Theorem 4.1]) *For one step of inexact bi-orthogonal two-sided JD, when the equations are solved inexactly according to* (2.5.3) *and* (2.5.4)*, we have (using the notation in* (2.3.3)*)*

$$\delta_{k+1} \leq \gamma_1 \delta_k + h.o.t. \quad \textit{and} \quad \varepsilon_{k+1} \leq \gamma_2 \varepsilon_k + h.o.t.$$

Here

$$\gamma_i = \xi_i \kappa((A - \lambda I)|_{y^\perp}) \qquad (i = 1, 2).$$

The orthogonal variant of two-sided JD has locally linear convergence as well.

Proof: For clarity, we leave out the index k. Write $P = I - \frac{uv^*}{v^*u}$. Let us first consider the bi-orthogonal variant of Section 2.4.1, where $\widetilde{s} \perp v$. From (2.5.3) we know that there exists a $\widetilde{\xi}$, $0 \leq \widetilde{\xi} \leq \xi_1$, and a unit vector $f \perp v$ such that

$$P(A - \theta I)P\widetilde{s} = -r_u + \widetilde{\xi}\,\|r_u\| f. \tag{2.5.5}$$

From (2.4.2) we can see that the "real" update $s \perp v$ satisfies

$$P(A - \lambda I)Ps = -r_u,$$

hence

$$P(A - \theta I)Ps = -r_u + (\lambda - \theta)s. \tag{2.5.6}$$

Both $u + s$ ($s \perp v$) and $u - \alpha\delta d$ ($d \perp y$) are multiples of the eigenvector x, and (2.5.2) and Lemma 2.5.1(b),(c) give that $\|s\| = \delta + \text{h.o.t.}$ Subtracting (2.5.6) from (2.5.5) gives

$$P(A - \theta I)P(\widetilde{s} - s) = \widetilde{\xi}\,\|r_u\| f - (\lambda - \theta)s.$$

The operator $P(A-\theta I)P$ is a bijection from $v^\perp$ to $v^\perp$ and

$$\| \left((P(A-\theta I)P)|_{v^\perp}\right)^{-1} \| = \| \left((A-\lambda I)|_{y^\perp}\right)^{-1} \| + \text{h.o.t.}$$

For the norm of the residual we have by (2.3.4)

$$\begin{aligned} \|r_u\| &= \|\alpha((\lambda-\theta)x + \delta(A-\theta I)d)\| \\ &= \delta \|(A-\theta I)d\| + \text{h.o.t.} \\ &\leq \delta \|(A-\theta I)|_{y^\perp}\| + \text{h.o.t.} \\ &= \delta \|(A-\lambda I)|_{y^\perp}\| + \text{h.o.t.} \end{aligned}$$

So

$$\|\widetilde{s} - s\| \leq \delta \xi_1 \kappa((A-\lambda I)|_{y^\perp}) + \text{h.o.t.}$$

The term $(\widetilde{s}-s) \perp v$ represents the error in the updated vector $u+\widetilde{s}$. Again using Lemma 2.5.1(b),(c), we get $\delta_{k+1} = \|\widetilde{s}-s\| + \text{h.o.t.}$ This proves the statement for the bi-orthogonal variant. Now consider the orthogonal variant of Section 2.4.2. The essential difference is that $\widetilde{s} \perp u$. Along the same lines, it can be shown that

$$\|r_u\| \leq \|(A-\lambda I)|_{x^\perp}\| \|s\| + \text{h.o.t.}$$

In the same way as in the proof for the bi-orthogonal case, we get

$$\|\widetilde{s} - s\| \leq \xi_1 \kappa((A-\lambda I)|_{x^\perp}) \|s\| + \mathcal{O}(\|s\|^2),$$

where $(A-\lambda I)|_{x^\perp}$ is interpreted as operator from $x^\perp$ to $y^\perp$. This estimate means locally linear convergence. □

Comparing inexact two-sided RQI with inexact two-sided JD, we remark that it is by no means possible to conclude from Theorems 2.5.2 and 2.5.3 that "inexact two-sided RQI is faster that inexact two-sided JD". Firstly, the theorems only make a statement about the local, not the global, rate of convergence. In fact, it can happen that inexact two-sided RQI does not converge at all (see the numerical experiments), while inexact two-sided JD trivially converges in a finite number of steps. Secondly and more importantly, the theorems do not tell how much effort it takes to solve the equations in question ((2.5.1) versus (2.5.3) and (2.5.4)) to a certain precision. In the proof of the previous theorem we have seen that the effective condition number of $\left(I - \frac{uv^*}{v^*u}\right)(A-\theta I)\left(I - \frac{uv^*}{v^*u}\right)$ approaches $\kappa((A-\lambda I)|_{y^\perp})$ as $\theta \to \lambda$, $u \to x$, $v \to y$, while the condition number of $A-\theta I$ is unbounded as $\theta \to \lambda$. Therefore, it may be much more difficult to solve (2.5.1). Thirdly, in the next section we show that the solutions to the linear systems are the same if the systems are solved by unpreconditioned BiCG with a fixed number of steps.

2.5.3 Relation between inexact two-sided JD and inexact two-sided RQI

We have already seen that two-sided JD is equivalent to accelerated two-sided RQI if all linear systems ((2.3.2) and (2.4.3), (2.4.4)) are solved exactly. In [72], the somewhat

surprizing result is proved that for Hermitian matrices, standard JD is equivalent to accelerated RQI when all linear equations are solved by a certain number of steps of conjugate gradients. We generalize this result, and show that two-sided JD for nonnormal matrices is also equivalent to accelerated two-sided RQI, if the linear systems are solved by a certain number of steps of BiCG. In the next lemma, C plays the role of $A - \theta I$.

Lemma 2.5.4 (Generalization of [81, Lemma 4.1].) *Let $P = I - \frac{uv^*}{v^*u}$ and $r = Cu$. Then for all $m \geq 1$,*

$$\text{span}\{u, r, (PCP)r, \ldots, (PCP)^{m-1}r\} = \text{span}\{u, r, Cr, \ldots, C^{m-1}r\}.$$

Proof: Let $\mathcal{K}_m = \text{span}\{u,\ r,\ Cr,\ \ldots,\ C^{m-1}r\}$ and $\mathcal{L}_m = \text{span}\{u,\ r,\ (PCP)r,\ \ldots,\ (PCP)^{m-1}r\}$. The proof is by induction. For $m = 1$, the claim is evidently true. Now assume that $\mathcal{K}_j = \mathcal{L}_j$ for all $j < m$. If $a \in \mathcal{L}_m$, then there exist $b \in \mathcal{L}_1 = \mathcal{K}_1$, and $c \in \mathcal{L}_{m-1} = \mathcal{K}_{m-1}$ such that $a = b + (PCP)c$. Writing out the projection P, we get

$$\begin{aligned} a &= b + \left(I - \frac{uv^*}{v^*u}\right) C \left(I - \frac{uv^*}{v^*u}\right) c \\ &= b + Cc - \frac{v^*c}{v^*u}Cu + \left(\frac{(v^*Cu)(v^*c)}{(v^*u)^2} - \frac{v^*Cc}{v^*u}\right) u. \end{aligned}$$

Now $Cc \in \mathcal{K}_m$, and all other terms are in $\mathcal{K}_1$, so $a \in \mathcal{K}_m$ and $\mathcal{L}_m \subset \mathcal{K}_m$.

If $\mathcal{L}_m$ is of full rank, then the lemma is proved. Otherwise, let j be the largest index such that $\mathcal{L}_j$ is full rank. Then $\mathcal{L}_{j+1} = \mathcal{L}_j = \mathcal{K}_j$. Now let $c \in \mathcal{K}_j$. Then we deduce that also $PCPc \in \mathcal{K}_j$. From an equation similar to the one displayed above, we see that $Cc \in \mathcal{K}_j$, so $\mathcal{K}_{j+1} = \mathcal{K}_j$. By induction we have $\mathcal{L}_m = \mathcal{L}_j = \mathcal{K}_j = \mathcal{K}_m$ for all $m \geq j$. □

Proposition 2.5.5 (Generalization of [72, Proposition 3.2].) *Let u and v be approximate eigenvectors. Let $\widetilde{s}_m$, $\widetilde{t}_m$ be the approximate solutions to the right and left JD correction equation ((2.4.3) and (2.4.4)) respectively, obtained by m steps of the BiCG method, without suffering from a breakdown. Let $\widetilde{u}_{m+1}$ and $\widetilde{v}_{m+1}$ be the approximate solutions to the two-sided RQI equations (2.3.2), obtained by $m+1$ steps of BiCG. Then there exist μ_1, μ_2 such that*

$$\widetilde{u}_{m+1} = \mu_1(u + \widetilde{s}_m), \quad \text{and} \quad \widetilde{v}_{m+1} = \mu_2(v + \widetilde{t}_m).$$

Proof: Let the columns of W_m and Z_m be bi-orthogonal bases for respectively $\text{span}\{r_u,\ P(A - \theta I)Pr_u,\ \ldots,\ (P(A - \theta I)P)^{m-1}r_u\}$ and $\text{span}\{r_v,\ P^*(A - \theta I)^*P^*r_v,\ \ldots,\ (P^*(A - \theta I)^*P^*)^{m-1}r_v\}$, respectively. Apply BiCG to the JD correction equations; then $\widetilde{s}_m$ and $\widetilde{t}_m$ are of the form $\widetilde{s}_m = W_m w$, $\widetilde{t}_m = Z_m z$, where w, z are solutions of

$$Z_m^*P(A - \theta I)PW_m w = -Z_m^*r_u \quad \text{and} \quad W_m^*P^*(A - \theta I)^*P^*Z_m z = -W_m^*r_v.$$

Now note that $W_m^*v = 0$ and $Z_m^*u = 0$, so $PW_m = W_m$ and $P^*Z_m = Z_m$. Hence w, z solve

$$Z_m^*(A-\theta I)W_m w = -Z_m^* r_u \quad \text{and} \quad W_m^*(A-\theta I)^* Z_m z = -W_m^* r_v.$$

On the other hand, according to Lemma 2.5.4, the columns of $[\, u \;\; W_m \,]$ and $[\, v \;\; Z_m \,]$ are bi-orthogonal bases for $\mathrm{span}\{u, (A-\theta I)u, \ldots, (A-\theta I)^{m-1}u\}$ and $\mathrm{span}\{v, (A-\theta I)^*v, \ldots, ((A-\theta I)^*)^{m-1}v\}$, respectively. Hence, BiCG applied to the two-sided RQI equations gives approximations $\widetilde{u}_{m+1}$ and $\widetilde{v}_{m+1}$ of the form $\widetilde{u}_{m+1} = W_m p + \mu_1 u$, $\widetilde{v}_{m+1} = Z_m q + \mu_2 v$, where p, q, μ_1, μ_2 are determined by the Petrov–Galerkin conditions

$$\widehat{A}\begin{bmatrix} \mu_1 \\ p \end{bmatrix} = \begin{bmatrix} v^*u \\ Z_m^*u \end{bmatrix} \quad \text{and} \quad \widehat{A}^*\begin{bmatrix} \mu_2 \\ q \end{bmatrix} = \begin{bmatrix} u^*v \\ W_m^*v \end{bmatrix}, \tag{2.5.7}$$

where

$$\widehat{A} := \begin{bmatrix} v^*(A-\theta I)u & v^*(A-\theta I)W_m \\ Z_m^*(A-\theta I)u & Z_m^*(A-\theta I)W_m \end{bmatrix}.$$

The terms Z_m^*u and W_m^*v in (2.5.7) vanish. From the (twice) last $n-1$ equations in (2.5.7) we get

$$\begin{aligned} Z_m^*(A-\theta I)W_m p &= -\mu_1 Z_m^*(A-\theta I)u, \\ W_m^*(A-\theta I)^* Z_m q &= -\mu_2 W_m^*(A-\theta I)^* v. \end{aligned}$$

Because of the assumption that no breakdown is encountered, $Z_m^*(A-\theta I)W_m$ is invertible and the proposition is proved. □

Commenting on the number of BiCG-steps in the previous proposition, we note that it is natural that two-sided RQI needs one step more ($m+1$ versus m), because two-sided JD already uses a matrix-vector multiplication to compute the residual. Based on this proposition, it is tempting to conclude that two-sided JD and two-sided RQI are equivalent. But the proposition only gives a statement for the situation when the linear systems ((2.3.2) and (2.4.3), (2.4.4)) are solved by unpreconditioned BiCG; and even then JD uses subspace acceleration, and RQI does not. Preconditioning can be included in JD and RQI in such a way that Proposition 2.5.5 still holds for the preconditioned methods. However, this preconditioning for JD will not be the one described in [73, 75] and seems to be less effective.

2.6 Alternating Jacobi–Davidson

Theoretically, RQI does not need to converge globally. Parlett [60] proposes a different generalization of RQI to ensure global convergence, which he calls *alternating Rayleigh quotient iteration*; see Algorithm 2.6.4.

This method is somewhat counter-intuitive, because $(A-\theta I)^{-1}$ and $(A-\theta I)^{-*}$ are used alternatingly on the iterates. For fixed θ, every two steps of the algorithm result in one action with $((A-\theta I)^*(A-\theta I))^{-1}$. This method could therefore be interpreted as an attempt to find the smallest singular value and corresponding singular vectors of $A-\theta I$. As such, it can be considered as a method for the singular value problem, rather

Input: initial vector u_1 with unit norm
Output: an eigenvalue of A with its right and left eigenvector

for $k = 1, 3, \ldots$
1. Compute $\theta_k = \theta(u_k) = u_k^* A u_k$
2. Solve $(A - \theta_k I)u_{k+1} = u_k$ and normalize u_{k+1}
3. Compute $\theta_{k+1} = \theta(u_{k+1}) = u_{k+1}^* A u_{k+1}$
4. Solve $(A - \theta_{k+1} I)^* u_{k+2} = u_{k+1}$ and normalize u_{k+2}

If $A - \theta_k I$ or $(A - \theta_{k+1} I)^*$ happen to be singular, solve the eigenvectors.

ALGORITHM 2.6.4: Parlett's alternating Rayleigh quotient iteration [60]

than one for the eigenvalue problem. But because a matrix has a zero eigenvalue if and only if it has a zero singular value, the method can asymptotically (that is, for $\theta \approx \lambda$) also be regarded as an eigenvalue method. Parlett [60, p. 692] shows that the (only) advantage of this process is that it converges for all starting vectors, while it has a big drawback: the asymptotic convergence is in general only linear with factor close to one $(1 - \kappa(\lambda)^{-2})$ when applied to nonnormal matrices.

Alternating RQI gives us inspiration for a new JD variant, which we call *alternating Jacobi–Davidson*. The idea is to accelerate Parlett's process, building up one (orthogonal) search space for *both* the left *and* the right eigenvector. Every odd step focuses on approximating the right eigenvector, every even step on approximating the left eigenvector, see Algorithm 2.6.5.

Because of the subspace acceleration, the convergence behavior of alternating JD is much better than that of alternating RQI. For nonnormal matrices, the odd or even steps alone garantee us quadratic convergence (when the correction equations are solved exactly). For normal matrices, one can check that alternating JD does exactly the same as standard JD, so with the same amount of work we get cubic convergence. Our hope is that alternating JD will have fast convergence for (slightly) nonnormal matrices with only a modest amount of extra work. Numerical experiments show that alternating JD can be faster that standard JD (see Section 2.9).

2.7 Extensions

In this section we extend the two-sided methods to the generalized and polynomial eigenproblem, and we discuss the application of the methods to the complex symmetric eigenproblem, which can be seen as a special case.

2.7.1 The generalized eigenproblem

Two-sided and alternating JD can easily be generalized. Let us examine the adaptations to apply two-sided JD to the generalized eigenproblem $Ax = \lambda Bx$. The Galerkin conditions $(A - \theta B)u \perp \mathcal{V}$ and $(A - \theta B)^* v \perp \mathcal{U}$ lead to the *two-sided Rayleigh quotient for*

Input: a device to compute Ax and A^*x for arbitrary x,
a starting vector y_1, and a tolerance ε.
Output: an approximation (θ, u, v) to the largest eigenvalue of A and its
left and right eigenvector satisfying $\min\{\|(A-\theta I)u\|, \|(A-\theta I)^*v\|\} \leq \varepsilon$.

1. $s = y_1$, $Y_0 = [\,]$
   for $k = 1, 2, \ldots$
2. $Y_k = \text{MGS}\ (Y_{k-1}, s)$
3. Compute kth column of $Z_k = AY_k$
   Compute kth row and column of $M_k = Y_k^* Z_k$
4. Compute eigenpairs (and select one)
   (θ, c) of $M_k = Y_k^* A Y_k$ (**k even**)
   $(\overline{\theta}, c)$ of $M_k^* = Y_k^* A^* Y_k$ (**k odd**)
5. $y = Y_k c$
6. $r = (A-\theta I)y = Z_k x - \theta y$ (**k even**)
   $r = (A-\theta I)^* y$ (**k odd**)
7. Stop if $\|r\| \leq \varepsilon$ (and compute second vector at will)
8. Solve (approximately) $s \perp u$ from
   $(I-uu^*)(A-\theta I)(I-uu^*)s = -r$ (**k even**)
   $(I-uu^*)(A-\theta I)^*(I-uu^*)s = -r$ (**k odd**)

ALGORITHM 2.6.5: Alternating Jacobi–Davidson

the generalized eigenvalue problem

$$\theta = \frac{v^*Au}{v^*Bu},$$

where u and v are the backtransformed right and left eigenvectors of the projected pencil $(V^*AU,\ V^*BU)$. For bi-orthogonal two-sided JD, one possibility for the right correction equation is

$$\left(I - \frac{Buv^*}{v^*Bu}\right)(A-\theta B)\left(I - \frac{u(B^*v)^*}{(B^*v)^*u}\right)s = -(A-\theta B)u \qquad (s \perp B^*v). \tag{2.7.1}$$

(For other options, see [73].) If we solve this correction equations exactly, then we get

$$s = -u + \zeta\,(A-\theta B)^{-1}\,Bu$$

(ζ such that $s \perp B^*v$), so that exact two-sided JD can also in this case be viewed as accelerated "generalized two-sided RQI" (see e.g. [61, Theorem 15.9.3] for the symmetric case), leading to cubic convergence:

Proposition 2.7.1 *Let B be nonsingular, and let λ be a simple eigenvalue of $B^{-1}A$. Then exact two-sided JD converges locally cubically.*

Proof: Note that $(A-\theta B)^{-1}\,Bu = (B^{-1}A - \theta I)^{-1}\,u$ and apply Theorem 2.3.1. □

Note that we get cubic convergence using $\alpha Au + \beta Bu$ instead of Bu as well, because Au and Bu are asymptotically linear dependent.

2.7.2 The complex symmetric eigenvalue problem

Let us now apply the methods to the complex symmetric eigenvalue problem

$$Ax = \lambda x,$$

where the large and sparse matrix A is *complex symmetric*: $A = A^T \in \mathbb{C}^{n\times n}$. Eigenvalue problems of this type, and of the related generalized complex symmetric eigenvalue problem

$$Ax = \lambda Bx, \qquad B \text{ invertible},$$

where both A and B are complex symmetric are becoming of increasing importance in applications, most notably in the field of electro-magnetic simulations.

Notice that complex symmetric matrices are not Hermitian. So, they do not possess the favorable properties of Hermitian matrices. In particular, complex symmetric matrices may have complex eigenvalues, and can be arbitrarily nonnormal. In fact, *every* matrix is similar to a complex symmetric matrix [27, 40], whence it may be arbitrarily difficult to a (standard or generalized) complex symmetric eigenproblem.

Nevertheless, complex symmetric matrices do have special properties. If x is a right eigenvector of A, $Ax = \lambda x$, then it is also a left eigenvector, in the sense that $x^T A = \lambda x^T$. Eigenvectors x, y corresponding to different eigenvalues $\lambda \neq \mu$ are *complex orthogonal*, i.e., they satisfy

$$(x, y)_T := y^T x = 0. \tag{2.7.2}$$

If A is diagonalizable then the diagonalization can be realized by a complex orthogonal matrix Q, $Q^T Q = I$ [40].

We call the (indefinite) bilinear form $(x, y)_T$ in (2.7.2)—somewhat abusively—an *"inner product"*. For brevity, we write $x \perp_T y$ if two vectors x and y are complex orthogonal. A vector x is called *quasi-null* if $(x, x)_T = 0$.

When treating the generalized complex symmetric eigenvalue problem it is natural to use the indefinite bilinear form

$$[x, y]_T := (x, By)_T = y^T Bx. \tag{2.7.3}$$

The matrix $B^{-1}A$ is then complex symmetric with respect to $[x, y]_T$ as A is complex symmetric with respect to $(x, y)_T$. We therefore restrict the discussion to the standard complex symmetric eigenvalue problem.

A number of algorithms have been designed for solving complex symmetric linear systems of equations. In [93], the bi-conjugate gradient algorithm is modified to obtain the complex conjugate gradient algorithm COCG. The idea is to set the initial shadow vector equal to the initial residual. (If one works with the Euclidean inner product, the shadow vector has to be the complex conjugate of the initial residual, see [93].) With regard to the relation among right and left eigenvectors mentioned before this choice of the shadow vector is natural. The same idea is used to adapt the quasi-minimal residual (QMR) algorithm to the complex symmetric case [26]. In COCG and QMR, the same Krylov subspaces are generated. However, the approximate solutions are extracted differently from these subspaces. In [16] an algorithm, CSYM, is introduced that is

closely related to the special form that the singular value decomposition (or Takagi factorization) takes on for complex symmetric matrices [40]. Every complex symmetric matrix is unitarily similar to a complex symmetric tridiagonal matrix. CSYM constructs the three-term recurrence that holds among the columns of the unitary matrix that realizes the similarity transformation. Notice that CSYM is not a Krylov subspace method.

With respect to methods for solving complex symmetric eigenvalue problems, a Lanczos type eigensolver employing the bilinear form (2.7.2) is proposed in [18, Chapter 6]. We apply the two-sided JD method to the complex symmetric eigenvalue problem. We give a short summary of the results, for a more extensive discussion, as well as some experiments for complex symmetric generalized eigenvalue problems, see [1]. In contrast to the complex symmetric methods mentioned before, our Jacobi–Davidson algorithm for the complex symmetric eigenproblem, which we denote by JDCS, can be transcribed quite easily into a solver for the generalized complex symmetric eigenvalue problem.

Assume that λ is a simple eigenvalue, then it has a finite condition $\kappa(\lambda)$. Because

$$\infty > \kappa(\lambda) = |x^T x|^{-1} = |(x,x)_T|^{-1},$$

an eigenvector corresponding to a simple eigenvalue is not quasi-null whence it can be "normalized" such that $(x,x)_T = 1$ [31, p. 323].

Given an approximate eigenvector u with Euclidean norm one, the corresponding eigenvalue is usually approximated by the Rayleigh quotient $\theta(u)$ (see Section 2.2). Alternatively, with regard to the "inner product" (2.7.2), we can also define the Rayleigh quotient by

$$\rho = \rho(u) := \frac{u^T A u}{u^T u}.$$

One may check that for complex symmetric A, the latter definition has the desirable property (cf. (2.3.1))

$$\rho(u) \text{ is stationary} \iff u \text{ is an eigenvector of } A, \text{ and } u \text{ not quasi-null.} \tag{2.7.4}$$

By writing

$$u = \left(\frac{xx^T}{x^T x}\right) u + \left(I - \frac{xx^T}{x^T x}\right) u,$$

we see that u can be written in the form (cf. (2.3.3))

$$u = \alpha x + \delta d,$$

where $\alpha^2 + \delta^2 = 1$, $(d,d)_T = 1$ and $x \perp_T d = 0$. Direct computation shows that

$$\lambda - \rho = \delta^2\, d^T(\lambda I - A)d.$$

So, we conclude that (cf. (2.3.4))

$$|\lambda - \rho| = \mathcal{O}(\delta^2), \tag{2.7.5}$$

while $|\lambda - \theta|$ is in general "only" $\mathcal{O}(\delta)$. (The reason for the last statement is that in general the eigenvectors are not stationary points of $\theta(u)$.) Therefore, the Rayleigh

quotient ρ is asymptotically (i.e., when u converges to x) more accurate than the usual Rayleigh quotient θ.

The crucial observation in this subsection is that if $\mathcal{U}$ is the search space for the (right) eigenvector, then with regard to the "inner product" (2.7.2), $\mathcal{U}$ forms a search space for the left eigenvector of equal quality. So, the fundamental difference with the two-sided Jacobi–Davidson algorithm is that as we build up a right search space (i.e., a search space for the right eigenvector), we get a reasonable left search space for free. We do not have to (approximately) solve a left correction equation as in the two-sided Jacobi–Davidson algorithm.

In view of (2.7.4) and (2.7.5), we take, instead of the usual Ritz–Galerkin condition on the residual (2.2.1), the same condition but with respect to the "inner product" (2.7.2):

$$r = Au - \rho u \perp_T \mathcal{U},$$

Writing $u = Uc$, $c \in \mathbb{C}^k$, we find that (ρ, c) should be a solution of the projected eigenproblem

$$U^T A U c = \rho\, U^T U c.$$

Therefore, it is practical that the search matrix U has complex orthogonal columns, $U^T U = I$ (note that this will not always be possible). This is achieved by a variant of MGS, denoted by MGS-CS. We have two possible choices for a correction equation for JDCS: one looking for an orthogonal update $s_1 \perp u$, or one looking for a complex orthogonal update $s_2 \perp_T u$. The first option leads to a correction equation with orthogonal projections (cf. (2.2.3))

$$(I - uu^*)\,(A - \rho I)\,(I - uu^*)\,s_1 = -(A - \theta I)u, \qquad s_1 \perp u.$$

Note that the right-hand side contains θ instead of ρ to ensure its orthogonality to u. The constraint $s_2 \perp_T u$ gives a correction equation with oblique projections; in this case, the left and right equation (2.4.3) and (2.4.4) reduce to one equation:

$$\left(I - \frac{uu^T}{u^T u}\right)(A - \rho I)\left(I - \frac{uu^T}{u^T u}\right)s_2 = -(A - \rho I)u, \qquad s_2 \perp_T u. \tag{2.7.6}$$

The operator in this equation is complex symmetric. So, we can try to solve (2.7.6) by a linear solver that is especially designed for complex symmetric systems, such as CSYM [16], complex symmetric QMR [26], or COCG [93]. The Jacobi–Davidson type algorithm JDCS is summarized in Algorithm 2.7.6.

The results of Section 2.5 can easily be carried over to the complex symmetric eigenvalue problem. For instance, the asymptotic convergence of exact JDCS (where one solves the correction equation exactly) is cubic, while the convergence of the inexact variant (fixed norm reduction in inner iteration) is linear.

2.7.3 The polynomial eigenproblem

We now derive the right correction equation of two-sided JD for the polynomial eigenproblem

$$p(\lambda)x = 0, \tag{2.7.7}$$

Input: a device to compute Ax for arbitrary x, a starting vector u_1, and a tolerance ε
Output: an approximation (ρ, u) to an eigenpair of A satisfying $\|Au - \rho u\| \leq \varepsilon$

1. $s = u_1$, $U_0 = [\,]$
   for $k = 1, 2, \ldots$
2. $U_k = \text{MGS-CS}\,(U_{k-1}, s)$
3. Compute kth column of $W_k = AU_k$
   Compute kth row and column of $H_k = U_k^T W_k$
4. Compute the eigenpair (ρ, c) of $U_k^T A U_k$ that is closest to the target τ
5. $u = U_k c \,/\, \|U_k c\|$
6. $r = (A - \rho I)u = W_k c \,/\, \|U_k c\| - \rho u$
7. Stop if $\|r\| \leq \varepsilon$
8. Solve (approximately) for either $s_1 \perp u$ or $s_2 \perp_T u$ from
$$(I - uu^*)\,(A - \rho I)\,(I - uu^*)\,s_1 \;=\; -(A - \theta I)u,$$
$$\left(I - \tfrac{uu^T}{u^T u}\right)(A - \rho I)\left(I - \tfrac{uu^T}{u^T u}\right)s_2 \;=\; -(A - \rho I)u,$$
   respectively

ALGORITHM 2.7.6: Jacobi–Davidson for the complex symmetric eigenvalue problem

where

$$p(\lambda) = \lambda^l A_l + \lambda^{l-1} A_{l-1} + \cdots + \lambda A_1 + A_0.$$

Suppose that we have approximate right and left eigenvector $u \in \mathcal{U}$ and $v \in \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are, as before, the right and left search spaces. The Petrov–Galerkin condition $p(\theta)u \perp \mathcal{V}$ implies that $\theta = \theta(u, v)$ satisfies

$$\sum_l (v^* A_l u)\,\theta^l = 0. \tag{2.7.8}$$

To derive Newton's method for (2.7.7), consider

$$p(\theta)(u + h) - p(\theta)(u) = p(\theta)(h) + \sum_l l\,\theta^{l-1} A_l u \frac{\partial \theta}{\partial u} h + \mathcal{O}(\|h\|^2).$$

Differentiating (2.7.8) with respect to u gives

$$\sum_l l\,(v^* A_l u)\,\theta^{l-1} \frac{\partial \theta}{\partial u} + \sum_l \theta^l v^* A_l = 0,$$

so with the notation $z := p'(\theta)u = \sum_l l\,\theta^{l-1} A_l u$ we find that, if $v^* z \neq 0$,

$$\frac{\partial \theta}{\partial u} = -\left(\sum_l l\,(v^* A_l u)\theta^{l-1}\right)^{-1} v^* \left(\sum_l \theta^l A_l\right) = -\,(v^* z)^{-1}\, v^* p(\theta).$$

Hence, the Jacobian $\frac{\partial p(\theta)}{\partial u}$ is equal to $\left(I - \frac{zv^*}{v^* z}\right) p(\theta)$, and a Newton step solves s from

$$\left(I - \frac{zv^*}{v^* z}\right) p(\theta) s = -p(\theta)u,$$

where $p(\theta)u$ can be seen as the (right) residual for the polynomial eigenvalue problem. This is the right correction equation for the polynomial eigenvalue problem; see also [73, (8.4)], where the result is stated without derivation. For the special case $l = 1$ this leads to (2.7.1), and if, in addition, $A_1 = -I$, we get (2.4.3). Because this two-sided process is a Newton method, we expect locally quadratic convergence. This is, under some conditions, true indeed [50, Theorem 2].

2.8 Various issues

2.8.1 Time complexity

We examine the time complexity of one outer iteration step of the methods introduced in this chapter (applied to the standard eigenvalue problem). Let k be the dimension of the search spaces, and let m be the number of steps with a linear solver (e.g., GMRES or BiCG) to solve the correction equation. One may check that in each of the methods we need $\mathcal{O}(k^3)$ time to solve the small projected eigenproblems. The number of matrix-vector multiplications (MVs) with A and A^* per outer iteration are summarized in the following table. For comparison, we also display standard JD in the table.

TABLE 2.1: Number of matrix-vector multiplications per outer iteration of each of the methods.

method	# MVs with A	# MVs with A^*
standard JD	$m+1$	0
two-sided JD (bi-orthogonal and orthogonal)	$m+1$	$m+1$
alternating JD (even step)	$m+1$	0
alternating JD (odd step)	1	$m+1$
JDCS	$m+1$	0

For the number of actions with a preconditioner, replace $m+1$ by m. Hence, two-sided Jacobi–Davidson is approximately twice as expensive as standard Jacobi–Davidson, alternating Jacobi–Davidson and JDCS. The same statement holds with respect to the storage requirements.

2.8.2 Deflation

If we have found one or more eigentriples of A, and we want to find another, we can deflate to avoid finding the same value again. Suppose that we have already found the right eigenvectors x_i and corresponding left eigenvectors y_i. Then it can be verified that, if we found the exact vectors,

$$\widetilde{A} = \prod_i \left(I - \frac{x_i y_i^*}{y_i^* x_i}\right) \cdot A \cdot \prod_i \left(I - \frac{x_i y_i^*}{y_i^* x_i}\right)$$

has the same eigentriples as A, except that the found eigenvalues are transformed to zeros.

2.8.3 Comparison with two-sided Lanczos

Suppose that we do not solve the corrections equations (2.4.3) and (2.4.4), but just take $\widetilde{s} = r_u = (A - \theta I)u$ and $\widetilde{t} = r_v = (A - \theta I)^* v$. Because of the orthogonalization at Step 2 of Algorithm 2.4.2, this is equivalent to taking $\widetilde{s} = Au$ and $\widetilde{t} = A^* v$, which is the subspace expansion of two-sided Lanczos. Therefore two-sided JD may, besides as a generalization of standard JD, also be regarded as a generalization of two-sided Lanczos.

2.8.4 Breakdown

Like two-sided Lanczos, two-sided JD may suffer from a breakdown, but in two-sided JD this can easily be overcome. First, BiCG (which we may use to solve the correction equations) may break down. Second, in the bi-orthogonal variant, the computed updates $\widetilde{s}$ and $\widetilde{t}$ may be (nearly) orthogonal. Realizing that our aim is to compute an eigenvalue and not to solve the correction equation accurately, we see that these breakdowns are not an intrinsic problem. In both cases, we can simply restart the method, or take different (e.g., random) approximate solutions to the correction equation.

2.9 Numerical experiments

Our experiments are coded in MATLAB and executed on a SUN workstation. We have already seen that JD has different convergence behavior for normal (cubic convergence) and nonnormal matrices (quadratic convergence); this in contrary to two-sided JD. The following lemma implies that two-sided JD does "feel" a difference, but this is only noticeable in the norm of the residuals, and not in the aproximations to the eigenvalue.

Lemma 2.9.1 *Let $A = X\Lambda Y^*$ be a diagonalizable matrix (so $Y^* = X^{-1}$). If there are no rounding errors, and two-sided JD's correction equations* (2.4.3) *and* (2.4.4) *in step k are solved by m_k steps of a Krylov method (without preconditioning), then two-sided JD applied to*

(a) *A, with starting vectors u_1 and v_1, and*

(b) *Λ, with starting vectors, $\widehat{u}_1 := Y^* u_1$ and $\widehat{v}_1 := X^* v_1$*

gives "the same" approximations: $\widehat{\theta}_k = \theta_k$. Moreover, $\widehat{u}_k = Y^* u_k$ and $\widehat{v}_k = X^* v_k$. In particular, if A is normal, then $\|\widehat{u}_k\| = \|u_k\|$ and $\|\widehat{v}_k\| = \|v_k\|$.

Proof: The first approximate eigenvalues are the same in both cases:

$$\widehat{\theta}_1 := \frac{\widehat{v}_1^* \Lambda \widehat{u}_1}{\widehat{v}_1^* \widehat{u}_1} = \frac{v_1^* X \Lambda Y^* u_1}{v_1^* (XY^*) u_1} = \frac{v_1^* A u_1}{v_1^* u_1} =: \theta_1.$$

For the right residuals in the first step of the method we have $\widehat{r}_u^{(1)} = (\Lambda - \widehat{\theta}_1 I)\widehat{u}_1 = (\Lambda - \theta_1 I) Y^* u_1$, so $X\widehat{r}_u^{(1)} = r_u^{(1)}$. In the same way we find a similar relation for the left residuals: $Y\widehat{r}_v^{(1)} = r_v^{(1)}$. So $\widehat{r}_u^{(1)} = Y^* r_u^{(1)}$ and $\widehat{r}_v^{(1)} = X^* r_v^{(1)}$. Denote by $\mathcal{K}_m(A, r)$ the

Krylov subspace of dimension m, generated by A and r. For the Krylov subspaces we have (generalization of [61, p. 264]):

$$\mathcal{K}_m(A, r_u^{(1)}) = \mathcal{K}_m(X\Lambda Y^*, X\widehat{r}_u^{(1)}) = X\mathcal{K}_m(\Lambda, \widehat{r}_u^{(1)}),$$

and likewise $\mathcal{K}_m(A^*, r_v^{(1)}) = \mathcal{K}_m(Y\Lambda^* X^*, Y\widehat{r}_v^{(1)}) = Y\mathcal{K}_m(\Lambda^*, \widehat{r}_v^{(1)})$. With little extra work one can check that same relations hold for the shifted and projected matrices that are present in the correction equations, for instance

$$\left(I - \frac{uv^*}{v^*u}\right)(A - \theta I)\left(I - \frac{uv^*}{v^*u}\right) = X\left(I - \frac{\widehat{u}\widehat{v}^*}{\widehat{v}^*\widehat{u}}\right)(\Lambda - \theta I)\left(I - \frac{\widehat{u}\widehat{v}^*}{\widehat{v}^*\widehat{u}}\right)Y^*.$$

So, using the notation $\widehat{P} = I - \frac{\widehat{u}\widehat{v}^*}{\widehat{v}^*\widehat{u}}$,

$$\mathcal{K}_m(P(A - \theta I)P, r_u^{(1)}) = X\mathcal{K}_m(\widehat{P}(\Lambda - \theta I)\widehat{P}, \widehat{r}_u^{(1)}).$$

We conclude that the approximate solutions from the first correction equations satisfy $\widehat{s}^{(1)} = Y^* s^{(1)}$ and $\widehat{t}^{(1)} = X^* t^{(1)}$. By induction we can prove that $\widehat{U}_k = Y^*X$ and $\widehat{V}_k = X^*V_k$, so the projected matrices are the same in both cases: $\widehat{H}_k := \widehat{V}_k^*\Lambda\widehat{U}_k = V_k^*AU_k = H_k$. In particular, the approximations to the eigenvalues are the same, and the approximate eigenvectors (u_k, v_k) and $(\widehat{u}_k, \widehat{v}_k)$ are transformations of each other: $\widehat{u}_k = Y^*u_k$ and $\widehat{v}_k = X^*v_k$. In particular, if A is normal, then X and Y are orthogonal, and so $\|\widehat{r}_k\| = \|r_k\|$. □

In the same way one may verify the next lemma.

Lemma 2.9.2 *With the assumptions and notations of the previous lemma, if the equations of two-sided RQI (see* (2.3.2)*) in step k are solved by m_k steps of a Krylov method (without preconditioning), then two-sided RQI applied to*

(a) *A, with starting vectors u_1 and v_1, and*

(b) *Λ, with starting vectors, $\widehat{u}_1 := Y^*u_1$ and $\widehat{v}_1 := X^*v_1$*

gives "the same" approximations: $\widehat{\theta}_k = \theta_k$. Moreover, $\widehat{u}_k = Y^*u_k$ and $\widehat{v}_k = X^*v_k$. In particular, if A is normal, then $\|\widehat{u}_k\| = \|u_k\|$ and $\|\widehat{v}_k\| = \|v_k\|$.

Experiment 2.9.3 Because of these results, our first example is $A = \mathrm{diag}(1:100)$. In Figure 2.1(a) we compare exact two-sided RQI (solid line) and inexact two-sided RQI (dashed line). We take for u_1 and v_1 the 100th basisvector plus 0.2 times a random vector (MATLAB's function `rand`, 'seed' 0), and take $\xi_1 = \xi_2 = 0.5$ in (2.5.1). In this figure, we show the error $|\lambda - \theta|$ in the approximation to the eigenvalue λ. One may see the somewhat faster convergence for ordinary RQI. What we do not see in the figure is that inexact two-sided RQI converges to $\lambda = 100$, while exact two-sided RQI converges to $\lambda = 79$. Apperently, without subpace acceleration it is impossible to guide the process to the desired eigenvalue.

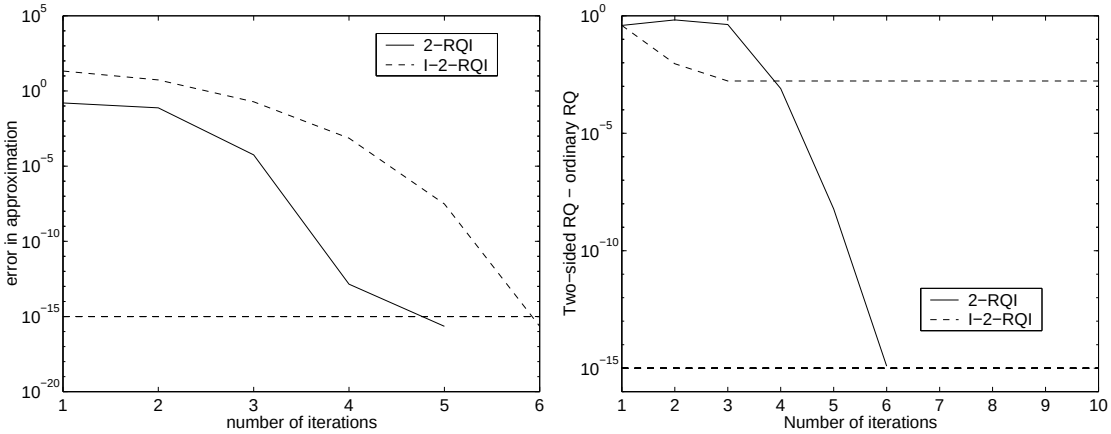


FIGURE 2.1: (a) The convergence of exact (solid line) and inexact (tolerance 0.5, dashed line) two-sided RQI for diag(1 : 100). (b) The difference of the two-sided Rayleigh quotient ($\frac{v^*Au}{v^*u}$) and the right Rayleigh quotient ($\frac{u^*Au}{u^*u}$) for exact (solid line) and inexact (tolerance 0.5, dashed line) two-sided RQI for $A = \text{tridiag}(1, -2, 1.2)$ of size 100×100.

Figure 2.1(b) is an example of the fact that inexact two-sided RQI does not need to converge. Here $A = \text{tridiag}(1, -2, 1.2)$. that is, A is the 100×100 tridiagonal matrix with stencil $[\,1\ -2\ 1.2\,]$, u_1 and v_1 are random vectors, and $\xi_1 = \xi_2 = 0.5$. We plot the difference between the two-sided Rayleigh quotient ($\theta(u, v) = \frac{v^*Au}{v^*u}$) and the right Rayleigh quotient ($\theta(u) = \frac{u^*Au}{u^*u}$). For inexact two-sided RQI, this difference (and the difference $\frac{v^*Au}{v^*u} - \frac{v^*A^*v}{v^*v}$) stabilizes. A small comfort is the fact that two-sided RQI can diagnose itself that there is a misconvergence. ⊘

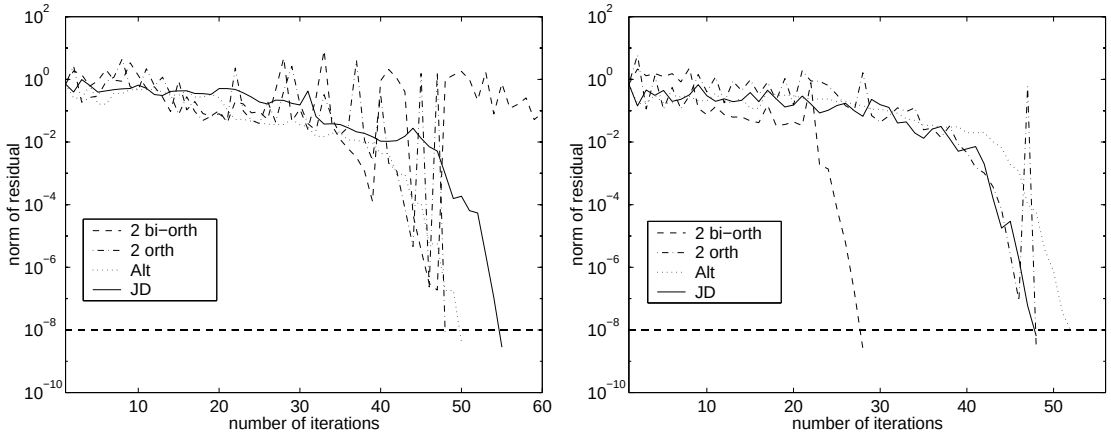


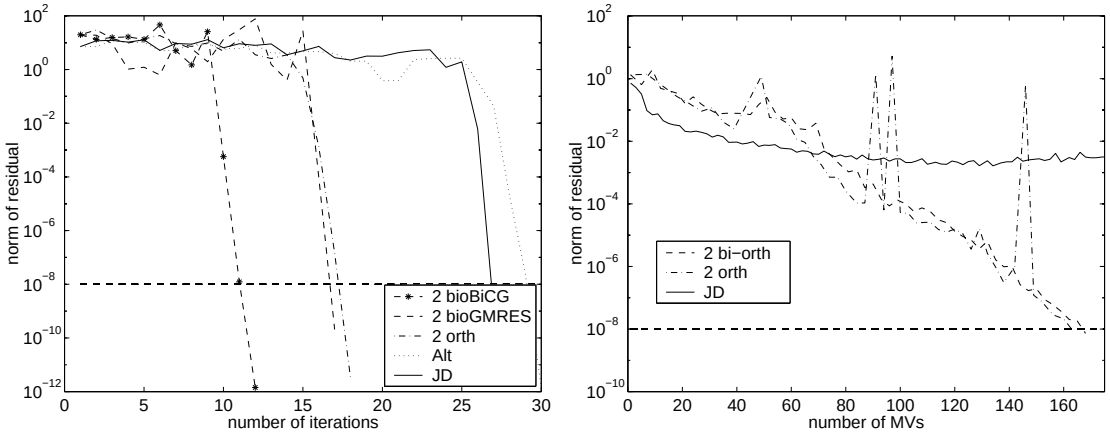
FIGURE 2.2: (a) The convergence histories of bi-orthogonal two-sided JD (dash), orthogonal two-sided JD (dash-dot), alternating JD (dot), and JD (solid) for the tridiagonal matrix with stencil $[-1\ 2\ 1.2]$ of size 100×100. All correction equations are solved by five steps of GMRES. (b) The same as (a), but now 10 steps of GMRES.

Experiment 2.9.4 Next, we experiment with two-sided JD variants. For Figure 2.2(a), we look for the eigenvalue with the largest magnitude for the 100×100 tridiagonal matrix with stencil $[\,-1 \;\; 2 \;\; 1.2\,]$. The starting vectors are random, and the correction equations are solved by 5 steps of unpreconditioned GMRES. All eigenvalues have real part equal to 2, and come in complex conjugate pairs. Note that, for the two-sided methods, the plotted line always represents $\min\{\|r_u\|, \|r_v\|\}$. The horizontal dashed line shows the stopping tolerance. We see that alternating JD is faster (also measured in matrix-vector products (MVs)) than standard JD. For orthogonal two-sided JD we choose the variant of (2.4.7). The method uses fewer iterations, but more MVs than standard JD. The convergence is very irregular; this might be improved using a target when one suspects that the process is converging. Bi-orthogonal two-sided JD almost converges, but then shows irregular behavior, and does not converge within 60 iterations. Using a target may be a good idea here as well.

For Figure 2.2(b), we change only the number of inner iteration steps to 10. Bi-orthogonal two-sided JD uses the fewest number of iterations. It uses slightly more MVs than JD. However, earlier in the process we already have more information. For instance, after 21 iterations of bi-orthogonal two-sided JD, $\kappa(\lambda) \approx 56.45$ is already approximated to a relative error of 0.5%: the condition number is well approximated before the method starts to converge. (Twenty-one iteration steps may not seem to be "early in the process". However, with an initial space that is "rich" in the direction of the desired eigenvector, the initial stage of slow convergence will be absent. We will have such a situation when we continue the process for the second eigenvalue after detection of the first one.) Upon termination, the norms of the residuals are $\|r_u\| \approx 3.9 \cdot 10^{-8}$ and $\|r_v\| \approx 2.5 \cdot 10^{-9}$. Using only four extra MVs to find u more accurately (see (2.4.5)), we have $\|r_u\| \approx 9.3 \cdot 10^{-9}$. This experiment is also an illustration of the situation that $\theta(u,v)$ is often more accurate than $\theta(u)$ and $\overline{\theta(v)}$: we have $|\lambda - \theta(u,v)| \approx 3.6 \cdot 10^{-15}$, while $|\lambda - \theta(u)| \approx 3.4 \cdot 10^{-10}$ and $|\lambda - \overline{\theta(v)}| \approx 1.6 \cdot 10^{-11}$. Alternating JD uses slightly more MVs than standard JD, but approximates the condition of the eigenvalue after 47 iterations up to 0.1% relative accuracy. Moreover, upon termination, the norms of both residuals are small ($\|r_u\| \approx 3.8 \cdot 10^{-8}$ and $\|r_v\| \approx 8.9 \cdot 10^{-9}$). Note the irregular convergence of the orthogonal variant of two-sided JD. ⊘

Experiment 2.9.5 As the next example, we take SHERMAN4 (size 1104, available from the Matrix Market [53]), u_1 random, and $v_1 = Au_1$. We solve the correction equations by 25 steps, see Figure 2.3(a). Now two-sided bi-orthogonal JD with BiCG is (also measured in MVs) the fastest method. Bi-orthogonal two-sided JD with GMRES and orthogonal two-sided JD with GMRES use fewer iterations, but more MVs than standard JD. Alternating JD is somewhat slower than standard JD, but finds the two eigenvectors with $\|r_u\| \approx 2.4 \cdot 10^{-8}$ and $\|r_v\| \approx 1.5 \cdot 10^{-12}$. Also in this example, the two-sided methods approximate $\kappa(\lambda)$ well already a few steps before termination.

For Figure 2.3(b), we take a symmetric matrix, the 1000×1000 matrix SHERMAN1. The starting vectors are the same as for (a). We solve the correction equations such that the relative residuals (ξ_1 and ξ_2 in (2.5.1), (2.5.3), and (2.5.4)) are less than 0.7. The convergence of the two-sided methods looks roughly linear (cf. Section 2.5, the number of



(a) (b)

FIGURE 2.3: (a) The convergence histories of bi-orthogonal two-sided JD (with BiCG, dash-star), bi-orthogonal two-sided JD (GMRES, dash), orthogonal two-sided JD (GMRES, dash-dot), alternating JD (GMRES, dot), and JD (GMRES, solid) for the matrix SHERMAN4. All correction equations are solved by 5 steps of GMRES. (b) The convergence histories of bi-orthogonal two-sided JD (dash), orthogonal two-sided JD (dash-dot), and JD (solid) for the matrix SHERMAN1, as a function of the MVs. All correction equations are solved to precision ($\xi_1 = \xi_2$) 0.7 by GMRES.

MVs per iteration is also almost constant), while standard JD does not converge within 175 MVs. The history of alternating JD is the same as that of JD, since the matrix is normal. ⊘

2.10 Conclusions

We have discussed an alternative approach to find eigenvalues and eigenvectors of a nonnormal matrix. Two-sided JD is a natural generalization of standard JD for nonnormal matrices. Without further demonstration, we mention that most of the techniques known in JD (such as preconditioning the correction equation, using a target, restarting, and using refined Ritz vectors) carry over to two-sided JD.

At the introduction of two-sided JD, we have focussed on the fast convergence of the method: exact two-sided JD has asymptotically cubic convergence for simple eigenvalues of nonnormal matrices. However, in practice this might not be the most important advantage of the method. Another benefit is the fact that already *during the process*, we have approximations to *both* the left *and* and the right eigenvector. We can use this information for an estimation of the condition of the eigenvalue $\kappa(\lambda)$. This, in turn, can be used as an error estimation

$$|\lambda - \theta| \lesssim \kappa(\lambda)\|r\|,$$

which can serve as a stopping criterion. Moreover, when we spot an eigenvalue with (possibly) a high condition, we may want to try to avoid it (using a target) when we are not interested in it, or stop the method and continue with standard JD.

During or after the process we can compare the three Rayleigh quotients $\theta(u)$, $\overline{\theta(v)}$, $\theta(u,v)$ to check for misconvergence, that is, check to see if they converge to the same

value. Moreover, from (2.3.1) it is clear that $\theta(u, v)$ can be more accurate ($\mathcal{O}(\delta_k \varepsilon_k)$) than $\theta(u)$ and $\overline{\theta(v)}$ ($\mathcal{O}(\delta_k)$ or $\mathcal{O}(\varepsilon_k)$), and this is confirmed by numerical experiments.

Compared with two-sided Lanczos, two-sided JD is more flexible, in the sense that we can restart with any vectors we like, and add some extra vectors to the subspaces. Two-sided JD is also more stable than two-sided Lanczos, in the sense that it can easily cope with breakdown, no look-ahead versions are necessary (see Section 2.8.4).

Of course, compared with standard JD, two-sided JD has also disadvantages. First of all, we need the action of multiplication by A^*. Two-sided JD costs approximately twice the work per iteration compared with standard JD, and also roughly twice the storage. One could argue that by two steps of ordinary RQI (or JD) one gets the fourth degree of the error, in contrast to the third degree by one step of two-sided RQI (or JD). Ostrowski [56, p. 472] states that

> "from this point of view, even in the case of a non-Hermitian matrix, the use of the ordinary Rayleigh quotient iteration appears to be not only permissible but even advisable".

However, Parlett [60, Remark 3, p. 689] criticizes this statement (in the context of dense methods).

Because of the two-sided Rayleigh quotient and the oblique projections, two-sided JD may have difficulties with eigenvalues with a large condition, affecting the stability of the method. This can result in *loss* of accuracy in determining λ; the order remarks above have little significance if $\kappa(\lambda)$ is huge.

In conclusion, two-sided JD is a natural alternative to standard JD and two-sided Lanczos for nonnormal matrices, especially in situations where the matrix is nonnormal (but not pathetically so) and when it is of interest to have approximations to the left eigenvector and condition of the eigenvalue during the process. Alternating JD may also give good results, especially if the matrix is slightly nonnormal.

The methods can be extended to the complex symmetric, generalized, and polynomial eigenvalue problem.

Acknowledgments The largest part of this chapter has been reprinted from Lin. Alg. Appl. 358(1-3), M. E. Hochstenbach and G. L. G. Sleijpen, Two-sided and alternating Jacobi–Davidson, pp. 145–172, Copyright (2003), with permission from Elsevier.

Chapter 3

A Jacobi–Davidson type SVD method

Abstract. We discuss a new method for the iterative computation of a portion of the singular values and vectors of a large sparse matrix. Similar to the Jacobi–Davidson method for the eigenvalue problem, we compute in each step a correction by (approximately) solving a correction equation. We give a few variants of this Jacobi–Davidson SVD (JDSVD) method with their theoretical properties. It is shown that JDSVD can be seen as an (inexact) accelerated Newton scheme. We experimentally compare the method with some other iterative SVD methods.

Key words: Jacobi–Davidson, singular value decomposition (SVD), singular values, singular vectors, norm, augmented matrix, Rayleigh quotient, correction equation, accelerated inexact Newton, refining singular values.

AMS subject classification: 65F15, 65F50 (65F35).

3.1 Introduction

Suppose that we want to compute one or more singular values, and the corresponding singular vectors, of the real $m \times n$ matrix A. (For convenience, we first consider real matrices, see Section 3.7.9 for complex matrices.) This subject has already been studied from a number of different viewpoints [28, 29, 17, 96, 97, 63], for example, to determine a few of the largest or smallest singular triples. This partial SVD can be computed in two different ways using equivalent eigenvalue decompositions.

The first is to compute some eigenvalues and eigenvectors of the $n \times n$ matrix A^TA or the $m \times m$ matrix AA^T. For large (sparse) matrices, direct methods like the QR method are unattractive, but there exist several iterative methods. In [63], for example, (block) Lanczos [51] and Davidson [19] are applied to A^TA. Another candidate is Jacobi–Davidson [75]. Note that it is in general not advisable (or necessary) to explicitly form the product A^TA. The nonzero eigenvalues of A^TA and AA^T are the squares of the nonzero singular values of A. This works positively for the separation of large singular values, but it forces a clustering of small ones. Moreover, it can be hard to find very

small singular values (relative to the largest singular value) accurately. Apart from this, the approaches via A^TA or AA^T are asymmetric: in the process we approximate only one of the two singular vectors. The second vector can be obtained from the first by a multiplication by A or A^T, but this may introduce extra loss of accuracy. Besides, when we have approximations to both the left and right singular vector, we can use only one of them as a starting vector for an iterative method.

A second approach is to compute some eigenvalues and eigenvectors of the *augmented matrix*

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}. \tag{3.1.1}$$

This approach has its own advantages and disadvantages. The eigenvalues of the augmented matrix are plus and minus the singular values of A, and we can extract the left and right singular vectors from the eigenvectors by just taking the first and second part (see Section 3.2). This makes an extra multiplication by A or A^T unnecessary. We do not have the drawback of squaring small singular values. On the negative side, the augmented matrix is larger in size, and the smallest singular values are in the interior of the spectrum.

The Lanczos method for the augmented matrix has been studied by a number of authors [28, 29, 17]. The Lanczos process does not exploit the special (block or "two-cyclic") structure of the matrix, unless the starting vector is of the form $(u, 0)$ or $(0, v)$. This is essentially Lanczos bidiagonalization of A; see [31, p. 495].

We can also consider the Jacobi–Davidson method [75] for the augmented matrix. This is an efficient method for the computation of a few eigenpairs, and it is of a different nature in comparison to Lanczos. The essence of Jacobi–Davidson is its correction equation, where the shifted operator is restricted to the subspace orthogonal to the current approximation to an eigenvector. When we solve this equation exactly, we can show that the updated vector is the same as the one we would get by one step of Rayleigh quotient iteration (RQI). But in practice one solves the Jacobi–Davidson correction equation only approximately, and one accelerates the convergence by projecting the matrix onto the subspace spanned by all iterates. Therefore, Jacobi–Davidson can also be viewed as an accelerated inexact RQI.

"Standard" Jacobi–Davidson does not make use of the structure of the augmented matrix. In this chapter we propose a Jacobi–Davidson variant that *does* take advantage of the special structure of the matrix. Instead of searching the eigenvector in one subspace, we search the left and right singular vectors in separate subspaces. We still solve a correction equation for the augmented matrix, but we use different projections, and we split the approximate solution of this equation for the expansion of the two search spaces. More similarities and differences are discussed in Section 3.7.5.

After some preparations in Section 3.2, we introduce the new approach, which we call the Jacobi–Davidson SVD (JDSVD), in Section 3.3. In Section 3.4, a few variants of the algorithm with their properties are presented. In Section 3.5, we show that the JDSVD process can be viewed as an accelerated (inexact) Newton scheme, and in Section 3.6 we focus on convergence. Various aspects of the method are discussed in Section 3.7, and after numerical examples in Section 3.8, we finish with conclusions in Section 3.9.

3.2 Preliminaries

Let A be a real $m \times n$ matrix with SVD $A = X\Sigma Y^T$ and singular values

$$0 \leq \sigma_{\min} = \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_2 \leq \sigma_1 = \sigma_{\max},$$

where $p := \min\{m, n\}$. Denote the corresponding left and right singular vectors by x_j $(1 \leq j \leq m)$ and y_j $(1 \leq j \leq n)$, respectively. If $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$, then, for convenience, we write $[\, a^T \;\; b^T \,]^T \in \mathbb{R}^{m+n}$ also as (a, b).

Definition 3.2.1 Let $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, $\mathcal{X} \subset \mathbb{R}^m$, and $\mathcal{Y} \subset \mathbb{R}^n$. We say that $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^{m+n}$ is *double-orthogonal* to the pair of subspaces $(\mathcal{X}, \mathcal{Y})$ if both $u \perp \mathcal{X}$ and $v \perp \mathcal{Y}$, which is denoted by $\begin{bmatrix} u \\ v \end{bmatrix} \perp\!\!\perp \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \end{bmatrix}$. The subspace $\{(a, b) \in \mathbb{R}^m \times \mathbb{R}^n : u^T a = v^T b = 0\}$ is denoted by $(u, v)^{\perp\perp}$. ⊘

The following lemma gives a relation between the singular triples of A and the eigenpairs of the augmented matrix.

Lemma 3.2.2 (Jordan–Wielandt; see [83, Theorem I.4.2]) *The augmented matrix* (3.1.1) *has eigenvalues*

$$-\sigma_1, \ldots, -\sigma_p, \underbrace{0, \ldots, 0}_{|m-n|}, \sigma_p, \ldots, \sigma_1$$

and eigenvectors

$$\begin{bmatrix} x_j \\ \pm y_j \end{bmatrix} \quad (1 \leq j \leq p)$$

corresponding to the $\pm\sigma_j$ *and, if* $m \neq n$, *additionally,*

$$\textit{either} \begin{bmatrix} x_j \\ 0 \end{bmatrix} \; (n+1 \leq j \leq m) \quad \textit{or} \quad \begin{bmatrix} 0 \\ y_j \end{bmatrix} \; (m+1 \leq j \leq n),$$

depending on whether $m > n$ *or* $n > m$.

The next definition is the natural analogue of the definition of a simple eigenvalue (see, e.g., [83, p. 15]). It is also defined in [82, p. 205].

Definition 3.2.3 We call σ_i a *simple singular value* of A if $\sigma_j \neq \sigma_i$ for all $j \neq i$. ⊘

The following lemma gives a link between a simple singular value of A and a simple eigenvalue of A^TA and AA^T.

Lemma 3.2.4 *Let* $\sigma > 0$. *Then* σ *is a simple singular value of* A *if and only if* σ^2 *is a simple eigenvalue of* A^TA *and* AA^T.

Proof: The nonzero eigenvalues of A^TA and AA^T are just the squares of the nonzero singular values of A (see, for example, [83, p. 31]). □

Note that the condition $\sigma > 0$ in the previous lemma is necessary. For example, 0 is a simple singular value of the 1×2 matrix $A = [\,0\ \ 0\,]$, but it is not a simple eigenvalue of A^TA.

For future use, we mention the following well-known results. Recall from Section 1.6 that eigenvalues are ordered increasingly: $\lambda_{\min} = \lambda_1 \leq \cdots \leq \lambda_n = \lambda_{\max}$.

Lemma 3.2.5 (Weyl; see [101, pp. 101–102], [83, Corollary IV.4.9], and [61, Theorem 10.3.1]) *Let B and E be real symmetric $n \times n$ matrices. Then for all $1 \leq j \leq n$*

$$\lambda_j(B) + \lambda_{\min}(E) \leq \lambda_j(B+E) \leq \lambda_j(B) + \lambda_{\max}(E).$$

Lemma 3.2.6 (see [41, (3.3.17)]) *If B and E are $m \times n$ matrices, then for $1 \leq i,\ j \leq p$, and $i + j\ \leq p + 1$,*

$$\sigma_{i+j-1}(B+E) \leq \sigma_i(B) + \sigma_j(E).$$

In particular, for $j = 1$ this yields $\sigma_i(B+E) \leq \sigma_i(B) + \sigma_1(E)$ for $i = 1, \ldots, p$.

Lemma 3.2.7 (see [40, (7.3.8)]) *Let B and E be real $m \times n$ matrices. Then*

$$\sum_{j=1}^{p} (\sigma_j(B+E) - \sigma_j(B))^2 \leq \|E\|_F^2 .$$

Lemma 3.2.8 (Unitary invariance of the singular values) *If U and V are orthogonal $m \times m$ and $n \times n$ matrices, respectively, then for all $1 \leq j \leq p$ we have $\sigma_j(U^TAV) = \sigma_j(A)$. In particular, $\|U^TAV\| = \|A\|$.*

Proof: The SVD of U^TAV is just $(U^TX)\Sigma(V^TY)^T$. The final statement follows from the characterization of the matrix two-norm as the largest singular value. □

Lemma 3.2.9 (see [41, (3.1.3)]) *Let B be an $m \times n$ matrix, and let B_l denote a submatrix of B obtained by deleting a total of l rows and/or columns from B. Then*

$$\sigma_j(B) \geq \sigma_j(B_l) \geq \sigma_{j+l}(B)$$

for $1 \leq j \leq p$, where for a $q \times r$ matrix X we set $\sigma_j(X) = 0$ if $j > \min\{q, r\}$.

3.3 The JDSVD correction equation

Suppose that we have k-dimensional *search spaces* $\mathcal{U} \subset \mathbb{R}^m$ and $\mathcal{V} \subset \mathbb{R}^n$ and *test spaces* $\widetilde{\mathcal{U}} \subset \mathbb{R}^m$ and $\widetilde{\mathcal{V}} \subset \mathbb{R}^n$. To determine approximations θ, η to a singular value, and $u \in \mathcal{U}, v \in \mathcal{V}$ (of unit norm) to the corresponding left and right singular vectors, we impose the *double Galerkin condition* with respect to $\widetilde{\mathcal{U}}$ and $\widetilde{\mathcal{V}}$ on the *residual* r:

$$r = r(\theta, \eta) := \begin{bmatrix} Av - \theta u \\ A^T u - \eta v \end{bmatrix} \perp\!\!\!\perp \begin{bmatrix} \widetilde{\mathcal{U}} \\ \widetilde{\mathcal{V}} \end{bmatrix}. \tag{3.3.1}$$

Because $u \in \mathcal{U}$ and $v \in \mathcal{V}$, we can write $u = Uc$ and $v = Vd$, where the columns of the $m \times k$ matrix U and the columns of the $n \times k$ matrix V form bases for $\mathcal{U}$ and $\mathcal{V}$, respectively, and $c, d \in \mathbb{R}^k$. Then we want to find θ, η, c, and d that are solutions of

$$\begin{cases} \widetilde{U}^T A V d &= \theta\, \widetilde{U}^T U c, \\ \widetilde{V}^T A^T U c &= \eta\, \widetilde{V}^T V d, \end{cases} \tag{3.3.2}$$

where $\widetilde{U}$ and $\widetilde{V}$ are matrices with columns that form bases for $\widetilde{\mathcal{U}}$ and $\widetilde{\mathcal{V}}$. For *test vectors* $\widetilde{u} \in \widetilde{\mathcal{U}}$ and $\widetilde{v} \in \widetilde{\mathcal{V}}$, we have, in particular, that $r \perp\!\!\!\perp (\widetilde{u}, \widetilde{v})$; so if $\widetilde{u}^T u \neq 0$ and $\widetilde{v}^T v \neq 0$,

$$\theta = \frac{\widetilde{u}^T A v}{\widetilde{u}^T u}, \qquad \eta = \frac{\widetilde{v}^T A^T u}{\widetilde{v}^T v}. \tag{3.3.3}$$

This shows that the approximations θ and η may differ. We discuss possible choices for $\widetilde{\mathcal{U}}$ and $\widetilde{\mathcal{V}}$ and the resulting relations for u and v in the following section. For now, suppose that we have approximations (u, v, θ, η). We would like to have a double-orthogonal correction $(s, t) \perp\!\!\!\perp (u, v)$ to (u, v) such that

$$\begin{cases} A(v + t) &= \sigma(u + s), \\ A^T(u + s) &= \tau(v + t), \end{cases} \tag{3.3.4}$$

where $\sigma \geq 0$ and $\tau \geq 0$ need not be equal because the vectors are not normalized. However, since $A^T A(v + t) = \sigma\tau(v + t)$, we have $\sigma\tau = \sigma_i^2$ for some $1 \leq i \leq p$. Equations (3.3.4) can be rearranged to obtain

$$\begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r + \begin{bmatrix} (\sigma - \theta)u \\ (\tau - \eta)v \end{bmatrix} + \begin{bmatrix} (\sigma - \theta)s \\ (\tau - \eta)t \end{bmatrix}.$$

Now neglect the last term on the right-hand side. This can be considered as "throwing away second order terms" (asymptotically, $\sigma - \theta$, $\tau - \eta$, s, and t will all be small), and suggests that JDSVD is in fact a Newton method, which is true indeed (see Section 3.5). In fact, the disregarded terms are even of third order, from which we may expect cubic convergence, see Sections 3.4.2 and 3.6.1. Because σ and τ are unknown, we do not know the differences $(\sigma - \theta)u$ and $(\tau - \eta)v$ either. Therefore, we can consider the projection of the last equation onto $(\widetilde{u}, \widetilde{v})^{\perp\perp}$ along (u, v). This projection is given by

$$\begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix},$$

and it fixes r. Projecting the previous equation, we get

$$\begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix} \begin{bmatrix} -\sigma I_m & A \\ A^T & -\tau I_n \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r, \qquad (s,t) \perp\!\!\!\perp (u,v). \tag{3.3.5}$$

Furthermore, since for every $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$ such that $u^T a \neq 0$ and $v^T b \neq 0$

$$\begin{bmatrix} I_m - \frac{au^T}{u^T a} & 0 \\ 0 & I_n - \frac{bv^T}{v^T b} \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix},$$

(3.3.5) leads to the *JDSVD correction equation*

$$\begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix} \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix} \begin{bmatrix} I_m - \frac{au^T}{u^T a} & 0 \\ 0 & I_n - \frac{bv^T}{v^T b} \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r, \tag{3.3.6}$$

where $(s,t) \perp\!\!\!\perp (u,v)$. We see that in general, the operator in (3.3.6) is symmetric if and only if a and b are a nonzero multiple of $\widetilde{u}$ and $\widetilde{v}$. It maps $(u,v)^{\perp\perp}$ to $(\widetilde{u},\widetilde{v})^{\perp\perp}$. In Sections 3.5 and 3.6 we explain why this process may lead to fast convergence, and we will come to a generalized version of the correction equation. In the next section we examine several choices for the Galerkin conditions (3.3.1).

3.4 Choices for the Galerkin conditions

Consider the eigenvalue problem for a symmetric matrix B, where we have one subspace $\mathcal{W}$ that is used both as search space and test space. If the columns of W form an orthonormal basis for $\mathcal{W}$, then the projected matrix $W^T B W$ has some nice properties; see [61, Section 11.4]. We will see that searching in two spaces, as in JDSVD, spreads those properties over a few Galerkin choices. In this section we examine some obvious choices.

3.4.1 The standard choice

Let us first take the test spaces $\widetilde{\mathcal{U}}$ and $\widetilde{\mathcal{V}}$ equal to the search spaces $\mathcal{U}$ and $\mathcal{V}$, which we will call the standard choice. If the columns of U and V form orthonormal bases for $\mathcal{U}$ and $\mathcal{V}$, then with the notation $H := U^T A V$, (3.3.2) reduces to

$$Hd = \theta c \quad \text{and} \quad H^T c = \eta d. \tag{3.4.1}$$

This gives approximations $u = Uc$ and $v = Vd$, where c and d are, respectively, left and right singular vectors of H. With the requirement $\|c\| = \|d\| = 1$ and test vectors $\widetilde{u} = u$ and $\widetilde{v} = v$, we get

$$\theta = \eta = u^T A v. \tag{3.4.2}$$

For reasons of symmetry, we choose $a = \widetilde{u}$ $(= u)$ and $b = \widetilde{v}$ $(= v)$ in (3.3.6). The resulting algorithm for the computation of a singular triple (in particular the largest triple) is given in Algorithm 3.4.1.

Input: a device to compute Av and A^Tu for arbitrary u and v, starting vectors u_1 and v_1, and a tolerance ε

Output: an approximation (θ, u, v) to a singular triple of A satisfying $\left\| \begin{bmatrix} Av - \theta u \\ A^Tu - \theta v \end{bmatrix} \right\| \leq \varepsilon$

1. $s = u_1,\ t = v_1,\ U_0 = [\,],\ V_0 = [\,]$
   for $k = 1, \ldots$
2. $U_k = \mathrm{MGS}(U_{k-1}, s)$
   $V_k = \mathrm{MGS}(V_{k-1}, t)$
3. Compute kth column of $W_k = AV_k$
   Compute kth row and column of $H_k = U_k^T A V_k = U_k^T W_k$
4. Compute singular triples (θ, c, d) of H_k, $(\|c\| = \|d\| = 1)$
   and select one (for instance the largest)
5. $u = U_k c,\ v = V_k d$
6. $r = \begin{bmatrix} Av - \theta u \\ A^Tu - \theta v \end{bmatrix} = \begin{bmatrix} W_k d - \theta u \\ A^Tu - \theta v \end{bmatrix}$
7. Stop if $\|r\| \leq \varepsilon$
8. Solve (approximately) an $(s, t) \perp\!\!\perp (u, v)$ from
$$\begin{bmatrix} I_m - uu^T & 0 \\ 0 & I_n - vv^T \end{bmatrix} \begin{bmatrix} -\theta I_m & A \\ A^T & -\theta I_n \end{bmatrix} \begin{bmatrix} I_m - uu^T & 0 \\ 0 & I_n - vv^T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r$$

ALGORITHM 3.4.1: The standard JDSVD algorithm for the singular value problem

In Step 2 of the algorithm, repeated (modified) Gram–Schmidt is used to make s and t orthogonal to U_{k-1} and V_{k-1}, and to expand the search spaces with the normalized vectors. We omit the index k of all variables that are overwritten in every step. If we are interested in another singular value, for example, the smallest, or the one closest to a specific target, we should adjust our choice in Step 4 of the algorithm accordingly (Chapter 4 will contain more details). The variant of Algorithm 3.4.1 is the only variant of JDSVD for which the operator in (3.3.6) is symmetric and maps $(u, v)^{\perp\perp}$ in itself. Other choices imply that the operator is not symmetric or maps $(u, v)^{\perp\perp}$ to a different space. See also Section 3.7.4.

3.4.2 Optimality of this choice

In this section, we treat two theorems that indicate that the method resulting from this standard Galerkin choice is optimal in some sense.

Suppose we have an $m \times k$ matrix U and an $n \times k$ matrix V. Then for any $k \times k$ matrices K and L there are associated an $m \times k$ *residual matrix* $R_1(K)$ and an $n \times k$ *residual matrix* $R_2(L)$:

$$R_1(K) := AV - UK \quad \text{and} \quad R_2(L) := A^TU - VL.$$

Definition 3.4.1 (cf. [5, p. 19]) $\mathcal{U}$ and $\mathcal{V}$ are *invariant singular subspaces* if

$$A\mathcal{V} \subset \mathcal{U} \quad \text{and} \quad A^T\mathcal{U} \subset \mathcal{V}.$$

⊘

If there exist K and L such that the residual matrices are zero, then we have found left and right invariant singular subspaces, i.e., invariant subspaces of A^TA and AA^T. The following theorem states that if both U and V have orthonormal columns, then $H := U^TAV$ and $H^T = V^TA^TU$ minimize the norm of these residual matrices, which is a desirable property. It is a generalization of a result in the theory for eigenproblems (see [61, Theorem 11.4.2] and [83, Theorem IV.1.15]), which deals with residuals of the form $AV - VK$.

Theorem 3.4.2 *For given $m \times k$ matrix U and $n \times k$ matrix V, let $H = U^TAV$.*

(a) *If the columns of U are orthonormal, then for all $k \times k$ matrices K we have $\|R_1(H)\| \leq \|R_1(K)\|$. Moreover, H is unique with respect to the Frobenius norm $\|R_1(H)\|_F \leq \|R_1(K)\|_F$ with equality only when $K = H$.*

(b) *If the columns of V are orthonormal, then $H^T = V^TA^TU$ minimizes the norm of $R_2(L)$, and H^T is unique with respect to the Frobenius norm.*

Proof: Suppose that the columns of U are orthonormal; then $U^TU = I$, so

$$\begin{aligned} R_1(K)^TR_1(K) &= V^TA^TAV + K^TK - K^TH - H^TK \\ &= V^TA^TAV - H^TH + (K-H)^T(K-H) \\ &= R_1(H)^TR_1(H) + (K-H)^T(K-H). \end{aligned}$$

Since $(K-H)^T(K-H)$ is positive semidefinite, it follows that

$$\|R_1(K)\|^2 = \lambda_{\max}(R_1(K)^TR_1(K)) \geq \lambda_{\max}(R_1(H)^TR_1(H)) = \|R_1(H)\|^2,$$

where we used Lemma 3.2.5 in the inequality. For uniqueness, we realize that $\|B\|_F^2 = \mathrm{tr}(B^TB)$ for every real matrix B. Part (b) can be proved using the same methods. □

Now we focus on the singular values of H, which satisfy $\theta = u^TAv$, where u and v are approximate left and right singular vectors. This motivates the following definition.

Definition 3.4.3 For $u, v \neq 0$, we define the *Rayleigh quotient for the singular value problem* of u and v by

$$\theta(u,v) := \frac{u^TAv}{\|u\|\,\|v\|}.$$

⊘

This Rayleigh quotient has the following (attactive) properties.

Theorem 3.4.4 *Let $u, v \neq 0$. Then $\theta = \theta(u,v)$ has the following properties:*

(a) *$Av - \theta u \perp u$ and $A^Tu - \theta v \perp v$ iff $\|u\| = \|v\|$* (Ritz–Galerkin)*;*

(b) *$\theta = \mathrm{argmin}_\alpha \|Av - \alpha u\|$ and $\theta = \mathrm{argmin}_\alpha \|A^Tu - \alpha v\|$ iff $\|u\| = \|v\|$* (minimum residual)*;*

(c) (u, v) *is a stationary point of* θ *iff* u *is a left singular vector,* v *is a corresponding right singular vector, and* $\|u\| = \|v\|$;

(d) $\sigma_1 = \max_{u,v\neq 0} \theta(u, v), \quad (x_1, y_1) = \operatorname{argmax}_{u,v\neq 0} \theta(u, v)$;

(e) *Let* $u = x + \delta e$ *and* $v = y + \varepsilon f$ *be approximate singular vectors corresponding to the singular value* σ*, where* $e \perp x$ *and* $f \perp y$ *are vectors of unit length. Then* $|\theta(u, v) - \sigma| = \mathcal{O}((\delta + \varepsilon)^2)$.

Proof: Part (a) is obvious, (b) follows from (a), and for (c) we have

$$\begin{aligned} \frac{\partial \theta}{\partial u}(u, v) &= \frac{1}{\|u\|\,\|v\|}\left(u^T A - \theta(u, v)\frac{\|u\|}{\|v\|}v^T\right), \\ \frac{\partial \theta}{\partial v}(u, v) &= \frac{1}{\|u\|\,\|v\|}\left(v^T A^T - \theta(u, v)\frac{\|v\|}{\|u\|}u^T\right). \end{aligned}$$

Stationary means that all directional derivatives are zero, from which (c) follows. For part (d), cf. [31, p. 74]. From $\theta(u, v) = (\sigma + \delta\varepsilon e^T A f)/\sqrt{(1 + \delta^2)(1 + \varepsilon^2)}$, part (e) follows directly. □

In particular, the Rayleigh quotient minimizes the residual (b), and if u and v are first order approximations to the singular vectors, their Rayleigh quotient is a second order approximation to the singular value (e).

Motivated by Theorems 3.4.2 and 3.4.4, it seems attractive to take the k singular values $\theta_j^{(k)}$ of H_k as approximations to the singular values of A. When U_k and V_k have orthonormal columns, we see by Lemma 3.2.8 that these approximations converge in a finite number of steps to the singular values of A. In the following theorem we show that the approximations to the singular values converge monotonically increasing.

Theorem 3.4.5 *Let* $\theta_k^{(k)} \leq \cdots \leq \theta_1^{(k)}$ *be the singular values of* $H_k := U_k^T A V_k$*, where* U_k *and* V_k *have orthonormal columns. Then for all fixed* j *and increasing* k*, the* $\theta_j^{(k)}$ *converge monotonically increasing to the* σ_j.

Proof: H_k is a submatrix of H_{k+1}, so Lemma 3.2.9 gives $\theta_j^{(k+1)} \geq \theta_j^{(k)}$ for $1 \leq j \leq k$. Because of the orthogonality of U_k and V_k, the $\theta_j^{(k)}$ converge to the σ_j. □

Remark 3.4.6 In practice, one often observes that the $\theta_j^{(k)}$ converge strictly monotonically to the σ_j. With the aid of [101, pp. 94–98], conditions could be formulated under which the convergence is strict. ⊘

Note that the theorem does *not* say that the smallest approximations $\theta_k^{(k)}$ converge monotonically (decreasing) to σ_p, because Lemma 3.2.9 only gives us $\theta_{k+1}^{(k+1)} \leq \theta_{k-1}^{(k)}$. For example, if $u_k \approx x_p$ and $v_k \approx y_{p-1}$, then $\theta_k^{(k)} \approx 0$, so we see that the smallest approximation can in fact be (much) smaller than σ_p. Experiments show that the convergence of the $\theta_k^{(k)}$ can be irregular and slow (see Section 3.8). This is a serious difficulty of

working with the augmented matrix, because the smallest singular values are in the interior of its spectrum. We discuss this matter further in Sections 3.4.3, 3.7.4, 3.7.8, and Chapter 4. The following theorem gives some relations between the singular values of H_k and those of A. It is a generalization of [61, Theorems 11.5.1 and 11.5.2] and [83, Corollary IV.4.15]. For clarity, we leave out the index k as much as possible.

Theorem 3.4.7 *For $j = 1, \ldots, k$, there exist singular values $\sigma_{j'}$ of A which can be put in one-one correspondence with the singular values θ_j of H in such a way that*

$$|\sigma_{j'} - \theta_j| \leq \max\{\|R_1(H)\|, \|R_2(H^T)\|\} \qquad (1 \leq j \leq k).$$

Moreover,

$$\sum_{j=1}^{k} (\sigma_{j'} - \theta_j)^2 \leq \|R_1(H)\|_F^2 + \left\|R_2(H^T)\right\|_F^2.$$

Proof: Let the columns of $U_\perp$ and $V_\perp$ be orthonormal bases for the orthogonal complements of U and V, respectively. Then both $[U\ U_\perp]$ and $[V\ V_\perp]$ are orthogonal and

$$[U\ U_\perp]^T A [V\ V_\perp] = \begin{bmatrix} H & 0 \\ 0 & U_\perp^T A V_\perp \end{bmatrix} + \begin{bmatrix} 0 & U^T A V_\perp \\ U_\perp^T A V & 0 \end{bmatrix}. \tag{3.4.3}$$

Using Lemmas 3.2.8 and 3.2.6, respectively, we obtain for $1 \leq j \leq p = \min\{m, n\}$

$$\sigma_j(A) = \sigma_j\left([U\ U_\perp]^T A [V\ V_\perp]\right) \leq \sigma_j\left(\begin{bmatrix} H & 0 \\ 0 & U_\perp^T A V_\perp \end{bmatrix}\right) + \sigma_{\max}\left(\begin{bmatrix} 0 & U^T A V_\perp \\ U_\perp^T A V & 0 \end{bmatrix}\right).$$

Now

$$[U\ U_\perp]^T R_1(H) = \begin{bmatrix} 0 \\ U_\perp^T A V \end{bmatrix} \qquad \text{and} \qquad [V\ V_\perp]^T R_2(H^T) = \begin{bmatrix} 0 \\ V_\perp^T A^T U \end{bmatrix},$$

so, because of the orthogonal invariance of the norm (see Lemma 3.2.8), $\|R_1(H)\| = \|U_\perp^T A V\|$ and $\|R_2(H^T)\| = \|V_\perp^T A^T U\| = \|U^T A V_\perp\|$. Because

$$\Sigma\left(\begin{bmatrix} H & 0 \\ 0 & U_\perp^T A V_\perp \end{bmatrix}\right) = \Sigma(H) \bigcup \Sigma(U_\perp^T A V_\perp),$$

there exist indices j' such that

$$\sigma_{j'}\left(\begin{bmatrix} H & 0 \\ 0 & U_\perp^T A V_\perp \end{bmatrix}\right) = \theta_j.$$

So the theorem's first inequality is obtained by

$$\begin{aligned} \sigma_{\max}\left(\begin{bmatrix} 0 & U^T A V_\perp \\ U_\perp^T A V & 0 \end{bmatrix}\right) &= \max\{\|U_\perp^T A V\|, \|U^T A V_\perp\|\} \\ &= \max\{\|R_1(H)\|, \|R_2(H^T)\|\}. \end{aligned}$$

For the second inequality, apply Lemma 3.2.7 to the splitting of (3.4.3). □

For the following proposition, we need the minimax theorem for singular values [41, Theorem 3.1.2]

$$\sigma_j = \max_{\mathcal{X}^j \subset \mathbb{R}^n} \min_{0 \neq x \in \mathcal{X}} \frac{\|Ax\|}{\|x\|}, \tag{3.4.4}$$

where $\mathcal{X}^j$ ranges over all subspaces of $\mathbb{R}^n$ of dimension j.

The following proposition states that the singular values of $U_k^T A V_k$, as approximations to the largest singular values, are also *not* optimal in another sense.

Proposition 3.4.8 *Let U_k and V_k have orthonormal columns. For $1 \leq j \leq k$,*

$$\sigma_j(U_k^T A V_k) \leq \sigma_j(AV_k) \qquad \text{and} \qquad \sigma_j(U_k^T A V_k) \leq \sigma_j(A^T U_k).$$

Proof: This follows from the inequalities $\|U_k^T A V_k y\| \leq \|AV_k y\|$ and $\|V_k^T A^T U_k x\| \leq \|A^T U_k x\|$, and (3.4.4) □

We have seen that the $\sigma_j(U_k^T A V_k)$ increase monotonically and that they are bounded above by both $\sigma_j(AV_k) = \lambda_j^{1/2}(V_k^T A^T A V_k)$ and $\sigma_j(A^T U_k) = \lambda_j^{1/2}(U_k^T A A^T U_k)$. This forms one motivation to study other Galerkin choices. A second is the possibly irregular convergence of the smallest singular value of $U_k^T A V_k$.

3.4.3 Other choices

We will briefly mention some other choices for the test spaces, but refer to Chapter 4 for an extensive discussion. Suppose that the columns of V form an orthonormal basis for $\mathcal{V}$. By the Galerkin choice $\widetilde{\mathcal{U}} = A\mathcal{V}$, $\widetilde{\mathcal{V}} = \mathcal{V}$, with test vectors $\widetilde{u} = Av$, $\widetilde{v} = v$, and $u = Uc$, $v = Vd$, and $\|v\| = 1$, (3.3.2) reduces to

$$\begin{cases} V^T A^T A V d &= \theta\, V^T A^T U c, \\ V^T A^T U c &= \eta\, d. \end{cases} \tag{3.4.5}$$

One can check that to satisfy the Galerkin conditions, $(\theta\eta, d)$ should be an eigenpair of $V^T A^T A V$. Now first suppose that $V^T A^T U$ is nonsingular. Note that in this case $\eta \neq 0$; otherwise, $V^T A^T U$ would be singular. It follows that $c = \eta(V^T A^T U)^{-1} d$, $\eta = v^T A^T u$, and $\theta = v^T A^T A v / v^T A u$. When $V^T A^T U$ is singular, then this construction is impossible, but in this case we can simply restart the process or add extra vectors to the search spaces (see Section 3.7.2 and also Section 4.7).

With this Galerkin choice, θ and η do not converge monotonically in general, but we can apply well-known results from eigenvalue theory to ensure that their product does converge monotonically to the squares of the singular values and also to the smallest. In Section 3.7.4 we discuss the resulting correction equation.

Likewise, if the columns of U form an orthonormal basis for $\mathcal{U}$, the Galerkin choice $\widetilde{\mathcal{U}} = \mathcal{U}$, $\widetilde{\mathcal{V}} = A^T\mathcal{U}$ leads to the determination of $(\theta\eta, c)$, an eigenpair of $U^T A A^T U$. These two approaches are natural with respect to minimax considerations, as we will see now.

Lemma 3.4.9 *Let $\xi \in [0,1]$. Then we have the following minimax property for singular values:*

$$\sigma_j = \max_{\substack{\mathcal{S}^j \subset \mathbb{R}^m \\ \mathcal{T}^j \subset \mathbb{R}^n}} \min_{\substack{0 \neq s \in \mathcal{S}^j \\ 0 \neq t \in \mathcal{T}^j}} \xi \frac{\|At\|}{\|t\|} + (1-\xi)\frac{\|A^T s\|}{\|s\|} \qquad (1 \leq j \leq p). \tag{3.4.6}$$

Proof: This follows from (3.4.4) and the observation that A and A^T have the same singular values. □

When we have search spaces $\mathcal{U}$ and $\mathcal{V}$, it is a natural idea to substitute $\mathcal{U}$ for $\mathbb{R}^m$ and $\mathcal{V}$ for $\mathbb{R}^n$ in (3.4.6), as a generalization of a similar idea in the theory of eigenproblems; see [61, p. 236]. This gives the following approximations to the singular values:

$$\tau_j = \max_{\substack{\mathcal{S}^j \subset \mathcal{U} \\ \mathcal{T}^j \subset \mathcal{V}}} \min_{\substack{0 \neq s \in \mathcal{S}^j \\ 0 \neq t \in \mathcal{T}^j}} \xi \frac{\|At\|}{\|t\|} + (1-\xi)\frac{\|A^T s\|}{\|s\|}. \tag{3.4.7}$$

The following theorem relates these approximations to the Ritz values of A^TA and AA^T.

Theorem 3.4.10 $\tau_j = \xi(\lambda_j^{1/2}(V^TA^TAV)) + (1-\xi)(\lambda_j^{1/2}(U^TAA^TU))$.

Proof: We have that $\mathcal{T}^j \subset \mathcal{V}$ if and only if $\mathcal{T}^j = V\widetilde{\mathcal{T}}^j := \{Vt : t \in \widetilde{\mathcal{T}}^j\}$ and $\widetilde{\mathcal{T}}^j \subset \mathbb{R}^k$. So for the first term of the expression for the τ_j we have that

$$\max_{\mathcal{T}^j \subset \mathcal{V}} \min_{0 \neq t \in \mathcal{T}^j} \frac{\|At\|^2}{\|t\|^2} = \max_{\widetilde{\mathcal{T}}^j \subset \mathbb{R}^k} \min_{0 \neq t \in \widetilde{\mathcal{T}}^j} \frac{t^TV^TA^TAVt}{\|t\|^2} = \lambda_j(V^TA^TAV).$$

For the second term we have a similar expression. □

When we take $\xi = 0$ and $\xi = 1$ in Theorem 3.4.10, we recognize the Galerkin approaches described in (3.4.5) and the discussion after that. They can essentially be viewed as a two-sided approach to A^TA or AA^T, in the sense that we have approximations to both the left and the right singular vector during the process. In Chapter 4, these methods are called the $\mathcal{U}$-harmonic and $\mathcal{V}$-harmonic extraction, and will be discussed in more details.

As observed in Section 3.4.2, the standard Galerkin choice leads to monotone convergence for the largest singular value, but it can imply irregular behavior for the smallest singular value. As we will see in Section 4.2, a related problem is how to select the best approximate vectors. Suppose for the moment that A is square and invertible. If the minimal singular value is the one of interest, the above observation suggests to study the singular values of A^{-1}. In Chapter 4, we will pursue this approach, departing from Galerkin conditions on A^{-1}.

3.5 JDSVD as accelerated inexact Newton scheme

In [76], it is shown that the Jacobi–Davidson method can be interpreted as an accelerated inexact Newton scheme [22] for the eigenvalue problem. Here we show that the same is

true for JDSVD applied to the singular value problem. Define $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ as

$$F(u,v) := \begin{bmatrix} Av - \theta u \\ A^T u - \eta v \end{bmatrix},$$

where $\theta = \theta(u,v)$ and $\eta = \eta(u,v)$ are as in (3.3.3). Thus the function F is nonlinear. Consider the singular value problem where we require the singular vectors x, y to be scaled such that $x^T\widetilde{a} = 1$ and $y^T\widetilde{b} = 1$ for certain vectors $\widetilde{a} \in \mathbb{R}^m$ and $\widetilde{b} \in \mathbb{R}^n$. So we look for solutions x, y of the equation $F(u,v) = 0$ in the "hyperplane"

$$\left\{(u,v) \in \mathbb{R}^m \times \mathbb{R}^n : u^T\widetilde{a} = 1,\ v^T\widetilde{b} = 1\right\}.$$

We introduce these $\widetilde{a}$ and $\widetilde{b}$ to derive a more general form of the correction equation (3.3.6). If (u_k, v_k) are approximations to the singular vectors, then the next Newton approximations (u_{k+1}, v_{k+1}) are given by $(u_{k+1}, v_{k+1}) = (u_k, v_k) + (s_k, t_k)$, where $(s_k, t_k) \perp\!\!\!\perp (\widetilde{a}, \widetilde{b})$ satisfies

$$DF(u_k, v_k)(s_k, t_k) = -F(u_k, v_k) = -r_k.$$

Omitting the index k, one may check (remembering that $\theta = \theta(u,v)$ and $\eta = \eta(u,v)$ are as in (3.3.3)) that the Jacobian $DF(u,v)$ of F is given by

$$DF(u,v) = \begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix} \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix}.$$

Hence the correction equation of the Newton step is given by

$$\begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix} \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r, \qquad \text{where } (s,t) \perp\!\!\!\perp (\widetilde{a}, \widetilde{b}).$$

For every a, b so that $\widetilde{a}^T a \neq 0$ and $\widetilde{b}^T b \neq 0$, this is equivalent to the slightly more general form of the JDSVD correction equation (in comparison with (3.3.6)),

$$\begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix} \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix} \begin{bmatrix} I_m - \frac{a\widetilde{a}^T}{\widetilde{a}^T a} & 0 \\ 0 & I_n - \frac{b\widetilde{b}^T}{\widetilde{b}^T b} \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r, \quad (3.5.1)$$

where $(s,t) \perp\!\!\!\perp (\widetilde{a}, \widetilde{b})$. Note that the substitution $\widetilde{a} = u$ and $\widetilde{b} = v$ gives (3.3.6).

If we keep a, b, $\widetilde{a}$, $\widetilde{b}$, $\widetilde{u}$, and $\widetilde{v}$ fixed during the process, and if $\widetilde{u}^T u$, $\widetilde{v}^T v$, $\widetilde{a}^T a$, and $\widetilde{b}^T b$ are nonzero, then Newton iteration produces a series (u_k, v_k) that converges asymptotically quadratically towards (x, y) if the starting vector (u_1, v_1) is sufficiently close to (x, y).

But if we take a, b, $\widetilde{a}$, $\widetilde{b}$, $\widetilde{u}$, and $\widetilde{v}$ variable but converging to certain vectors, such that the denominators in (3.5.1) do not vanish, we get asymptotically quadratic convergence as well. The choice $a = \widetilde{a} = \widetilde{u} = u_k$ and $b = \widetilde{b} = \widetilde{v} = v_k$ leads to Algorithm 3.4.1. With other Galerkin choices described in Section 3.4, the test vectors $(\widetilde{u}, \widetilde{v})$ are, in general,

not equal to the approximations (u, v), and in this situation the vectors $\widetilde{a}$ and $\widetilde{b}$ can be useful; see Sections 3.6 and 3.7.4.

We see that JDSVD is a Newton scheme, accelerated by the usage of all previous iterates and the projection of A on the subspace that they span. This *subspace acceleration* accelerates the "prequadratic" phase of the method and ensures that we find a singular triple in a finite number of steps. It may be expensive to solve the correction equation exactly. Instead we may solve (3.5.1) approximately (see Section 3.6.2 and 3.7.1); the resulting method is an *accelerated inexact Newton* scheme.

In [73], it is proved that if the correction equation is solved exactly, then Jacobi–Davidson applied to a symmetric matrix has asymptotically cubic convergence. Because the augmented matrix (3.1.1) is symmetric, we expect that JDSVD can also reach cubic convergence. The next section shows that this expectation is correct indeed.

3.6 Convergence

In the previous section we have already seen that the correction equation represents a Jacobian system in a Newton step. Now we focus on the asymptotic convergence (see Section 1.3.4 for an informal definition). In Section 3.6.1, we study the convergence rate of exact JDSVD (see [73] for similar observations for Jacobi–Davidson applied to the eigenvalue problem), and in Section 3.6.2 the convergence rate of inexact JDSVD.

3.6.1 Exact JDSVD

In the correction equation (3.5.1), u and v are the current approximations and $\widetilde{u}$ and $\widetilde{v}$ are test vectors, but we have not said much about choosing a, b, $\widetilde{a}$, and $\widetilde{b}$. These vectors can vary per step. The next lemma and theorem show that exact JDSVD (that is, JDSVD where we solve the correction equation exactly) has asymptotically cubic convergence for specific choices of the test vectors $\widetilde{u}$ and $\widetilde{v}$ and the vectors $\widetilde{a}$ and $\widetilde{b}$. To be precise, with ε small enough, if

$$\angle(u_k, x) = \mathcal{O}(\varepsilon) \quad \text{and} \quad \angle(v_k, y) = \mathcal{O}(\varepsilon) \tag{3.6.1}$$

and if

$$\angle(\widetilde{a}, x) = \mathcal{O}(\varepsilon), \quad \angle(\widetilde{b}, y) = \mathcal{O}(\varepsilon), \quad \angle(\widetilde{u}, x) = \mathcal{O}(\varepsilon), \quad \text{and} \quad \angle(\widetilde{v}, y) = \mathcal{O}(\varepsilon), \tag{3.6.2}$$

then $\angle(u_{k+1}, x) = \mathcal{O}(\varepsilon^3)$ and $\angle(v_{k+1}, y) = \mathcal{O}(\varepsilon^3)$. Then the approximate singular values (see (3.3.3)) converge cubically as well. The following lemma is a generalization of [73, Lemma 3.1].

Lemma 3.6.1 *Assume that $Ay = \sigma x$ and $A^T x = \tau y$, where $\sigma, \tau > 0$, and that $\sqrt{\sigma\tau}$ is a simple singular value of A. Let a, b, $\widetilde{a}$, $\widetilde{b}$, $\widetilde{u}$, and $\widetilde{v}$ be such that $\widetilde{u}^T x$, $\widetilde{v}^T y$, $\widetilde{a}^T a$, $\widetilde{b}^T b$, $\widetilde{a}^T x$, and $\widetilde{b}^T y$ are all nonzero. Then the map*

$$G := \begin{bmatrix} I_m - \frac{x\widetilde{u}^T}{\widetilde{u}^T x} & 0 \\ 0 & I_n - \frac{y\widetilde{v}^T}{\widetilde{v}^T y} \end{bmatrix} \begin{bmatrix} -\sigma I_m & A \\ A^T & -\tau I_n \end{bmatrix} \begin{bmatrix} I_m - \frac{a\widetilde{a}^T}{\widetilde{a}^T a} & 0 \\ 0 & I_n - \frac{b\widetilde{b}^T}{\widetilde{b}^T b} \end{bmatrix}$$

is a bijection from $(\widetilde{a},\widetilde{b})^{\perp\perp}$ *onto* $(\widetilde{u},\widetilde{v})^{\perp\perp}$.

Proof: Suppose $(z_1,z_2) \perp\!\!\!\perp (\widetilde{a},\widetilde{b})$ and $G(z_1,z_2)=0$. We show that $z_1=z_2=0$. We have

$$\begin{bmatrix} -\sigma I_m & A \\ A^T & -\tau I_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \mu x \\ \nu y \end{bmatrix}$$

for certain μ,ν. Then

$$\begin{cases} Az_2 &= \sigma z_1 + \mu x, \\ A^T z_1 &= \tau z_2 + \nu y. \end{cases}$$

Multiplying the first equation by A^T and the second by A, we find

$$\begin{cases} (A^TA - \sigma\tau I)z_2 &= (\sigma\nu + \tau\mu)y, \\ (AA^T - \sigma\tau I)z_1 &= (\sigma\nu + \tau\mu)x. \end{cases}$$

So both z_1 and x belong to the kernel of $(AA^T-\sigma\tau I)^2$, and both z_2 and y belong to the kernel of $(A^TA-\sigma\tau I)^2$. From the simplicity of $\sigma\tau$ using Lemma 3.2.4, we have that z_1 and z_2 are multiples of x and y, respectively. Because $z_1 \perp \widetilde{a}$, $z_2 \perp \widetilde{b}$, and $\widetilde{a}^Tx \neq 0$, $\widetilde{b}^Ty \neq 0$, we conclude $z_1=z_2=0$. The bijectivity follows from comparing dimensions. □

The next theorem, a generalization of [73, Theorem 3.2], shows the cubic convergence.

Theorem 3.6.2 *With the assumptions of* Lemma 3.6.1*, if the initial vectors are close enough to the singular vectors corresponding to a simple nonzero singular value (i.e., if* (3.6.1) *holds), and if the correction equation is solved exactly, then for fixed vectors* $\widetilde{u}$*,* $\widetilde{v}$*,* $\widetilde{a}$*, and* $\widetilde{b}$*, the JDSVD process has quadratic convergence. Moreover, if* (3.6.2) *holds, then JDSVD has even cubic convergence.*

Proof: For convenience write

$$P = \begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^Tu} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^Tv} \end{bmatrix},\ B = \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix},\ Q = \begin{bmatrix} I_m - \frac{a\widetilde{a}^T}{\widetilde{a}^Ta} & 0 \\ 0 & I_n - \frac{b\widetilde{b}^T}{\widetilde{b}^Tb} \end{bmatrix}.$$

Then the correction equation (3.5.1) reads, for $(\widetilde{s},\widetilde{t}) \perp\!\!\!\perp (\widetilde{a},\widetilde{b})$,

$$PBQ(\widetilde{s},\widetilde{t}) = PB(\widetilde{s},\widetilde{t}) = -r = -B(u,v).$$

Suppose that $\widetilde{x}$ and $\widetilde{y}$ are scalar multiples of the singular vectors x and y and that $(\widetilde{x},\widetilde{y}) = (u,v)+(s,t)$, where $(s,t) \perp\!\!\!\perp (\widetilde{a},\widetilde{b})$, and $\|(s,t)\| = \mathcal{O}(\varepsilon)$. Our first goal is to show that $\|(s-\widetilde{s},t-\widetilde{t})\| = \mathcal{O}(\varepsilon^2)$. We know that there are $\sigma,\tau>0$ such that

$$0 = \begin{bmatrix} -\sigma I_m & A \\ A^T & -\tau I_n \end{bmatrix} \begin{bmatrix} \widetilde{x} \\ \widetilde{y} \end{bmatrix} = \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix} \begin{bmatrix} \widetilde{x} \\ \widetilde{y} \end{bmatrix} - \begin{bmatrix} (\sigma-\theta)\widetilde{x} \\ (\tau-\eta)\widetilde{y} \end{bmatrix}.$$

Therefore, we have

$$B(s,t) = -B(u,v) + ((\sigma-\theta)\widetilde{x},(\tau-\eta)\widetilde{y}). \tag{3.6.3}$$

We multiply this on the left side by P and use the facts $PB(u,v) = B(u,v)$ and $P(u,v) = 0$:

$$PB(s,t) = -B(u,v) + P((\sigma-\theta)s, (\tau-\eta)t). \tag{3.6.4}$$

Subtracting $PB(\widetilde{s},\widetilde{t}) = -B(u,v)$ from (3.6.4), we get

$$PB(s-\widetilde{s}, t-\widetilde{t}) = P((\sigma-\theta)s, (\tau-\eta)t). \tag{3.6.5}$$

Multiplying (3.6.3) on the left by $\begin{bmatrix} \widetilde{u} & 0 \\ 0 & \widetilde{v} \end{bmatrix}^T$ leads to

$$\begin{bmatrix} \sigma-\theta \\ \tau-\eta \end{bmatrix} = \begin{bmatrix} (\widetilde{u}^T\widetilde{x})^{-1} & 0 \\ 0 & (\widetilde{v}^T\widetilde{y})^{-1} \end{bmatrix} \begin{bmatrix} -\theta\widetilde{u}^T & \widetilde{u}^TA \\ \widetilde{v}^TA^T & -\eta\widetilde{v}^T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix}. \tag{3.6.6}$$

So for fixed $\widetilde{u}$, $\widetilde{v}$, $\widetilde{a}$, and $\widetilde{b}$ we have $\|PB(s-\widetilde{s}, t-\widetilde{t})\| = \mathcal{O}(\varepsilon^2)$. Using Lemma 3.6.1 and the assumption that the initial vectors are close enough to the singular vectors, we see that PB in (3.6.5) is invertible, so $\|(s-\widetilde{s}, t-\widetilde{t})\| = \mathcal{O}(\varepsilon^2)$, which implies quadratic convergence. But, if additionally (3.6.2) holds, then

$$\left\| \begin{bmatrix} -\theta\widetilde{u}^T & \widetilde{u}^TA \\ \widetilde{v}^TA^T & -\eta\widetilde{v}^T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} \right\| = \left\| \begin{bmatrix} x^TAt \\ y^TA^Ts \end{bmatrix} \right\| + \mathcal{O}(\varepsilon^2) = \sigma \left\| \begin{bmatrix} y^Tt \\ x^Ts \end{bmatrix} \right\| + \mathcal{O}(\varepsilon^2) = \mathcal{O}(\varepsilon^2),$$

so from (3.6.6) we see that $\|(\sigma-\theta, \tau-\eta)\| = \mathcal{O}(\varepsilon^2)$. We conclude that in this case the convergence is even cubic. □

One may check that the hypotheses on $\widetilde{u}$, $\widetilde{v}$, $\widetilde{a}$ and $\widetilde{b}$ in the theorem are true when we choose $\widetilde{u}$ and $\widetilde{a}$ equal to u_k or Av_k, and $\widetilde{v}_k$ and $\widetilde{b}$ equal to v_k or A^Tu_k in the process. The cubic convergence can be observed in practice; see Section 3.8.

3.6.2 Inexact JDSVD

Similar to Section 2.5.2, we can show that inexact JDSVD will typically result in asymptotically linear convergence. Suppose that in every step, we solve the correction equation inexactly, such that for $0 < \xi < 1$, $\widetilde{s} \perp \widetilde{a}$ and $\widetilde{t} \perp \widetilde{b}$ satisfy

$$\left\| \begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^Tu} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^Tv} \end{bmatrix} \begin{bmatrix} -\theta I_m & A \\ A^T & -\eta I_n \end{bmatrix} \begin{bmatrix} I_m - \frac{a\widetilde{a}^T}{\widetilde{a}^Ta} & 0 \\ 0 & I_n - \frac{b\widetilde{b}^T}{\widetilde{b}^Tb} \end{bmatrix} \begin{bmatrix} \widetilde{s} \\ \widetilde{t} \end{bmatrix} + r \right\| \leq \xi \|r\|. \tag{3.6.7}$$

Theorem 3.6.3 (Locally linear convergence of inexact JDSVD) *Suppose we solve JDSVD's correction equation inexactly according to* (3.6.7)*. Then the method has locally linear convergence.*

Proof: With the same notation as in the proof of Theorem 3.6.2, we know by definition that there exists a $\widetilde{\xi}$, $0 \leq \widetilde{\xi} \leq \xi$, and a unit vector $(e,f) \perp\perp (\widetilde{u},\widetilde{v})$, such that $(\widetilde{s},\widetilde{t}) \perp\perp (\widetilde{a},\widetilde{b})$ satisfies

$$PBQ(\widetilde{s},\widetilde{t}) = -r + \widetilde{\xi}\|r\|(e,f).$$

When we subtract (3.6.4) from this equation, this gives

$$PBQ(\widetilde{s}-s,\widetilde{t}-t) = \widetilde{\xi}\,\|r\|(e,f) - P((\sigma-\theta)s,(\tau-\eta)t).$$

From (3.6.4), we can deduce

$$\|r\| \leq \left\| \begin{bmatrix} -\sigma I_m & A \\ A^T & -\tau I_n \end{bmatrix} \right\| \left\| \begin{bmatrix} s \\ t \end{bmatrix} \right\|.$$

The result follows from

$$\|PBQ(\widetilde{s}-s,\widetilde{t}-t)\| \leq \widetilde{\xi} \left\| \begin{bmatrix} -\sigma I_m & A \\ A^T & -\tau I_n \end{bmatrix} \right\| \left\| \begin{bmatrix} s \\ t \end{bmatrix} \right\| + \mathcal{O}(\varepsilon^2),$$

and the fact that PBQ is invertible according to Lemma 3.6.1. □

3.7 Various issues

3.7.1 Solving the correction equation

We now translate a number of observations for Jacobi–Davidson in [75, 73] to the JDSVD context. Consider the situation after k steps of the JDSVD algorithm. For easy reading, we again leave out the index k. In this section we take for simplicity the Galerkin spaces used in Section 3.4.1, but most arguments carry over to other choices. First we rewrite the correction equation. Because of $(s,t) \perp\!\!\!\perp (u,v)$, we can eliminate the projections and write (3.3.6) as

$$\begin{bmatrix} -\theta I_m & A \\ A^T & -\theta I_n \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r + \begin{bmatrix} \alpha u \\ \beta v \end{bmatrix},$$

where α and β are determined by the requirement that $(s,t) \perp\!\!\!\perp (u,v)$. If we have a nonsingular preconditioner $M \approx \left[\begin{smallmatrix} -\theta I_m & A \\ A^T & -\theta I_n \end{smallmatrix}\right]$, then we can take an approximation

$$(\widetilde{s},\widetilde{t}) = -M^{-1}r + M^{-1}(\alpha u, \beta v). \tag{3.7.1}$$

1. (cf. [75, p. 406, point 1]) If we approximate (s,t) simply by $\pm r$ (by taking $M = \mp I$ and $\alpha = \beta = 0$), then, because of the orthogonalization at Step 2 of Algorithm 3.4.1, this is equivalent to taking $(\widetilde{s},\widetilde{t}) = (Av, A^Tu)$. By induction one can prove that for the special case where we take this simple approximation in every step, we have

   $$\mathcal{U}_{2k} = \mathcal{K}_k(AA^T, u_1) \oplus \mathcal{K}_k(AA^T, Av_1), \qquad \mathcal{V}_{2k} = \mathcal{K}_k(A^TA, v_1) \oplus \mathcal{K}_k(A^TA, A^Tu_1),$$

   as long as the Krylov subspaces have a trivial intersection. Compare this with Lanczos bidiagonalization, where

   $$\mathcal{U}_k = \mathcal{K}_k(AA^T, Av_1), \qquad \mathcal{V}_k = \mathcal{K}_k(A^TA, v_1).$$

2. (cf. [75, p. 408, point 3]) If θ is not equal to a singular value, then $M = \left[\begin{smallmatrix} -\theta I_m & A \\ A^T & -\theta I_n \end{smallmatrix}\right]$ is nonsingular and $M^{-1}r = (u, v)$. So for the updated vectors we have

$$\begin{bmatrix} u+s \\ v+t \end{bmatrix} = \begin{bmatrix} -\theta I_m & A \\ A^T & -\theta I_n \end{bmatrix}^{-1} \begin{bmatrix} \alpha u \\ \beta v \end{bmatrix}. \tag{3.7.2}$$

   We conclude that the exact JDSVD can be seen as an accelerated scaled RQI.

3. (cf. [75, p. 409, point 4]) If we take $M \neq \left[\begin{smallmatrix} -\theta I_m & A \\ A^T & -\theta I_n \end{smallmatrix}\right]$, M nonsingular, then with $(\widetilde{s}, \widetilde{t}) = M^{-1}(\alpha u, \beta v)$ we obtain an inexact shift and invert method. This may be an attractive alternative if (3.7.2) is expensive.

4. When we are interested in a singular value close to a specific *target* τ, we can replace this in the left-hand side of the correction equation (3.3.6):

$$\begin{bmatrix} I_m - uu^T & 0 \\ 0 & I_n - vv^T \end{bmatrix} \begin{bmatrix} -\tau I_m & A \\ A^T & -\tau I_n \end{bmatrix} \begin{bmatrix} I_m - uu^T & 0 \\ 0 & I_n - vv^T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = -r.$$

   An advantage of this approach is that we avoid misconvergence to some unwanted singular value "on the way." For example, if we want to compute the largest singular value, we can use a known approximation of $\sigma_{\max}$ as a target. In practice, $\tau \approx \|A\|_\infty$ may be a good guess (see Section 3.8). For the minimal singular value, we can take $\tau = 0$ or a small positive number as target. As soon as we notice that the process starts to converge, we may replace the target in the correction equation by the current approximation to the singular value again. A further advantage of using a target for the largest singular value is that the resulting system is (almost) definite, which is a favorable circumstance for the use of a preconditioner. Unfortunately, for the smallest singular value we have a ("severely") indefinite system.

5. In practice we often solve (3.5.1) approximately by an iterative method: for example, a few steps of GMRES or MINRES if the operator is symmetric (in case of the standard Galerkin choice). We may use a (projected) preconditioner; see Section 3.7.7.

3.7.2 Restart

A nice property of Jacobi–Davidson is its flexibility in restarting. JDSVD, too, has this advantage: we can restart at every moment in the process with any number of vectors, only keeping those parts of the search spaces that look promising, or possibly adding some extra vectors. This is practical when the search spaces become large or to avoid a breakdown in case of the nonstandard Galerkin choices. Of course, JDSVD can also be *started* with search spaces of dimension larger than one. This may be favorable when we look successively for singular triples of A and of a perturbed matrix $A + E$, for instance in the computation of pseudospectra.

3.7.3 Deflation

We can compute multiple singular triples of A by using a deflation technique. If we have found a singular triple of A, and we want to find another, we can deflate the augmented matrix to avoid finding the same triple again. For JDSVD, this can be done as follows. Suppose that X and Y contain the already found left and right singular vectors. Then it can be checked that, if we found the exact vectors,

$$\begin{bmatrix} I_m - XX^T & 0 \\ 0 & I_n - YY^T \end{bmatrix} \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} I_m - XX^T & 0 \\ 0 & I_n - YY^T \end{bmatrix}$$

has the same eigenvalues as the original augmented matrix, except that the found eigenvalues are transformed to zeros. The method can then be restarted with another approximate triple.

3.7.4 Correction equation with nonstandard Galerkin choices

In the case of nonstandard Galerkin choices (see Section 3.4.3 and Chapter 4), we may have the situation that $(\widetilde{u}, \widetilde{v}) \neq (u, v)$. Now we exploit the flexibility of (a, b) in (3.5.1): by the choice

$$(a, b) = (u, v) \text{ and } (\widetilde{a}, \widetilde{b}) = (\widetilde{u}, \widetilde{v}) \tag{3.7.3}$$

we ensure that the operator in (3.5.1) maps $(\widetilde{u}, \widetilde{v})^{\perp\perp}$ onto itself, and that the asymptotic convergence is cubic according to Theorem 3.6.2 (if the correction equation is solved exactly). Another option is

$$(a, b) = (\widetilde{u}, \widetilde{v}) \text{ and } (\widetilde{a}, \widetilde{b}) = (u, v), \tag{3.7.4}$$

to make the operator in (3.5.1) symmetric. In this case the operator maps $(u, v)^{\perp\perp}$ to $(\widetilde{u}, \widetilde{v})^{\perp\perp}$. Therefore, we should use a left "preconditioner" that maps the image space $(\widetilde{u}, \widetilde{v})^{\perp\perp}$ bijectively onto the domain space $(u, v)^{\perp\perp}$ (see also Section 3.8 and [73, 78]).

3.7.5 Comparison with Jacobi–Davidson on the augmented matrix

It is interesting to compare JDSVD with Jacobi–Davidson on the augmented matrix, starting with the "same" starting vector $w_1 = (u_1, v_1)/\sqrt{2}$.

There are some analogies between Jacobi–Davidson and JDSVD. When their correction equations are solved exactly, both converge asymptotically cubically to a simple eigenvalue of the augmented matrix. Moreover, the costs per iteration are almost the same; the only difference is that in each step JDSVD needs a small SVD, while Jacobi–Davidson needs a small eigenvalue decomposition. The storage requirements are also comparable.

The main difference is the fact that JDSVD, by construction, searches in two (smaller) subspaces, while Jacobi–Davidson has one search space. If Jacobi–Davidson solves its correction equation exactly, then in fact it solves (3.7.2) with $\alpha = \beta$ [75]. This suggests that JDSVD may cope better with "unbalanced" vectors, that is, vectors (u, v), where

$\|u\| \neq \|v\|$. An extreme example of this can be seen by taking a starting vector of the form $(x, \delta y)$ for $0 < \delta < 1$. In contrast to Jacobi–Davidson, JDSVD terminates after computing a zero residual.

Another (mostly theoretical) difference is the fact that JDSVD terminates for every starting vector after at most $\max\{m, n\}$ iterations, and Jacobi–Davidson terminates on the augmented matrix after at most $m + n$ iterations. In Sections 3.8 and 4.10, we compare the methods experimentally.

3.7.6 Refinement procedure

Suppose that we have found an approximate minimal right singular vector $v = (1 - \varepsilon^2)^{1/2} v_{\min} + \varepsilon v_{\max}$ by an iterative method applied to $A^T A$, so that $\sin \angle(v, v_{\min}) = \varepsilon$. Then, in the absence of other information, $u = Av = (1-\varepsilon^2)^{1/2} \sigma_{\min} u_{\min} + \varepsilon \sigma_{\max} u_{\max}$ is the best approximation to the left singular vector we have at our disposal. But $\tan \angle(u, u_{\min}) \approx \varepsilon \frac{\sigma_{\max}}{\sigma_{\min}} = \kappa(A)\varepsilon$, and this can be large. Moreover, $\|u\|^2 = (1 - \varepsilon^2)\sigma_{\min}^2 + \varepsilon^2 \sigma_{\max}^2$ can be an inaccurate approximation to $\sigma_{\min}^2$, and so may $\|A^T u\|^2 / \|u\|^2$ be. See also Lemma 4.3.1.

Hence the approximations to small singular values, resulting from working with $A^T A$, may be inaccurate. In this situation, we may try to improve (or *refine*) the approximate singular triple by a two-sided approach like JDSVD. The following lemma, a generalization of [73, Theorem 3.5], gives a link with [20], where a system with a matrix of the form

$$\begin{bmatrix} -\theta I_m & A & -u & 0 \\ A^T & -\theta I_n & 0 & -v \\ 2u^T & 0 & 0 & 0 \\ 0 & 2v^T & 0 & 0 \end{bmatrix} \tag{3.7.5}$$

is used for improving an approximate singular triple.

Lemma 3.7.1 *The JDSVD correction equation* (3.5.1) *is equivalent to*

$$\begin{bmatrix} -\theta I_m & A & -u & 0 \\ A^T & -\eta I_n & 0 & -v \\ \widetilde{a}^T & 0 & 0 & 0 \\ 0 & \widetilde{b}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} s \\ t \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \theta u - Av \\ \eta v - A^T u \\ 0 \\ 0 \end{bmatrix}; \tag{3.7.6}$$

that is, if (s, t, α, β) *is a solution of* (3.7.6)*, then* (s, t) *is a solution of the correction equation* (3.5.1)*, and if* (s, t) *is a solution of* (3.5.1)*, then there exist unique* α, β *such that* (s, t, α, β) *is a solution of* (3.7.6)*.*

Proof: We use the same notation as in the proof of Theorem 3.6.2. System (3.7.6) is equivalent to

$$B(s, t) - (\alpha u, \beta v) = -r \qquad \text{and} \qquad (s, t) \perp\!\!\!\perp (\widetilde{a}, \widetilde{b}).$$

By splitting the first equation in $(\widetilde{u}, \widetilde{v})^{\perp\!\!\!\perp}$ and its complement, we obtain

$$\begin{cases} PB(s, t) = -r, \\ \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} (\widetilde{u}^T u)^{-1} & 0 \\ 0 & (\widetilde{v}^T v)^{-1} \end{bmatrix} \begin{bmatrix} \widetilde{u}^T & 0 \\ 0 & \widetilde{v}^T \end{bmatrix} B \begin{bmatrix} s \\ t \end{bmatrix}, \\ (s, t) \perp\!\!\!\perp (\widetilde{a}, \widetilde{b}). \end{cases}$$

Note that we have used $Pr = r$, $P(\alpha u, \beta v) = 0$, and $r \perp\!\!\!\perp (\widetilde{u}, \widetilde{v})$. The first and third equation together are equivalent to the correction equation (3.5.1), and the second equation determines α, β uniquely. □

Of course, this equivalence is valid only when both (3.7.6) and (3.5.1) are solved exactly, not when we solve them approximately. In particular, when we substitute $\eta = \theta$ and $(\widetilde{a}, \widetilde{b}) = 2(u, v)$, the matrix in (3.7.6) becomes the one in (3.7.5).

3.7.7 Preconditioning the correction equation

The correction equation of JDSVD can be preconditioned in a manner similar to Jacobi–Davidson (see, for example, [78]). We use the same notation as in the proof of Theorem 3.6.2. Suppose that we have a preconditioner M for B. For left preconditioning we are given $(b_1, b_2) \perp\!\!\!\perp (\widetilde{u}, \widetilde{v})$, and we have to solve for $(z_1, z_2) \perp\!\!\!\perp (\widetilde{a}, \widetilde{b})$ from

$$PMQ(z_1, z_2) = (b_1, b_2).$$

Note that we project the preconditioner as well. Hence, for some α, β,

$$(z_1, z_2) = M^{-1}(b_1, b_2) - M^{-1}(\alpha u, \beta v),$$

and by left multiplication by $\widetilde{a}^T$ and $\widetilde{b}^T$ we obtain

$$\begin{bmatrix} \widetilde{a} & 0 \\ 0 & \widetilde{b} \end{bmatrix}^T M^{-1} \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \widetilde{a} & 0 \\ 0 & \widetilde{b} \end{bmatrix}^T M^{-1} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Thus, we have

$$(z_1, z_2) = \left(I - M^{-1} \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix} \left(\begin{bmatrix} \widetilde{a} & 0 \\ 0 & \widetilde{b} \end{bmatrix}^T M^{-1} \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix} \right)^{-1} \begin{bmatrix} \widetilde{a} & 0 \\ 0 & \widetilde{b} \end{bmatrix}^T \right) M^{-1}(b_1, b_2).$$

A recipe for computing (z_1, z_2) is given by the following four steps.

(1) Compute $(\widehat{u}_1, \widehat{u}_2) = M^{-1}(u, 0)$ and $(\widehat{v}_1, \widehat{v}_2) = M^{-1}(0, v)$.

(2) Compute $(\widehat{b}_1, \widehat{b}_2) = M^{-1}(b_1, b_2)$.

(3) Compute (α, β) from $\begin{bmatrix} \widetilde{a}^T\widehat{u}_1 & \widetilde{a}^T\widehat{v}_1 \\ \widetilde{b}^T\widehat{u}_2 & \widetilde{b}^T\widehat{v}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \widetilde{a}^T\widehat{b}_1 \\ \widetilde{b}^T\widehat{b}_2 \end{bmatrix}$.

(4) Compute $(z_1, z_2) = (\widehat{b}_1, \widehat{b}_2) - \alpha(\widehat{u}_1, \widehat{u}_2) - \beta(\widehat{v}_1, \widehat{v}_2)$.

An important observation is that Step (1) and the computation of the 2×2 matrix in Step (3) have to be performed only once at the start of the iterative solution process of the correction equation.

3.7.8 Smallest singular value

As mentioned in Section 3.4.1, the standard variant of JDSVD may have difficulties with finding the smallest singular value of a matrix. This is not surprising, because the small singular values of A correspond to the interior eigenvalues of the augmented matrix. But in many applications, e.g., the computation of pseudospectra, the smallest singular value is just what we want to compute.

Although Chapter 4 will be largely devoted to this subject, here we already give a clue. We can use JDSVD with the nonstandard Galerkin (harmonic) variants, mentioned in Section 3.4.3, starting with zero, or a small positive number as a target, and solve the correction equation rather accurately, possibly with the aid of a preconditioner; see Section 3.8. In this way the method is close to a shift and invert iteration but less expensive. Of course it is hereby advantageous to have a good initial triple (e.g., coming from an iterative method on A^TA); JDSVD (with nonstandard Galerkin) can then be used as refinement procedure.

Suppose A is square and invertible, and we are interested in the smallest singular value. If we have a preconditioner $M \approx A$, then, since

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & A^{-T} \\ A^{-1} & 0 \end{bmatrix}, \tag{3.7.7}$$

we may use this M to form a preconditioner for the augmented matrix.

3.7.9 JDSVD for complex matrices

When A is a complex matrix, then all methods in this chapter can be used when we replace the transpose by the conjugate transpose. An alternative approach, to avoid complex arithmetic, is to consider

$$\mathbf{A} = \begin{bmatrix} \mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A) \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathrm{Re}(u) \\ \mathrm{Im}(u) \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \mathrm{Re}(v) \\ \mathrm{Im}(v) \end{bmatrix}.$$

Then $\mathbf{Av} = \sigma\mathbf{u}$ and $\mathbf{A}^T\mathbf{u} = \sigma\mathbf{v}$. However, a drawback of this system is that $\Sigma(\mathbf{A})$ is the multiset $\Sigma(A) \bigcup \Sigma(A)$, which means that no singular value of $\mathbf{A}$ is simple, which is not favorable for convergence in view of Section 3.6.

3.7.10 Time complexity

With k the dimension of the search spaces and m the number of steps with a linear solver (e.g., GMRES or MINRES) to solve the correction equation, one outer iteration of JDSVD consumes $\mathcal{O}(k^3)$ time to solve the small projected singular value problem, and $2m + 2$ matrix-vector multiplications, half of which with A and half with A^T. We have to store bases for $\mathcal{U}$ and $\mathcal{V}$; storage can be reduced by a restart.

3.8 Numerical experiments

Our experiments are coded in MATLAB and are executed on a SUN workstation. The following lemma implies that up to rounding errors, it is not a loss of generality to consider (rectangular) diagonal matrices.

Lemma 3.8.1 *If there are no rounding errors, and JDSVD's correction equation* (3.5.1) *in step k is solved by l_k steps of GMRES, then JDSVD applied to*

(a) $A = X\Sigma Y^T$, *with starting vectors u_1 and v_1,*

(b) Σ, *with starting vectors $\widehat{u}_1 := X^T u_1$ and $\widehat{v}_1 := Y^T v_1$,*

gives "the same" results; that is,

$$\widehat{\theta}_k = \theta_k \qquad \text{and} \qquad \|\widehat{r}_k\| = \|r_k\|.$$

Proof: Define

$$Q = \begin{bmatrix} X^T & 0 \\ 0 & Y^T \end{bmatrix};$$

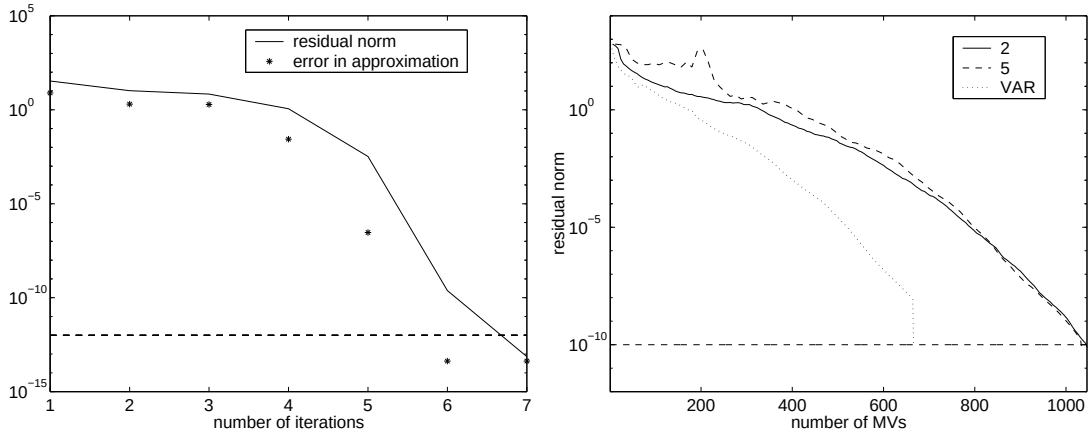
then Q is orthogonal, and one may verify that $(\widehat{u}_1, \widehat{v}_1) = Q(u_1, v_1)$, $\widehat{\theta}_1 := \widehat{u}_1^T \Sigma \widehat{v}_1 = u_1^T A v_1 =: \theta_1$, and $\widehat{r}_1 = Q r_1$. A well-known property of Krylov subspaces ensures that (see [61, p. 264])

$$Q^T \mathcal{K}_l\left(\begin{bmatrix} 0 & \Sigma \\ \Sigma^T & 0 \end{bmatrix}, \widehat{r}\right) = \mathcal{K}_l\left(Q^T \begin{bmatrix} 0 & \Sigma \\ \Sigma^T & 0 \end{bmatrix}, Q, Q^T \widehat{r}\right) = \mathcal{K}_l\left(\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}, r\right).$$

With little extra work one can check that the same relation holds for the shifted and projected matrices that are present in the correction equation (3.5.1), where one should bear in mind that all other vectors involved in the projectors ($\widetilde{a}$, $\widetilde{b}$, $\widetilde{u}$, $\widetilde{v}$, a, and b) must also be altered for the Σ-system in the obvious way. So the approximate solutions from the correction equations satisfy $(\widehat{s}_1, \widehat{t}_1) = Q(s_1, t_1)$. By induction we can prove that $\widehat{U}_k = X^T U_k$ and $\widehat{V}_k = Y^T V_k$, so the projected matrices are the same in both cases: $\widehat{H}_k := \widehat{U}_k^T \Sigma \widehat{V}_k = U_k^T A V_k = H_k$. In particular, the approximations to the singular values are the same, and the approximations (u_k, v_k) and $(\widehat{u}_k, \widehat{v}_k)$ are orthogonal transformations of each other: $(\widehat{u}_k, \widehat{v}_k) = Q(u_k, v_k)$ and $\widehat{r}_k = Q r_k$, so $\|\widehat{r}_k\| = \|r_k\|$. □

For this reason, we first study some phenomena on $A = \mathrm{diag}(1:100)$ and $A = \mathrm{diag}(1:1000)$.

Experiment 3.8.2 In Figure 3.1(a), the solid line is the convergence history of (the standard variant of Algorithm 3.4.1 of) JDSVD for the computation of the largest singular triple of $A = \mathrm{diag}(1:100)$. The starting vectors are the normalized $v_1 = v_{\max} + 0.1r$, where r is a vector with random entries, chosen from a uniform distribution on the unit interval, and $u_1 = Av_1/\|Av_1\|$. The dots represent the error in the approximation $\sigma_{\max} - \theta_k^{(k)}$. In all figures, a horizontal dotted line indicates the stopping tolerance. We solve the correction equation by 200 steps of (unpreconditioned) GMRES. Because the (augmented) matrices in the correction equation (Step 8 of Algorithm 3.4.1) are



(a) (b)

FIGURE 3.1: (a) The convergence history of the exact JDSVD algorithm for diag(1 : 100) as in Algorithm 3.4.1: residual norm (solid line) and error in the approximations to $\sigma_{\max}$ (dots). The horizontal dotted line indicates the stopping tolerance. (b) Convergence for diag(1 : 1000) using, respectively, 5, 2, and a variable number of GMRES steps to solve the correction equation.

of size 200×200, this means (theoretically) exactly, so according to Theorem 3.6.2 we expect cubic convergence. In Figure 3.1(a) we see, for instance, that the error in the approximation in iteration number 5 decreases from $\approx 10^{-2}$ to $\approx 10^{-7}$.

In Figure 3.1(b), we take $A = \text{diag}(1 : 1000)$, and u_1 and v_1 random vectors (as described above) with unit norm. We experiment with the number of GMRES steps. For the solid line, we solve the correction equation approximately by five steps of GMRES, which we denote by GMRES_5, for the dashed line by GMRES_2, and for the dotted line by a variable number equal to $\max\{2 \cdot (\lceil -\log \|r\| \rceil + 1), 0\}$. Measured in terms of matrix-vector products (MVs), the variable choice is best, while GMRES_2 and GMRES_5 are comparable. An explanation of this is that when the initial approximations are not good (as in this case), it is of no use to try hard to solve the correction equation in the beginning. When we are almost converging, it may make sense to solve it more accurately to get fast convergence. See also [78]. ⊘

Experiment 3.8.3 In Figure 3.2(a) we compare, for $A = \text{diag}(1 : 1000)$, the standard JDSVD method for the three largest singular triples (solid), with Jacobi–Davidson on the augmented matrix for the computation of the three largest eigenpairs (dashed), each with GMRES_5. For JDSVD, we take v_1 as a random vector, and $u_1 = Av_1/\|Av_1\|$. For Jacobi–Davidson we take the "same" starting vector $(u_1, v_1)/\sqrt{2}$. We see that JDSVD is faster for the first triple; for the second and third we restart with a good approximation, and then the histories are similar.

In Figure 3.2(b), we do the same, but now using GMRES_2. For the first two triples, JDSVD is somewhat faster than Jacobi–Davidson, for the third JDSVD in the first instance (mis)converges to the fourth largest singular value 997. Other experiments also suggest that JDSVD is generally (somewhat) faster than Jacobi–Davidson on the augmented matrix. In the following chapter we (practically) compare JDSVD and JD to compute the smallest singular values. There JDSVD seems much more advantageous. ⊘

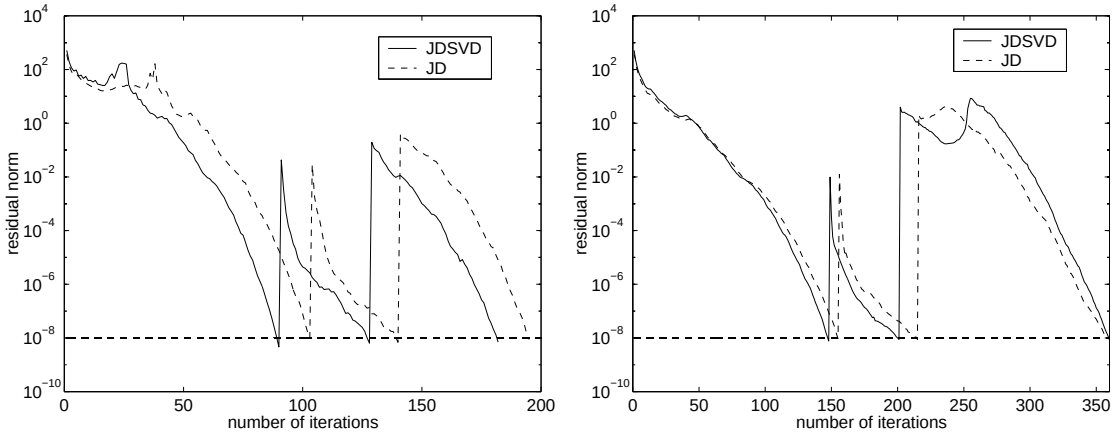


FIGURE 3.2: (a) JDSVD (solid) and Jacobi–Davidson (dashed) for the three largest σs of diag(1 : 1000). (b) The same as Figure 3.2(a), only with GMRES_2 to solve the correction equation.

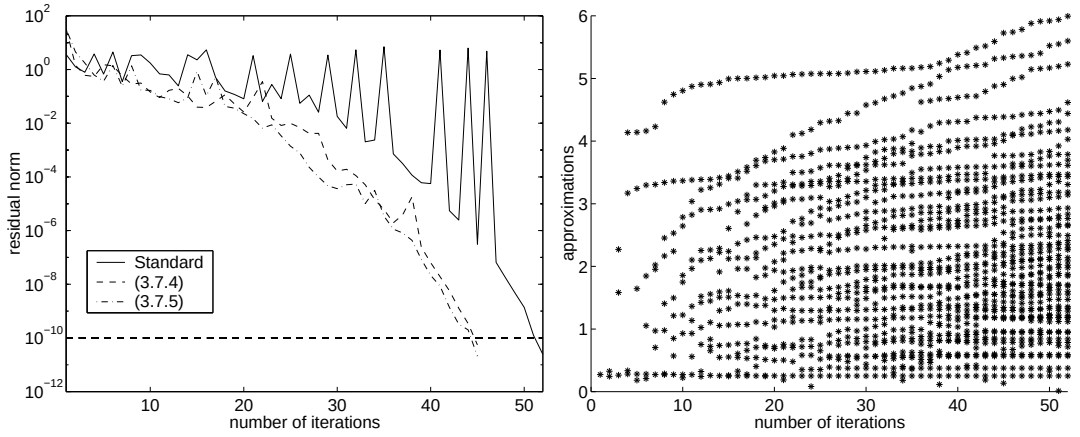
Experiment 3.8.4 Next, we take some examples from the Matrix Market [53]. For Figure 3.3(a), we apply different JDSVD variants to find the smallest singular triple of PDE225 ($\sigma_{\min} \approx 2.5 \cdot 10^{-1}$), using two random starting vectors and GMRES_{10} (no preconditioning). In all variants, we take initially target 0, but when $\|r\| < 10^{-3}$, we replace the target by the best approximations again (see Section 3.7.1, point 4). The solid line is the standard choice; we see an irregular convergence history, as could be expected (see Section 3.4). The dashed line represents the Galerkin choice (3.4.5), where in the correction equation (3.5.1) we substitute (3.7.3). Finally, the dash-dotted line is (3.4.5) with (3.7.4) substituted in (3.5.1). In the last case, as seen in Section 3.7.4, the operator in (3.5.1) maps $(u, v)^{\perp\perp}$ to $(\widetilde{u}, \widetilde{v})^{\perp\perp}$. Since in this case $v = \widetilde{v}$ but $u \neq \widetilde{u}$, we use a left "preconditioner" to handle the correction equation correctly. The preconditioned identity

$$\begin{bmatrix} I_m - \frac{\widetilde{u}u^T}{u^T\widetilde{u}} & 0 \\ 0 & I_n \end{bmatrix} I_{m+n} \begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n \end{bmatrix}$$

maps $(\widetilde{u}, \widetilde{v})^{\perp\perp}$ back to $(u, v)^{\perp\perp}$.

In Figure 3.3(b), standard JDSVD's approximations to the singular values during this process are plotted. These are "standard", nonharmonic estimates. Note the monotone convergence of the approximations to the largest singular values but the irregular behavior of the approximations to the smallest singular value. ⊘

Experiment 3.8.5 Next, we compare JDSVD with Lanczos bidiagonalization for the computation of $\sigma_{\max}$. These methods are of a different nature. Lanczos bidiagonalization can be viewed as an accelerated power method, while JDSVD can be seen as an accelerated inexact RQI. An advantage of JDSVD is that we may use preconditioning for the correction equation. Therefore, we expect that if we have a reasonable preconditioner, and if preconditioning is relatively cheap in comparison to a multiplication by A or A^T, then JDSVD can be cheaper than Lanczos bidiagonalization. On the other hand,



(a) (b)

FIGURE 3.3: (a) Three different JDSVD variants for the computation of $\sigma_{\min}$ of PDE225: standard, (3.4.5) + (3.5.1) + (3.7.3), and (3.4.5) + (3.5.1) + (3.7.4). (b) (Nonharmonic) approximations to the singular values by the standard variant.

if $m \gg n$, or if there is no good or cheap preconditioner available, then we expect that Lanczos bidiagonalization will be better. Table 3.1 shows some test results.

For JDSVD, we take a target $\tau \approx \|A\|_\infty$, in the hope that $\tau \approx \sigma_{\max}$. We make an incomplete LU-decomposition (using a drop tolerance displayed in the table) of the augmented matrix (3.1.1) minus τ times the identity, and we use $M = LU$ as a preconditioner. The starting vector v_1 is the vector with all coordinates equal to one, and is then normalized, and u_1 is a random vector. We solve the correction equation by only preconditioning the residual ("0 steps of GMRES"). Lanczos bidiagonalization uses v_1 as starting vector. Both methods stop if $\|r\| < 10^{-8}$. The matrix A_1 stands for $\mathrm{diag}(1:100) + 0.1 \cdot \mathrm{rand}(100, 100)$, where $\mathrm{rand}(100, 100)$ denotes an 100×100 matrix with random entries, chosen from a uniform distribution on the unit interval. See [63] for more information on the origin and singular values of the other matrices. For JDSVD, a pair is given, consisting of the number of MVs and the number of solves with L or U. For Lanczos bidiagonalization we show the number of MVs. We use a straightforward implementation of Lanczos bidiagonalization without any (partial) reorthogonalization; for a more sophisticated implementation see, e.g., [52].

For the first two examples, the target τ is relatively far from the largest singular value ($\sigma_{\max} \approx 0.66$ for HOR131, $\sigma_{\max} \approx 1.5 \cdot 10^5$ for PORES3). We see that Lanczos bidiagonalization is cheaper than JDSVD when we take the preconditioning into account. For SHERMAN1, the target is a reasonable approximation to $\sigma_{\max} \approx 5.05$. When we take the preconditioning into account, bidiagonalization is still somewhat cheaper than JDSVD. The last row of the table is an example where preconditioning is relatively cheap. The reason for this is that we now take the diagonal of A, instead A itself, to form an augmented matrix of the form (3.1.1) and to make an ILU-decomposition. Using far more MVs, Lanczos bidiagonalization is (also counting the preconditioning) much more expensive. ⊘

Experiment 3.8.6 Finally, in Table 3.2, we compare JDSVD for the computation of

TABLE 3.1: Some experiments with JDSVD to compute $\sigma_{\max}$, using incomplete LU-factorizations of the shifted augmented matrix. The number of MVs, and the number of solves with L or U is displayed in the 5th column. The shift (or target) τ (6th column) for the preconditioning is roughly taken to be $\|A\|_\infty$. The last three columns give information on the incomplete LU-factorization: the drop tolerance of ILU, and the resulting number of nonzeros of L and U. We compare JDSVD's results with the MVs of Lanczos bidiagonalization applied to A (4th column).

Matrix	Size	nnz(A)	bidiag	JDSVD	τ	droptol	nnz(L)	nnz(U)
HOR131	434×434	4182	30	(30, 84)	0.90	$1e-2$	1792	1792
PORES3	532×532	3474	80	(72, 210)	$2e5$	$1e-1$	1301	1300
SHERMAN1	1000×1000	3750	70	(18, 48)	5	$1e-2$	4805	4803
A_1	100×100	10000	74	(40, 114)	106	$1e-2$	299	299

$\sigma_{\min}$ with Lanczos bidiagonalization applied to A^{-1}. We use the Galerkin choice (3.4.5) for JDSVD. Note that the comparison with bidiagonalization is mainly meant to get an idea of how well JDSVD performs. In practice, for large (sparse) A, it is often too expensive to work with A^{-1} and A^{-T} or $\left(A^TA\right)^{-1}$. For JDSVD, we take a small target $\tau = 10^{-5}$, drop tolerance 10^{-3}, and we test two different kinds of preconditioners. The first type (odd rows in Table 3.2), represents an incomplete LU-decomposition for the augmented matrix based on the target, just as in the previous example. The second type of preconditioner (even rows in Table 3.2) is based on an ILU of A, see (3.7.7). The starting vectors are the same as for Table 3.1. We solve the correction equation by preconditioning only the residual ("0 steps of GMRES"). Both processes are continued until $\|r\| < 10^{-7}$.

TABLE 3.2: Some experiments with JDSVD to compute $\sigma_{\min}$. The numbers of MVs and solves with L or U (fourth column), and the number of nonzeros of L and U are displayed. The odd rows use an ILU of the augmented matrix, while the even rows exploit an ILU of A. We compare JDSVD's results with the number of MVs of Lanczos bidiagonalization applied to A^{-1}.

Matrix	$\sigma_{\min}(A)$	bidiag(A^{-1})	JDSVD	nnz(L)	nnz(U)
HOR131	$1.5e-5$	30	(24, 66)	20593	21167
			(34, 192)	3623	8117
PORES3	$2.7e-1$	12	(38, 108)	3683	5491
			(32, 180)	1727	2919
SHERMAN1	$3.2e-4$	16	(20, 54)	11575	11738
			(24, 132)	5777	5853
A_1	$9.5e-1$	14	(26, 72)	200	200
			(26, 144)	3000	1430

We see that although JDSVD may in general use more MVs, it may be much cheaper than Lanczos bidiagonalization applied to A^{-1}, due to the sparsity of A, L, and U. Again A_1 serves as an example for the situation where preconditioning is relatively cheap, which makes JDSVD attractive. The second kind of preconditioner strategy (ILU for square A) looks worthwhile as an alternative to a preconditioner for the augmented matrix. We also tried Lanczos bidiagonalization applied to A for the computation of $\sigma_{\min}$, but the

results were bad (262 MVs for A_1, and more than 500 MVs for the other matrices). ⊘

3.9 Conclusions

We have discussed an alternative approach for the computation of a few singular values and vectors of a matrix. The JDSVD method searches in two separate subspaces, and it can be interpreted as an inexact Newton method for the singular value problem. JDSVD can also be seen as an accelerated inexact scaled RQI method. Therefore, the best results may be expected when we have a good initial starting triple (refinement), but we can start with any approximations. While the asymptotic convergence is cubic if the correction equation is solved exactly, in practice we solve it approximately, and then the convergence typically looks (super)linear. Although we mainly discussed the application of JDSVD for the largest and smallest singular value, the method is in principle suitable for all singular values (see also the next chapter). We may use preconditioning for the solution of the correction equation. This can be a decisive factor for fast convergence. Experiments indicate that JDSVD is a good competitor to other iterative SVD methods, in particular when A is (almost) square and we have a reasonable, relatively cheap preconditioner for the correction equation, or when the smallest singular triples are sought.

Chapter 4

Harmonic and refined extraction methods for the singular value problem, with applications in least squares problems

Abstract. For the accurate approximation of the minimal singular triple (singular value and left and right singular vector), we may use two separate search spaces, one for the left, and one for the right singular vector. In Lanczos bidiagonalization, for example, such search spaces are constructed. In Chapter 3, we have proposed a Jacobi–Davidson type method for the singular value problem, where solutions to certain correction equations are used to expand the search spaces.

As noted in the previous chapter, the standard Galerkin subspace extraction works well for the computation of large singular triples, but may lead to unsatisfactory approximations to small and interior triples. To overcome this problem for the smallest triples, we propose three harmonic and a refined approach. All methods are derived in a number of different ways. Two of these methods can also be applied when we are interested in interior singular triples. Theoretical results as well as numerical experiments indicate that the results of the alternative extraction processes are often better than the standard approach. We show that when Lanczos bidiagonalization is used to approximate the smallest singular triples, the standard, harmonic, and refined extraction methods are essentially equivalent. This gives more insight in the success of the use of Lanczos bidiagonalization to find the smallest singular triples.

Finally, we present a novel method for the least squares problem, the success of which is based on a good extraction process for the smallest singular triples. The truncated SVD is also discussed in this context.

Key words: SVD, singular value problem, subspace method, subspace extraction, two-sided approach, harmonic extraction, refined extraction, Rayleigh quotient, Lanczos bidiagonalization, Saad's theorem, least squares problem, truncated SVD.

AMS subject classification: 65F15, 65F50, (65F35, 93E24).

4.1 Introduction

We study subspace methods for the computation of some singular triples (i.e., singular values and their corresponding singular left and right vectors) for large sparse matrices. The methods we consider are *two-sided*, i.e., they work with two search spaces: a search space $\mathcal{U}$ for the left singular vector, and a search space $\mathcal{V}$ for the right singular vector.

A well-known example of a two-sided subspace method for the singular value problem is the (Golub–Kahan–)Lanczos bidiagonalization ([28], see also [31, p. 495]). Chapter 3 has introduced a new method, JDSVD, based on Jacobi–Davidson type expansion techniques. Besides a new subspace expansion process, some new nonstandard extraction techniques have been suggested. In this chapter we analyze these, and other alternative extractions in more detail.

Let us first introduce some notations. Let A be a real (large sparse) $m \times n$ matrix with singular value decomposition (SVD) $A = X\Sigma Y^T$ and singular values

$$0 \leq \sigma_{\min} = \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_2 \leq \sigma_1 = \sigma_{\max},$$

where $p := \min\{m, n\}$. The assumption that A is real is made for convenience only, the adaptations for complex A are not difficult. Denote the left and right singular vectors by x_j $(1 \leq j \leq m)$ and y_j $(1 \leq j \leq n)$, respectively. For later use, we also introduce a second labeling for the singular values:

$$\sigma_{-1} \leq \sigma_{-2} \leq \cdots \leq \sigma_{-p+1} \leq \sigma_{-p}, \tag{4.1.1}$$

so that, if $\sigma_{\min} > 0$,

$$\sigma_{-j}(A) = \sigma_j^{-1}(A^+),$$

where A^+ is the pseudoinverse of A. (Such a labeling is also used for eigenvalues in [61]; one of its benefits is the matrix size independency of the indices of the smallest values. This facilitates the formulation of results for those values.) By $\|\cdot\|$ we denote the Euclidean norm, while $\kappa(A)$ is the condition number of A. For a subspace $\mathcal{U}$, let $P_{\mathcal{U}}$ denote the orthogonal projection onto $\mathcal{U}$; U denotes a "*search matrix*" whose columns form an orthonormal basis for $\mathcal{U}$. We write $\mathcal{N}(A)$ for the nullspace of A, and e_j for the jth canonical vector. For a positive definite matrix B, the B-inner product is defined by

$$(x, y)_B := y^T B x.$$

We denote the situation where $(x, y)_B = 0$ as $x \perp_B y$.

This chapter has been organized as follows. Section 4.2 recalls the standard subspace extraction and some of its properties from Chapter 3, and presents a theorem like Saad's theorem on Rayleigh–Ritz approximations, that gives insight in the strength and weakness of the standard extraction. Section 4.3 explores three variants of harmonic extraction, based on certain Galerkin conditions on the inverse of the matrix. In Section 4.4, we propose a refined extraction method. Section 4.5 discusses the Rayleigh quotient for the singular value problem. We discuss the possibilities for interior singular values in Section 4.6, and those for nonsquare or singular matrices in Section 4.7. In

Section 4.8, we study the extraction methods for the special case of the Lanczos bidiagonalization, and show that all extraction methods for the smallest singular triples are essentially equivalent. The application of the methods to the least squares problem and the truncated SVD is the subject of Section 4.9. Numerical experiments are presented in Section 4.10, and some conclusions are collected in Section 4.11.

4.2 Standard extraction

We first repeat part of the setting of Chapter 3. Given a left search space $\mathcal{U}$ and a right search space $\mathcal{V}$, we would like to determine an approximate left singular vector $u \in \mathcal{U}$ and an approximate right singular vector $v \in \mathcal{V}$. Throughout this chapter, we assume that both AV and A^TU are of full rank. (When they are not, the search spaces $\mathcal{U}$ or $\mathcal{V}$ contain a singular vector corresponding to the singular value 0.)

Given (possibly different) approximations θ and η to the same singular value, and approximate left and right singular vectors u and v, the *residual* r is defined as

$$r = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} Av - \theta u \\ A^Tu - \eta v \end{bmatrix}.$$

These approximate values and vectors are determined by a *double Galerkin condition* as follows. Write $u = Uc$ and $v = Vd$, where it is understood that $c, d \neq 0$. The *standard extraction* process is now derived from one of the two following equivalent conditions:

$$\text{(i)} \quad \begin{cases} AVd - \theta Uc & \perp & \mathcal{U}, \\ A^TUc - \eta Vd & \perp & \mathcal{V}. \end{cases}$$

$$\text{(ii)} \quad \begin{cases} U^TAVd & = & \theta c, \\ V^TA^TUc & = & \eta d. \end{cases}$$

Here, (i) can be regarded as the standard Galerkin conditions for the singular value problem; the word "standard" reflects the fact that we choose the test spaces (on the right-hand side of (i)) equal to the search spaces. Choosing the scaling $\|c\| = \|d\| = 1$, we see from (ii) that c and d are left and right singular vectors of

$$H := U^TAV$$

with singular value $\theta = \eta$; stated differently: $u = Uc$ and $v = Vd$ are left and right singular vectors of A with singular value $\theta = \eta$ with respect to the subspaces $\mathcal{U}$ and $\mathcal{V}$. Note that in this extraction process we have $\theta = \eta$, for some other methods in Section 4.3 and 4.4 this will not be the case.

Recall from Chapter 3 that this H is optimal in the sense that it minimizes the *residual matrices*

$$R_1(K) := AV - UK \quad \text{and} \quad R_2(L) := A^TU - VL, \tag{4.2.1}$$

and that the singular values $\theta_k^{(k)} \leq \cdots \leq \theta_1^{(k)}$ of $H_k := U_k^TAV_k$ converge monotonically to the singular values of A. However, the smallest singular values of H_k may converge very

irregularly to the smallest singular values of A. If the norms of the residual matrices $R_1(H)$ and $R_2(H^T)$ are small, then the largest singular values of H must be good approximations to the largest singular values of A (Theorem 3.4.7). The smallest singular values of H are not necessarily good approximations to the smallest singular values of A. So H will tend to be a better approximation to the "top" of the singular spectrum of A than to the "bottom". See also Section 4.9.2.

Now, we present another result that sheds more light on the standard extraction: a theorem that expresses the quality of the approximate singular vectors produced by this extraction in terms of the quality of the search spaces. For the Hermitian eigenvalue problem, such a result has been proved by Saad (see [69, p. 136]). This result can be extended to non-Hermitian matrices ([82, p. 286]). We first give a new short proof of Saad's result in terms of orthogonal projections, before we prove a similar result for the singular value problem along the same lines.

Let P_u denote the orthogonal projection onto span(u). Note that because $u \in \mathcal{U}$, the projections satisfy

$$P_{\mathcal{U}} P_u = P_u P_{\mathcal{U}} = P_u.$$

Suppose that we have a search space $\mathcal{U}$ for the Hermitian eigenproblem. The Rayleigh–Ritz process for the eigenvalue problem $Bx = \lambda x$ (see, for instance, [61]) ensures that

$$P_{\mathcal{U}} B P_{\mathcal{U}} P_u = \theta P_u. \tag{4.2.2}$$

Theorem 4.2.1 (Saad, [69, p. 136]) *Suppose that B is a Hermitian matrix with eigenpair (λ, x). Let (θ, u) be the Ritz pair (with respect to the search space $\mathcal{U}$), for which θ is the Ritz value closest to λ. Then*

$$\sin(u, x) \leq \sqrt{1 + \frac{\gamma^2}{\delta^2}}\, \sin(\mathcal{U}, x),$$

where

$$\begin{aligned} \gamma &= \|P_{\mathcal{U}}(B - \lambda I)(I - P_{\mathcal{U}})\|, \\ \delta &= \min_{\theta_j \neq \theta} |\theta_j - \lambda|, \end{aligned}$$

where θ_j ranges over all Ritz values not equal to θ.

Proof: We start with

$$x = P_u x + (P_{\mathcal{U}} - P_u)x + (I - P_{\mathcal{U}})x.$$

Apply $P_{\mathcal{U}}(B - \lambda I)$ on both sides, and use (4.2.2) to get

$$0 = (\theta - \lambda)P_u x + P_{\mathcal{U}}(B - \lambda I)(P_{\mathcal{U}} - P_u)x + P_{\mathcal{U}}(B - \lambda I)(I - P_{\mathcal{U}})x, \tag{4.2.3}$$

so

$$\begin{aligned} -P_{\mathcal{U}}(B - \lambda I)(I - P_{\mathcal{U}})x &= (P_{\mathcal{U}} - P_u)(B - \lambda I)(P_{\mathcal{U}} - P_u)x \\ &\quad + P_u\left((\theta - \lambda) + (B - \lambda I)(P_{\mathcal{U}} - P_u)\right)x. \end{aligned}$$

Taking the square of the norms and using Pythagoras' theorem leads to

$$\delta^2 \|(P_{\mathcal{U}} - P_u)x\|^2 \leq \|(P_{\mathcal{U}} - P_u)(B - \lambda I)(P_{\mathcal{U}} - P_u)x\|^2 \leq \gamma^2 \|(I - P_{\mathcal{U}})x\|^2.$$

Since $\|(I - P_{\mathcal{U}})x\| = \sin(\mathcal{U}, x)$ and $\|(I - P_u)x\| = \sin(u, x)$, the result now follows from

$$\|(I - P_u)x\|^2 = \|(I - P_{\mathcal{U}})x\|^2 + \|(P_{\mathcal{U}} - P_u)x\|^2.$$

□

For the singular value problem, the standard extraction gives

$$P_{\mathcal{U}} A P_{\mathcal{V}} P_v = \theta P_u \quad \text{and} \quad P_{\mathcal{V}} A^T P_{\mathcal{U}} P_u = \theta P_v.$$

We are now in a position to prove a similar result for the standard extraction for the singular value problem.

Theorem 4.2.2 (cf. Theorem 4.2.1) *Let (σ, x, y) be a singular triple of A, and (θ, u, v) be the approximate triple (derived with the standard extraction with respect to the search spaces $\mathcal{U}$ and $\mathcal{V}$), for which θ is the value closest to σ. Then*

$$\max\{\sin(u, x), \sin(v, y)\} \leq \sqrt{1 + 2\frac{\widetilde{\gamma}^2}{\widetilde{\delta}^2}}\, \max\{\sin(\mathcal{U}, x), \sin(\mathcal{V}, y)\},$$

where

$$\begin{aligned} \widetilde{\gamma} &= \max\{\|P_{\mathcal{U}} A(I - P_{\mathcal{V}})\|, \|(I - P_{\mathcal{U}})AP_{\mathcal{V}})\|\}, \\ \widetilde{\delta} &= \begin{cases} \min_{\theta_j \neq \theta} |\theta_j - \sigma| & \text{when } H \text{ is square,} \\ \min(\min_{\theta_j \neq \theta} |\theta_j - \sigma|, \sigma) & \text{when } H \text{ is nonsquare,} \end{cases} \end{aligned}$$

where θ_j ranges over all approximate singular values of H not equal to θ.

Proof: The proof follows the same line as the proof of Theorem 4.2.1, where we take for B the augmented matrix (3.1.1), and make the following other substitutions in the proof of Theorem 4.2.1: replace

$$x \text{ by } \begin{bmatrix} x \\ y \end{bmatrix}, \qquad P_{\mathcal{U}} \text{ by } \begin{bmatrix} P_{\mathcal{U}} & 0 \\ 0 & P_{\mathcal{V}} \end{bmatrix}, \quad \text{and} \quad P_u \text{ by } \begin{bmatrix} P_u & 0 \\ 0 & P_v \end{bmatrix}.$$

One may check that we get (cf. (4.2.3))

$$\begin{aligned} 0 = (\theta - \sigma) \begin{bmatrix} P_u x \\ P_v y \end{bmatrix} &+ \begin{bmatrix} P_{\mathcal{U}} & 0 \\ 0 & P_{\mathcal{V}} \end{bmatrix} \begin{bmatrix} -\sigma I & A \\ A^T & -\sigma I \end{bmatrix} \begin{bmatrix} P_{\mathcal{U}} - P_u & 0 \\ 0 & P_{\mathcal{V}} - P_v \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &+ \begin{bmatrix} P_{\mathcal{U}} & 0 \\ 0 & P_{\mathcal{V}} \end{bmatrix} \begin{bmatrix} -\sigma I & A \\ A^T & -\sigma I \end{bmatrix} \begin{bmatrix} I - P_{\mathcal{U}} & 0 \\ 0 & I - P_{\mathcal{V}} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \end{aligned}$$

Since the last term on the right-hand side can be written as

$$\begin{bmatrix} 0 & P_{\mathcal{U}}A(I-P_{\mathcal{V}}) \\ (I-P_{\mathcal{U}})A^TP_{\mathcal{V}} & 0 \end{bmatrix} \begin{bmatrix} (I-P_{\mathcal{U}})x \\ (I-P_{\mathcal{V}})y \end{bmatrix},$$

the norm of this term is bounded by

$$\sqrt{2}\,\widetilde{\gamma}\max\{\|(I-P_{\mathcal{U}})x\|, \|(I-P_{\mathcal{V}})y\|\}.$$

Furthermore, the smallest singular value of

$$\begin{bmatrix} P_{\mathcal{U}}-P_u & 0 \\ 0 & P_{\mathcal{V}}-P_v \end{bmatrix} \begin{bmatrix} -\sigma I & A \\ A^T & -\sigma I \end{bmatrix} \begin{bmatrix} P_{\mathcal{U}}-P_u & 0 \\ 0 & P_{\mathcal{V}}-P_v \end{bmatrix}$$

is $\widetilde{\delta}$. The result now follows in the same way as in the proof of Theorem 4.2.1. □

The theorem states that when $\angle(\mathcal{U},x) \to 0$ and $\angle(\mathcal{V},y) \to 0$, the standard extraction gives good Ritz vectors, unless the singular values of H are poorly separated ($\widetilde{\delta} \approx 0$). In fact, $\widetilde{\delta}$ becomes 0 in case of double singular values of H, see also Example 4.7.1. The phenomenon of poorly separated Ritz values is also encountered in the Rayleigh–Ritz method for eigenvalue problem; it is often not very serious: we may just continue with the subspace method (by expanding the search spaces), at the next step the singular values of H may be well separated.

A more serious problem of the standard extraction is that the theorem does not predict *which* singular triple is the best: it is a problem of *selection*. Suppose that $u = \sum_{j=1}^{m} \gamma_j x_j$ and $v = \sum_{j=1}^{n} \delta_j y_j$ are approximate singular vectors of unit length; then $\theta = u^TAv = \sum_{j=1}^{p} \gamma_j\delta_j\sigma_j$. (We may assume θ is nonnegative; otherwise, take $-u$ instead of u.) Now suppose that $\theta \approx \sigma_1$, in the sense that $\sigma_2 < \theta < \sigma_1$, and that $\sigma_1 - \theta$ is (much) smaller than $\theta - \sigma_2$. Then we conclude that $\gamma_1 \approx 1$ and $\delta_1 \approx 1$, so u and v are good approximations to x_1 and y_1. But when $\theta \approx \sigma_p$, u and v are not necessarily good approximations to x_p and y_p. For example, u could have a large component of x_{p-1} and a small component of x_1, and v could have a large component of y_{p-2} and a small component of y_1. In conclusion, when we search for the largest singular value, it is asymptotically safe to select the largest singular triple of H, but for the smallest triple it is not safe to select the smallest approximate triple. See also Example 4.7.1.

Failure to select the best approximate vectors is especially dangerous when we use *restarts*. At the moment of restart, selection of bad approximate singular vectors may spoil the whole process.

4.3 Harmonic extractions

As seen in the previous section, the standard extraction process is satisfactory in the quest for the largest singular values, but the approximations to the smallest singular values often display a irregular convergence.

> Here and in the next section we assume that A is nonsingular (which implies that A is square);

we will treat the general case in Section 4.7. Based on the observation that the smallest singular values of A are the largest ones of A^{-1}, it was briefly suggested in Section 3.4.3 to consider modified Galerkin conditions on A^{-T} and A^{-1}. Here we work out this idea in detail.

The following are equivalent:

$$\text{(i)} \quad \begin{cases} A^{-T}V\widetilde{d} - \widetilde{\eta}^{-1}U\widetilde{c} & \perp & \widetilde{\mathcal{U}}, \\ A^{-1}U\widetilde{c} - \widetilde{\theta}^{-1}V\widetilde{d} & \perp & \widetilde{\mathcal{V}}. \end{cases}$$

$$\text{(ii)} \quad \begin{cases} \widetilde{\eta}\widetilde{U}^T A^{-T}V\widetilde{d} & = & \widetilde{U}^T U\widetilde{c}, \\ \widetilde{\theta}\widetilde{V}^T A^{-1}U\widetilde{c} & = & \widetilde{V}^T V\widetilde{d}. \end{cases}$$

The idea is now to choose the test spaces $\widetilde{\mathcal{U}}$ and $\widetilde{\mathcal{V}}$ in such a way, that we do not have to work with the inverse of the (large sparse) matrices A and A^T. Some terminology: in line with nomenclature for the Rayleigh–Ritz procedure, $\widetilde{u} := U\widetilde{c}$ and $\widetilde{v} := V\widetilde{d}$ are called *left and right harmonic singular vectors*, $\widetilde{\theta}$ and $\widetilde{\eta}$ *harmonic singular values*, and $(\widetilde{\theta}, \widetilde{\eta}, \widetilde{u}, \widetilde{v})$ a *harmonic singular tuple*. We remark that, similar to the harmonic Ritz values for the eigenvalue problem, the harmonic Ritz values may be ∞, see Example 4.7.1, and Section 4.5 for a way to overcome this difficulty.

Now we can make the following four choices for $(\widetilde{\mathcal{U}}, \widetilde{\mathcal{V}})$:

- $(A\mathcal{V}, A^T\mathcal{U})$ gives the standard extraction of Section 4.2.
- $(AA^T\mathcal{U}, A^T\mathcal{U})$ leads to the *$\mathcal{U}$-harmonic extraction*, see Section 4.3.1.
- $(A\mathcal{V}, A^TA\mathcal{V})$ is the *$\mathcal{V}$-harmonic extraction* to be discussed in Section 4.3.1.
- $(AA^T\mathcal{U}, A^TA\mathcal{V})$ gives the *double harmonic extraction*, examined in Section 4.3.2.

4.3.1 $\mathcal{U}$-harmonic and $\mathcal{V}$-harmonic extraction

In this subsection we only treat the $\mathcal{V}$-harmonic extraction, the $\mathcal{U}$-harmonic extraction is derived by interchanging the roles of $\mathcal{U}$ and $\mathcal{V}$, and those of A and A^T. The following are equivalent:

$$\text{(i)} \quad \begin{cases} A^{-T}V\widetilde{d} - \widetilde{\eta}^{-1}U\widetilde{c} & \perp & A\mathcal{V}, \\ A^{-1}U\widetilde{c} - \widetilde{\theta}^{-1}V\widetilde{d} & \perp & A^TA\mathcal{V}. \end{cases}$$

$$\text{(ii)} \quad \begin{cases} AV\widetilde{d} - \widetilde{\theta}U\widetilde{c} & \perp & A\mathcal{V}, \\ A^TU\widetilde{c} - \widetilde{\eta}V\widetilde{d} & \perp & \mathcal{V}. \end{cases}$$

$$\text{(iii)} \quad \begin{cases} V^TA^TAV\widetilde{d} & = & \widetilde{\theta}V^TA^TU\widetilde{c}, \\ V^TA^TU\widetilde{c} & = & \widetilde{\eta}\widetilde{d}. \end{cases}$$

Here, (i) expresses that the $\mathcal{V}$-harmonic method arises from Galerkin conditions on A^{-1} and A^{-T} with respect to modified test spaces. Item (ii) gives a derivation in terms of Galerkin conditions on A and A^T, but with different test spaces, compared with the standard extraction of Section 4.2. From (iii), we see that $V^TA^TAV\widetilde{d} = (\widetilde{\theta\eta})\widetilde{d}$, that is, $(\widetilde{\theta\eta}, \widetilde{v})$ is a Ritz pair of A^TA with respect to the search space $\mathcal{V}$. This is just the well-known Raleigh–Ritz approach on A^TA. The "secret", however, is in the formation of the vector $\widetilde{c}$, and hence the approximate left singular vector $\widetilde{u}$. We have $\widetilde{c} = H^{-T}\widetilde{d}$, up to scaling (see Section 4.7 for the case that H is nonsquare or singular). Since H^{-T} can be considered as a projected A^{-T}, this suggests that the vector $\widetilde{u} = UH^{-T}\widetilde{d}$ may be a much better approximation to the left singular vector than the "usual" approximation $A\widetilde{v} = AV\widetilde{d}$. The following lemma substantiates this.

Lemma 4.3.1 *Suppose that v is an approximation to the "smallest" right singular vector $y_{\min}$. Then, denoting $\varepsilon := \tan(v, y_{\min})$, we have*

$$\kappa(A)^{-1}\varepsilon \leq \tan(A^{-T}v, x_{\min}) \leq \varepsilon \leq \tan(Av, x_{\min}) \leq \kappa(A)\varepsilon,$$

where the inequalities are sharp.

Proof: Write $v = y_{\min} + \varepsilon e$, where $e \perp y_{\min}$ and $\|e\| = 1$. Note that $\tan(v, y_{\min}) = \varepsilon$. Then for $Av = \sigma_{\min}x_{\min} + \varepsilon Ae$, we have $\tan(Av, x_{\min}) = \varepsilon\sigma_{\min}^{-1}\|Ae\|$, from which the last two inequalities follow (sharp if $e = y_{p-1}$ and $\sigma_{\min} = \sigma_{p-1}$, respectively $e = y_{\max}$). Since $A^{-T}v = \sigma_{\min}^{-1}x_{\min} + \varepsilon A^{-T}e$, we have $\tan(A^{-T}v, x_{\min}) = \varepsilon\sigma_{\min}\|A^{-T}e\|$, from which we get the first two inequalities (sharp in the same circumstances as above). □

Concluding from the lemma, it would be ideal, given an approximate right vector $\widetilde{v}$, to take $A^{-T}\widetilde{v}$ as approximate left vector. In practice, the action with A^{-T} is often too expensive, but H^{-T} is a, much cheaper, projected approximation to A^{-T}. Numerical experiments (see Section 4.10) confirm that $\widetilde{u} = UH^{-T}\widetilde{d}$ may be a much more accurate than $AV\widetilde{d}$.

Since the $(\widetilde{\theta\eta})_j$ are Ritz values of A^TA with respect to the subspace $\mathcal{V}$, we can invoke well-known results. We label the values in two different ways (cf. (4.1.1)):

$$(\widetilde{\theta\eta})_k \leq \cdots \leq (\widetilde{\theta\eta})_1 \quad \text{and} \quad (\widetilde{\theta\eta})_{-1} \leq \cdots \leq (\widetilde{\theta\eta})_{-k}.$$

Theorem 4.3.2 *Let $R_V = A^TAV - V(V^TA^TAV) = (I - VV^T)A^TAV$. Then:*

(a) *for fixed $j \leq p$, and $k \geq j$, the $((\theta\eta)_j^{(k)})^2$ converge monotonically (up)to σ_j^2:*

$$((\theta\eta)_j^{(k)})^2 \leq ((\theta\eta)_j^{(k+1)})^2 \leq \sigma_j^2;$$

the $((\theta\eta)_{-j}^{(k)})^2$ converge monotonically (down)to σ_{-j}^2:

$$\sigma_{-j}^2 \leq ((\theta\eta)_{-j}^{(k+1)})^2 \leq ((\theta\eta)_{-j}^{(k)})^2;$$

(b) *for each* $j = 1, \ldots, k$, *there exist singular values* $\sigma_{j'}$ *of* A *which can be put in one-one correspondence with the* $(\theta\eta)_j$ *in such a way that*

$$|(\widetilde{\theta\eta})_j - \sigma_{j'}^2| \leq \|R_V\| \quad \textit{and} \quad \sum_{j=1}^{k} ((\widetilde{\theta\eta})_j - \sigma_{j'}^2)^2 \leq \|R_V\|_F^2;$$

(c) *let* $(\theta\eta, v)$ *be the Ritz pair of* A^TA *where* $\theta\eta$ *is the Ritz value closest to* σ^2. *Then*

$$\sin(v, y) \leq \sqrt{1 + \frac{\gamma_V^2}{\delta_V^2}}\, \sin(\mathcal{V}, y),$$

where

$$\begin{aligned} \gamma_V &= \|P_{\mathcal{V}}(A^TA - \sigma^2 I)(I - P_{\mathcal{V}})\|, \\ \delta_V &= \min_{(\theta\eta)_j \neq \theta\eta} |(\theta\eta)_j - \sigma^2|. \end{aligned}$$

Proof: For (a) and (b), apply [61, Theorems 11.5.1 and 11.5.2] to A^TA and AA^T. Part (c) is a corollary to Theorem 4.2.1, when we take $B = A^TA$. □

4.3.2 Double harmonic extraction

We now give a number of possible derivations for the double harmonic extraction. The following are equivalent:

(i) $$\begin{cases} A^{-T}V\widetilde{d} - \widetilde{\eta}^{-1}U\widetilde{c} & \perp & AA^T\mathcal{U}, \\ A^{-1}U\widetilde{c} - \widetilde{\theta}^{-1}V\widetilde{d} & \perp & A^TA\mathcal{V}. \end{cases}$$

(ii) $$\begin{cases} A^{-T}V\widetilde{d} - \widetilde{\eta}^{-1}U\widetilde{c} & \perp_{AA^T} & \mathcal{U}, \\ A^{-1}U\widetilde{c} - \widetilde{\theta}^{-1}V\widetilde{d} & \perp_{A^TA} & \mathcal{V}. \end{cases}$$

(iii) $$\begin{cases} A^{-T}\widetilde{V}\widetilde{c} - \widetilde{\theta}^{-1}\widetilde{U}\widetilde{d} & \perp & \widetilde{\mathcal{U}} = A\mathcal{V}, \\ A^{-1}\widetilde{U}\widetilde{d} - \widetilde{\eta}^{-1}\widetilde{V}\widetilde{c} & \perp & \widetilde{\mathcal{V}} = A^T\mathcal{U}. \end{cases}$$

(iv) $$\begin{cases} AV\widetilde{d} - \widetilde{\theta}U\widetilde{c} & \perp & A\mathcal{V}, \\ A^TU\widetilde{c} - \widetilde{\eta}V\widetilde{d} & \perp & A^T\mathcal{U}. \end{cases}$$

(v) $$\begin{cases} \widetilde{\eta}U^TAV\widetilde{d} & = & U^TAA^TU\widetilde{c}, \\ \widetilde{\theta}V^TA^TU\widetilde{c} & = & V^TA^TAV\widetilde{d}. \end{cases}$$

Here, (i) states that the double harmonic method arises from Galerkin conditions on A^{-1} and A^{-T} with respect to modified test spaces. Item (ii) formulates the result with respect to the standard search and test spaces, but with respect to a different inner product;

note that both $(\cdot,\cdot)_{A^TA}$ and $(\cdot,\cdot)_{AA^T}$ are inner products because of the assumption that A is nonsingular. Item (iii) derives the approach from the situation where we take modified search and test spaces; as in the standard extraction, the test spaces are equal to the search spaces. This notation in item is different from the other items (exchange of $\widetilde{c}$ and $\widetilde{d}$, and of $\widetilde{\theta}$ and $\widetilde{\eta}$). The reason for this is the following. The vectors $\widetilde{u} = \widetilde{U}\widetilde{d} = AV\widetilde{d}$ and $\widetilde{v} = \widetilde{V}\widetilde{d} = A^TU\widetilde{c}$ will tend to be deficient in the direction of the smallest singular vectors. So it is better to do a "free" step of inverse iteration and take $\widetilde{u} = A^{-T}(A^TU\widetilde{c}) = U\widetilde{c}$ and $\widetilde{v} = A^{-1}(AV\widetilde{d}) = V\widetilde{d}$ as approximate singular vectors. Similar remarks for the harmonic vectors in the eigenvalue problem can be found in [54]. Item (iv) gives a derivation in terms of Galerkin conditions on A and A^T, but with respect to different test spaces. Finally, (v) can be interpreted as: $(\widetilde{\theta}^{-1} = \widetilde{\eta}^{-1}, \widetilde{u}, \widetilde{v})$ is a singular tuple of U^TAV, where A^TU and AV are orthogonal, as the following analysis shows.

We introduce the QR decompositions

$$AV = Q_U G_U \quad \text{and} \quad A^TU = Q_V^T G_V.$$

(We choose for the letter "G", because the letter "R" is already "overloaded", and because $G_U^TG_U$ and $G_V^TG_V$ are Choleski decompositions of V^TA^TAV and U^TAA^TU, respectively. We choose for these subscripts since AV "lives in the U-space", i.e., A maps right to left singular vectors. Similarly, A^T maps left to right singular vectors.) Note that G_U and G_V are nonsingular because of the assumption that AV and A^TU are of full rank. Then

$$Q_U = AVG_U^{-1} \quad \text{and} \quad Q_V = A^TUG_V^{-1}$$

are orthogonal and span $A\mathcal{V}$ and $A^T\mathcal{U}$, respectively. Then characterization (v) can be written as

$$\begin{cases} \widetilde{\eta}G_V^{-T}U^TAVG_U^{-1}(G_U\widetilde{d}) &= (G_V\widetilde{c}), \\ \widetilde{\theta}G_U^{-T}V^TA^TUG_V^{-1}(G_V\widetilde{c}) &= (G_U\widetilde{d}). \end{cases}$$

When we normalize $\widetilde{c}$ and $\widetilde{d}$ such that $\|G_V\widetilde{c}\| = \|G_U\widetilde{d}\| = 1$, we see that $G_V\widetilde{c}$ and $G_U\widetilde{d}$ are left and right singular vectors of

$$\widetilde{H}^{-1} := Q_V^TA^{-1}Q_U = G_V^{-T}U^TAVG_U^{-1} = G_V^{-T}HG_U^{-1}$$

corresponding to singular value $\widetilde{\theta}^{-1} = \widetilde{\eta}^{-1}$. Analogously to (4.2.1), we define the *residual matrices*

$$\widetilde{R}_1(K) := A^{-1}Q_U - Q_VK, \qquad \widetilde{R}_2(L) := A^{-T}Q_V - Q_UL.$$

Then

$$\begin{aligned} \widetilde{R}_1(\widetilde{H}^{-1}) &= (I - Q_VQ_V^T)A^{-1}Q_U &&= (I - Q_VQ_V^T)VG_U^{-1}, \\ \widetilde{R}_2(\widetilde{H}^{-T}) &= (I - Q_UQ_U^T)A^{-T}Q_V &&= (I - Q_UQ_U^T)UG_V^{-1}. \end{aligned}$$

Note that a multiplication by A^{-1} or A^{-T} is not necessary to compute the residual matrices (which, in practice, will not be done anyway). The following theorem can be proved applying Theorem 3.4.2 to A^{-1} instead of A. Informally, it states that $\widetilde{H}^{-1}$ can be considered as the best approximation to A^{-1} over $A\mathcal{V}$ and $A^T\mathcal{U}$.

Theorem 4.3.3 (cf. Theorem 3.4.2) *For given $m \times k$ matrix Q_U and $n \times k$ matrix Q_V with orthogonal columns, let $\widetilde{H}^{-1} = Q_V^T A^{-1} Q_U$.*

(a) *For all $k \times k$ matrices K we have $\|\widetilde{R}_1(\widetilde{H}^{-1})\| \leq \|\widetilde{R}_1(K)\|$. Moreover, $\widetilde{H}^{-1}$ is unique with respect to the Frobenius norm: $\|\widetilde{R}_1(\widetilde{H}^{-1})\|_F \leq \|\widetilde{R}_1(K)\|_F$ with equality only when $K = \widetilde{H}^{-1}$.*

(b) *$\widetilde{H}^{-T}$ minimizes the norm of $\widetilde{R}_2(L)$, and $\widetilde{H}^{-T}$ is unique with respect to the Frobenius norm.*

Since $\widetilde{H}^{-1}$ approximates A^{-1}, we may take the singular values $\widetilde{\theta}_{-1}^{(k)} \leq \cdots \leq \widetilde{\theta}_{-k}^{(k)}$ (notation: cf. (4.1.1)) of

$$\widetilde{H}_k = G_U H_k^{-1} G_V^T \tag{4.3.1}$$

as approximations to the $\sigma_j(A^{-1}) = \sigma_{-j}(A)$. By applying Theorems 3.4.5 and 3.4.7 to A^{-1}, we get the following result, which states that all $\widetilde{\theta}_{-j}^{(k)}$ converge monotonically decreasing to the σ_{-j}.

Theorem 4.3.4 (cf. Theorems 3.4.5 and 3.4.7)

(a) *For fixed $j \leq p$, and $k \geq j$, the $\widetilde{\theta}_{-j}^{(k)}$ converge monotonically (down)to σ_{-j}:*

$$\sigma_{-j} \leq \widetilde{\theta}_{-j}^{(k+1)} \leq \widetilde{\theta}_{-j}^{(k)}.$$

(b) *For each $j = 1, \ldots, k$, there exist singular values $\sigma_{-j'}$ of A which can be put in one-one correspondence with the singular values $\widetilde{\theta}_{-j}$ of $\widetilde{H}$ in such a way that*

$$|\sigma_{-j'}^{-1} - \widetilde{\theta}_{-j}^{-1}| \leq \max\{\|\widetilde{R}_1(\widetilde{H}^{-1})\|, \|\widetilde{R}_2(\widetilde{H}^{-T})\|\}.$$

Moreover,

$$\sum_{j=1}^{k} (\sigma_{-j'}^{-1} - \widetilde{\theta}_{-j}^{-1})^2 \leq \|\widetilde{R}_1(\widetilde{H}^{-1})\|_F^2 + \|\widetilde{R}_2(\widetilde{H}^{-T})\|_F^2.$$

The previous two theorems indicate that the double harmonic approach indeed has favorable properties in the quest for the smallest singular triples. For instance, if the norms of the residual matrices $\widetilde{R}_1(\widetilde{H}^{-1})$ and $\widetilde{R}_2(\widetilde{H}^{-T})$ are small, then, since

$$|\sigma_{-j'}^{-1} - \widetilde{\theta}_{-j}^{-1}| = \frac{|\sigma_{-j'} - \widetilde{\theta}_{-j}|}{\sigma_{-j'}\widetilde{\theta}_{-j}},$$

the smallest singular values of $\widetilde{H}$ must be good approximations to the small singular values of A. So while H tends to approximate A well with respect to the largest singular values, $\widetilde{H}$ tends to approximate A well with respect to the smallest singular values. See also Section 4.9.2.

4.4 Refined extraction

In this section we introduce another extraction process. The key idea is that the "minimal" left and right singular vector of A minimize $\|A^T x\|$ and $\|Ay\|$, respectively. When we have search spaces $\mathcal{U}$ and $\mathcal{V}$, it is a natural idea to extract those vectors that minimize the norm of the matrix and its transpose over the search spaces. This approach amounts to computing two SVDs of tall skinny matrices AV and A^TU. It is somewhat similar to the refined extraction in the Ritz method in the eigenvalue problem (see, for instance [43] and [82, p. 289]), in the sense that we look for the minimal singular vector in a search space; therefore we choose the name *refined extraction.*

With $\widehat{u} = U\widehat{c}$ and $\widehat{v} = V\widehat{d}$, the following are equivalent (remind that we assume in this section that A is nonsingular):

(i) $$\begin{cases} \widehat{\theta} &= \min\limits_{\substack{u \in \mathcal{U} \\ \|u\| = 1}} \|A^T u\| \\ \widehat{\eta} &= \min\limits_{\substack{v \in \mathcal{V} \\ \|v\| = 1}} \|Av\| \end{cases}$$ with $\widehat{u}, \widehat{v}$, respectively, as minimizing argument,

(ii) $$\begin{cases} \left(A^TA\right)^{-1}\widehat{V}\widehat{c} - \widehat{\theta}^{-2}\widehat{V}\widehat{c} \;\perp\; \widehat{\mathcal{V}} := A^T\mathcal{U} \\ \left(AA^T\right)^{-1}\widehat{U}\widehat{d} - \widehat{\eta}^{-2}\widehat{U}\widehat{d} \;\perp\; \widehat{\mathcal{U}} := A\mathcal{V} \end{cases}$$ where $\widehat{\theta}^{-2}$ and $\widehat{\eta}^{-2}$ are maximal,

(iii) $$\begin{cases} U^TAA^TU\widehat{c} &= \widehat{\theta}^2\widehat{c} \\ V^TA^TAV\widehat{d} &= \widehat{\eta}^2\widehat{d} \end{cases}$$ where $\widehat{\theta}^2$ and $\widehat{\eta}^2$ are minimal.

Here, (i) expresses that $\widehat{u}$ and $\widehat{u}$ minimize the matrix norm over $\mathcal{U}$ and $\mathcal{V}$. Item (ii) gives a derivation in terms of a Galerkin condition on $\left(A^TA\right)^{-1}$ and $\left(AA^T\right)^{-1}$. Item (iii) states that $\widehat{u}$ and $\widehat{v}$ are the "smallest" Ritz vectors of AA^T and A^TA with respect to $\mathcal{U}$ and $\mathcal{V}$, respectively.

We call $\widehat{\theta}$ and $\widehat{\eta}$ *refined singular values*, $\widehat{u}$ and $\widehat{v}$ *refined singular vectors*, and $(\widehat{\theta}, \widehat{\eta}, \widehat{u}, \widehat{v})$ a *refined singular tuple.*

A difference between the refined approach and the harmonic approaches is that the first leads to only one approximate triple instead of k ones; however, this can easily be modified by computing more than just one smallest singular value and corresponding vectors of AV and A^TU, respectively. We label these values in two different ways (cf. (4.1.1)):

$$\widehat{\theta} = \widehat{\theta}_{-1} \leq \cdots \leq \widehat{\theta}_{-k} \quad \text{and} \quad \widehat{\theta} = \widehat{\theta}_k \leq \cdots \leq \widehat{\theta}_1,$$

and the $\widehat{\eta}$'s similarly. These approximations have the desirable property of monotonic convergence.

Theorem 4.4.1 *Define $R_U = (I - UU^T)AA^TU$. Then:*

(a) *for fixed $j \leq p$, and $k \geq j$, both $\widehat{\theta}_{-j}^{(k)}$ and $\widehat{\eta}_{-j}^{(k)}$ converge monotonically (down)to σ_{-j}:*

$$\sigma_{-j} \leq \widehat{\theta}_{-j}^{(k+1)} \leq \widehat{\theta}_{-j}^{(k)} \quad \text{and} \quad \sigma_{-j} \leq \widehat{\eta}_{-j}^{(k+1)} \leq \widehat{\eta}_{-j}^{(k)};$$

both $\widehat{\theta}_j^{(k)}$ and $\widehat{\eta}_j^{(k)}$ converge monotonically (up)to σ_j:

$$\widehat{\theta}_j^{(k)} \leq \widehat{\theta}_j^{(k+1)} \leq \sigma_j \quad \text{and} \quad \widehat{\eta}_j^{(k)} \leq \widehat{\eta}_j^{(k+1)} \leq \sigma_j;$$

(b) *for each* $j = 1, \ldots, k$, *there exist singular values* $\sigma_{j'}$ *and* $\sigma_{j''}$ *of* A *which can be put in one-one correspondence with the* $\widehat{\theta}_j$ *and* $\widehat{\eta}_j$ *in such a way that*

$$|\widehat{\theta}_j^2 - \sigma_{j'}^2| \leq \|R_U\| \quad \textit{and} \quad \sum_{j=1}^{k} (\widehat{\theta}_j^2 - \sigma_{j'}^2)^2 \leq \|R_U\|_F^2,$$

$$|\widehat{\eta}_j^2 - \sigma_{j''}^2| \leq \|R_V\| \quad \textit{and} \quad \sum_{j=1}^{k} (\widehat{\eta}_j^2 - \sigma_{j''}^2)^2 \leq \|R_V\|_F^2;$$

(c) $\sin(\widehat{u}, x_{\min}) \leq \sqrt{1 + \frac{\gamma_U^2}{\delta_U^2}} \sin(\mathcal{U}, x_{\min})$ *and* $\sin(\widehat{v}, y_{\min}) \leq \sqrt{1 + \frac{\gamma_V^2}{\delta_V^2}} \sin(\mathcal{V}, y_{\min})$, *where*

$$\gamma_U = \|P_{\mathcal{U}}(AA^T - \sigma_{\min}^2 I)(I - P_{\mathcal{U}})\|, \quad \gamma_V = \|P_{\mathcal{V}}(A^TA - \sigma_{\min}^2 I)(I - P_{\mathcal{V}})\|,$$
$$\delta_U = |\widehat{\theta}_{-2}^2 - \sigma_{\min}^2|, \quad \delta_V = |\widehat{\eta}_{-2}^2 - \sigma_{\min}^2|.$$

Proof: Follows from characterizations (i) or (iii), using the same techniques as in Theorem 4.3.2. □

Advantages of this refined approach are good asymptotic vector extraction ($\sigma_{\min}^2$ is an exterior eigenvalue of A^TA), and the fact that we have upper bounds for $\sigma_{\min}$. On the other hand, the left and right singular vector are approximated completely independently. It may thus happen that $u \approx x$ and $v \approx y$ are approximate vectors to singular vectors corresponding to different singular values. In this case the Rayleigh quotient of the vectors (see Section 4.5) is meaningless as approximate singular value.

Note that we can also formulate a result similar to part (c) of Theorem 4.4.1 for the largest singular vectors. This also shows that the refined approach is equally useful for the largest singular triples as for the smallest ones: just take the maximum instead of the minimum in characterization (i) and (iii).

The next theorem, that can be seen as generalization of a result in [82, p. 290], gives some idea how fast the refined approach converges to the minimal singular value. In particular, it shows that, since $\angle(\mathcal{U}, x) \to 0$ and $\angle(\mathcal{V}, y) \to 0$ as the search spaces expand, convergence is guaranteed.

Theorem 4.4.2

$$\|A\widehat{v}\| \leq \frac{\sigma_{\min} + \sin(\mathcal{V}, y_{\min})\sigma_{\max}}{\sqrt{1 - \sin^2(\mathcal{V}, y_{\min})}}, \quad \textit{and} \quad \|A^T\widehat{u}\| \leq \frac{\sigma_{\min} + \sin(\mathcal{U}, x_{\min})\sigma_{\max}}{\sqrt{1 - \sin^2(\mathcal{U}, x_{\min})}}.$$

Proof: We only prove the first statement, the proof of the second one being similar. Decompose $y_{\min} = c_V y_V + s_V f_V$, where $y_V := VV^* y_{\min} / \|VV^* y_{\min}\|$ is the orthogonal projection of $y_{\min}$ onto $\mathcal{V}$, $c_V = \cos(\mathcal{V}, y_{\min})$, and $s_V = \sin(\mathcal{V}, y_{\min})$. Since $Ay_V = (\sigma_{\min} x_{\min} - s_V A f_V)/c_V$, we have by definition of a refined singular vector

$$\|A\widehat{v}\| \leq \|Ay_V\| \leq (\sigma_{\min} + s_V \|A\|)/c_V.$$

□

The next theorem gives a justification of the new methods: they retrieve singular triples that are exactly present in $\mathcal{U}$ and $\mathcal{V}$.

Theorem 4.4.3 *Let (σ, x, y) be a singular triple of A with $x = Uc$ and $y = Vd$. Then (σ, σ, Uc, Vd) is both a harmonic and refined singular tuple.*

Proof: This can be verified by calculating the left and right-hand sides of (iii) of Section 4.3.1, (v) of Section 4.3.2, and (iii) of Section 4.4. □

The standard extraction method in principle also finds singular triples that are exactly present in the search spaces. However, it may have difficulties selecting them, see Example 4.7.1.

In the following table we summarize the properties of convergence of the different extraction methods (Theorems 3.4.5, 4.3.2, 4.3.4, 4.4.1). A "+" stands for monotonic convergence. A "–" means that we do not have monotonic convergence; as a result the convergence can in practice be irregular and very slow.

TABLE 4.1: Properties of monotonic convergence of the extraction methods to $\sigma_{\max}$ and $\sigma_{\min}$.

extraction	$\sigma_{\max}$	$\sigma_{\min}$
standard	+	–
$\mathcal{U}$-, $\mathcal{V}$-harmonic	+	+
double-harmonic	–	+
refined	+	+

The table suggests that to find the smallest singular triples, one can use all methods except the standard extraction, while for the largest singular values one can use all methods except the double-harmonic extraction. This appears to be a good rule of thumb indeed, see also Sections 4.10 and 4.11.

4.5 Rayleigh quotient for the singular value problem

In the harmonic Ritz approach for the eigenproblem, the harmonic Ritz value may be a bad approximation to the eigenvalue (it can even be ∞), while the harmonic Ritz vector may be of good quality; see, e.g., [78]. Therefore, it is advisable to discard the harmonic Ritz value and, instead, approximate the eigenvalue by the Rayleigh quotient of the harmonic Ritz vector.

In our case, we also encounter the situation that while the harmonic vectors may be good, the harmonic value can be bad, see also Example 4.7.1. Therefore, we propose to take the Rayleigh quotient (in the sense of the singular value problem), as defined in Section 3.4.2 of the left and right approximate singular vector as an approximate singular value.

One may check that we have the expressions as in Table 4.2 for the Rayleigh quotient of approximate vectors in the standard and harmonic approaches.

So for the standard and harmonic methods we can obtain the Rayleigh quotient of the approximate singular vectors at little (double-harmonic approach) or no (standard, $\mathcal{U}$-,

TABLE 4.2: Rayleigh quotients for the standard and harmonic extraction methods.

approach	Rayleigh quotient of approximate vectors
standard	$\theta = \eta$
$\mathcal{V}$-harmonic	$\widetilde{\eta}$
double-harmonic	$\frac{\|AVd\|^2}{\widetilde{\theta}} = \frac{\|A^TUc\|^2}{\widetilde{\eta}} = \frac{\|AVd\|\|A^TUc\|}{\sqrt{\widetilde{\theta}\widetilde{\eta}}}$

and $\mathcal{V}$-harmonic approaches) additional cost. Of course, we can also take the Rayleigh quotient in the refined extraction process. For complex matrices, we should scale u and v in such a way that their Rayleigh quotient is real and nonnegative.

4.6 Interior singular values

The extraction processes introduced in Section 4.3 and 4.4 are tailored for the smallest (and, with exception of the double-harmonic approach, also useful for the largest) singular triples. Now we study the situation where we are interested in interior singular values (and corresponding vectors) near a target $\tau \geq 0$. We present two methods that can be seen as generalizations of the double-harmonic approach (Section 4.3.2) and the refined approach (Section 4.4). We will see that for $\tau = 0$, the methods in this section deduce to those of Sections 4.3 and 4.4.

For a Hermitian matrix B and a search space $\mathcal{W}$, it is well known that instead of standard Raleigh–Ritz, better results may be expected from the harmonic Ritz approach (see, e.g., [82, p. 292])

$$W^T(B - \tau I)^2 Wc = (\widetilde{\theta} - \tau) W^T(B - \tau I)Wc,$$

where we are interested in the $\widetilde{\theta}$ closest to some (interior) target τ. When we take the augmented matrix (3.1.1) for B, and "split up" $\mathcal{W}$ into a left and right space $\mathcal{U}$ and $\mathcal{V}$—i.e., we take $W = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$—then we get the *harmonic extraction for target* $\tau \geq 0$ for the singular value problem: find the eigenpair(s) of the generalized symmetric eigenvalue problem

$$\begin{bmatrix} U^TAA^TU + \tau^2 I_m & -2\tau H \\ -2\tau H^T & V^TA^TAV + \tau^2 I_n \end{bmatrix} \begin{bmatrix} \widetilde{c} \\ \widetilde{d} \end{bmatrix} = (\widetilde{\theta} - \tau) \begin{bmatrix} -\tau I & H \\ H^T & -\tau I \end{bmatrix} \begin{bmatrix} \widetilde{c} \\ \widetilde{d} \end{bmatrix},$$

for which $\widetilde{\theta}$ is closest to τ. Here the matrix on the left-hand side is positive semidefinite. Restated, the problem is to find the smallest eigenpairs of

$$\begin{bmatrix} U^TAA^TU & -\tau H \\ -\tau H^T & V^TA^TAV \end{bmatrix} \begin{bmatrix} \widetilde{c} \\ \widetilde{d} \end{bmatrix} = \widetilde{\theta} \begin{bmatrix} -\tau I & H \\ H^T & -\tau I \end{bmatrix} \begin{bmatrix} \widetilde{c} \\ \widetilde{d} \end{bmatrix}.$$

One may check that for $\tau \to \infty$ we get the standard Galerkin approach of Section 4.2, while $\tau = 0$ gives the double-harmonic approach of Section 4.3.2.

A *refined method* for the approximation of interior singular triples is the minimization over $\widehat{c}, \widehat{d} \in \mathbb{R}^k$, $\|\widehat{c}\| = \|\widehat{d}\| = 1$ of

$$\left\| \begin{bmatrix} -\tau I_m & A \\ A^T & -\tau I_n \end{bmatrix} \begin{bmatrix} U\widehat{c} \\ V\widehat{d} \end{bmatrix} \right\| = \left\| \begin{bmatrix} -\tau U & AV \\ A^T U & -\tau V \end{bmatrix} \begin{bmatrix} \widehat{c} \\ \widehat{d} \end{bmatrix} \right\|,$$

which amounts to the SVD of a tall skinny $(m+n) \times 2k$ matrix. For $\tau = 0$, we get back the refined approach of Section 4.4.

We note that both approaches may fail to produce approximations to both the left and right singular vectors, since it is not guaranteed that the computed $\widetilde{c}, \widetilde{d}, \widehat{c}$, and $\widehat{d}$ are nonzero. Suppose, for example, that $\widetilde{c} = 0$ and $\widetilde{d} \neq 0$. Then a possibility is to solve $\widetilde{c}$ from $H^T\widetilde{c} = \widetilde{d}$ (as in the $\mathcal{V}$-harmonic method). See the next section for the case of a nonsquare or singular H.

4.7 Nonsquare or singular matrices

Most characterizations of the extraction processes given above use A^{-1} and A^{-T}. Moreover, in the $\mathcal{U}$- and $\mathcal{V}$-harmonic approach, we have to solve a system with H or H^T. We now show that no difficulties arise from a nonsquare or singular A, or from a singular H. We first show that such a nonsquare or singular A gives no problems for the double-harmonic approach, using characterizations (iv) or (v) of Section 4.3.2. Let

$$U' = (I - P_{\mathcal{N}(A^T)})U \quad \text{and} \quad V' = (I - P_{\mathcal{N}(A)})V. \tag{4.7.1}$$

Notice that $A' := (I - P_{\mathcal{N}(A^T)})A(I - P_{\mathcal{N}(A)}) = A$. By using characterization (iv) of the double-harmonic method we obtain that

$$\begin{cases} AV\widetilde{d} - \widetilde{\eta}U\widetilde{c} & \perp & A\mathcal{V}, \\ A^TU\widetilde{c} - \widetilde{\theta}V\widetilde{d} & \perp & A^T\mathcal{U}, \end{cases}$$

if and only if

$$\begin{cases} A'V'\widetilde{d} - \widetilde{\eta}U'\widetilde{c} & \perp & A'\mathcal{V}', \\ (A')^TU'\widetilde{c} - \widetilde{\theta}V'\widetilde{d} & \perp & (A')^T\mathcal{U}'. \end{cases}$$

So for the double-harmonic approach we may assume without loss of generality that A is nonsingular. Another way to see this is via item (v) in Section 4.3.2: since we assumed that AV and A^TU are of full rank, we see that nonsquare or singular A form no difficulty.

Though characterization (ii) of the refined approach (see Section 4.4) is suitable only for nonsingular matrices, it can be seen from the items (i) and (iii) that nonsquare or singular A give no problems.

In the $\mathcal{U}$- and $\mathcal{V}$-harmonic method, we have to solve a system of the form $H\widetilde{d} = \widetilde{c}$ or $H^T\widetilde{c} = \widetilde{d}$. Independent of the properties of A, the matrix H may be singular. In this circumstance, the systems involving H or H^T do not have a unique solution. This is precisely the situation where infinite harmonic Ritz values may occur in the $\mathcal{U}$- and $\mathcal{V}$-harmonic method, see (iii) in Section 4.3.1.

A possibility is then to solve the small system in a least squares sense, that is, take the pseudoinverse: $\widetilde{d} = H^{+}\widetilde{c}$ or $\widetilde{c} = (H^T)^{+}\widetilde{d}$. However, these vectors may be zero, in that case the $\mathcal{U}$- and $\mathcal{V}$-harmonic approaches fail to give an approximation to one of the two singular vectors. In numerical experiments, however, we have not encountered this situation.

We mention that for nonsquare or singular matrices, the theorems in Sections 4.3 and 4.4 still hold, as far as the nonzero singular values are concerned.

The following examples illustrate some properties of the extraction methods.

Example 4.7.1 Take $A = \mathrm{diag}(1, 2, 3)$, and suppose that we try to find an approximation to the smallest singular triple of A from the search spaces $\mathcal{U} = \mathrm{span}(e_1, e_3)$, and $\mathcal{V} = \mathrm{span}(e_1, e_2)$. First we consider the standard extraction: the singular triples of $H = \mathrm{diag}(1, 0)$ lead to approximate singular triples $(\theta, u, v) = (0, e_3, e_2)$ and $(1, e_1, e_1)$. So, although the standard extraction finds the smallest triple of A, it is not safe to take the smallest $(0, e_3, e_2)$ as an approximation to the smallest triple of A. Both the $\mathcal{U}$-harmonic and $\mathcal{V}$-harmonic approach have to deal with a singular H. As discussed, we could enlarge $\mathcal{U}$ and $\mathcal{V}$ (by for instance the residual vectors r_1 and r_2) to avoid the singularity, or take the pseudoinverse for the systems involving H or H^T. The latter option gives $(\theta, \eta, u, v) = (1, 1, e_1, e_1)$ for both the $\mathcal{U}$- and $\mathcal{V}$-harmonic approach. The double-harmonic approach finds the tuples $(\theta, \eta, u, v) = (1, 1, e_1, e_1)$ and $(\infty, \infty, e_3, e_2)$. Here it is safe to take the smallest tuple. (We can get rid of the infinite harmonic value by taking the Rayleigh quotient of e_3 and e_2: this gives the approximate value 0. This is an example where it can be seen that the Rayleigh quotients often make more sense than the harmonic values.) Finally, the refined approach gives the correct solution $(\theta, u, v) = (1, e_1, e_1)$. The conclusion is that in this example, the standard approach is the only one having difficulties to determine the "smallest" singular vectors that are present in the search spaces. ⊘

Example 4.7.2 Let A be as in Example 4.7.1, and let $\mathcal{U} = \mathcal{V} = \mathrm{span}(e_2, (e_1 + e_3)/\sqrt{2})$. Then $H = \mathrm{diag}(2, 2)$ has a double singular value and the standard extraction does not know which approximate vectors to take. The other methods do know how to decide: for target $\tau = 0$, all three harmonic approaches and the refined approach take $(u, v) = (e_2, e_2)$ as approximate vectors with approximate singular value $\theta = \eta = 2$. Also for target $\tau = 2$, the double-harmonic and refined method (see Section 4.6) yield $(2, e_2, e_2)$ as approximate triple. ⊘

4.8 Lanczos bidiagonalization

We now study the different extraction methods in the context of Lanczos bidiagonalization. After k steps of Lanczos bidiagonalization with starting vector v_1 we have the relations [31, p. 495]:

$$\begin{cases} AV_k &= U_k B_{k,k}, \\ A^T U_k &= V_{k+1} B_{k+1,k}^T, \end{cases}$$

where $B_{k,k}$ and $B_{k+1,k}^T$ are a $k \times k$ upper, and a $(k+1) \times k$ lower bidiagonal matrix, respectively. This implies that the standard extraction process (see Section 4.2) takes the singular triples of

$$U_k^T A V_k = B_{k,k}$$

as approximations to the singular triples of A. The $\mathcal{V}$-harmonic method (see Section 4.3.2, characterization (iii)) reduces to

$$\left\{ \begin{array}{lcl} B_{k,k}^T B_{k,k} \widetilde{d} & = & \widetilde{\theta} B_{k,k}^T \widetilde{c}, \\ B_{k,k}^T \widetilde{c} & = & \widetilde{\eta} \widetilde{d}. \end{array} \right.$$

Here $B_{k,k}$ is nonsingular due to the assumption that AV and A^TU are of full rank. Hence, one may check that the $\mathcal{V}$-harmonic approach also takes the singular triples of $B_{k,k}$ as approximate singular triples. Similar remarks can be made for the $\mathcal{U}$-harmonic method.

Item (v) of the double-harmonic approach (Section 4.4) deduces to

$$\left\{ \begin{array}{lcl} \widetilde{\eta} B_{k,k} \widetilde{d} & = & B_{k,k+1} B_{k+1,k}^T \widetilde{c}, \\ \widetilde{\theta} B_{k,k}^T \widetilde{c} & = & B_{k,k}^T B_{k,k} \widetilde{d}. \end{array} \right.$$

Again we may assume that $B_{k,k}$ is nonsingular. Then $B_{k,k}\widetilde{d} = \widetilde{\theta}\widetilde{c}$ and $B_{k,k+1}B_{k+1,k}^T\widetilde{c} = (\widetilde{\theta}\widetilde{\eta})\widetilde{c}$. Therefore, $\widetilde{c}$ is a left singular vector of $B_{k,k+1}$. When $B_{k,k+1}$ and $B_{k,k}$ do not differ much (as will be true when $\mathcal{U}$ and $\mathcal{V}$ are nearly invariant singular subspaces), $\widetilde{c}$ is close to a left singular vector of $B_{k,k}$, and hence $\widetilde{d}$ is close to a right left singular vector of $B_{k,k}$.

The refined approach considers

$$\left\{ \begin{array}{lcl} \min\limits_{\substack{c \in \mathbb{R}^k \\ \|c\| = 1}} \|A^T U_k c\| & = & \min\limits_{\substack{c \in \mathbb{R}^k \\ \|c\| = 1}} \|B_{k+1,k}^T c\|, \\ \min\limits_{\substack{d \in \mathbb{R}^k \\ \|d\| = 1}} \|A V_k d\| & = & \min\limits_{\substack{d \in \mathbb{R}^k \\ \|d\| = 1}} \|B_{k,k} d\|, \end{array} \right.$$

to which the smallest (left and right, respectively) singular vectors of $B_{k,k+1}$ and $B_{k,k}$ are the solutions.

We conclude that in Lanczos bidiagonalization (a two-sided subspace method where we choose a specific subspace expansion), all extraction processes do essentially the same: approximating singular triples of A by those of $B_{k,k}$ or $B_{k,k+1}$. Since the new extraction processes in this chapter are often good for the minimal singular triple, the standard extraction is also fine in this case. This may be seen as an explanation why Lanczos bidiagonalization is, besides for the largest singular values, also successful for the approximation of the smallest singular triples. For other two-sided subspace methods, such as JDSVD, the extraction processes may differ much, see also the numerical experiments.

Finally, for completeness, we give the extraction processes of Section 4.6 for interior singular values in the case of Lanczos bidiagonalization. The double harmonic approach attempts to determine (the "smallest") eigenpairs of

$$\begin{bmatrix} B_{k,k+1}B_{k+1,k}^T & -\tau B_{k,k} \\ -\tau B_{k,k} & B_{k,k}^T B_{k,k} \end{bmatrix} \begin{bmatrix} \widetilde{c} \\ \widetilde{d} \end{bmatrix} = \widetilde{\theta} \begin{bmatrix} -\tau I_k & B_{k,k} \\ B_{k,k}^T & -\tau I_k \end{bmatrix} \begin{bmatrix} \widetilde{c} \\ \widetilde{d} \end{bmatrix}.$$

The refined extraction for target $\tau \geq 0$ considers

$$\min_{\substack{c,\,d\,\in\,\mathbb{R}^k \\ \|c\| = \|d\| = 1}} \left\| \begin{bmatrix} AV_k d - \tau U_k c \\ AU_k^T c - \tau V_k d \end{bmatrix} \right\| = \min_{\substack{c,\,d\,\in\,\mathbb{R}^k \\ \|c\| = \|d\| = 1}} \left\| \begin{bmatrix} B_{k,k} d - \tau c \\ B_{k+1,k}^T c - \tau I_{k+1,k} d \end{bmatrix} \right\|,$$

where $I_{k+1,k}$ is the identity with an extra $(k+1)$th zero row. We conclude that in Lanczos bidiagonalization, the extraction methods for the smallest singular triples are all more or less equivalent, but the extraction methods for interior singular triples differ.

4.9 Applications

In this section, we study two applications where the new extraction processes can be used: the least squares problem and the truncated SVD.

4.9.1 The least squares problem

The *least squares problem*

$$\min_v \|b - Av\|,$$

with minimal norm solution

$$A^+ b = \sum_{\sigma_j \neq 0} \sigma_j^{-1} (x_j^T b) y_j, \tag{4.9.1}$$

has often been successfully attacked by methods based on Lanczos bidiagonalization. For example, LSQR [59] chooses $u_1 = b/\beta$, where $\beta = \|b\|$, and forms U_k and V_k such that $A^T U_k = V_k B_{k,k}^T$ and $AV_k = U_{k+1} B_{k+1,k}$. Suppose we look for a solution $v \in \mathcal{V}_k$, say $v = V_k d$. Then

$$\|b - Av\| = \|\beta U_{k+1} \underline{e}_1 - AV_k d\| = \|\beta U_{k+1} \underline{e}_1 - U_{k+1} B_{k+1,k} d\| = \|\beta \underline{e}_1 - B_{k+1,k} d\|,$$

where $\underline{e}_1$ is the first unit vector in $\mathbb{R}^{k+1}$. Now LSQR takes the approximation

$$v = \beta V_k B_{k+1,k}^+ \underline{e}_1. \tag{4.9.2}$$

For other two-sided SVD methods, such as JDSVD (Chapter 3), we can use a similar idea, although in general, we will not have short recurrences as in LSQR. Let $\widetilde{\mathcal{U}}$ a test space, yet to be determined, and let $[\,\widetilde{U}\ \widetilde{U}_\perp\,]$ form an orthogonal basis. For ease omit the index k. Then, again with $\beta u_1 = b$ and $v = Vd$ we get

$$\|b - Av\| = \left\| \begin{bmatrix} \widetilde{U}^T (\beta U e_1 - AVd) \\ \widetilde{U}_\perp^T (\beta U e_1 - AVd) \end{bmatrix} \right\|.$$

Now we neglect the second part $\|\widetilde{U}_\perp^T (\beta U e_1 - AVd)\|$, this is equivalent to requiring

$$b - Av \perp \widetilde{\mathcal{U}}.$$

The minimal norm solution to $\min_d \|\beta\widetilde{U}^T U e_1 - \widetilde{U}^T AVd\|$ leads to

$$\widetilde{v} = \beta V(\widetilde{U}^T AV)^+ \widetilde{U}^T U e_1.$$

To get a good solution, $V(\widetilde{U}^T AV)^+\widetilde{U}^T$ should be a good approximation to A^+. Since a pseudoinverse is mainly determined by its smallest singular values and vectors, we realize that the extraction of those small singular triples from the search spaces $\mathcal{U}$ and $\mathcal{V}$ is crucial. We have already seen that the choice $\widetilde{U} = U$ (leading to $v = \beta V H^+ e_1$) is often not satisfactory for the smallest singular triples. The choice $\widetilde{U} = AV$, as in the $\mathcal{V}$-harmonic and double-harmonic approaches, is more promising, since

$$\widetilde{v} = \operatorname*{argmin}_{v\in\mathcal{V}} \|b - Av\| \quad \text{iff} \quad b - A\widetilde{v} \perp A\mathcal{V}.$$

With this choice, we have $d = \beta(V^T A^T AV)^{-1} V^T A^T U e_1 = \beta(AV)^+ U e_1$, so d is the least squares solution to $\min_d \|\beta U e_1 - AVd\|$. Then $v = \beta V(AV)^+ U e_1$. For Lanczos bidiagonalization, this gives $v = \beta V_k B_{k,k}^+ e_1$, which resembles the LSQR solution (4.9.2).

Since AV and $H^T = V^T A^T U$ are already computed in the $\mathcal{V}$-harmonic and double-harmonic methods, these approaches can therefore also be useful for least square problems: they may give an approximate solution of the least squares problem at low additional costs during the process.

As already mentioned, methods such as JDSVD may need restarts and deflation from time to time. When we would like to use these methods for the least squares problem, special care has to be taken when the maximum dimension of the search space has been reached (restart), or when a singular triple has been found (deflation). With restarts, we have to ensure that $b = \beta u_1 \in \mathcal{U}$. Therefore, we restart with the span of the best (say) $l-1$ left vectors in $\mathcal{U}$, together with u_1 as the new left search space $\mathcal{U}$ (what "the best" means, depends on the extraction method). For the new right search space $\mathcal{V}$, we take the span of the best $l-1$ right vectors in $\mathcal{V}$, together with $V(AV)^+ U e_1$, the minimal norm solution to $\min_{v\in\mathcal{V}} \|b - Av\|$. Since we include the best approximation so far to the least squares problem in the new search space $\mathcal{V}$, we get monotonic convergence for the least squares solution, that is,

$$\|b - Av_{k+1}\| \leq \|b - Av_k\|.$$

This is trivial when we expand the search space $\mathcal{V}$, but by restarting in this way, it is also valid at restarts.

Now consider deflation. Suppose we have detected a singular triple (σ, x, y), where $\sigma \neq 0$. By decomposing

$$A = (I - xx^T)A(I - yy^T) + \sigma xy^T \quad \text{and} \quad b = (I - xx^T)b + (xx^T)b,$$

we get

$$\min_v \|b - Av\|^2 = \min_v \|(I - xx^T)(b - A(I - yy^T)v)\|^2 + |x^T b - \sigma y^T v|^2.$$

So we may conclude that the (minimal norm) solution v has a component $\sigma^{-1}x^Tb$ in the direction of y. This may also be seen from (4.9.1). With $\widetilde{v} = (I - yy^T)v$, we are left with a *deflated least squares problem*

$$\min_{\widetilde{v}\perp y} \|(I - xx^T)(b - A\widetilde{v})\|.$$

Hence, if a triple has been found, we restart with the best $l-1$ left vectors in $\mathcal{U}$ and $(I-xx^T)b$ as the new $\mathcal{U}$. (In this case, "the best" means the best vectors to find the next singular triple.) For the new $\mathcal{V}$, we take the best $l-1$ right vectors in $\mathcal{V}$, together with $(I - yy^T)v$, where, as before, $v = V(AV)^+Ue_1$ is the current best approximation to the least squares problem. Of course, this procedure can be repeated when more singular triples are found. See Section 4.10 for numerical experiments.

4.9.2 The truncated SVD

We may also use the standard and double-harmonic methods to give an approximation to the *truncated SVD* of A. The solution to

$$\min_{\text{rank}(B)=k} \|A - B\|$$

is given by $B = A_k := \sum_{j=1}^{k} \sigma_j x_j y_j^T$ (unique if σ_k is simple). Analogously, the solution to

$$\min_{\text{rank}(B)=k} \|A^+ - B^+\|$$

is given by $B = A_{-k} := \sum_{j=1}^{k} \sigma_{-j} x_{-j} y_{-j}^T$, where the sum is over nonzero singular values of A (unique if σ_{-k} is simple).

In view of the discussed extraction processes, we expect that when $\mathcal{U}_k$ and $\mathcal{V}_k$ are search spaces for the largest triples,

$$P_{\mathcal{U}_k} A P_{\mathcal{V}_k} = U_k U_k^T A V_k V_k^T = U_k H_k V_k^T$$

may be a reasonable approximation to A_k. Compare this with [71], where the authors use the Lanczos bidiagonalization to approximate A_k by $U_k B_{k,k} V_k^T$. Although the Lanczos process has in principle short recurrences, (some) reorthogonalization of the vectors appears to be necessary [71].

Now consider the situation that $\mathcal{U}_k$ and $\mathcal{V}_k$ are search spaces for the smallest triples. Then from (4.3.1), we know that $\widetilde{H} = G_U H^{-1} G_V^T$ can be viewed as a projected approximation to A, which attempts to approximate the smallest portion of the singular spectrum well. Hence, as a reasonable approximation to A_{-k}, we may take

$$U_k \widetilde{H}_k V_k^T = U_k G_U H_k^{-1} G_V^T V_k^T.$$

See Section 4.10 for numerical experiments.

4.10 Numerical experiments

The following experiments were carried out in MATLAB. For all experiments where the random generator is used, we first put the "seed" to 0 so that our results are reproducible (see Section 1.4.3).

Experiment 4.10.1 Up to rounding errors, it is not a loss of generality to consider diagonal matrices (see Lemma 3.8.1). For the first example we take $A = \mathrm{diag}(1:100)$. We build up four-dimensional search spaces $\mathcal{U}$ and $\mathcal{V}$ to find the minimal singular triple. The first basis vector of the left search space $\mathcal{U}$ is $u_1 = e_1 + \varepsilon_U w_U$, where w_U is a random vector of unit length. We complement $\mathcal{U}$ by three random vectors. The right search space $\mathcal{V}$ is formed in a similar way: take $v_1 = e_1 + \varepsilon_V w_V$, and add three random vectors.

We consider two cases. For the first we take $\varepsilon_U = 10^{-3}$ and $\varepsilon_V = 10^{-1}$. This means that the left search space is good, while the right search space is not very accurate. It appears that

$$\angle(\mathcal{U}, e_1) \approx 3.5 \cdot 10^{-3} \quad \text{and} \quad \angle(\mathcal{V}, e_1) \approx 3.0 \cdot 10^{-1},$$

these angles also give the best possible approximate vectors in $\mathcal{U}$ and $\mathcal{V}$. For the second case we take $\varepsilon_U = \varepsilon_V = 10^{-3}$, in other words: left and right search spaces of good quality. In this case

$$\angle(\mathcal{U}, e_1) \approx 3.5 \cdot 10^{-3} \quad \text{and} \quad \angle(\mathcal{V}, e_1) \approx 3.2 \cdot 10^{-3}.$$

Table 4.3 gives the results of the 5 different extraction processes. We display the error in the approximate vectors u and v, and the error in the approximate value ρ, the Rayleigh quotient of u and v.

TABLE 4.3: The 5 different extraction processes for the minimal singular triple of $A = \mathrm{diag}(1:100)$. Column 2 to 4 are for $\varepsilon_U = 10^{-3}$ and $\varepsilon_V = 10^{-1}$, while column 5 to 7 represent $\varepsilon_U = \varepsilon_V = 10^{-3}$

	$\varepsilon_U = 10^{-3}, \varepsilon_V = 10^{-1}$			$\varepsilon_U = 10^{-3}, \varepsilon_V = 10^{-3}$		
method	$\angle(u, e_1)$	$\angle(v, e_1)$	$\lvert\sigma_{\min} - \rho\rvert$	$\angle(u, e_1)$	$\angle(v, e_1)$	$\lvert\sigma_{\min} - \rho\rvert$
standard	$1.9e-1$	$8.0e-1$	$2.8e-1$	$4.0e-3$	$1.1e-2$	$2.5e-5$
$\mathcal{U}$-harmonic	$3.5e-3$	$7.6e-1$	$2.7e-1$	$3.5e-3$	$1.0e-2$	$2.6e-5$
$\mathcal{V}$-harmonic	$1.1e-2$	$3.1e-1$	$2.3e-2$	$4.4e-3$	$3.3e-3$	$6.8e-5$
double-harmonic	$3.6e-3$	$3.1e-1$	$4.4e-2$	$3.5e-3$	$3.3e-3$	$2.3e-7$
refined	$3.5e-3$	$3.1e-1$	$4.8e-2$	$3.5e-3$	$3.3e-3$	$8.1e-7$

Almost all errors of the new methods are smaller than those of the standard approach. Moreover, the new approaches are almost optimal in most cases, by which we mean that the extracted vectors are almost the best possible ones, given the search spaces. In view of the factors $\sqrt{1 + \frac{\gamma^2}{\delta^2}} \approx 1.4$ for both the $\mathcal{U}$- and $\mathcal{V}$-harmonic approach (see Theorem 4.3.2), and $\sqrt{1 + 2\frac{\widetilde{\gamma}^2}{\widetilde{\delta}^2}} \approx 11$ for the standard method (see Theorem 4.2.2), we already could suspect that the harmonic approaches would be superior. We mention that in the $\mathcal{V}$-harmonic method, the approximate left vector $\widetilde{u} = UH^{-T}\widetilde{d}$ is indeed much better that $A\widetilde{v}$ (cf. Lemma 4.3.1 and discussion). Similar remarks hold for the $\mathcal{U}$-harmonic method. ⊘

Experiment 4.10.2 For the other experiments, we use JDSVD, the Jacobi–Davidson type method for the singular value problem introduced in Chapter 3. Unless mentioned otherwise, we set the following parameters for JDSVD: the dimension of the search spaces is at most 20, after which we restart with the best 10 vectors (remember that the meaning of "the best" depends on the extraction method). We use the JDSVD correction equation of the form

$$\begin{bmatrix} I_m - \frac{u\widetilde{u}^T}{\widetilde{u}^T u} & 0 \\ 0 & I_n - \frac{v\widetilde{v}^T}{\widetilde{v}^T v} \end{bmatrix} \begin{bmatrix} -\zeta I_m & A \\ A^T & -\zeta I_n \end{bmatrix} \begin{bmatrix} I_m - uu^T & 0 \\ 0 & I_n - vv^T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = - \begin{bmatrix} Av - \rho u \\ A^T u - \rho v \end{bmatrix}.$$

Here u and v are the current approximate vectors, $\widetilde{u}$ and $\widetilde{v}$ are the test vectors (depending on the extraction method; for example, in the $\mathcal{V}$-harmonic approach we have $\widetilde{u} = Av$, $\widetilde{v} = v$), and we solve for orthogonal updates $s \perp u$ and $t \perp v$. On the right-hand side of the correction equation, we take the Rayleigh quotient $\rho(u, v)$ as approximate singular value. In the beginning, the shift ζ in the left-hand side of the correction equation is taken to be the target τ. The reason for this is that the Rayleigh quotient ρ is not likely to be accurate at that stage. So initially, the method behaves as an inexact inverse iteration with target τ. When we are close to convergence, in the sense that $\|r\| < 0.01$, we take the shift ζ equal to the Rayleigh quotient ρ. Then the methods works as an inexact Rayleigh quotient iteration. We solve the correction equation (the so-called "inner iteration") approximately, by 10 steps of GMRES, preconditioned by the projected identity (see Chapter 3; in practice, it is advisable to use a projected nontrivial preconditioner if one is available). We continue with the method until $\|r\| < 1e-6$.

We take the 1850×712 matrix `well1850` from the Matrix Market [53], with $\sigma_{\min} \approx 1.6 \cdot 10^{-2}$. We perform 70 steps of JDSVD with double-harmonic extraction and target $\tau = 0$ to find the smallest singular triple. The starting vectors u_1 and v_1 are the vector of all ones. For the (20-dimensional) search spaces after 70 steps of the method, it appears that

$$\angle(\mathcal{U}, x) = 4.9 \cdot 10^{-3} \quad \text{and} \quad \angle(\mathcal{V}, y) = 3.8 \cdot 10^{-3}.$$

With these search spaces, we test the different extraction processes, see the first four columns of Table 4.4.

TABLE 4.4: The 5 different extraction processes for the minimal singular triple of `well1850`. Column 2 to 4 are the extraction results for the 20-dimensional search spaces produced after 70 steps of JDSVD. The last three columns give the number of outer steps needed for the computation of the smallest singular triple, with 10, 20, and 30 steps of GMRES to solve the correction equations, respectively.

method	$\angle(u, e_1)$	$\angle(v, e_1)$	$\lvert\sigma_{\min} - \rho\rvert$	GMRES_{10}	GMRES_{20}	GMRES_{30}
standard	$1.6e+0$	$1.6e+0$	$1.0e-2$	> 200	67	41
$\mathcal{U}$-harmonic	$1.0e-2$	$8.0e-3$	$8.7e-7$	155	72	41
$\mathcal{V}$-harmonic	$1.2e-2$	$6.7e-3$	$2.1e-7$	171	67	41
double-harmonic	$1.2e-2$	$1.1e-2$	$1.3e-6$	97	48	34
refined	$1.0e-2$	$6.7e-3$	$8.9e-7$	97	51	36

We see that the standard approach fails completely, apparently due to the selection

of the wrong triple (a situation similar to that in Example 4.7.1). The new extraction methods perform reasonably well.

In the last three columns of Table 4.4, we give the number of outer iterations it takes before JDSVD with the specific extraction method has detected the smallest singular triple. For column 5, we use 10 steps of GMRES, for column 6 we perform 20 steps, and for the last column 30 steps to solve the correction equations. It appears that the less accurate we solve the correction equation (which is, of course, cheaper), the more advantageous the new extraction methods are, especially the double-harmonic and refined approach. ⊘

Experiment 4.10.3 Next, for $A = \mathrm{diag}(1 : 100)$, we perform 100 (outer) steps of JDSVD with each of the extraction processes, and count how many singular triples we find. See Table 4.5. For the second and third column we take target $\tau = 0$ (i.e., we look for the smallest singular triples), for the fourth and fifth column $\tau = 50.1$ (interior triples closest to 50.1), and for the last two columns $\tau = 105$ and $\tau = \infty$ (largest triples).

TABLE 4.5: The number of singular triples found within 100 outer iterations by the 5 different extraction processes for $A = \mathrm{diag}(1 : 100)$. Columns 2 and 3 are for $\tau = 0$ (minimal singular triples), respectively with no fix and fix = 0.01. Column 4 and 5 represent $\tau = 50.1$ (interior singular triples), with 10 and 20 steps of GMRES, respectively, while for the last two columns $\tau = 105$ and $\tau = \infty$ (largest singular triples).

	$\tau = 0$		$\tau = 50.1$		$\tau = 105$	$\tau = \infty$
method	no fix	fix	GMRES_{10}	GMRES_{20}		
standard	1	2	–	3	20	20
$\mathcal{U}$-harmonic	1	4	–	3	17	20
$\mathcal{V}$-harmonic	3	3	–	3	17	20
double-harmonic	3	7	1	4	20	–
refined	5	7	2	8	19	21

The "no fix" in the second column means that we take the shift ζ in the left-hand side of the correction equation equal to the Rayleigh quotient from the beginning. For the results of column 3 through 6 we fix the target in the left-hand side of the correction equation until $\|r\| < 0.01$. As can be seen from the second and third column, and as already has been suggested in Experiment 4.10.2, a "fix" gives better results than "no fix". Except for column 5, all correction equations are solved by 10 steps of unpreconditioned GMRES.

We see that for small and interior triples the $\mathcal{U}$- and $\mathcal{V}$-harmonic, and especially the double-harmonic and refined methods are superior compared with the standard approach. For the largest triples all methods are fine, with the exception that the double-harmonic needs a target $\tau < \infty$. ⊘

Experiment 4.10.4 In Figure 4.1 we compare the quality of extraction of the standard (a) and double-harmonic (b) method during the search for the smallest singular triple in Experiment 4.10.3. We plot $\angle(\mathcal{U}, e_1)$ (solid), $\angle(u, e_1)$ (dashed), $\angle(\mathcal{V}, e_1)$ (dots), and $\angle(v, e_1)$ (dash-dot).

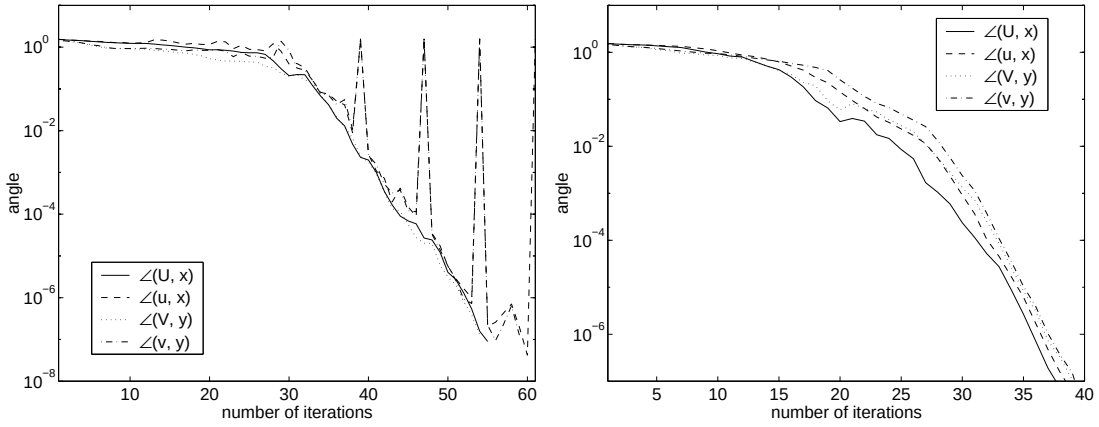


FIGURE 4.1: The extraction results of the standard (a) and double-harmonic approach (b) for the computation of the smallest singular triple of $A = \mathrm{diag}(1:100)$.

In the double-harmonic approach (b), in every step $\angle(u, e_1) \approx \angle(\mathcal{U}, e_1)$ and $\angle(v, e_1) \approx \angle(\mathcal{V}, e_1)$, so this extraction process is almost optimal. We might say that the extraction is as good as the search spaces allow. In the standard extraction (a), the extraction is often good, but sometimes bad. In the expansion step (solving the correction equation), it is then unlikely to get a reasonable update, this can be regarded as the loss of one outer iteration. But what is worse: if we restart in a situation of bad extraction with few vectors, we may throw away the best part of the search space. ⊘

Experiment 4.10.5 Next, we compare JDSVD with Jacobi–Davidson (JD) to compute the smallest singular triples. We compare these methods with two different extraction processes. For Figure 4.2(a), we use JD with harmonic Rayleigh–Ritz (see, for instance, [82, p. 292]) to compute the eigenpairs of the augmented matrix (3.1.1) closest to target $\tau = 0$. Recall that finding an eigenpair of the augmented matrix gives full information on a singular triple of A, and vice versa. For JDSVD, we take the double-harmonic extraction method. All correction equations (JD and JDSVD) are solved approximately by 5 steps of GMRES. As the initial vectors we take v_1 random and $u_1 = Av_1$.

For Figure 4.2(b), JD uses refined Ritz vectors (see, for instance, [82, p. 289]), while JDSVD uses the refined Ritz extraction of Section 4.4.

From Figure 4.2, it is clear that JDSVD easily beats JD with both extraction methods. Part of an explanation of this fact could be that JD sees $\sigma_{\min}$ and $-\sigma_{\min}$ as two different eigenvalues of the augmented matrix. JDSVD avoids this "doubling". ⊘

Experiment 4.10.6 Now we illustrate the use the new extraction methods may have in producing approximate solutions to least squares problems with a simple example. Suppose we are interested in the problem

$$\min_v \|b - Av\|,$$

where $A = \mathrm{diag}(1:100)$ and b is the vector of unit length with all entries equal. For Figure 4.3, we run JDSVD with $\mathcal{V}$-harmonic (a) and double-harmonic extraction (b) with target $\tau = 0$ and starting vectors $u_1 = v_1 = b$.

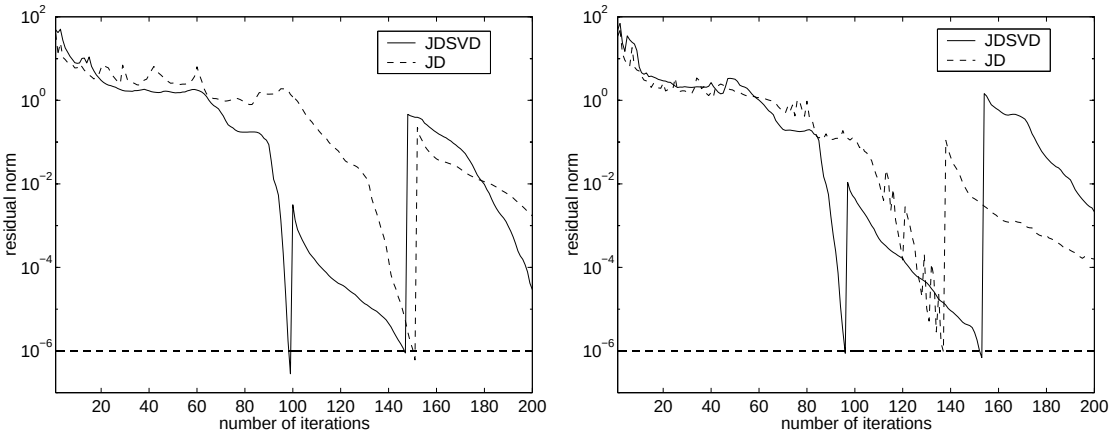


FIGURE 4.2: JDSVD with double-harmonic approach versus JD with the harmonic Ritz approach (a), and JDSVD with refined extraction versus JD with refined Ritz (b) for the computation of the smallest singular triples of $A = \text{diag}(1:100)$.

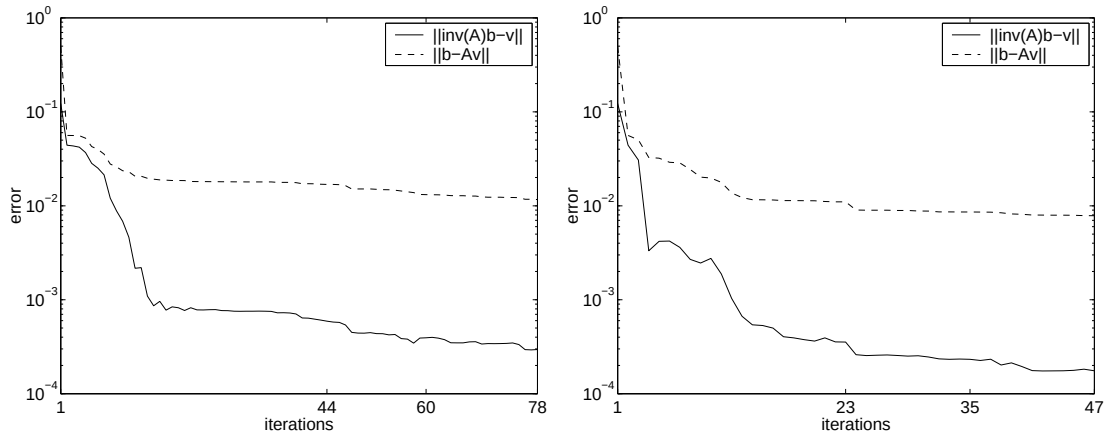
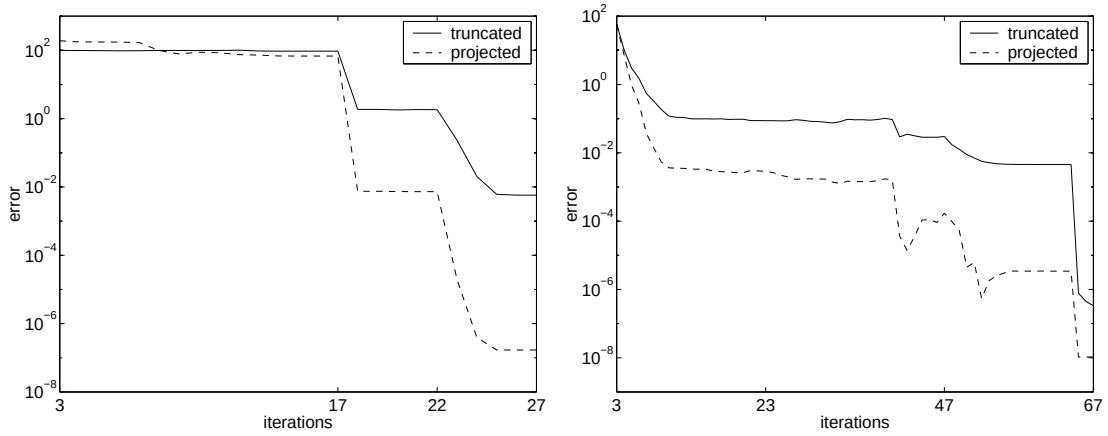


FIGURE 4.3: The error and residual norm of the least squares solution produced by the $\mathcal{V}$-harmonic (a) and double-harmonic approach (b) during the computation of the three smallest singular triples of $A = \text{diag}(1:100)$.

Depicted are the residual norm $\|b - Av\|$ and the error $\|A^{-1}b - v\|$. The numbers on the horizontal axis (except the number 1) indicate that at that iteration step, a singular triple is detected. We see that the errors decrease rapidly in the beginning; afterwards the convergence is slower but for the residual norm still monotonic, as predicted in Section 4.9.1. In particular, the true error seems to behave even more favorable than the residual norm. ⊘

Experiment 4.10.7 Finally, we illustrate the use of the extraction methods for the approximation of truncated SVDs. We run JDSVD to compute the three largest (a), respectively smallest (b) singular triples of $A = \text{diag}(1:100)$. For (a) we use the standard extraction, for (b) the double-harmonic extraction. The starting vectors u_1 and v_1 are the vector of all ones, the target τ is 0 for (a) and ∞ for (b).



(a) (b)

FIGURE 4.4: JDSVD with standard (a) and double-harmonic (b) to approximate the truncated SVDs A_3 (a) and A_{-3} (b) of $A = \text{diag}(1:100)$.

In Figure 4.4(a), we depict the error in the truncated SVD $\|P_{\mathcal{U}_3} A P_{\mathcal{V}_3} - A_3\|$, (here $A_3 = \sum_{j=98}^{100} j e_j e_j^T$, and in every step $\mathcal{U}_3$ and $\mathcal{V}_3$ represent the best three-dimensional part of $\mathcal{U}$ and $\mathcal{V}$), and the error in the "projected" truncated SVD $\|U_3^T A V_3 - \text{diag}(98:100)\|$. For Figure 4.4(b), we give the error in the truncated SVD $\|P_{A\mathcal{V}_3} A^{-T} P_{A^T\mathcal{U}_3} - A_{-3}\|$ (here $A_{-3} = \sum_{j=1}^{3} j e_j e_j^T$) and the error in the "projected" truncated SVD $\|\widetilde{H}_3 - \text{diag}(1:3)\|$ (b), where (cf. (4.3.1)) $\widetilde{H}_3 = G_U H_3^{-1} G_V^T$, with $H_3 = U_3^T A V_3$, where U_3 and V_3 represent the best three-dimensional part of $\mathcal{U}$ and $\mathcal{V}$. The numbers (except the number 3) on the horizontal axis in Figure 4.4 indicate the detection of a singular triple. For the "top" truncated SVD (a), the convergence corresponds, not surprisingly, with the detection of the singular triples. ⊘

4.11 Conclusions

For the accurate approximation of the minimal singular triple, we may use two separate subspaces. With respect to the *subspace expansion*, the Jacobi–Davidson SVD (inexact

scaled RQI) is a competitor to Lanczos bidiagonalization when a (good) preconditioner is available (see the previous chapter).

With respect to the *subspace extraction*, the standard approach (via $H = U^T AV$) is fine for large singular triples, while for small and interior triples, the harmonic or refined approaches are recommended. For the extraction of the smallest singular triple in Lanczos bidiagonalization we have seen that the standard, harmonic and refined approach are essentially equivalent.

Based on the theory and supported by numerical experiments, we can do the following recommendations:

- for the largest singular triples, we can choose any extraction method, except the double-harmonic (although this method may also perform well if we have a target $\tau < \infty$); the standard extraction is preferable because it is the cheapest;
- for interior singular triples we opt for the double-harmonic or refined approach;
- for the smallest singular triples we suggest any method except the standard; the double-harmonic and refined approach seem to be the most promising.

The $\mathcal{V}$-harmonic and double-harmonic method can also serve to give an approximate solution to a least squares problem; the standard and double-harmonic method can approximate the truncated SVDs.

Chapter 5

A Jacobi–Davidson type method for the right definite two-parameter eigenvalue problem

Abstract. We present a new numerical iterative method for computing selected eigenpairs of a right definite two-parameter eigenvalue problem. The method does not need good initial approximations and is able to tackle large problems that are too expensive for existing methods. The new method is similar to the Jacobi–Davidson method for the eigenvalue problem. In each step, we first compute Ritz pairs of a small projected right definite two-parameter eigenvalue problem and then expand the search spaces using approximate solutions of appropriate correction equations. We present two alternatives for the correction equations, introduce a selection technique that makes it possible to compute more than one eigenpair, and give some numerical results.

Key words: right definite two-parameter eigenvalue problem, subspace method, Jacobi–Davidson, correction equation, Ritz pair, accelerated inexact Newton.

AMS subject classification: 65F15, 15A18, 15A69, 65F50.

5.1 Introduction

We are interested in computing one or more eigenpairs of a right definite two-parameter eigenvalue problem

$$\begin{aligned} A_1 x &= \lambda B_1 x + \mu C_1 x, \\ A_2 y &= \lambda B_2 y + \mu C_2 y, \end{aligned} \tag{5.1.1}$$

where A_i, B_i, and C_i are given real symmetric $n_i \times n_i$ matrices for $i = 1, 2$ and $\lambda, \mu \in \mathbb{R}$, $x \in \mathbb{R}^{n_1}$, $y \in \mathbb{R}^{n_2}$. A pair (λ, μ) is called an eigenvalue if it satisfies (5.1.1) for nonzero

*Based on joint work with Bor Plestenjak, see Section 1.5.

vectors x, y. The tensor product $x \otimes y$ is the corresponding eigenvector. The condition for *right definiteness* is that the determinant

$$\begin{vmatrix} x^T B_1 x & x^T C_1 x \\ y^T B_2 y & y^T C_2 y \end{vmatrix} \tag{5.1.2}$$

is strictly positive for all nonzero vectors $x \in \mathbb{R}^{n_1}$, $y \in \mathbb{R}^{n_2}$. Right definiteness and symmetry of the matrices A_i, B_i, and C_i imply that there exist $n_1 n_2$ linearly independent eigenvectors for the problem (5.1.1) [4].

As mentioned in Section 1.1, multiparameter eigenvalue problems of this kind arise in a variety of applications [3], particularly in mathematical physics when the method of separation of variables is used to solve boundary value problems [100]. The result of the separation is a two-parameter system of ordinary differential equations.

Two-parameter problems can be expressed as two coupled generalized eigenvalue problems as follows. On the tensor product space $S := \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ of the dimension $N := n_1 n_2$ we define the matrices

$$\begin{aligned} \Delta_0 &= B_1 \otimes C_2 - C_1 \otimes B_2, \\ \Delta_1 &= A_1 \otimes C_2 - C_1 \otimes A_2, \\ \Delta_2 &= B_1 \otimes A_2 - A_1 \otimes B_2 \end{aligned} \tag{5.1.3}$$

(For details on the tensor product and the relation to the multiparameter eigenvalue problem, see, for example, [4].) Since the tensor product of symmetric matrices is symmetric, Δ_i is symmetric for $i = 0, 1, 2$. By noting that

$$(x \otimes y)^T \Delta_0 (x \otimes y) = \begin{vmatrix} x^T B_1 x & x^T C_1 x \\ y^T B_2 y & y^T C_2 y \end{vmatrix},$$

it can be seen that right definiteness of (5.1.1) is equivalent to the condition that Δ_0 is positive definite, see [4, Theorem 7.8.2]. It is also shown in [4] that $\Delta_0^{-1}\Delta_1$ and $\Delta_0^{-1}\Delta_2$ commute and that the problem (5.1.1) is equivalent to the associated problem

$$\begin{aligned} \Delta_1 z &= \lambda \Delta_0 z, \\ \Delta_2 z &= \mu \Delta_0 z \end{aligned} \tag{5.1.4}$$

for decomposable tensors $z \in S$, $z = x \otimes y$. The eigenvectors of (5.1.1) are Δ_0-orthogonal, i.e. if $x_1 \otimes y_1$ and $x_2 \otimes y_2$ are eigenvectors of (5.1.1) corresponding to different eigenvalues, then

$$(x_1 \otimes y_1)^T \Delta_0 (x_2 \otimes y_2) = \begin{vmatrix} x_1^T B_1 x_2 & x_1^T C_1 x_2 \\ y_1^T B_2 y_2 & y_1^T C_2 y_2 \end{vmatrix} = 0. \tag{5.1.5}$$

Decomposable tensors $x_i \otimes y_i$ for $i = 1, \ldots, N$ form a complete basis for S.

There exist numerical methods for right definite two-parameter eigenvalue problems. First of all, the associated problem (5.1.4) can be transformed in such a way that it can be solved by numerical methods for simultaneous diagonalization of commutative symmetric matrices [79, 42, 14]. This is only feasible for problems of low dimension as the size of the matrices of the associated problem is $N \times N$. Among other methods, we

mention those based on Newton's method [11], the gradient method [9, 10, 13], and the minimal residual quotient iteration [8]. A deficiency of these methods is that they require initial approximations close enough to the solution in order to avoid misconvergence.

The continuation method [70, 64] overcomes problems with initial approximations but, since the ordering of the eigenvalues is not necessarily preserved in a continuation step, we have to compute all eigenvalues even if we are interested only in a small portion. In this chapter, we introduce a new numerical method which is similar to the Jacobi–Davidson method for the one-parameter eigenvalue problem [75]. The method can be used to compute selected eigenpairs and does not need good initial approximations.

Our method computes the exterior eigenvalue (λ, μ) of (5.1.1) which has the maximum value of $\lambda \cos\alpha + \mu \sin\alpha$ for a given α. We also present a version that computes the interior eigenpair closest to a given pair (λ_0, μ_0), i.e., the one with minimum $(\lambda - \lambda_0)^2 + (\mu - \mu_0)^2$.

The outline of this chapter is as follows. We generalize the Rayleigh–Ritz approach to right definite two-parameter eigenvalue problems in Section 5.2. In Section 5.3 we present a Jacobi–Davidson type method for right definite two-parameter eigenvalue problems and introduce two alternatives for the correction equations. We discuss how the method can be used for exterior and interior eigenvalues in Section 5.4. In Section 5.5, we present a selection technique that allows us to compute more than one eigenpair. The time complexity is given in Section 5.6 and the methods are extended to more than two parameters in Section 5.7. Some numerical examples are presented in Section 5.8. Conclusions are summarized in Section 5.9.

5.2 Subspace methods and Ritz pairs

The Jacobi–Davidson method [75] is one of the subspace methods that may be used for the numerical solution of one-parameter eigenvalue problems. We will apply a Jacobi–Davidson type method to (5.1.1). Recall from Section 1.3 that extraction and expansion are two main themes in a subspace method. In this section we discuss the extraction, in the next section the algorithm and the expansion.

For the standard eigenvalue problem, the Rayleigh–Ritz approach defines approximations to the eigenpairs that can be extracted from the given subspace (see for instance [61]). We generalize the Rayleigh–Ritz approach for the two-parameter eigenvalue problem as follows. Suppose that the k-dimensional search subspaces $\mathcal{U}_k$ of $\mathbb{R}^{n_1}$ and $\mathcal{V}_k$ of $\mathbb{R}^{n_2}$ are represented by matrices $U_k \in \mathbb{R}^{n_1 \times k}$ and $V_k \in \mathbb{R}^{n_2 \times k}$ with orthonormal columns, respectively. The *Ritz–Galerkin conditions* on the *residuals*

$$\begin{aligned} r_1 &:= (A_1 - \sigma B_1 - \tau C_1)u \quad \perp \quad \mathcal{U}_k, \\ r_2 &:= (A_2 - \sigma B_2 - \tau C_2)v \quad \perp \quad \mathcal{V}_k, \end{aligned} \tag{5.2.1}$$

where $u \in \mathcal{U}_k \backslash \{0\}$ and $v \in \mathcal{V}_k \backslash \{0\}$, lead to the smaller projected right definite two-parameter problem

$$\begin{aligned} U_k^T A_1 U_k c &= \sigma U_k^T B_1 U_k c + \tau U_k^T C_1 U_k c, \\ V_k^T A_2 V_k d &= \sigma V_k^T B_2 V_k d + \tau V_k^T C_2 V_k d, \end{aligned} \tag{5.2.2}$$

where $u = U_k c \neq 0$, $v = V_k d \neq 0$, $c, d \in \mathbb{R}^k$, and $\sigma, \tau \in \mathbb{R}$.

We say that an eigenvalue (σ, τ) of (5.2.2) is a *Ritz value* for the two-parameter eigenvalue problem (5.1.1) and subspaces $\mathcal{U}_k$ and $\mathcal{V}_k$. If (σ, τ) is an eigenvalue of (5.2.2) and $c \otimes d$ is the corresponding eigenvector, then $u \otimes v$ is a *Ritz vector*, where $u = U_k c$ and $v = V_k d$. Altogether, we obtain k^2 *Ritz pairs* that are approximations to the eigenpairs of (5.1.1). It is easy to check that, if $u \otimes v$ is a Ritz vector corresponding to the Ritz value (σ, τ), then σ and τ are equal to the *tensor Rayleigh quotients* [64]

$$\begin{aligned} \sigma &= \frac{(u \otimes v)^T \Delta_1 (u \otimes v)}{(u \otimes v)^T \Delta_0 (u \otimes v)} = \frac{(u^T A_1 u)(v^T C_2 v) - (u^T C_1 u)(v^T A_2 v)}{(u^T B_1 u)(v^T C_2 v) - (u^T C_1 u)(v^T B_2 v)}, \\ \tau &= \frac{(u \otimes v)^T \Delta_2 (u \otimes v)}{(u \otimes v)^T \Delta_0 (u \otimes v)} = \frac{(u^T B_1 u)(v^T A_2 v) - (u^T A_1 u)(v^T B_2 v)}{(u^T B_1 u)(v^T C_2 v) - (u^T C_1 u)(v^T B_2 v)}. \end{aligned} \tag{5.2.3}$$

In order to obtain Ritz values, we have to solve small right definite two-parameter eigenvalue problems. For this purpose, one of the available numerical methods that computes all eigenpairs of a small right definite two-parameter eigenvalue problem can be used. For instance, the associated problem (5.1.4) can be solved using methods for simultaneous diagonalization of two commutative symmetric matrices [79, 42, 14].

5.3 A Jacobi–Davidson type method

The Jacobi–Davidson method [75] is a subspace method where approximate solutions of certain correction equations are used to expand the search space. Jacobi–Davidson type methods restrict the search for a new direction to the subspace that is orthogonal or oblique to the last chosen Ritz vector.

In this chapter, we show that a Jacobi–Davidson type method can be applied to the right definite two-parameter problem as well. A brief sketch of the method is presented in Algorithm 5.3.1. In Step 4, we have to decide which Ritz pair to select. We give details of this step in Section 5.4 where we discuss how to deal with exterior and interior eigenvalues. In Step 8, we have to find new search directions to expand the search subspaces. We will discuss two possible correction equations for Step 8 later in this section.

To apply this algorithm we need to specify a tolerance ε, a maximum number of steps $k_{\max}$, a maximum dimension of the search subspaces $l_{\max}$, and a number $l_{\min} < l_{\max}$ that specifies the dimension of the search subspaces after a restart.

A larger search space involves a larger projected problem (5.2.2). The existing methods are able to solve only low-dimensional two-parameter problems in a reasonable time. Therefore, we expand the search spaces up to the preselected dimension $l_{\max}$ and then restart the algorithm. For a restart, we take the most promising $l_{\min}$ eigenvector approximations as a basis for the initial search space.

Suppose that we have computed new directions s and t for the search spaces $\mathcal{U}_{k+1}$ and $\mathcal{V}_{k+1}$, respectively. We expand the search spaces simply by adding new columns to the matrices U_k and V_k. For reasons of efficiency and stability we want orthonormal columns, and, therefore, we orthonormalize s against U_k and t against V_k by a stable form of the Gram–Schmidt orthonormalization.

Input: initial vectors u_1 and v_1 with unit norm and a tolerance ε
Output: an approximate eigenpair satisfying $(\|r_1\|^2 + \|r_2\|^2)^{1/2} \leq \varepsilon$

1. $s = u_1$, $t = v_1$, $U_0 = [\,]$, $V_0 = [\,]$
   for $k = 1, \ldots, k_{\max}$
2. Expand the search subspaces.
   $U_k = \text{MGS}\ (U_{k-1}, s)$,
   $V_k = \text{MGS}\ (V_{k-1}, t)$
3. Solve the projected right definite two-parameter eigenvalue problem
   $U_k^T A_1 U_k c = \sigma U_k^T B_1 U_k c + \tau U_k^T C_1 U_k c$,
   $V_k^T A_2 V_k d = \sigma V_k^T B_2 V_k d + \tau V_k^T C_2 V_k d$.
4. Select an appropriate Ritz value (σ, τ) and the corresponding Ritz vector $u \otimes v$, where $u = U_k c$, $v = V_k d$.
5. Compute the residuals
   $r_1 = (A_1 - \sigma B_1 - \tau C_1)u$,
   $r_2 = (A_2 - \sigma B_2 - \tau C_2)v$.
6. Stop if $\rho_k = (\|r_1\|^2 + \|r_2\|^2)^{1/2} \leq \varepsilon$
7. Restart. If the dimension of U_k and V_k exceeds $l_{\max}$
   then replace U_k, V_k with new orthonormal bases of dimension $l_{\min}$.
8. Compute new search directions s and t.

ALGORITHM 5.3.1: A Jacobi–Davidson type method for the right definite two-parameter eigenvalue problem

The next theorem expresses that, if the residuals (5.2.1) are small, then the Ritz value (σ, τ) is a good approximation to an eigenvalue of (5.1.1). This justifies the criterion in Step 6.

Theorem 5.3.1 *If (σ, τ) is a Ritz value and r_1, r_2 are the residuals (5.2.1), then there exists an eigenvalue (λ, μ) of the right definite two-parameter problem (5.1.1) such that*

$$\begin{aligned} |\lambda - \sigma| &\leq \|\Delta_0^{-1}\|(\|C_1\|\|r_2\| + \|C_2\|\|r_1\|), \\ |\mu - \tau| &\leq \|\Delta_0^{-1}\|(\|B_1\|\|r_2\| + \|B_2\|\|r_1\|). \end{aligned}$$

Proof: To prove the theorem, we consider the associated problem (5.1.4). First, we derive a relation between the residuals (5.2.1) and the residuals of the associated problem. We denote

$$\begin{aligned} p_1 &= \Delta_1(u \otimes v) - \sigma\Delta_0(u \otimes v), \\ p_2 &= \Delta_2(u \otimes v) - \tau\Delta_0(u \otimes v), \end{aligned} \tag{5.3.1}$$

where u, v are the normalized Ritz vectors from Step 4. From (5.1.3) and (5.2.1), it follows that

$$\begin{aligned} p_1 &= -C_1 u \otimes r_2 + r_1 \otimes C_2 v, \\ p_2 &= B_1 u \otimes r_2 - r_1 \otimes B_2 v \end{aligned}$$

and we have the bounds

$$\begin{aligned} \|p_1\| &\leq \|C_1\|\|r_2\| + \|C_2\|\|r_1\|, \\ \|p_2\| &\leq \|B_1\|\|r_2\| + \|B_2\|\|r_1\|. \end{aligned} \tag{5.3.2}$$

Now we return to the residuals (5.3.1). As Δ_0 is a symmetric positive definite matrix, we can transform (5.3.1) into

$$\begin{aligned} \Delta_0^{-1/2} p_1 &= G_1 w - \sigma w, \\ \Delta_0^{-1/2} p_2 &= G_2 w - \tau w, \end{aligned} \tag{5.3.3}$$

where $w = \Delta_0^{1/2}(u \otimes v)$ and $G_i = \Delta_0^{-1/2}\Delta_i\Delta_0^{-1/2}$ for $i = 1, 2$. The matrices G_1 and G_2 are symmetric and an application of Bauer–Fike for the first equation of (5.3.3) gives

$$\min_{\lambda \in \Lambda(G_1)} |\lambda - \sigma| \leq \|\Delta_0^{-1/2} p_1\| / \|w\| \leq \|\Delta_0^{-1/2}\|^2 \|p_1\| = \|\Delta_0^{-1}\|\|p_1\|.$$

Similarly, we get $\min_{\mu \in \Lambda(G_2)} |\mu - \tau| \leq \|\Delta_0^{-1}\|\|p_2\|$. When we insert (5.3.2) into these two inequalities, we have proved the theorem. □

In the next theorem, we show that, if the Ritz vector $u \otimes v$ is close to an eigenvector $x \otimes y$ of problem (5.1.1), then the residuals r_1 and r_2 from (5.2.1) are of order $\mathcal{O}(\|u - x\|)$ and $\mathcal{O}(\|v - y\|)$, respectively. This shows that the criterion in Step 6 will be fulfilled if the Ritz vector $u \otimes v$ approximates an eigenvector of (5.1.1) well enough.

Theorem 5.3.2 *Let (σ, τ) be a Ritz value of (5.1.1) with corresponding Ritz vector $u \otimes v$, where u and v are normalized. If $(u + s) \otimes (v + t)$ is an eigenvector of (5.1.1) with corresponding eigenvalue (λ, μ), then we can bound the error of (σ, τ) as*

$$\sqrt{(\lambda - \sigma)^2 + (\mu - \tau)^2} = \mathcal{O}(\|s\|^2 + \|t\|^2) \tag{5.3.4}$$

and the norm of the residuals r_1, r_2 from (5.2.1) as

$$\begin{aligned} \|r_1\| &\leq \|A_1 - \lambda B_1 - \mu C_1\|\|s\| + \mathcal{O}(\|s\|^2 + \|t\|^2), \\ \|r_2\| &\leq \|A_2 - \lambda B_2 - \mu C_2\|\|t\| + \mathcal{O}(\|s\|^2 + \|t\|^2). \end{aligned} \tag{5.3.5}$$

Proof: We write the residuals (5.2.1) as

$$\begin{aligned} r_1 &= -(A_1 - \lambda B_1 - \mu C_1)s + (\lambda - \sigma)B_1 u + (\mu - \tau)C_1 u, \\ r_2 &= -(A_2 - \lambda B_2 - \mu C_2)t + (\lambda - \sigma)B_2 v + (\mu - \tau)C_2 v. \end{aligned} \tag{5.3.6}$$

When we multiply equations (5.3.6) by u^T and v^T, respectively, and take into account that $u^T r_1 = v^T r_2 = 0$, then we obtain

$$\begin{bmatrix} u^T B_1 u & u^T C_1 u \\ u^T B_2 v & v^T C_2 v \end{bmatrix} \begin{bmatrix} \lambda - \sigma \\ \mu - \tau \end{bmatrix} = - \begin{bmatrix} s^T (A_1 - \lambda B_1 - \mu C_1)s \\ u^T (A_2 - \lambda B_2 - \mu C_2)t \end{bmatrix}. \tag{5.3.7}$$

The system (5.3.7) is nonsingular because of right definiteness. From (5.3.7), it follows that

$$\left\| \begin{bmatrix} \lambda-\sigma \\ \mu-\tau \end{bmatrix} \right\| = \left\| \begin{bmatrix} u^TB_1u & u^TC_1u \\ v^TB_2v & v^TC_2v \end{bmatrix}^{-1} \begin{bmatrix} s^T(A_1-\lambda B_1-\mu C_1)s \\ t^T(A_2-\lambda B_2-\mu C_2)t \end{bmatrix} \right\| = \mathcal{O}(\|s\|^2+\|t\|^2),$$

and we get (5.3.4). The bound (5.3.5) is now a result of (5.3.6) and (5.3.4). □

In the following two subsections, the expansion for our Jacobi–Davidson method is discussed. We present two alternatives for the correction equations for the right definite two-parameter eigenvalue problem. Let (σ,τ) be a Ritz value that approximates the eigenvalue (λ,μ) of (5.1.1), and let $u\otimes v$ be its corresponding Ritz vector. Let us assume that u and v are normalized.

5.3.1 Correction equations with orthogonal projections

The first alternative for the correction equations is a generalization of the approach used in [75] for the one-parameter eigenvalue problem. We are searching for orthogonal improvements of the vectors u and v of the form

$$\begin{aligned} A_1(u+s) &= \lambda B_1(u+s)+\mu C_1(u+s), && (5.3.8)\\ A_2(v+t) &= \lambda B_2(v+t)+\mu C_2(v+t), && (5.3.9) \end{aligned}$$

where $s\perp u$ and $t\perp v$. Using the residuals of the Ritz vector $u\otimes v$ and Ritz value (σ,τ) (5.2.1), we can rewrite (5.3.8) and (5.3.9) as

$$\begin{aligned} (A_1-\sigma B_1-\tau C_1)s &= -r_1+(\lambda-\sigma)B_1u+(\mu-\tau)C_1u && (5.3.10)\\ &\quad +(\lambda-\sigma)B_1s+(\mu-\tau)C_1s, \\ (A_2-\sigma B_2-\tau C_2)t &= -r_2+(\lambda-\sigma)B_2v+(\mu-\tau)C_2v && (5.3.11)\\ &\quad +(\lambda-\sigma)B_2t+(\mu-\tau)C_2t. \end{aligned}$$

In this subsection, we treat the equations (5.3.10) and (5.3.12) separately. From Theorem 5.3.2, it follows that $\|(\lambda-\sigma)B_1u+(\mu-\tau)C_1u\|=\mathcal{O}(\|s\|^2+\|t\|^2)$. Asymptotically (i.e., when $u\otimes v$ is close to an eigenvector of (5.1.1)), s and t are first order corrections and $(\lambda-\sigma)B_1u+(\mu-\tau)C_1u$ represents some second order correction. In the same sense, the term $(\lambda-\sigma)B_1s+(\mu-\tau)C_1s$ can be interpreted as a third order correction.

If we ignore second and higher order terms in (5.3.10), then we obtain the equation

$$(A_1-\sigma B_1-\tau C_1)s=-r_1. \qquad (5.3.12)$$

Because r_1 and s are orthogonal to u, we can multiply (5.3.12) with the orthogonal projection $I-uu^T$ and write $(I-uu^T)s$ instead of s. Thus we obtain the correction equation for the vector u

$$(I-uu^T)(A_1-\sigma B_1-\tau C_1)(I-uu^T)s=-r_1. \qquad (5.3.13)$$

In a similar way, we obtain from (5.3.12) the correction equation for the vector v

$$(I - vv^T)(A_2 - \sigma B_2 - \tau C_2)(I - vv^T)t = -r_2. \tag{5.3.14}$$

From (5.3.13) and (5.3.14), it is clear that the orthogonal projections preserve the symmetry of the matrices. Another advantage of orthogonal projections is that they are stable and easy to implement. The systems (5.3.13) and (5.3.14) for s and t are not of full rank but they are consistent. We solve them only approximately with a Krylov subspace method with initial guess 0, for instance, by a few steps of MINRES. If we do just one step of MINRES, then s and t are scalar multiples of r_1 and r_2, respectively, and then, in the sense that we expand the search spaces by the residuals, we have an Arnoldi type method, similar to the situation for the standard eigenproblem [75].

5.3.2 Correction equation with oblique projections

As in the correction equations with orthogonal projections we start with the equations (5.3.10) and (5.3.12). We neglect the third order correction terms $(\lambda - \sigma)B_1 s + (\mu - \tau)C_1 s$ and $(\lambda - \sigma)B_2 t + (\mu - \tau)C_2 t$, but rather than neglecting the second order terms $(\lambda - \sigma)B_1 u + (\mu - \tau)C_1 u$ and $(\lambda - \sigma)B_2 v + (\mu - \tau)C_2 v$, we project them to 0 using an oblique projection.

If we define

$$D = \begin{bmatrix} A_1 - \sigma B_1 - \tau C_1 & 0 \\ 0 & A_2 - \sigma B_2 - \tau C_2 \end{bmatrix}$$

and

$$r = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix},$$

then we can reformulate (5.3.10) and (5.3.12) (without the neglected third order correction terms) as

$$D \begin{bmatrix} s \\ t \end{bmatrix} = -r + (\lambda - \sigma) \begin{bmatrix} B_1 u \\ B_2 v \end{bmatrix} + (\mu - \tau) \begin{bmatrix} C_1 u \\ C_2 v \end{bmatrix}. \tag{5.3.15}$$

Let $V \in \mathbb{R}^{(n_1+n_2)\times 2}$ be a matrix with columns (for reasons of stability, preferably orthonormal) such that

$$\operatorname{span}(V) = \operatorname{span}\left(\begin{bmatrix} B_1 u \\ B_2 v \end{bmatrix}, \begin{bmatrix} C_1 u \\ C_2 v \end{bmatrix} \right),$$

and let $W \in \mathbb{R}^{(n_1+n_2)\times 2}$ be

$$W = \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix}.$$

With the oblique projection

$$P = I - V(W^T V)^{-1} W^T$$

onto $\mathrm{span}(V)^{\perp}$ along $\mathrm{span}(W)$, it follows that

$$Pr = r \quad \text{and} \quad P \begin{bmatrix} B_1 u \\ B_2 v \end{bmatrix} = P \begin{bmatrix} C_1 u \\ C_2 v \end{bmatrix} = 0. \tag{5.3.16}$$

Therefore, from multiplying (5.3.15) by P we obtain

$$PD \begin{bmatrix} s \\ t \end{bmatrix} = -r.$$

Furthermore, since $s \perp u$ and $t \perp v$ it follows that

$$P \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix} \tag{5.3.17}$$

and the result is the correction equation

$$PDP \begin{bmatrix} s \\ t \end{bmatrix} = -r \tag{5.3.18}$$

for $s \perp u$ and $t \perp v$.

The correction equation (5.3.18) is again not of full rank but consistent, and it is often sufficient to solve it only approximately (e.g., by a few steps of GMRES). As before, if we do one step of GMRES, then s and t are scalar multiples of r_1 and r_2, respectively.

The Jacobi–Davidson method for the one-parameter problem can be viewed as an accelerated inexact Newton scheme [76]. In a similar manner, we now show that there is a connection between the Jacobi–Davidson correction equation (5.3.18) and Newton's method for the right definite two-parameter eigenvalue problem in [64].

Eigenpairs of the two-parameter problem (5.1.1) are solutions of the equation

$$G(x, y, \lambda, \mu) := \begin{bmatrix} A_1 x - \lambda B_1 x - \mu C_1 x \\ A_2 y - \lambda B_2 y - \mu C_2 y \\ (x^T x - 1)/2 \\ (y^T y - 1)/2 \end{bmatrix} = 0. \tag{5.3.19}$$

If we apply Newton's method to (5.3.19) and use u, v, σ, τ with $\|u\| = \|v\| = 1$ as an initial approximation, then, in order to obtain the improved approximation $u+s, v+t, \lambda, \mu$ we have to solve the system

$$\begin{bmatrix} A_1 - \sigma B_1 - \tau C_1 & 0 & -B_1 u & -C_1 u \\ 0 & A_2 - \sigma B_2 - \tau C_2 & -B_2 v & -C_2 v \\ u^T & 0 & 0 & 0 \\ 0 & v^T & 0 & 0 \end{bmatrix} \begin{bmatrix} s \\ t \\ \lambda - \sigma \\ \mu - \tau \end{bmatrix} = \begin{bmatrix} -r_1 \\ -r_2 \\ 0 \\ 0 \end{bmatrix}. \tag{5.3.20}$$

Lemma 5.3.3 *The Jacobi–Davidson correction equation (5.3.18), where $s \perp u$ and $t \perp v$, is equivalent to Newton's equation (5.3.20). That is, if (s, t) is a solution of (5.3.18), then there exist unique λ, μ such that $(s, t, \lambda - \sigma, \mu - \tau)$ is a solution of (5.3.20), and, if $(s, t, \lambda - \sigma, \mu - \tau)$ is a solution of (5.3.20), then (s, t) is a solution of (5.3.18).*

Proof: We can rewrite the equation (5.3.20) as

$$D \begin{bmatrix} s \\ t \end{bmatrix} = -r + (\lambda - \sigma) \begin{bmatrix} B_1 u \\ B_2 v \end{bmatrix} + (\mu - \tau) \begin{bmatrix} C_1 u \\ C_2 v \end{bmatrix}$$

and $s \perp u$, $t \perp v$, which is exactly the equation (5.3.15) that appears in the derivation of the Jacobi–Davidson correction equation (5.3.18). The proof now follows from the relations (5.3.16) and (5.3.17), and the fact that $\ker(P) = \mathrm{span}(V)$. □

This shows that the Jacobi–Davidson type method with the correction equation (5.3.18) is a Newton scheme, accelerated by the projection of (5.1.1) onto the subspace of all previous approximations. Therefore, we expect locally at least quadratic convergence of the Jacobi–Davidson method when the correction equations are solved exactly.

5.4 Selection of Ritz values

In this section we present different options for the selection of Ritz values in Step 4 of Algorithm 5.3.1.

5.4.1 Exterior eigenvalues

First, we discuss how to obtain the eigenvalue (λ, μ) of (5.1.1) with the maximum value of λ. We denote such an eigenvalue by $(\lambda_{\max}, \mu_{\max})$. We show that, if we select the Ritz value (σ, τ) with the maximum value of σ in each Step 4 of Algorithm 5.3.1, then the Ritz pairs will converge monotonically to an eigenpair of (5.1.1).

Lemma 5.4.1 *Let (σ, τ) be the Ritz value for problem (5.1.1) and subspaces $\mathcal{U}, \mathcal{V}$ with the maximum value of σ. Then*

$$\sigma = \max_{\substack{u \in \mathcal{U}, v \in \mathcal{V} \\ u, v \neq 0}} \frac{(u \otimes v)^T \Delta_1 (u \otimes v)}{(u \otimes v)^T \Delta_0 (u \otimes v)}. \tag{5.4.1}$$

Proof: Let the columns of U and V be orthonormal bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. It follows from (5.1.1), (5.1.4), and (5.2.2) that, if (σ, τ) is a Ritz value, then σ is an eigenvalue of a symmetric definite pencil

$$(U \otimes V)^T \Delta_1 (U \otimes V) - \sigma (U \otimes V)^T \Delta_0 (U \otimes V). \tag{5.4.2}$$

From the minimax theorem (cf. [31, p. 394]) it follows that

$$\sigma = \max_{\substack{w \in \mathcal{U} \otimes \mathcal{V} \\ w \neq 0}} \frac{w^T \Delta_1 w}{w^T \Delta_0 w}.$$

Since pencil (5.4.2) is related to the two-parameter problem (5.2.2), we can restrict w to a decomposable tensor $w = u \otimes v$, where $u \in \mathcal{U}$ and $v \in \mathcal{V}$. From this, (5.4.1) follows. □

If we select the Ritz value (σ_k, τ_k) in Step 4 of Algorithm 5.3.1 with the maximum σ_k, then it follows from Lemma 5.4.1 that

$$\sigma_k \leq \sigma_{k+1} \leq \lambda_{\max}.$$

We cannot guarantee that the eigenvalue (λ, μ) of (5.1.1) to which (σ_k, τ_k) converges is equal to $(\lambda_{\max}, \mu_{\max})$, but convergence to a local optimum also may happen in the Jacobi–Davidson method for the symmetric eigenproblem and in all projection methods. Our numerical examples indicate that we usually do obtain the eigenvalue with the largest value of λ.

We can use the algorithm to obtain the eigenvalue (λ, μ) of (5.1.1) with the maximum value of $\lambda \cos\alpha + \mu \sin\alpha$ for a given parameter α if we apply the orthogonal linear substitution

$$\begin{aligned} \lambda &= \lambda' \cos\alpha - \mu' \sin\alpha, \\ \mu &= \lambda' \sin\alpha + \mu' \cos\alpha \end{aligned}$$

to the problem (5.1.1). The associated two-parameter eigenproblem with this substitution is now

$$\begin{aligned} A_1 x &= \lambda'(\cos\alpha B_1 + \sin\alpha C_1)x + \mu'(-\sin\alpha B_1 + \cos\alpha C_1)x, \\ A_2 y &= \lambda'(\cos\alpha B_2 + \sin\alpha C_2)y + \mu'(-\sin\alpha B_2 + \cos\alpha C_2)y. \end{aligned} \tag{5.4.3}$$

The operator determinant Δ_0 remains unchanged, and the substituted problem (5.4.3) is right definite as well. Using orthogonal linear substitutions we can thus obtain exterior eigenvalues of (5.1.1) in chosen directions in the (λ, μ)-plane.

Step 4 of Algorithm 5.3.1 can be modified in an obvious manner if we are interested in the eigenvalue (λ, μ) of (5.1.1) with the maximum value of $\lambda^2 + \mu^2$.

5.4.2 Interior eigenvalues

Suppose that we are interested in the eigenvalue (λ, μ) of (5.1.1) closest to a specific target (λ_0, μ_0). Let us denote such an eigenvalue as $(\lambda_{\text{int}}, \mu_{\text{int}})$.

Similar to the algorithm for exterior eigenvalues, we decide to select the Ritz value nearest to the target in each Step 4 of Algorithm 5.3.1. The convergence for interior Ritz values is not as favorable as for the exterior ones. If a Ritz value (σ, τ) is close enough to $(\lambda_{\max}, \mu_{\max})$, then the Ritz vector corresponding to (σ, τ) is a good approximation to the eigenvector corresponding to $(\lambda_{\max}, \mu_{\max})$. On the contrary, if (σ, τ) is close to $(\lambda_{\text{int}}, \mu_{\text{int}})$ then the Ritz vector corresponding to (σ, τ) may be a poor approximation to the eigenvector corresponding to $(\lambda_{\text{int}}, \mu_{\text{int}})$, just as in the real symmetric eigenproblem.

Numerical examples in Section 5.8 show that, although the convergence is very irregular, the method can still be used to compute the eigenvalue closest to the target. It

turns out that for interior eigenvalues, good search directions are needed, which may be obtained by solving the correction equation more accurately. The number of GMRES steps is of large influence. The more steps of GMRES we take, the better updates for the approximate eigenvectors will be added to the search spaces. If we take too many steps, then the method often converges to an eigenvalue $(\lambda, \mu) \neq (\lambda_{\text{int}}, \mu_{\text{int}})$. On the other hand, if we take too few GMRES steps, then we need many outer iterations or we have no convergence at all. Two alternative approaches are considered in the next two subsections.

5.4.3 Harmonic Rayleigh–Ritz

If we are interested in interior eigenvalues of the standard or generalized eigenproblem then one of the possible tools is harmonic Rayleigh–Ritz. This extraction method can be derived from certain Galerkin conditions on the matrix $(A-\tau I)^{-1}$, see, for instance, [75]. The shift-and-invert transformation $t \mapsto (t-\tau)^{-1}$, a Möbius transform on the projective line $\mathbb{P}^1(\mathbb{R})$ or $\mathbb{P}^1(\mathbb{C})$, maps ∞ to 0 and τ to ∞.

In a two-parameter problem, the (λ, μ)-plane is embedded in the projective plane $\mathbb{P}^2(\mathbb{R})$ (the equivalence classes $\langle \lambda, \mu, \nu \rangle$ where, for $\alpha \neq 0$, $\langle \lambda, \mu, \nu \rangle \sim \langle \alpha\lambda, \alpha\mu, \alpha\nu \rangle$) by the identification

$$(\lambda, \mu) \leftrightarrow \langle \lambda, \mu, 1 \rangle.$$

As our method works best for exterior eigenvalues ("the ones closest to the line on ∞"), for interior eigenvalues we can try to map the line on ∞, i.e., the line $\nu = 0$, to any other line in the (λ, μ)-plane. For instance, when we are interested in the eigenvalues with minimal $|\mu|$, then to map the line on ∞ onto the line $\mu = 0$. In our homogeneous projective two-parameter problem

$$\begin{aligned} \nu A_1 x &= \lambda B_1 x + \mu C_1 x, \\ \nu A_2 y &= \lambda B_2 y + \mu C_2 y, \end{aligned}$$

this map is achieved by interchanging the roles of μ and ν. The resulting non-projective two-parameter problem is

$$\begin{aligned} C_1 x &= -\widetilde{\lambda} B_1 x + \widetilde{\mu} A_1 x, \\ C_2 y &= -\widetilde{\lambda} B_2 y + \widetilde{\mu} A_2 y, \end{aligned} \tag{5.4.4}$$

where $\widetilde{\lambda}$ corresponds to $\lambda\mu^{-1}$ and $\widetilde{\mu}$ corresponds to μ^{-1}. A problem with this approach is that (5.4.4) is in general not right-definite. We may try to tackle this problem with the method for more general multiparameter problems, developed in Chapter 6.

5.4.4 Refined Ritz vectors

Another possible extraction process for interior eigenvalues generalizes refined Ritz vectors. As usual, we perform the Rayleigh–Ritz process to get a Ritz pair $((\sigma, \tau), u \otimes v)$.

But then we discard the Ritz vector, and instead take the *refined Ritz vector* $\widehat{u}\otimes\widehat{v}$, where $\widehat{u} = U\widehat{c}$ and $\widehat{v} = V\widehat{d}$ are such that

$$\begin{aligned} \widehat{c} &= \operatorname*{argmin}_{c} \|(A_1 - \sigma B_1 - \tau C_1)Uc\|, \\ \widehat{d} &= \operatorname*{argmin}_{d} \|(A_2 - \sigma B_2 - \tau C_2)Vd\|. \end{aligned} \tag{5.4.5}$$

Taking this vector ensures that the norms of the residuals (quantities related to the backward error, see Chapter 7) are minimized over the search spaces. Therefore, the refined Ritz vector may be better than the (ordinary) Ritz vector.

When interested in the target (λ_0, μ_0), one can also replace the Ritz value (σ, τ) in (5.4.5) by the target. When the Ritz value is not very accurate (as will often be the case in the beginning of the search process), then the target is probably a better point to focus on.

5.5 Computing more eigenpairs

Suppose that we are interested in $p > 1$ eigenpairs of (5.1.1). In a one-parameter problem, various deflation techniques can be applied in order to compute more than one eigenpair. In this section, we first show difficulties that are met when we try to translate standard deflation ideas from one-parameter problems to two-parameter problems. We then propose a selection method for Ritz vectors that makes it possible to obtain more than one eigenpair for two-parameter problems.

If (ξ, z) is an eigenpair of a symmetric matrix A, then all other eigenpairs can be computed from the projection of A onto the subspace $z^{\perp}$. Similarly, if (λ, μ) is an eigenvalue of (5.1.1) and $x\otimes y$ is the corresponding eigenvector, then all other eigenvectors lie in the subspace

$$(x \otimes y)^{\perp_{\Delta_0}} := \{z \in S : z^T \Delta_0 (x \otimes y) = 0\}$$

of dimension $n_1 n_2 - 1$. By comparing the dimensions, it is clear that the subspace $(x \otimes y)^{\perp_{\Delta_0}}$ cannot be written as $\mathcal{U} \otimes \mathcal{V}$, where $\mathcal{U} \subset \mathbb{R}^{n_1}$ and $\mathcal{V} \subset \mathbb{R}^{n_2}$. Therefore, this kind of deflation cannot be applied to Algorithm 5.3.1.

Another way of deflation of a symmetric matrix A is to shift the eigenvalue to an unwanted part of the spectrum using the matrix $A' = A - (\xi - \widetilde{\xi})zz^T$. Matrix A' has the same eigenvalues as A except for ξ, which is transformed into $\widetilde{\xi}$. A generalization of this approach would be to transform the two-parameter problem (5.1.1) into a two-parameter problem with the same eigenvalues as of (5.1.1) except for the eigenvalue (λ, μ) which, should be transformed into $(\widetilde{\lambda}, \widetilde{\mu})$. Since in a two-parameter problem, there can exist eigenvalues (λ, μ) and (λ', μ') with eigenvectors $x \otimes y$ and $x' \otimes y'$, respectively, such that $(\lambda, \mu) \neq (\lambda', \mu')$ and $x = x'$, this approach would again work only if we apply the associated problem (5.1.4) in the tensor product space S. However, then we have to work with large Δ_i matrices, and this is too expensive.

We propose the following approach. Suppose that we have already found p eigenvalues (λ_i, μ_i) and eigenvectors $x_i \otimes y_i$, $i = 1, \ldots, p$. Based on the fact that eigenvectors are

Δ_0-orthogonal (see (5.1.5)), we adjust Algorithm 5.3.1 so that, in Step 4, we consider only those Ritz vectors $u \otimes v$ which satisfy

$$|(u \otimes v)^T \Delta_0 (x_i \otimes y_i)| < \eta \text{ for } i = 1, \ldots, p \tag{5.5.1}$$

for an $\eta > 0$. Suppose that we are interested in eigenvalues with the maximum values of λ. Then, in Step 4, we first order Ritz pairs (σ_i, τ_i), $u_i \otimes v_i$ by their σ values so that $\sigma_i \geq \sigma_j$ for $i < j$, and then we select the Ritz pair that satisfies (5.5.1) and has the minimal index. In the case of interior eigenvalues, a different ordering is used.

If none of the Ritz pairs meet (5.5.1), then we take the Ritz pair with index 1, but, in this case, the algorithm is not allowed to stop. This is achieved by a change of the stopping criterion in Step 6, where, in addition to a small residual norm

$$\rho := (\|r_1\|^2 + \|r_2\|^2)^{1/2}, \tag{5.5.2}$$

we now also require that the Ritz vector $u \otimes v$ satisfies (5.5.1). This guarantees that the method does not converge to the already computed eigenpairs.

The bound η should not be taken too small to avoid the situation that none of the Ritz vectors are sufficiently Δ_0-orthogonal to the set of already computed eigenvectors. In numerical experiments in Section 5.8, we use

$$\eta = \frac{1}{2} \min_{i=1,\ldots,p} |(x_i \otimes y_i)^T \Delta_0 (x_i \otimes y_i)|,$$

and that value successfully prevents the method from converging to the already computed eigenpairs.

All other steps of Algorithm 5.3.1 remain unchanged. Numerical results in Section 5.8 show that this approach enables us to compute more than one eigenpair.

5.6 Time complexity

We examine the time complexity of one outer iteration step of Algorithm 5.3.1. Let $n = n_1 = n_2$, let k be the dimension of the search spaces, and let m be the number of GMRES (MINRES) steps for a correction equation. The two steps that largely determine the time complexity are Step 3 and Step 8. In Step 3 we first construct the smaller projected problem (5.2.2). We need to compute only the last row (and column) of the matrices in (5.2.2). In the second part of Step 3, we solve (5.2.2) by solving its associated problem with matrices of size k^2, and thus we need $\mathcal{O}(k^6)$ [14].

First we assume that A_i, B_i, and C_i are sparse. This is true in many applications, for instance when two-parameter Sturm–Liouville problems [21] are discretized. Because MINRES and GMRES are methods intended for sparse matrices, the Jacobi–Davidson type method can in principle handle very large sparse problems. For such problems, the time complexities of Step 3 and Step 8 can be expressed as 6 MV + $\mathcal{O}(k^6)$ and $6m$ MV, respectively, where MV stands for a matrix-vector multiplication with an $n \times n$ matrix.

The analysis for dense matrices A_i, B_i, and C_i is as follows. In Step 3, we need $\mathcal{O}(n^2)$ for the construction of the smaller problem (5.2.2) and additional $\mathcal{O}(k^6)$ for the solution

of (5.2.2). As, in practice, only very small values of k are used, we can assume that $k = \mathcal{O}(n^{1/3})$ and thus the time complexity of Step 3 is $\mathcal{O}(n^2)$. If we use correction equations (5.3.13), (5.3.14) with orthogonal projections and perform m steps of MINRES, then the time complexity of Step 8 is $\mathcal{O}(mn^2)$ when we perform m matrix-vector multiplications. We obtain the same time complexity for Step 8 when we use the correction equation (5.3.18) with oblique projections and do m steps of GMRES. The only difference is that we are working with one matrix of size $2n$, while we are working with two matrices of size n if we use orthogonal projections.

Based on the above assumptions, the time complexity of one outer step of Algorithm 5.3.1 for dense matrices is $\mathcal{O}(mn^2)$. Also important is the storage requirement. If an algorithm works with matrices A_i, B_i, and C_i as Algorithm 5.3.1 does, then it requires $\mathcal{O}(n^2)$ memory. The methods that work with the associated system (5.1.4) need $\mathcal{O}(n^4)$ memory, which may exceed memory rapidly, even for modest values of n.

5.7 Generalization to multiparameter problems

The methods in this chapter can be generalized to p-parameter problems, where $p > 2$. We give a sketch of the method in this case. Consider the p-parameter eigenvalue problem

$$\left(V_{i0} - \sum_{j=1}^{p} \lambda_j V_{ij} \right) x_i = 0, \qquad i = 1, \ldots, p$$

(see Chapter 7 for more details on p-parameter problems). With $u_i^{(k)} = U_k c_i$, the *Ritz–Galerkin conditions*

$$\left(V_{i0} - \sum_{j=1}^{p} \theta_j^{(k)} V_{ij} \right) U_k c_i \perp U_k, \qquad i = 1, \ldots, p$$

lead to a *subspace extraction* defined by the projected right definite p-parameter problem

$$U_k^T \left(V_{i0} - \sum_{j=1}^{p} \theta_j^{(k)} V_{ij} \right) U_k c_i = 0, \qquad i = 1, \ldots, p.$$

For the *subspace expansion*, we first define the *residuals*

$$r_i := \left(V_{i0} - \sum_{j=1}^{p} \theta_j^{(k)} V_{ij} \right) u_i, \qquad i = 1, \ldots, p.$$

We would like to update the current approximation $u_1 \otimes \cdots \otimes u_p$ by $s_1, \ldots, s_p$ such that

$$\left(V_{i0} - \sum_{j=1}^{p} \lambda_j V_{ij} \right) (u_i + s_i) = 0, \qquad s_i \perp u_i.$$

We rewrite these equations as

$$\left(V_{i0} - \sum_{j=1}^{p} \theta_j^{(k)} V_{ij} \right) s_i = -r_i + \sum_{j=1}^{p} (\lambda_j - \theta_j^{(k)}) V_{ij} u_j + \sum_{j=1}^{p} (\lambda_j - \theta_j^{(k)}) V_{ij} s_j. \qquad (5.7.1)$$

Then, by the following generalization of Theorem 5.3.2, the middle p terms on the right-hand side are of second order, and the last p terms on the right-hand side are of third order. For convenience, we omit the index k.

Theorem 5.7.1 (cf. Theorem 5.3.2) *Suppose that $x_i = u_i + s_i$, for $i = 1, \ldots, p$. Then for all $i = 1, \ldots, p$ we have*

$$|\lambda_i - \theta_i| = \mathcal{O}\left(\sum_{i=1}^{p} \|s_i\|^2\right).$$

Proof: From $r_i \perp u_i$, it follows from left multiplying (5.7.1) by u_i^T that

$$\begin{bmatrix} u_1^T V_{11} u_1 & \cdots & u_1^T V_{1p} u_1 \\ \vdots & & \vdots \\ u_p^T V_{p1} u_1 & \cdots & u_p^T V_{pp} u_p \end{bmatrix} \begin{bmatrix} \theta_1 - \lambda_1 \\ \vdots \\ \theta_p - \lambda_p \end{bmatrix} = \begin{bmatrix} s_1^T (V_{10} - \sum_{j=1}^{p} \lambda_j V_{1j}) s_1 \\ \vdots \\ s_p^T (V_{p0} - \sum_{j=1}^{p} \lambda_j V_{pj}) s_p \end{bmatrix}$$

The result now follows by taking norms, and noting that the matrix in the previous equation is invertible because of right definiteness. □

Along the same lines as in Section 5.3.1, neglecting the second and third order terms in (5.7.1) gives p correction equations with one-dimensional orthogonal projections (cf. (5.3.13) and (5.3.14))

$$(I - u_i u_i^T)\left(V_{i0} - \sum_{j=1}^{p} \theta_j^{(k)} V_{ij}\right)(I - u_i u_i^T) s_i = -r_i, \qquad s_i \perp u_i.$$

Inclusion of the second order terms in (5.7.1) leads to a generalization of Section 5.3.2; we get one correction equation with a p-dimensional oblique projector

$$PDPs = -r.$$

Here $s = [\, s_1^T \ \ldots \ s_p^T \,]^T$, $r = [\, r_1^T \ \ldots \ r_p^T \,]^T$, and D is a $(\sum_{j=1}^{p} n_j) \times (\sum_{j=1}^{p} n_j)$ block diagonal matrix with $(V_{i0} - \sum_{j=1}^{p} \theta_j^{(k)} V_{ij})$ as its blocks. Furthermore, $P = I - V(W^T V)^{-1} W^T$, where W is the $(\sum_{j=1}^{p} n_j) \times p$ matrix

$$W = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & u_p \end{bmatrix},$$

and V is an orthonormal matrix with columns spanning the column space of

$$\begin{bmatrix} V_{11} u_1 & \cdots & V_{1p} u_1 \\ \vdots & & \vdots \\ V_{p1} u_1 & \cdots & V_{pp} u_p \end{bmatrix}.$$

Thus, the method can be extended to right definite multiparameter problems with more than two parameters.

5.8 Numerical experiments

We present some numerical examples obtained with MATLAB 5.3. If the dimension of the matrices is $n = n_1 = n_2 = 100$, then none of the existing methods that work in the tensor product space are able to compute all eigenpairs in a reasonable time [64]. Therefore, we construct right definite two-parameter examples where the exact eigenpairs are known, which enables us to check the obtained results.

We construct our right definite two-parameter examples in the following way. We take matrices

$$A_i = Q_i F_i Q_i^T, \quad B_i = Q_i G_i Q_i^T, \quad C_i = Q_i H_i Q_i^T,$$

where F_i, G_i, and H_i are diagonal matrices and Q_i is a random orthogonal matrix for $i = 1, 2$. We select diagonal elements of matrices F_1, F_2, G_2, and H_1 as uniformly distributed random numbers from the interval $(0, 1)$ and diagonal elements of G_1 and H_2 as uniformly distributed random numbers from the interval $(1, 2)$. The determinant (5.1.2) is clearly strictly positive for nonzero x, y, and the obtained two-parameter problem is right definite. All matrices are of dimension $n \times n$.

Write $F_i = \operatorname{diag}(f_{i1}, \ldots, f_{in})$, $G_i = \operatorname{diag}(g_{i1}, \ldots, g_{in})$, and $H_i = \operatorname{diag}(h_{i1}, \ldots, h_{in})$. It is easy to see that eigenvalues of the two-parameter problem (5.1.1) are solutions of linear systems

$$\begin{aligned} f_{1i} &= \lambda g_{1i} + \mu h_{1i}, \\ f_{2j} &= \lambda g_{2j} + \mu h_{2j} \end{aligned}$$

for $i, j = 1, \ldots, n$. This enables us to compute all the eigenvalues from the diagonal elements of F_i, G_i, H_i, for $i = 1, 2$. In order to construct a two-parameter problem that has the point $(0, 0)$ in the interior of the convex hull of all the eigenvalues, we take the shifted problem

$$\begin{aligned} (A_1 - \lambda_0 B_1 - \mu_0 C_1)x &= (\lambda - \lambda_0)B_1 x + (\mu - \mu_0)C_1 x, \\ (A_2 - \lambda_0 B_2 - \mu_0 C_2)y &= (\lambda - \lambda_0)B_2 y + (\mu - \mu_0)C_2 y, \end{aligned}$$

where the shift (λ_0, μ_0) is the arithmetic mean of all the eigenvalues. Figure 5.1 shows the distribution of eigenvalues obtained for $n = 100$.

For the following numerical examples, we use GMRES instead of MINRES in the correction equation with orthogonal projections because MINRES is not standardly available in MATLAB 5.3.

Example 5.8.1 In the first example we use the Jacobi–Davidson type method for the exterior eigenvalues. Our goal is to compute the eigenvalue $(\lambda_{\max}, \mu_{\max})$ with the maximum value of λ. We are interested in the number of iterations that the Jacobi–Davidson method needs for sufficiently accurate approximations and also in the percentage of the convergence to the eigenvalue $(\lambda_{\max}, \mu_{\max})$ for a test set of 250 different initial vectors.

We test both alternatives for the correction equations using various numbers of GMRES steps. Each combination is tested on the same set of 250 random initial vectors.

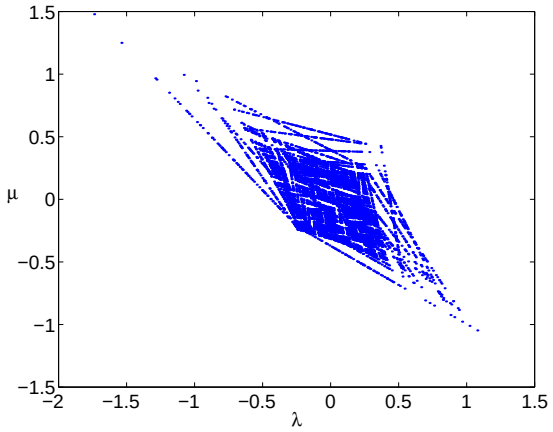


FIGURE 5.1: Distribution of eigenvalues for a right definite two-parameter problem of size $n = 100$.

The algorithm is restarted after every 10 iterations with the current eigenvector approximation, so $l_{\max} = 10$ and $l_{\min} = 1$. The value $\varepsilon = 10^{-8}$ is used for the test of convergence, and flops count in MATLAB are used for a measure of time complexity.

TABLE 5.1: Statistics of the Jacobi–Davidson type method for the eigenvalue $(\lambda_{\max}, \mu_{\max})$ using different correction equations and number of GMRES steps for right definite two-parameter problems of size $n = 100$ and $n = 200$: average number of outer iterations, percentage of convergence to $(\lambda_{\max}, \mu_{\max})$, and average number of flops over 250 trials with different random initial vectors. Correction equations: Orth(m) stands for orthogonal projections and m steps of GMRES, Obli(m) stands for oblique projections and m steps of GMRES.

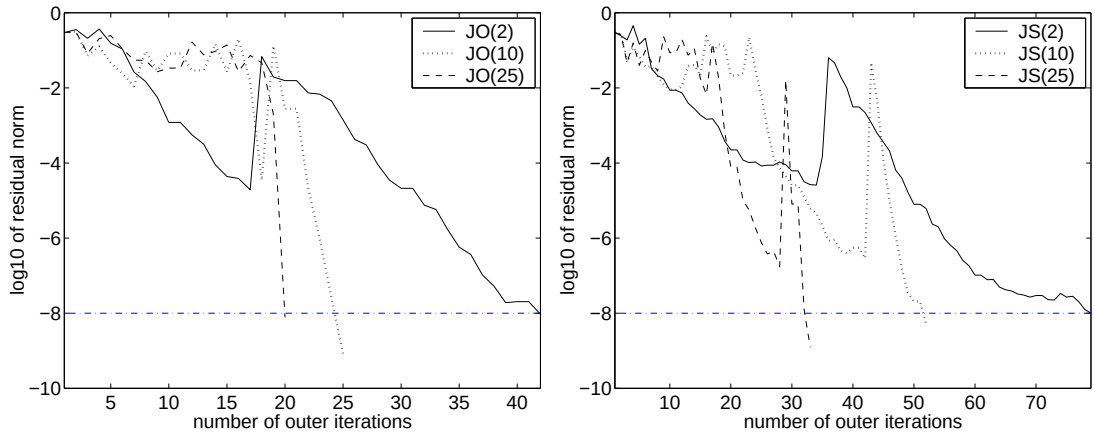
method	$n = 100$			$n = 200$		
	iter	%	flops	iter	%	flops
Orth(1)=Obli(1)	105.4	100.0 %	$4.6 \cdot 10^8$	68.9	100.0 %	$3.4 \cdot 10^8$
Orth(2)	50.0	100.0 %	$2.2 \cdot 10^8$	35.6	100.0 %	$2.0 \cdot 10^8$
Orth(4)	26.7	100.0 %	$1.1 \cdot 10^8$	25.7	100.0 %	$1.6 \cdot 10^8$
Orth(8)	23.3	99.2 %	$1.1 \cdot 10^8$	27.7	99.2 %	$2.1 \cdot 10^8$
Orth(16)	25.4	30.0 %	$1.4 \cdot 10^8$	34.0	48.4 %	$3.6 \cdot 10^8$
Orth(32)	29.8	38.0 %	$2.2 \cdot 10^8$	42.8	10.4 %	$7.2 \cdot 10^8$
Orth(64)	33.1	28.0 %	$4.0 \cdot 10^8$	51.6	9.6 %	$16.0 \cdot 10^8$
Obli(2)	96.4	100.0 %	$4.6 \cdot 10^8$	94.4	100.0 %	$6.1 \cdot 10^8$
Obli(4)	99.9	100.0 %	$5.0 \cdot 10^8$	92.9	100.0 %	$6.6 \cdot 10^8$
Obli(8)	63.9	100.0 %	$3.3 \cdot 10^8$	62.4	100.0 %	$5.2 \cdot 10^8$
Obli(16)	45.2	94.0 %	$2.6 \cdot 10^8$	53.5	98.4 %	$6.0 \cdot 10^8$
Obli(32)	41.9	82.4 %	$3.2 \cdot 10^8$	55.4	70.8 %	$9.6 \cdot 10^8$
Obli(64)	39.7	66.0 %	$4.9 \cdot 10^8$	56.0	35.6 %	$17.6 \cdot 10^8$

Table 5.1 contains results obtained for $n = 100$ and $n = 200$. Orth(m) and Obli(m) denote that m steps of GMRES are used for the correction equation with orthogonal projections or with oblique projections, respectively. For each combination, we list the average number of outer iterations for convergence, the percentage of eigenvalues that converged to the eigenvalue $(\lambda_{\max}, \mu_{\max})$, and the average number of flops in MATLAB, all obtained on the same set of 250 different initial vectors.

The results in Table 5.1 indicate that the method is likely to converge to an unwanted eigenvalue if we solve the correction equation too accurately, i.e., if too many GMRES steps are used to solve the correction equation. A comparison of the flops suggests that the best approach is to do a few steps of GMRES. We also see that, for larger n, the number of GMRES steps has more impact on the time complexity than the number of outer iterations. The reason is that for larger n the factor k^6 becomes relatively smaller compared with mn^2.

The correction equations with orthogonal projections behave similarly to the one with oblique projections but require fewer operations. The experiments suggest to use the correction equations with orthogonal projections in combination with a small number of GMRES steps in each outer iteration for $(\lambda_{\max}, \mu_{\max})$. ⊘

Example 5.8.2 In the second example, the convergence to the exterior eigenvalue for the two-parameter problem of dimension $n = 100$ and initial vectors $u = v = [1 \ \cdots \ 1]^T$ is examined. We compare the convergence for 2, 10, and 25 GMRES steps per iteration for the correction equation with orthogonal and the one with oblique projections, respectively. Figure 5.2 shows the residual norm ρ_k (5.5.2) versus the outer iteration number k. In all six cases, the Ritz values converge to the eigenvalue $(\lambda_{\max}, \mu_{\max})$.



(a) (b)

FIGURE 5.2: Convergence plot for the exterior eigenvalue $(\lambda_{\max}, \mu_{\max})$ for $n = 100$ and $u = v = [1 \ \cdots \ 1]^T$. The plots show the residual norm ρ_k (5.5.2) versus the outer iteration number k for the Jacobi–Davidson type method for the eigenvalue $(\lambda_{\max}, \mu_{\max})$ using 2 (solid line), 10 (dotted line), and 25 (dashed line) GMRES steps to solve the correction equation with orthogonal projections (left plot) and oblique projections (right plot), respectively.

It is clear from Figure 5.2 that convergence near the solution is faster if more GMRES steps are used. Experiments indicate that, if only a few steps of GMRES are applied, then the convergence near the solution is about linear; this is similar to the Jacobi–Davidson method for the standard eigenvalue problem [75, p. 419]. ⊘

Example 5.8.3 In this example, we examine the convergence of the Jacobi–Davidson type method for the interior eigenvalues. We look for the eigenvalue closest to $(0, 0)$. We

use the same $n = 100$ two-parameter problem as in Example 5.8.1 and again test both correction equations with different number of GMRES steps on a set of 250 different initial vectors. The algorithm is restarted after every 10 iterations with the current eigenvector approximation. For the convergence test, we take $\varepsilon = 10^{-6}$. The reason for a more relaxed criterion is an irregular convergence of the interior eigenvalues (see the peaks in Figure 5.3).

TABLE 5.2: Statistics of the Jacobi–Davidson type method for the eigenvalue closest to $(0, 0)$ using different correction equations and different inner iteration processes for a right definite two-parameter problem of size $n = 100$: average number of iterations, percentage of convergence to the eigenvalue closest to $(0, 0)$, and average number of flops over 250 trials with different random initial vectors. Correction equations: Orth(m) stands for orthogonal projections and m steps of GMRES, Obli(m) stands for oblique projections and m steps of GMRES.

method	iter	%	flops
Orth(90)	15.2	80.8 %	$2.4 \cdot 10^8$
Orth(80)	15.9	89.2 %	$2.2 \cdot 10^8$
Orth(70)	18.9	90.0 %	$2.4 \cdot 10^8$
Orth(60)	23.3	91.2 %	$2.5 \cdot 10^8$
Orth(50)	32.8	79.6 %	$3.2 \cdot 10^8$
Orth(40)	41.4	81.6 %	$3.5 \cdot 10^8$
Orth(30)	76.5	72.8 %	$5.8 \cdot 10^8$
Orth(20)	219.2	63.2 %	$14.4 \cdot 10^8$
Obli(90)	20.2	92.4 %	$4.7 \cdot 10^8$
Obli(80)	21.1	96.4 %	$4.3 \cdot 10^8$
Obli(70)	24.2	95.6 %	$4.4 \cdot 10^8$
Obli(60)	29.0	94.4 %	$4.7 \cdot 10^8$
Obli(50)	38.1	93.2 %	$5.4 \cdot 10^8$
Obli(40)	47.0	93.2 %	$5.7 \cdot 10^8$
Obli(30)	82.9	94.0 %	$8.5 \cdot 10^8$
Obli(20)	239.7	84.0 %	$20.5 \cdot 10^8$

The results, presented in Table 5.2, show that the method may also be used effectively for interior eigenvalues. In contrast to Example 5.8.1, more GMRES steps are required for one outer iteration step. If too many steps are applied, then the process converges to an unwanted eigenvalue, similar to Example 5.8.1. On the other hand, if we do not take enough GMRES steps, then we need many outer iteration steps, and the results may be worse. This is different from Example 5.8.1, where the process converges in reasonable time even if only one GMRES step is applied per Jacobi–Davidson iteration step. The correction equation with oblique projections is more effective than the one with orthogonal projections. It is more expensive, but the probability of coming close to the eigenvalue closest to $(0, 0)$ is higher.

$\oslash$

Example 5.8.4 We examine the convergence to the eigenvalue closest to $(0, 0)$ for the two-parameter problem of size $n = 100$ and initial vectors $u = v = [1 \ \cdots \ 1]^T$. Figure 5.3 shows the residual norm ρ_k (5.5.2) versus the outer iteration number k. We compare 40,

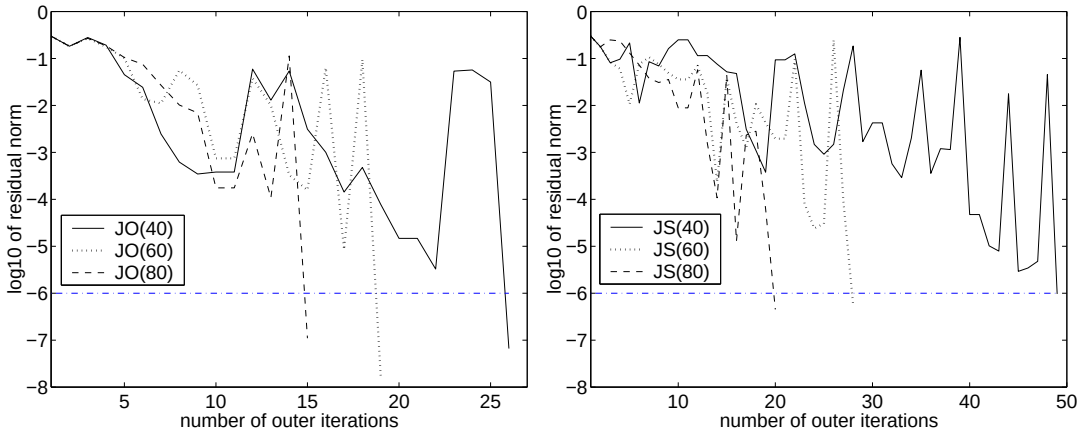


FIGURE 5.3: Convergence plot for the eigenvalue closest to $(0,0)$ for $n = 100$ and $u = v = [1 \cdots 1]^T$. The plots show the residual norm ρ_k (5.5.2) versus the outer iteration number k for the Jacobi–Davidson type method for the eigenvalue closest to $(0,0)$ using 40 (solid line), 60 (dotted line), and 80 (dashed line) GMRES steps to solve the correction equation with orthogonal projections (left plot) and oblique projections (right plot), respectively.

60, and 80 GMRES steps for the correction equation with orthogonal and with oblique projections, respectively. In all six cases, the Ritz values converge to the eigenvalue closest to $(0,0)$. We observe that the more GMRES steps are taken, the fewer iteration steps are needed. The convergence is not as smooth as in Figure 5.2 for Example 5.8.2, but the algorithm is clearly useful for interior eigenvalues. ⊘

Example 5.8.5 In the last example, we test the selection technique from Section 5.5 for computing more eigenpairs for the two-parameter problem of dimension $n = 100$. With 5 GMRES steps for the correction equation with orthogonal projections, we try to compute 30 successive eigenvalues with the maximum value of λ. Figure 5.4 shows how well the first 15 and all 30 computed eigenvalues agree with the desired eigenvalues, respectively.

The eigenvalues are not necessarily computed in the same order as their λ values. This explains the situation in Figure 5.4, where some eigenvalues that are in the top 30 by their λ values are not among the 30 computed eigenvalues. In order to obtain the top k eigenvalues with high probability, it is therefore advisable to always compute more than k eigenvalues. ⊘

5.9 Conclusions

We have presented a new Jacobi–Davidson type method for the right definite two-parameter eigenvalue problem. It has several advantages over the existing methods. It can compute selected eigenpairs, and it does not require good initial approximations. Probably the most important advantage is that it can tackle very large two-parameter problems, especially if the matrices A_i, B_i, and C_i are sparse.

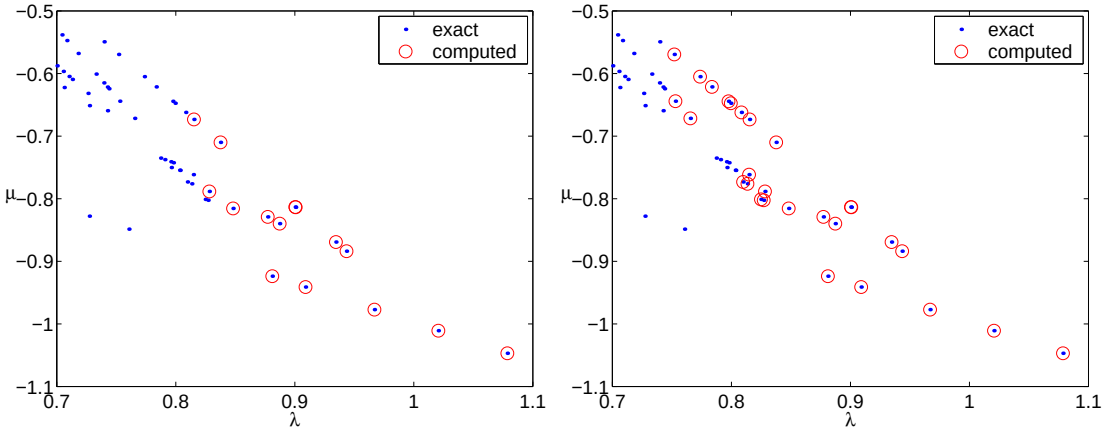


FIGURE 5.4: First 15 (left plot) and first 30 (right plot) computed eigenvalues with maximum value of λ for a two-parameter problem of size $n = 100$ computed using selection for Ritz vectors. The Jacobi–Davidson type method used 5 GMRES steps for the correction equation with orthogonal projections.

We have proposed two correction equations. On the one hand, orthogonal projections are generally more stable than oblique projections, and, in addition, orthogonal projections preserve symmetry. On the other hand, the correction equation with oblique projections can be viewed as an inexact Newton scheme which guarantees asymptotically quadratic convergence. Numerical results indicate that the correction equation with oblique projections is more reliable but more expensive. It is therefore more suitable for the interior eigenvalues, while the one with orthogonal projections may be used for the exterior eigenvalues.

Numerical results indicate that the probability of misconvergence is low when parameters are optimal. The number of GMRES steps is important. Experiments suggest to take up to 5 GMRES steps for exterior eigenvalues and more GMRES steps for interior eigenvalues. Restarts also impact the behavior of the method. In our experiments, we restart the method after every 10 iterations with the current eigenvector approximations, but a different setting may further improve the method.

Because standard deflation techniques for an one-parameter problem cannot be applied to two-parameter problems, we came up with a new selection technique for Ritz vectors.

Chapter 6

A Jacobi–Davidson type method for the two-parameter eigenvalue problem

Abstract. We present a new numerical method for computing selected eigenvalues and eigenvectors of the two-parameter eigenvalue problem. The method does not require good initial approximations and is able to tackle large problems that are too expensive for methods that compute all eigenvalues. The new method uses a two-sided approach and is a generalization of the Jacobi–Davidson type method for the right definite two-parameter eigenvalue problems (Chapter 5). In this chapter, we consider the much wider class of nonsingular problems. In each step we first compute Petrov triples of a small projected two-parameter eigenvalue problem and then expand the left and right search spaces using approximate solutions of appropriate correction equations. The use of a selection enables us to compute more than one eigenpair. Some numerical examples are presented.

Key words: two-parameter eigenvalue problem, subspace method, Jacobi–Davidson method, correction equation, Petrov–Galerkin, two-sided approach.

AMS subject classification: 65F15, 15A18, 15A69.

6.1 Introduction

In this section, we partly repeat the setting of the previous chapter. We are interested in computing one or more eigenpairs of the *two-parameter eigenvalue problem*

$$\begin{aligned} A_1 x_1 &= \lambda B_1 x_1 + \mu C_1 x_1, \\ A_2 x_2 &= \lambda B_2 x_2 + \mu C_2 x_2, \end{aligned} \tag{6.1.1}$$

where A_i, B_i, and C_i are given $n_i \times n_i$ matrices over $\mathbb{C}$, $\lambda, \mu \in \mathbb{C}$ and $x_i \in \mathbb{C}^{n_i}$ for $i = 1, 2$. A pair (λ, μ) is called an *eigenvalue* if it satisfies (6.1.1) for nonzero vectors x_1, x_2. The

*Based on joint work with Tomaž Košir and Bor Plestenjak, see Section 1.5.

tensor product $x_1 \otimes x_2$ is then the corresponding *right eigenvector*. Similarly, $y_1 \otimes y_2$ is the corresponding *left eigenvector* if $0 \neq y_i \in \mathbb{C}^{n_i}$ and $y_i^*(A_i - \lambda B_i - \mu C_i) = 0$ for $i = 1, 2$.

Two-parameter problems can be expressed as two coupled generalized eigenvalue problems as follows. On the tensor product space $S := \mathbb{C}^{n_1} \otimes \mathbb{C}^{n_2}$ of the dimension $N := n_1 n_2$ we define (see (5.1.3))

$$\begin{aligned} \Delta_0 &= B_1 \otimes C_2 - C_1 \otimes B_2, \\ \Delta_1 &= A_1 \otimes C_2 - C_1 \otimes A_2, \\ \Delta_2 &= B_1 \otimes A_2 - A_1 \otimes B_2 \end{aligned}$$

We assume that the two-parameter problem (6.1.1) is *nonsingular*, that is, the corresponding operator determinant Δ_0 is invertible. In this case $\Gamma_1 := \Delta_0^{-1}\Delta_1$ and $\Gamma_2 := \Delta_0^{-1}\Delta_2$ commute and problem (6.1.1) is equivalent to the associated problem

$$\begin{aligned} \Delta_1 z &= \lambda \Delta_0 z, \\ \Delta_2 z &= \mu \Delta_0 z \end{aligned} \tag{6.1.2}$$

for decomposable tensors $z \in S$, $z = x \otimes y$. The left and right eigenvectors of (6.1.1) are Δ_0-orthogonal; i.e., if $x_1 \otimes x_2$ and $y_1 \otimes y_2$ are right and left eigenvector of (6.1.1), respectively, corresponding to distinct eigenvalues, then (cf. (5.1.5))

$$(y_1 \otimes y_2)^* \Delta_0 (x_1 \otimes x_2) = \begin{vmatrix} y_1^* B_1 x_1 & y_1^* C_1 x_1 \\ y_2^* B_2 x_2 & y_2^* C_2 x_2 \end{vmatrix} = 0.$$

If (λ, μ) is an eigenvalue of (6.1.1) then

$$\dim \left(\bigcap_{\substack{i_1 + i_2 = N \\ i_1, i_2 \geq 0}} \ker \left[(\Gamma_1 - \lambda I)^{i_1} (\Gamma_2 - \mu I)^{i_2} \right] \right) \tag{6.1.3}$$

is the *algebraic multiplicity* of (λ, μ). We say that (λ, μ) is *algebraically simple* when its algebraic multiplicity is one. The following lemma is a consequence of in [47, Lemma 3].

Lemma 6.1.1 *If $\boldsymbol{\lambda}$ is an algebraically simple eigenvalue of the two-parameter eigenvalue problem (6.1.1) and $x_1 \otimes x_2$ and $y_1 \otimes y_2$ are the corresponding right and left eigenvector, respectively, then the matrix*

$$\begin{bmatrix} y_1^* B_1 x_1 & y_1^* C_1 x_1 \\ y_2^* B_2 x_2 & y_2^* C_2 x_2 \end{bmatrix}$$

is nonsingular.

There exist some numerical methods for two-parameter eigenvalue problems. Most of them require that the problem is real and *right definite*, i.e., that all matrices A_i, B_i, and C_i are real symmetric and that Δ_0 is positive definite. One of the algorithms (also useful for large sparse matrices) for the right definite two-parameter problem is a Jacobi–Davidson type method (see Chapter 5) and ideas from this method are generalized in this chapter to handle all nonsingular two-parameter problems.

One possible approach to solve (6.1.1) is to solve the associated couple of generalized problems (6.1.2). In the right definite case this can be achieved by numerical methods for simultaneous diagonalization of commutative symmetric matrices [79, 42, 14], while an algorithm for the general nonsingular case using the QZ algorithm is presented in this chapter. Solving the problem via the associated problem is only feasible for problems of low dimension as the size of the matrices of the associated problem is $N \times N$.

Another method that can be used for non right definite two-parameter problems of moderate size is Newton's method [11], which has the deficiency that it requires initial approximations close enough to the solution in order to avoid misconvergence. The continuation method [65] can be used for *weakly elliptic* problems, i.e. such that A_i, B_i and C_i are real symmetric and one of B_i, C_i is positive definite. We mention that right definite two-parameter problems are also weakly elliptic [64, Lemma 2.1].

In this chapter, we introduce a new Jacobi–Davidson type method that can be used to compute selected eigenpairs. The method works does not need close initial approximations and is suitable for large sparse matrices. Our method computes the eigenvalue (λ, μ) of (6.1.1), which is closest to a given target $(\lambda_{\mathrm{T}}, \mu_{\mathrm{T}})$, i.e., the one with minimum $(\lambda - \lambda_{\mathrm{T}})^2 + (\mu - \mu_{\mathrm{T}})^2$.

The outline of this chapter is as follows. In Section 6.2, we present a new algorithm for the computation of eigenpairs using the associated problem. This method is only suitable for matrices of moderate size, so we combine it with a subspace method. We generalize the Petrov–Galerkin approach to two-parameter eigenvalue problems in Section 6.3. In Section 6.4, we present a two-sided Jacobi–Davidson type method for two-parameter eigenvalue problems. Several possible correction equations are discussed in Section 6.5. In Section 6.6, we present a selection technique that allows the computation of more than one eigenpair. The time complexity is given in Section 6.7, and some numerical examples are presented in Section 6.8. We give some conclusions in Section 6.9.

6.2 Algorithm based on the associated problem

We propose the following method to solve the associated problem (6.1.2). First we compute a QZ decomposition (generalized Schur form) of the matrix pencil (Δ_1, Δ_0). We obtain unitary matrices Q and Z such that $Q^*\Delta_0 Z = R$ and $Q^*\Delta_1 Z = S$ are upper triangular. Since Δ_0 is nonsingular, the same is true for R. From

$$\Delta_0^{-1}\Delta_1 = ZR^{-1}SZ^*$$

it follows that the eigenvalues of the first generalized eigenvalue problem in (6.1.2) are the quotients s_{ii}/r_{ii} of the diagonal elements of matrices S and R.

Next, we sort the generalized Schur form so that multiple eigenvalues of the first generalized eigenvalue problem in (6.1.2) appear in blocks (see for instance [94]). Let us assume that the generalized Schur form is sorted to meet this requirement and let

matrix $R^{-1}S$ be partitioned accordingly as

$$R^{-1}S = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1p} \\ 0 & L_{22} & \cdots & L_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_{pp} \end{bmatrix}. \tag{6.2.1}$$

In the above partition, multiple eigenvalues of $\Delta_0^{-1}\Delta_1$ are clustered in upper triangular matrices $L_{11}, \ldots, L_{pp}$ along the diagonal so that $\lambda(L_{ii}) \neq \lambda(L_{jj})$ for $i \neq j$, where $\lambda(L_{kk})$ is the eigenvalue of a block L_{kk}. Let us denote the size of L_{ii} by m_i for $i = 1, \ldots, p$.

Lemma 6.2.1 *Let*

$$L = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1p} \\ 0 & L_{22} & \cdots & L_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_{pp} \end{bmatrix}$$

be a partitioning of a block upper triangular matrix L such that $\Lambda(L_{11}), \ldots, \Lambda(L_{pp})$ are mutually disjoint, where $\Lambda(L_{kk})$ is the set of eigenvalues of L_{kk}. If M commutes with L then M is block upper triangular partitioned conformally with L.

Proof: First we study the case $p = 2$. Let M be partitioned conformally with L as

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

From $LM - ML = 0$ and the above assumption we obtain the equation $L_{22}M_{21} - M_{21}L_{11} = 0$. Because L_{11} and L_{22} have no eigenvalues in common, this is a nonsingular homogeneous Sylvester equation for M_{21} (see for example [83, p. 223]). Therefore, the unique solution is $M_{21} = 0$.

In case $p > 2$ one can see that M is block upper triangular by applying the above argument on all appropriate 2×2 block partitions of L and M. □

Lemma 6.2.2 *$T = Q^*\Delta_2 Z$ partitioned conformally with (6.2.1) is block upper triangular.*

Proof: As $\Delta_0^{-1}\Delta_1$ and $\Delta_0^{-1}\Delta_2$ commute, so do $R^{-1}S$ and $R^{-1}T$. It follows from Lemma 6.2.1 that $R^{-1}T$ is block upper triangular partitioned conformally to (6.2.1). As block upper triangular matrices keep their shape when multiplied by a triangular matrix, it follows from $T = R(R^{-1}T)$ that T is block upper triangular as well. □

Once R, S and T are partitioned conformally with (6.2.1) as

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ 0 & R_{22} & \cdots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{pp} \end{bmatrix}, \; S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ 0 & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_{pp} \end{bmatrix}, \; T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1p} \\ 0 & T_{22} & \cdots & T_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_{pp} \end{bmatrix},$$

it is straightforward to compute eigenvalues of (6.1.1). To each diagonal block L_{ii} of size m_i in $R^{-1}S$ correspond m_i eigenvalues $(\lambda_i, \mu_{i1}), \ldots, (\lambda_i, \mu_{im_i})$, where λ_i is the eigenvalue of L_{ii} and $\mu_{i1}, \ldots, \mu_{im_i}$ are eigenvalues of the generalized eigenvalue problem $T_{ii}w = \mu R_{ii}w$.

Now that we have all eigenvalues (λ_j, μ_j), $j = 1, \ldots, N$, of (6.1.1) we compute the corresponding eigenvectors $x_{j1} \otimes x_{j2}$. We do this by solving $(A_i - \lambda_j B_i - \mu_j C_i)x_{ji} = 0$, where x_{ji} is normalized, for $i = 1, 2$. In a similar way we can obtain left eigenvectors $y_{j1} \otimes y_{j2}$ when they are required. The complete procedure is summarized in Algorithm 6.2.1.

Input: A nonsingular two-parameter eigenvalue problem (6.1.1)
Output: Eigenpairs $((\lambda_j, \mu_j), x_j \otimes y_j)$ $(j = 1, \ldots, N)$

1. Compute Δ_0, Δ_1 and Δ_2 of the associated problem (6.1.2)
2. Compute the sorted generalized Schur decomposition $Q^*\Delta_0 Z = R$ and $Q^*\Delta_1 Z = S$ (multiple values of $\lambda_i := s_{ii}/r_{ii}$ clustered along the diagonal of $R^{-1}S$)
3. Compute diagonal blocks $T_{11}, \ldots, T_{pp}$ of $T = Q^*\Delta_2 Z$, partitioned conformally with R and S
4. Compute the eigenvalues $\mu_{i1}, \ldots, \mu_{im_i}$ of $T_{ii}w = \mu R_{ii}w$ for $i = 1, \ldots, p$
5. The eigenvalues of (6.1.1) are $(\lambda_1, \mu_{11}), \ldots, (\lambda_1, \mu_{1m_1}); \ldots; (\lambda_p, \mu_{p1}), \ldots, (\lambda_p, \mu_{pm_p})$, reindex them as $(\lambda_1, \mu_1), \ldots, (\lambda_N, \mu_N)$.
6. For each eigenvalue (λ_j, μ_j), $j = 1, \ldots, N$, take for x_{ji} and y_{ji} the smallest right and left singular vector of $A_i - \lambda_j B_i - \mu_j C_i$, respectively, for $i = 1, 2$

ALGORITHM 6.2.1: An algorithm for the nonsingular two-parameter eigenvalue problem

Remark 6.2.3 In numerical computation we may cluster not only multiple eigenvalues but also clustered eigenvalues of $R^{-1}S$. After clustering we take the mean of all eigenvalues in the cluster of size m_i as a multiple eigenvalue of order m_i. This means that we take λ_i as a mean of all eigenvalues of the generalized eigenvalue problem

$$S_{ii}w = \lambda R_{ii}w$$

for $i = 1, \ldots, p$. ⊘

Remark 6.2.4 In practice there will be an error in a detected eigenvalue (λ_j, μ_j). Because of that, we take, in Step 6 of Algorithm 6.2.1, the smallest left and right singular vector to find an approximation to the eigenvectors x_{ji} and y_{ji}. ⊘

Let as assume that A_i, B_i, C_i are dense and that $n_1 = n_2 = n$. The time complexity of Algorithm 6.2.1 is $\mathcal{O}(n^6)$ for the computation of eigenvalues using the QZ decomposition of matrices of size n^2. The maximum additional work for eigenvectors is $\mathcal{O}(n^5)$ as we have to compute $\mathcal{O}(n^2)$ singular value decompositions of matrices of size n. If we are not interested in all eigenvectors (as is often the case for large sparse matrices) then the additional work can be substantially smaller.

The large time complexity is the reason that Algorithm 6.2.1 is useful only for matrices of a modest size. For larger problems we embed this method in a subspace method and use Algorithm 6.2.1 for the small projected problems.

6.3 Subspace methods and Petrov triples

Now we study a Jacobi–Davidson type subspace method for the two-parameter eigenvalue problem. In this section we discuss the extraction, in the next section the algorithm and the expansion.

Suppose that we have k-dimensional search spaces $\mathcal{U}_{ik} \subset \mathbb{C}^{n_i}$ and k-dimensional test spaces $\mathcal{V}_{ik} \subset \mathbb{C}^{n_i}$ for $i = 1, 2$. Let the columns of the $n_i \times k$ matrices U_{ik} and V_{ik} form orthogonal bases for $\mathcal{U}_{ik}$ and $\mathcal{V}_{ik}$, respectively, for $i = 1, 2$. The *Petrov–Galerkin conditions* on the *residuals* (cf. (5.2.1))

$$\begin{aligned} r_1 &:= (A_1 - \sigma B_1 - \tau C_1)u_1 \quad \perp \quad \mathcal{V}_{1k}, \\ r_2 &:= (A_2 - \sigma B_2 - \tau C_2)u_2 \quad \perp \quad \mathcal{V}_{2k}, \end{aligned} \tag{6.3.1}$$

where $u_i \in \mathcal{U}_{ik} \backslash \{0\}$ for $i = 1, 2$, lead to the smaller projected two-parameter problem (cf. (5.2.2))

$$\begin{aligned} V_{1k}^* A_1 U_{1k} c_1 &= \sigma V_{1k}^* B_1 U_{1k} c_1 + \tau V_{1k}^* C_1 U_{1k} c_1, \\ V_{2k}^* A_2 U_{2k} c_2 &= \sigma V_{2k}^* B_2 U_{2k} c_2 + \tau V_{2k}^* C_2 U_{2k} c_2, \end{aligned} \tag{6.3.2}$$

where $u_i = U_{ik} c_i \neq 0$ for $i = 1, 2$ and $\sigma, \tau \in \mathbb{C}$.

We say that an eigenvalue (σ, τ) of (6.3.2) is a *Petrov value* for the two-parameter eigenvalue problem (6.1.1) with respect to the search spaces $\mathcal{U}_{1k}$ and $\mathcal{U}_{2k}$ and test spaces $\mathcal{V}_{1k}$ and $\mathcal{V}_{2k}$. If (σ, τ) is an eigenvalue of (6.3.2) and $c_1 \otimes c_2$ is the corresponding right eigenvector, then $u_1 \otimes u_2$ is a *right Petrov vector*, where $u_i = U_{ik} c_i$ for $i = 1, 2$. Similarly, if $d_1 \otimes d_2$ is the corresponding left eigenvector of (6.3.2) then $v_1 \otimes v_2$ is a *left Petrov vector*, where $v_i = V_{ik} d_i$ for $i = 1, 2$. It is easy to check that σ and τ are equal to the *two-sided tensor Rayleigh quotients* (cf. (5.2.3))

$$\begin{aligned} \sigma = \rho_1(u, v) &= \frac{(v_1 \otimes v_2)^* \Delta_1 (u_1 \otimes u_2)}{(v_1 \otimes v_2)^* \Delta_0 (u_1 \otimes u_2)} = \frac{(v_1^* A_1 u_1)(v_2^* C_2 u_2) - (v_1^* C_1 u_1)(v_2^* A_2 u_2)}{(v_1^* B_1 u_1)(v_2^* C_2 u_2) - (v_1^* C_1 u_1)(v_2^* B_2 u_2)}, \\ \tau = \rho_2(u, v) &= \frac{(v_1 \otimes v_2)^* \Delta_2 (u_1 \otimes u_2)}{(v_1 \otimes v_2)^* \Delta_0 (u_1 \otimes u_2)} = \frac{(v_1^* B_1 u_1)(v_2^* A_2 u_2) - (v_1^* A_1 u_1)(v_2^* B_2 u_2)}{(v_1^* B_1 u_1)(v_2^* C_2 u_2) - (v_1^* C_1 u_1)(v_2^* B_2 u_2)}. \end{aligned} \tag{6.3.3}$$

In order to obtain Petrov values, we have to solve small two-parameter eigenvalue problems. For this purpose, we use Algorithm 6.2.1. Altogether, we obtain k^2 *Petrov triples* $((\sigma_j, \tau_j), u_{j1} \otimes u_{j2}, v_{j1} \otimes v_{j2})$ that are approximations to eigentriples $((\lambda_j, \mu_j), x_{j1} \otimes x_{j2}, y_{j1} \otimes y_{j2})$ of (6.1.1) for $j = 1, \ldots, k^2$.

6.4 A Jacobi–Davidson type method

The Jacobi–Davidson method [75] is one of the subspace methods that may be used for the numerical solution of one-parameter eigenvalue problems. In the Jacobi–Davidson method approximate solutions of certain correction equations are used to expand the search space. The search for a new direction is restricted to the subspace that is orthogonal or oblique to the last chosen right (or left) Petrov vector.

A Jacobi–Davidson type method has been successfully applied to the right definite two-parameter eigenvalue problem (see Chapter 5). In this chapter we show that a Jacobi–Davidson type method can be applied to a general two-parameter eigenvalue problem as well. Numerical experiments (see Example 6.8.1) indicate that one-sided Jacobi–Davidson (where, as in Chapter 5, the search spaces $\mathcal{V}_i$ in (6.2.1) are the same as the test spaces $\mathcal{U}_i$) is not accurate enough for non right definite two-parameter eigenvalue problems. Therefore, we generalize the two-sided Jacobi–Davidson method (see Chapter 2) to two-parameter eigenvalue problems. The idea is to take $\mathcal{U}_i$ as search spaces for the right eigenvectors and $\mathcal{V}_i$ as search spaces for the left eigenvectors.

A brief sketch of the two-sided Jacobi–Davidson type method for the two-parameter problem is presented in Algorithm 6.4.2. In Step 4 we have to choose a Petrov triple. Some options are given later in this section. In Step 8, we have to find new search directions in order to expand the search and test subspaces. We discuss several possible correction equations in Section 6.5.

Input: initial vectors u_1, u_2, v_1, and v_2 with unit norm
Output: an approximate eigenpair satisfying $(\|r_1^R\|^2 + \|r_2^R\|^2)^{1/2} \leq \varepsilon$

1. $s_i = u_i$, $t_i = v_i$, $U_{i,0} = [\,]$, $V_{i,0} = [\,]$, for $i = 1, 2$.
   for $k = 1, \ldots, k_{\max}$ do:
2. Expand the search subspaces for $i = 1, 2$
   $U_{i,k} = \text{MGS}\ (U_{i,k-1}, s_i)$,
   $V_{i,k} = \text{MGS}\ (V_{i,k-1}, t_i)$
3. Solve the projected two-parameter eigenvalue problem
   $V_{1k}^* A_1 U_{1k} c_1 = \sigma V_{1k}^* B_1 U_{1k} c_1 + \tau V_{1k}^* C_1 U_{1k} c_1$,
   $V_{2k}^* A_2 U_{2k} c_2 = \sigma V_{2k}^* B_2 U_{2k} c_2 + \tau V_{2k}^* C_2 U_{2k} c_2$
   by Algorithm 6.2.1
4. Select an appropriate Petrov value (σ, τ) and the corresponding right and left Petrov vectors $u_1 \otimes u_2$ and $v_1 \otimes v_2$, where $u_i = U_{ik} c_i$, $v_i = V_{ik} d_i$ for $i = 1, 2$, respectively
5. Compute the right and left residuals for $i = 1, 2$
   $r_i^R = (A_i - \sigma B_i - \tau C_i) u_i$,
   $r_i^L = (A_i - \sigma B_i - \tau C_i)^* v_i$
6. Stop if $\rho_k \leq \varepsilon$, where $\rho_k = \left(\|r_1^R\|^2 + \|r_2^R\|^2\right)^{1/2}$
7. Restart. If the dimension of the image of U_{ik} and V_{ik} exceeds $l_{\max}$, then replace U_{ik}, V_{ik} with new orthonormal bases of dimension $l_{\min}$.
8. Solve approximately one of the proposed correction equations (see Section 6.5) and obtain new directions s_i and t_i for $i = 1, 2$

ALGORITHM 6.4.2: Two-sided Jacobi–Davidson for the nonsingular two-parameter eigenvalue problem

To apply this algorithm, we need to specify a target (λ_T, μ_T), a tolerance ε, a maximum number of steps $k_{\max}$, a maximum dimension of the search subspaces $l_{\max}$, and a number $l_{\min} < l_{\max}$ that specifies the dimension of the search subspaces after a restart.

We also have to specify a criterion for Step 4. Suppose that we are looking for the

eigenvalue closest to the target (λ_T, μ_T). We suggest to combine two approaches. In the first part we select the Petrov value (σ, τ) closest to the target until the residual ρ_k drops below $\varepsilon_{\text{change}}$. Then, in the second part, we take the Petrov triple with the smallest residual norm

$$\left(\|r_1^R\|^2 + \|r_2^R\|^2\right)^{1/2}. \tag{6.4.1}$$

Both stages can be seen as an accelerated inexact Rayleigh quotient iteration.

As Algorithm 6.2.1 is able to solve only low-dimensional two-parameter problems (6.3.2) in a reasonable time, we expand the search spaces up to the preselected dimension $l_{\max}$ and then restart the algorithm. For a restart, we take the $l_{\min}$ eigenvector approximations with the smallest residuals (6.4.1) as a basis for the initial search space.

Remark 6.4.1 In Step 6 we could also stop the algorithm if the norms of the left residuals r_1^L and r_2^L are small enough. If either left or right residuals are small then we can expect (σ, τ) to be a good approximation to an eigenvalue and we can compute the corresponding right or left eigenvectors by solving one (orthogonal) correction equation, see also the discussion in Section 2.4.1. ⊘

In the following section we discuss the expansion in Step 8 and derive several correction equations.

6.5 Correction equations

Let (σ, τ) be a Petrov value that approximates the eigenvalue (λ, μ) of (6.1.1) and let $u_1 \otimes u_2$ and $v_1 \otimes v_2$ be its corresponding left and right Petrov vector, respectively. Let us assume that u_1, u_2, v_1, and v_2 are normalized.

We are searching for improvements of the left and right Petrov vectors of the form (cf. (5.3.8) and (5.3.9))

$$(A_i - \lambda B_i - \mu C_i)\,(u_i + s_i) = 0, \tag{6.5.1}$$
$$(A_i - \lambda B_i - \mu C_i)^*(v_i + t_i) = 0, \tag{6.5.2}$$

where $s_i \perp a_i$ and $t_i \perp b_i$ for $i = 1, 2$. We will discuss the choices for a_i and b_i later, at this time we require just that $a_i \not\perp u_i$ and $b_i \not\perp v_i$.

Using (6.3.1), we can rewrite (6.5.1) and (6.5.2) as (cf. (5.3.10) and (5.3.12))

$$\begin{aligned}(A_i - \sigma B_i - \tau C_i)\, s_i = -r_i^R &+ (\lambda - \sigma)B_i u_i + (\mu - \tau)C_i u_i \\ &+ (\lambda - \sigma)B_i s_i + (\mu - \tau)C_i s_i, \end{aligned}\tag{6.5.3}$$

$$\begin{aligned}(A_i - \sigma B_i - \tau C_i)^* t_i = -r_i^L &+ (\lambda - \sigma)^* B_i^* v_i + (\mu - \tau)^* C_i^* v_i \\ &+ (\lambda - \sigma)^* B_i^* t_i + (\mu - \tau)^* C_i^* t_i. \end{aligned}\tag{6.5.4}$$

Theorem 6.5.1 (cf. Theorem 5.3.2) *If $u_i = x_i - s_i$ and $v_i = y_i - t_i$, for $i = 1, 2$, are close enough approximations to a left and a right eigenvector of (6.1.1) for the same algebraically simple eigenvalue (λ, μ) then the two-sided Rayleigh quotient $(\sigma, \tau) = (\rho_1(u, v), \rho_2(u, v))$ is an $\mathcal{O}(\|s_1\|\|t_1\| + \|s_2\|\|t_2\|)$ approximation to (λ, μ), i.e.,*

$$\left\| \begin{bmatrix} \lambda - \sigma \\ \mu - \tau \end{bmatrix} \right\| = \mathcal{O}(\|s_1\|\|t_1\| + \|s_2\|\|t_2\|). \tag{6.5.5}$$

Proof: We write the residual (6.3.1) as

$$r_i^R = -(A_i - \lambda B_i - \mu C_i)s_i + (\lambda - \sigma)B_i u_i + (\mu - \tau)C_i u_i. \tag{6.5.6}$$

When we multiply (6.5.6) by v_i^* and take into account that $v_i^* r_i^R = 0$ and

$$v_i^*(A_i - \lambda B_i - \mu C_i) = -t_i^*(A_i - \lambda B_i - \mu C_i)$$

for $i = 1, 2$, then we obtain

$$\begin{bmatrix} v_1^* B_1 u_1 & v_1^* C_1 u_1 \\ v_2^* B_2 u_2 & v_2^* C_2 u_2 \end{bmatrix} \begin{bmatrix} \lambda - \sigma \\ \mu - \tau \end{bmatrix} = - \begin{bmatrix} t_1^*(A_1 - \lambda B_1 - \mu C_1)s_1 \\ t_2^*(A_2 - \lambda B_2 - \mu C_2)s_2 \end{bmatrix}. \tag{6.5.7}$$

If $\|s_i\|$ and $\|t_i\|$ are small enough then (6.5.7) is a nonsingular system because of Lemma 6.1.1 and continuity. We can deduce from (6.5.7) that

$$\left\| \begin{bmatrix} \lambda - \sigma \\ \mu - \tau \end{bmatrix} \right\| = \left\| \begin{bmatrix} v_1^* B_1 u_1 & v_1^* C_1 u_1 \\ v_2^* B_2 u_2 & v_2^* C_2 u_2 \end{bmatrix}^{-1} \begin{bmatrix} t_i^*(A_1 - \lambda B_1 - \mu C_1)s_i \\ t_2^*(A_2 - \lambda B_2 - \mu C_2)s_i \end{bmatrix} \right\|$$

and so obtain (6.5.5). □

It follows from Theorem 6.5.1 that asymptotically (i.e., when we have good approximate right and left eigenvectors), we can consider s_i and t_i as first order corrections, $(\lambda - \sigma)B_i u_i + (\mu - \tau)C_i u_i$ and $(\lambda - \sigma)^* B_i^* v_i + (\mu - \tau)^* C_i^* v_i$ as second order corrections, and finally, $(\lambda - \sigma)B_i s_i + (\mu - \tau)C_i s_i$ and $(\lambda - \sigma)^* B_i^* t_i + (\mu - \tau)^* C_i^* t_i$ can be interpreted as third order corrections.

6.5.1 First order based correction equations

If we ignore the second and higher order terms in (6.5.3) then we obtain the equation

$$(A_i - \sigma B_i - \tau C_i)s_i = -r_i^R. \tag{6.5.8}$$

Because r_i^R is orthogonal to v_i, we can multiply (6.5.8) with an oblique projection $\left(I - \frac{c_i v_i^*}{v_i^* c_i}\right)$, where $c_i \not\perp v_i$, that fixes r_i^R. Secondly, since s_i is orthogonal to a_i, we can write $\left(I - \frac{u_i a_i^*}{a_i^* u_i}\right) s_i$ instead of s_i. Thus we obtain the correction equation for the vector u_i

$$\left(I - \frac{c_i v_i^*}{v_i^* c_i}\right)(A_i - \sigma B_i - \tau C_i)\left(I - \frac{u_i a_i^*}{a_i^* u_i}\right) s_i = -r_i^R \tag{6.5.9}$$

for $i = 1, 2$. In a similar way we obtain from (6.5.4) the correction equation for the vector v_i

$$\left(I - \frac{d_i u_i^*}{u_i^* d_i}\right)(A_i - \sigma B_i - \tau C_i)^*\left(I - \frac{v_i b_i^*}{b_i^* v_i}\right) t_i = -r_i^L \tag{6.5.10}$$

for $i = 1, 2$, where $d_i \not\perp u_i$.

We solve these correction equations only approximately, for instance using some Krylov subspace method. Since the operator in (6.5.9) maps $a_i^\perp$ onto $v_i^\perp$, it is suitable to

take $a_i = v_i$ in order to apply Krylov solver without a preconditioner (see, for example, the discussion in Section 2.4.2). If $a_i \neq v_i$, then we need a preconditioner that maps the image space $v_i^{\perp}$ bijectively onto $a_i^{\perp}$. Similarly, we need a preconditioner for (6.5.10) when $b_i \neq u_i$.

Different choices of vectors a_i, b_i, c_i, d_i lead to different correction equations. We discuss some options.

1. For the first correction equation we take $a_i = d_i = v_i$, $b_i = c_i = u_i$. We obtain a pair of correction equations

$$\begin{aligned} \left(I - \frac{u_i v_i^*}{v_i^* u_i}\right)(A_i - \sigma B_i - \tau C_i)\left(I - \frac{u_i v_i^*}{v_i^* u_i}\right) s_i &= -r_i^R, \\ \left(I - \frac{v_i u_i^*}{u_i^* v_i}\right)(A_i - \sigma B_i - \tau C_i)^*\left(I - \frac{v_i u_i^*}{u_i^* v_i}\right) t_i &= -r_i^L \end{aligned} \tag{6.5.11}$$

   for $i = 1, 2$. The operator in the first equation is the conjugate transpose of the operator in the second equation and we can solve these equations simultaneously by bi-conjugate gradients (BiCG). It is also possible to solve equations in (6.5.11) separately by GMRES.

2. For this correction equation we take $a_i = c_i = u_i$, $b_i = d_i = v_i$.

   It is a natural approach for (6.5.9) and (6.5.10) to take $a_i = u_i$ and $b_i = v_i$ as in this case we are looking for updates orthogonal to the current approximation. As it turns out later in Section 6.5.2, when we use preconditioning, an interesting choice for c_i and d_i is to take $c_i = u_i$ and $d_i = v_i$, which leads to a pair of correction equations

$$\begin{aligned} \left(I - \frac{u_i v_i^*}{v_i^* u_i}\right)(A_i - \sigma B_i - \tau C_i)(I - u_i u_i^*) s_i &= -r_i^R, \\ \left(I - \frac{v_i u_i^*}{u_i^* v_i}\right)(A_i - \sigma B_i - \tau C_i)^*(I - v_i v_i^*) t_i &= -r_i^L \end{aligned} \tag{6.5.12}$$

   for $i = 1, 2$.

   In order to solve (6.5.12) approximately by a Krylov solver we need a preconditioner because $a_i^{\perp}$ and $v_i^{\perp}$ do not agree, see Section 6.5.2.

3. In this case we take $a_i = u_i$, $b_i = v_i$, $c_i = g_i$, $d_i = h_i$, where

$$\begin{aligned} g_i &= (\lambda_{\mathrm{T}} - \sigma) B_i u_i + (\mu_{\mathrm{T}} - \tau) C_i u_i, \\ h_i &= (\lambda_{\mathrm{T}} - \sigma)^* B_i^* v_i + (\mu_{\mathrm{T}} - \tau)^* C_i^* v_i. \end{aligned}$$

   The idea behind the choice of c_i and d_i is that when the target $(\lambda_{\mathrm{T}}, \mu_{\mathrm{T}})$ is close to the eigenvalue then the projections with g_i and h_i almost annihilate the second

order terms in equations (6.5.3) and (6.5.4) and thus reduce the neglected quantity. We derive the correction equations

$$\begin{aligned}\left(I - \frac{g_i {v_i}^*}{{v_i}^* g_i}\right)(A_i - \sigma B_i - \tau C_i)(I - u_i {u_i}^*)\, s_i &= -r_i^R, \\ \left(I - \frac{h_i {u_i}^*}{{u_i}^* h_i}\right)(A_i - \sigma B_i - \tau C_i)^*(I - v_i {v_i}^*)\, t_i &= -r_i^L\end{aligned} \tag{6.5.13}$$

for $i = 1, 2$.

Again, if we want to solve (6.5.13) approximately by a Krylov solver then we need a preconditioner as $a_i \neq v_i$, see the next section.

6.5.2 Preconditioned first order based correction equations

We mentioned that we need a preconditioner for a Krylov solver when the domain and the range of the operator in the correction equation do not agree. But we can also use a preconditioner when domain and range do agree to speed up the convergence.

Suppose that a left preconditioner M_i is available for $A_i - \sigma B_i - \mu_i C_i$ such that $M_i^{-1}(A_i - \sigma B_i - \mu_i C_i) \approx I$. A calculation shows that if we assume that $a_i^* M_i^{-1} c_i \neq 0$ then the inverse of the map

$$\left(I - \frac{c_i {v_i}^*}{{v_i}^* c_i}\right) M_i \left(I - \frac{u_i {a_i}^*}{{a_i}^* u_i}\right)$$

from $a_i^\perp$ to $v_i^\perp$ is the map

$$\left(I - \frac{M_i^{-1} c_i a_i^*}{a_i^* M_i^{-1} c_i}\right) M_i^{-1} \left(I - \frac{c_i v_i^*}{v_i^* c_i}\right)$$

from $v_i^\perp$ to $a_i^\perp$. Therefore, using left preconditioning changes (6.5.9) into

$$\begin{aligned}\left(I - \frac{M_i^{-1} c_i a_i^*}{a_i^* M_i^{-1} c_i}\right) M_i^{-1} \left(I - \frac{c_i v_i^*}{v_i^* c_i}\right)(A_i - \sigma B_i - \tau C_i)\left(I - \frac{u_i a_i^*}{a_i^* u_i}\right) s_i = \\ - \left(I - \frac{M_i^{-1} c_i a_i^*}{a_i^* M_i^{-1} c_i}\right) M_i^{-1} r_i^R\end{aligned}$$

for $i = 1, 2$.

Correction equation (6.5.10) for the left eigenvector can be dealt with similarly. A preconditioner for $A_i - \sigma B_i - \tau C_i$ automatically suggests a preconditioner for $(A_i - \sigma B_i - \tau C_i)^*$.

We can combine different preconditioners with different correction equations. Here are some possibilities.

1. Our suggestion for the preconditioner is

$$M_i = A_i - \lambda_{\mathrm{T}} B_i - \mu_{\mathrm{T}} C_i, \tag{6.5.14}$$

where $(\lambda_{\mathrm{T}}, \mu_{\mathrm{T}})$ is the target. Instead of exact inversion we can also take an inexact inverse, for example one obtained using an incomplete LU decomposition.

2. The simplest option is to take the identity as a preconditioner in order to be able to use a Krylov solver for the correction equation. For example, if we take correction equation (6.5.12) and the identity as a preconditioner, then we have to multiply (6.5.9) and (6.5.10) by orthogonal projectors $I - u_i u_i^*$ and $I - v_i v_i^*$, respectively. From $(I - u_i u_i^*)\left(I - \frac{u_i v_i^*}{v_i^* u_i}\right) = I - u_i u_i^*$ and $(I - v_i v_i^*)\left(I - \frac{v_i u_i^*}{u_i^* v_i}\right) = I - v_i v_i^*$ we get

$$\begin{aligned}(I - u_i u_i^*)(A_i - \sigma B_i - \tau C_i)(I - u_i u_i^*)s_i &= -(I - u_i u_i^*)r_i^R, \\ (I - v_i v_i^*)(A_i - \sigma B_i - \tau C_i)^*(I - v_i v_i^*)t_i &= -(I - v_i v_i^*)r_i^L\end{aligned} \tag{6.5.15}$$

for $i = 1, 2$. One can recognize (6.5.15) as the correction equations of standard Jacobi–Davidson applied to $A_i - \sigma B_i - \tau C_i$ and $(A_i - \sigma B_i - \tau C_i)^*$.

6.5.3 Second order based correction equation

For this case we generalize the correction equation with oblique projections for the right definite two-parameter eigenvalue problem (see Section 5.3.2). If we define

$$D = \begin{bmatrix} A_1 - \sigma B_1 - \tau C_1 & 0 \\ 0 & A_2 - \sigma B_2 - \tau C_2 \end{bmatrix},$$

$$r^R = \begin{bmatrix} r_1^R \\ r_2^R \end{bmatrix}, \quad r^L = \begin{bmatrix} r_1^L \\ r_2^L \end{bmatrix},$$

then we can reformulate (6.5.3) and (6.5.4) (neglecting third order correction terms) as

$$D \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = -r^R + (\lambda - \sigma) \begin{bmatrix} B_1 u_1 \\ B_2 u_2 \end{bmatrix} + (\mu - \tau) \begin{bmatrix} C_1 u_1 \\ C_2 u_2 \end{bmatrix} \tag{6.5.16}$$

and

$$D^* \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = -r^L + (\lambda - \sigma)^* \begin{bmatrix} B_1^* v_1 \\ B_2^* v_2 \end{bmatrix} + (\mu - \tau)^* \begin{bmatrix} C_1^* v_1 \\ C_2^* v_2 \end{bmatrix}. \tag{6.5.17}$$

Let V_{R} be a $(n_1 + n_2) \times 2$ matrix with orthonormal columns such that

$$\operatorname{span}(V_{\mathrm{R}}) = \operatorname{span}\left(\begin{bmatrix} B_1 u_1 \\ B_2 u_2 \end{bmatrix}, \begin{bmatrix} C_1 u_1 \\ C_2 u_2 \end{bmatrix} \right)$$

and let

$$W_{\mathrm{R}} = \begin{bmatrix} v_1 & 0 \\ 0 & v_2 \end{bmatrix}.$$

With the oblique projection

$$P_{\mathrm{R}} = I - V_{\mathrm{R}}(W_{\mathrm{R}}^* V_{\mathrm{R}})^{-1} W_{\mathrm{R}}^*$$

onto $\operatorname{span}(W_{\mathrm{R}})^{\perp}$ along $\operatorname{span}(V_{\mathrm{R}})$, it follows that

$$P_{\mathrm{R}} r^R = r^R \quad \text{and} \quad P_{\mathrm{R}} \begin{bmatrix} B_1 u_1 \\ B_2 u_2 \end{bmatrix} = P_{\mathrm{R}} \begin{bmatrix} C_1 u_1 \\ C_2 u_2 \end{bmatrix} = 0.$$

Therefore, from multiplying (6.5.16) by P_{R} we obtain

$$P_{\mathrm{R}} D \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = -r^R.$$

Suppose that we are looking for corrections such that $s_i \perp v_i$ and $t_i \perp u_i$. Then

$$P_{\mathrm{R}} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

and the result is the correction equation

$$P_{\mathrm{R}} D P_{\mathrm{R}} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = -r^R. \tag{6.5.18}$$

Remark 6.5.2 If $u_1 \otimes u_2$ and $v_1 \otimes v_2$ are close approximations to eigenvectors $x_1 \otimes x_2$ and $y_1 \otimes y_2$, corresponding to a single eigenvalue of (6.1.1), then it follows from Lemma 6.1.1 that $W_R^* V_R$ is nonsingular. During the process, it is possible that V_R does not exist or that $W_R^* V_R$ is singular. In either of these two cases we can use one of the correction equations from Section 6.5.1 to expand the search and test spaces. ⊘

In a similar manner we obtain a correction equation for t_1 and t_2. If V_{L}, W_{L}, and P_{L} are defined similarly for (6.5.17), then we have

$$P_{\mathrm{L}} D^* P_{\mathrm{L}} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = -r^L. \tag{6.5.19}$$

We separately solve (6.5.18) and (6.5.19) approximately using a few steps of GMRES.

Better results can be obtained if we use preconditioners. Suppose that M is a left preconditioner for D. One can show that if $W_{\mathrm{R}}^* M^{-1} V_{\mathrm{R}}$ is nonsingular then the inverse of a map $P_{\mathrm{R}} M P_{\mathrm{R}}$ from $\mathrm{span}(W_{\mathrm{R}})^\perp$ to $\mathrm{span}(W_{\mathrm{R}})^\perp$ is

$$\left(I - M^{-1} V_{\mathrm{R}} (W_{\mathrm{R}}^* M^{-1} V_{\mathrm{R}})^{-1} W_{\mathrm{R}}^*\right) M^{-1} P_{\mathrm{R}}.$$

Thus we obtain a preconditioned correction equation

$$\left(I - M^{-1} V_{\mathrm{R}} (W_{\mathrm{R}}^* M^{-1} V_{\mathrm{R}})^{-1} W_{\mathrm{R}}^*\right) M^{-1} P_{\mathrm{R}} D P_{\mathrm{R}} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \left(I - M^{-1} V_{\mathrm{R}} (W_{\mathrm{R}}^* M^{-1} V_{\mathrm{R}})^{-1} W_{\mathrm{R}}^*\right) M^{-1} r^R. \tag{6.5.20}$$

In a similar manner we get a preconditioned equation for t_1 and t_2.

6.6 Computing more eigenpairs

Suppose that we are interested in $p > 1$ eigenpairs of (6.1.1). In one-parameter eigenvalue problems various deflation techniques can be applied in order to compute more than one

eigenpair. The difficulties that are met when we try to translate standard deflation ideas from one-parameter problems to two-parameter problems are discussed in Section 5.5.

For a general two-parameter eigenvalue problem we can apply a similar technique as in Section 5.5 for the right definite problem using the Δ_0-orthogonality of left and right eigenvectors. Suppose that we have already found p eigenvalues (λ_i, μ_i) with the corresponding left and right eigenvectors $x_{1i} \otimes x_{2i}$ and $y_{1i} \otimes y_{2i}$ for $i = 1, \ldots, p$. Now we adjust Algorithm 6.4.2 so that in Step 4 we consider only those Petrov triples for which $u_1 \otimes u_2$ and $v_1 \otimes v_2$ satisfy

$$\min\left(|(v_1 \otimes v_2)^* \Delta_0 (x_{1i} \otimes x_{2i})|,\ |(y_{1i} \otimes y_{2i})^* \Delta_0 (u_1 \otimes u_2)|\right) < \eta \text{ for } i = 1, \ldots, p \quad (6.6.1)$$

for an $\eta > 0$. A suggestion for η (used in Example 6.8.6 in Section 6.8) is

$$\eta = \frac{1}{2} \min_{i=1,\ldots,p} \left((y_{1i} \otimes y_{2i})^* \Delta_0 (x_{1i} \otimes x_{2i})\right).$$

If no triple satisfies this condition then we take the one with the smallest left side of (6.6.1). Let us mention that an efficient way to compute (6.6.1) is to apply the relation (cf. (6.3.3))

$$(x_1 \otimes x_2)^* \Delta_0 (y_1 \otimes y_2) = (x_1^* B_1 y_1)(x_2^* C_2 y_2) - (x_1^* C_1 y_1)(x_2^* B_2 y_2).$$

6.7 Time complexity

The analysis of time complexity of Algorithm 6.4.2 is similar to the analysis for the Jacobi–Davidson algorithm for right definite two-parameter eigenvalue in Section 5.6. Because of that the details are omitted and the main results are stated.

If we assume that $n = n_1 = n_2$ and that m steps of GMRES are used for the approximate solutions of the correction equations, then the time complexity of one outer step of Algorithm 6.4.2 for dense matrices is $\mathcal{O}(mn^2)$. Also important is the storage requirement. If an algorithm works with matrices A_i, B_i, and C_i as Algorithm 6.4.2 does then it requires $\mathcal{O}(n^2)$ memory. On the other hand, Algorithm 6.2.1 that works with the associated system (6.1.2) needs $\mathcal{O}(n^4)$ memory, which may quickly exceed the available memory, even for modest values of n. Therefore, restarts are no luxury.

If the matrices A_i, B_i, and C_i are sparse, then the time complexity of the outer step of Algorithm 6.4.2 is of order $\mathcal{O}(m\mathrm{MV})$, where MV stands for a matrix-vector multiplication by an $n \times n$ matrix.

6.8 Numerical examples

The following numerical results were obtained with Matlab 5.3. In order to be able to compare the results of the direct method of Algorithm 6.2.1 to the results of the subspace method of Algorithm 6.4.2, we use a small two-parameter eigenvalue problem with random matrices of size $n = 15$.

In all numerical examples we use the same two-parameter eigenvalue problem which we construct in Matlab by the following commands:

```
rand('seed',0)
A1=rand(15)-0.5; B1=rand(15)-0.5; C1=rand(15)-0.5;
A2=rand(15)-0.5; B2=rand(15)-0.5; C2=rand(15)-0.5;
```

The five eigenvalues of the obtained two-parameter problem that are closest to the origin are

$$\begin{aligned}(\lambda_1,\mu_1) &= (-0.12446, 0.24740),\\ (\lambda_2,\mu_2) &= (-0.09509 + 0.25002i, 0.11122 - 0.13857i),\\ (\lambda_3,\mu_3) &= (-0.09509 - 0.25002i, 0.11122 + 0.13857i),\\ (\lambda_4,\mu_4) &= (-0.19895, 0.27873),\\ (\lambda_5,\mu_5) &= (-0.00020 + 0.36828i, 0.00029 + 0.12196i).\end{aligned}$$

Example 6.8.1 The results in this first example suggest that for non right definite problems, the two-sided approach (different test and search spaces) is superior to the one-sided approach (the same test and search spaces). We perturb the eigenvectors $x_{11}, x_{12}, y_{11}, y_{12}$ into u_1, u_2, v_1, v_2, respectively, by adding random vectors of small norm and then compute the difference between the two-sided Rayleigh quotient (6.3.3) of u_1, u_2, v_1, v_2 and (λ_1, μ_1). If we take $v_1 = u_1$ and $v_2 = u_2$ and apply formula (6.3.3) then we obtain the one-sided Rayleigh quotient. It is equal to the Ritz value in the one-sided Jacobi–Davidson type method where the search subspaces are equal to test subspaces (Chapter 5).

Table 6.1 shows the errors of one-sided and two-sided Rayleigh quotients (σ, τ) as approximations to the eigenvalue (λ_1, μ_1). The results indicate that the order of the error of the two-sided Rayleigh quotient is equal to the square of the error of eigenvector approximations u_i, v_i, which agrees with Theorem 6.5.1. On the other hand, the error of the one-sided Rayleigh quotient depends on the error of eigenvector approximation in a linear way.

TABLE 6.1: Comparison of errors $((\lambda_1-\sigma)^2+(\mu_1-\tau)^2)^{1/2}$ for the one-sided and two-sided tensor Rayleigh quotients, related to the norm of the eigenvector perturbations $\|x_i - u_i\|$ and $\|y_i - v_i\|$.

perturbation	one-sided RQ error	two-sided RQ error
10^{-3}	$2.4 \cdot 10^{-3}$	$3.1 \cdot 10^{-6}$
10^{-4}	$3.0 \cdot 10^{-4}$	$1.2 \cdot 10^{-8}$
10^{-5}	$2.2 \cdot 10^{-5}$	$2.4 \cdot 10^{-10}$
10^{-6}	$3.3 \cdot 10^{-6}$	$2.2 \cdot 10^{-12}$

⊘

Example 6.8.2 In the second example we compare different correction equations without preconditioning. For the initial vectors we take $u_i = x_{1i} + 10^{-3}e$, $v_i = y_{1i} + 10^{-3}e$ for $i = 1, 2$, where $e = [\,1\ \cdots\ 1\,]^T$. In each Step 4 of Algorithm 6.4.2 we take the Petrov triple with the smallest residual (6.4.1).

Table 6.2 contains the number of steps required for the residual (6.4.1) to become smaller than 10^{-8}. The other parameters are $l_{\max} = 8$, $l_{\min} = 2$ and $k_{\max} = 500$. We compared three correction equations without preconditioning:

TABLE 6.2: Comparison of three correction equations NP1, NP2, and NP3 without preconditioning for the initial vectors $u_i = x_{1i} + 10^{-3}e$ and $v_i = y_{1i} + 10^{-3}e$, where $e = [1 \ \cdots \ 1]^T$. GMRES: the number of steps used in GMRES for the approximate solution of the correction equation; iterations: the number of outer iterations for convergence.

NP1		NP2		NP3	
GMRES	iterations	GMRES	iterations	GMRES	iterations
10	> 500	10	> 500	17	> 500
11	70	11	50	20	155
12	24	12	27	23	36
13	14	13	6	26	5
14	3	14	3	29	3

- NP1 - first order correction equation (6.5.11).
- NP2 - first order correction equation (6.5.15). Although it is preconditioned, we treat this equation as an unpreconditioned one because the preconditioner is the (projected) identity.
- NP3 - second order correction equation (6.5.18) and (6.5.19).

The results in the table indicate that the convergence is slow or we have no convergence at all if the correction equations are not solved accurately. Let us remark that the number of GMRES steps for the second order correction equation is larger because the size of matrices is twice the size of the matrices in the first order correction equations. ⊘

Example 6.8.3 For the third example we take the same initial vectors and parameters as in Example 6.8.2, but, this time we use preconditioned correction equations. For a preconditioner we take (6.5.14). We compared the following three preconditioned correction equations:

- P1 - preconditioned NP1 from Example 6.8.2.
- P2 - first order correction equation (6.5.13), left preconditioned by (6.5.14).
- P3 - (6.5.20) preconditioned NP3 from Example 6.8.2.

⊘

The results in Table 6.3 indicate that correction equations with preconditioners work much better than the ones that are not preconditioned.

Example 6.8.4 In this example we take initial vectors $u_1 = u_2 = v_1 = v_2 = [\, 1 \ \cdots \ 1 \,]^T$. Our goal is the eigenvalue closest to the origin. In Step 4 of Algorithm 6.4.2 we pick the

TABLE 6.3: Comparison of three correction equations P1, P2, and P3 with preconditioning for initial vectors $u_i = x_{1i} + 10^{-3}e$ and $v_i = y_{1i} + 10^{-3}e$, where $e = [1 \ \cdots \ 1]^T$. GMRES: the number of steps used in GMRES for the approximate solution of the correction equation; iterations: the number of outer iterations for convergence.

P1		P2		P3	
GMRES	iterations	GMRES	iterations	GMRES	iterations
1	32	1	22	1	32
3	43	3	12	3	25
5	11	5	6	5	12
7	5	7	4	7	7
9	4	9	4	9	6

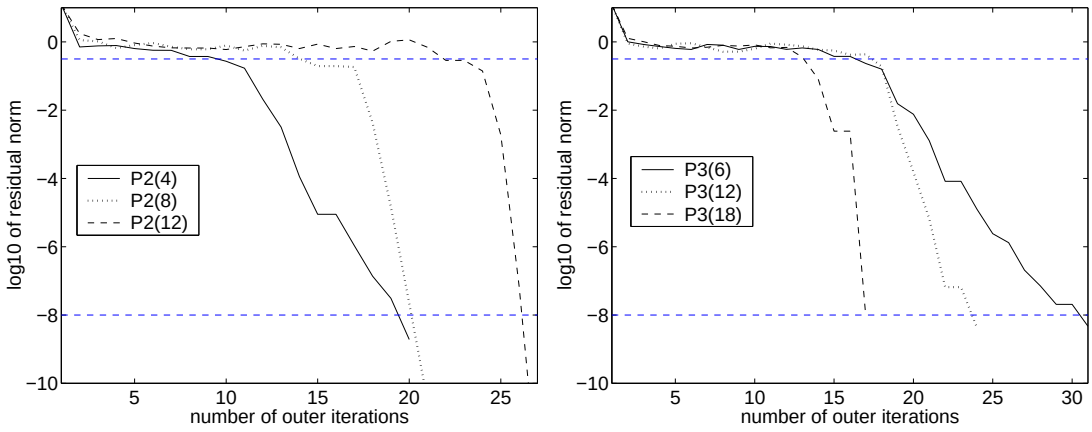


FIGURE 6.1: Convergence plot for the eigenvalue closest to $(0,0)$ for $u_i = v_i = [1 \ \cdots \ 1]^T$. The plots show the residual norm (6.4.1) versus the outer iteration number for the Jacobi–Davidson type method using correction equation P2 (a) with 3 (solid line), 8 (dotted line), and 12 (dashed line) GMRES steps, and correction equation P3 (b) with 6 (solid line), 12 (dotted line), and 18 (dashed line) GMRES steps to solve the correction equation.

Petrov triple with the Petrov value closest to the target $(0, 0)$ until the residual ρ_k is less than $\varepsilon_{\text{change}} = 0.5$. After that we take Petrov triple with the smallest residual (6.4.1).

Figure 6.1 shows convergence plot for correction equations P2 and P3 using various number of GMRES steps to solve the correction equation. One can see that once the residual becomes smaller than $\varepsilon_{\text{change}}$ (top horizontal dotted line in the figure) and we are close to the eigentriple, then the number of GMRES steps determines how fast the convergence is.

There is no guarantee that the process will converge to the eigenvalue closest to the target. In fact, the eigenvalue obtained using P3 with 12 GMRES steps is $(-0.33, 0.24)$, which is equal to (λ_7, μ_7). In all other 5 cases we get (λ_1, μ_1).

The statistics in the following example show that the probability of a successful convergence is high if we carefully tune the parameters of the method. ⊘

Example 6.8.5 In this example we are interested in the number of iterations that the Jacobi–Davidson type method needs for convergence and in the percentage of convergence to the eigenvalue (λ_1, μ_1) if random initial vectors are used.

TABLE 6.4: Statistics of the Jacobi–Davidson type method for the eigenvalue (λ_1, μ_1) using correction equations P2, P3 and various settings of GMRES steps and $\varepsilon_{\text{change}}$. GMRES: the number of steps used in GMRES for the approximate solution of the correction equation; $\varepsilon_{\text{change}}$: setting of $\varepsilon_{\text{change}}$ parameter; %: percentage of convergence to (λ_1, μ_1); iter: the average number of outer iterations for convergence.

Parameters		$\varepsilon_{\text{change}} = 0.5$		$\varepsilon_{\text{change}} = 0.1$		$\varepsilon_{\text{change}} = 10^{-2}$		$\varepsilon_{\text{change}} = 10^{-3}$	
Equation	GMRES	%	iter	%	iter	%	iter	%	iter
P2	3	77	34.4	87	65.5	73	90.1	38	108.3
P2	5	73	25.0	95	42.5	88	49.5	79	57.6
P2	7	67	24.5	94	39.8	88	52.8	83	59.6
P3	6	70	37.0	97	56.3	83	83.7	65	110.1
P3	10	73	24.9	95	36.6	89	41.2	82	49.9
P3	14	64	21.1	100	36.3	97	44.8	94	47.8

We test the preconditioned correction equations P2 and P3 on the same set of 100 random initial vectors. We use the combined method for selecting the Petrov triple: in the first part we select the closest Petrov value to the origin until the residual becomes smaller than $\varepsilon_{\text{change}}$ and in the remaining steps we select Petrov triple with the minimum residual. We set the maximum number of outer steps to 250.

The numbers in Table 6.4 show that the probability of computing the correct eigenvalue is high when the parameters are carefully chosen. A small value of $\varepsilon_{\text{change}}$ does not necessarily improve the probability. If $\varepsilon_{\text{change}}$ is too small then in the first phase, when we select the closest Petrov value to the origin, the method requires too many iterations until the residual is smaller than $\varepsilon_{\text{change}}$. On the other hand, if $\varepsilon_{\text{change}}$ is too large then the method is likely to converge fast, but to an unwanted eigenvalue. More GMRES steps reduce the number of outer iterations and enlarge the probability, but we must keep in mind that the total amount of work depends on the number of matrix-vector

multiplications, and thus roughly equal to the product of the number of GMRES steps and outer iterations. ⊘

Example 6.8.6 In the last example we test the selection technique from Section 6.6 that enables us to compute more than one eigenvalue. Figure 6.2 shows a convergence plot for five eigenvalues computed in a row. The approach works and we obtain five different eigentriples. Unfortunately, the obtained eigenvalues are not the five eigenvalues that are closest to the origin. If we order the eigenvalues on their distance from the origin then the obtained eigenvalues have indices $1, 10, 18, 11$, and 19 among 225 eigenvalues. Additional numerical experiments with different initial vectors and correction equations showed that this behavior is not an exception and we were not able to reliably compute a small number of eigenvalues closest to the target with this method. It remains future work to modify the method to enable this feature.

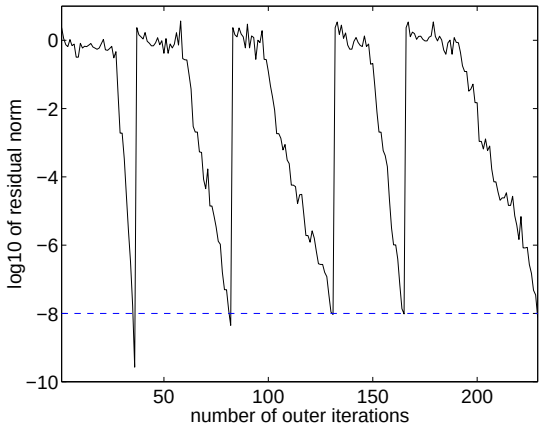


FIGURE 6.2: Convergence plot for the first five computed eigenvalues using the selection technique from Section 6.6. Used is correction equation P3 with 8 steps of GMRES and $\varepsilon_{\text{change}} = 5 \cdot 10^{-1}$.

⊘

6.9 Conclusions

We have presented a novel Jacobi–Davidson type method for the nonsingular two-parameter eigenvalue problem. This problem is a very challenging one, where we have to use many techniques to be successful: a two-sided subspace approach, preconditioning, selection techniques instead of deflating, and the use of a target. The new method can compute selected eigenpairs without good initial approximations and it can tackle very large two-parameter problems, especially if the matrices A_i, B_i, and C_i are sparse. In such situations, preconditioning is of great importance.

Let us also mention that Algorithm 6.2.1 and Algorithm 6.4.2 both offer a generalization to multiparameter problems with more than two parameters, in a way similar to Section 5.7.

Chapter 7

Backward error, condition numbers, and pseudospectrum for the multiparameter eigenvalue problem

Abstract. We define and evaluate the normwise backward error and condition numbers for the multiparameter eigenvalue problem (MEP). The pseudospectrum for the MEP is defined and characterized. We show that the distance from a right definite MEP to the closest non right definite MEP is related to the smallest unbounded pseudospectrum. Some numerical results are given.

Key words: multiparameter eigenvalue problem, right definiteness, backward error, condition number, pseudospectrum, nearness problem.

AMS subject classification: 65F15, 15A18, 15A69.

7.1 Introduction

We study the backward error, condition numbers and pseudospectrum for the multiparameter eigenvalue problem (MEP)

$$W_i(\boldsymbol{\lambda})x_i = 0, \quad 0 \neq x_i \in \mathbb{C}^{n_i}, \quad i = 1, \ldots, k, \tag{7.1.1}$$

where

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k) \in \mathbb{C}^k,$$

$$W_i(\boldsymbol{\lambda}) = V_{i0} - \sum_{j=1}^{k} \lambda_j V_{ij},$$

and V_{ij} are $n_i \times n_i$ matrices over $\mathbb{C}$. We will denote the MEP (7.1.1) by $\boldsymbol{W}$. For $k = 1$, a MEP is a generalized eigenvalue problem $V_{10}x_1 = \lambda_1 V_{11}x_1$. For $k = 2$, see also Chapters 5 and 6.

*Based on joint work with Bor Plestenjak, see Section 1.5.

A k-tuple $\boldsymbol{\lambda}$ that satisfies (7.1.1) is called an *eigenvalue* and the tensor product $\boldsymbol{x} = x_1 \otimes \cdots \otimes x_k$ is the corresponding *right eigenvector*. A *left eigenvector* corresponding to the eigenvalue $\boldsymbol{\lambda}$ is $\boldsymbol{y} = y_1 \otimes \cdots \otimes y_k$, where $0 \neq y_i \in \mathbb{C}^{n_i}$ and $y_i^* W_i(\boldsymbol{\lambda}) = 0$ for $i = 1, \ldots, k$.

The backward error and condition numbers are important tools in numerical linear algebra that reveal the quality and sensitivity of numerical solutions. The theory of backward error and conditioning for eigenproblems is well developed for the generalized eigenvalue problem (see, e.g., [32]) and the polynomial eigenvalue problem (see, e.g., [84]). See Chapter 5 for the origin of multiparameter eigenvalue problems.

To a MEP (7.1.1) which satisfies a certain regularity condition (nonsingularity, see below), a k-tuple of commuting linear transformations on a tensor product space is associated, as follows. The tensor product space $\mathbb{C}^{n_1} \otimes \cdots \otimes \mathbb{C}^{n_k}$ is isomorphic to $\mathbb{C}^N$, where $N = n_1 \cdots n_k$. Linear transformations $V_{ij}^\dagger$ on $\mathbb{C}^N$ are induced by the V_{ij}, $i = 1, 2, \ldots, k; j = 0, 1, \ldots, k$, and defined by

$$V_{ij}^\dagger(x_1 \otimes \cdots \otimes x_i \otimes \cdots \otimes x_k) = x_1 \otimes \cdots \otimes V_{ij}x_i \otimes \cdots \otimes x_k$$

and linearity. On $\mathbb{C}^N$ we define operator determinants (cf. (5.1.3) for $k = 2$)

$$\Delta_0 = \begin{vmatrix} V_{11}^\dagger & V_{12}^\dagger & \cdots & V_{1k}^\dagger \\ V_{21}^\dagger & V_{22}^\dagger & \cdots & V_{2k}^\dagger \\ \vdots & \vdots & & \vdots \\ V_{k1}^\dagger & V_{k2}^\dagger & \cdots & V_{kk}^\dagger \end{vmatrix}$$

and

$$\Delta_i = \begin{vmatrix} V_{11}^\dagger & \cdots & V_{1,i-1}^\dagger & V_{10}^\dagger & V_{1,i+1}^\dagger & \cdots & V_{1k}^\dagger \\ V_{21}^\dagger & \cdots & V_{2,i-1}^\dagger & V_{20}^\dagger & V_{2,i+1}^\dagger & \cdots & V_{2k}^\dagger \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ V_{k1}^\dagger & \cdots & V_{k,i-1}^\dagger & V_{k0}^\dagger & V_{k,i+1}^\dagger & \cdots & V_{kk}^\dagger \end{vmatrix}$$

for $i = 1, \ldots, k$.

A MEP is called *nonsingular* if the corresponding operator determinant Δ_0 is invertible. A nonsingular MEP is equivalent to the associated problem (cf. (5.1.4) for $k = 2$)

$$\Delta_i \boldsymbol{x} = \lambda_i \Delta_0 \boldsymbol{x}, \quad i = 1, \ldots, k,$$

for decomposable tensors $\boldsymbol{x} = x_1 \otimes \cdots \otimes x_k \in \mathbb{C}^N$, where the matrices $\Gamma_i := \Delta_0^{-1}\Delta_i$ commute for $i = 1, \ldots, k$ (see [4]).

If $\boldsymbol{\lambda}$ is an eigenvalue of $\boldsymbol{W}$ then (cf. (6.1.3) for $k = 2$)

$$d_a := \dim \left(\bigcap_{\substack{j_1 + \cdots + j_k = N \\ j_1, \ldots, j_k \geq 0}} \ker \left[(\Gamma_1 - \lambda_1 I)^{j_1} \cdots (\Gamma_k - \lambda_k I)^{j_k} \right] \right)$$

is the *algebraic multiplicity* (cf. 6.1.3) and

$$d_g := \dim \left(\bigcap_{i=1}^{k} \ker (\Gamma_i - \lambda_i I) \right) = \prod_{i=1}^{k} \dim \Big(\ker W_i(\boldsymbol{\lambda}) \Big)$$

is the *geometric multiplicity* of the eigenvalue (see [4]). We say that an eigenvalue $\boldsymbol{\lambda}$ is *geometrically* or *algebraically simple* when $d_g = 1$ or $d_a = 1$, respectively. It can be seen that $d_a \geq d_g$, so an eigenvalue that is algebraically simple is also geometrically simple.

Let $\boldsymbol{\lambda}$ be an eigenvalue of $\boldsymbol{W}$ with the corresponding left and right eigenvectors $\boldsymbol{x}$ and $\boldsymbol{y}$. We form a $k \times k$ matrix

$$B_0 = \begin{bmatrix} y_1^* V_{11} x_1 & y_1^* V_{12} x_1 & \cdots & y_1^* V_{1k} x_1 \\ y_2^* V_{21} x_2 & y_2^* V_{22} x_2 & \cdots & y_2^* V_{2k} x_2 \\ \vdots & \vdots & & \vdots \\ y_k^* V_{k1} x_k & y_k^* V_{k2} x_k & \cdots & y_k^* V_{kk} x_k \end{bmatrix}.$$

The following lemma is a consequence of [47, Lemma 3].

Lemma 7.1.1 (cf. Lemma 6.1.1) *If $\boldsymbol{\lambda}$ is an algebraically simple eigenvalue of the multiparameter eigenvalue problem $\boldsymbol{W}$ then B_0 is nonsingular.*

A MEP is called *Hermitian* when all matrices V_{ij} are Hermitian. Furthermore, a Hermitian MEP is called *right definite* if (cf. (5.1.2) for $k = 2$)

$$\begin{vmatrix} x_1^* V_{11} x_1 & x_1^* V_{12} x_1 & \cdots & x_1^* V_{1k} x_1 \\ x_2^* V_{21} x_2 & x_2^* V_{22} x_2 & \cdots & x_2^* V_{2k} x_2 \\ \vdots & \vdots & & \vdots \\ x_k^* V_{k1} x_k & x_k^* V_{k2} x_k & \cdots & x_k^* V_{kk} x_k \end{vmatrix} \geq \delta \tag{7.1.2}$$

for all vectors $x_i \in \mathbb{C}^{n_i}$, $\|x_i\| = 1$, $i = 1, \ldots, k$, and some $\delta > 0$. By noting that for decomposable tensors

$$\boldsymbol{x}^* \Delta_0 \boldsymbol{x} = \begin{vmatrix} x_1^* V_{11} x_1 & \cdots & x_1^* V_{1k} x_1 \\ \vdots & & \vdots \\ x_k^* V_{k1} x_k & \cdots & x_k^* V_{kk} x_k \end{vmatrix}, \tag{7.1.3}$$

we realize that right definiteness is equivalent to the positive definiteness of Δ_0 [4, Theorem 7.8.2] (we have $\boldsymbol{x}^* \Delta_0 \boldsymbol{x} > 0$ for decomposable tensors if and only if $\boldsymbol{x}^* \Delta_0 \boldsymbol{x} > 0$ for all tensors). This implies that if $\boldsymbol{W}$ is right definite then there exist N linearly independent eigenvectors. If $\boldsymbol{\lambda}$ is an eigenvalue of a right definite problem $\boldsymbol{W}$ then $\boldsymbol{\lambda} \in \mathbb{R}^k$. Furthermore, if all matrices V_{ij} of a right definite problem $\boldsymbol{W}$ are real, then the eigenvectors can be chosen real. For a real, geometrically simple eigenvalue of a Hermitian MEP, the corresponding left and right eigenvectors coincide.

After preliminaries in Section 7.2, we study the backward error in Section 7.3. The condition numbers for eigenvalues and eigenvectors are discussed in Section 7.4. The pseudospectrum, examined in Section 7.5, is another valuable tool for the study of the sensitivity of eigenvalues to perturbations of the matrices. In Section 7.6, we give some numerical experiments for right definite two-parameter eigenvalue problems, where pseudospectra can be visualized in $\mathbb{R}^2$. Conclusions are summarized in Section 7.7.

7.2 Preliminaries

Throughout this chapter we assume that the MEP $\boldsymbol{W}$ is nonsingular. The matrices E_{ij} for $i = 1, \ldots, k$; $j = 0, \ldots, k$ represent tolerances for the perturbations ΔV_{ij} of V_{ij}, defined by $\|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|$ for some $\varepsilon > 0$. Usually we take either $E_{ij} = V_{ij}$ considering *normwise relative perturbations*, or $E_{ij} = I$ considering *normwise absolute perturbations*. *Elementwise perturbations* $|\Delta V_{ij}| \leq \varepsilon |E_{ij}|$ can also be considered, see Remark 7.3.4. We define

$$\Delta W_i(\boldsymbol{\lambda}) := \Delta V_{i0} - \sum_{j=1}^{k} \lambda_j \Delta V_{ij}.$$

We will denote the perturbed MEP with matrices $V_{ij} + \Delta V_{ij}$ by $\boldsymbol{W} + \Delta \boldsymbol{W}$. For a complex λ the *sign of* λ is defined as (cf. [32, p. 495])

$$\operatorname{sign}(\lambda) := \begin{cases} \overline{\lambda}/|\lambda|, & \lambda \neq 0, \\ 0, & \lambda = 0. \end{cases}$$

Suppose that we are looking for the maximum Euclidean norm of Az where $A \in \mathbb{C}^{k \times k}$ and $z \in \mathbb{C}^k$ is such that $|z_i| \leq \beta_i$ for $i = 1, \ldots, k$, where $\beta_1, \ldots, \beta_k$ are given positive constants. According to Bauer's maximum principle (both the function $\|\cdot\|$ and its domain are convex), the maximum is attained by z for which $|z_i| = \beta_i$ for $i = 1, \ldots, k$. For $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_k]^T$ we define the *$\boldsymbol{\beta}$-weighted norm of* A as

$$\|A\|_{\boldsymbol{\beta}} := \max\{ \|Az\|_2 \ : \ z \in \mathbb{C}^k, \ |z_i| = \beta_i \text{ for } i = 1, \ldots, k \}. \tag{7.2.1}$$

Clearly,

$$\|A\|_{\boldsymbol{\beta}} \leq \|A\|_2 \cdot \|\boldsymbol{\beta}\|_2. \tag{7.2.2}$$

One may verify that $\|\cdot\|_{\boldsymbol{\beta}}$ is indeed a matrix norm. One may also see that $\|\cdot\|_{\boldsymbol{\beta}}$ is not a consistent norm as it does not necessarily satisfy $\|AB\|_{\boldsymbol{\beta}} \leq \|A\|_{\boldsymbol{\beta}} \|B\|_{\boldsymbol{\beta}}$ (for a counterexample, take $A = B = I$ and $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\|_2 < 1$).

From now on, $\|\cdot\|$ stands for $\|\cdot\|_2$. We say that a decomposable tensor $\boldsymbol{z} = z_1 \otimes \cdots \otimes z_k$ is *normalized* if $\|z_i\| = 1$ for $i = 1, \ldots, k$. From $\|\boldsymbol{z}\| = \|z_1\| \cdots \|z_k\|$ it follows that $\|\boldsymbol{z}\| = 1$. In this chapter we will assume that the eigenvectors are normalized.

7.3 Backward error

Let $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$ be an approximate eigenpair of $\boldsymbol{W}$ and let $\widetilde{\boldsymbol{x}}$ be normalized. We define the *normwise backward error of* $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$ by

$$\begin{aligned} \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) := \min \{ \quad & \varepsilon : (W_i(\widetilde{\boldsymbol{\lambda}}) + \Delta W_i(\widetilde{\boldsymbol{\lambda}}))\widetilde{x}_i = 0, \\ & \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|, \ i = 1, \ldots, k; \ j = 0, \ldots, k \ \}. \end{aligned} \tag{7.3.1}$$

The following theorem is a generalization of the backward errors for the case $k = 1$ (i.e., the generalized eigenproblem) given in [25, Lemma 2.1] and [32, Theorem 2.1].

Theorem 7.3.1 *For the normwise backward error $\eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$ we have*

$$\eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) = \max_{i=1,\ldots,k} \frac{\|r_i\|}{\widetilde{\beta}_i}, \tag{7.3.2}$$

where $r_i := W_i(\widetilde{\boldsymbol{\lambda}})\widetilde{x}_i$ are the residuals *and*

$$\widetilde{\beta}_i := \|E_{i0}\| + \sum_{j=1}^{k} |\widetilde{\lambda}_j| \|E_{ij}\|$$

for $i = 1, \ldots, k$.

Proof: From $r_i = -\Delta W_i(\widetilde{\boldsymbol{\lambda}})\widetilde{x}_i$ it follows that $\|r_i\| \leq \widetilde{\beta}_i \varepsilon$ for $i = 1, \ldots, k$. Therefore, the right-hand side of (7.3.2) is a lower bound for $\eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$. The lower bound is attained for the perturbations

$$\Delta V_{i0} = -\frac{1}{\widetilde{\beta}_i} \|E_{i0}\| r_i \widetilde{x}_i^*, \quad \Delta V_{ij} = \frac{\operatorname{sign}(\widetilde{\lambda}_j)}{\widetilde{\beta}_i} \|E_{ij}\| r_i \widetilde{x}_i^*$$

for $i, j = 1, \ldots, k$. □

If $\boldsymbol{W}$ is Hermitian then it is of interest to consider a backward error in which the perturbations ΔV_{ij} are Hermitian. The *backward error for a Hermitian MEP* can be defined as

$$\begin{aligned} \eta_{\mathrm{H}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) := \min \{ \quad & \varepsilon : (W_i(\widetilde{\boldsymbol{\lambda}}) + \Delta W_i(\widetilde{\boldsymbol{\lambda}}))\widetilde{x}_i = 0,\ \Delta V_{ij}^* = \Delta V_{ij}, \\ & \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|,\ i = 1, \ldots, k;\ j = 0, \ldots, k\}. \end{aligned} \tag{7.3.3}$$

It is clear that $\eta_{\mathrm{H}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) \geq \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$ and that the optimal perturbations in (7.3.1) are not Hermitian in general. The next lemma, which is is a generalization of [32, Lemma 2.6], shows that in the case when $\widetilde{\boldsymbol{\lambda}}$ is real requiring the perturbations to be Hermitian has no effect on the backward error.

Theorem 7.3.2 *If $\boldsymbol{W}$ is Hermitian and $\widetilde{\boldsymbol{\lambda}}$ is real then*

$$\eta_{\mathrm{H}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) = \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}). \tag{7.3.4}$$

Proof: It follows from $\widetilde{\boldsymbol{\lambda}}$ being real that $\widetilde{x}_i^* r_i$ is real. We are looking for a Hermitian matrix S_i such that $S_i \widetilde{x}_i = -r_i$. We take $S_i = \|r_i\| I$ if r_i is a negative multiple of $\widetilde{x}_i$; otherwise we take $S_i = \|r_i\| H_i$ where H_i is a Householder matrix that maps $\widetilde{x}_i$ to $-r_i/\|r_i\|$. Such an H_i exists because $\widetilde{x}_i^* r_i$ is real and is equal to $I - 2(w_i^* w_i)^{-1} w_i w_i^*$, where $w_i = \widetilde{x}_i + r_i/\|r_i\|$.

Let ΔV_{ij} be Hermitian matrices defined by

$$\Delta V_{i0} = \frac{1}{\widetilde{\beta}_i} \|E_{i0}\| H_i, \quad \Delta V_{ij} = -\frac{1}{\widetilde{\beta}_i} \operatorname{sign}(\widetilde{\lambda}_j) \|E_{ij}\| H_i \tag{7.3.5}$$

for $i, j = 1, \ldots, k$. It follows that $\Delta W_i(\widetilde{\boldsymbol{\lambda}}) = S_i$ and the first constraint in (7.3.3) is satisfied. Using (7.3.2), we get

$$\|S_i\| = \|r_i\| \leq \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})\widetilde{\beta}_i$$

for $i = 1, \ldots, k$. From (7.3.5) we deduce $\eta_{\mathrm{H}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) \leq \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$. Since $\eta_{\mathrm{H}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}}) \geq \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})$ by definition, equality (7.3.4) must hold. □

We remark that one can see from $\widetilde{x}_i^* S_i \widetilde{x}_i = -\widetilde{x}_i r_i$ that a Hermitian matrix S_i such that $S_i \widetilde{x}_i = -\widetilde{x}_i r_i$ exists only when $\widetilde{x}_i^* r_i$ is real. This is the reason why Lemma 7.3.2 cannot be generalized for nonreal approximations $\widetilde{\boldsymbol{\lambda}}$. As it is reasonable to assume that $\widetilde{\boldsymbol{\lambda}}$ is real if $\boldsymbol{\lambda}$ is real, Lemma 7.3.2 can also be applied for a right definite MEP.

If we are interested only in the approximate eigenvalue $\widetilde{\boldsymbol{\lambda}}$, then a more appropriate measure of the backward error may be

$$\eta(\widetilde{\boldsymbol{\lambda}}) := \min\{\ \eta(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\lambda}})\ :\ \widetilde{\boldsymbol{x}} \text{ normalized }\}.$$

Proposition 7.3.3

$$\eta(\widetilde{\boldsymbol{\lambda}}) = \max_{i=1,\ldots,k} \frac{1}{\widetilde{\beta}_i}\, \sigma_{\min}(W_i(\widetilde{\boldsymbol{\lambda}})).$$

Proof: The result follows from Theorem 7.3.1 by using the equality

$$\min_{\|x\|=1} \|Ax\| = \sigma_{\min}(A).$$

□

Remark 7.3.4 Although in this chapter we do not consider *componentwise backward errors*, componentwise results from [32] can be generalized as well. ⊘

7.4 Condition numbers

In this section, we assume that $\boldsymbol{\lambda}$ is a nonzero algebraically simple eigenvalue of a nonsingular MEP $\boldsymbol{W}$ with corresponding normalized right eigenvector $\boldsymbol{x}$ and left eigenvector $\boldsymbol{y}$.

7.4.1 Eigenvalue condition number

A *normwise condition number of* $\boldsymbol{\lambda}$ can be defined by

$$\begin{aligned}\kappa(\boldsymbol{\lambda}, \boldsymbol{W}) \quad := \limsup_{\varepsilon \downarrow 0} \Bigg\{ & \frac{\|\Delta\boldsymbol{\lambda}\|}{\varepsilon} : \\ & \left(V_{i0} + \Delta V_{i0} - \sum_{j=1}^{k} (\lambda_j + \Delta\lambda_j)(V_{ij} + \Delta V_{ij})\right)(x_i + \Delta x_i) = 0, \qquad (7.4.1) \\ & \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|,\ i = 1, \ldots, k;\ j = 0, \ldots, k \Bigg\}.\end{aligned}$$

The following results can be considered as generalizations of the theory in [32, Section 2.2].

Theorem 7.4.1 *The condition number $\kappa(\boldsymbol{\lambda}, \boldsymbol{W})$ is given by*

$$\kappa(\boldsymbol{\lambda}, \boldsymbol{W}) = \|B_0^{-1}\|_{\boldsymbol{\beta}}, \tag{7.4.2}$$

where

$$\beta_i := \|E_{i0}\| + \sum_{j=1}^{k} |\lambda_j| \|E_{ij}\|$$

for $i = 1, \ldots, k$, and $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_k]^T$.

Proof: If we expand the equality constraints in (7.4.1) and keep only the first order terms then we get

$$\Delta W_i(\boldsymbol{\lambda}) x_i + \sum_{j=1}^{k} \Delta\lambda_j V_{ij} x_i + W_i(\boldsymbol{\lambda}) \Delta x_i = \mathcal{O}(\varepsilon^2), \qquad i = 1, \ldots, k. \tag{7.4.3}$$

Premultiplying by y_i^* yields

$$y_i^* \Delta W_i(\boldsymbol{\lambda}) x_i + y_i^* \sum_{j=1}^{k} \Delta\lambda_j V_{ij} x_i = \mathcal{O}(\varepsilon^2)$$

for $i = 1, \ldots, k$. By rearranging the equations we obtain the linear system

$$\begin{bmatrix} y_1^* V_{11} x_1 & \cdots & y_1^* V_{1k} x_1 \\ \vdots & & \vdots \\ y_k^* V_{k1} x_k & \cdots & y_k^* V_{kk} x_k \end{bmatrix} \begin{bmatrix} \Delta\lambda_1 \\ \vdots \\ \Delta\lambda_k \end{bmatrix} = \begin{bmatrix} y_1^* \Delta W_1(\boldsymbol{\lambda}) x_1 \\ \vdots \\ y_k^* \Delta W_k(\boldsymbol{\lambda}) x_k \end{bmatrix} + \mathcal{O}(\varepsilon^2),$$

or in shorter form

$$B_0 \Delta\boldsymbol{\lambda} = \begin{bmatrix} y_1^* \Delta W_1(\boldsymbol{\lambda}) x_1 \\ \vdots \\ y_k^* \Delta W_k(\boldsymbol{\lambda}) x_k \end{bmatrix} + \mathcal{O}(\varepsilon^2).$$

Since $\boldsymbol{\lambda}$ is an algebraically simple eigenvalue, it follows from Lemma 7.1.1 that B_0 is nonsingular. Thus,

$$\Delta\boldsymbol{\lambda} = B_0^{-1} \begin{bmatrix} y_1^* \Delta W_1(\boldsymbol{\lambda}) x_1 \\ \vdots \\ y_k^* \Delta W_k(\boldsymbol{\lambda}) x_k \end{bmatrix} + \mathcal{O}(\varepsilon^2)$$

and we conclude

$$\|\Delta\boldsymbol{\lambda}\| \leq \|B_0^{-1}\|_{\varepsilon\boldsymbol{\beta}} + \mathcal{O}(\varepsilon^2) = \varepsilon \|B_0^{-1}\|_{\boldsymbol{\beta}} + \mathcal{O}(\varepsilon^2).$$

Hence, the expression in (7.4.2) is an upper bound for the condition number. To show that this bound can be attained we take the matrices

$$\Delta V_{i0} = \varepsilon \|E_{i0}\| y_i x_i^*, \quad \Delta V_{ij} = -\text{sign}(\widetilde{\lambda}_j) \varepsilon \|E_{ij}\| y_i x_i^*$$

for $i, j = 1, \ldots, k$. □

As for the backward error, if the MEP $\boldsymbol{W}$ is Hermitian then it is natural to restrict the perturbations ΔV_{ij} in (7.4.1) to be Hermitian. We denote

$$\begin{aligned}\kappa_{\mathrm{H}}(\boldsymbol{\lambda}, \boldsymbol{W}) \quad := \limsup_{\varepsilon \downarrow 0} \Bigg\{ & \frac{\|\Delta\boldsymbol{\lambda}\|}{\varepsilon} : \\ & \left(V_{i0} + \Delta V_{i0} - \sum_{j=1}^{n} (\lambda_j + \Delta\lambda_j)(V_{ij} + \Delta V_{ij})\right)(x_i + \Delta x_i) = 0, \\ & \Delta V_{ij}^* = \Delta V_{ij},\ \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|,\ i = 1, \ldots, k;\ j = 0, \ldots, k \Bigg\}.\end{aligned}$$

Lemma 7.4.2 *If $\boldsymbol{\lambda}$ is a real algebraically simple eigenvalue of a Hermitian multiparameter eigenvalue problem $\boldsymbol{W}$ then*

$$\kappa_{\mathrm{H}}(\boldsymbol{\lambda}, \boldsymbol{W}) = \kappa(\boldsymbol{\lambda}, \boldsymbol{W}).$$

Proof: For a Hermitian MEP and algebraically simple eigenvalue $\boldsymbol{\lambda}$ we can take $\boldsymbol{y} = \boldsymbol{x}$ and then the matrices H_i in the proof of Theorem 7.4.1 are Hermitian. It follows that the perturbations for which the bound is attained are also Hermitian. □

As in Section 7.3 let us remark that Lemma 7.4.2 can also be applied to a right definite MEP.

7.4.2 Eigenvector condition number

In order to study the condition number of the eigenvector of an algebraically simple eigenvalue we introduce the following approach. If an eigenvector $\boldsymbol{x} = x_1 \otimes \cdots \otimes x_k$ is perturbed to $\widetilde{\boldsymbol{x}} = (x_1 + \Delta x_1) \otimes \cdots \otimes (x_k + \Delta x_k)$, then we are interested in $\|\operatorname{vec}(\boldsymbol{\Delta x})\|$, where

$$\operatorname{vec}(\boldsymbol{\Delta x}) = [\Delta x_1^T \ \cdots \ \Delta x_k^T]^T$$

is a vector in $\mathbb{C}^{n_1 + \cdots + n_k}$. Therefore we define a *normwise condition number of* $\boldsymbol{x}$ by

$$\begin{aligned}\kappa(\boldsymbol{x}, \boldsymbol{W}) \quad := \limsup_{\varepsilon \downarrow 0} \Bigg\{ & \frac{\|\operatorname{vec}(\boldsymbol{\Delta x})\|}{\varepsilon} : \\ & \left(V_{i0} + \Delta V_{i0} - \sum_{j=1}^{k} (\lambda_j + \Delta\lambda_j)(V_{ij} + \Delta V_{ij})\right)(x_i + \Delta x_i) = 0, \\ & g_i^* x_i = g_i^*(x_i + \Delta x_i) = 1, \\ & \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|,\ i = 1, \ldots, k;\ j = 0, \ldots, k \Bigg\},\end{aligned}$$

where the vectors g_i that are used for the normalization of $\widetilde{\boldsymbol{x}}$ are such that $g_i^* x_i \neq 0$ for $i = 1, \ldots, k$ and that the matrix

$$\begin{bmatrix} g_1^* V_{11} x_1 & \cdots & g_1^* V_{1k} x_1 \\ \vdots & & \vdots \\ g_k^* V_{k1} x_k & \cdots & g_k^* V_{kk} x_k \end{bmatrix} \tag{7.4.4}$$

is nonsingular. We can for instance take $g_i = y_i$, since in this case the matrix (7.4.4) is equal to B_0, which is nonsingular for algebraically simple eigenvalues by Lemma 7.1.1, see also Remark 7.4.5.

Let $m = n_1 + \cdots + n_k$. We can combine all the equations (7.4.3) into one equation in $\mathbb{C}^m$ as

$$D \operatorname{vec}(\boldsymbol{\Delta x}) = -\operatorname{diag}(\Delta W_i(\boldsymbol{\lambda})) \operatorname{vec}(\boldsymbol{x}) - V \Delta \boldsymbol{\lambda} + \mathcal{O}(\varepsilon^2), \tag{7.4.5}$$

where

$$D = \begin{bmatrix} W_1(\boldsymbol{\lambda}) & & \\ & \ddots & \\ & & W_k(\boldsymbol{\lambda}) \end{bmatrix}, \quad \operatorname{diag}(\Delta W_i(\boldsymbol{\lambda})) = \begin{bmatrix} \Delta W_1(\boldsymbol{\lambda}) & & \\ & \ddots & \\ & & \Delta W_k(\boldsymbol{\lambda}) \end{bmatrix},$$

$$V = \begin{bmatrix} V_{11} x_1 & \cdots & V_{1k} x_1 \\ \vdots & & \vdots \\ V_{k1} x_k & \cdots & V_{kk} x_k \end{bmatrix},$$

$$\Delta \boldsymbol{\lambda} = [\Delta \lambda_1 \ \cdots \ \Delta \lambda_k]^T, \quad \text{and} \quad \operatorname{vec}(\boldsymbol{x}) = [x_1^T \ \cdots \ x_k^T]^T.$$

If we define the $m \times k$ matrix

$$G = \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & g_k \end{bmatrix}$$

then $G^* V$ is equal to (7.4.4). As a result $G^* V$ is nonsingular and we can define an oblique projection

$$P = I - V(G^* V)^{-1} G^*$$

onto $\operatorname{span}(G)^\perp$ along $\operatorname{span}(V)$. It follows that $PV = 0$ and when we left multiply (7.4.5) by P we obtain

$$PD \operatorname{vec}(\boldsymbol{\Delta x}) = -P \operatorname{diag}(\Delta W_i(\boldsymbol{\lambda})) \operatorname{vec}(\boldsymbol{x}) + \mathcal{O}(\varepsilon^2). \tag{7.4.6}$$

From $g_i^* \Delta x_i = 0$ for $i = 1, \ldots, k$ it follows that $G^* \operatorname{vec}(\boldsymbol{\Delta x}) = 0$ and thus $P \operatorname{vec}(\boldsymbol{\Delta x}) = \operatorname{vec}(\boldsymbol{\Delta x})$. Now we can rewrite (7.4.6) as

$$PDP \operatorname{vec}(\boldsymbol{\Delta x}) = -P \operatorname{diag}(\Delta W_i(\boldsymbol{\lambda})) \operatorname{vec}(\boldsymbol{x}) + \mathcal{O}(\varepsilon^2). \tag{7.4.7}$$

Lemma 7.4.3 *The operator T defined by $T := PDP$ is a bijection as an operator from $\mathcal{G}^\perp$ onto $\mathcal{G}^\perp$, where $\mathcal{G}^\perp := \operatorname{span}(G)^\perp$*

Proof: Since T clearly maps to $\mathcal{G}^\perp$, it is enough to show that T is injective. Suppose that there exists a $\boldsymbol{z} \in \mathcal{G}^\perp$ such that $T\boldsymbol{z} = 0$. Since $P\boldsymbol{z} = \boldsymbol{z}$, there exists an $h \in \mathbb{C}^k$ such that

$$D\boldsymbol{z} = Vh. \tag{7.4.8}$$

If we left-multiply (7.4.8) by Y^*, where Y is the $m \times k$ matrix

$$Y = \begin{bmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & y_k \end{bmatrix},$$

we obtain $Y^*Vh = 0$ and since Y^*V is nonsingular it follows that $h = 0$. As a result we have $W_i(\boldsymbol{\lambda})z_i = 0$ for $i = 1, \ldots, k$ where $\boldsymbol{z}$ is partitioned conformally with $\operatorname{vec}(\boldsymbol{x})$. Since $\boldsymbol{\lambda}$ is algebraically simple by assumption it follows that $\dim\ker W_i(\boldsymbol{\lambda}) = 1$ and therefore $z_i = \gamma_i x_i$ for certain $\gamma_i \in \mathbb{C}$. Now we know that $G^*\boldsymbol{z} = 0$ on the one hand, and on the other hand $G^*\boldsymbol{z} = [\gamma_1 \ \cdots \ \gamma_k]^T$ so $\gamma_i = 0$ for $i = 1, \ldots, k$ from which we conclude that $\boldsymbol{z} = 0$. □

It follows from Lemma 7.4.3 and (7.4.7) that

$$\operatorname{vec}(\boldsymbol{\Delta x}) = \left(PDP\big|_{\mathcal{G}^\perp}\right)^{-1} P\operatorname{diag}(\Delta W_i(\boldsymbol{\lambda}))\operatorname{vec}(\boldsymbol{x}),$$

where $PDP\big|_{\mathcal{G}^\perp}$ is a restriction of PDP to $\mathcal{G}^\perp$. This gives

$$\|\operatorname{vec}(\boldsymbol{\Delta x})\| \leq \varepsilon \left\| \left(PDP\big|_{\mathcal{G}^\perp}\right)^{-1} P \right\|_{\boldsymbol{\beta},\boldsymbol{n}} + \mathcal{O}(\varepsilon^2), \tag{7.4.9}$$

where

$$\|A\|_{\boldsymbol{\beta},\boldsymbol{n}} := \max\left\{ \|Az\| : z = [z_1^T \ \cdots \ z_k^T]^T,\ z_i \in \mathbb{C}^{n_i},\ \|z_i\| \leq \beta_i,\ i = 1, \ldots, k \right\}$$

and $n = [n_1 \ \cdots \ n_k]^T$. One can view this $(\boldsymbol{\beta},\boldsymbol{n})$-norm as a block version of (7.2.1). This leads to the next theorem.

Theorem 7.4.4

$$\kappa(\boldsymbol{x},\boldsymbol{W}) = \left\| \left(PDP\big|_{\mathcal{G}^\perp}\right)^{-1} P \right\|_{\boldsymbol{\beta},\boldsymbol{n}}. \tag{7.4.10}$$

Proof: In the discussion preceding the theorem we showed in (7.4.9) that

$$\kappa(\boldsymbol{x},\boldsymbol{W}) \leq \left\| \left(PDP\big|_{\mathcal{G}^\perp}\right)^{-1} P \right\|_{\boldsymbol{\beta},\boldsymbol{n}}.$$

What remains is to construct a perturbation for which equality is attained. Suppose that for $\boldsymbol{z} = [z_1^T \ \cdots \ z_k^T]^T$ such that $\|z_i\| \leq \beta_i$ for $i = 1, \ldots, k$ we have

$$\left\| \left(PDP|_{\mathcal{G}^\perp} \right)^{-1} P \right\|_{\boldsymbol{\beta}, \boldsymbol{n}} = \left\| \left(PDP|_{\mathcal{G}^\perp} \right)^{-1} P\boldsymbol{z} \right\|.$$

Equality in (7.4.9) is then attained if we take

$$\Delta V_{i0} = -\frac{\varepsilon \|E_{i0}\|}{\alpha_i} z_i x_i^*,$$

for $i, j = 1, \ldots, k$. □

Remark 7.4.5 If we take $g_i = y_i$ for $i = 1, \ldots, k$ then D is a bijection as an operator from $\mathcal{Y}^\perp$ to $\mathcal{Y}^\perp$, where $\mathcal{Y} := \operatorname{span}(Y)$, and we have $\left\| \left(PDP|_{\mathcal{Y}^\perp} \right)^{-1} P \right\|_{\boldsymbol{\beta}, \boldsymbol{n}} = \left\| P \left(D|_{\mathcal{Y}^\perp} \right)^{-1} P \right\|_{\boldsymbol{\beta}, \boldsymbol{n}}$. ⊘

From (7.4.9) we can produce an upper bound for the norm of $\widetilde{\boldsymbol{x}} - \boldsymbol{x}$. If we consider only first order terms then we have

$$\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq \|\Delta x_1\| + \cdots + \|\Delta x_k\| + \mathcal{O}(\varepsilon^2)$$

and it follows that

$$\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq \sqrt{k} \, \|\operatorname{vec}(\boldsymbol{\Delta x})\| + \mathcal{O}(\varepsilon^2).$$

If we apply (7.4.9) then we obtain the bound

$$\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq \varepsilon \sqrt{k} \left\| \left(PDP|_{\mathcal{G}^\perp} \right)^{-1} P \right\|_{\boldsymbol{\beta}, \boldsymbol{n}} + \mathcal{O}(\varepsilon^2).$$

7.5 Pseudospectra

Another tool for the study of the sensitivity of the eigenvalues to perturbations are pseudospectra. They have been studied for the standard (see, e.g., [87, 88], and [66]) and generalized eigenproblem [24] and for the polynomial eigenvalue problem (see, e.g., [85]). We extend the definition of pseudospectrum to the multiparameter eigenvalue problem.

We define the *ε-pseudospectrum of* $\boldsymbol{W}$ by

$$\Lambda_\varepsilon(\boldsymbol{W}) = \Big\{ \boldsymbol{\lambda} \in \mathbb{C}^k \quad : \quad W_i(\boldsymbol{\lambda}) + \Delta W_i(\boldsymbol{\lambda}) \text{ singular}, \qquad (7.5.1)$$
$$\|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|, \ i = 1, \ldots, k; \ j = 0, \ldots, k \Big\}.$$

If we define the *ε-pseudospectrum of W_i* by

$$\Lambda_\varepsilon(W_i) = \left\{ \boldsymbol{\lambda} \in \mathbb{C}^k \ : \ W_i(\boldsymbol{\lambda}) + \Delta W_i(\boldsymbol{\lambda}) \text{ singular}, \ \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|, \ j = 0, \ldots, k \right\},$$

then it is easy to see that

$$\Lambda_\varepsilon(\boldsymbol{W}) = \Lambda_\varepsilon(W_1) \cap \Lambda_\varepsilon(W_2) \cap \cdots \cap \Lambda_\varepsilon(W_k).$$

Theorem 7.5.1

$$\begin{aligned}
\Lambda_\varepsilon(\boldsymbol{W}) &= \{ \boldsymbol{\lambda} \in \mathbb{C}^k \ : \ \eta(\boldsymbol{\lambda}) \leq \varepsilon \text{ for } i = 1, \ldots, k \} \\
&= \{ \boldsymbol{\lambda} \in \mathbb{C}^k \ : \ \sigma_{\min}(W_i(\boldsymbol{\lambda})) \leq \varepsilon \widetilde{\beta}_i \text{ for } i = 1, \ldots, k \} \\
&= \{ \boldsymbol{\lambda} \in \mathbb{C}^k \ : \ \|W_i(\boldsymbol{\lambda})^{-1}\| \geq 1/(\varepsilon \widetilde{\beta}_i) \text{ for } i = 1, \ldots, k \} \\
&= \{ \boldsymbol{\lambda} \in \mathbb{C}^k \ : \ \exists u_i, \ \|u_i\| = 1 \text{ with } \|W_i(\boldsymbol{\lambda}) u_i\| \leq \varepsilon \widetilde{\beta}_i \text{ for } i = 1, \ldots, k \}.
\end{aligned}$$

Proof: The first equality follows readily from Definition (7.5.1). For the second equality Proposition 7.3.3 can be applied. The last two equalities follow from the identity $\min_{x \neq 0} \|Ax\|/\|x\| = \|A^{-1}\|^{-1} = \sigma_{\min}(A)$, with the convention that $\|A^{-1}\| = \infty$ if A is singular. □

Pseudospectra for the MEP have a property that is different from pseudospectra for the standard eigenvalue problem $Ax = \lambda x$: if ε is large enough then $\Lambda_\varepsilon(\boldsymbol{W})$ will be unbounded. This is the subject of the rest of this section.

If $\boldsymbol{W}$ is a right definite MEP, then we may be interested in the smallest perturbation that makes $\boldsymbol{W} + \Delta\boldsymbol{W}$ not right definite. Again, here we restrict the perturbations ΔV_{ij} to be Hermitian. We can define the distance to the closest non right definite MEP as

$$\begin{aligned}
\xi(\boldsymbol{W}) := \min\{ \varepsilon \ : \ & \boldsymbol{W} + \Delta\boldsymbol{W} \text{ is not right definite}, \Delta V_{ij}^* = \Delta V_{ij}, \\
& \|\Delta V_{ij}\| \leq \varepsilon \|E_{ij}\|, \ i = 1, \ldots, k; \ j = 0, \ldots, k \}.
\end{aligned}$$

In the next theorem we show that $\xi(\boldsymbol{W})$ is bounded by the minimal ε for which the pseudospectrum is unbounded.

Theorem 7.5.2

$$\xi(\boldsymbol{W}) \leq \min\{ \varepsilon : \Lambda_\varepsilon(\boldsymbol{W}) \text{ is unbounded} \}. \tag{7.5.2}$$

Proof: It is sufficient to prove that a right definite $\boldsymbol{W}$ cannot have "infinite" eigenvalues. If $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ is an eigenvalue of a right definite $\boldsymbol{W}$ with corresponding normalized eigenvector $\boldsymbol{x} = x_1 \otimes \cdots \otimes x_k$, then it follows that λ_i is equal to the tensor Rayleigh quotient [64]

$$\lambda_i = \frac{\boldsymbol{x}^* \Delta_i \boldsymbol{x}}{\boldsymbol{x}^* \Delta_0 \boldsymbol{x}}$$

for $i = 1, \ldots, k$. From (7.1.3) and right definiteness, we get the bound (7.5.2). □

7.6 Numerical experiments

We present some numerical examples obtained with MATLAB 5.3. For all examples we take $E_{ij} = V_{ij}$ for all i, j (corresponding to relative perturbations). We draw all pseudospectra by computing $\sigma_{\min}(W_i(\boldsymbol{\lambda}))$ in all grid points by MATLAB's `svd`. For more efficiency one could try to use similar ideas as mentioned in [87], but we will pay no further attention to this. The size of the grid used in the examples is 400×400.

Experiment 7.6.1 For the first numerical example we take the right definite two-parameter eigenvalue problem

$$W_1(\boldsymbol{\lambda}) = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} - \lambda_1 \begin{bmatrix} 2.2 & 1 \\ 1 & 2.3 \end{bmatrix} - \lambda_2 \begin{bmatrix} 0.1 & -1 \\ -1 & 0.1 \end{bmatrix},$$

$$W_2(\boldsymbol{\lambda}) = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} - \lambda_1 \begin{bmatrix} 1 & -0.2 \\ -0.2 & -0.1 \end{bmatrix} - \lambda_2 \begin{bmatrix} 2 & -0.1 \\ -0.1 & 4 \end{bmatrix}.$$

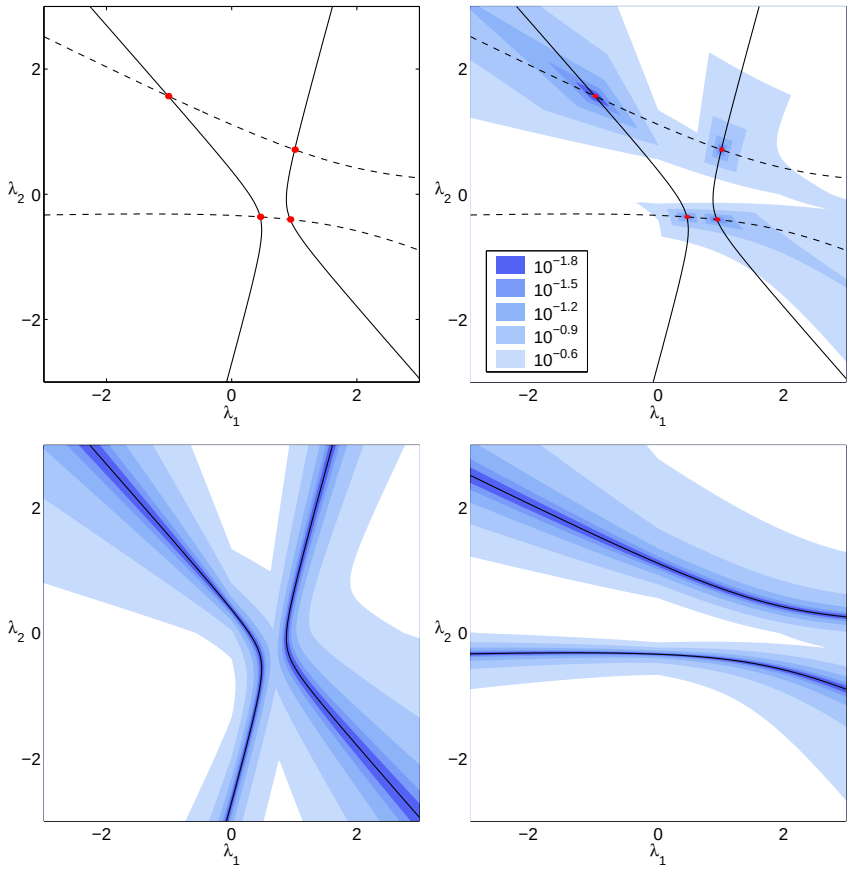


FIGURE 7.1: Pseudospectra for Example 7.6.1. Top left: The eigenvalues are intersections of the eigencurves $\det W_1(\boldsymbol{\lambda}) = 0$ (solid line) and $\det W_2(\boldsymbol{\lambda}) = 0$ (dashed line). Top right: pseudospectra for $\varepsilon = 10^{-1.8}$, $10^{-1.5}$, $10^{-1.2}$, $10^{-0.9}$, and $10^{-0.6}$. Bottom: pseudospectra for W_1 (left) and W_2 (right).

The eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ are intersection points of the eigenvalue curves defined by $\det(W_1(\boldsymbol{\lambda})) = 0$ and $\det(W_2(\boldsymbol{\lambda})) = 0$ as depicted in the top left picture in Figure 7.1.

The pseudospectra for $\varepsilon = 10^{-0.6}$, $10^{-0.3}$, 10^{0}, and $10^{0.3}$ are shown in the top right picture in Figure 7.1. One can see that the boundaries of the pseudospectra are not differentiable. The reason is that pseudospectra are intersections of pseudospectra for W_1 and W_2, which are shown on the bottom left and bottom right picture in Figure 7.1, respectively.

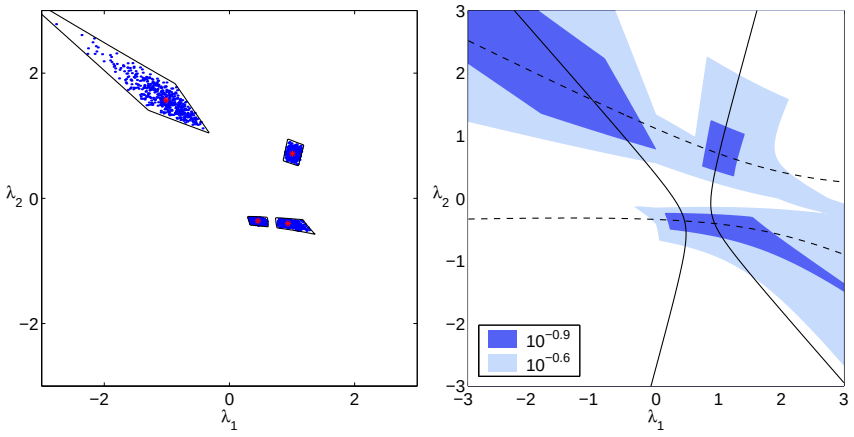
TABLE 7.1: Eigenvalues and their condition numbers for the right definite two-parameter problem in Example 7.6.1.

λ_1	λ_2	$\kappa(\boldsymbol{\lambda}, \boldsymbol{W})$
−1.0142	1.5688	4.66
0.4556	−0.3613	2.42
0.9360	−0.4025	3.34
1.0069	0.7125	3.37

The eigenvalues together with the corresponding condition numbers are presented in Table 7.1. To obtain the condition number of an eigenvalue we have to compute $\|B_0^{-1}\|_{\boldsymbol{\beta}}$. Since the problem is right definite and all matrices V_{ij} are real we have to consider only real vectors in definition (7.2.1) of $\|B_0^{-1}\|_{\boldsymbol{\beta}}$. This fact makes it easy to compute the $\boldsymbol{\beta}$-norm as we only have to compute a finite number of norms. In particular, for a right definite two-parameter case we have

$$\|B_0^{-1}\|_{\boldsymbol{\beta}} = \max\{\, \|B_0^{-1}z\| \;:\; z \in \mathbb{R}^2,\ |z_i| = \beta_i \text{ for } i = 1, 2\,\}.$$

By comparing the results of Table 7.1 and Figure 7.1 one can see that the eigenvalue with the largest condition number has the "largest pseudospectrum" as may be expected.



(a) (b)

FIGURE 7.2: Left: eigenvalues of 500 randomly perturbed two-parameter eigenvalue problems of Example 7.6.1, where each ΔV_{ij} is a symmetric matrix such that $\|\Delta V_{ij}\| = 10^{-1.2}\|V_{ij}\|$, and the pseudospectrum for $\varepsilon = 10^{-1.2}$. Right: pseudospectra for Example 7.6.1 for $\varepsilon = 10^{-0.9}$ and $\varepsilon = 10^{-0.6}$.

Figure 7.2(a) shows eigenvalues of 500 randomly perturbed problems, where each ΔV_{ij} is a random symmetric matrix such that $\|\Delta V_{ij}\| = 10^{-1.2}\|V_{ij}\|$. One can see that all dots in Figure 7.2 lie in the interior of the pseudospectrum for $\varepsilon = 10^{-1.2}$.

Figure 7.2(b) presents pseudospectra for $\varepsilon = 10^{-0.9}$ and $\varepsilon = 10^{-0.6}$ on a larger area. One may suspect that here, in contrast to the eigenvalue problem $Ax = \lambda x$, a pseudospectrum may be unbounded.

Figures 7.1 and 7.2 suggest that the sensitivity of the eigenvalue is related to the angle of the intersection between the curves $\det(W_1(\boldsymbol{\lambda})) = 0$ and $\det(W_2(\boldsymbol{\lambda})) = 0$. We observe that the pseudospectrum is large when the angle of the intersection is small. The following proposition (which can be easily generalized to MEPs with more than two parameters) justifies this observation.

Proposition 7.6.2 *Let $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathbb{R}^2$ be an algebraically simple eigenvalue of a real right definite two-parameter eigenvalue problem* $\mathbf{W}$ *and let $\boldsymbol{x} = x_1 \otimes x_2$ and $\boldsymbol{y} = y_1 \otimes y_2$ be the corresponding normalized right and left eigenvector, respectively. Then*

$$B_0 = -\begin{bmatrix} \pm\prod_{j=1}^{n_1-1}\sigma_j^{(1)}(\boldsymbol{\mu}) & 0 \\ 0 & \pm\prod_{j=1}^{n_2-1}\sigma_j^{(2)}(\boldsymbol{\mu}) \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f_1}{\partial\lambda_1}(\boldsymbol{\mu}) & \frac{\partial f_1}{\partial\lambda_2}(\boldsymbol{\mu}) \\ \frac{\partial f_2}{\partial\lambda_1}(\boldsymbol{\mu}) & \frac{\partial f_2}{\partial\lambda_2}(\boldsymbol{\mu}) \end{bmatrix},$$

where $f_i(\boldsymbol{\lambda}) = \det W_i(\boldsymbol{\lambda})$ and where $\sigma_1^{(i)}(\boldsymbol{\mu}) \geq \sigma_2^{(i)}(\boldsymbol{\mu}) \geq \cdots \geq \sigma_{n_i-1}^{(i)}(\boldsymbol{\mu}) > 0$ are nonzero singular values of $W_i(\boldsymbol{\mu})$ for $i = 1, 2$.

Proof: We define $Z(t) = V_{10} - tV_{11} - \mu_2 V_{12}$ and $g(t) = \det(Z(t))$. Since $Z(t)$ is a real analytic function of t, there exists an analytic singular value decomposition (see [15])

$$Z(t) = U(t)\Sigma(t)V(t)^T \tag{7.6.1}$$

such that

1. $U(t)$ and $V(t)$ are orthogonal matrices,
2. $\Sigma(t) = \operatorname{diag}(\sigma_1(t), \ldots, \sigma_{n_1}(t))$ is a diagonal matrix,
3. the elements of $U(t)$, $\Sigma(t)$, and $V(t)$ are analytic functions of t in a small neighborhood of μ_1, and
4. $Z(\mu_1) = U(\mu_1)\Sigma(\mu_1)V(\mu_1)^T$ is a singular value decomposition of $W_i(\boldsymbol{\mu})$.

We may consider (7.6.1) as a singular value decomposition of $Z(t)$ where the singular values are not necessarily nonnegative and ordered for all t. Let $u_{n_i}(t)$ and $v_{n_i}(t)$ denote the n_ith column of $U(t)$ and $V(t)$, respectively. Since $\boldsymbol{\mu}$ is an algebraically simple eigenvalue, $\sigma_{n_i}(\mu_1) = 0$, $\sigma_{n_i-1}(\mu_1) \neq 0$, $v_{n_i}(\mu_1) = x_i$, and $u_{n_i}(\mu_1) = y_i$.

If we differentiate $\sigma_{n_1}(t) = u_{n_1}(t)^T Z(t) v_{n_1}(t)$ then we obtain

$$\frac{d\sigma_{n_1}}{dt}(\mu_1) = -y_1^T V_{11} x_1 = -(B_0)_{11}. \tag{7.6.2}$$

From $g(t) = \mp\sigma_1(t)\sigma_2(t)\cdots\sigma_n(t)$ and (7.6.2) it follows that

$$\frac{\partial f_1}{\partial \lambda_1}(\boldsymbol{\mu}) = \frac{dg}{dt}(\mu_1) = \pm\sigma_1^{(1)}(\boldsymbol{\mu})\cdots\sigma_{n_1-1}^{(1)}(\boldsymbol{\mu})(B_0)_{11}.$$

In order to complete the proof one has to repeat the above procedure for all partial derivatives $\frac{\partial f_i}{\partial \lambda_j}(\boldsymbol{\mu})$ for $i, j = 1, 2$. □

It follows from Theorem 7.4.1 and (7.2.2) that $\|B_0^{-1}\|$ has a great impact on the sensitivity of the eigenvalue $\boldsymbol{\lambda}$. As follows from Proposition 7.6.2, $\|B_0^{-1}\|$ may be large when the angle of the intersection between the curves $\det(W_1(\boldsymbol{\lambda})) = 0$ and $\det(W_2(\boldsymbol{\lambda})) = 0$ is small. ⊘

Experiment 7.6.3 For the second example we take the two-parameter Sturm–Liouville problem

$$\begin{aligned} W_1(\boldsymbol{\lambda})x_1(t_1) &= -x_1''(t_1) - (\lambda_1 + \lambda_2\cos 2t_1)x_1(t_1), \\ W_2(\boldsymbol{\lambda})x_2(t_2) &= -x_2''(t_2) - \lambda_2 x_2(t_2) \end{aligned} \tag{7.6.3}$$

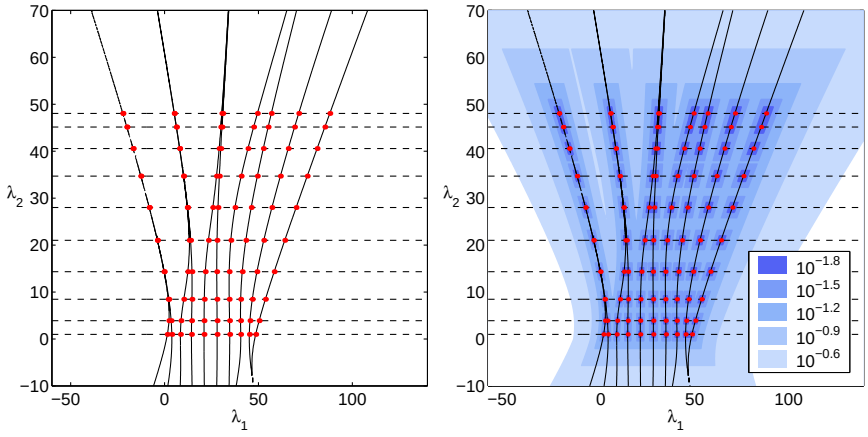
with boundary conditions $x_i(0) = x_i(\pi) = 0$ for $i = 1, 2$, studied in [6]. The second equation of (7.6.3) yields that $\lambda_2 = 1^2, 2^2, 3^2, \ldots$ and then it follows from the first equation of (7.6.3) that λ_1 is an eigenvalue of Mathieu's equation with parameter λ_2.

If we take $h = \pi/n$ and apply the finite-difference method to the two-parameter boundary-value problem (7.6.3) using symmetric differences $y_i' \approx (y_{i+1} - y_{i-1})/(2h)$ and $y_i'' \approx (y_{i+1} - 2y_i + y_{i-1})/h^2$ for the derivatives y' and y'', then we obtain an algebraic two–parameter problem where

$$\begin{gathered} V_{10} = V_{20} = \frac{1}{h^2}\,\mathrm{tridiag}(1, -2, 1), \\ V_{11} = I, \quad V_{21} = 0, \\ V_{12} = \mathrm{diag}\Big(\cos\frac{2\pi}{n+1}, \cos\frac{4\pi}{n+1}, \ldots, \cos\frac{2n\pi}{n+1}\Big), \quad V_{22} = I_n. \end{gathered} \tag{7.6.4}$$

The eigenvalues of the above algebraic two-parameter problem are approximations to the eigenvalues of (7.6.3) with order of approximation $\mathcal{O}(h^2)$.

Figure 7.3 shows eigenvalues and pseudospectra for the algebraic two-parameter approximation (7.6.4) of (7.6.3) for $n = 10$. The left figure shows eigenvalues as the points where eigencurves $\det(W_1(\boldsymbol{\lambda})) = 0$ (solid line) and $\det(W_2(\boldsymbol{\lambda})) = 0$ (dashed line) intersect. Note that the lines $\det(W_2(\boldsymbol{\lambda})) = 0$ do not agree with the known result $\lambda_2 = 1^2, 2^2, 3^2, \ldots$, since the eigenvalues in Figure 7.3 are the eigenvalues of the algebraic approximation (7.6.4) and not of the original problem (7.6.3). The eigenvalues occur in groups of two for a fixed λ_2. In some of these pairs the eigenvalues are so clustered that they look like a single eigenvalue in Figure 7.3, an example of such a pair is $(-12.6225, 34.7056)$ and $(-12.6215, 34.7056)$. The right figure with the pseudospectra for $\varepsilon = 10^{-1.8}, 10^{-1.5}, \ldots, 10^{-0.6}$ indicates that the fact that some of the eigenvalues are clustered does not have much influence on their pseudospectra; the eigenvalues are well conditioned. ⊘



(a) (b)

FIGURE 7.3: Spectrum and pseudospectra for the algebraic two-parameter approximation of Example 7.6.3, where $n = 10$ and $\varepsilon = 10^{-1.8}$, $10^{-1.5}$, $10^{-1.2}$, $10^{-0.9}$, and $10^{-0.6}$.

7.7 Conclusions

We have studied the backward error, condition numbers, and pseudospectra for the MEP. The results can be viewed as a generalization of the theory for the generalized eigenvalue problem and have similarities with the results for the polynomial eigenvalue problem. We also studied the nearness of a right definite MEP to a non right definite MEP and established that it is connected with unbounded pseudospectra.

Chapter 8

Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem

Abstract. We consider the quadratic eigenvalue problem $\lambda^2 Ax + \lambda Bx + Cx = 0$. Suppose that u is an approximation to an eigenvector x (for instance obtained by a subspace method), and that we want to determine an approximation to the corresponding eigenvalue λ. The usual approach is to impose the Galerkin condition $r(\theta, u) = (\theta^2 A + \theta B + C)u \perp u$ from which it follows that θ must be one of the two solutions to the quadratic equation $(u^*Au)\theta^2 + (u^*Bu)\theta + (u^*Cu) = 0$. An unnatural aspect is that if $u = x$, the second solution has in general no meaning. When u is not very accurate, it may not be clear which solution is the best. Moreover, when the discriminant of the equation is small, the solutions may be very sensitive to perturbations in u.

In this chapter we therefore examine alternative approximations to λ. We compare the approaches theoretically and by numerical experiments. The methods are extended to approximations from subspaces and to the polynomial eigenvalue problem.

Key words: quadratic eigenvalue problem, Rayleigh quotient, Galerkin, minimum residual, subspace method, polynomial eigenvalue problem, backward error, refining a Ritz pair.

AMS subject classification: 65F15 (65F50).

8.1 Introduction

First consider the eigenvalue problem $Ax = \lambda x$, with A a real symmetric $n \times n$ matrix, where $n \geq 2$. Suppose that we have an approximate eigenvector u with unit norm. The usual approximation to the corresponding eigenvalue is given by the *Rayleigh quotient*

*Based on joint work with Henk A. van der Vorst, see Section 1.5.

of u

$$\rho = \rho(u) := \frac{u^*Au}{u^*u}. \tag{8.1.1}$$

This Rayleigh quotient has the following attractive properties:

1. ρ satisfies the *Ritz–Galerkin condition* on the *residual* $r(\theta, u)$:

$$r(\rho, u) := Au - \rho u \perp u. \tag{8.1.2}$$

2. ρ satisfies the *minimum residual condition* on the residual

$$\rho = \operatorname*{argmin}_{\theta \in \mathbb{R}} \|Au - \theta u\|. \tag{8.1.3}$$

3. The function $\rho(u)$ has as its *stationary points* exactly the n eigenvectors x_i, and even

$$\frac{d\rho}{du}(x_i) = 0. \tag{8.1.4}$$

(Recall that stationary means that all directional derivatives are zero.) This implies that a first order perturbation of the eigenvector only gives a second order perturbation of the Rayleigh quotient: $\rho(x_i + h) = \lambda_i + \mathcal{O}(\|h\|^2)$.

Remark 8.1.1 When A is normal, (8.1.2) and (8.1.3) hold, and the eigenvectors are still stationary points, though the Rayleigh quotient is in general not differentiable. When A is nonnormal, (8.1.2) and (8.1.3) also hold, but the eigenvectors are in general not stationary points. One can show that instead of this, the *two-sided Rayleigh quotient* $\rho(u, v) := \frac{v^*Au}{v^*u}$ has as its stationary points exactly the right/left eigenvectors combinations (x_i, y_i), provided that $y_i^*x_i \neq 0$, see also (2.3.1). This suggests replacing the Ritz–Galerkin condition (8.1.2) by the *Petrov–Galerkin condition*

$$r(\theta, u) = Au - \theta u \perp v,$$

which is used in two-sided methods such as two-sided Lanczos [51] and two-sided Jacobi–Davidson (Chapter 2). However, we use no information about the left eigenvector in this chapter. ⊘

Now consider the quadratic eigenvalue problem

$$Q(\lambda)x := (\lambda^2 A + \lambda B + C)x = 0, \tag{8.1.5}$$

where A, B, and C are (complex) $n \times n$ matrices. In this chapter, we examine generalizations of the properties (8.1.2)–(8.1.4) for the quadratic eigenvalue problem, to derive different eigenvalue approximations. See [86] for an overview of the quadratic eigenvalue problem. For an eigenvector x we have either one of the following properties:

- Ax and Bx are dependent, then Cx is also dependent, and there are two eigenvalues (counting multiplicities) corresponding to x,

- Ax and Bx are independent, Cx lies in the span of Ax and Bx, and the corresponding eigenvalue λ is unique.

We will assume in the remainder of the chapter that x has the second property. For a motivation see Remark 8.2.3 at the end of Section 8.2.4.

Now let u be an approximation to an eigenvector x, for instance one obtained by a subspace method. We will also assume that Au and Bu are independent, which is not unnatural in view of the assumptions that Ax and Bx are independent, and $u \approx x$; see also Remark 8.2.3. We study ways to determine an approximation θ to the eigenvalue λ, from the information of u. In Section 8.2.1 we discuss the "classical" one-dimensional Galerkin method, while in Sections 8.2.2, 8.2.3, and 8.2.4 we introduce new approaches. The methods are extended to subspaces of dimension larger than one and to the polynomial eigenvalue problem in Section 8.3. Numerical experiments and a conclusion can be found in Section 8.4 and 8.5.

8.2 Approximations for the quadratic eigenproblem

8.2.1 One-dimensional Galerkin

For an approximate eigenpair $(\theta, u) \approx (\lambda, x)$ we define the *residual* $r(\theta, u)$ by

$$r(\theta, u) := Q(\theta)u = (\theta^2 A + \theta B + C)u.$$

The usual approach to derive an approximate eigenvalue θ from the approximate eigenvector u is to impose the Galerkin condition $r(\theta, u) \perp u$. Then it follows that $\theta = \theta(u)$ must be one of the two solutions to the quadratic equation

$$\alpha\theta^2 + \beta\theta + \gamma = 0, \tag{8.2.1}$$

where $\alpha = \alpha(u) = u^*Au$, $\beta = \beta(u) = u^*Bu$, and $\gamma = \gamma(u) = u^*Cu$. An unnatural aspect is that if $u = x$, the second solution of (8.2.1) has in general no meaning. If u is close to x, we will be able to decide which one is best by looking at the norms of the residuals. But if u is not very accurate, it may not be clear which solution is the best. For instance, this may happen when we try to solve (8.1.5) by a subspace method; in the beginning of the process, the search space may not contain good approximations to an eigenvector. This problem is also mentioned in [5, p. 282].

A second, related problem is the subject of the rest of this subsection. A nice property that an approximate eigenvalue can (or should) have is that it is close to the eigenvalue if the corresponding approximate eigenvector is close to the eigenvector. In other words, we like the situation where

$$|\theta(x + h) - \lambda| = |\theta(x + h) - \theta(x)| \text{ is small}$$

for small $\|h\|$. When θ is differentiable with respect to u in the point x this is equivalent to the condition

$$\left\| \frac{\partial \theta}{\partial u}(x) \right\| \text{ is small.} \tag{8.2.2}$$

The one-dimensional Galerkin approach (8.2.1) defines θ implicitly as a function of α, β, and γ, say $f(\theta, \alpha, \beta, \gamma) = 0$, with $f(\lambda, \alpha(x), \beta(x), \gamma(x)) = 0$. Define the discriminant δ by

$$\delta = \delta(u) := \beta^2 - 4\alpha\gamma. \tag{8.2.3}$$

When $\delta(x) \neq 0$, the Implicit Function Theorem states that locally θ is a function of α, β, and γ, say $\theta = \varphi(\alpha, \beta, \gamma)$, and that

$$\begin{aligned} D\varphi(\alpha(x), \beta(x), \gamma(x)) &= -((D_\theta f)^{-1} D_{(\alpha,\beta,\gamma)} f)(\lambda, \alpha(x), \beta(x), \gamma(x)) \\ &= \pm \frac{1}{\sqrt{\delta(x)}} \cdot (\lambda^2, \lambda, 1). \end{aligned}$$

So when δ is small, which means that (8.2.1) has two roots that are close, then the solutions of (8.2.1) may be very sensitive to perturbations in u (although in general, the coefficients α, β, and γ are not differentiable with respect to u). Therefore, we may expect that $|\theta - \lambda|$ is large for small perturbations of x, see also the numerical experiments. Thus, the second solution of (8.2.1) is not only useless, but it may also hinder the accuracy of the solution that is of interest!

We therefore examine alternative ways to approximate λ. We generalize the Galerkin property (8.1.2) and minimum residual property (8.1.3) for the quadratic eigenvalue problem in the following three subsections.

Remark 8.2.1 As in the standard eigenvalue problem, $\theta = \theta(u, v)$ as solution of

$$(v^*Au)\theta^2 + (v^*Bu)\theta + (v^*Cu) = 0$$

is stationary in the right/left eigenvector combinations (x_i, y_i). However, we use no information about the (approximate) left eigenvector in this chapter. ⊘

8.2.2 Two-dimensional Galerkin

In the standard eigenvalue problem, we deal with two vectors u and Au, which are asymptotically (by which we mean when $u \to x$) dependent. Therefore it is natural to take the length of the projection of Au onto the span of u as an approximation to the eigenvalue, which is exactly what the Rayleigh quotient $\rho(u)$ (see (8.1.1)) does. For the generalized eigenvalue problem we have a similar situation.

In the quadratic eigenvalue problem, however, we deal with three vectors Au, Bu, and Cu, which asymptotically lie in a plane. Therefore it is natural to consider the projection of these three vectors onto a certain plane, spanned by two independent vectors p and q. To generalize the approach of (8.1.2), define the *generalized residual* $r(\mu, \nu, u)$ by

$$r(\mu, \nu, u) := (\mu A + \nu B + C)u. \tag{8.2.4}$$

The idea behind this is that we want to impose conditions on r such that μ forms an approximation to λ^2, and ν an approximation to λ. Then both μ/ν and ν may be good approximations to the eigenvalue λ. A generalization of (8.1.2) is obtained by imposing

two Galerkin conditions $r(\mu,\nu,u) \perp p$ and $r(\mu,\nu,u) \perp q$ for specific independent vectors p, q. This leads to the system

$$W^*Z \begin{bmatrix} \mu \\ \nu \end{bmatrix} = -W^*Cu, \quad \text{where} \quad W = [\, p \;\; q \,], \quad Z = [\, Au \;\; Bu \,]. \tag{8.2.5}$$

When W^*Z is nonsingular, (8.2.5) defines a unique μ and ν. A logical choice for p and q is any linear combination of Au, Bu, and Cu. Specifically, one could take the "least-squares" plane such that

$$\|(I-\Pi)Au\|^2 + \|(I-\Pi)Bu\|^2 + \|(I-\Pi)Cu\|^2$$

is minimal, where Π is the orthogonal projection onto the plane. An advantage of this least-squares plane is that it takes the norm of the vectors Au, Bu, and Cu into account.

Let z be the normal in $\text{span}(Au, Bu, Cu)$ of the sought plane, then one may verify that $\|(I-\Pi)Au\|^2 = \|(z^*Au)z\|^2 = |z^*Au|^2$. If D denotes the $n \times 3$ matrix with Au, Bu, and Cu as its columns, then z is the vector of unit length such that $\|z^*D\|^2$ is minimal. So we conclude that z is the minimal left singular vector of D, and for p and q we can take the two "largest" left singular vectors. Another choice for p and q, as well as its meaning, are discussed in Section 8.2.4.

This two-dimensional Galerkin method yields three approximations to the eigenvalue. Besides the already mentioned possibilities μ/ν and ν, we can determine a third approximation by solving for $\theta \in \mathbb{C}$ such that

$$\left\| \begin{bmatrix} \theta^2 \\ \theta \end{bmatrix} - \begin{bmatrix} \mu \\ \nu \end{bmatrix} \right\|^2 \tag{8.2.6}$$

is minimal. We will indicate this solution as the "argmin" solution. Differentiating (8.2.6) with respect to $\text{Re}(\theta)$ and $\text{Im}(\theta)$ gives two mixed equations of degree three in $\text{Re}(\theta)$ and $\text{Im}(\theta)$. We may try to solve these equations by modern algorithms for systems of polynomials, see for instance [99, 102]. Another approach to solve the two coupled cubic equations is to form an equation, the so-called *resultant* in only $\text{Re}(\theta)$ or $\text{Im}(\theta)$ (see Section 8.4). It appears that in this case, the degree of the resultant is (only) five. Since we use this equation to find $\text{Re}(\theta)$ and $\text{Im}(\theta)$, only the real solutions are of interest. The approach via the resultant may be numerically somewhat less stable, but an advantage is that one can use widely available mathematical packages such as MAPLE, as is done for experiments in Section 8.4. We will summarize the two-dimensional Galerkin method, and the two-dimensional minimal residual method (to be discussed in Section 8.2.4) in Algorithm 8.2.1.

Let us consider the sensitivity of the approximations derived by the two-dimensional Galerkin method (8.2.5). As seen, we have three possible approximation to the eigenvalue: μ/ν, ν, and the "argmin" solution. Asymptotically, if we take $u = x$, then we may assume $W = Z = [\, Ax \;\; Bx \,]$. When we "freeze" $W = Z = [\, Ax \;\; Bx \,]$ and differentiate (8.2.5), it can be seen that the sensitivity of μ and ν is related to $\kappa(Z)$, the condition number of Z (cf. [31, Section 5.3.7]). For comparison, we will give $\kappa(Z)$ in the experiments.

When μ and ν are differentiable with respect to u (as is the case for the quasi-hyperbolic quadratic eigenvalue problem, see below), then we have that

$$\frac{\partial(\mu/\nu)}{\partial u}(x) = \frac{1}{\lambda} \cdot \frac{\partial \mu}{\partial u}(x) - \frac{\partial \nu}{\partial u}(x). \tag{8.2.7}$$

This suggests that μ/ν might give inaccurate approximations for small λ, which is confirmed by numerical experiments, see Experiment 8.4.1. In general, μ and ν are not differentiable with respect to u, but still μ/ν will be sensitive for small λ. The sensitivity of the "argmin" solution will depend on the coefficients of the two defining cubic equations; we omit a (difficult) analysis.

8.2.3 One-dimensional minimum residual

Another approach generalizes the minimum residual approach (8.1.3). We try to minimize the norm of the residual with respect to θ:

$$\min_{\theta \in \mathbb{C}} \|(\theta^2 A + \theta B + C)u\|^2. \tag{8.2.8}$$

For complex θ, differentiating (8.2.8) with respect to $\mathrm{Re}(\theta)$ and $\mathrm{Im}(\theta)$ gives two mixed equations of degree three in $\mathrm{Re}(\theta)$ and $\mathrm{Im}(\theta)$. As in the previous subsection, we may use available algorithms [99, 102], or form the resultant, which in this case has degree nine (in only $\mathrm{Re}(\theta)$ or $\mathrm{Im}(\theta)$), see also Section 8.4.

In the special case that we know that λ is real, we would like to have a real approximation θ. Then differentiating (8.2.8) with respect to θ gives the cubic equation with real coefficients

$$4\,\|Au\|^2\theta^3 + 6\ \mathrm{Re}((Au)^*Bu)\theta^2 + 2\,(\|Bu\|^2 + 2\ \mathrm{Re}((Cu)^*Au)\theta + 2\ \mathrm{Re}((Cu)^*Bu) = 0, \tag{8.2.9}$$

which may be solved analytically. For instance, this is the case for the important class of *quasi-hyperbolic quadratic eigenvalue problems*:

Definition 8.2.2 (cf. [86, p. 257]) A quadratic eigenvalue problem $Q(\lambda)x = 0$ is called *quasi-hyperbolic* if A is Hermitian positive definite, B and C are Hermitian, and for all eigenvectors of $Q(\lambda)$ we have

$$(x^*Bx)^2 > 4(x^*Ax)(x^*Cx).$$

$\oslash$

It is easy to see that all eigenvalues of quasi-hyperbolic quadratic eigenvalue problems are real.

We would like to stress that the one-dimensional minimum residual approach may also suffer from the same difficulties as the one-dimensional Galerkin method. In some cases, there may be more than one solution to choose from, and the irrelevant solutions may affect the accuracy of the relevant solution (cf. the discussion in Section 8.2.1). When there is more than one real solution, we take the one that minimizes the norm

of the residual (just as for the one-dimensional Galerkin method). However, in most numerical experiments of Section 8.4, the resultant has a unique real solution, making it unnecessary to choose. Moreover, the one-dimensional minimum residual method often gives (much) better results than the one-dimensional Galerkin method, see Section 8.4.

8.2.4 Two-dimensional minimum residual

Another idea is to minimize the norm of the generalized residual (8.2.4) with respect to μ, ν:

$$\operatorname*{argmin}_{(\mu,\nu)\in\mathbb{C}^2} \|(\mu A + \nu B + C)u\|. \tag{8.2.10}$$

To solve this, consider the corresponding overdetermined $n \times 2$ linear system

$$Z \begin{bmatrix} \mu \\ \nu \end{bmatrix} = -Cu,$$

with Z as in (8.2.5). By assumption Au and Bu are independent, so μ and ν are uniquely determined by

$$\begin{bmatrix} \mu \\ \nu \end{bmatrix} := -Z^+Cu = -\left(Z^*Z\right)^{-1} Z^*Cu,$$

where Z^+ denotes the pseudoinverse of Z. We see that (8.2.10) is a special case of (8.2.5), namely the case where we choose $p = Au$ and $q = Bu$, so $W = Z$.

As in Section 8.2.2, we can form three possible approximations to the eigenvalue from the computed μ and ν: μ/ν, ν, and $\operatorname{argmin} \|(\theta^2, \theta) - (\mu, \nu)\|^2$. From Section 8.2.2, it follows that the approximations derived by the two-dimensional methods depend on the plane of projection. The plane of the two-dimensional Galerkin method is contained in $\operatorname{span}\{Au, Bu, Cu\}$ (see Section 8.2.2), while the plane for the two-dimensional minimum residual method is $\operatorname{span}\{Au, Bu\}$. Since $\operatorname{span}\{Ax, Bx, Cx\} = \operatorname{span}\{Ax, Bx\}$, we conclude that when $u = x$, both two-dimensional methods yield the same approximations. As a consequence, the sensitivity of the approximations is also the same for both two-dimensional methods.

The two-dimensional methods are summarized in Algorithm 8.2.1.

The following remark explains why we assumed in Section 8.1 that both of the pairs Ax and Bx, and Au and Bu are independent.

Remark 8.2.3 When Au and Bu are dependent, then the one-dimensional minimum residual approach reduces to the one-dimensional Galerkin approach, while the two-dimensional methods are not uniquely determined. When Ax and Bx are dependent, then, though the approaches may be uniquely determined, the results may be bad. (In this case $\sigma_{\min}([\, Ax \;\; Bx \,]) = 0$, so μ and ν are "infinitely sensitive", see Section 8.2.2.) For example, the matrix Z in the two-dimensional methods is ill-conditioned if u is a good approximation to x. Of course, for Au and Bu to be independent, we must have $n \geq 2$. ⊘

Input: an approximate vector u
Output: three approximate eigenvalues

1. Choose a plane of projection spanned by p and q:
   (a) the two "largest" left singular vectors of $[\,Au\ Bu\ Cu\,]$ **(Galerkin)**,
   (b) or $p = Au$ and $q = Bu$ **(minimum residual)**
2. Compute $(\mu, \nu) = -(W^*Z)^{-1}W^*Cu$, where $W = [\,p\ q\,]$ and $Z = [\,Au\ Bu\,]$
3. Approximate λ by one of the following:
   (a) μ/ν
   (b) ν
   (c) $\operatorname{argmin}_\theta \|(\theta^2, \theta) - (\mu, \nu)\|^2$

ALGORITHM 8.2.1: The two-dimensional Galerkin and two-dimensional minimum residual method

8.3 Extensions

8.3.1 Approximations from subspaces

We can also use the techniques described in Section 8.2 for approximations to eigenpairs from subspaces of dimension larger than one. Let $\mathcal{U}$ be a k-dimensional subspace, where for subspace methods one typically has $k \ll n$, and let the columns of U form a basis for $\mathcal{U}$. The *Ritz–Galerkin condition*

$$\theta^2 Au + \theta Bu + Cu \perp \mathcal{U}, \qquad u \in \mathcal{U},$$

leads, with the substitution $u = Us$, to the projected quadratic eigenvalue problem

$$(\theta^2 U^*AU + \theta U^*BU + U^*CU)s = 0, \tag{8.3.1}$$

which in general yields $2k$ Ritz pairs (θ, u). For a specific pair, one can, as a first step, "refine" the value θ by the methods of Section 8.2. Although it is not guaranteed that the new $\widetilde{\theta}$ is better, it seems to be often the case, see the numerical experiments. Moreover, we can monitor the *backward error*.

Definition 8.3.1 (cf. [84]) The backward error of an approximate eigenpair (θ, u) of Q is defined as

$$\begin{aligned}\eta(\theta, u) := \min\{\varepsilon :\ & (\theta^2(A + \Delta A) + \theta(B + \Delta B) + (C + \Delta C))u = 0, \\ & \|\Delta A\| \le \varepsilon\zeta_1,\ \|\Delta B\| \le \varepsilon\zeta_2,\ \|\Delta C\| \le \varepsilon\zeta_3\ \}.\end{aligned}$$

The backward error of an approximate eigenvalue θ of Q is defined as

$$\eta(\theta) := \min_{\|u\|=1} \eta(\theta, u).$$

⊘

In [84, Theorems 1 and 2], the following results are proved:

$$\eta(\theta, u) = \frac{\|r\|}{\zeta_1|\theta|^2 + \zeta_2 \cdot |\theta| + \zeta_3}, \qquad \eta(\theta) = \frac{\sigma_{\min}(Q(\theta))}{\zeta_1|\theta|^2 + \zeta_2 \cdot |\theta| + \zeta_3}. \tag{8.3.2}$$

In the numerical experiments we therefore examine the quality of the computed θ by examining $\|r\|$ and $\sigma_{\min}(Q(\theta))$, which, for convenience, are also called backward errors. Note that the backward errors are related: $\sigma_{\min}(Q(\theta)) \leq \|r\|$.

Then, as a second step after refining the θ, one can "refine" the vector u by taking $\widetilde{u} = U\widetilde{s}$, where

$$\widetilde{s} = \text{the "smallest" right singular vector of } \quad \widetilde{\theta}^2 AU + \widetilde{\theta} BU + CU$$

(For the Arnoldi method for the standard eigenvalue problem, a similar refinement of a Ritz vector has been proposed in [43].) This step is relatively cheap, because all matrices are "skinny". Given $\widetilde{\theta}$, the vector $\widetilde{u}$ minimizes the backward error $\eta(\widetilde{\theta}, u)$, see (8.3.2). It is also possible to repeat these two steps to get better and better approximations, leading to Algorithm 8.3.2.

Input: a search space $\mathcal{U}$
Output: an approximate eigenpair (θ, u) with $u \in \mathcal{U}$

1. Compute an approximate eigenpair (θ, u) with the standard Ritz–Galerkin method
   for $k = 1, 2, \ldots$
2. Compute a new θ_k choosing one of the methods of Section 8.2
3. Compute the "smallest" singular vector s_k of $\theta_k^2 AU + \theta_k BU + CU$
4. $u_k = U s_k$

ALGORITHM 8.3.2: Refinement of an approximate eigenpair for the quadratic eigenproblem

During this algorithm, we do not know the (forward) error $|\theta_k - \lambda|$, but the backward errors $\|r\|$ and $\sigma_{\min}(\theta_k^2 AU + \theta_k BU + CU)$ are cheaply available; they can be used to decide whether or not to continue the algorithm. When we take the one-dimensional minimum residual method in each step, we are certain that the backward error $\|r\|$ decreases monotonically. In Experiment 8.4.3 we use the two-dimensional Galerkin approach in every step.

Remark 8.3.2 For the symmetric eigenvalue problem, the possibility of an iterative procedure to minimize $\|Au - \rho(u)u\|$ over the subspace $\mathcal{U}$ is mentioned in [74], in the context of finding inclusion intervals for eigenvalues. Moreover, a relation between the minimization of $\|Au - \rho(u)u\|$ and the smallest possible Lehmann interval is given. ⊘

8.3.2 The polynomial eigenvalue problem

Consider the polynomial eigenvalue problem

$$(\lambda^l A_l + \lambda^{l-1} A_{l-1} + \cdots + \lambda A_1 + A_0)x = 0,$$

where the size n of the matrices satisfies $n \geq l$. Define the *generalized residual* as

$$r(\mu_1, \ldots, \mu_l, u) := (\mu_l A_l + \mu_{l-1} A_{l-1} + \cdots + \mu_1 A_1 + A_0)u.$$

Both the l-dimensional Galerkin method

$$r(\mu_1, \ldots, \mu_l, u) \perp \{p_1, \ldots, p_l\}$$

and the l-dimensional minimum residual method

$$\min_{\mu_1, \ldots, \mu_l} \|r(\mu_1, \ldots, \mu_l, u)\|$$

lead to a system of the form

$$W^* Z \begin{bmatrix} \mu_l \\ \vdots \\ \mu_1 \end{bmatrix} = -W^* A_0 u, \tag{8.3.3}$$

where $Z = [A_l u \ \cdots \ A_1 u]$. For the l-dimensional minimum residual method we have $W = Z$; for the l-dimensional Galerkin approach with "least-squares" l-dimensional plane, W consists of the l largest left singular vectors of $[\, Z \ A_0 u \,]$. Assuming that the vectors $A_1 u, \ldots, A_l u$ are independent, (8.3.3) has a unique solution. In principle we can try every quotient μ_l/μ_{l-1}, μ_{l-1}/μ_{l-2}, $\ldots$, $\mu_2/\mu_1, \mu_1$, and also some other combinations like $\mu_l/(\mu_{l-2}\mu_1)$, as an approximation to λ. When λ is small, μ_1 will probably be the best. In principle, an "argmin" solution is also possible, although the degree of the associated polynomials will get larger quickly.

The one-dimensional minimum residual approach may be less attractive for the polynomial eigenvalue problem, as the degree of the associated polynomials (cf. (8.2.9) and (8.4.1)) increases fast. This results in more irrelevant solutions, while the relevant solution will likely to be more sensitive to perturbations in the (approximate) eigenvector.

8.4 Numerical experiments

The experiments are carried out in MATLAB and MAPLE. First a word on solving (8.2.8) and (8.2.6) for the one-dimensional minimum residual approach, and the "argmin" solution of the two-dimensional Galerkin and minimum residual approach, respectively. Write $\theta = \theta_1 + i\theta_2$, $\mu = \mu_1 + i\mu_2$, and $\nu = \nu_1 + i\nu_2$. Differentiating (8.2.8) with respect to θ_1 and θ_2 leads to two mixed equations (in θ_1 and θ_2) of degree three. With MAPLE the equations are manipulated so that we have two equations of degree nine in θ_1 or θ_2 only, which are called the resultants. When we know that λ is real, then we get the cubic equation (8.2.9).

Differentiation of (8.2.6) with respect to θ_1 and θ_2, leads to

$$\begin{aligned} \theta_1^3 + (\theta_2^2 - \mu_1 + \tfrac{1}{2})\theta_1 - \mu_2\theta_2 - \tfrac{1}{2}\nu_1 &= 0, \\ \theta_2^3 + (\theta_1^2 + \mu_1 + \tfrac{1}{2})\theta_2 - \mu_2\theta_1 - \tfrac{1}{2}\nu_2 &= 0. \end{aligned} \tag{8.4.1}$$

Because of the missing θ_1^2 and θ_2^2 terms, in the first and second equation respectively, the corresponding resultants have degree only five. All equations were solved numerically by a MAPLE command of the form

`solve(resultant(` *equation*$_1$ `(x, y),` *equation*$_2$ `(x, y), y), x).`

Of course, we only have to solve one resultant, say for $\mathrm{Re}(\theta)$, then $\mathrm{Im}(\theta)$ can be solved from a cubic equation. In our experiments, many equations have a unique real solution, making it unnecessary to choose. When there is more than one real solution, we take the one that minimizes the norm of the residual.

Experiment 8.4.1 Our first example is taken from [86, p. 250]:

$$A = \begin{bmatrix} 0 & 6 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & -6 & 0 \\ 2 & -7 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad C = I_3.$$

This problem has two eigenvectors for each of which there exist two eigenvalues: $[\,1\ 1\ 0\,]^T$ corresponds to $\lambda = 1/2$ and $\lambda = 1/3$, while $[\,0\ 0\ 1\,]^T$ corresponds to $\lambda = \pm i$. In line with our assumptions, we do not consider these. Instead, we focus on the other eigenpairs $(\lambda, x) = \big(1, [\,0\ 1\ 0\,]^T\big)$ and $(\lambda, x) = \big(\infty, [\,1\ 0\ 0\,]^T\big)$. For the last pair we consider the problem for $\lambda^{-1} = 0$. We simulate the situation of having a good approximation $u \approx x$ by adding a random (complex) perturbation to x:

$$u := (x + \varepsilon \cdot w) \,/\, \|x + \varepsilon \cdot w\|, \tag{8.4.2}$$

where w is a normalized vector of the form `rand(3,1)` $+\, i \cdot$ `rand(3,1)`. (For all experiments, we take "`seed=0`" so that our results are reproducible.) Table 8.2 gives the results of the four approaches for $\varepsilon = 0.01$. The first row of the two-dimensional Galerkin (Gal-2) and two-dimensional minimum residual (MR-2) approaches represents μ/ν, the second row gives ν, while the third row indicates the "argmin" solution as approximate eigenvalue. For clarity, the meaning of the different rows is first summarized in Table 8.1.

TABLE 8.1: The rows of Tables 8.2 to 8.4, with their meaning.

row nr.	label	meaning
1	Gal-1	best approximation (of the two) of the one-dimensional Galerkin method
2	Gal-2	μ/ν approximation of the two-dimensional Galerkin method
3		ν approximation of the two-dimensional Galerkin method
4		"argmin" approximation of the two-dimensional Galerkin method
5	MR-1	best approximation of the one-dimensional minimum residual method
6	MR-2	μ/ν approximation of the one-dimensional minimum residual method
7		ν approximation of the one-dimensional minimum residual method
8		"argmin" approximation of the two-dimensional minimum residual method

For $\lambda = 1$, all other approaches (Gal-2, MR-1, and MR-2) give a smaller (forward) error than the classical one-dimensional Galerkin method (Gal-1). The "ν" approximation of the two-dimensional approaches Gal-2 (row 3) and MR-2 (row 7) is particularly

TABLE 8.2: The one-dimensional Galerkin (Gal-1), two-dimensional Galerkin (Gal-2: μ/ν, ν, and "argmin"), one-dimensional minimum residual (MR-1), and two-dimensional minimum residual (MR-2: μ/ν, ν, and "argmin") approaches for $\lambda = 1$ (columns 2 to 4) and $\lambda^{-1} = 0$ (columns 5 to 7). The columns give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	error	$\|r\|$	$\sigma_{\min}$	error	$\|r\|$	$\sigma_{\min}$
Gal-1	0.00202	0.0112	0.00143	0.03958	0.0399	0.0280
Gal-2	0.00168	0.0112	0.00119	1.00009	2.8285	0.0016
	0.00004	0.0181	0.00003	0.01987	0.0206	0.0141
	0.00067	0.0143	0.00047	0.01988	0.0206	0.0141
MR-1	0.00182	0.0111	0.00128	0.02384	0.0186	0.0168
MR-2	0.00169	0.0112	0.00119	1.00006	2.8284	0.0016
	0.00013	0.0178	0.00009	0.01987	0.0206	0.0141
	0.00070	0.0142	0.00050	0.01988	0.0206	0.0141

good. For the sensitivity of the ν-solution for the two-dimensional approaches, the modest value $\kappa([\, Ax \;\; Bx \,]) \approx 26$ already indicates this. Of all approximations, the MR-1 solution has the smallest backward error $\|r\|$, as expected, but not the smallest forward error. For the discriminant (8.2.3) we have $\delta = 25$.

For $\lambda^{-1} = 0$, the "μ/ν" approximations (rows 2 and 6) are bad, which was already predicted by (8.2.7); $\kappa([\, Ax \;\; Bx \,]) \approx 2.6$ is again modest, and for the discriminant we have $\delta = 1$. ⊘

Experiment 8.4.2 For the second example we construct matrices such that the discriminant δ is small and hence the zeros of (8.2.1) are close. For small $\zeta > 0$ define

$$A = I_3, \qquad B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -2 & 2 \\ 0 & 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & -1-\sqrt{\zeta} & 0 \\ 0 & 1-\zeta & 2 \\ 0 & 0 & 1 \end{bmatrix}.$$

One may check that $x = [\, 0 \;\; 1 \;\; 0 \,]^T$ is an eigenvector with corresponding eigenvalue $1+\sqrt{\zeta}$. (The second solution $1 - \sqrt{\zeta}$ to (8.2.1) is close to the eigenvalue, but has no meaning.) The discriminant is equal to 4ζ. We take $\zeta = 10^{-4}$, so $\lambda = 1.01$. We test the approaches for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$, see Table 8.3.

For the sensitivity of μ and ν of the two-dimensional methods Gal-2 and MR-2 we note that $\kappa([\, Ax \;\; Bx \,]) \approx 5.8$. Because the discriminant $\delta = 4 \cdot 10^{-4}$ is small, and the sensitivity is modest, it is no surprise that all other approximations are much better (measured in forward or backward error) than Gal-1. ⊘

Experiment 8.4.3 For the following example we take A, B, and C random symmetric matrices of size 100×100. We try to approximate the eigenvalue $\lambda \approx 7.2288 + 2.7803i$, for $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$ in (8.4.2), see Table 8.4.

For the sensitivity for Gal-2 and MR-2 we have $\kappa([\, Ax \;\; Bx \,]) \approx 63$; $|\delta| \approx 2.4 \cdot 10^{-5}$. We see that the two "$\mu/\nu$" approximations (row 2 and 6) are the best, together with the MR-1 solution (row 5). Note that for larger matrices, the computation of $\sigma_{\min}(Q(\theta))$ is expensive. In practice, one does not compute it, but it is shown here to compare the methods. ⊘

TABLE 8.3: The one-dimensional Galerkin (Gal-1), two-dimensional Galerkin (Gal-2: μ/ν, ν, and "argmin"), one-dimensional minimum residual (MR-1), and two-dimensional minimum residual (MR-2: μ/ν, ν, and "argmin") approaches for $\lambda = 1.01$, for $\varepsilon = 10^{-2}$ (columns 2 to 4) and $\varepsilon = 10^{-3}$ (columns 5 to 7), respectively. The columns give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	error	$\|r\|$	$\sigma_{\min}$	error	$\|r\|$	$\sigma_{\min}$
Gal-1	0.1459	0.1344	0.01327	0.0445	0.0431	0.001326
Gal-2	0.0346	0.0349	0.00052	0.0035	0.0035	0.000037
	0.0151	0.0274	0.00019	0.0015	0.0028	0.000018
	0.0225	0.0286	0.00027	0.0022	0.0029	0.000026
MR-1	0.0159	0.0274	0.00020	0.0015	0.0027	0.000018
MR-2	0.0348	0.0351	0.00053	0.0035	0.0035	0.000037
	0.0152	0.0274	0.00019	0.0015	0.0028	0.000018
	0.0226	0.0287	0.00027	0.0023	0.0029	0.000026

TABLE 8.4: The approximations of the one-dimensional Galerkin (Gal-1), two-dimensional Galerkin (Gal-2: μ/ν, ν, and "argmin"), one-dimensional minimum residual (MR-1), and two-dimensional minimum residual (MR-2: μ/ν, ν, and "argmin") approaches for $\lambda \approx 7.2288 + 2.7803i$, and $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$, respectively. The other columns give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	appr. ($\varepsilon = 10^{-3}$)	error	$\|r\|$	$\sigma_{\min}$	appr. ($\varepsilon = 10^{-4}$)	error	$\|r\|$	$\sigma_{\min}$
Gal-1	6.86+2.71i	0.37	2.89	0.186	7.218+2.739i	0.0428	0.308	0.0221
Gal-2	7.26+2.68i	0.10	2.90	0.054	7.231+2.769i	0.0110	0.290	0.0057
	6.87+3.04i	0.44	3.16	0.234	7.189+2.801i	0.0446	0.330	0.0232
	7.07+2.86i	0.18	2.93	0.096	7.210+2.785i	0.0195	0.300	0.0101
MR-1	7.04+2.61i	0.24	2.81	0.123	7.227+2.769i	0.0107	0.290	0.0055
MR-2	7.23+2.65i	0.13	2.88	0.064	7.231+2.769i	0.0112	0.290	0.0058
	3.67+1.63i	3.74	6.34	0.709	7.123+2.775i	0.1057	0.437	0.0545
	5.14+2.08i	2.20	5.23	0.822	7.177+2.772i	0.0529	0.332	0.0247

Experiment 8.4.4 Next, we test Algorithm 8.3.2. We start with a three-dimensional subspace $\mathcal{U}$, consisting of the same vector as in the previous experiment ($\varepsilon = 10^{-3}$), completed by two random (independent) vectors. We determine six Ritz pairs according to (8.3.1), and refine the one with θ approximating the eigenvalue $\lambda \approx 7.2288 + 2.7803i$ by Algorithm 8.3.2, where in every step we choose the μ/ν-approximation of the two-dimensional Galerkin method. The results, shown in Table 8.5, reveal that both u and θ are improved four times, after which they keep fixed in the decimals shown. Note that the smallest possible angle of a vector in $\mathcal{U}$ with x is

$$\angle(\mathcal{U}, x) = \angle(x_U, x) \approx 6.2809 \cdot 10^{-4},$$

where $x_U = UU^*x/\|UU^*x\|$ is the eigenvector projected onto $\mathcal{U}$.

TABLE 8.5: Refinement of an approximate eigenvalue by Algorithm 8.3.2 for $\lambda \approx 7.2288+2.7803i$. The columns give the iteration number, angle between u and x, (forward) error $|\theta - \lambda|$, and $\|r\|$, $\tau_{\min} := \sigma_{\min}(\theta^2 AU + \theta BU + CU)$ and $\sigma_{\min}(Q(\theta)) = \sigma_{\min}(\theta^2 A + \theta B + C)$ for the backward errors.

iteration	$\angle(u,x)$	θ	error	$\|r\|$	$\tau_{\min}$	$\sigma_{\min}$
0	$7.192\,e-4$	$7.222+2.778i$	$6.112\,e-3$	$1.234\,e-1$	$1.166\,e-3$	$3.178\,e-3$
1	$6.542\,e-4$	$7.231+2.783i$	$4.113\,e-3$	$1.196\,e-1$	$1.137\,e-3$	$2.142\,e-3$
2	$6.529\,e-4$	$7.231+2.781i$	$2.627\,e-3$	$1.137\,e-1$	$1.137\,e-3$	$1.368\,e-3$
3	$6.528\,e-4$	$7.231+2.781i$	$2.597\,e-3$	$1.137\,e-1$	$1.137\,e-3$	$1.352\,e-3$
≥ 4	$6.528\,e-4$	$7.231+2.781i$	$2.596\,e-3$	$1.137\,e-1$	$1.137\,e-3$	$1.351\,e-3$

We see that in particular the first step of the algorithm considerably improves the approximate eigenpair. After four steps, the angle of the refined approximate eigenvector with the optimal vector in $\mathcal{U}$ is less than 30% of the angle that the Ritz vector makes with the optimal vector. The error in θ is more than halved. Note again that $\sigma_{\min}(\theta^2 A + \theta B + C)$ is expensive, but $\tau_{\min} := \sigma_{\min}(\theta^2 AU + \theta BU + CU)$ is readily available in the algorithm. ⊘

Experiment 8.4.5 Finally, we test the ideas of Section 8.3.2. Consider the polynomial eigenvalue problem of degree four

$$(\lambda^4 A_4 + \lambda^3 A_3 + \lambda^2 A_2 + \lambda A_1 + A_0)x = 0,$$

where the A_i are random 5×5 matrices. We try to approximate the eigenpair corresponding to $\lambda \approx -2.2009 - 1.5366i$ and take $\varepsilon = 10^{-4}$ in (8.4.2). The μ_i in the generalized residual

$$r(\mu_1, \mu_2, \mu_3, \mu_4, u) := (\mu_4 A_4 + \mu_3 A_3 + \mu_2 A_2 + \mu_1 A_1 + A_0)u$$

are determined by the 4-dimensional Galerkin method with "least-squares" plane. The results of the μ_4/μ_3, μ_3/μ_2, μ_2/μ_1, and μ_1 approximations are summarized in Table 8.6.

Note that both the μ_4/μ_3 and μ_3/μ_2 approximations give better results than the standard approach. We mention that the results of the 4-dimensional minimum residual method were roughly the same. ⊘

TABLE 8.6: Approximations of the one-dimensional Galerkin (Gal-1) and 4-dimensional Galerkin (μ_4/μ_3, μ_3/μ_2, μ_2/μ_1, and μ_1) approaches for $\lambda \approx -2.2009 - 1.5366i$. The other columns give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	approximation	error	$\|r\|$	$\sigma_{\min}$
Gal-1	-2.1891-1.5404i	0.0123	0.0894	0.0388
μ_4/μ_3	-2.2010-1.5370i	0.0004	0.0171	0.0014
μ_3/μ_2	-2.2057-1.5318i	0.0067	0.0422	0.0215
μ_2/μ_1	-2.2054-1.5238i	0.0135	0.0863	0.0429
μ_1	-2.1693-1.5346i	0.0317	0.2121	0.0972

8.5 Conclusions

The usual one-dimensional Galerkin approach for the determination of an approximate eigenvalue corresponding to an approximate eigenvector may give inaccurate results, especially when the discriminant of equation (8.2.1) is small. We have proposed several alternative ways that often give better results with little extra effort (all methods require three matrix-vector multiplications Au, Bu, and Cu, and additionally $\mathcal{O}(n)$ time). Based on our analysis and the numerical experiments, we recommend in particular the μ/ν and ν approximations of the two-dimensional approaches Gal-2 and MR-2, because they are cheap to compute and give good results. For small eigenvalues, one should take the "ν" approximations. The MR-1 method ensures a minimal residual (backward error).

The approaches are also useful for approximations from a subspace and for polynomial eigenvalue problems of higher degree.

Chapter 9

Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method

Abstract. We study the Lanczos method for computing extreme eigenvalues of a symmetric or Hermitian matrix. It is not guaranteed that the extreme Ritz values are close to the extreme eigenvalues—even when the norms of the corresponding residual vectors are small. Assuming that the starting vector has been chosen randomly, we compute probabilistic bounds for the extreme eigenvalues from data available during the execution of the Lanczos process. Four different types of bounds are obtained using Lanczos, Ritz, and Chebyshev polynomials. These bounds are compared theoretically and numerically. Furthermore we show how one can determine, after each Lanczos step, a probabilistic upper bound for the number of steps still needed (without performing these steps) to obtain an approximation to the largest or smallest eigenvalue within a prescribed tolerance.

Key words: symmetric matrix, Hermitian matrix, Lanczos, Ritz values, misconvergence, Lanczos polynomial, Ritz polynomial, Chebyshev polynomial.

AMS subject classification: 65F15, 65F50.

9.1 Introduction

Knowledge about the extreme eigenvalues of symmetric or Hermitian matrices is important in many applications. For example, the stability of processes involving such matrices is often governed by the location of their eigenvalues. The extreme eigenvalues can also be used to determine condition numbers, the field of values, and ε-pseudospectra of arbitrary matrices (see, e.g., [12, 87]). For small-sized matrices the eigenvalues can be computed by the QR-method (see, e.g., [31]), but this is not feasible for large matrices. A method which is often used in practice to compute a few extreme eigenvalues of large sparse symmetric or Hermitian matrices is the Lanczos method (see, e.g., [31, 61, 101]).

*Based on joint work with Jos L. M. van Dorsselaer and Henk A. van der Vorst, see Section 1.5.

The approximations of the eigenvalues obtained with the Lanczos method (the Ritz values) lie between the smallest and largest eigenvalue of the original matrix and one would like to know whether the largest (or smallest) Ritz value is sufficiently close to the largest (or smallest) eigenvalue of that matrix.

The classical a priori error estimates for the Lanczos method, established by Kaniel, Paige, and Saad (see, e.g., [31, 44, 57, 61, 68]) are not applicable in practice to obtain bounds on the spectrum of Hermitian matrices, because they involve knowledge about the eigenvalues and angles between the eigenvectors and the starting vector. Furthermore one should note that small residuals for the Ritz values *only* imply that these Ritz values are close to an eigenvalue, but it is not guaranteed that this eigenvalue is indeed the one we are looking for (cf., e.g., [62]). In fact, it is not possible to derive rigorous bounds on the spectrum from *any* possible starting vector: if the starting vector is perpendicular to the eigenvector (or eigenspace in case of multiple eigenvalues) corresponding to the largest or smallest eigenvalue, it is impossible to obtain any information regarding this eigenvalue from the Lanczos process.

In this chapter we derive various a posteriori bounds for the spectrum of real symmetric matrices using a probabilistic approach. Assuming that the starting vector of the Lanczos process is chosen randomly from the uniform distribution over the unit sphere, we derive, using data available while executing the Lanczos process, for every $\varepsilon \in (0,1)$ bounds for the spectrum with probability at least $1-\varepsilon$. No intrinsic properties of the matrix (apart from being symmetric) are required to compute our bounds. Polynomials related to the Lanczos process, namely the Lanczos polynomials and Ritz polynomials, are used to derive two types of such bounds. For symmetric positive definite matrices Kuczyński and Woźniakowski [49, Theorem 3] give, for arbitrary $t > 1$, an a priori upper bound for the probability that the largest eigenvalue is greater than t times the largest Ritz value; Chebyshev polynomials of the second kind are used to obtain these bounds. This result can be used to compute a posteriori probabilistic bounds for the spectrum while executing the Lanczos process, and bounds based on [49, Theorem 3] can be used for symmetric indefinite matrices as well. The fourth kind of bounds for the spectrum is obtained with Chebyshev polynomials of the first kind. The sharpness of the different bounds is analyzed theoretically and compared numerically. It turns out that the bounds based on Lanczos polynomials are the sharpest ones in most cases; however, the Ritz polynomials sometimes provide better bounds when the Lanczos method suffers from a misconvergence (i.e., the largest (or smallest) Ritz values in consecutive Lanczos steps seem to converge, but not to an extreme eigenvalue).

Apart from the bounds on the spectrum, we also study probabilistic bounds for the number of Lanczos steps needed to get an error (or relative error) in the largest or smallest eigenvalue that is smaller than a given tolerance. In [48, Theorem 4.2] the authors present a probabilistic upper bound for the number of Lanczos steps needed to yield a relative error in the largest eigenvalue of a symmetric positive definite matrix that is smaller than a given tolerance. For this special case numerical experiments demonstrate that our bound and the one from [48, Theorem 4.2] are almost the same. Furthermore, we provide upper bounds for the number of Lanczos steps needed to guarantee with probability at least $1-\varepsilon$ that either the spectrum lies between certain prescribed bounds, or that a misconvergence has occurred.

The results in this chapter deal with the Lanczos process applied to real symmetric matrices and real starting vectors. This includes the case of Hermitian matrices, because the Lanczos method applied to a complex Hermitian matrix (with a complex starting vector) can be written as the application of the Lanczos method to a related real symmetric matrix of double size with a real starting vector (see Remark 9.2.1 for details). In Remark 9.2.2, we discuss an application to Lanczos bidiagonalization.

All bounds discussed in this chapter are easily implemented and can be computed with little effort while executing the Lanczos process.

The chapter has been organized as follows. In Section 9.2 some notations and definitions are introduced. Bounds based on Lanczos polynomials are presented in Section 9.3, and bounds obtained with Ritz polynomials can be found in Section 9.4. In Section 9.5 we derive bounds from Chebyshev polynomials. The estimates for the number of Lanczos steps still to be done for sufficiently accurate approximations can be found in Section 9.6.1, and the estimates for the number of Lanczos steps needed to obtain prescribed bounds for the spectrum or to detect misconvergence are given in Section 9.6.2. Numerical experiments are presented in Section 9.7, and the conclusions can be found in Section 9.8.

9.2 Preliminaries

In this section we introduce some notations and present relevant properties of the Lanczos method. For an introduction to the Lanczos method and more details, as well as implementation issues, the reader may consult, e.g., [31, 61]. Throughout this chapter we do not consider the effect of rounding errors.

The standard inner product on $\mathbb{R}^n$ will be denoted by $(\cdot,\cdot)$, and $\|\cdot\|$ stands for the Euclidean norm, and I is the $n \times n$ identity matrix.

Let A be a real symmetric $n \times n$ matrix with eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

The corresponding normalized eigenvectors x_j form an orthonormal basis of $\mathbb{R}^n$. We use the Lanczos method to approximate one or a few extreme eigenvalues of A. The unit starting vector is denoted by v_1 and can be written as

$$v_1 = \sum_{j=1}^{n} \gamma_j x_j. \tag{9.2.1}$$

If v_1 is chosen randomly from the uniform distribution with respect to the unit sphere, the dimension of the Krylov subspace

$$K_k(A, v_1) = \operatorname{span}\{v_1, Av_1, \ldots, A^{k-1}v_1\}$$

is equal to k with probability one for k less than the number of distinct eigenvalues of A.

In the Lanczos process vectors v_k are generated by the three-term recurrence

$$\delta_k v_{k+1} = Av_k - \alpha_k v_k - \beta_{k-1} v_{k-1} \qquad \text{for} \quad k = 1, 2, 3, \ldots, \tag{9.2.2}$$

where $v_0 = 0$, $\beta_0 = 1$, $\alpha_k = (Av_k, v_k)$, $\beta_{k-1} = (Av_k, v_{k-1})$, and $\delta_k > 0$ is chosen such that $\|v_{k+1}\| = 1$. With this choice one has $\delta_k = \beta_k$ for $k \geq 1$. The vectors $v_1, v_2, \ldots, v_k$ form an orthonormal basis of the Krylov subspace $K_k(A, v_1)$. Let V_k be the $n \times k$ matrix of which v_j is the jth column. The *Ritz values* occurring in step k of the Lanczos process are the eigenvalues of the tridiagonal $k \times k$ matrix $T_k = V_k^{\mathrm{T}} A V_k$, and are denoted by

$$\theta_1^{(k)} < \theta_2^{(k)} < \cdots < \theta_k^{(k)};$$

the Ritz values satisfy $\theta_j^{(k)} > \lambda_j$ and $\theta_{k+1-j}^{(k)} < \lambda_{n+1-j}$ $(1 \leq j \leq k)$. We denote the eigenvectors of T_k by $s_j^{(k)}$: $T_k s_j^{(k)} = \theta_j^{(k)} s_j^{(k)}$ and the *Ritz vectors* by $y_j^{(k)} = V_k s_j^{(k)}$, where we assume that these Ritz vectors are normalized. We also introduce the *residuals*

$$r_j^{(k)} = A y_j^{(k)} - \theta_j^{(k)} y_j^{(k)}.$$

Related to the three-term recursion (9.2.2) are the polynomials p_k of degree k defined by $p_{-1}(t) = 0$, $p_0(t) = 1$, and

$$\beta_k p_k(t) = (t - \alpha_k) p_{k-1}(t) - \beta_{k-1} p_{k-2}(t) \qquad \text{for} \quad k = 1, 2, 3, \ldots. \tag{9.2.3}$$

From (9.2.2) with $\delta_k = \beta_k$ and (9.2.3) it follows that

$$v_{k+1} = p_k(A) v_1 \qquad \text{for} \quad k = 1, 2, 3, \ldots.$$

The polynomials p_k are called the *Lanczos polynomials* with respect to A and v_1. Other polynomials related to the Lanczos method are the *Ritz polynomials* $q_j^{(k)}$ of degree $k-1$, which are characterized by the fact that

$$y_j^{(k)} = q_j^{(k)}(A) v_1 \qquad \text{for} \quad j = 1, 2, \ldots, k. \tag{9.2.4}$$

In the following sections estimates for the eigenvalues of A, based on Lanczos and Ritz polynomials, will be studied and compared. Therefore it is important to understand the relation between these polynomials. The polynomial p_k is a scalar multiple of the characteristic polynomial of the matrix T_k (cf., e.g., [61, Section 7.3]), which implies that $\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_k^{(k)}$ are the zeros of p_k. From [61, Section 12.3] it follows that these Ritz values without $\theta_j^{(k)}$ are the zeros of $q_j^{(k)}$. Hence $p_k(t) = c_j^{(k)} (t - \theta_j^{(k)}) q_j^{(k)}(t)$ for a certain constant $c_j^{(k)}$. (From this relation it follows that $q_j^{(k)}$ is a scalar multiple of $\prod_{i \neq j} (t - \theta_i^{(k)})$ and that polynomial is called a reduced Ritz polynomial in [90]. The relation between these polynomials and (9.2.4) also follows from [90, Formula (5.14)].) Because $v_{k+1} = p_k(A) v_1 = c_j^{(k)} (A - \theta_j^{(k)} I) q_j^{(k)}(A) v_1 = c_j^{(k)} r_j^{(k)}$, we have $c_j^{(k)} = 1/\|r_j^{(k)}\|$, which yields the following relation between the Lanczos and Ritz polynomials:

$$p_k(t) = (t - \theta_j^{(k)}) q_j^{(k)}(t) \,/\, \|r_j^{(k)}\| \qquad \text{for} \quad j = 1, 2, \ldots, k. \tag{9.2.5}$$

Remark 9.2.1 The Lanczos method described above can also be used to determine a few extreme eigenvalues of a complex Hermitian matrix A. The results in this chapter are only valid for real symmetric matrices, but the Lanczos method for Hermitian matrices

can be formulated in terms of real matrices and vectors. Let $\mathrm{Re}(A)$ and $\mathrm{Im}(A)$ be the real and imaginary part of A, respectively. The Lanczos method applied to the $2n \times 2n$ real symmetric matrix

$$B = \begin{bmatrix} \mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A) \end{bmatrix}$$

with starting vector $\begin{bmatrix} \mathrm{Re}(v_1) \\ \mathrm{Im}(v_1) \end{bmatrix}$ yields the same tridiagonal matrices T_k as the Lanczos method applied to A with starting vector v_1; this can be seen from taking the real and imaginary part of the three-term recurrence (9.2.2). The numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of B, but with multiplicity twice as large as for the matrix A. Therefore (probabilistic) bounds for the spectrum of B are (probabilistic) bounds for the spectrum of A as well. ⊘

Remark 9.2.2 The methods in this chapter can also be applied to Lanczos bidiagonalization, by applying Lanczos to the augmented matrix (3.1.1) with starting vector $(0, v_1)$. Partly because the treatment of this subject would lead to a clutch of notation in this chapter, this matter is left for future work. ⊘

9.3 Spectral bounds using the Lanczos polynomial

In this section we will give probabilistic upper and lower bounds for the spectrum of A, based on Lanczos polynomials. For each step of the Lanczos process we obtain these bounds based on the information computed so far. No assumptions on the location or separation of the eigenvalues are required.

The Lanczos polynomials p_k are a byproduct of the process. They are usually small between $\theta_1^{(k)}$ and $\theta_k^{(k)}$ and increase rapidly outside this interval. We can exploit this fact: assuming that the starting vector has components in the direction of x_1 and x_n, we can provide upper and lower bounds for the spectrum of A.

From

$$1 = \|v_{k+1}\|^2 = \|p_k(A)v_1\|^2 = \sum_{j=1}^{n} \gamma_j^2 \, p_k(\lambda_j)^2$$

and $p_k(\lambda_n) > 0$ it follows that

$$1 \geq |\gamma_n| \, p_k(\lambda_n).$$

If γ_n is known, this estimate provides an upper bound λ^{up} for λ_n: let λ^{up} be the largest real zero of

$$f_L(t) = p_k(t) - 1/|\gamma_n|. \tag{9.3.1}$$

This number λ^{up} exists and satisfies $\lambda^{\mathrm{up}} > \theta_k^{(k)}$ because p_k is strictly increasing on $(\theta_k^{(k)}, \infty)$. The number λ^{up} can be determined by Newton's method or bisection. As a starting point for the Newton process one can take, for instance, $\|A\|_\infty$, $\|A\|_1$, or a previously computed upper bound for λ_n.

In practice we do not know γ_n, but we can determine the probability that $|\gamma_n|$ is smaller than a given (small) constant. Let S^{n-1} denote the $(n-1)$-dimensional unit

sphere in $\mathbb{R}^n$. We assume that v_1 is chosen randomly with respect to the uniform distribution over S^{n-1}. Then, as a result, $(\gamma_1, \gamma_2, \ldots, \gamma_n)$ is also random with respect to the uniform distribution over S^{n-1} (cf., e.g., [48, p. 1116]). In the following lemma we compute the probability that $|\gamma_n|$ is smaller than δ.

Lemma 9.3.1 *Assume that the starting vector v_1 has been chosen randomly with respect to the uniform distribution over the unit sphere S^{n-1} and let $\delta \in [0,1]$. Then*

$$P(|\gamma_n| \leq \delta) \;=\; 2\,B(\tfrac{n-1}{2}, \tfrac{1}{2})^{-1} \cdot \int_0^{\arcsin\delta} \cos^{n-2} t \,\mathrm{d}t,$$

where B denotes Euler's Beta function*: $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$ and P stands for probability.*

Proof: Define $S_\delta = \{\gamma \in S^{n-1} : |\gamma_n| < \delta\}$; we want to determine the ratio of the areas of the sets S_δ and S^{n-1}. The image of the map

$$\varphi \;:\; (-\pi, \pi) \;\times\; (-\tfrac{\pi}{2}, \tfrac{\pi}{2})^{n-2} \;\to\; S^{n-1}$$

defined by

$$\varphi : \begin{bmatrix} \alpha \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{n-2} \end{bmatrix} \mapsto \begin{bmatrix} \cos\alpha\cos\psi_1\cos\psi_2\cdots\cos\psi_{n-3}\cos\psi_{n-2} \\ \sin\alpha\cos\psi_1\cos\psi_2\cdots\cos\psi_{n-3}\cos\psi_{n-2} \\ \sin\psi_1\cos\psi_2\cdots\cos\psi_{n-3}\cos\psi_{n-2} \\ \vdots \\ \sin\psi_{n-3}\cos\psi_{n-2} \\ \sin\psi_{n-2} \end{bmatrix}$$

equals the sphere up to a negligible set. One can check that the associated Euclidean density is given by

$$\omega(\alpha, \psi_1, \psi_2, \ldots, \psi_{n-2}) \;=\; \cos\psi_1 \cdot \cos^2\psi_2 \cdots \cos^{n-2}\psi_{n-2}.$$

Therefore we can compute the areas of S_δ and S^{n-1} by integrating this density over the respective domains. Taking the ratio of the two results, we get

$$\begin{aligned} P(|\gamma_n| \leq \delta) \;&=\; P(|\psi_{n-2}| \leq \arcsin\delta) \\ &=\; 2\int_0^{\arcsin\delta} \cos^{n-2} t \,\mathrm{d}t \,\Big/ \int_{-\pi/2}^{\pi/2} \cos^{n-2} t \,\mathrm{d}t \\ &=\; 2\int_0^{\arcsin\delta} \cos^{n-2} t \,\mathrm{d}t \,\Big/\, B(\tfrac{n-1}{2}, \tfrac{1}{2}), \end{aligned}$$

which proves the lemma. □

Now suppose we would like to have an upper bound for the spectrum of A that is correct with probability at least $1-\varepsilon$. Then we determine the value of δ for which

$$\int_0^{\arcsin\delta} \cos^{n-2} t \,\mathrm{d}t \;=\; \tfrac{\varepsilon}{2} B(\tfrac{n-1}{2}, \tfrac{1}{2}) \quad \left(= \varepsilon \int_0^{\pi/2} \cos^{n-2} t \,\mathrm{d}t\right) \tag{9.3.2}$$

holds, e.g., by using Newton's method. The integrals in (9.3.2) can be computed using an appropriate quadrature formula. We replace $|\gamma_n|$ in (9.3.1) by the value δ computed from (9.3.2) and determine the zero $\lambda^{\text{up}} > \theta_k^{(k)}$. This λ^{up} is an upper bound for the spectrum of A with probability at least $1-\varepsilon$, and we call λ^{up} a probabilistic upper bound.

A lower bound λ^{low} for the spectrum of A with probability at least $1-\varepsilon$ can be obtained in a similar way. (Note that Lemma 9.3.1 remains valid if $|\gamma_n|$ is replaced by $|\gamma_1|$.) The only difference is that we have to separate the cases where k, the degree of p_k, is even ($p_k(t) \to +\infty$ for $t \to -\infty$) or odd ($p_k(t) \to -\infty$ for $t \to -\infty$). Hence we have proved the following theorem.

Theorem 9.3.2 *Assume that the starting vector v_1 has been chosen randomly with respect to the uniform distribution over S^{n-1} and let $\varepsilon \in (0,1)$. Then λ^{up}, the largest zero of the polynomial*

$$f_L(t) = p_k(t) - 1/\delta \tag{9.3.3}$$

with δ given by (9.3.2), is an upper bound for the spectrum of A with probability at least $1-\varepsilon$, and λ^{low}, the smallest zero of

$$f_L(t) = (-1)^k p_k(t) - 1/\delta, \tag{9.3.4}$$

is a lower bound for the spectrum of A with probability at least $1-\varepsilon$.

Note that if we are unlucky in choosing v_1, so that $|\gamma_n| < \delta$, then the computed bounds may or may not be correct; see Section 9.7 for an illustration.

The determination of the lower and upper bounds from Theorem 9.3.2 is rather cheap in general (compared with a matrix-vector multiplication with A); the computation of $f_L(t)$ (using (9.2.3)) costs approximately $5k$ floating point operations. Note that the Ritz values and vectors are not needed to obtain these bounds of the spectrum. For very small k one cannot expect to obtain tight bounds, so it only makes sense to compute the zeros of (9.3.3) and (9.3.4) for k of moderate size. In practice one could, e.g., compute these zeros only every second or third Lanczos step until the bounds become sufficiently sharp.

9.4 Spectral bounds using Ritz polynomials

We can also try to obtain probabilistic upper and lower bounds for the spectrum of A using some Ritz polynomials $q_j^{(k)}$. The degree of these polynomials is one less than the degree of p_k, but while $p_k(\theta_k^{(k)}) = 0$, the polynomial $q_k^{(k)}$ has its last zero in $\theta_{k-1}^{(k)}$ and could be a competitor of p_k to give a possibly tighter upper bound. Similarly, $q_1^{(k)}$ may be used to obtain another lower bound.

We write $\theta_j^{(k)}$ as a Rayleigh quotient:

$$\theta_j^{(k)} = (Ay_j^{(k)}, y_j^{(k)}) = \sum_{i=1}^{n} \lambda_i \, \gamma_i^2 \, q_j^{(k)}(\lambda_i)^2. \tag{9.4.1}$$

First suppose that A is positive semidefinite. Then set $j = k$ to derive the inequality $\theta_k^{(k)} \geq \lambda_n \gamma_n^2 q_k^{(k)}(\lambda_n)^2$. Hence the zero $\lambda^{\rm up} > \theta_k^{(k)}$ of

$$f_R(t) = t q_k^{(k)}(t)^2 \; - \; \theta_k^{(k)}/\gamma_n^2 \tag{9.4.2}$$

is an upper bound for λ_n. If γ_n is not known, one can obtain a probabilistic upper bound $\lambda^{\rm up}$ of λ_n with probability at least $1-\varepsilon$, as in the previous section. (Replace γ_n in (9.4.2) by δ where δ satisfies (9.3.2).)

As in the previous section, if we happen to choose a v_1 so that $|\gamma_n| < \delta$, then we are not certain that the computed upper bound is correct. It can even happen that the largest zero $\lambda^{\rm up}$ of f_R with γ_n replaced by δ satisfies $\lambda^{\rm up} < \theta_k^{(k)}$! See Section 9.7 for an illustration.

When it is not known whether A is positive definite, we can obtain a probabilistic upper bound in the following way. Let $-\sigma < 0$ be a known lower bound for the spectrum of A: then the matrix $A + \sigma I$ is positive semidefinite. We get

$$\theta_k^{(k)} + \sigma = \sum_{i=1}^{n} (\lambda_i + \sigma)\, \gamma_i^2\, q_k^{(k)}(\lambda_i)^2$$

with $\lambda_i + \sigma \geq 0$ for all i. The rightmost zero of

$$f_R(t) = (t+\sigma) q_k^{(k)}(t)^2 \; - \; (\theta_k^{(k)} + \sigma)/\gamma_n^2$$

is an upper bound for the spectrum of A. Again, we can replace γ_n by the δ that satisfies (9.3.2) to compute a probabilistic upper bound.

For a lower bound, we use the polynomial $q_1^{(k)}$. If A is negative semidefinite, it follows from $\theta_1^{(k)} \leq \lambda_1 \gamma_1^2 q_1^{(k)}(\lambda_1)^2$ (cf. (9.4.1)) that the unique zero $\lambda^{\rm low} < \theta_1^{(k)}$ of

$$f_R(t) = t q_1^{(k)}(t)^2 \; - \; \theta_1^{(k)}/\gamma_1^2 \tag{9.4.3}$$

is a lower bound for λ_1. Otherwise one has to use a shift $\tau > 0$ such that $A - \tau I$ becomes negative semidefinite and modify f_R in (9.4.3) accordingly. Of course the shifts σ and τ should be chosen as small as possible to get the best results.

The bounds discussed in this section can be determined for example by Newton's method or bisection. In order to compute $f_R(t)$ one has to know the largest or smallest Ritz value and the corresponding eigenvector of the tridiagonal matrix T_k. Apart from that, the computation of $f_R(t)$ is cheap. The determination of the bounds based on Ritz polynomials will be more expensive in general than the determination of the bounds based on the Lanczos polynomials. (The Ritz values and vectors are not needed in the latter case.)

It is interesting to compare the sharpness of the bounds based on Ritz polynomials and those based on Lanczos polynomials. For simplicity we assume that A is positive semidefinite and compare the largest zero of (9.4.2) with the largest zero of (9.3.1). (The other cases, including those where shifts are used, can be analyzed in a similar way.) Consider the function

$$g(t) = \sqrt{t/\theta_k^{(k)}}\, q_k^{(k)}(t) \; - \; 1/|\gamma_n|; \tag{9.4.4}$$

the largest zero of g is the largest zero of f_R from (9.4.2). After some straightforward calculations, using (9.2.5) with $j = k$, one obtains that (with f_L as in (9.3.1) and g as in (9.4.4))

$$f_L(t) \;<\; g(t) \qquad \text{for} \quad \theta_k^{(k)} \;\le\; t \;\le\; (1+\zeta)\,\theta_k^{(k)}$$

and

$$f_L(t) \;>\; g(t) \qquad \text{for} \quad t \;\ge\; (1+\zeta+\zeta^2)\,\theta_k^{(k)},$$

where $\zeta = \|r_k^{(k)}\|/\theta_k^{(k)}$. The quantity ζ can be interpreted as an approximation of the relative error for the largest eigenvalue, and ζ will be small after sufficiently many Lanczos steps. For small ζ the Ritz polynomial provides a smaller upper bound for λ_n *only* when this upper bound is very close to $\theta_k^{(k)}$—but in that case the Lanczos polynomial yields a very tight upper bound as well. Hence, it is not likely that the bounds based on Ritz polynomials are sharper than the bounds obtained with the Lanczos polynomials—unless ζ is large. Numerical experiments illustrating these observations can be found in Section 9.7.

9.5 Spectral bounds using Chebyshev polynomials

Chebyshev polynomials are often used to obtain error bounds for the Lanczos method; cf., e.g., [31, 49, 61]. In this section we explain how these polynomials can be used to obtain probabilistic upper and lower bounds for the spectrum of A, based on computations with the Lanczos method. One type of bounds follows easily from a result by Kuczyński and Woźniakowski [49, Theorem 3].

Let $c_j(t) = \cos(j \arccos t)$ be the *Chebyshev polynomial (of the first kind)* of degree j, with the usual extension outside the interval $[-1, 1]$. The polynomial

$$u_{j-1}(t) = \tfrac{1}{j}\, c_j'(t)$$

of degree $j-1$ is a *Chebyshev polynomial of the second kind* (cf. [67, p. 7]).

In [49, Theorem 3], the following result has been derived for symmetric positive definite matrices. Let $t > 1$ and v_1 be chosen randomly from the uniform distribution over S^{n-1}. Then

$$P(\lambda_n \le t\,\theta_k^{(k)}) \;\ge\; 1 \;-\; 2\left(B(\tfrac{n-1}{2}, \tfrac{1}{2})\,\sqrt{t-1}\,u_{2(k-1)}(\sqrt{t}\,)\right)^{-1}. \tag{9.5.1}$$

(where B is the Euler Beta function.) The estimate (9.5.1) can be generalized for symmetric indefinite matrices by using a shift σ such that $A + \sigma I$ is positive definite. Probability estimates for lower bounds of λ_1 can be obtained similarly. Along these lines we can derive bounds for the spectrum of A with probability at least $1-\varepsilon$, and these results are presented in the following theorem.

Theorem 9.5.1 *Let $\varepsilon \in (0,1)$ and $\sigma, \tau \in \mathbb{R}$ be such that $A + \sigma I$ is positive semidefinite, and $A - \tau I$ is negative semidefinite. Consider for $t \ge 1$ the function*

$$f(t) \;=\; \tfrac{\varepsilon}{2}\,B(\tfrac{n-1}{2}, \tfrac{1}{2})\,\sqrt{t-1}\,u_{2(k-1)}(\sqrt{t}\,) \;-\; 1 \tag{9.5.2}$$

(B is the Euler Beta function) and let $t_k > 1$ be the (unique) zero of f. Furthermore, let v_1 be chosen randomly from the uniform distribution over S^{n-1}. Then

$$\lambda^{\mathrm{up}} \;=\; t_k\,\theta_k^{(k)} \;+\; (t_k - 1)\sigma \tag{9.5.3}$$

is an upper bound for the spectrum of A with probability at least $1-\varepsilon$, and

$$\lambda^{\mathrm{low}} \;=\; t_k\,\theta_1^{(k)} \;-\; (t_k - 1)\tau \tag{9.5.4}$$

is a lower bound for the spectrum of A with probability at least $1-\varepsilon$.

The quantity t_k can be determined numerically. The numbers $u_j(t)$ can be computed from the three-term recurrence $u_j(t) = 2tu_{j-1}(t) - u_{j-2}(t)$ for $j \geq 2$, $u_0(t) = 1$, $u_1(t) = 2t$ (see, e.g., [67, p. 40]). From (9.5.3) and (9.5.4) it is clear that the shifts σ and τ should be chosen as small as possible (cf. Section 9.4).

Other bounds for the spectrum of A can be obtained as follows, using Chebyshev polynomials of the first kind. Let $a < b$ and $c_j(t;a,b) = c_j(1 + 2(t-b)/(b-a))$ be the Chebyshev polynomial of degree j with respect to the interval $[a,b]$. With σ such that $A + \sigma I$ is positive semidefinite, we define the polynomial $h(t) = c_{k-1}(t; -\sigma, \theta_k^{(k)})$ and the vector $x = h(A)v_1 \in K_k(A, v_1)$. From $\theta_k^{(k)}(x,x) \geq (Ax,x)$ it follows that the largest zero of

$$f_C(t) \;=\; (t - \theta_k^{(k)})c_{k-1}(t; -\sigma, \theta_k^{(k)})^2 \;-\; (\theta_k^{(k)} + \sigma)/\gamma_n^2 \tag{9.5.5}$$

is an upper bound for λ_n. (Invoke (9.2.1): use $\sum \gamma_j^2 \leq 1$ where the summation is with respect to those j satisfying $\lambda_j \leq \theta_k^{(k)}$ and $h(\lambda_j)^2 \leq 1$ for $\lambda_j \leq \theta_k^{(k)}$.)

With γ_n replaced by the δ computed from (9.3.2), as in the previous sections, one obtains an upper bound λ^{up} for the spectrum of A with probability at least $1-\varepsilon$. A lower bound for the spectrum of A can be obtained in a similar way, using $\theta_1^{(k)}(x,x) \leq (Ax,x)$ with $x = c_{k-1}(A; \theta_1^{(k)}, \tau)v_1$, where τ is such that $A - \tau I$ is negative semidefinite.

In order to compare the bounds derived along these lines with those obtained from Theorem 9.5.1, we first replace γ_n in (9.5.5) by δ and scale the interval $[-\sigma, \theta_k^{(k)}]$ to $[0,1]$. The largest zero λ^{up} of (9.5.5) satisfies the equality $\lambda^{\mathrm{up}} = \widehat{t}\,\theta_k^{(k)} + (\widehat{t} - 1)\sigma$, where $\widehat{t} > 1$ is the unique zero of

$$g(t) \;=\; \delta\sqrt{t-1}\,c_{k-1}(t;0,1) \;-\; 1.$$

One can show that $c_{k-1}(t;0,1) = c_{2(k-1)}(\sqrt{t}\,;-1,1)$ $(= c_{2(k-1)}(\sqrt{t}\,))$ for $t > 0$. This means that we have to compare the zeros of (9.5.2) and those of

$$g(t) \;=\; \delta\sqrt{t-1}\,c_{2(k-1)}(\sqrt{t}\,) \;-\; 1. \tag{9.5.6}$$

The relation between δ and $\frac{\varepsilon}{2}B(\frac{n-1}{2}, \frac{1}{2})$ is given by (9.3.2). One has $\delta > \frac{\varepsilon}{2}B(\frac{n-1}{2}, \frac{1}{2})$ for all $\varepsilon \in (0,1)$ and $n > 3$, but $\delta \approx \frac{\varepsilon}{2}B(\frac{n-1}{2}, \frac{1}{2})$ for ε and n of practical interest. For instance, $(\delta - \frac{\varepsilon}{2}B(\frac{n-1}{2}, \frac{1}{2}))/\delta \approx 2.6 \cdot 10^{-5}$ for $\varepsilon = 1.0 \cdot 10^{-2}$ and $n = 10^3, 10^4, 10^5, 10^6$. On the other hand one has the relation

$$u_{2(k-1)}(\sqrt{t}\,) \;=\; 2\,c_{2(k-1)}(\sqrt{t}\,) \;+\; u_{2(k-2)}(\sqrt{t}\,) \qquad \text{for} \quad t > 0$$

(cf., e.g., [67, p. 9]) so that $u_{2(k-1)}(\sqrt{t}) > 2c_{2(k-1)}(\sqrt{t})$ for $t \geq 1$ and this implies, together with $\delta \approx \frac{\varepsilon}{2}B(\frac{n-1}{2}, \frac{1}{2})$, that the zero of (9.5.6) is larger than the zero of (9.5.2) in most applications. Hence, the upper bound λ^{up} from (9.5.3) is in general smaller than the upper bound obtained from (9.5.5), so Theorem 9.5.1 will produce sharper bounds than the construction described above. These observations are supported by numerical experiments in Section 9.7.

9.6 Upper bounds for the number of Lanczos steps

9.6.1 Bounds based on Theorem 9.5.1

Theorem 9.5.1 can also be used to compute a probabilistic upper bound for the number of Lanczos steps necessary to obtain a Ritz value close enough to λ_n in a relative or absolute sense. These estimates can be obtained while executing the Lanczos process. First we investigate how many Lanczos steps are needed to obtain a relative error that is smaller than a prescribed tolerance `tol` with probability at least $1 - \varepsilon$.

Suppose k steps of the Lanczos method have been performed and $\theta_k^{(k)} > 0$; if $\theta_k^{(k)} \leq 0$ the eigenvalue λ_n can be arbitrarily close to zero and the relative error $(\lambda_n - \theta_m^{(m)})/\lambda_n$ cannot be estimated properly. Let $m \geq k$ and let t_m be the zero of the function f in (9.5.2) with k replaced by m. It follows from (9.5.3) that

$$\frac{\lambda_n - \theta_m^{(m)}}{\lambda_n} \leq \frac{(t_m - 1)(\theta_m^{(m)} + \sigma)}{\lambda_n} \leq \frac{(t_m - 1)(\lambda_n + \sigma)}{\lambda_n} \leq \frac{(t_m - 1)(\mu + \sigma)}{\mu} \tag{9.6.1}$$

holds with probability at least $1 - \varepsilon$; here $\mu = \theta_k^{(k)}$ if $\sigma \geq 0$, and $\mu \geq \lambda_n$ (e.g., $\mu = \|A\|_\infty$; one should not take a probabilistic upper bound for λ_n) whenever $\sigma < 0$; σ is as in Theorem 9.5.1. The requirement $(t_m - 1)(\mu + \sigma)/\mu \leq \mathtt{tol}$ is equivalent to $t_m \leq 1 + \mathtt{tol} \cdot \mu/(\mu + \sigma)$, and the smallest integer m, for which the quantity t_m from (9.5.2) satisfies

$$t_m \leq 1 + \mathtt{tol} \cdot \mu/(\mu + \sigma), \tag{9.6.2}$$

is an upper bound for the number of Lanczos steps necessary to provide an approximation $\theta_m^{(m)}$ to λ_n that satisfies $(\lambda_n - \theta_m^{(m)})/\lambda_n \leq \mathtt{tol}$ with probability at least $1 - \varepsilon$. Note that in case $\sigma > 0$ the right-hand side of (9.6.2) increases with k, so that the smallest number m satisfying (9.6.2) may decrease during the execution of the Lanczos process.

For symmetric positive definite matrices an upper bound m for the number of Lanczos steps which yields an approximation to the largest eigenvalue, such that the relative error is bounded by `tol` with probability at least $1 - \varepsilon$, has been given in [48, Theorem 4.2]: the number m should satisfy

$$1.648\sqrt{n}\,\mathrm{e}^{-(2m-1)\sqrt{\mathtt{tol}}} \leq \varepsilon. \tag{9.6.3}$$

Numerical experiments show that (9.6.3) yields almost the same upper bound as (9.6.2) with $\sigma = 0$ (in most cases the bounds were exactly the same, while the difference was at most two steps); this is not surprising in view of the discussion in [49, p. 679]. However,

(9.6.2) can be used for indefinite matrices as well, as long as $\theta_k^{(k)} > 0$. Furthermore, for symmetric positive definite matrices smaller numbers m may be obtained when (9.6.2) is applied with $\sigma < 0$.

To estimate the number of steps, still necessary to have the absolute error $\lambda_n - \theta_m^{(m)} \leq$ `tol` with probability at least $1 - \varepsilon$, we proceed as follows. If m satisfies the requirement (cf. (9.6.1))

$$(t_m - 1)(\mu + \sigma) \ \leq \ \texttt{tol}, \tag{9.6.4}$$

with $\mu \geq \lambda_n$ (μ should not be a probabilistic upper bound), the equality $\lambda_n - \theta_m^{(m)} \leq$ `tol` holds with probability at least $1 - \varepsilon$. The smallest integer m satisfying (9.6.4) can be computed. Note that (9.6.4) is also valid when $\theta_k^{(k)} \leq 0$ and we do not have to distinguish between the cases $\sigma \geq 0$ and $\sigma < 0$.

Estimates for the number of Lanczos steps, to be done so that the (relative) error in the smallest eigenvalue is less than `tol` with probability at least $1 - \varepsilon$, can be derived in a similar way.

9.6.2 Bounds for the number of Lanczos steps in case of misconvergence

Suppose that after sufficiently many Lanczos steps the largest Ritz value seems to have converged to an eigenvalue: $\theta_k^{(k)} \approx \theta_{k-1}^{(k-1)}$ for several consecutive k and $\|r_k^{(k)}\|$ is small. It is known that $|\theta_k^{(k)} - \lambda_j| \leq \|r_k^{(k)}\|$ for a certain eigenvalue λ_j (see, e.g., [61, Section 4.5]), and in most cases the largest Ritz value will have converged to the largest eigenvalue λ_n, but it may also happen that $\theta_k^{(k)}$ is not close to λ_n (misconvergence); this can happen, e.g., if $|\gamma_n|$ is very small. Below we show how one can determine a probabilistic upper bound for the number of Lanczos steps needed after which one can conclude that either $\lambda_n < \lambda$ holds for a given constant λ, or a misconvergence has been detected, i.e., $\lambda_n > \theta_k^{(k)} + \|r_k^{(k)}\|$.

Let $m > k$ and g be a polynomial of degree $m - 1$, and $x = g(A)v_1 \in K_m(A, v_1)$. If $\lambda_n > \theta_k^{(k)} + \|r_k^{(k)}\|$, the inequality

$$(Ag(A)v_1, g(A)v_1) \ > \ \left(\theta_k^{(k)} + \|r_k^{(k)}\|\right)(g(A)v_1, g(A)v_1) \tag{9.6.5}$$

is satisfied for a certain m and a suitable polynomial g: the Ritz polynomial $q_m^{(m)}$ maximizes the Rayleigh quotient $(Ag(A)v_1, g(A)v_1)/(g(A)v_1, g(A)v_1)$ but $q_m^{(m)}$ is not available after k steps of the Lanczos process, so we will consider another polynomial of degree $m - 1$. Rewriting (9.6.5) using (9.2.1) gives

$$\begin{aligned} \left(\lambda_n - (\theta_k^{(k)} + \|r_k^{(k)}\|)\right)\gamma_n^2\, g(\lambda_n)^2 \ &> \ \left(\theta_k^{(k)} + \|r_k^{(k)}\| - \lambda_{n-1}\right)\gamma_{n-1}^2\, g(\lambda_{n-1})^2 \\ &+ \ \sum_{j=1}^{n-2}\left(\theta_k^{(k)} + \|r_k^{(k)}\| - \lambda_j\right)\gamma_j^2\, g(\lambda_j)^2. \end{aligned} \tag{9.6.6}$$

In order to satisfy (9.6.6) with m as small as possible we search for a polynomial g that resembles the Ritz polynomial $q_m^{(m)}$. We have $q_k^{(k)}$ at our disposal, and therefore we take $g(t) = q_k^{(k)}(t)\, h(t)$ with h a suitable polynomial of degree $m - k$. We assume that

$|\theta_k^{(k)} - \lambda_{n-1}| \leq \|r_k^{(k)}\|$ (with $\|r_k^{(k)}\|$ small); this assumption is likely to be realistic in case of a misconvergence. In order to amplify the effect of $q_k^{(k)}$ in (9.6.6) we choose h such that h is large in λ_n and small in $\lambda_1, \ldots, \lambda_{n-2}$. Hence $h(t) = c_{m-k}(t; \lambda_1, \lambda_{n-2})$ would be a proper choice, but λ_1 and λ_{n-2} are not known, so we replace both quantities. Again let $-\sigma \leq \lambda_1$, and assume that $\lambda_{n-2} \leq \theta_{k-1}^{(k)} + \|r_{k-1}^{(k)}\|$; we now define

$$g(t) \;=\; q_k^{(k)}(t)\, c_{m-k}\big(t; -\sigma, \theta_{k-1}^{(k)} + \|r_{k-1}^{(k)}\|\big).$$

If we replace in the right-hand side of (9.6.6) the quantity $\theta_k^{(k)} + \|r_k^{(k)}\| - \lambda_{n-1}$ by $2\,\|r_k^{(k)}\|$, γ_{n-1}^2 by 1, $g(\lambda_{n-1})$ by $g(\theta_k^{(k)} + \|r_k^{(k)}\|)$, and $g(\lambda_j)$ by M, where

$$M \;=\; \max\{\, |q_k^{(k)}(t)| \;:\; -\sigma \leq t \leq \theta_{k-1}^{(k)} + \|r_{k-1}^{(k)}\| \,\},$$

then the inequality

$$\begin{aligned} \big(\lambda_n - (\theta_k^{(k)} + \|r_k^{(k)}\|)\big)\, g(\lambda_n)^2 \;&>\; 2\,\|r_k^{(k)}\|\, g\big(\theta_k^{(k)} + \|r_k^{(k)}\|\big)^2 / \gamma_n^2 \\ &+\; M^2\big(\theta_k^{(k)} + \|r_k^{(k)}\| + \sigma\big) / \gamma_n^2 \end{aligned} \tag{9.6.7}$$

implies (9.6.6) (cf. the derivation of (9.5.5), which is based on the same ideas). We now replace λ_n in (9.6.7) by the given constant λ and γ_n by δ, where $|\gamma_n| \geq \delta$ holds with probability $1 - \varepsilon$. We determine the smallest integer $m > k$ such that

$$\begin{aligned} \big(\lambda - (\theta_k^{(k)} + \|r_k^{(k)}\|)\big)\, g(\lambda)^2 \;&>\; 2\,\|r_k^{(k)}\|\, g\big(\theta_k^{(k)} + \|r_k^{(k)}\|\big)^2 / \delta^2 \\ &+\; M^2\big(\theta_k^{(k)} + \|r_k^{(k)}\| + \sigma\big) / \delta^2 \end{aligned} \tag{9.6.8}$$

is satisfied and perform $m - k$ Lanczos steps to obtain $\theta_m^{(m)}$. If $\theta_m^{(m)} < \theta_k^{(k)} + \|r_k^{(k)}\|$, then (9.6.5) and (9.6.6) are violated. This implies that (9.6.7) does not hold if, e.g., $\lambda_{n-1} \leq \theta_{k-1}^{(k)} + \|r_{k-1}^{(k)}\|$. (This will be satisfied in most cases.) From the fact that (9.6.7) is violated and (9.6.8) holds we conclude that $\lambda_n < \lambda$ holds with probability at least $1 - \varepsilon$.

If $\theta_m^{(m)} > \theta_k^{(k)} + \|r_k^{(k)}\|$, we know that a misconvergence has occurred and we do not know whether $\lambda_n < \lambda$ is satisfied or not. In the latter case one may repeat the above construction with k replaced by m.

These ideas can also be used to investigate whether or not the smallest Ritz value has converged to λ_1.

9.7 Numerical experiments

In this section we compare the different bounds derived in the previous sections. All experiments are carried out with MATLAB. Without loss of generality we can restrict ourselves to diagonal matrices A (cf. [48, Section 6]): this will reduce the influence of rounding errors on our computations. For analysis it is also convenient to know the eigenvalues and eigenvectors of A. The vector v_1 is chosen randomly from the uniform distribution over the unit sphere S^{n-1}; in [48, p. 1116] it is explained how this can be done.

Experiment 9.7.1 In our first example we take

$$n = 1000, \quad A = \text{diag}(1:1000).$$

Let $\varepsilon = 0.01$, i.e., we are looking for bounds of the spectrum that are 99% reliable. From (9.3.2) one obtains $\delta \approx 3.97 \cdot 10^{-4}$. We checked that our randomly chosen starting vector v_1 satisfied $|\gamma_1| > \delta$ and $|\gamma_n| > \delta$, so the computed probabilistic bounds are true bounds for the spectrum of A. We have performed 100 Lanczos steps. The shifts (see Sections 9.4 and 9.4) used in our computations are $\sigma = 0$ and $\tau = \lambda_n = 1000$. The results are displayed in Figure 9.1.

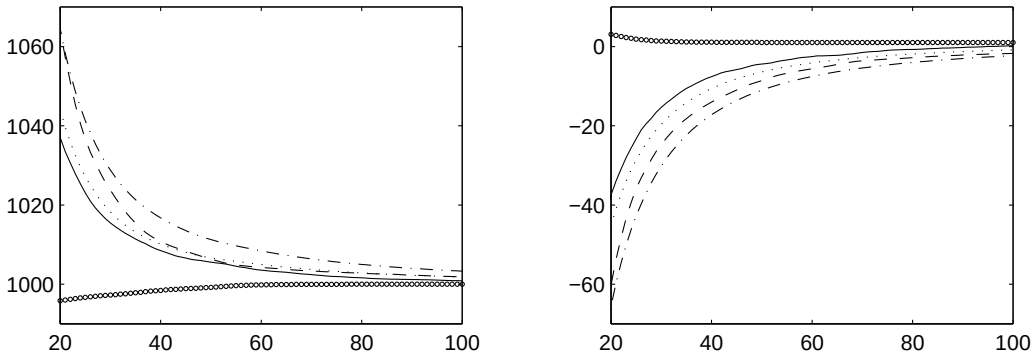


FIGURE 9.1: Probabilistic bounds for the spectrum of A. Solid curves correspond to the bounds based on Lanczos polynomials, the dashed curves correspond to bounds based on Ritz polynomials, the dotted curves correspond to bounds obtained from Theorem 9.5.1, and the dash-dotted curves correspond to (9.5.5). The left figure shows the upper bounds and the right figure the lower bounds. The largest Ritz values (left picture) and smallest Ritz values (right picture) are indicated by small circles.

We see that the Lanczos polynomials provide the sharpest bounds and (9.5.5) yields the worst bounds. In Section 9.4 it has already been explained why the Lanczos polynomials may provide better bounds than the Ritz polynomials. Furthermore, it may not be a surprise that the Lanczos polynomials produce better bounds than the Chebyshev polynomials, because more information regarding the actual Lanczos process is used in the construction of the Lanczos polynomials. The relationship between the different bounds based on Chebyshev polynomials is in agreement with the discussion on this topic in Section 9.5. We repeated the same experiment with other random starting vectors v_1, and the bounds behaved similarly as those displayed in Figure 9.1.

We also investigated how many Lanczos steps are needed to obtain an approximation to λ_n with a relative error less than a prescribed tolerance `tol`. Again we set $\sigma = 0$, so that (9.6.2) reduces to $t_m \leq 1 + \texttt{tol}$; the upper bound m for the number of Lanczos steps does not depend on the matrix A or the starting vector v_1 and can be computed in advance. The results are displayed in Table 9.1. We see that the upper bound m from (9.6.2) is much larger than k_1, the actual number of steps needed to obtain a relative error smaller than `tol`; this has already been observed in other examples for the upper bound obtained with (9.6.3) [48, 49]. We also observe that $m > k_2$, the number of steps needed to obtain $(\lambda^{\text{up}} - \theta_k^{(k)})/\lambda^{\text{up}} \leq \texttt{tol}$ with λ^{up} the upper bound obtained from

the Lanczos polynomial of degree k. This is not surprising in view of the results from Figure 9.1, because m is related to the upper bound determined with Theorem 9.5.1, and these bounds are not as sharp as those based on Lanczos polynomials. Instead of performing m Lanczos steps, it may be useful in practice to compute $(\lambda^{\mathrm{up}} - \theta_k^{(k)})/\lambda^{\mathrm{up}}$ while executing the Lanczos method and check whether this quantity is smaller than `tol` or not.

TABLE 9.1: The second column displays the smallest integer m satisfying (9.6.2) with $\sigma = 0$. The smallest integer k_1 for which $(\lambda_n - \theta_k^{(k)})/\lambda_n \leq$ `tol` is shown in the third column, and the smallest integer k_2 with $(\lambda^{\mathrm{up}} - \theta_k^{(k)})/\lambda^{\mathrm{up}} \leq$ `tol`, where λ^{up} is the upper bound for λ_n obtained with the Lanczos polynomial of degree k, is listed in the fourth column of the table.

`tol`	m	k_1	k_2
$5.0 \cdot 10^{-2}$	20	5	18
$1.0 \cdot 10^{-2}$	44	11	40
$5.0 \cdot 10^{-3}$	61	17	55
$1.0 \cdot 10^{-3}$	136	48	97

We have repeated the experiments described above with $\varepsilon = 0.001$ (instead of $\varepsilon = 0.01$). The behavior of the bounds is the same as for $\varepsilon = 0.01$, but of course the bounds are further away from the spectrum of A. In order to compare the different bounds, let λ^{up} be an upper bound corresponding to $\varepsilon = 0.01$ (determined with one of the four techniques discussed here), and let $\widetilde{\lambda}^{\mathrm{up}}$ be the upper bound determined with the same technique but with $\varepsilon = 0.001$. For all four techniques we observed that $1 < (\widetilde{\lambda}^{\mathrm{up}} - \lambda_n)/(\lambda^{\mathrm{up}} - \lambda_n) < 2.2$ for $20 \leq k \leq 100$ (k denotes the number of Lanczos steps) and the same holds for $(\lambda_1 - \widetilde{\lambda}^{\mathrm{low}})/(\lambda_1 - \lambda^{\mathrm{low}})$, where the lower bounds λ^{low} and $\widetilde{\lambda}^{\mathrm{low}}$ are defined analogously. Hence the behavior of the bounds for the spectrum of A does not change much when ε is decreased from 0.01 to 0.001, which is reasonable because the polynomials used to derive the bounds grow fast outside the spectrum of A. $\oslash$

Experiment 9.7.2 The second example comes from the discretization of the Laplace operator on the unit square with homogeneous Dirichlet boundary conditions. When the standard second order finite difference scheme with uniform meshwidth equal to $1/33$ (in both directions) is used, one obtains a symmetric matrix of order $n = 32^2 = 1024$ with eigenvalues

$$33^2(-4 + 2\cos(\tfrac{i\pi}{33}) + 2\cos(\tfrac{j\pi}{33})), \qquad i, j = 1, 2, \ldots, 32.$$

Let A be the diagonal matrix of order 1024 with these eigenvalues on its diagonal in increasing order. Note that A is negative definite.

We have computed bounds for the spectrum of A with $\varepsilon = 0.01$ (which yields $\delta \approx 3.92 \cdot 10^{-4}$ by (9.3.2)), $\sigma = -\lambda_1$ and $\tau = 0$, using different randomly chosen starting vectors. For most starting vectors the bounds behave similarly as in the first example and we will not consider this further. Instead we deal with two different starting vectors that provide a different behavior for the upper bounds (similar results can be obtained for lower bounds as well), and the results can be found in Figure 9.2.

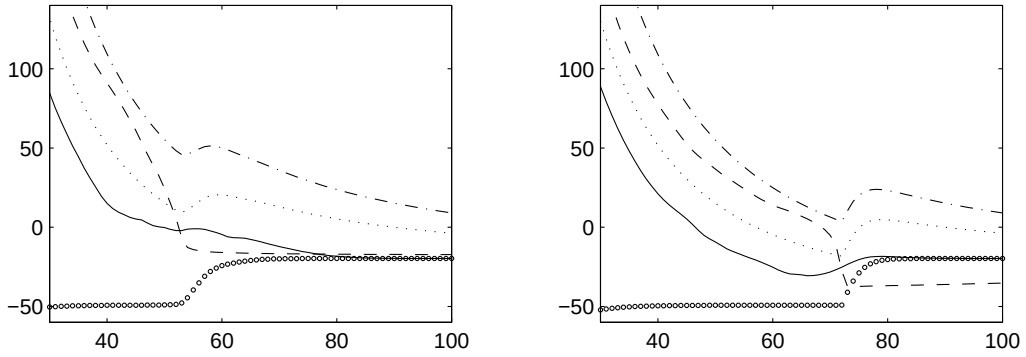


FIGURE 9.2: "Upper bounds" for the spectrum of A, obtained with two different starting vectors; the starting vector for the left picture satisfies $|\gamma_n| > \delta$, while $|\gamma_n| < \delta$ for the starting vector used to produce the right picture. Solid curves correspond to the bounds based on Lanczos polynomials, the dashed curves correspond to bounds based on Ritz polynomials, the dotted curves correspond to bounds obtained from Theorem 9.5.1, and the dash-dotted curves correspond to (9.5.5). The largest Ritz values are indicated by small circles.

In the left picture we see what can happen if $|\gamma_n|$ is small ($|\gamma_n| = 5.46 \cdot 10^{-4}$ for this example), but still greater than δ. The Ritz polynomials provide the sharpest bounds at a certain stage of the Lanczos process. At that stage the misconvergence behavior of the Lanczos process (cf., e.g., [62]) is discovered: for $37 \leq k \leq 49$ one has $|\lambda_{n-1} - \theta_k^{(k)}| \leq 0.15$ ($\lambda_{n-1} \approx -49.22$), and the largest Ritz values seem to converge to a number close to the (double) eigenvalue λ_{n-1}. For larger values k the Lanczos process notices the existence of a larger eigenvalue ($\lambda_n \approx -19.72$) and starts to converge to this eigenvalue. At the stage of the Lanczos process where the misconvergence behavior is discovered, the norm of the residual usually increases strongly (for example, $\|r_{42}^{(42)}\| = 5.65$ and $\|r_{55}^{(55)}\| = 102$) and a large residual norm may explain why the Ritz polynomials provide sharper bounds than the Lanczos polynomials (see the discussion at the end of Section 9.4). However, for larger k the bounds based on Lanczos polynomials are again the sharpest ones. The misconvergence of the Lanczos process also causes a hump in the upper bounds obtained with the Chebyshev polynomials. Finally we note that the upper bounds obtained with the Lanczos polynomials are much sharper than those obtained with the Chebyshev polynomials.

In the right figure the behavior is shown for a starting vector for which, in contrary to our assumption, $|\gamma_n| < \delta$ ($|\gamma_n| = 3.13 \cdot 10^{-5}$). This means that the probabilistic upper bounds for λ_n need not to be true bounds, and the right picture in Figure 9.2 shows that at certain stages of the Lanczos process the Lanczos and Ritz polynomials provide bounds that are actually smaller than λ_n. The Chebyshev bounds follow the jump of the Ritz values at the discovering of the misconvergence, as in the left picture. At that stage the Lanczos bound corrects its value to give a tight bound, but the Ritz bound fails completely: the upper bound stays far below the largest Ritz value. ⊘

Experiment 9.7.3 In the third example we illustrate the theory of Section 9.6.2. We take

$$n = 1000, \quad A = \mathrm{diag}(1, 2, \ldots, 999, 1020).$$

We set $\sigma = -\lambda_1$ and the starting vector v_1 is chosen as follows: $\gamma_1 = \gamma_2 = \gamma_{n-2} = \gamma_{n-1} = \zeta$, $\gamma_j = 10^{-3}\zeta$ $(3 \leq j \leq n-3)$, $\gamma_n = 10^{-6}\zeta$, and the constant ζ is such that $\sum \gamma_j^2 = 1$. For $k = 34$ we have $\theta_k^{(k)} = \lambda_{n-1} - 3.20 \cdot 10^{-5}$, $\|r_k^{(k)}\| = 7.3 \cdot 10^{-2}$ so that $\lambda_n > \theta_k^{(k)} + \|r_k^{(k)}\|$. We now determine the smallest integer m for which (9.6.8) holds. We take $k = 34$, $\lambda = \lambda_n$, $\delta = \gamma_n = 5.0 \cdot 10^{-7}$ and $M = 2.11$. The smallest m satisfying (9.6.8) is $m = 69$. The Lanczos process finds the largest eigenvalue λ_n earlier: one has, e.g., $\theta_{50}^{(50)} = \lambda_n - 2.4 \cdot 10^{-2}$, $\theta_{60}^{(60)} = \lambda_n - 5.5 \cdot 10^{-5}$ and $\theta_{69}^{(69)} = \lambda_n - 2.4 \cdot 10^{-7}$. This behavior is not surprising: the Ritz polynomial $q_m^{(m)}$ maximizes the Rayleigh quotient $(Ag(A)v_1, g(A)v_1)/(g(A)v_1, g(A)v_1)$ and several other estimates used in the derivation of (9.6.8) may not be sharp as well. ⊘

9.8 Conclusions

Using the fact that the Lanczos, Ritz, and Chebyshev polynomials increase rapidly outside the smallest interval containing the Ritz values, we have derived probabilistic bounds for the spectrum of a symmetric matrix. These bounds can be computed while executing the Lanczos process. From theoretical arguments supported by experiments, we conclude that the bounds obtained with the Lanczos polynomials are generally sharper than those derived from Chebyshev polynomials (more information regarding the actual Lanczos process is used in the construction of the Lanczos polynomials). In most cases the bounds based on Lanczos polynomials are also sharper than the bounds found with Ritz polynomials—unless the norm of the corresponding residual is relatively large (which occurs if the Lanczos method suffers from a misconvergence).

The bounds corresponding to the Lanczos polynomials are cheap to compute, because the Ritz values are not required. When the Ritz values are available, it is useful to compute the bounds based on these polynomials as well, because they might be sharper; in that case it can indicate a misconvergence of the Lanczos method. The bounds based on Theorem 9.5.1, using Chebyshev polynomials of the second kind, may be determined as well because they can be computed cheaply when the Ritz values are known. The bounds obtained from Theorem 9.5.1 are sharper than those derived from (9.5.5), which are based on Chebyshev polynomials of the first kind, in all cases of practical interest; hence it seems not useful to determine the latter ones.

Chebyshev polynomials may also be used to determine probabilistic bounds for the number of Lanczos steps still to be done to get bounds for the (relative) error which are smaller than the desired tolerance. However, our experiments suggest that these bounds are much larger than the actual number of Lanczos steps still necessary to get an approximation which is sufficiently accurate. From their derivation (9.6.1) it is clear that one cannot expect a proper estimation of the number of steps required if the bounds from Theorem 9.5.1 are far from sharp.

A combination of Ritz and Chebyshev polynomials can be used to obtain probabilistic bounds for the number of Lanczos steps needed such that one can decide that either the spectrum lies between certain prescribed bounds or a misconvergence has occurred.

Bibliography

[1] Arbenz, P., and Hochstenbach, M. E. A Jacobi–Davidson method for solving complex symmetric eigenvalue problems. Preprint 1255, Dept. Math., University Utrecht, Utrecht, the Netherlands, September 2002. See Section 2.7.2 of this thesis.

[2] Arnoldi, W. E. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math. 9* (1951), 17–29.

[3] Atkinson, F. V. Multiparameter spectral theory. *Bull. Amer. Math. Soc. 74* (1968), 1–27.

[4] Atkinson, F. V. *Multiparameter Eigenvalue Problems.* Academic Press, New York, 1972.

[5] Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., and van der Vorst, H., Eds. *Templates for the Solution of Algebraic Eigenvalue Problems.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[6] Binding, P. A., Browne, P. J., and Ji, X. Z. A numerical method using the Prüfer transformation for the calculation of eigenpairs of two-parameter Sturm-Liouville problems. *IMA J. Numer. Anal. 13*, 4 (1993), 559–569.

[7] Björck, Å. *Numerical methods for least squares problems.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[8] Blum, E. K., and Chang, A. F. A numerical method for the solution of the double eigenvalue problem. *J. Inst. Math. Appl. 22*, 1 (1978), 29–42.

[9] Blum, E. K., and Curtis, A. R. A convergent gradient method for matrix eigenvector-eigentuple problems. *Numer. Math. 31*, 3 (1978/79), 247–263.

[10] Blum, E. K., and Geltner, P. B. Numerical solution of eigentuple-eigenvector problems in Hilbert spaces by a gradient method. *Numer. Math. 31*, 3 (1978/79), 231–246.

[11] Bohte, Z. Numerical solution of some two-parameter eigenvalue problems. *Slov. Acad. Sci. Art. Anton Kuhelj Memorial Volume* (1982), 17–28.

[12] Braconnier, T., and Higham, N. J. Computing the field of values and pseudospectra using the Lanczos method with continuation. *BIT 36*, 3 (1996), 422–440.

[13] Browne, P. J., and Sleeman, B. D. A numerical technique for multiparameter eigenvalue problems. *IMA J. Numer. Anal. 2*, 4 (1982), 451–457.

[14] Bunse-Gerstner, A., Byers, R., and Mehrmann, V. Numerical methods for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl. 14*, 4 (1993), 927–949.

[15] Bunse-Gerstner, A., Byers, R., Mehrmann, V., and Nichols, N. K. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numer. Math. 60*, 1 (1991), 1–39.

[16] Bunse-Gerstner, A., and Stöver, R. On a conjugate gradient-type method for solving complex symmetric linear systems. *Linear Algebra Appl. 287(1–3)* (1999), 105–123.

[17] Cullum, J., Willoughby, R. A., and Lake, M. A Lanczos algorithm for computing singular values and vectors of large matrices. *SIAM J. Sci. Stat. Comput. 4*, 2 (1983), 197–215.

[18] Cullum, J. K., and Willoughby, R. A. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Volume 1, Theory.* Birkhäuser, Boston, 1985.

[19] Davidson, E. R. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. *J. Comput. Phys. 17* (1975), 87–94.

[20] Dongarra, J. J. Improving the accuracy of computed singular values. *SIAM J. Sci. Stat. Comput. 4*, 4 (1983), 712–719.

[21] Faierman, M. *Two-parameter Eigenvalue Problems in Ordinary Differential Equations.* Longman Scientific & Technical, Harlow, 1991.

[22] Fokkema, D. R., Sleijpen, G. L. G., and van der Vorst, H. A. Accelerated inexact Newton schemes for large systems of nonlinear equations. *SIAM J. Sci. Comput. 19*, 2 (1998), 657–674.

[23] Fokkema, D. R., Sleijpen, G. L. G., and van der Vorst, H. A. Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils. *SIAM J. Sci. Comput. 20*, 1 (1999), 94–125.

[24] Frayssé, V., Gueury, M., Nicoud, F., and Toumazou, V. Spectral portraits for matrix pencils. Preprint Tech. Report TR/PA/96/19, CERFACS, Toulouse, France, 1996.

[25] Frayssé, V., and Toumazou, V. A note on the normwise perturbation theory for the regular generalized eigenproblem. *Numer. Linear Algebra Appl. 5*, 1 (1998), 1–10.

[26] Freund, R. W. Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices. *SIAM J. Sci. Stat. Comput. 13*, 1 (1992), 425–448.

[27] Gantmacher, F. R. *The theory of matrices. Vols. 1, 2.* Chelsea Publishing Co., New York, 1959.

[28] Golub, G. H., and Kahan, W. Calculating the singular values and pseudo-inverse of a matrix. *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. 2* (1965), 205–224.

[29] Golub, G. H., Luk, F. T., and Overton, M. L. A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. *ACM Trans. Math. Software 7*, 2 (1981), 149–169.

[30] Golub, G. H., and van der Vorst, H. A. Eigenvalue computation in the 20th century. *J. Comput. Appl. Math. 123(1–2)* (2000), 35–65. Numerical analysis 2000, Vol. III. Linear algebra.

[31] Golub, G. H., and Van Loan, C. F. *Matrix Computations*, 3rd ed. The John Hopkins University Press, Baltimore, London, 1996.

[32] Higham, D. J., and Higham, N. J. Structured backward error and condition of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl. 20*, 2 (1999), 493–512.

[33] Hochstenbach, M. E. A Jacobi–Davidson type SVD method. *SIAM J. Sci. Comput. 23*, 2 (2001), 606–628. See Chapter 3 of this thesis.

[34] Hochstenbach, M. E. Harmonic and refined extraction methods for the singular value problem, with applications in least squares problems. Preprint 1263, Dept. Math., University Utrecht, Utrecht, the Netherlands, December 2002. See Chapter 4 of this thesis.

[35] Hochstenbach, M. E., Košir, T., and Plestenjak, B. A Jacobi–Davidson type method for general two-parameter eigenvalue problem. Preprint 1262, Dept. Math., University Utrecht, Utrecht, the Netherlands, November 2002. See Chapter 6 of this thesis.

[36] Hochstenbach, M. E., and Plestenjak, B. A Jacobi–Davidson type method for a right definite two-parameter eigenvalue problem. *SIAM J. Matrix Anal. Appl. 24*, 2 (2002), 392–410. See Chapter 5 of this thesis.

[37] Hochstenbach, M. E., and Plestenjak, B. Backward error, condition numbers, and pseudospectrum for the multiparameter eigenvalue problem. Preprint 1225, Dept. Math., University Utrecht, Utrecht, the Netherlands, February 2002. Submitted. See Chapter 7 of this thesis.

[38] Hochstenbach, M. E., and Sleijpen, G. L. G. Two-sided and alternating Jacobi–Davidson. *Linear Algebra Appl. 358(1-3)* (2003), 145–172. See Chapter 2 of this thesis.

[39] Hochstenbach, M. E., and van der Vorst, H. A. Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem. Preprint 1212, Dept. Math., University Utrecht, Utrecht, the Netherlands, November 2001. See Chapter 8 of this thesis.

[40] Horn, R. A., and Johnson, C. R. *Matrix Analysis.* Cambridge University Press, Cambridge, UK, 1985.

[41] Horn, R. A., and Johnson, C. R. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1991.

[42] Ji, X. Z. Numerical solution of joint eigenpairs of a family of commutative matrices. *Appl. Math. Lett. 4*, 3 (1991), 57–60.

[43] Jia, Z. Refined iterative algorithms based on Arnoldi's process for large unsymmetric eigenproblems. *Linear Algebra Appl. 259* (1997), 1–23.

[44] Kaniel, S. Estimates for some computational techniques in linear algebra. *Math. Comp. 20* (1966), 369–378.

[45] Kato, T. Estimation of iterated matrices, with application to the von Neumann condition. *Numer. Math. 2* (1960), 22–29.

[46] Kato, T. *Perturbation theory for linear operators*, second ed. Springer-Verlag, Berlin, 1976. Grundlehren der Mathematischen Wissenschaften, Band 132.

[47] Košir, T. Finite-dimensional multiparameter spectral theory: the nonderogatory case. In *Proceedings of the 3rd ILAS Conference (Pensacola, FL, 1993)* (1994), vol. 212/213, pp. 45–70.

[48] Kuczyński, J., and Woźniakowski, H. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl. 13*, 4 (1992), 1094–1122.

[49] Kuczyński, J., and Woźniakowski, H. Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm. *SIAM J. Matrix Anal. Appl. 15*, 2 (1994), 672–691.

[50] Lancaster, P. A generalized Rayleigh quotient iteration for lambda-matrices. *Arch. Rational Mech. Anal. 8* (1961), 309–322.

[51] Lanczos, C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand. 45*, 4 (1950), 255–282.

[52] Larsen, R. M. Lanczos bidiagonalization with partial reorthogonalization. Technical report DAIMI PB-357, Department of Computer Science, University of Aarhus, September 1998. See also `http://sun.stanford.edu/~rmunk/PROPACK`.

[53] The Matrix Market. `http://math.nist.gov/MatrixMarket`, a repository for test matrices.

[54] Morgan, R. B. Computing interior eigenvalues of large matrices. *Linear Algebra Appl. 154/156* (1991), 289–309.

[55] Ostrowski, A. M. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. III. (Generalized Rayleigh quotient characteristic roots with linear elementary divisors). *Arch. Rational Mech. Anal. 3* (1959), 325–340.

[56] Ostrowski, A. M. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. V. (Usual Rayleigh quotient for non-Hermitian matrices and linear elementary divisors). *Arch. Rational Mech. Anal. 3* (1959), 472–481.

[57] Paige, C. C. *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices.* Ph.D. thesis, London University, London, England, 1971.

[58] Paige, C. C., Parlett, B. N., and van der Vorst, H. A. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Num. Lin. Alg. Appl. 2*, 2 (1995), 115–133.

[59] Paige, C. C., and Saunders, M. A. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software 8*, 1 (1982), 43–71.

[60] Parlett, B. N. The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Math. Comp. 28* (1974), 679–693.

[61] Parlett, B. N. *The Symmetric Eigenvalue Problem.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[62] Parlett, B. N., Simon, H., and Stringer, L. M. On estimating the largest eigenvalue with the Lanczos algorithm. *Math. Comp. 38*, 157 (1982), 153–165.

[63] Philippe, B., and Sadkane, M. Computation of the fundamental singular subspace of a large matrix. *Linear Alg. Appl. 257* (1997), 77–104.

[64] Plestenjak, B. A continuation method for a right definite two-parameter eigenvalue problem. *SIAM J. Matrix Anal. Appl. 21*, 4 (2000), 1163–1184.

[65] Plestenjak, B. A continuation method for a weakly elliptic two-parameter eigenvalue problem. *IMA J. Numer. Anal. 21*, 1 (2001), 199–216.

[66] `http://web.comlab.ox.ac.uk/projects/pseudospectra/`, the Pseudospectum Gateway.

[67] Rivlin, T. J. *Chebyshev Polynomials*, second ed. John Wiley & Sons Inc., New York, 1990.

[68] Saad, Y. On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM J. Numer. Anal. 17*, 5 (1980), 687–706.

[69] Saad, Y. *Numerical Methods for Large Eigenvalue Problems.* Manchester University Press, Manchester, UK, 1992.

[70] Shimasaki, M. A homotopy algorithm for two-parameter eigenvalue problems. *Zeitschrift Fur Angewandte Mathematik und Mechanik 76* (1996), 675–676, Suppl. 2.

[71] Simon, H. D., and Zha, H. Low-rank matrix approximation using the Lanczos bidiagonalization process with applications. *SIAM J. Sci. Comput. 21*, 6 (2000), 2257–2274.

[72] Simoncini, V., and Eldén, L. Inexact Rayleigh quotient-type methods for eigenvalue computations. *BIT 42*, 2 (2002), 159–182.

[73] Sleijpen, G. L. G., Booten, A. G. L., Fokkema, D. R., and van der Vorst, H. A. Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT 36*, 3 (1996), 595–633.

[74] Sleijpen, G. L. G., and van den Eshof, J. On the use of harmonic Ritz pairs in approximating internal eigenpairs. *Linear Algebra Appl. 358(1–3)* (2003), 115–137.

[75] Sleijpen, G. L. G., and van der Vorst, H. A. A Jacobi–Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl. 17*, 2 (1996), 401–425.

[76] Sleijpen, G. L. G., and van der Vorst, H. A. The Jacobi–Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes. In *Iterative Methods in Linear Algebra, II.* (New Brunswick, NJ, U.S.A., 1996), S. D. Margenov and P. S. Vassilevski, Eds., vol. 3 of *IMACS Series in Computational and Applied Mathematics*, IMACS, pp. 377–389.

[77] Sleijpen, G. L. G., and van der Vorst, H. A. A Jacobi–Davidson iteration method for linear eigenvalue problems. *SIAM Review 42*, 2 (2000), 267–293.

[78] Sleijpen, G. L. G., van der Vorst, H. A., and Meijerink, E. Efficient expansion of subspaces in the Jacobi–Davidson method for standard and generalized eigenproblems. *Electron. Trans. Numer. Anal. 7* (1998), 75–89.

[79] Slivnik, T., and Tomšič, G. A numerical method for the solution of two-parameter eigenvalue problems. *J. Comput. Appl. Math. 15*, 1 (1986), 109–115.

[80] Smit, P., and Paardekooper, M. H. C. The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra Appl. 287(1–3)* (1999), 337–357.

[81] Stathopoulos, A., and Saad, Y. Restarting techniques for the (Jacobi–)Davidson symmetric eigenvalue methods. *Electron. Trans. Numer. Anal. 7* (1998), 163–181.

[82] Stewart, G. W. *Matrix algorithms. Vol. II.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[83] Stewart, G. W., and Sun, J. G. *Matrix perturbation theory.* Academic Press Inc., Boston, MA, 1990.

[84] Tisseur, F. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl. 309(1–3)* (2000), 339–361.

[85] Tisseur, F., and Higham, N. J. Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM J. Matrix Anal. Appl. 23*, 1 (2001), 187–208.

[86] Tisseur, F., and Meerbergen, K. The quadratic eigenvalue problem. *SIAM Rev. 43*, 2 (2001), 235–286.

[87] Trefethen, L. N. Computation of pseudospectra. In *Acta numerica, 1999.* Cambridge Univ. Press, Cambridge, 1999, pp. 247–295.

[88] Trefethen, L. N. Spectra and pseudospectra. In *The Graduate Student's Guide to Numerical Analysis '98*, M. Ainsworth, J. Levesley, and M. Marletta, Eds. Springer, Berlin, 1999, pp. 217–250.

[89] van den Eshof, J. The convergence of Jacobi–Davidson iterations for Hermitian eigenproblems. *Numer. Linear Algebra Appl. 9*, 2 (2002), 163–179.

[90] van der Sluis, A., and van der Vorst, H. A. The convergence behavior of Ritz values in the presence of close eigenvalues. *Linear Algebra Appl. 88/89* (1987), 651–694.

[91] van der Vorst, H. A. Computational methods for large eigenvalue problems. In *Handbook of Numerical Analysis*, P. G. Ciarlet and J. L. Lions, Eds., vol. VIII. North-Holland, Amsterdam, 2002, pp. 3–179.

[92] van der Vorst, H. A., and Golub, G. H. 150 years old and still alive: eigenproblems. In *The state of the art in numerical analysis (York, 1996)*, vol. 63 of *Inst. Math. Appl. Conf. Ser. New Ser.* Oxford Univ. Press, New York, 1997, pp. 93–119.

[93] VAN DER VORST, H. A., AND MELISSEN, J. A Petrov–Galerkin type method for solving $Ax = b$, where a is a symmetric complex matrix. *IEEE Trans. on Magnetics 26*, 2 (1990), 706–708.

[94] VAN DOOREN, P. A generalized eigenvalue approach for solving Riccati equations. *SIAM J. Sci. Statist. Comput. 2*, 2 (1981), 121–135.

[95] VAN DORSSELAER, J. L. M., HOCHSTENBACH, M. E., AND VAN DER VORST, H. A. Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method. *SIAM J. Matrix Anal. Appl. 22*, 3 (2000), 837–852. See Chapter 9 of this thesis.

[96] VAN HUFFEL, S. Iterative algorithms for computing the singular subspace of a matrix associated with its smallest singular values. *Linear Algebra Appl. 154/156* (1991), 675–709.

[97] VARADHAN, S., BERRY, M. W., AND GOLUB, G. H. Approximating dominant singular triplets of large sparse matrices via modified moments. *Numer. Algorithms 13(1–2)* (1996), 123–152.

[98] VARGA, R. S. *Matrix Iterative Analysis*, expanded ed. Springer-Verlag, Berlin, 2000.

[99] VERSCHELDE, J. Algorithm 795: PHCPACK: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Transactions on Mathematical Software 25*, 2 (1999), 251–276.

[100] VOLKMER, H. *Multiparameter Eigenvalue Problems and Expansion Theorems.* Springer-Verlag, Berlin, 1988.

[101] WILKINSON, J. H. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

[102] WISE, S. M., SOMMESE, A. J., AND WATSON, L. T. Algorithm 801: POLSYS_PLP: A partitioned linear product homotopy code for solving polynomial systems of equations. *ACM Transactions on Mathematical Software 26*, 1 (2000), 176–200.

Index

Notations

notation	meaning
$\mathbb{R},\ \mathbb{C},\ \mathbb{R}^n,\ \mathbb{C}^n$	real, complex numbers, n-space
$\overline{\alpha},\ \overline{u}$	complex conjugate of number α, vector u
$A^T,\ A^*$	transpose, conjugate transpose of A
$A^{-1},\ A^+,\ A^{-T}$	inverse, pseudoinverse, inverse of transpose of A
$u^T v,\ u^* v$	real, complex vector dot product
$u \perp v$	vector u is orthogonal to vector v
$\angle(u,v),\ \angle(\mathcal{U},v)$	angle between vector u (resp. subspace $\mathcal{U}$) and v
$\mathrm{span}(a_1,\ldots,a_k),\ \mathrm{span}(A)$	spanning space of vectors $a_1,\ldots,a_k$, resp. of columns of A
$\ker(A),\ \mathrm{tr}(A)$	kernel, trace of A
$\mathrm{vec}(A)$	$[A_1^T\ \cdots\ A_n^T]^T$, where the A_i are the columns of A
$\mathrm{diag}(\alpha_1,\ldots,\alpha_k),\ \mathrm{diag}(1:n)$	diagonal matrix from scalars $\alpha_1,\ldots,\alpha_k$; $1,\ldots,n$
$\mathrm{tridiag}(\alpha,\beta,\gamma)$	tridiagonal matrix with stencil $(\alpha,\ \beta,\ \gamma)$
$\mathrm{rand}(m,n)$	$m \times n$ matrix, random entries, uniformly distributed on (0,1)
$\|\cdot\| = \|\cdot\|_2$	Euclidean (or spectral or 2-) norm
$\|\cdot\|_F,\ \|\cdot\|_1,\ \|\cdot\|_\infty$	Frobenius, 1-, inf-norm
$\Lambda(A),\ \Lambda_\varepsilon(A)$	spectrum, ε-pseudospectrum of A
$\lambda_j(A),\ \lambda_{\max}(A),\ \lambda_{\min}(A)$	jth *smallest*, largest, smallest eigenvalue of a Hermitian A
$\Sigma(A)$	singular values of A
$\sigma_j(A),\ \sigma_{\max}(A),\ \sigma_{\min}(A)$	jth *largest*, largest, smallest singular value of A
$\mathcal{K}_k(A,v)$	Krylov subspace of dimension k: $\mathrm{span}\{v, Av, \ldots, A^{k-1}v\}$
$\kappa(\lambda)$	condition number of eigenvalue λ
$\kappa(A)$	condition number of A: $\kappa(A) = \|A\| \cdot \|A^{-1}\|$
$\kappa_e(A)$	effective condition number of A: $\kappa_e(A) = \|A\| \cdot \|A^+\|$
$\mathrm{Re}(\zeta),\ \mathrm{Im}(\zeta)$	real, imaginary part of number or matrix
$\det(A)$	determinant of A
$Df(u),\ \frac{\partial f}{\partial u}(u)$	total, partial derivative of f
$B(\alpha,\beta)$	Euler's beta function
MGS	numerically stable Gram–Schmidt orthogonalization
$\mathrm{nnz}(A)$	number of non-zeros of the matrix A
MV	matrix-vector product
[ ]	empty matrix
$\sum,\ \prod$	sum, product
$\bigcup,\ \bigcap$	union, intersection
$\otimes,\ \oplus$	tensor product, direct sum
$\lesssim$	less than or approximately equal to
$\mathcal{O}(\cdot)$	order ("big-oh")
h.o.t.	higher order terms
$\square$	end of proof
$\oslash$	end of definition, remark, example, or experiment

Samenvatting

Dit proefschrift behandelt een aantal aspecten van deelruimte methoden voor verschillende eigenwaarde problemen. Trillingen en de bijbehorende eigenwaarden (of frequenties) komen voor in de wetenschap, techniek en het dagelijks leven. Eigenwaarde problemen van matrices zijn afkomstig uit een groot aantal gebieden, zoals

- de chemie (chemische reacties, energienivo's van een molecuul),
- de mechanica (ontwerp van aardbeving bestendige gebouwen)
- dynamische systemen (stabiliteit, bifurcatie analyse van systemen die afhangen van een parameter),
- Markov ketens (stationaire verdeling van random processen),
- magneto-hydrodynamica,
- de oceanografie,
- de economie,
- de signaal- en beeldverwerking,
- de control theorie,
- de patroonherkenning,
- en de statistiek.

Eigenwaarden en eigenvectoren geven waardevolle informatie over het gedrag en de eigenschappen van een matrix; daarom is het niet verbazendwekkend dat eigenwaarde problemen al meer dan anderhalve eeuw onderwerp van studie zijn, deels voordat de huidige matrix notatie standaard werd. Afhankelijk van de toepassing, is men geïnteresseerd in een of meerdere eigenwaarden aan het eind van het spectrum, of juist in eigenwaarden in het midden van het spectrum of in het aantal eigenwaarden in een interval.

Methoden voor eigenwaarde problemen worden vaak ingedeeld in twee categorieën. De eerste categorie, de *directe methoden* zoals de QR-methode en de verdeel-en-heers methode, hebben als doel alle eigenwaarden (nauwkeurig) te vinden van relatief kleine (zeg orde 10^3) matrices. Hoewel deze aanpakken op een iteratieve manier werken, worden ze "direct" genoemd, omdat ze (bijna) gegarandeerd in een vast aantal stappen convergeren. Deze methodes zijn efficiënt, en de onderliggende wiskunde is goed ontwikkeld.

Veel toepassingen, bijvoorbeeld die uit de chemie, geven echter aanleiding tot eigenwaarde problemen waar de afmeting van de matrix gemakkelijk een miljoen overschrijdt. Deze problemen komen dikwijls van gediscretiseerde partiële differentiaalvergelijkingen; meestal is slechts een kleine deel van de eigenwaarden interessant. Bovendien zijn de

matrices vaak *ijl*, dit betekent dat de matrix relatief veel elementen bevat die nul zijn. Daarom kan men goedkoop, dat wil zeggen snel, een matrix-vector product berekenen, ook voor grote matrices. Voor deze matrices zijn de directe aanpakken vaak niet mogelijk omdat ze teveel computer tijd en/of geheugen consumeren, zelfs op moderne (en toekomstige) computers. Om al deze redenen verdienen *iteratieve methoden*, en in het bijzonder de belangrijke deelklasse van *deelruimte methoden*, vaak de voorkeur voor grote ijle matrices. In een deelruimte methode wordt de matrix geprojecteerd op een laag-dimensionale deelruimte; de geprojecteerde matrix wordt dan opgelost met behulp van directe methoden. Op deze manier krijgen we benaderingen voor eigenparen uit een laag-dimensionale deelruimte.

Voor grote ijle problemen bestaat vaak niet "*de* beste methode". De keuze voor een methode kan afhangen van bepaalde eigenschappen van de matrix (structuur, afmeting), de data die gevraagd wordt (wat, met welke nauwkeurigheid) en de beschikbare operaties (getransponeerde van de matrix, preconditioneerder). Dit proefschrift hoopt een bijdrage te leveren aan het interessante en actieve gebied van deelruimte methoden voor eigenwaarde problemen. We bestuderen verschillende eigenwaarde problemen, te weten

- het (standaard) eigenwaarde probleem,
- het gegeneraliseerde eigenwaarde probleem,
- het singuliere waarde probleem,
- het polynomiale eigenwaarde probleem
- en het multiparameter eigenwaarde probleem.

Van deze problemen zijn het standaard en het gegeneraliseerde eigenwaarde probleem, afkomstig uit talrijke toepassingen, het meest bekend. Het singuliere waarde probleem speelt een belangrijke rol in toepassingen als signaal- en beeldverwerking, control theorie, patroonherkenning, statistiek en zoekmachines op het internet. Maar het heeft ook een centrale positie in de numerieke lineaire algebra zelf, bijvoorbeeld voor het kleinste kwadraten probleem, de numerieke rang van een matrix, de hoeken tussen deelruimtes, de gevoeligheid (conditie) van de oplossing van lineaire systemen, het pseudospectrum en de (Euclidische) norm van een matrix.

Het polynomiale eigenwaarde probleem komt onder andere voort uit de studie van trillingen van een mechanisch systeem veroorzaakt door een externe kracht (het effect van de wind op een brug), bij het simuleren van electronische circuits en in de vloeistof mechanica.

Een voorbeeld van de oorsprong van het multiparameter eigenwaarde probleem is de mathematische fysica, wanneer scheiding van variabelen wordt gebruikt om randwaarde problemen op te lossen.

Een deel van dit proefschrift wordt gevormd door vier hoofdstukken die Jacobi–Davidson achtige methoden introduceren voor verschillende eigenwaarde problemen:

- voor het (niet-normale) standaard, complex symmetrische, gegeneraliseerde en polynomiale eigenwaarde probleem in hoofdstuk 2,

- voor het singuliere waarde probleem in hoofdstuk 3 (met hoofdstuk 4 als vervolg),

- en voor het multiparameter eigenwaarde probleem (in het bijzonder het geval van twee parameters) in hoofdstuk 5 en 6.

Om te beginnen bestudeert hoofdstuk 2 twee Jacobi–Davidson achtige methoden voor *niet-normale* matrices, die we *tweezijdig* en *alternerend Jacobi–Davidson* noemen. Voor deze matrices zijn de rechts en links eigenvectoren in het algemeen niet identiek, zoals het geval is voor normale matrices. Dit vormt een motivatie om twee zoekruimtes bij te houden, een voor de rechts en een voor de links eigenvector. De zoekruimte voor de linker vector is de testruimte voor de rechtse vector en vice versa. De correctievergelijking, die dient voor de expansie van de zoekruimtes, bevat scheve projecties, in plaats van de orthogonale projecties die kenmerkend zijn voor de standaard Jacobi–Davidson methode. Deze methoden worden toegepast op het standaard, complex symmetrische, gegeneraliseerde en polynomiale eigenwaarde probleem.

Hoofdstuk 3 introduceert een Jacobi–Davidson achtige methode voor het *singuliere waarde probleem*. Net als in hoofdstuk 2 hebben we twee zoekruimtes, nu een voor de rechts en een voor de links singuliere vector. Dit geeft aanleiding tot een methode met cubische convergentie wanneer de correctievergelijking exact wordt opgelost. In de praktijk zal deze vergelijking vaak inexact worden opgelost, met lineaire convergentie als resultaat. De methode kan gezien worden als een versneld inexact Newton proces en als een versnelde inexacte Rayleigh quotient iteratie. Speciale aandacht wordt in hoofdstuk 4 gegeven aan het benaderen van de *kleinste* en *inwendige singuliere waarden*. Hierbij is de standaard Galerkin deelruimte extractie niet meer bevredigend. Net als bij het standaard eigenwaarde probleem zijn een harmonische en "verfijnde" aanpak meer belovend. We bespreken ook toepassingen van de methode op het kleinste kwadraten probleem en de benadering van een matrix door middel van een afgekapte singuliere waarde ontbinding.

Hoofdstuk 5 en 6 behandelen een Jacobi–Davidson achtige methode voor het *multiparameter eigenwaarde probleem*, in het bijzonder het geval van twee parameters. In hoofdstuk 5 bekijken we het zogenaamde *rechts-definiete* multiparameter eigenwaarde probleem. Voor het geval van twee parameters hebben we wederom twee zoekruimtes, een voor elke component van de ontbindbare tensor. De extractie van de zoekruimte gebeurt met een generalisatie van de Rayleigh–Ritz methode, die monotone convergentie naar de extreme eigenwaarden verzekert. Voor de uitbreiding van de zoekruimtes presenteren we twee verschillende correctievergelijkingen: een met orthogonale een-dimensionale projecties die tweedegraads termen verwaarloost, en een met twee-dimensionale scheve projecties die alleen derdegraads termen weggooit. Omdat standaard deflatietechnieken niet opgaan in dit probleem, wordt een selectiecriterium op de Ritzwaarden toegepast wanneer we geïnteresseerd zijn in meerdere eigenparen.

In hoofdstuk 6 behandelen we de wijdere klasse van de *niet-singuliere* multiparameter eigenwaarde problemen. Dit is een uitdagend probleem, waar we vele technieken nodig hebben om het te kraken. Zo kiezen we hier voor een tweezijdige aanpak (verschillende test- en zoekruimtes), vergelijkbaar met hoofdstuk 2.

Hoofdstuk 7 bekijkt numeriek belangrijke aspecten van het multiparameter probleem: *terugwaartse fouten* en de *conditie* van eigenwaarden en eigenvectoren. Deze begrippen geven een indicatie hoe goed een bepaalde verkregen benadering is, en hoe gevoelig de

eigenwaarden en eigenvectoren voor perturbaties van het probleem zijn. Ook wordt het *pseudospectrum* voor multiparameter problemen geïntroduceerd. Dit kan een fraai grafisch beeld geven van de gevoeligheid van een aantal of alle eigenwaarden.

Voor het standaard eigenwaarde probleem is de extractie van Ritzparen van een zoekruimte al goed onderzocht. Voor het *polynomiale eigenwaarde probleem* is de situatie minder duidelijk. Hoofdstuk 8 beschouwt benaderingen van een eigenwaarde die verkregen kunnen worden uit een zoekruimte. De nadruk ligt op het quadratisch eigenwaarde probleem en een-dimensionale zoekruimtes. Er worden drie nieuwe methoden gegeven, gebaseerd op een Galerkin- of minimum residu-aanpak. De methoden worden met behulp van perturbatie resultaten en terugwaartse fouten vergeleken, en vervolgens gegeneraliseerd naar algemene polynomiale problemen en extractie van meer-dimensionale zoekruimtes.

In hoofdstuk 9 ontwikkelen we *probabilistische grenzen* voor de extreme eigenwaarden van een Hermietse matrix met behulp van de Lanczos methode. Deze grenzen worden verkregen met Lanczos-, Ritz- en Chebyshevpolynomen. Omdat we er vanuit gaan dat de startvector een bepaalde component in de richting van de gezocht eigenrichting bevat, verkrijgen we zo grenzen die met een bepaalde (grote) waarschijnlijkheid inderdaad juist zijn. De grenzen kunnen gebruikt worden als stopcriterium. Een tweede toepassing van de technieken is het maken van een schatting voor het aantal stappen van de Lanczos methode die nog nodig zijn om een extreme eigenwaarde met een bepaalde tolerantie te verkrijgen.

Dankwoord / Acknowledgments

Allereerst wil ik Henk van der Vorst en Gerard Sleijpen hartelijk bedanken voor de leuke begeleiding. Ze maakten altijd tijd voor me en gaven me veel vrijheid; ik heb veel van ze geleerd. Als promovendus is het ideaal om deel uit te maken van een topgroep zoals de numerieke wiskunde in Utrecht.

Het NWO via Wim Aspers, en het Mathematisch Instituut waren prima werkgevers met veel aandacht voor "de mens achter de werknemer". Daarnaast was het leuk om mijn vroegere docenten in Utrecht nu als collega te leren kennen. Met mijn collega-AIO's, en numerici in het bijzonder, heb ik genoten van vele interessante praatjes en informele bijeenkomsten.

Geertje Hek, die ik ook al kende van onze studietijd, was mijn gezellige kamergenoot tijdens de eerste twee jaren. Samen hebben we vele interessante gesprekken gevoerd. Toen zij promoveerde werd Jasper van den Eshof de nieuwe bewoner van kamer 803. Naast de gezelligheid hebben we ook vele vakinhoudelijke discussies gevoerd. Ook wil ik Geertje en Jasper bedanken voor het geduldig aanhoren van vele uren (voornamelijk Mozart-)muziek.

I enjoyed the hospitality of Prof. Hubert Schwetlick (Dresden), Prof.Marlis Hochbruck (Düsseldorf) and Dr. Bor Plestenjak (Ljubljana) very much, and their ideas were very inspiring for me. I also would like to thank Bor for the very pleasant and fruitful cooperation that we had during the last two years. It was a nice coincidence that we both became a father in the same period. I also had an interesting cooperation with Dr. Peter Arbenz (Zürich), Dr. Jos van Dorsselaer (Utrecht), and Dr. Tomaž Košir (Ljubljana). I thank Prof. Wil Schilders, Prof. Valeria Simoncini, Prof. Ferdinand Verhulst, and Dr. Gerard Sleijpen for their diligent reading of the thesis and their constructive comments. The referees of the papers have also given many to-the-point suggestions.

Naast mijn werk heb ik genoten van vele contacten en hobbies. Schaken in het eerste van Paul Keres was altijd weer een uitje. In de Nieuwe Kerk heb ik me als een vis in het water gevoeld. Met mijn huisgenoten Willem, Hans, Jan-Willem, Gert-Jan en Emma heb ik vaak gegeten en bordspelen en toto's gedaan.

Ferrie van de Oudeweetering, mijn wiskunde-leraar van het VWO, heeft me verder geholpen in het plezier krijgen in de wiskunde. Jan Hogendijk heeft me aangespoord AIO te worden. Michaël Brouwers en José Lieshout hebben me enorm geholpen met hun deskundigheid.

Uiteraard is mijn familie het belangrijkst geweest. Mijn ouders en Bas bedank ik voor mijn leuke, positieve jeugd, mijn schoonfamilie voor hun belangstelling, en Ineke voor haar liefde en warmte. Joël en Mirjam hebben ons beiden een enorme vreugde in ons leven gebracht. Tot slot bedank ik mijn hemelse Vader voor zijn liefde die alle verstand te boven gaat.

List of publications

List of publications Michiel E. Hochstenbach

Numerical Linear Algebra publications

1. J.L.M. van Dorsselaer, M.E. Hochstenbach, H.A. van der Vorst, *Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method,* SIAM J. on Matrix Anal. Appl. 22(3), pp. 837–852, 2000.

2. M.E. Hochstenbach, *A Jacobi–Davidson type SVD method,* SIAM J. on Sci. Comp. 23(2), pp. 606–628, 2001. Second place student paper competition 6th Copper Mountain Conference 2000.

3. M.E. Hochstenbach, B. Plestenjak, *A Jacobi–Davidson type method for a right definite two-parameter eigenvalue problem,* SIAM J. on Matrix Anal. Appl. 24(2), pp. 392–410, 2002. Winner of poster competition at IWASEP 4.

4. M.E. Hochstenbach, G.L.G. Sleijpen, *Two-sided and alternating Jacobi–Davidson,* Lin. Alg. Appl. 358(1-3), pp. 145–172, 2003.

5. M.E. Hochstenbach, H.A. van der Vorst, *Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem,* Preprint 1212, Dept. of Math., Utrecht University, November 2001. Accepted for publication in SIAM J. on Sci. Comp. Winner student/new PhD paper competition 7th Copper Mountain Conference 2002.

6. M.E. Hochstenbach, B. Plestenjak, *Backward error, condition and pseudospectra for the multiparameter eigenvalue problem,* Preprint 1225, Dept. of Math., Utrecht University, February 2002. Accepted for publication in Lin. Alg. Appl.

7. P. Arbenz, M.E. Hochstenbach, *A Jacobi–Davidson method for complex symmetric matrices,* Preprint 1255, Dept. of Math., Utrecht University, September 2002. Submitted.

8. M.E. Hochstenbach, B. Plestenjak, T. Košir, *A Jacobi–Davidson type method for the two-parameter eigenvalue problem,* Preprint 1262, Dept. of Math., Utrecht University, November 2002. Submitted.

9. M.E. Hochstenbach, *Harmonic and refined extraction methods for the singular value problem, with applications in least squares problems,* Preprint 1263, Dept. of Math., Utrecht University, December 2002. Submitted. Winner SIAM student paper competition 2003. Winner travel award student/new PhD paper competition 6th International Symposium on Iterative Methods in Scientific Computing.

Industrial / other publications

1. W.J. Grootjans, M.E. Hochstenbach, J. Hurink, W. Kern, M. Luczak, Q. Puite, J. Resing, F. Spieksma, *Cache as cash can,* In: Proceedings of the Thirty-sixth European Study Group with Industry (ESGI36), July 2000.

2. M.E. Hochstenbach, R.D. van der Meij, *Caching op het Internet* (Dutch), Vakidioot Utrecht University, July 2000.

3. R.D. van der Meij, M.E. Hochstenbach, Q. Puite, *World Wide Wait - ruimtelijke ordening op internet* (Dutch), Natuur & Techniek, February 2001.

4. E.Cahyono, M.E.Hochstenbach, J.Molenaar, W.Schilders, G.Terra, *Velocity estimation in mixtures using tomography,* In: Proceedings of the Thirty-ninth European Study Group with Industry (ESGI39), April 2001.

5. P. van Blokland, L. Booth, K. Hiremath, M.E. Hochstenbach, G. Koole, S. Pop, M. Quant, D. Wirosoetisno, *Euro diffusie* (Dutch), Vakidioot Utrecht University, July 2002.

6. P. van Blokland, L. Booth, K. Hiremath, M.E. Hochstenbach, G. Koole, S. Pop, M. Quant, D. Wirosoetisno, *The euro diffusion project,* In: Proceedings of the forty-second European study group with industry, CWI Syllabus 51, pp. 41–57, October 2002.

7. M.E. Hochstenbach, *Euro diffusie* (Dutch), Nieuw Archief voor de Wiskunde, March 2003.

Curriculum Vitae

Michiel Hochstenbach is op 24 maart 1973 geboren te Gouda. Hij behaalde in 1991 het VWO-diploma aan het Gymnasium Apeldoorn. Hierna ging hij wiskunde met bijvak informatica studeren aan de Universiteit Utrecht. In 1992 haalde Michiel cum laude propedeuses in wiskunde en informatica. In 1996 studeerde hij cum laude af bij Prof. Hans Duistermaat en Dr. Joop Kolk met de sciptie "De Hankel transformatie en generalisaties". Tijdens zijn studie was Michiel lid van het dagelijks bestuur van de vakgroep wiskunde, voorzitter van de studenten overleggroep, en was hij studentassistent bij werkcolleges.

Na van 1996 tot 1998 bij Logica B.V. te hebben gewerkt, werd Michiel onderzoeker in opleiding, en in 1999 assistent in opleiding bij prof. Henk van der Vorst in Utrecht. Het onderzoek wat hij deed resulteerde in dit proefschrift. In het kader van zijn onderzoek bezocht hij Prof. Hubert Schwetlick in Dresden, Prof. Marlis Hochbruck in Düsseldorf en Dr. Bor Plestenjak in Ljubljana en gaf voordrachten op diverse conferenties en universiteiten. Daarnaast gaf hij met veel plezier onderwijs.

In mei 2001 werd hij de gelukkige vader van Joël en in november 2002 van Mirjam. Vanaf april 2003 is Michiel werkzaam als postdoc aan de Heinrich Heine Universität te Düsseldorf en vanaf augustus 2003 als assistant professor aan de Case Western Reserve University in Cleveland, Ohio, USA.