BERGISCHE
UNIVERSITÄT
WUPPERTAL

Integration based solvers for standard and generalized Hermitian eigenvalue problems

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Naturwissenschaften der
Bergischen Universität Wuppertal vorgelegte und genehmigte

Dissertation

von

Lukas Krämer

Gutachter:	Prof. Dr. Bruno Lang
Gutachter:	Prof. Dr. Matthias Bolten
Gutachter:	Prof. Dr. Thomas Huckle

Dissertation eingereicht am:	30.01.2014
Tag der Disputation:	28.04.2014

Diese Dissertation kann wie folgt zitiert werden:

urn:nbn:de:hbz:468-20140701-112141-6
[http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hbz:468-20140701-112141-6]

Danksagung

Diese Arbeit entstand im Zeitraum von April 2009 bis Januar 2014. In dieser Zeit war ich als wissenschaftlicher Mitarbeiter in der Arbeitsgruppe ,,Angewandte Informatik“ im Fachbereich C – Mathematik und Naturwissenschaften der Bergischen Universität Wuppertal tätig. Allen Kolleginnen und Kollegen danke ich für die angenehme Arbeitsatmosphäre.

Zahlreiche Personen haben zum Gelingen der Promotion beigetragen. Als erstes möchte ich Prof. Dr. Bruno Lang danken. Er hat die Arbeit betreut und in unzähligen Diskussionen zum Gelingen beigetragen.

Ich danke Prof. Dr. Matthias Bolten für die Erstellung des Zweitgutachtens und Prof. Dr. Thomas Huckle von der TU München für die kurzfristige Bereitschaft, ein drittes Gutachten anzufertigen. Außerdem bedanke ich mich bei Prof. Dr. Andreas Frommer und Prof. Dr. Markus Reineke für die Mitwirkung bei der Prüfungskommission.

Mein Kollege Martin Galgon hat in vielen Diskussionen zum Verständnis der hier behandelten Verfahren beigetragen. Er hat mir bei vielen technischen Problemen geholfen und außerdem frühere Versionen der Arbeit Korrektur gelesen.

Ich bedanke mich bei Sebastian Meyer, Thomas Pawlaschyk, Matthias Rottmann und Sonja Sokolović dafür, dass sie Teile der Arbeit Korrektur gelesen haben. Mein ehemaliger Kollege Dr. Paul Willems hat mir eine sehr nützliche Latex-Vorlage gegeben.

Ich danke meiner Familie und vor allem Daniela für ihre Unterstützung.

Wuppertal, Juni 2014 Lukas Krämer

Abstract

This thesis is about the computation of eigenvalues and eigenvectors of large Hermitian matrices and of Hermitian/Hermitian positive definite matrix pairs. The core technique employed is numerical integration of the resolvent of the matrix (pair). It turns out that the problem of integrating the resolvent is equivalent to a certain approximation problem which can be solved in several ways. A number of contributions to theory of this class of algorithms are made, together with practical considerations. Furthermore, some results concerning the general theory of generalized eigenvalue problems are presented.

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Berechnung von Eigenwerten und Eigenvektoren von großen Hermiteschen Matrizen und von Hermitesch/Hermitesch positiv definiten Matrix-Paaren. Die zentrale Technik, die dabei zum Einsatz kommt, ist die numerische Integration der Resolvente der Matrix (bzw. des Matrix-Paares). Es stellt sich heraus, dass das Problem, die Resolvente zu integrieren, äquivalent zu einem gewissen Approximationsproblem ist. Es werden einige Beiträge zur Theorie dieser Algorithmenklasse geleistet, zusammen mit praktischen Betrachtungen. Außerdem werden einige Ergebnisse zur allgemeinen Theorie des verallgemeinerten Eigenwertproblemes vorgestellt.

◇

Die Arbeit des Autors wurde unterstützt vom Bundesministerium für Bildung und Forschung innerhalb des Projektes „ELPA – Hochskalierbare Eigenwert-Löser für Petaflop-Großanwendungen“, Förderkennzeichen 01IH08007B, sowie von der Deutschen Forschungsgemeinschaft im Rahmen des Schwerpunktprogrammes „Software for Exascale Computing“ (SPP 1648).

Contents

Motivation and outline

This work is about the solution of eigenvalue problems involving a Hermitian matrix A and a Hermitian positive definite matrix B. We aim at solving the problem of finding all solutions of

$$\mathsf{Ax} = \mathsf{Bx}\lambda,\ \lambda \in I_\lambda,$$

where $I_\lambda \subset \mathbb{R}$ is an interval. This interval is also allowed to contain the complete spectrum of the pair (A, B), in this case the problem is to find all eigenpairs of the pair. To the best of our knowledge, no method is available for computing large fractions of the eigensystem of very large, sparse matrix pairs at the moment.

We investigate a passably new method based on numerical integration and approximation, which has not yet been completely dissected and understood. This method, introduced by Polizzi [85] and named FEAST, computes a subspace approximating the eigenspace belonging to the eigenvalues in I_λ. The subspace is constructed using an approximate B-orthogonal projector onto the eigenspace. While the exact projector can be expressed using a Cauchy integral, an approximation to this projector can be computed using numerical integration of the Cauchy integral (however, this is not the only way as we will see).

When starting the research, we found that FEAST in "most cases" works well, probably better than many other iterative eigensolvers, e. g., such based on Krylov subspaces. However, we found that there are some shortcomings, a (mainly) test based analysis of the algorithm was published in [60].

This work is devoted to a further theoretical assessment of the method and to some algorithmical improvements, yielding e. g., faster convergence, results of higher accuracy or a better perspective for the use in connection with high performance computing.

Here is a list of our goals.

- Understand the general mechanics of an integration based eigensolver.
- Separate the integration from the Rayleigh–Ritz-process.
- Obtain theoretical error bounds for the approximation error in the eigenvalues and bounds for the angles between exact and computed eigenvectors,
- Give a convergence proof for an integration based eigensolver.
- Dissect the integration process and derive error bounds.
- Use the theoretical results for a more robust algorithm.
- Make clear the connection between numerical integration and approximation.
- Show perspectives for a solver that does not need any auxiliary sparse matrix routines (e. g., linear solvers) but only the matrix-vector product.

The focus is not so much on high performance computing but more on the numerical properties of the methods. However, the developed techniques could be the basis for new HPC methods.

Structure of this work

In Chapter 1, we start by introducing the required notions from linear algebra and fix notation. In particular we introduce our terminology concerning eigenvalues, eigenvectors and eigenspaces as well as angles between subspaces.

In Chapter 2, we show the theoretical foundations of eigenvalue solvers based on integration. We explain that such methods consist of (i) numerical integration and (ii) subspace extraction, and study the theoretical properties that give cause for the assumption that the method works in practice. Several general results on error bounds of eigenvalues and eigenvectors are carried over to the generalized eigenvalue problem. It turns out that the useful geometry for these results is that one induced by B. The convergence of the trapezoidal rule and the Gauß–Legendre rule for numerical integration is investigated (and proven).

In Chapter 3, we come to algorithmics. First, techniques for eigenvalue counting are introduced and tested. They can be used stand alone or as a preprocessor for the FEAST algorithm. Then, the numerical integration is examined from a different point of view, leading to approximation methods. Particular emphasis is put on the approximation by polynomials, which is extensively studied and tested. Next, a rather analytical method is explained for transforming integration regions, leading to better results in some cases. At the end of the chapter we treat some smaller questions that occurred when analyzing and testing the methods from this thesis. Besides, we present further numerical experiments.

An index of the most important key words can be found at the end of the thesis.

Chapter 1

Introduction

Synopsis

In this chapter, we introduce some basic notions from (numerical) linear algebra. It should be mostly self-contained, while of course a basic knowledge of numerical linear algebra is quite useful. Very good introductions to numerical linear algebra can be found in the books [21, 108] as well as in Golub and Van Loan's book [36], which has recently been newly edited [37]. Books dedicated especially to eigenvalue problems include Parlett [80], Stewart [100] and Wilkinson's early work [118].

The structure of this chapter is as follows. In Section 1.1 we fix our notation for matrices, vectors and norms and introduce projectors and the SVD. In Section 1.2 we introduce the quantities at the core of this work, eigenvalues and eigenvectors, as well as eigenspaces, which are sometimes better to handle than single vectors. Section 1.3 deals with angles between vectors and subspaces, as it turns out that angles are the right measure for assessing the accuracy of eigenvectors. The same is true for subspaces, where the Euclidean distance does not make any sense. Angles are introduced for standard Euclidean geometry as well as for the geometry induced by the positive definite matrix B.

The rest of the chapter, involving Sections 1.4 and 1.5, is devoted to different eigenproblems, an overview of eigenvalue solvers and on how to assess the quality of a numerical method for eigenvalue computations.

1.1 Basics: (Computational) Linear Algebra

1.1.1 Matrices and vectors

In this section, we recall basic notions from linear algebra and fix important parts of our notation concerning these notions.

Throughout this thesis, the symbols $\mathbb{R}^{n\times m}$, $\mathbb{C}^{n\times m}$ denote the spaces of real and complex $n \times m$ matrices, respectively. At the core of this thesis are square matrices with complex entries, i. e., from $\mathbb{C}^{n\times n}$. All matrices are named by capital letters A, B, ...,. The integer n denotes the size of the square matrix that is at the heart of our discussion, this matrix is called A. The entries of an arbitrary matrix M can be accessed via the parenthesis operator; the entry (i, j) then is given by $\mathsf{M}(i, j)$. The colon-notation denotes a range of indices, e. g., $1 : k$ means all consecutive indices from 1 to k.

The *transpose* M^T of a matrix is the matrix itself with rows and columns interchanged, i. e., $\mathsf{M}^\mathsf{T}(i, j) = \mathsf{M}(j, i)$ for all i, j. If $\mathsf{M} \in \mathbb{C}^{n\times m}$, we have $\mathsf{M}^\mathsf{T} \in \mathbb{C}^{m\times n}$. If the matrix is complex, the simple transpose is not a very useful operator and should be replaced by the Hermitian transpose M^H that also conjugates the entries of the matrix, i. e., $\mathsf{M}^\mathsf{H}(i, j) = \overline{\mathsf{M}(j, i)}$. In order to simplify notation, we will use the adjoint operator * for $^\mathsf{T}$ as well as for $^\mathsf{H}$. It has the wanted effect in the real as well as in the complex case. A *Hermitian* matrix is a matrix with $\mathsf{M}^\star = \mathsf{M}$. If M is real we say *symmetric* instead of Hermitian.

Vectors from some space $\mathbb{R}^n$ or $\mathbb{C}^n$ are denoted by a, b, ...,. All vectors are column vectors in the first place, the corresponding row vector is the transpose of the vector and hence denoted by x^T. Note that in the complex case the row vector differs from the complex adjoint vector $\mathsf{x}^\star = \bar{\mathsf{x}}^\mathsf{T}$. Vectors are also accessed via the parenthesis operator, $\mathsf{x}(j)$ is the j-th entry of the vector x. The zero vector is denoted by o.

A real square matrix M is said to be *symmetric positive definite* if it is symmetric and if further $\mathsf{x}^\star\mathsf{M}\mathsf{x} > 0$ for all $\mathsf{x} \neq \mathsf{o}$. Similarly, a Hermitian matrix $\mathsf{M} \in \mathbb{C}^{n\times n}$ is called *Hermitian positive definite (hpd)* if $\mathsf{x}^\star\mathsf{M}\mathsf{x} > 0$ for all $\mathsf{x} \neq \mathsf{o}$. The following lemma/definition concerning Hermitian positive definite matrices is important when dealing with norms and scalar products induced by matrices (see below).

Lemma 1.1 (Square root [36, p. 149])
Let M *be Hermitian positive definite. Then there is a unique Hermitian positive definite matrix* $\mathsf{M}^{1/2}$ *such that* $\mathsf{M}^{1/2}\mathsf{M}^{1/2} = \mathsf{M}$. *This matrix is for obvious reasons called* square root of M.

We also note that all eigenvalues (see Section 1.2) of a Hermitian positive definite matrix are positive.

By I_n we denote as usual the identity matrix of size n, we might omit the subscript if the size is clear from context. The k-th column of I_n is denoted by e_k.

1.1.2 Norms

The notion of a normed vector space is supposed to be known and can be found in many textbooks, see, e. g., [36]. By $\|\cdot\|$ we denote an arbitrary vector norm or the corresponding matrix norm. From now on, let $\mathsf{M} \in \mathbb{C}^{n\times m}$ be any matrix.

The norm of M corresponding to the vector norm $\|\cdot\|$ is in general defined via the equation

$$\|\mathsf{M}\| = \max_{\mathsf{x}\neq\mathsf{o},\mathsf{x}\in\mathbb{C}^m} \frac{\|\mathsf{M}\mathsf{x}\|}{\|\mathsf{x}\|} = \max_{\|\mathsf{x}\|=1,\mathsf{x}\in\mathbb{C}^m} \|\mathsf{M}\mathsf{x}\| \,. \tag{1.1}$$

Note that $\|\mathsf{M}\mathsf{x}\|$ is a norm on $\mathbb{C}^n$ while $\|\mathsf{x}\|$ is a norm on $\mathbb{C}^m$; they might be defined in different ways. If the matrix M is real, it corresponds to a map from $\mathbb{R}^m$ to $\mathbb{R}^n$, hence the maxima in (1.1) have to be taken only over real vectors $\mathsf{x} \in \mathbb{R}^m$ to obtain the correct notion. In [46] a simple example can be found where

$$\max_{\|\mathsf{x}\|=1,\mathsf{x}\in\mathbb{C}^m} \|\mathsf{M}\mathsf{x}\| \neq \max_{\|\mathsf{x}\|=1,\mathsf{x}\in\mathbb{R}^m} \|\mathsf{M}\mathsf{x}\|$$

for a real matrix M.

By $\|\mathsf{x}\|_2$ we designate the 2-norm of the vector x,

$$\|\mathsf{x}\|_2 = \sqrt{\sum_{j=1}^{n} |\mathsf{x}(j)|^2}.$$

If not stated otherwise, the symbol $\|\cdot\|$ will denote the 2-norm of a vector or matrix, respectively. Sometimes we write $\|\cdot\|_2$ explicitly to clarify that the 2-norm is used. The 2-norm of a matrix is defined by equation (1.1), where both norms in the definition are taken as the 2-norm.

Occasionally, we will make use of the *Frobenius norm* of M. It arises when considering the space of $n \times m$ matrices as $\mathbb{C}^{n\cdot m}$ and equipping it with the Euclidean norm. The result is the norm

$$\|\mathsf{M}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |\mathsf{M}(i,j)|^2}.$$

Norms enjoy numerous useful properties we will make use of and which can be found in most textbooks, e. g., [36]. However, it is worth to mention that for a square positive definite matrix M it holds $\left\|\mathsf{M}^{1/2}\right\|_2 = \|\mathsf{M}\|_2^{1/2}$.

A quantity that can, for square matrices, be defined by means of norms is the *condition number* κ of the matrix. Its definition depends on a given norm. The condition number describes the sensitivity of the solution of linear systems involving the matrix, with respect to changes in the input data. For a square, nonsingular matrix M its condition number is defined to be

$$\kappa(\mathsf{M}) = \|\mathsf{M}\| \cdot \|\mathsf{M}^{-1}\|.$$

The value of this formula equals, when using the 2-norm, the quotient of largest and smallest singular value, $\sigma_{\max}$ and $\sigma_{\min}$ of M, see Section 1.1.6 below. We have

$$\kappa(\mathsf{M}) = \sigma_{\max}/\sigma_{\min}.$$

The formulation of the condition number by singular values can also be used for non square matrices, as long as they have full rank. The condition number of a rank deficient matrix can formally be defined as ∞.

1.1.3 Scalar products and orthogonality

A *scalar product* on a complex vector space V is a map $\langle \cdot, \cdot \rangle : V \times V \longrightarrow \mathbb{C}$ that fulfills:

1. Positive definiteness: $\langle \mathsf{x}, \mathsf{x} \rangle > 0$ for all $\mathsf{x} \neq \mathsf{o}$.
2. Conjugate symmetry: $\langle \mathsf{x}, \mathsf{y} \rangle = \overline{\langle \mathsf{y}, \mathsf{x} \rangle}$ for all $\mathsf{x}, \mathsf{y} \in V$.
3. Sesquilinearity: For any vectors x, y, z and any scalars α, $\beta \in \mathbb{C}$ it holds $\langle \mathsf{x}, \alpha\mathsf{y} + \beta\mathsf{z} \rangle = \alpha \langle \mathsf{x}, \mathsf{y} \rangle + \beta \langle \mathsf{x}, \mathsf{z} \rangle$.

It follows immediately that $\langle \mathsf{o}, \mathsf{o} \rangle = 0$. We largely use the *standard* or *Euclidean* scalar product on $\mathbb{C}^n$,

$$\langle \mathsf{x}, \mathsf{y} \rangle := \mathsf{x}^\star \mathsf{y} = \sum_{j=1}^{n} \overline{\mathsf{x}(j)} \mathsf{y}(j).$$

Note that any scalar product on V induces a norm on V via $\|\mathsf{x}\| := \sqrt{\langle \mathsf{x}, \mathsf{x} \rangle}$. In particular, the Euclidean scalar product induces the 2-norm.

Two nonzero vectors from a space that is equipped with a scalar product are called *orthogonal* if they satisfy $\langle \mathsf{x}, \mathsf{y} \rangle = 0$. If further $\|\mathsf{x}\| = \|\mathsf{y}\| = 1$, they are said to be *orthonormal*. These definitions naturally generalize to whole sets of vectors if those are pairwise orthogonal.

A matrix X is said to be orthonormal if it has orthonormal columns, in matrix notation this is expressed as $\mathsf{X}^\star \mathsf{X} = \mathsf{I}$. A bit confusingly, a real square matrix is usually called orthogonal if its columns are orthonormal, but we will abide by our definition of an orthonormal matrix. A complex square matrix is called *unitary* if its columns are orthonormal.

1.1.4 Matrix induced scalar products and norms

Given a scalar product $\langle \cdot, \cdot \rangle$ on $\mathbb{C}^n$ and a Hermitian positive definite matrix B, it can easily be established that

$$\langle \mathsf{x}, \mathsf{y} \rangle_{\mathsf{B}} = \langle \mathsf{x}, \mathsf{B} \cdot \mathsf{y} \rangle$$

is also a scalar product on $\mathbb{C}^n$. In particular, if we consider the standard scalar product, we have a new scalar product

$$(\mathsf{x}, \mathsf{y}) \mapsto \mathsf{x}^\star \mathsf{B} \mathsf{y}$$

induced by B. Then, the terms B-*orthogonality* and B-*orthonormality* are well defined. Also, the B-*norm* $\|\mathsf{x}\|_{\mathsf{B}} := \sqrt{\mathsf{x}^\star \mathsf{B} \mathsf{x}}$ of a vector is defined. These definitions are often needed in the remainder of the thesis when dealing with generalized eigenproblems. Note that if X is a square, B-orthonormal matrix, we have $\mathsf{X}^{-1} = \mathsf{X}^\star \mathsf{B}$.

Let K be any factorization of B such that $\mathsf{B} = \mathsf{K}^\star\mathsf{K}$. The matrix K can be the square root of B or its Cholesky factor [36]. Then it is easy to see that $\|\mathsf{x}\|_\mathsf{B} = \|\mathsf{Kx}\|$. For a matrix $\mathsf{M} \in \mathbb{C}^{n\times n}$ we have $\|\mathsf{M}\|_\mathsf{B} = \|\mathsf{KMK}^{-1}\|$, where both norms in the definition of the matrix norm (1.1) for $\|\mathsf{M}\|_\mathsf{B}$ are taken as the B-norm.

A norm similar to the B-norm can for a rectangular matrix $\mathsf{M} \in \mathbb{C}^{n\times m}$ be defined by thinking of the space $\mathbb{C}^n$ equipped with B-norm and $\mathbb{C}^m$ equipped with the 2-norm. We then have

$$\|\mathsf{M}\|_{\mathsf{B}2} := \max_{\|\mathsf{y}\|=1} \|\mathsf{My}\|_\mathsf{B} = \max_{\|\mathsf{y}\|=1} \|\mathsf{KMy}\| = \|\mathsf{KM}\| \,.$$

The notation $\|\cdot\|_{\mathsf{B}2}$ means that $\mathbb{C}^m$ is equipped with the 2-norm while $\mathbb{C}^n$ is equipped with the B-norm. We will call the norm $\|\cdot\|_{\mathsf{B}2}$ the $\mathsf{B}2$-*norm*.

1.1.5 Projectors

We will make use of *projectors* onto subspaces. A square matrix P is called *projector* if $\mathsf{P}^2 = \mathsf{P}$. Then, $\mathsf{I} - \mathsf{P}$ is also a projector and we have $\mathrm{null}(\mathsf{P}) = \mathrm{range}(\mathsf{I} - \mathsf{P})$ and vice versa; $\mathbb{C}^n$ is the direct sum $\mathrm{null}(\mathsf{P}) \oplus \mathrm{range}(\mathsf{P})$. We say the projector is *orthogonal* if null space and range are orthogonal. The notion of a B-orthogonal projector is then well defined.

It can be shown that for every subspace $\mathcal{U}$ there is a unique orthogonal projector (with respect to a given inner product) P with $\mathrm{range}(\mathsf{P}) = \mathcal{U}$. Clearly the converse is true, we hence have a one-to-one correspondence between orthogonal projectors and subspaces.

Given a subspace $\mathcal{U}$ with orthonormal basis U, the matrix $\mathsf{P} = \mathsf{UU}^\star$ is the orthogonal projector onto $\mathcal{U}$ with respect to the standard scalar product. If the basis U is B-orthogonal for some Hermitian positive definite matrix B, the matrix $\mathsf{UU}^\star\mathsf{B}$ is the B-orthogonal projector onto $\mathcal{U}$.

1.1.6 Singular value decomposition

The *singular value decomposition (SVD)* of a matrix plays an important role in this work. Let $\mathsf{M} \in \mathbb{C}^{n\times m}$, $m \leq n$, then it can be shown [36] that there is a matrix $\mathsf{U} \in \mathbb{C}^{n\times m}$ with orthonormal columns, a diagonal matrix $\mathsf{\Sigma} \in \mathbb{R}^{m\times m}$ and a unitary matrix $\mathsf{V} \in \mathbb{C}^{m\times m}$ such that

$$\mathsf{M} = \mathsf{U\Sigma V}^\star.$$

Furthermore, the diagonal entries of $\mathsf{\Sigma}$ are all non-negative and ordered descendingly down the diagonal. These entries are called *singular values* of M. The individual singular values of M, e. g., the diagonal entries of $\mathsf{\Sigma}$, will be denoted by $\sigma_1(\mathsf{M}) \geq \sigma_2(\mathsf{M}) \geq \cdots \geq \sigma_m(\mathsf{M})$. If the matrix is clear from context, we may just write σ_j for $\sigma_j(\mathsf{M})$. The SVD is an outstandingly important matrix factorization that enjoys many useful properties which will not be repeated here (see e. g., [36]). Note that the SVD as defined here is the "thin" or reduced SVD [36].

1.1.7 Computer arithmetic

The symbol ε_M will denote the machine precision in this work, i. e., the smallest machine number ε such that $1 + \varepsilon > 1$ in floating point representation. For IEEE 754 double precision [47, 48] we have $\varepsilon_M = 2^{-53} \approx 1.1 \times 10^{-16}$.

1.2 Eigenvalues and eigenvectors

As this work is concerned with the computation of eigenvalues and eigenvectors, let us define them.

1.2.1 Basic notions

In this subsection, we introduce eigenvectors and eigenvalues, both for the standard and the generalized problem.

Standard problem

Given a square matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$, we consider the equation

$$\mathsf{A}\mathsf{x} = \mathsf{x}\lambda, \tag{1.2}$$

where $\mathsf{x} \in \mathbb{C}^n$ is a vector and $\lambda \in \mathbb{C}$. If (1.2) is fulfilled with some nonzero vector $\mathsf{x} \in \mathbb{C}^n$, this vector is called *eigenvector* and the number λ is called *eigenvalue*. The pair (x, λ) is accordingly called *eigenpair*. Obviously, for any $0 \neq \alpha \in \mathbb{C}$, $\alpha\mathsf{x}$ is also an eigenvector. We will also call a matrix pair (X, Λ) an eigenpair of A if X has full rank, Λ is a diagonal matrix and the equation $\mathsf{A}\mathsf{X} = \mathsf{X}\Lambda$ is fulfilled, i. e., if the columns of X are eigenvectors. Sporadically, we will even call a matrix pair (X, H) an eigenpair of A if $\mathsf{A}\mathsf{X} = \mathsf{X}\mathsf{H}$ for a non-diagonal matrix H and a matrix X of full rank. In this case, the eigenvalues of H are also eigenvalues of A (see below). Note that the columns of X are not necessarily eigenvectors in this case.

The set of eigenvalues of A will be denoted by $\operatorname{spec}(\mathsf{A})$ and is called *spectrum*. The number $\rho(\mathsf{A}) := \max\{|\lambda| : \lambda \text{ is eigenvalue of } \mathsf{A}\}$ is called the *spectral radius* of A.

Many simple statements can be made about eigenvalues and eigenvectors (and found in any textbook on linear algebra, such as [36, 108, 112]). Let us collect the most important ones.

- The eigenvalues are the roots of the *characteristic polynomial* $\det(\mathsf{A} - \lambda\mathsf{I}_n)$. In particular, $\det(\mathsf{A} - \lambda\mathsf{I}_n)$ is a polynomial in λ of degree n.
- Consequently, A has n eigenvalues. Those are not necessarily distinct which actually means that A has at least one eigenvalue.
- If A is Hermitian, all eigenvalues are real.

- Even more notably, all eigenvectors are *orthogonal* in that case, i. e., if x, y are eigenvectors to different eigenvalues, we have $\mathsf{x}^\star\mathsf{y} = 0$.
- $\rho(\mathsf{A}) \leq \|\mathsf{A}\|$ for *any* norm.

Generalized problem

Most of this thesis is dealing with the so called *generalized eigenvalue problem* that consists of finding solutions of the equation

$$\mathsf{A}\mathsf{x} = \mathsf{B}\mathsf{x}\lambda. \tag{1.3}$$

A solution of (1.3) involves a nonzero vector x and a number λ, called *(generalized) eigenvector* and *(generalized) eigenvalue* of the pair (A, B), respectively. Consequently, the pair (x, λ) is called a *(generalized) eigenpair* of the matrix pair (A, B). The blockwise eigenpairs (X, Λ), (X, H) as for the standard problem are declared accordingly. The set of generalized eigenvalues is also called spectrum and is denoted by spec(A, B).

We usually require B to be nonsingular, consequently we have

$$\mathsf{A}\mathsf{x} = \mathsf{B}\mathsf{x}\lambda \iff \mathsf{B}^{-1}\mathsf{A}\mathsf{x} = \mathsf{x}\lambda. \tag{1.4}$$

In other words, generalized eigenvalues and eigenvectors of (A, B) are eigenvalues and eigenvectors of the matrix $\mathsf{B}^{-1}\mathsf{A}$. We will use this relation from time to time for theoretical considerations; it is common sense that it should not be used in a numerical algorithm. In [36, Example 7.7.1] an example is stated that illustrates the danger of forming $\mathsf{B}^{-1}\mathsf{A}$ and then solving the eigenvalue problem of that matrix. Difficulties typically appear if B has a high condition number. Another interesting example why the inversion of B should be avoided can be found in [109]. Looking simple, much of the theory for the standard Hermitian problem is not valid any more in case of the generalized problem. This can best be explained by a simple example, borrowed from Parlett [80, Sec. 15.2].

Example 1.2 (Parlett). In this example, three cases are considered which illustrate that the solution of the generalized eigenproblem can result in more difficulties than the solution of the Hermitian standard problem does. Note that point 1 and 2 include a singular matrix B, meaning (1.4) does not make sense in these cases. In all three examples, A and B are real and symmetric.

1. The pair
$$\mathsf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$
has obviously eigenvalue 1 belonging to eigenvector e_1, but *any* number $\lambda \in \mathbb{C}$ fulfills $\mathsf{A}\mathsf{e}_2 = \mathsf{B}\mathsf{e}_2\lambda$.

2. $$\mathsf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$
   has $(0, \mathsf{e}_2)$ as eigenpair. Solving $\mathsf{A}\mathsf{e}_1 = \mathsf{B}\mathsf{e}_1\lambda$ for λ yields $1 \cdot \mathsf{e}_1 = 0 \cdot \mathsf{e}_1\lambda$ and hence formally $\lambda = 1/0$. Such an eigenvalue is called *infinite* eigenvalue.

3. Let
   $$\mathsf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathsf{B} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$
   Finding the zeros of the quadratic equation $\det(\mathsf{A} - \lambda\mathsf{B})$ yields $+\mathbf{i}, -\mathbf{i}$ ($\mathbf{i}$ denotes the complex unit). This shows that even in the case of two real symmetric matrices the eigenvalues of the corresponding generalized eigenvalue problem need not be real.

$\Diamond$

The example illustrates nicely the three problems with generalized eigenvalue problems: unbounded spectra, infinite eigenvalues and complex eigenvalues, even though all matrices that appear are real symmetric (the examples in [36, 109] include non-symmetric matrices).

The key is the Hermitian positive definiteness of B. If this is fulfilled and A is Hermitian, the eigenvalues of (A, B) are real again. In the following, we will refer to such a pair as a *definite pair*. In [103] and possibly elsewhere in the literature, the definition of a definite pair is slightly weaker, however we will use it exactly as described. We then have the following theorem whose statement is the analogue to the orthogonality of eigenvectors in the standard Hermitian case.

Theorem 1.3 ([103, Thm. VI.1.15])
Let (A, B) *be a definite matrix pair of size* n*. Then there is a nonsingular matrix* X *satisfying* $\mathsf{X}^\star\mathsf{B}\mathsf{X} = \mathsf{I}_n$ *and* $\mathsf{X}^\star\mathsf{A}\mathsf{X} = \Lambda$*, where* Λ *is real and diagonal.*

It can then easily be seen that the columns of X are eigenvectors of the pair (A, B) corresponding to the eigenvalues on the diagonal of Λ. The property $\mathsf{X}^\star\mathsf{B}\mathsf{X} = \mathsf{I}_n$ is nothing but the B-orthonormality of the vectors collected in X, expressed in matrix terms.

"Standardizing" of generalized eigenproblems

Hermitian definite problems have an important advantage for theoretical considerations; they can be transformed to standard problems. Let $\mathsf{K}^\star\mathsf{K} = \mathsf{B}$ be a factorization of B. We then have $\mathsf{A}\mathsf{x} = \mathsf{K}^\star\mathsf{K}\mathsf{x}\lambda$ and hence $(\mathsf{K}^\star)^{-1}\mathsf{A}\mathsf{x} = \mathsf{K}\mathsf{x}\lambda$. Let $\mathsf{y} = \mathsf{K}\mathsf{x}$, i. e., $\mathsf{x} = \mathsf{K}^{-1}\mathsf{y}$, we obtain

$$\mathsf{A}\mathsf{x} = \mathsf{K}^\star\mathsf{K}\mathsf{x}\lambda \iff (\mathsf{K}^\star)^{-1}\mathsf{A}\mathsf{K}^{-1}\mathsf{y} = \mathsf{y}\lambda.$$

This is a standard Hermitian eigenvalue problem with the same eigenvalues as the original problem. The eigenvectors transform in a simple manner.

Two important examples for K may be mentioned. First, the Cholesky factorization [36] where K is an upper triangular matrix. This factorization is computable, at least for small systems. Linear systems with K as system matrix are easy to solve and the transformation of a generalized problem via the Cholesky factors is sometimes actually done in practice [8]. The second example is $\mathsf{K} = \mathsf{B}^{1/2}$ that will sometimes be useful in theoretical considerations. We will often use $\mathsf{B}^{1/2}$, in most cases it can be replaced by a general factor K of B with $\mathsf{K}^\star\mathsf{K} = \mathsf{B}$.

Notation

We make the following conventions on notation:

- $\lambda_{\min}$ is the smallest, $\lambda_{\max}$ the largest eigenvalue of the matrix (pair) of the current discussion.
- Whenever eigenvalues are numbered, they are implicitly assumed to be ordered according to their index, i. e., if $j_1 < j_2$, we have $\lambda_{j_1} \leq \lambda_{j_2}$ (unless otherwise stated). Eigenvalues with different index need not be distinct: although $j_1 \neq j_2$, we can have $\lambda_{j_1} = \lambda_{j_2}$. However, if we make statements about a collection Θ of eigenvalues, we implicitly assume the following:

  $$\lambda \in \Theta \Rightarrow \mu \neq \lambda \text{ for all } \mu \in \operatorname{spec}(\mathsf{A}, \mathsf{B}) \setminus \Theta.$$

  In words, a certain collection of eigenvalues is always assumed to contain all eigenvalues with a particular value.
- Sometimes we write, for instance, $\lambda(\mathsf{A})$ in order to clarify that λ is an eigenvalue of A.

1.2.2 Eigenspaces

An eigenvector x of the standard equation $\mathsf{Ax} = \mathsf{x}\lambda$ fulfills in particular

$$\operatorname{span}(\mathsf{Ax}) \subseteq \operatorname{span}(\mathsf{x}),$$

with equality if and only if $\lambda \neq 0$. Next, let $\mathsf{X} = [\mathsf{x}_1, \ldots, \mathsf{x}_m]$, $m \leq n$ be an orthonormal matrix such that

$$\mathsf{AX} = \mathsf{XH}, \tag{1.5}$$

where H is some square matrix of size m. If G is another matrix than H fulfilling (1.5), we have $\mathsf{XH} = \mathsf{XG}$, consequently $\mathsf{X}^\star\mathsf{XH} = \mathsf{X}^\star\mathsf{XG}$ and hence $\mathsf{H} = \mathsf{G}$. In other words, the factor H in (1.5) is uniquely determined.

Now, let (W, Λ) be an eigenpair of H, where Λ is a diagonal matrix and W has orthonormal columns. We then have

$$(\mathsf{XW})^\star\mathsf{AXW} = \Lambda. \tag{1.6}$$

This shows that XW is a matrix consisting of (orthonormal) eigenvectors of A and that the eigenvalues of H are also eigenvalues of A, i. e., $\mathrm{spec}(\mathsf{H}) \subseteq \mathrm{spec}(\mathsf{A})$.

The space $\mathcal{X} = \mathrm{span}(\mathsf{X})$ is called *invariant subspace* of A since $\mathsf{A}\mathcal{X} \subseteq \mathcal{X}$. Because it is related to certain eigenvalues of A (those of $\mathsf{X}^\star\mathsf{AX} = \mathsf{H}$) we will call it *eigenspace* in the following. We will identify the space $\mathrm{span}(\mathsf{X})$ with the matrix X and vice versa (we will from now on always mean the space $\mathrm{span}(\mathsf{X})$ if we say "the subspace X" with some matrix X). We say that the invariant subspace X *belongs* to the eigenvalues of H. On the other hand, to each eigenspace belongs a set of eigenvalues, namely those of H, the uniquely determined matrix. We hence have an important one-to-one correspondence between subsets of the spectrum and eigen subspaces. The fact that some eigenvalues of A can be computed from the often much smaller matrix H is the key of eigenvalue computations of large matrices. Eigenvectors of A can be extracted from information in X and H via equation (1.6).

The definition of an eigenspace is slightly more complicated in the case of a matrix pair (A, B). Let a vector x fulfill the eigenvector equation

$$\mathsf{Ax} = \mathsf{Bx}\lambda. \tag{1.7}$$

We then obviously have that $\mathrm{span}(\mathsf{Ax}) \subseteq \mathrm{span}(\mathsf{Bx})$. Generalizing (1.7) to a similar equation as (1.5) yields

$$\mathsf{AX} = \mathsf{BXH}$$

with a matrix $\mathsf{X} \in \mathbb{C}^{n\times m}$. Let $\mathcal{X} = \mathrm{span}(\mathsf{X})$. Obviously, both spaces $\mathsf{A}\mathcal{X}$ and $\mathsf{B}\mathcal{X}$ reside in one subspace of dimension $\leq m$. If—as supposed—B is nonsingular, we also have $\mathsf{A}\mathcal{X} \subseteq \mathsf{B}\mathcal{X}$. Such a subspace $\mathcal{X}$ is called an *eigenspace* of the pair (A, B). If B is a general matrix, we have to require $\dim(\mathsf{A}\mathcal{X}+\mathsf{B}\mathcal{X}) \leq \dim(\mathcal{X})$ [103]. Like in the standard case, the computation of eigenspaces is the key in computing eigenvalues and eigenvectors of matrix pairs, as is stated in the following theorem.

Theorem 1.4 ([103, Thm. VI.3.5])
Let (A, B) *be a definite matrix pair and let the columns of* X_1 *be a basis of an eigenspace of* (A, B). *Then there is a matrix* X_2 *such that* $[\mathsf{X}_1, \mathsf{X}_2]$ *is nonsingular and the equations*

$$\begin{bmatrix}\mathsf{X}_1^\star\\ \mathsf{X}_2^\star\end{bmatrix} \mathsf{A}\,[\mathsf{X}_1, \mathsf{X}_2] = \begin{bmatrix}\mathsf{A}_1 & 0\\ 0 & \mathsf{A}_2\end{bmatrix},$$
$$\begin{bmatrix}\mathsf{X}_1^\star\\ \mathsf{X}_2^\star\end{bmatrix} \mathsf{B}\,[\mathsf{X}_1, \mathsf{X}_2] = \begin{bmatrix}\mathsf{B}_1 & 0\\ 0 & \mathsf{B}_2\end{bmatrix}$$

hold. Moreover, $\mathsf{X}_1, \mathsf{X}_2$ *may be chosen such that* $\mathsf{A}_1, \mathsf{A}_2$ *are diagonal and* $\mathsf{B}_1, \mathsf{B}_2$ *are identity matrices of proper dimension, meaning* $\mathsf{X}_1, \mathsf{X}_2$ *are eigenvectors.*

Theorem 1.4 states that the basis of an eigenspace of (A, B) is sufficient to compute a subset of the spectrum of (A, B). Note that bases of eigenspaces need not be formed from eigenvectors.

1.3 Angles between vectors and subspaces

A common and meaningful measure for the "distance" of eigenvectors and subspaces is the angle between those objects rather than the Euclidean distance (which is zero for subspaces, since the zero vector is contained in every subspace). In this section, we discuss different notions of angles for scalar product based geometries.

1.3.1 Scalar products and geometry

Let the vector space $\mathbb{C}^n$ be equipped with a (so far abstract) scalar product $\langle \cdot, \cdot \rangle$. This scalar product defines a geometry on $\mathbb{C}^n$. Via the norm $\|\mathsf{x}\| = \sqrt{\langle \mathsf{x}, \mathsf{x} \rangle}$, the length of a vector and distances between vectors can be measured. The following well-known result gives the possibility to define angles between vectors and subsequently between whole subspaces.

Lemma 1.5 (Cauchy–Schwartz)
For a vector space V equipped with a scalar product $\langle \cdot, \cdot \rangle$ and $\mathsf{x}, \mathsf{y} \in V$ it holds

$$|\langle \mathsf{x}, \mathsf{y} \rangle| \leq \sqrt{\langle \mathsf{x}, \mathsf{x} \rangle} \sqrt{\langle \mathsf{y}, \mathsf{y} \rangle}.$$

The lemma allows us to define the unique acute angle $\theta := \angle(\mathsf{x}, \mathsf{y}) \in [0, \pi/2]$ between x and y as

$$\cos \theta = \frac{|\langle \mathsf{x}, \mathsf{y} \rangle|}{\|\mathsf{x}\| \, \|\mathsf{y}\|}. \tag{1.8}$$

Note that with (1.8) the angle between x and $-\mathsf{x}$ is 0. The definition naturally extends to the angle between two one-dimensional subspaces that can be written as $\operatorname{span}(\mathsf{x})$, $\operatorname{span}(\mathsf{y})$ as the angle between the two basis vectors.

1.3.2 Angles between subspaces

It is not trivial to extend the notion of angles between vectors to the angle between two subspaces $\mathcal{U}, \mathcal{V} \subset \mathbb{C}^n$ of arbitrary dimension. A first approach is to define "the" angle as the maximum of all angles $\angle(\mathsf{u}, \mathsf{v})$ between vectors $\mathsf{u} \in \mathcal{U}$, $\mathsf{v} \in \mathcal{V}$, which will not give useful results as one can easily see.

The right track is along the so called *principal angles*; a comprehensive overview can be found in [12, 123] and [58] (which we closely follow) but also in textbooks such as [36]. In the following, we will use the standard scalar product $\langle \mathsf{x}, \mathsf{y} \rangle = \mathsf{x}^\star \mathsf{y}$.

Let $p = \dim \mathcal{U} \geq \dim \mathcal{V} = q \geq 1$, then we can define the principal angles $\theta_1, \ldots, \theta_q$ between $\mathcal{U}$ and $\mathcal{V}$ algorithmically as (see [58])

$$\cos \theta_k = \max_{\mathsf{u} \in \mathcal{U}_k, \|\mathsf{u}\|=1} \max_{\mathsf{v} \in \mathcal{V}_k, \|\mathsf{v}\|=1} |\mathsf{u}^\star \mathsf{v}| =: |\mathsf{u}_k^\star \mathsf{v}_k|, \; k = 1, \ldots, q \tag{1.9}$$

where

$$\begin{aligned} \mathcal{U}_k &= \{\mathsf{u} \in \mathcal{U} : \mathsf{u}^\star \mathsf{u}_j = 0,\ j = 1, \ldots, k-1\}, \\ \mathcal{V}_k &= \{\mathsf{v} \in \mathcal{V} : \mathsf{v}^\star \mathsf{v}_j = 0,\ j = 1, \ldots, k-1\}. \end{aligned} \tag{1.10}$$

In (1.9), the vectors u_k, v_k are implicitly defined. They are given as vectors for which in the left hand side of the equation the maxima are attained. Clearly, the angle θ_q has to fulfill most restrictions and hence $\cos\theta_q$ is the smallest among all cosines; we consequently call θ_q the *largest canonical angle* between $\mathcal{U}$ and $\mathcal{V}$ and define the angle between the two subspaces as

$$\angle(\mathcal{U}, \mathcal{V}) := \theta_q.$$

The definition (1.9)–(1.10) of canonical angles is not very handy, but there is an alternative formulation that pleases the linear algebra scientist more. Let U, V denote orthonormal bases of $\mathcal{U}$, $\mathcal{V}$ respectively, we may write (1.9) as (see [36, p. 603]

$$\max_{\mathsf{u}\in\mathcal{U},\ \|\mathsf{u}\|=1} \max_{\mathsf{v}\in\mathcal{V},\ \|\mathsf{v}\|=1} |\mathsf{u}^\star \mathsf{v}| = \max_{\mathsf{y}\in\mathbb{C}^p,\ \|\mathsf{y}\|=1} \max_{\mathsf{z}\in\mathbb{C}^q,\ \|\mathsf{z}\|=1} |\mathsf{y}^\star \mathsf{U}^\star \mathsf{V} \mathsf{z}|. \tag{1.11}$$

Together with the orthogonality constraints given in the definitions in (1.10), equation (1.11) characterizes the singular values of $\mathsf{U}^\star\mathsf{V}$, i. e., $\cos\theta_k = \sigma_k(\mathsf{U}^\star\mathsf{V})$ [12]. We formulate this important relation as a theorem.

Theorem 1.6 (Canonical angles as singular values, [12, Thm. 1])
Let $\mathcal{U}$, $\mathcal{V}$ be subspaces as above with orthonormal bases U, V, then the canonical angles $0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_q \le \pi/2$ between $\mathcal{U}$ and $\mathcal{V}$ are given by

$$\theta_k = \arccos(\sigma_k),\ k = 1, \ldots, q,$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_q \ge 0$ are the first q (potentially zero) singular values of $\mathsf{U}^\star\mathsf{V}$.

Note that by the foregoing theorem the angle $\angle(\mathcal{U}, \mathcal{V})$ is symmetric in its arguments, since $\mathsf{U}^\star\mathsf{V}$ and $\mathsf{V}^\star\mathsf{U}$ have the same singular values. It can hence also be defined for $p \le q$.

Small angles

Small angles θ, as they appear in the convergence analysis of Ritz values, are not well determined by $\cos\theta$. It would be convenient to have an expression for the sine, since $\sin\theta \approx \theta$ near 0.

In the following, the symbol $\perp$ stands for the orthogonal complement; $\mathcal{U}_\perp$ denotes the orthogonal complement of a subspace $\mathcal{U}$. If U is an orthonormal matrix, then $\mathsf{U}_\perp$ is a matrix such that $[\mathsf{U}, \mathsf{U}_\perp]$ is unitary. Motivated by the simple case of two vectors in $\mathbb{R}^2$ and simple trigonometry, we expect that $\sin\angle(\mathcal{U}, \mathcal{V}) = \|\mathsf{U}_\perp^\star \mathsf{V}\|$. This is true, indeed, as the following theorem and its consequences show. For completeness, we also state its proof. For brevity, let in the following $\mathsf{P}_\mathsf{U} = \mathsf{U}\mathsf{U}^\star$ and $\mathsf{P}_\mathsf{V} = \mathsf{V}\mathsf{V}^\star$ denote the orthogonal projectors onto $\mathcal{U}$ and $\mathcal{V}$, respectively.

Theorem 1.7 (Sines of canonical angles, [58, Thm. 3.1])
The singular values $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_q$ *of* $(\mathsf{I} - \mathsf{P}_\mathsf{U})\mathsf{V}$ *are given by* $\mu_k = \sqrt{1 - \sigma_k^2}$, *where* σ_k *are the singular values of* $\mathsf{U}^\star\mathsf{V}$ *(i. e., the cosines of the canonical angles between* $\mathcal{U}$ *and* $\mathcal{V}$*). We have* $\sin\angle(\mathcal{U}, \mathcal{V}) = \|(\mathsf{I} - \mathsf{P}_\mathsf{U})\mathsf{V}\|$.

Proof. Recall that $\mathsf{I} - \mathsf{U}\mathsf{U}^\star$ is the orthogonal projector onto $\mathcal{U}_\perp$ and let $\mathsf{C} = (\mathsf{I} - \mathsf{U}\mathsf{U}^\star)\mathsf{V}$. We then have

$$\begin{aligned}\mathsf{C}^\star\mathsf{C} &= \mathsf{V}^\star(\mathsf{I}_n - \mathsf{U}\mathsf{U}^\star)(\mathsf{I}_n - \mathsf{U}\mathsf{U}^\star)\mathsf{V} \\ &= \mathsf{V}^\star(\mathsf{I}_n - \mathsf{U}\mathsf{U}^\star)\mathsf{V} \\ &= \mathsf{I}_q - \mathsf{V}^\star\mathsf{U}\mathsf{U}^\star\mathsf{V}.\end{aligned}$$

Next, let $\mathsf{Y}\mathsf{\Sigma}\mathsf{Z}^\star = \mathsf{U}^\star\mathsf{V}$ be the reduced SVD of $\mathsf{U}^\star\mathsf{V}$ with a square matrix $\mathsf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_q)$. With Z we have

$$\begin{aligned}\mathsf{Z}^\star\mathsf{C}^\star\mathsf{C}\mathsf{Z} &= \mathsf{Z}^\star(\mathsf{I}_q - \mathsf{V}^\star\mathsf{U}\mathsf{U}^\star\mathsf{V})\mathsf{Z} \\ &= \mathsf{I}_q - \mathsf{Z}^\star(\mathsf{Z}\mathsf{\Sigma}\mathsf{Y}^\star\mathsf{Y}\mathsf{\Sigma}\mathsf{Z}^\star)\mathsf{Z} \\ &= \mathsf{I}_q - \mathsf{\Sigma}^2.\end{aligned}$$

Hence, the singular values of C are the numbers

$$\mu_k = \sqrt{1 - \sigma_k^2} = \sqrt{1 - \cos^2\theta_k} = \sin\theta_k.$$

It follows $\sin\angle(\mathcal{U}, \mathcal{V}) = \mu_q = \|(\mathsf{I}_n - \mathsf{P}_\mathsf{U})\mathsf{V}\|$. □

We can also easily prove the following result, see [58].

Theorem 1.8
Let σ_k, $k = 1, \ldots, q$ *denote the singular values of* $\mathsf{U}^\star\mathsf{V}$. *We then have the relations*

1. $\sigma_k = \sigma_k(\mathsf{P}_\mathsf{U}\mathsf{V})$
2. *The numbers* σ_k, $k = 1, \ldots, q$ *are the* q *largest singular values of* $\mathsf{P}_\mathsf{U}\mathsf{P}_\mathsf{V}$*; all other singular values of this matrix are zero.*
3. *The* q *largest singular values of* $(\mathsf{I} - \mathsf{P}_\mathsf{U})\mathsf{P}_\mathsf{V}$ *are the numbers* $\mu_1, \ldots, \mu_q$ *from Theorem* 1.7.

Proof.

1. Just replace C in the proof of Theorem 1.7 with $\mathsf{P}_\mathsf{U}\mathsf{V}$.
2. Use the maximum characterization (1.11) and observe

$$\max_{\mathsf{y}\in\mathbb{C}^p,\ \|\mathsf{y}\|=1}\ \max_{\mathsf{z}\in\mathbb{C}^q,\ \|\mathsf{z}\|=1} |\mathsf{y}^\star\mathsf{U}^\star\mathsf{V}\mathsf{z}| = \max_{\tilde{\mathsf{y}}\in\mathbb{C}^n,\ \|\tilde{\mathsf{y}}\|=1}\ \max_{\tilde{\mathsf{z}}\in\mathbb{C}^n,\ \|\tilde{\mathsf{z}}\|=1} |\tilde{\mathsf{y}}^\star\mathsf{U}\mathsf{U}^\star\mathsf{V}\mathsf{V}^\star\tilde{\mathsf{z}}|\,.$$

   Due to the rank of $\mathsf{P}_\mathsf{U}\mathsf{P}_\mathsf{V}$, it is clear that the last $n - q$ singular values are zero.

3. follows from Theorem 1.7, 1. and 2.

□

Combining the facts on canonical angles, we see that $\sin\angle(\mathcal{U},\mathcal{V}) = \cos\angle(\mathcal{U}_\perp^\star,\mathcal{V}) = \|\mathsf{U}_\perp\mathsf{V}\|$, which was the desired result. Let us close this section with a note on the case $p = q$ (see [58]). In that case, we have

$$\sin\angle(\mathcal{U},\mathcal{V}) = \|\mathsf{P}_\mathsf{U} - \mathsf{P}_\mathsf{V}\|\,,$$

yielding a metric on the set of all p-dimensional subspaces of $\mathbb{C}^n$.

1.3.3 Angles in B-induced scalar products

When dealing with generalized eigenproblems of a matrix pair (A, B), we will often express angles in terms of the scalar product induced by B. This is natural when both matrices are Hermitian and B is positive definite, in addition. Our goal will be to compute B-orthonormal eigenvectors, hence we should measure all angles in the B-scalar product $\langle \mathsf{x}, \mathsf{y}\rangle_\mathsf{B} := \mathsf{x}^\star\mathsf{B}\mathsf{y}$. Angles between two vectors then are defined via

$$\cos\theta := \frac{|\langle \mathsf{x}, \mathsf{y}\rangle_\mathsf{B}|}{\|\mathsf{x}\|_\mathsf{B}\,\|\mathsf{y}\|_\mathsf{B}},$$

which is a reasonable definition because the Cauchy-Schwartz inequality holds for all scalar products.

Next, we will define canonical angles in the B-geometry. Knyazev and Argentati [58] gave a comprehensive overview of the subject; we will follow their lines closely. Subsequently, all angles are expressed in the B scalar product, unless stated otherwise. We write $\angle_\mathsf{B}$ for the angle to emphasize this fact. Again, let $p = \dim\mathcal{U} \geq \dim\mathcal{V} = q \geq 1$, then we can define the principal angles $\theta_1, \ldots, \theta_q$ between $\mathcal{U}$ and $\mathcal{V}$ in the B-geometry algorithmically as (see [58])

$$\cos\theta_k = \max_{\mathsf{u}\in\mathcal{U}_k,\,\|\mathsf{u}\|_\mathsf{B}=1}\ \max_{\mathsf{v}\in\mathcal{V}_k\,\|\mathsf{v}\|_\mathsf{B}=1} |\mathsf{u}^\star\mathsf{B}\mathsf{v}| =: |\mathsf{u}_k{}^\star\mathsf{B}\mathsf{v}_k|\,,\ k=1,\ldots,q \tag{1.12}$$

where

$$\begin{aligned}\mathcal{U}_k &= \{\mathsf{u}\in\mathcal{U} : \mathsf{u}^\star\mathsf{B}\mathsf{u}_j = 0,\ j=1,\ldots,k-1\}\,,\\ \mathcal{V}_k &= \{\mathsf{v}\in\mathcal{V} : \mathsf{v}^\star\mathsf{B}\mathsf{v}_j = 0,\ j=1,\ldots,k-1\}\,.\end{aligned} \tag{1.13}$$

Like in the case of the standard scalar product, the vectors u_k, v_k in (1.12) are defined implicitly as vectors for which the maxima at the left hand side of the equation are attained. Our goal now again is to relate the canonical angles defined via (1.12)–(1.13) to certain singular values.

Theorem 1.9 (Cosines of canonical angles; [58])
Let U, V *be* B*-orthonormal bases of* $\mathcal{U}$, $\mathcal{V}$, *respectively. Let* $\sigma_1 \geq \ldots \geq \sigma_q$ *denote the singular values of* $\mathsf{U}^\star\mathsf{B}\mathsf{V}$. *Then, the canonical angles defined via* (1.12)–(1.13) *fulfill*

$$\theta_k = \arccos\sigma_k \in [0,\pi/2],\ k=1,\ldots,q.$$

In particular, we have

$$\cos\angle_\mathsf{B}(\mathcal{U},\mathcal{V}) = \|\mathsf{U}^\star\mathsf{B}\mathsf{V}\|\,.$$

Proof. The proof is again based on the maximum characterization of singular values, (1.11). We can express the vectors u, v from (1.12) in terms of the bases U and V and obtain

$$\cos\theta_k = \max_{\mathsf{y}\in\mathbb{C}^p}\max_{\mathsf{z}\in\mathbb{C}^q} |\mathsf{y}^\star\mathsf{U}^\star\mathsf{B}\mathsf{V}\mathsf{z}| = |\mathsf{y}_k^\star\mathsf{U}^\star\mathsf{B}\mathsf{V}\mathsf{z}_k|\,,\ k=1,\ldots,q$$

with the constraints

$$\|\mathsf{y}\| = \|\mathsf{z}\| = 1,\ \mathsf{y}^\star\mathsf{y}_j = 0,\ \mathsf{z}^\star\mathsf{z}_j = 0,\ j=1,\ldots,k-1.$$

The vectors $\mathsf{u}_k = \mathsf{U}\mathsf{y}_k$, $\mathsf{u} = \mathsf{U}\mathsf{y}$ and $\mathsf{v}_k = \mathsf{V}\mathsf{z}_k$, $\mathsf{v} = \mathsf{V}\mathsf{z}$ fulfill the orthogonality constraints from (1.13). Simple computations show that $\|\mathsf{u}_k\|_\mathsf{B} = \|\mathsf{u}\|_\mathsf{B} = 1$ and $\|\mathsf{v}_k\|_\mathsf{B} = \|\mathsf{v}\|_\mathsf{B} = 1$.

The statement of the theorem follows with the maximum characterization of singular values and we obtain $\cos\theta_k = \sigma_k,\ k=1,\ldots,q$. □

The canonical angles in the B- and in the standard scalar product are simply related. Let $\mathsf{B} = \mathsf{K}^\star\mathsf{K}$, e. g., let $\mathsf{K} = \mathsf{B}^{1/2}$ or let K be a Cholesky factor of B. Then obviously $\mathsf{K}\mathsf{U}$ and $\mathsf{K}\mathsf{V}$ are orthonormal bases for the spaces $\mathsf{K}\mathcal{U}$ and $\mathsf{K}\mathcal{V}$, respectively. Writing $\mathsf{U}^\star\mathsf{B}\mathsf{V} = (\mathsf{K}\mathsf{U})^\star(\mathsf{K}\mathsf{V})$ and invoking Theorems 1.6 and 1.9 shows that the canonical angles between $\mathcal{U}$ and $\mathcal{V}$ in the B scalar product and the canonical angles between $\mathsf{K}\mathcal{U}, \mathsf{K}\mathcal{V}$ in the standard scalar product coincide. For the largest canonical angle we have

$$\angle_\mathsf{B}(\mathcal{U},\mathcal{V}) = \angle(\mathsf{K}\mathcal{U},\mathsf{K}\mathcal{V}),$$

see [58, Thm. 4.2].

Small angles

Similarly to the standard scalar product, we can express angles in the B scalar product in terms of their sines, which is important since $\sin\theta \approx \theta$ for small angles θ. In the following, let $\mathsf{P_U} = \mathsf{UU^\star B}$, $\mathsf{P_V} = \mathsf{VV^\star B}$ denote the B-orthogonal projectors onto $\mathcal{U}$ and $\mathcal{V}$, respectively. Similarly to Theorem 1.7, the following can be proven [58, Thm. 4.3].

Theorem 1.10
Let K *be such that* $\mathsf{K^\star K} = \mathsf{B}$. *Then, the singular values* $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_q$ *of* $\mathsf{K}(\mathsf{I} - \mathsf{P_U})\mathsf{V}$ *are given by* $\mu_k = \sqrt{1-\sigma_k^2}$, *where* σ_k *are the singular values of* $\mathsf{V^\star BU}$ *(i. e., the cosines of the canonical angles between* $\mathcal{U}$ *and* $\mathcal{V}$ *in the* B *scalar product). We have* $\theta_k = \arcsin(\mu_k)$, $k = 1, \ldots, q$.

As a direct consequence we obtain, avoiding a factorization of B:

Theorem 1.11 ([58, Thm. 4.4])
Let $\mathsf{S} = (\mathsf{I} - \mathsf{P_U})\mathsf{V}$ *and let* $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_q$ *be the eigenvalues of* $\mathsf{S^\star BS}$. *Then we have* $\nu_k = 1 - \sigma_k^2$, $k = 1, \ldots, q$, *where* σ_k *are the singular values of* $\mathsf{V^\star BU}$. *We have* $\theta_k = \arcsin(\sqrt{\nu_k})$, $k = 1, \ldots, q$.

Finally, let us note:

Theorem 1.12 ([58, Thm. 4.6])
The singular values $\sigma_1 \geq \cdots \geq \sigma_q$ *from Theorem* 1.9 *are the* q *largest singular values of* $\mathsf{KP_U P_V K^{-1}}$.

Theorems 1.10 and 1.12 are slight and simple generalizations of [58, Thms. 4.3, 4.6] since they allow some factor K of B (with $\mathsf{K^\star K} = \mathsf{B}$) instead of only the square root.

Now, it can be seen that the cosines of the canonical angles between $\mathcal{U}, \mathcal{V}$ are the sines of the canonical angles between $\mathcal{U}_\perp, \mathcal{V}$,

$$\cos\angle_{\mathsf{B}}(\mathcal{U}, \mathcal{V}) = \sin\angle_{\mathsf{B}}(\mathcal{U}_\perp, \mathcal{V})$$

(and vice versa). We have

$$\sin\angle_{\mathsf{B}}(\mathcal{U}, \mathcal{V}) = \|\mathsf{U_\perp^\star BV}\|$$

due to Theorem 1.9. If $p = q$, it can also be shown that

$$\sin\angle_{\mathsf{B}}(\mathcal{U}, \mathcal{V}) = \|\mathsf{P_U} - \mathsf{P_V}\|_{\mathsf{B}}\,.$$

This shows that $\sin\angle_{\mathsf{B}}(\cdot,\cdot)$ is a metric on the set of all q-dimensional subspaces of $\mathbb{C}^n$.

Connection between angles in standard and B-scalar product

Finally, let us see how the angles in standard and in B-geometry are related. Such relations are hard to find in the corresponding literature. However, in one of

Knyazev's early works [57] they can be found[1]. The result can easily be formulated in terms of the sines of the respective angles. We have

$$\frac{\sin^2 \angle_{\mathsf{B}}(\mathcal{U},\mathcal{V})}{\sin^2 \angle(\mathcal{U},\mathcal{V})} \in \left[\frac{\lambda_{\min}(\mathsf{B})}{\lambda_{\max}(\mathsf{B})}, \frac{\lambda_{\max}(\mathsf{B})}{\lambda_{\min}(\mathsf{B})}\right]. \tag{1.14}$$

Analyzing the interval boundaries in (1.14) shows that the interval is $[1/\kappa(\mathsf{B}), \kappa(\mathsf{B})]$ since all eigenvalues of B are positive. We consequently have that

$$\frac{1}{\kappa(\mathsf{B})} \leq \frac{\sin^2 \angle_{\mathsf{B}}(\mathcal{U},\mathcal{V})}{\sin^2 \angle(\mathcal{U},\mathcal{V})} \leq \kappa(\mathsf{B}). \tag{1.15}$$

Multiplying (1.15) by $\sin^2 \angle(\mathcal{U},\mathcal{V})$ yields

$$\frac{1}{\kappa(\mathsf{B})} \sin^2 \angle(\mathcal{U},\mathcal{V}) \leq \sin^2 \angle_{\mathsf{B}}(\mathcal{U},\mathcal{V}) \leq \kappa(\mathsf{B}) \sin^2 \angle(\mathcal{U},\mathcal{V}),$$

then taking square roots finally gives

$$\frac{1}{\sqrt{\kappa(\mathsf{B})}} \sin \angle(\mathcal{U},\mathcal{V}) \leq \sin \angle_{\mathsf{B}}(\mathcal{U},\mathcal{V}) \leq \sqrt{\kappa(\mathsf{B})} \sin \angle(\mathcal{U},\mathcal{V}).$$

Note that $\sqrt{\kappa(\mathsf{B})} = \kappa(\mathsf{K})$, where K is the Cholesky factor or square root of B.

In [57] sharper bounds can be found for measuring angles between an eigenspace of a matrix pair (A,B) and another subspace. The results are expressed in terms of the tangent of the angle and hence do not fit into the framework for angles we used here. It is however worth noting that recently expressions for the tangents of angles between subspaces based on the singular values of certain matrices have been published [123].

1.4 Eigenproblems and their numerical solution

This section is about different kinds of eigenproblems. It also includes a short overview of numerical methods for the solution of some eigenproblems.

1.4.1 Types of eigenproblems

Equations (1.2) or (1.3) do not represent "problems" in the first instance. They become problems when trying to actually solve them numerically, given the matrix or matrix pair as input data. Still, it is not clear which data actually should be computed. Here is an incomplete list of what one could ask for when solving a generalized definite "eigenproblem"

$$\mathsf{A}\mathsf{x} = \mathsf{B}\mathsf{x}\lambda,\ \mathsf{A},\mathsf{B} \in \mathbb{C}^{n\times n}: \tag{1.16}$$

[1] Thanks to Andrew Knyazev for explaining parts of the Russian text.

- Find one solution (x, λ) of (1.16).
- Find n solutions of (1.16), i. e., a matrix $\mathsf{X} \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\mathsf{\Lambda} \in \mathbb{C}^{n\times n}$ such that $\mathsf{AX} = \mathsf{BX\Lambda}$ (the so called full eigenproblem).
- Find all eigenvalues of (A, B).
- Find $k < n$ eigenpairs with smallest/largest eigenvalues, with eigenvalues with largest/smallest magnitude,... (partial eigenproblem).
- Find the $k < n$ eigenvalues closest to a given "target" value τ.
- ...

Of course, the choice of a numerical method for the solution of one of the above problems depends on several other factors. Is the matrix (pair) in question real or complex? Is it dense or sparse (how many zeros do the matrices contain and can we make use of them)? On which hardware should the method run? How large is the matrix dimension n?

This thesis is concerned with a problem slightly different than those mentioned before. Specifically, the considered problem is to find all eigenpairs of a matrix pair (A, B) where the eigenvalue resides in a given region $I_\lambda \subset \mathbb{C}$. We will deal with the definite problem having real eigenvalues. This means that I_λ can be a compact interval with boundaries $\underline{\lambda}$, $\overline{\lambda}$, i. e.,

$$I_\lambda = \left[\, \underline{\lambda}, \overline{\lambda} \,\right].$$

Consequently, we are confronted with the following eigenproblem.

$$\text{Find all solutions of the equation } \mathsf{Ax} = \mathsf{Bx}\lambda \text{ such that } \lambda \in I_\lambda. \tag{1.17}$$

More precisely, by (1.17) we mean the search for a system $\{\mathsf{x}_1, \ldots, \mathsf{x}_k\}$ consisting of B-orthonormal eigenvectors, i. e., $\mathsf{x}_i^\star \mathsf{B} \mathsf{x}_j = \delta_{ij}$, and the corresponding eigenvalues.

In order to distinguish between a specific computational eigenvalue problem such as (1.17) and the corresponding equation such as (1.3), we will call the equation itself an *eigenequation*. The terms standard (eigen)equation, generalized (eigen)equation, definite (eigen)equation and so on are then defined. Hence, for every eigenproblem there is one eigenequation, but for an eigenequation there are many eigenproblems, see the list above. When talking about the generalized problem or the standard problem we mean one arbitrary instance of the corresponding problem.

The fact that the eigenvalues of the definite equation are real leads to a significant benefit when solving such problems. Suppose we have a method for solving (1.17). We then can divide our so called *search interval* I_λ into K smaller parts $I_\lambda^{(k)}$, resulting in

$$I_\lambda = I_\lambda^{(1)} \cup I_\lambda^{(2)} \cup \cdots \cup I_\lambda^{(K)},$$

and solve problem (1.17) *independently* for every $I_\lambda^{(k)}$, $k = 1, \ldots, K$. See also Section 3.6.2. If further information on lower and upper bounds of the spectrum is available i. e., numbers $\underline{\tau}$, $\overline{\tau}$ such that

$$\begin{aligned} \underline{\tau} &\leq \min \operatorname{spec}(\mathsf{A}, \mathsf{B}) \\ \overline{\tau} &\geq \max \operatorname{spec}(\mathsf{A}, \mathsf{B}), \end{aligned}$$

we can choose the search interval I_λ to contain the whole spectrum, as well.

Summary

We can define eigenproblems for an eigenequation. The solution of an eigenproblem depends heavily on the nature of the problem and on that one of the equation.

1.4.2 Types of eigensolvers

This section is intended to give a short overview of the two classes of eigensolvers, direct and iterative methods.

Direct solvers

A very broad class of eigensolvers implies the so called *direct* eigensolvers. They are characterized by the fact that they "almost" always work correctly and in predictable runtime. Possibly the term "direct" is misleading, since every eigensolver has to be iterative in a sense, as it computes the zeros of a certain polynomial.

Important members of the class of direct solvers are the classical QR/QZ algorithms for the solution of the full eigenproblem of (1.2), (1.3), respectively. Developed over 50 years ago [31], the QR algorithm is still one of the most frequently used methods for the full eigenproblem of the standard equation with an unsymmetric matrix [113]. Often Wilkinson's monograph [118] is mentioned as one of the first books giving a comprehensive overview of the numerical solution of eigenvalue problems, including a convergence analysis of the QR algorithm. A very robust and fast implementation of the QR algorithm can nowadays be found in LAPACK [5]. Similar applies to the QZ algorithm for the full problem of the generalized equation, see [71]. Both algorithms have time complexity $\mathcal{O}(n^3)$ for the solution of the full eigenproblem; it does not decrease when only parts of the eigensystem are wanted.

Concerning the symmetric/Hermitian equation or generalized definite equation, more efficient methods than QR are available. After the reduction to tridiagonal form, the Divide and Conquer algorithm [15, 39] can be applied, which has proven to be a very fast and stable method with $\mathcal{O}(n^{2.5})$ time complexity in practice, see [22]. D&C in its original form is only able to compute the complete

eigensystem but recently a way for computing only parts of the eigensystem has been published [8].

In recent years, the so called MRRR algorithm became more popular since it gained speed and robustness. It was first presented in [23], for recent developments see, e. g., [119]. It can be numbered among the direct solvers due to its robustness, while performing modified inverse iteration (which in turn could be counted to the iterative methods).

For very large matrix sizes n (at the time of this writing, "very large" is several million, $\approx 10^7 - 10^8$), some direct methods are not applicable due to their nature or due to the nature of the eigenproblem that is to solve. The full eigenproblem requires the storage of the full matrix of eigenvectors, which has storage complexity $\mathcal{O}(n^2)$. Also, a reduction to tridiagonal form is needed for most of the methods mentioned before, which also requires n^2 storage for the transformation matrices and $\mathcal{O}(n^3)$ runtime. Some methods, such as QR, have a runtime that is cubic with the matrix size, which is far too much for very large matrices.

All direct solvers for the symmetric problem have in common that they are able to compute a full eigen-decomposition, i. e., an unitary matrix X and a diagonal matrix Λ such that

$$\mathsf{X}^\star \mathsf{A} \mathsf{X} = \Lambda.$$

Direct solvers also play an important role as auxiliary methods in the development and implementation of iterative solvers, this is why they were mentioned here.

Iterative solvers

The core topic of this thesis are the so called iterative solvers. The name can be explained in several ways. First, the term "iterative" implies that one has to care about convergence issues, meaning the method might not converge or not converge to desired accuracy. Further, iterative methods can be characterized by the fact that they compute approximate solutions to the problem at every step of the iteration. Hence, if such a method is stopped at any point in time, one already can hope for a meaningful output, which in general is not the case for direct solvers.

A prominent member of the class of iterative solvers is, for instance, the implicitly restarted Arnoldi method [96] that is also implemented in the software package ARPACK [66]. Another well-known method is the so called Jacobi–Davidson method, first introduced by Sleijpen and Van der Vorst in 1996 [95]. All these methods were first introduced for the partial problem of the standard equation. They have been adapted for the solution of the corresponding problem of the generalized equation, see e. g., [94].

Alternatively, we can define iterative eigensolvers as *subspace methods*. This means they are based on the approximation of an eigenspace, which then yields

Equation	Full problem/ large fraction of eigenpairs, large n	partial problem large n	Full problem, small n	partial problem, small n
$\mathsf{A}\mathsf{x} = \mathsf{x}\lambda$	??	Arnoldi/Lanczos, JD	QR	MRRR, D&C
$\mathsf{A}\mathsf{x} = \mathsf{B}\mathsf{x}\lambda$	??	Arnoldi/Lanczos, JD	QZ	MRRR, D&C

Table 1.1: Different methods for different eigenproblems. The MRRR and D&C methods are also applicable to the full problem, while the QR and QZ methods can be used in the non-Hermitian case, too. A large fraction of eigenpairs could for instance be 50% of all pairs.

approximate eigenvalues and eigenvectors via a Rayleigh–Ritz process. The details are left to Chapter 2.

For some eigenproblems we collected suitable methods in Table 1.1. The question marks in the first column indicate that this problem is, to be vague, difficult to solve. The direct methods are not applicable due to memory limitations and a reliable iterative method does not exist, at least to the best of our knowledge. The QZ method for the full problem $\mathsf{A}\mathsf{x} = \mathsf{B}\mathsf{x}\lambda$ is also applicable to matrix pairs that are not definite, A and B can be any two square matrices. The MRRR and D&C methods are only applicable to the generalized problem after having it brought to standard form.

1.5 Measures for the quality of an eigensolver

Whenever designing a numerical method one has to apply criteria that measure the quality of the method. This is necessary in order to evaluate if the method is correct at all (i. e., if it computes the correct quantities) and in order to compare it to similar methods. Another important measure is the speed or efficiency and to which extent a method exploits the hardware.

For eigensolvers that solve one of the problems from Section 1.4.1 it is not that easy to identify unified criteria, since different methods solve different problems as indicated. Furthermore, some criteria depend on the underlying computer architecture. These include, amongst others, the speed of the method and the achievable accuracy (though we suppose that all computations are performed in IEEE double precision [47, 48]). A good numerical method for eigenvalue/eigenvector computations will—like all numerical methods—perform in a good balance of speed, accuracy and reliability. Additionally, memory requirements have to be considered in practice.

1.5.1 Accuracy

In eigenvalue methods we have to assess whether the computed quantities are "correct enough", in a sense. An *accurate* result is one that does not deviate too much from the quantity that we actually wanted to compute. In the best case, a concrete error bound for the difference between the computed and the exact quantity can be found. Here, it has to be specified what is meant by "difference". For instance, for eigenvalues (eigenvectors) we want to measure the distance to the next exact eigenvalue (eigenvector), which is not available since the exact quantities are unknown. However, upper bounds for such errors can be formulated, see Section 2.1.5. When solving eigenvalue problems, one usually has to revert to *residuals* and measure orthogonality.

In the rest of this text we distinguish between exact and computed quantities whenever necessary. We might then talk of, e. g., an exact eigenvalue or a computed eigenvalue. The corresponding symbol of the computed quantity is marked with a tilde on top, for instance $\widetilde{\lambda}$ is the computed counterpart of λ and so on.

Residuals

Let us discuss which quantities can be measured directly. Suppose we are given a definite matrix pair (A, B) of order n and a certain eigenproblem whose solution is a pair $(\widetilde{\mathsf{X}}, \widetilde{\Lambda})$ with a matrix $\widetilde{\mathsf{X}}$ containing $m \leq n$ (computed) eigenvectors and a diagonal matrix $\widetilde{\Lambda}$ containing the corresponding m (computed) eigenvalues.

We then want our computed quantities to fulfill $\mathsf{A}\widetilde{\mathsf{X}} = \mathsf{B}\widetilde{\mathsf{X}}\widetilde{\Lambda}$ or, equivalently, $\|\mathsf{A}\widetilde{\mathsf{X}} - \mathsf{B}\widetilde{\mathsf{X}}\widetilde{\Lambda}\| = 0$ in any matrix norm. This will in general not be the case, hence we measure the absolute, blockwise residual norm

$$\mathsf{res} := \left\| \mathsf{A}\widetilde{\mathsf{X}} - \mathsf{B}\widetilde{\mathsf{X}}\widetilde{\Lambda} \right\|$$

Also, the residual for each single eigenpair can be computed. As the residual is an absolute number and is expected to grow with m, n, $\|\mathsf{A}\|$, $\|\mathsf{B}\|$ or the absolute values of the eigenvalues that are computed, we can replace it by some relative value, e. g., $\|\mathsf{A}\widetilde{\mathsf{X}} - \mathsf{B}\widetilde{\mathsf{X}}\widetilde{\Lambda}\| / \|\mathsf{A}\|$.

In a computer implementation of an eigenvalue solver we need some stopping criterion. It is usually based on the per-eigenpair residual, e. g., we can check for the condition

$$\left\| \mathsf{A}\widetilde{\mathsf{x}} - \mathsf{B}\widetilde{\mathsf{x}}\widetilde{\lambda} \right\|_2 \leq \|\mathsf{A}\|_2 \times n \times \mathsf{tol},$$

where $\mathsf{tol} \geq \varepsilon_M$ is some tolerance supplied by the user. See also [60] and Section 3.6.4 below.

Orthogonality

When solving the definite generalized eigenproblem for the matrix pair (A, B), we wish to compute eigenvectors that are as orthogonal as possible. Supposing

all computed vectors $\widetilde{\mathsf{x}}_j,\ j = 1, \ldots, m$ are normalized, i. e., $\|\widetilde{\mathsf{x}}_j\|_{\mathsf{B}} = 1$ for all j, we want to have

$$\widetilde{\mathsf{x}}_i^\star \mathsf{B} \widetilde{\mathsf{x}}_j = \delta_{ij}, \tag{1.18}$$

as the theory suggests. As expected, this will rarely be the case so we have to revert to the requirement that the left hand side of (1.18) is "small" for $i \neq j$. In practice, we here can hope for

$$|\widetilde{\mathsf{x}}_i^\star \mathsf{B} \widetilde{\mathsf{x}}_j| = \mathcal{O}(n\varepsilon_M),\ i \neq j \tag{1.19}$$

in the very best case. In the standard case ($\mathsf{B} = \mathsf{I}$) we only have $|\widetilde{\mathsf{x}}_i^\star \widetilde{\mathsf{x}}_j| = \mathcal{O}(n\varepsilon_M)$ even if the vectors $\widetilde{\mathsf{x}}_i, \widetilde{\mathsf{x}}_j$ are orthogonal in *exact* arithmetic, at least as long as $n\varepsilon_M \leq 1.01$ [36, p. 63]. Hence, the numerical evaluation of the right hand side of (1.19) will cause errors of order $n\varepsilon_M$ in that case. In the generalized case with $\mathsf{B} \neq \mathsf{I}$, the norm of B appears as factor on the right hand side of (1.19). In this consideration the actual errors in the computed eigenvectors are not taken into account. Normally, the computed eigenvectors are of course not orthogonal in exact arithmetic. We will refine our measures for orthogonality later (Section 3.6.3).

1.5.2 Reliability

The reliability, also called robustness, of a method is something that cannot be measured as simply as the accuracy just with some numbers. Reliability means the overall capability to deliver correct results for correct inputs, detect wrong inputs and flag wrongly computed outputs with a clear error message. Maybe the last point is the most important one because "the unpardonable sin is for a method to lie, to deliver results which appear to be reasonable but which are utterly wrong", [80, p. 14].

Also, a method should be robust against "hard" problems. Of course, these problems should still be solvable in reasonable time and with sufficient accuracy. "Hard" can mean, for instance, in the context of eigenvalue problems, matrices with very small eigenvalues or eigenvalues that are very close. Very close eigenvalues will informally be called "clustered" in the following.

Reliability can hardly be quantified. It is usually assessed with tests based on statistics, meaning a method is applied to a broad class of problems. Then for instance the overall number of failures is measured. Furthermore, it can be assessed by applying the method to outstanding hard problems.

Chapter 2

General theory of contour integration based eigensolvers

Synopsis

This chapter deals with general techniques for the solution of eigenvalue problems by means of contour integration. As stated in Chapter 1, we focus on the eigenvalue problem

$$\mathsf{A}\mathsf{X} = \mathsf{B}\mathsf{X}\Lambda,\ \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_k),\ \lambda_j \in I_\lambda, \tag{2.1}$$

where $I_\lambda \subset \mathbb{C}$ is some subset. Because we suppose (A, B) to be a definite pair, the set I_λ will be chosen as a closed real interval if not stated otherwise. The aim of this chapter is to review and analyze techniques for solving (2.1) that are based on numerical integration and subspace iteration.

The main ingredient of all techniques is the evaluation of a contour integral

$$\mathsf{U} := \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{Y}\mathrm{d}z,$$

where $\mathsf{Y} \in \mathbb{C}^{n\times m}$ and $\mathcal{C}$ is a contour in the complex plane around the desired part of the spectrum of the matrix pair (A, B). The integral then is used to form a subspace $\mathcal{U} = \operatorname{span}(\mathsf{U})$ of modest dimension, from which eigenvectors and eigenvalues of (A, B) are extracted.

To understand the method, in Section 2.1 we first introduce subspace methods in detail, while not supposing a special structure of the subspace. We generalize some well-known results that can only be found for a single matrix in literature to the case of a generalized eigenproblem with a matrix pair (A, B). Here, most geometric notions are in terms of the B-norms and angles from Sections 1.1 and 1.3.

To make the connection to the special eigensolvers based on numerical integration, we review some notions and results from complex analysis in Section 2.2.

Year	Method	Selected reference
1950	Lanczos Algorithm	[64]
1951	Arnoldi Algorithm	[6]
1992	Implicitly restarted Arnoldi	[96]
1996	Jacobi-Davidson	[95]
2002	Krylov-Schur	[101]

Table 2.1: Milestones in subspace eigenvalue algorithms.

In Section 2.3 we recall some well-known techniques from the field of numerical integration that we will make use of later. The exposition includes some error bounds.

In Section 2.4 we come to the actual integral based eigensolver, combining the methods from subspace eigensolvers and numerical integration. In Section 2.5 we analyze the errors that occur in the course of the integration based algorithm. Ultimately, in Section 2.6 we give a conclusion of the chapter.

2.1 Subspace eigensolvers

In this section, we give an overview of subspace based eigensolvers, one of which is the integration algorithm at the core of this thesis. We named such methods "iterative methods" in Chapter 1. Here, we will introduce them in detail in a general framework based on the so called Rayleigh–Ritz theorem.

Subspace eigensolvers have a long history that dates back at least to Lanczos (1950, [64]) and Arnoldi (1951, [6]). They are therefore even older than, e. g., the QR algorithm.[1] In particular the descendants of Lanczos' and Arnoldi's method still are widely used and there exists a broad literature. To spare the reader details, we compiled some important milestones in the development of subspace eigenvalue algorithms in Table 2.1.

All these algorithms have in common that they try to approximate eigenspaces of the matrix (pair) under inspection as a first and crucial step. The second step is the extraction of eigenpairs from that space. The subspace might be a Krylov subspace, generated by vectors of the form $\mathsf{A}^j\mathsf{v}$ for some starting vector v. This is the case in the Lanczos and Arnoldi methods. Krylov subspaces are not part of this treatise, while they often can be found in the literature. The interested reader can find an introduction in [100].

In [100, Chap. 4.4] as well as in [80] a general theory of subspace methods can

[1] Jacobi introduced methods for the solution of eigenvalue problems already over 100 years earlier [50,51]. Though, he did not have the chance to implement these methods on a computer. However, Jacobi's methods still play an important role in numerical linear algebra.

be found that does not require that the subspace is a Krylov subspace.

2.1.1 Rayleigh–Ritz-method

As mentioned above, subspace methods rely on the Rayleigh–Ritz theory and method, originally published by Rayleigh [87] and Ritz [88], as the name suggests. Rayleigh and Ritz published their respective methods in the context of physics.

We start by stating the so called Rayleigh–Ritz theorem, where we follow the presentation in [100, p. 283]. There it is stated for the standard equation, whereas in [60] we adapted it to the generalized equation.

Theorem 2.1 (Rayleigh–Ritz, [100], [60])
Let $\mathcal{U}$ *be a subspace containing an eigenspace* $\mathcal{X}$ *of the matrix pair* (A, B)*. Let* U *be a basis of* $\mathcal{U}$*. Define*

$$\mathsf{A}_\mathsf{U} = \mathsf{U}^\star \mathsf{A} \mathsf{U},$$
$$\mathsf{B}_\mathsf{U} = \mathsf{U}^\star \mathsf{B} \mathsf{U},$$

the so-called Rayleigh quotients for A *and* B*.*

Then there is an eigenpair (W, Λ) *of* $(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$ *such that* (UW, Λ) *is an eigenpair of* (A, B) *and* $\operatorname{span}(\mathsf{UW}) = \mathcal{X}$*.*

Proof. Let (X, Λ) be an eigenpair of (A, B) corresponding to $\mathcal{X}$, i. e., $\mathcal{X} = \operatorname{span}(\mathsf{X})$. Since $\mathcal{U} \supset \mathcal{X}$ we can express X via U as $\mathsf{X} = \mathsf{UW}$. By definition of W we have

$$\mathsf{AUW} = \mathsf{BUW}\Lambda$$

and hence

$$\mathsf{A}_\mathsf{U}\mathsf{W} = \mathsf{U}^\star\mathsf{AUW} = \mathsf{U}^\star\mathsf{BUW}\Lambda = \mathsf{B}_\mathsf{U}\mathsf{W}\Lambda,$$

meaning that (W, Λ) is an eigenpair of $(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$. □

Note that in the proof we do not use that we left-multiply by the transpose of U. In principle, we could use any other matrix as the left factor of the Rayleigh quotients, but the choice $\mathsf{U}^\star$ keeps the Rayleigh quotients Hermitian if the original matrices were so. This is important for practical reasons. Note further that if U was chosen to be B-orthogonal, it reduces the original generalized eigenequation to a small scale *standard* one, since then $\mathsf{B}_\mathsf{U} = \mathsf{I}$.

In practice one will rarely find a subspace that contains an exact eigenspace since there are only 2^n of those (one for each subset of the spectrum). The idea of the Rayleigh–Ritz method is to use the theorem as basis for an approximation with spaces $\mathcal{U}$ that only contain approximate eigenspaces [100]. The procedure which can be derived [60, 100] is presented in Algorithm 2.1.

We already defined the terms Rayleigh quotient and (primitive) Ritz pair implicitly in the algorithm. A *Ritz vector* is a vector of the form Uw where w is a *primitive Ritz vector*, meaning an eigenvector of the Rayleigh quotients. The

Algorithm 2.1 Rayleigh–Ritz method

1: Find a suitable basis U for $\mathcal{U}$.
2: Compute the *Rayleigh quotients* $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star\mathsf{A}\mathsf{U}$, $\mathsf{B}_\mathsf{U} = \mathsf{U}^\star\mathsf{B}\mathsf{U}$.
3: Compute the *primitive Ritz pairs* $(\widetilde{\mathsf{W}}, \widetilde{\Lambda})$ of $\mathsf{A}_\mathsf{U}\mathsf{W} = \mathsf{B}_\mathsf{U}\mathsf{W}\Lambda$.
4: Return the approximate *Ritz pairs* $(\mathsf{U}\widetilde{\mathsf{W}}, \widetilde{\Lambda})$ of $\mathsf{A}\mathsf{X} = \mathsf{B}\mathsf{X}\Lambda$.
5: Check convergence criterion; if not satisfied, go back to Step 1.

corresponding eigenvalue is called Ritz value. Note that all quantities are *with respect to* U. Actually, they do not depend on the concrete choice for the basis U but on the subspace that is spanned by it. A procedure of the type above is, following Stewart [100], a *Rayleigh-Ritz procedure*.

2.1.2 Subspace iteration

In the description of the Rayleigh–Ritz method, we naively wrote about a subspace $\mathcal{U}$ that is chosen somehow. Now, let us present a very simple method for computing such a subspace, the so called *subspace iteration*. It will play an important role later on. As motivation, consider the power method [37, Sec. 7.3.1]. Given a (not necessarily Hermitian) matrix A and a vector q with $\|\mathsf{q}\|_2 = 1$, the iteration

$$\begin{aligned} \mathsf{z} &:= \mathsf{A}\mathsf{q}, \\ \mathsf{q} &:= \mathsf{z}/\left\|\mathsf{z}\right\|_2, \\ \lambda &:= \mathsf{q}^\star\mathsf{A}\mathsf{q} \end{aligned}$$

is repeatedly performed. If the eigenvalues of A can be ordered as $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$ and the initial vector q has components in the direction of the eigenvector belonging to λ_1 it can be shown that λ converges towards λ_1. For the k-th iterate $\lambda^{(k)}$ we then have $\left|\lambda_1 - \lambda^{(k)}\right| = \mathcal{O}(|\lambda_2/\lambda_1|^k)$ [37].

This iteration can be performed for whole subspaces, too. This leads to subspace iteration, which can be found basically in any book on numerical linear algebra and can also be seen as the basis for the QR algorithm [108]. In Algorithm 2.2 we present subspace iteration as it can be found in similar form in [91, p. 115] (we present it for the generalized case).

Comparing Algorithm 2.2 with the Rayleigh–Ritz method from Algorithm 2.1 shows that the two methods are very similar. Indeed, in Lines 2–4 of Algorithm 2.1 and lines 5–7 of Algorithm 2.2 basically the same computations are performed. Eigenvalues of (A, B) can of course also be approximated in subspace iteration by the eigenvalues of $(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$. The difference to Rayleigh–Ritz is that subspace iteration provides a way to actually compute the basis, while Rayleigh–Ritz is more a general framework.

Algorithm 2.2 Subspace iteration

1: Choose initial vectors $\mathsf{U}^{(0)} \in \mathbb{C}^{n\times m},\ m \leq n$.
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad \mathsf{Z} := \mathsf{A}\mathsf{U}^{(k-1)}$
4: $\quad \mathsf{U}^{(k)}\mathsf{R} = \mathsf{Z}$ (QR factorization w.r.t. B-scalar product)
5: $\quad$ Set $\mathsf{A}_\mathsf{U} := \mathsf{U}^{(k)\star}\mathsf{A}\mathsf{U}^{(k)}$, $\mathsf{B}_\mathsf{U} := \mathsf{U}^{(k)\star}\mathsf{B}\mathsf{U}^{(k)}$ $(= \mathsf{I}_m)$
6: $\quad$ Compute eigenvectors W of $(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$
7: $\quad$ Set $\mathsf{U}^{(k)} := \mathsf{U}^{(k)}\mathsf{W}$

If the pair $(\mathsf{A},\, \mathsf{B})$ has m eigenvalues that are dominant in absolute value, say,

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_m| > |\lambda_{m+1}| \geq \cdots \geq |\lambda_n|\,,$$

it can be shown that subspace iteration converges, for details see [91, Thm. 5.2].

2.1.3 Eigenvalue bounds

We discuss convergence of Ritz values for generalized eigenvalue problems.

A famous theorem by Hermann Weyl [116] is the basis for our estimates of the error in the eigenvalues. It asserts that the eigenvalues of a perturbed matrix do not differ from those of the original matrix more than the norm of the perturbation. It can be formulated as follows (see [103, Cor. 4.10]).

Theorem 2.2 (Weyl)
Let A *and* $\mathsf{A}+\mathsf{E}$ *be Hermitian. Let* $\lambda_j,\ \widetilde{\lambda}_j,\ j = 1, \ldots, n$ *denote the eigenvalues of* A *and* $\mathsf{A}+\mathsf{E}$*, respectively and let both sequences be ordered ascendingly. Then*

$$\max_j \left|\widetilde{\lambda}_j - \lambda_j\right| \leq \|\mathsf{E}\|\,.$$

Let us continue with a backward perturbation result, adapted from [100]. Suppose a subspace $\mathcal{U} = \operatorname{span}(\mathsf{U})$ with B-orthonormal U is given and an eigenvector of $(\mathsf{A},\, \mathsf{B})$ is near the subspace in the sense that the angle $\theta := \angle_\mathsf{B}(\mathsf{x}, \mathsf{U})$ is small.

Theorem 2.3 (Adapted from [100, Ch. 4, Thm. 4.4])
Let (x, λ) *be an eigenpair of* $(\mathsf{A},\, \mathsf{B})$*. Let the Rayleigh quotients* $\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U}$ *be given (*$\mathsf{B}_\mathsf{U} = \mathsf{I}$*) and* $\theta = \angle_\mathsf{B}(\mathsf{x}, \mathsf{U})$*. Let* K *be a matrix with* $\mathsf{K}^\star\mathsf{K} = \mathsf{B}$*. Then there is a matrix* E_θ *such that*

$$\|\mathsf{E}_\theta\|_2 \leq \frac{\sin(\theta)}{\sqrt{1-\sin^2(\theta)}}\,\|\mathsf{A}\|\,\left\|\mathsf{K}^{-1}\right\|^2$$

and such that λ *is an eigenvalue of the pair* $(\mathsf{A}_\mathsf{U} + \mathsf{E}_\theta, \mathsf{B}_\mathsf{U})$*.*

Proof. We follow the proof of Theorem 4.4 in Chapter 4 of [100]. Complement U to a B-unitary matrix $[\mathsf{U}, \mathsf{U}_\perp]$ and let

$$\mathsf{y} = \mathsf{U}^\star \mathsf{B}\mathsf{x},\ \mathsf{z} = \mathsf{U}_\perp^\star \mathsf{B}\mathsf{x}.$$

We then have that $\|\mathsf{y}\|_2 = \cos\theta = \sqrt{1-\sin^2\theta},\ \|\mathsf{z}\|_2 = \sin\theta$. By the prerequisites the eigenequation $\mathsf{A}\mathsf{x} - \mathsf{B}\mathsf{x}\lambda = \mathsf{o}$ holds. Multiplying this by $\mathsf{U}^\star$ from the left and plugging in the identity $\mathsf{I}_n = [\mathsf{U},\, \mathsf{U}_\perp][\mathsf{U},\, \mathsf{U}_\perp]^\star \mathsf{B}$, we obtain

$$\mathsf{U}^\star \mathsf{A}\,[\mathsf{U}, \mathsf{U}_\perp] \begin{bmatrix} \mathsf{U}^\star \\ \mathsf{U}_\perp^\star \end{bmatrix} \mathsf{B}\mathsf{x} - \mathsf{U}^\star \mathsf{B}\mathsf{x}\lambda = \mathsf{o}.$$

By forming the matrix products, we obtain

$$\mathsf{U}^\star \mathsf{A}(\mathsf{U}\mathsf{U}^\star \mathsf{B} + \mathsf{U}_\perp \mathsf{U}_\perp^\star \mathsf{B})\mathsf{x} - \mathsf{U}^\star \mathsf{B}\mathsf{x}\lambda = \mathsf{o}.$$

This results in (remember definitions of $\mathsf{y},\ \mathsf{z},\ \mathsf{A}_\mathsf{U}$)

$$\mathsf{A}_\mathsf{U}\mathsf{y} + \mathsf{U}^\star \mathsf{A}\mathsf{U}_\perp \mathsf{z} - \mathsf{y}\lambda = \mathsf{o}.$$

By normalizing y to have 2-norm 1, $\hat{\mathsf{y}} = \mathsf{y}/\sqrt{1-\sin^2\theta}$, the residual is hence

$$\begin{aligned} \mathsf{r} &= \mathsf{A}_\mathsf{U}\hat{\mathsf{y}} - \hat{\mathsf{y}}\lambda \\ &= -\frac{1}{\sqrt{1-\sin^2\theta}} \mathsf{U}^\star \mathsf{A}\mathsf{U}_\perp \mathsf{z}. \end{aligned} \tag{2.2}$$

Next, let K be a matrix with $\mathsf{K}^\star\mathsf{K} = \mathsf{B}$. Note that $\mathsf{U} = \mathsf{K}^{-1}(\mathsf{K}\mathsf{U})$, where $\mathsf{K}\mathsf{U}$ is orthonormal. Consequently, $\|\mathsf{U}\| \leq \|\mathsf{K}^{-1}\|$. The same holds for $\mathsf{U}_\perp$. Now, taking norms on both sides of (2.2) yields

$$\|\mathsf{r}\| \leq \frac{\sin\theta}{\sqrt{1-\sin^2\theta}} \|\mathsf{A}\| \cdot \left\|\mathsf{K}^{-1}\right\|^2.$$

Setting $\mathsf{E}_\theta = -\mathsf{r}\hat{\mathsf{y}}^\star$ yields a matrix with $\|\mathsf{E}_\theta\| \leq \|\mathsf{r}\|$ and $(\mathsf{A}_\mathsf{U} + \mathsf{E}_\theta)\hat{\mathsf{y}} = \mathsf{A}_\mathsf{U}\hat{\mathsf{y}} - \mathsf{r}\hat{\mathsf{y}}^\star\hat{\mathsf{y}} = \mathsf{A}_\mathsf{U}\hat{\mathsf{y}} - \mathsf{r} = \lambda\hat{\mathsf{y}}$. This is the desired result. □

With this backward perturbation result an approximation bound for λ in the spirit of Elsner's famous theorem [29] can be derived. By using a generalization of the theorem [102], we can obtain the following error bound.

Corollary 2.4
Let the prerequisites be as in Theorem 2.3. Then there is an eigenvalue $\mu \in \operatorname{spec}(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$ *such that*

$$|\lambda - \mu| \leq \frac{\sqrt{\|\mathsf{A}_\mathsf{U}\|_2^2 + \|\mathsf{B}_\mathsf{U}\|_2^2}^{1-\frac{1}{m}} \|\mathsf{E}_\theta\|_2^{\frac{1}{m}}}{\max_{\|(\alpha,\beta)\|=1} \sigma_{\min}(\beta\mathsf{A}_\mathsf{U} - \alpha\mathsf{B}_\mathsf{U})}, \tag{2.3}$$

where m *denotes the order of* A_U *and* B_U *and* $\sigma_{\min}$ *denotes the smallest singular value.*

In practice the complicated looking denominator of (2.3) can be replaced by $\min_j \sqrt{|\mu_j|^2 + 1}$, where μ_j, $j = 1, \ldots, m$ denote the eigenvalues of $(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$, see [102]. Note that the derived bounds are valid for general matrices A.

In our case with A Hermitian and B Hermitian positive definite, we can use Weyl's theorem (Theorem 2.2) to ensure the existence of a Ritz value μ with

$$|\mu - \lambda| \leq \|\mathsf{E}_\theta\| ,$$

similar to the standard case [100, p. 288]. See Theorem 2.9 below with $\mathsf{F} = \mathsf{0}$.

Additive perturbations of the subspace

Now, suppose we have a subspace $\mathcal{U}$ with basis U at hand that is to approximate an eigenspace $\mathcal{X}$ with basis X. In this thesis, we will state bounds for $\|\mathsf{X} - \mathsf{U}\|$ in Section 2.5. It is consequently important to know how the computed quantities—Ritz values and Ritz vectors—behave if the basis is changed. This knowledge can of course be best used if the computed basis U is interpreted as a perturbed exact basis X.

Let us begin with results that depend on the difference $\mathsf{X} - \mathsf{U}$. One of those is by Knyazev and Argentati [59, Thm. 9]: Let A be a Hermitian $n \times n$ matrix and X, U be full rank matrices of size $n \times m$ where $m \leq n$. Let α_j, β_j, $j = 1, \ldots, m$ denote the Ritz values of A with respect to X and U, respectively, both ordered ascendingly. Then,

$$\max_{j=1,\ldots,m} |\alpha_j - \beta_j| \leq (\lambda_{\max} - \lambda_{\min}) \kappa(\mathsf{X}) \frac{\|\mathsf{X} - \mathsf{U}\|}{\|\mathsf{X}\|} .$$

If X is orthonormal, the statement of this theorem boils down to:

Theorem 2.5 (Knyazev and Argentati, [59])
Let A be Hermitian and the notation as above. Then

$$\max_{j=1,\ldots,m} |\alpha_j - \beta_j| \leq (\lambda_{\max} - \lambda_{\min}) \|\mathsf{X} - \mathsf{U}\| ,$$

if X is a matrix with orthonormal columns.

As a simple consequence we obtain the following result.

Corollary 2.6 (Knyazev and Argentati—Generalized version)
Let (A, B) be a definite matrix pair, let X be a matrix with B-orthonormal columns. Let α_j, β_j, $j = 1, \ldots, m$ be the Ritz values of (A, B) with respect to X and U, respectively. Let both sequences of Ritz values be ordered ascendingly. Then

$$\max_{j=1,\ldots,m} |\alpha_j - \beta_j| \leq (\lambda_{\max} - \lambda_{\min}) \|\mathsf{X} - \mathsf{U}\|_{\mathsf{B}2} .$$

Proof. Just apply Theorem 2.5 to the matrix $\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2}$ with the matrices $\mathsf{B}^{1/2}\mathsf{U}$, $\mathsf{B}^{1/2}\mathsf{X}$. □

If X is also a basis for the eigenspace $\mathcal{X}$, the theorem limits the maximum approximation error of the Ritz values compared to the exact eigenvalues, provided an upper bound for $\|\mathsf{X} - \mathsf{U}\|$ is known.

Next, suppose we have an error bound $\varepsilon := \|\mathsf{X} - \mathsf{U}\|$ accessible. Let us interpret the Rayleigh quotient $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star\mathsf{A}\mathsf{U}$ as a perturbed Rayleigh quotient of the "exact" quotient $\mathsf{A}_\mathsf{X} = \mathsf{X}^\star\mathsf{A}\mathsf{X}$. We make the ansatz $\mathsf{U}^\star\mathsf{A}\mathsf{U} = \mathsf{X}^\star\mathsf{A}\mathsf{X} + \mathsf{E}$, where E denotes the error as the symbol suggests. We obtain

$$\begin{aligned}\mathsf{E} &= \mathsf{U}^\star\mathsf{A}\mathsf{U} - \mathsf{A}_\mathsf{X} \\ &= \mathsf{U}^\star\mathsf{A}\mathsf{U} - \mathsf{X}^\star\mathsf{A}\mathsf{X} \\ &= \mathsf{A}_{\mathsf{X}-\mathsf{U}} - 2\mathsf{A}_\mathsf{X} + \mathsf{U}^\star\mathsf{A}\mathsf{X} + \mathsf{X}^\star\mathsf{A}\mathsf{U} \\ &= \mathsf{A}_{\mathsf{X}-\mathsf{U}} + (\mathsf{U}-\mathsf{X})^\star\mathsf{A}\mathsf{X} + \mathsf{X}^\star\mathsf{A}(\mathsf{U}-\mathsf{X}). \end{aligned} \tag{2.4}$$

Hence, we have

$$\|\mathsf{E}\| \leq \|\mathsf{A}\| \, (\varepsilon^2 + 2\varepsilon \, \|\mathsf{X}\|). \tag{2.5}$$

Note that E is Hermitian if A is so. In particular we have, assuming that X is of small norm, e. g., orthonormal, that $\|\mathsf{E}\| = \mathcal{O}(\varepsilon \, \|\mathsf{A}\|)$ for $\varepsilon \to 0$.

Using Theorem 2.2, we obtain the following perturbation bound on the Ritz values of A with respect to U.

Theorem 2.7
Let A be Hermitian and consider the standard equation. Suppose an error bound for the subspace, $\varepsilon := \|\mathsf{X} - \mathsf{U}\|$, is at hand. Let $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k$ denote the Ritz values of A with respect to U and $\lambda_1, \ldots, \lambda_k$ the eigenvalues of A belonging to the space $\mathcal{X}$, both ordered ascendingly. We then have

$$\max_j \left| \widetilde{\lambda}_j - \lambda_j \right| \leq \|\mathsf{A}\| \, (\varepsilon^2 + 2\varepsilon \, \|\mathsf{X}\|).$$

Proof. Use the bound for E (2.5) and the fact that the eigenvalues of A belonging to the space $\mathcal{X}$ are the eigenvalues of A_X. Then use Weyl's Theorem 2.2. □

Note that the error in the theorem does not differ significantly from the one of Theorem 2.5, where the number $\lambda_{\max} - \lambda_{\min}$ might grow up to $2 \, \|\mathsf{A}\|$.

Next, let us come to the case of the generalized definite equation $\mathsf{A}\mathsf{x} = \mathsf{B}\mathsf{x}\lambda$. Here, the same analysis for the perturbation $\mathsf{B}_\mathsf{X} \mapsto \mathsf{B}_\mathsf{U}$ as in eq. (2.4) applies; we obtain $\mathsf{B}_\mathsf{U} = \mathsf{B}_\mathsf{X} + \mathsf{F}$ with

$$\mathsf{F} = \mathsf{B}_{\mathsf{X}-\mathsf{U}} + (\mathsf{U}-\mathsf{X})^\star\mathsf{B}\mathsf{X} + \mathsf{X}^\star\mathsf{B}(\mathsf{U}-\mathsf{X}).$$

Consequently, we have

$$\|\mathsf{F}\| \leq \|\mathsf{B}\| \, (\varepsilon^2 + 2\varepsilon \, \|\mathsf{X}\|). \tag{2.6}$$

Perturbation analysis for perturbations of matrix pairs $(\mathsf{A}, \mathsf{B}) \mapsto (\mathsf{A} + \mathsf{E}, \mathsf{B} + \mathsf{F})$ is available, see e. g., the book [103]. Also Sun's report [104] is a rich source

of information. In most of the literature, for instance in the references above, the analysis is performed for general (possibly not positive definite) B. Then, for perturbation analysis of the eigenvalues, the so called *chordal metric* [103] is employed which allows to treat finite and infinite eigenvalues in a unified way.

Nakatsukasa published a Weyl-style perturbation bound for definite matrix pairs [74]. He notes that the use of the chordal metric is not a very natural choice in this case. The new bound comprises what one would intuitively expect from such a bound. It contains information about B's smallest eigenvalues, since the eigenvalues of $(\mathsf{A},\, \mathsf{B})$ coincide with those of $\mathsf{B}^{-1}\mathsf{A}$.

Theorem 2.8 (Weyl for generalized eigenvalues; Nakatsukasa)
Let $(\mathsf{A},\, \mathsf{B})$ *be a definite matrix pair with eigenvalues* $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. *Let* $\mathsf{E},\, \mathsf{F}$ *be Hermitian and* $\|\mathsf{F}\| < \lambda_{\min}(\mathsf{B})$. *Then the perturbed pair* $(\mathsf{A}+\mathsf{E},\, \mathsf{B}+\mathsf{F})$ *is Hermitian definite. Its eigenvalues* $\widetilde{\lambda}_1 \leq \widetilde{\lambda}_2 \leq \cdots \leq \widetilde{\lambda}_n$ *satisfy*

$$\left|\lambda_j - \widetilde{\lambda}_j\right| \leq \frac{\|\mathsf{E}\|}{\lambda_{\min}(\mathsf{B})} + \frac{\|\mathsf{A}\| + \|\mathsf{E}\|}{\lambda_{\min}(\mathsf{B})(\lambda_{\min}(\mathsf{B}) - \|\mathsf{F}\|)} \|\mathsf{F}\| \,. \tag{2.7}$$

Note that the right hand side of (2.7) is monotonic with $\|\mathsf{E}\|$ and $\|\mathsf{F}\|$. If the eigenequation in question is the standard one, i. e., $\mathsf{F} = \mathsf{0}$ and $\mathsf{B} = \mathsf{I}$, the theorem boils down to Weyl's classic theorem.

Now, we can derive perturbation bounds for Ritz values in the case of a perturbed subspace. Replace $\|\mathsf{E}\|$, $\|\mathsf{F}\|$ in (2.7) by their respective upper bounds (2.5), (2.6). We then obtain, due to the aforementioned monotonicity, the following theorem. The notation is as above.

Theorem 2.9
Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$ *denote the eigenvalues of* $(\mathsf{A}_\mathsf{X},\, \mathsf{B}_\mathsf{X})$. *Let* $\|\mathsf{F}\| < \lambda_{\min}(\mathsf{B}_\mathsf{X})$. *Then, the perturbed pair* $(\mathsf{A}_\mathsf{X}+\mathsf{E},\, \mathsf{B}_\mathsf{X}+\mathsf{F})$ *is Hermitian definite and its eigenvalues* $\widetilde{\lambda}_1 \leq \widetilde{\lambda}_2 \leq \cdots \leq \widetilde{\lambda}_k$ *satisfy*

$$\begin{aligned}&\left|\lambda_j - \widetilde{\lambda}_j\right| \leq \\ &\frac{\|\mathsf{A}\|\,(\varepsilon^2 + 2\varepsilon\,\|\mathsf{X}\|)}{\lambda_{\min}(\mathsf{B}_\mathsf{X})} + \frac{\|\mathsf{A}_\mathsf{X}\| + \|\mathsf{A}\|\,(\varepsilon^2 + 2\varepsilon\,\|\mathsf{X}\|)}{\lambda_{\min}(\mathsf{B}_\mathsf{X})\,(\lambda_{\min}(\mathsf{B}_\mathsf{X}) - \|\mathsf{B}\|\,(\varepsilon^2 + 2\varepsilon\,\|\mathsf{X}\|))} \cdot \|\mathsf{B}\|\,(\varepsilon^2 + 2\varepsilon\,\|\mathsf{X}\|).\end{aligned} \tag{2.8}$$

If X is an eigenspace of $(\mathsf{A},\, \mathsf{B})$, the numbers λ_j from Theorem 2.9 are also eigenvalues of $(\mathsf{A},\, \mathsf{B})$. The theorem states that the approximation error in the Ritz values can be expected to be $\mathcal{O}(\varepsilon)$ (if all other quantities are considered fixed). Looking not very handy at first glance, notice that only eigenvalues of small scale matrices appear in the right hand side of (2.8). The minimum eigenvalue of B_X can be computed with low effort. If X is supposed to be B-orthonormal, we even have $\mathsf{B}_\mathsf{X} = \mathsf{I}$. For the norms of A and B estimates are sufficient as well as for X. The upper bound could hence be monitored in a numerical algorithm in order to

implement a stopping criterion. The computation of the quantity ε for certain subspaces U is the subject of section 2.5.

2.1.4 Convergence of Ritz vectors

So far, we have established some error bounds for eigenvalues. The convergence of Ritz vectors is a more subtle thing. In this context, convergence is not to be understood as the result of an iterative process, but rather as the continuity of Ritz vectors (and complete subspaces) as functions of certain other quantities. For instance, the first question we will address is under which conditions a Ritz vector converges to an eigenvector. The following is independent of the actual method for computing the subspace U.

Convergence of single vectors

Let us fix an eigenpair (x, λ) and suppose we have computed a basis U. In order to emphasize the angle to x let us rename $\mathsf{U}_\theta = \mathsf{U}$, in the style of Stewart [100] and define $\theta := \angle(\mathsf{x}, \mathsf{U}_\theta)$. The important question is if there is a Ritz vector $\mathsf{u} \in \mathsf{U}_\theta$ such that $\angle(\mathsf{x}, \mathsf{u}) \to 0$ as $\theta \to 0$. A small angle θ or even $\theta = 0$ is not sufficient for the answer.

To begin with, a simple bound will be derived that describes the quality of the computed Ritz vector. It is a generalization of Theorem 4.6 of Saad [91] which we state first for a better understanding.

Theorem 2.10
Let P be the orthogonal projector onto the subspace $\mathcal{U}$ used in the Rayleigh–Ritz-procedure. Let $\gamma = \|\mathsf{PA}(\mathsf{I} - \mathsf{P})\|$ and let (x, λ) be any eigenpair of A. Let $\widetilde{\lambda}$ be an approximate eigenvalue extracted from $\mathcal{U}$ and let δ be the distance between λ and the approximate eigenvalues other than $\widetilde{\lambda}$. Then there is an approximate eigenvector $\mathsf{u} \in \mathcal{U}$ associated with $\widetilde{\lambda}$ such that

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

Using this theorem, we can relate the angle between corresponding eigenvector and Ritz vectors to the angles between eigenvectors and the subspace $\mathcal{U}$. Obviously, $\angle(\mathsf{u}, \mathcal{U})$ decreases (at least it does not increase) if $\delta > 0$ when $\dim(\mathcal{U})$ increases. Hence $\angle(\mathsf{x}, \mathsf{u})$ decreases, this is what one would expect. Note that $\gamma \leq \|\mathsf{A}\|$ since (nonzero) projectors have norm 1.

The following theorem is the equivalent to Theorem 2.10 for the generalized case.

Theorem 2.11
Let (x, λ) *be any eigenpair of* (A, B). *Let* $\widetilde{\lambda}$ *be an approximate eigenvalue extracted from* $\mathcal{U}$ *and let* δ *be the distance between* λ *and the approximate eigenvalues other than* $\widetilde{\lambda}$. *Then there is an approximate eigenvector* u *associated with* $\widetilde{\lambda}$ *such that*

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

where $\gamma \leq \|\mathsf{A}\|$.

Proof. We write the generalized eigenequation in standard form as

$$\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2}\mathsf{y} = \mathsf{y}\lambda,$$

obtaining $\mathsf{B}^{1/2}\mathsf{x} = \mathsf{y}$ as eigenvector of the pair (A, B) corresponding to eigenvalue λ. Similarly we obtain the Ritz vector $\mathsf{B}^{1/2}\mathsf{u}$ belonging to the space $\mathsf{B}^{1/2}\mathcal{U}$. Applying Theorem 2.10 yields

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

for some number γ in the first place. Next, let U be a B-orthonormal basis of $\mathcal{U}$. For γ we obtain, similar to Theorem 2.10,

$$\gamma = \|\mathsf{P}\mathsf{A}(\mathsf{I} - \mathsf{P})\|$$

where $\mathsf{P} = (\mathsf{B}^{1/2}\mathsf{U})^{\star}\mathsf{B}^{1/2}\mathsf{U}$ is the orthogonal projector onto the space $\mathsf{B}^{1/2}\mathcal{U}$. Hence, $\gamma \leq \|\mathsf{A}\|$. Using $\angle(\mathsf{B}^{1/2}\mathsf{x}, \mathsf{B}^{1/2}\mathsf{u}) = \angle_{\mathsf{B}}(\mathsf{x}, \mathsf{u})$ and $\angle(\mathsf{B}^{1/2}\mathsf{x}, \mathsf{B}^{1/2}\mathcal{U}) = \angle_{\mathsf{B}}(\mathsf{x}, \mathcal{U})$ (see section 1.3.3) finishes the proof. □

The theorem expresses the angle between approximate and exact eigenvector by means of the angle between exact eigenvector and approximate eigenspace. The key ingredient of the convergence of single vectors is always that the corresponding eigenvalue is well separated from the other eigenvalues. This separation is captured by δ in the preceding theorem. Stewart [100] finds the catchy formula "convergence of the desired eigenvalues + separation of the desired eigenvalues = convergence of the Ritz space". The first point has already been treated in Section 2.1.3 above. Let us now further discuss the second summand of Stewart's formula; for the moment we stay with the standard eigenequation and follow [100].

Let w_θ be a primitive Ritz vector with Ritz value λ_θ and complement w_θ to a unitary matrix $[\mathsf{w}_\theta, \mathsf{W}_\theta]$; it follows

$$\begin{bmatrix} \mathsf{w}_\theta^\star \\ \mathsf{W}_\theta^\star \end{bmatrix} \mathsf{A}_{\mathsf{U}} \, [\mathsf{w}_\theta, \mathsf{W}_\theta] = \begin{bmatrix} \lambda_\theta & \mathsf{o} \\ \mathsf{o} & \mathsf{N}_\theta \end{bmatrix}$$

with some Hermitian matrix N_θ (note that A_U is Hermitian). Suppose that λ_θ is separated from the eigenvalues of N_θ for all values of θ,

$$\min_{\lambda_\theta \neq \lambda \in \operatorname{spec} \mathsf{N}_\theta} |\lambda_\theta - \lambda| \geq \alpha > 0. \tag{2.9}$$

The property (2.9) is called *uniform separation property* (with α) in [100, p. 289]. The following theorem states the convergence of Ritz vectors under this condition.

Theorem 2.12 (Convergence of Ritz vectors, [100, p. 289])
Let (x, λ) *be an eigenpair of* A *and let* $(\mathsf{U}_\theta \mathsf{w}_\theta, \lambda_\theta)$ *be a Ritz pair such that* λ_θ *converges to* λ*. Let the uniform separation property be fulfilled with* $\alpha > 0$*. Then*

$$\sin \angle(\mathsf{x}, \mathsf{U}_\theta \mathsf{w}_\theta) \lesssim \sin\theta \sqrt{1 + \frac{\|\mathsf{A}\|^2}{\alpha^2}}$$

asymptotically.

Convergence of Subspaces

As Theorem 2.12 declares, convergence of single Ritz vectors can only be expected if the corresponding Ritz value is well separated (the theorem does not state that convergence of Ritz vectors with badly separated Ritz value is impossible). It might therefore sometimes be better to ask for a basis of an eigenspace whose Ritz values are well separated from *all other* Ritz values. Theorem 2.13 gives such a bound, even independently of the separation of eigenvalues. It depends on the normwise difference of the chosen bases of the subspaces. We adapted the notation to ours.

Theorem 2.13 ([58, Lem. 5.5], Knyazev, Argentati)
Let B *be a Hermitian positive definite matrix, let* $\mathcal{U} = \operatorname{span}(\mathsf{U})$, $\widetilde{\mathcal{U}} = \operatorname{span}(\widetilde{\mathsf{U}})$. *Then*

$$\sin \angle_\mathsf{B}(\mathcal{U}, \widetilde{\mathcal{U}}) \leq \kappa_\mathsf{B}(\mathsf{U}) \frac{\left\| \mathsf{U} - \widetilde{\mathsf{U}} \right\|_{\mathsf{B}2}}{\|\mathsf{U}\|_{\mathsf{B}2}}, \tag{2.10}$$

where $\kappa_\mathsf{B} = \sigma_{\max}(\mathsf{B}^{1/2}\mathsf{U}) / \sigma_{\min}(\mathsf{B}^{1/2}\mathsf{U})$ *denotes the condition number with respect to the* B*-norm.*

The theorem is true for any two subspaces, but if $\mathcal{U}$ is an eigenspace and U is chosen B-orthogonal, the denominator in (2.10) is 1 as well as $\kappa_\mathsf{B}(\mathsf{U})$. The same of course is true in the context of the standard eigenvalue equation with $\mathsf{B} = \mathsf{I}$. We consequently have in that case

$$\sin \angle_\mathsf{B}(\mathcal{U}, \widetilde{\mathcal{U}}) \leq \left\| \mathsf{U} - \widetilde{\mathsf{U}} \right\|_{\mathsf{B}2} \leq \varepsilon,$$

if ε is an upper bound for $\left\| \mathsf{U} - \widetilde{\mathsf{U}} \right\|_{\mathsf{B}2}$. In the case of the standard equation, we have

$$\sin \angle(\mathcal{U}, \widetilde{\mathcal{U}}) \leq \left\| \mathsf{U} - \widetilde{\mathsf{U}} \right\| \leq \varepsilon$$

if U was chosen orthonormal.

The next theorem, which is a generalization of Theorem 2 in [99], also gives a quantitative statement about the angle between subspaces. It depends on another angle and on a quantity "sep" that captures the separation of spectra, in a sense; it is discussed below.

In order to motivate the definition of the quantities that appear, we derive the theorem step by step rather than stating it and then proving it. The subsequent analysis closely follows [99]. The difference is that in [99] the result was proven for the standard eigenvalue problem, we now extend it to the generalized eigenvalue problem.

Let $\mathcal{K}$ be a subspace of $\mathbb{C}^n$ with B-orthonormal basis K. Let (U, M) be a Ritz pair belonging to $\mathcal{K}$, i. e., $\mathsf{U} = \mathsf{KG}$ and (G, M) is an eigenpair of $(\mathsf{K}^\star\mathsf{AK}, \mathsf{K}^\star\mathsf{BK})$. Define $\mathcal{U} = \operatorname{span}(\mathsf{U})$ and suppose $\mathcal{U} \neq \mathcal{K}$. Let $\mathcal{X}$ be an eigenspace of (A, B) with the same dimension as $\mathcal{U}$ and corresponding eigenpair (X, L). Let $\mathcal{V} = \operatorname{span}(\mathsf{V})$ be the B-orthogonal complement of $\mathcal{U}$ in $\mathcal{K}$ and let W be chosen such that $[\mathsf{U}, \mathsf{V}, \mathsf{W}]$ is B-unitary (in particular, $\mathsf{U}, \mathsf{V}, \mathsf{W}$ are B-orthonormal). Then we have

$$\begin{bmatrix} \mathsf{U}^\star \\ \mathsf{V}^\star \\ \mathsf{W}^\star \end{bmatrix} \mathsf{A} \, [\mathsf{U}, \mathsf{V}, \mathsf{W}] = \begin{bmatrix} \mathsf{M} & \mathsf{A}_{12} & \mathsf{A}_{13} \\ \mathsf{A}_{21} & \mathsf{N} & \mathsf{A}_{23} \\ \mathsf{A}_{31} & \mathsf{A}_{32} & \mathsf{A}_{33} \end{bmatrix},$$
$$\begin{bmatrix} \mathsf{U}^\star \\ \mathsf{V}^\star \\ \mathsf{W}^\star \end{bmatrix} \mathsf{B} \, [\mathsf{U}, \mathsf{V}, \mathsf{W}] = \mathsf{I}.$$

Since U spans a Ritz space, $\mathsf{A}_{21} = \mathsf{0}$ (let us neglect that A is Hermitian for the moment, actually we have $\mathsf{A}_{12} = \mathsf{0}$, as well). Next, consider the eigenpair (X, L) of (A, B). Let us express X in the basis $\mathsf{Y} := [\mathsf{U}, \mathsf{V}, \mathsf{W}]$. By definition, the inverse of Y is $\mathsf{Y}^\star\mathsf{B}$, consequently the change of bases yields

$$\begin{bmatrix} \mathsf{U}^\star \\ \mathsf{V}^\star \\ \mathsf{W}^\star \end{bmatrix} \mathsf{BX} =: \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \\ \mathsf{R} \end{bmatrix}.$$

By definition of angles we have $\|\mathsf{R}\| = \|\mathsf{W}^\star\mathsf{BX}\| = \sin\angle_{\mathsf{B}}(\mathcal{X}, \mathcal{K})$ (recall that $\operatorname{span}(\mathsf{W})$ is B-orthogonal to $\mathcal{K}$). The norm $\left\| \left[\begin{smallmatrix} \mathsf{Q} \\ \mathsf{R} \end{smallmatrix} \right] \right\|$ is $\sin\angle_{\mathsf{B}}(\mathcal{X}, \mathcal{U})$ since $[\mathsf{V}, \mathsf{W}]$ spans the space B-orthogonal to $\mathcal{U}$. The norm of the Q, R-block is hence the quantity that we want to bound. Recall $\mathsf{Y} = [\mathsf{U}, \mathsf{V}, \mathsf{W}]$, $\mathsf{Y}^{-1} = \mathsf{Y}^\star\mathsf{B}$ and use the equivalence

$$\mathsf{AX} = \mathsf{BXL} \iff \mathsf{Y}^\star\mathsf{AYY}^\star\mathsf{BX} = \mathsf{Y}^\star\mathsf{BXL}. \tag{2.11}$$

Forming the right hand side equation of (2.11) yields

$$\begin{bmatrix} \mathsf{M} & \mathsf{A}_{12} & \mathsf{A}_{13} \\ \mathsf{0} & \mathsf{N} & \mathsf{A}_{23} \\ \mathsf{A}_{31} & \mathsf{A}_{32} & \mathsf{A}_{33} \end{bmatrix} \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \\ \mathsf{R} \end{bmatrix} = \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \\ \mathsf{R} \end{bmatrix} \mathsf{L}.$$

Now, everything reduced to a standard problem and the rest of the analysis is essentially the proof in [99]. For completeness, we give it in full length.

Neglecting the last rows of the square matrix leads to

$$\begin{bmatrix} \mathsf{M} & \mathsf{A}_{12} \\ 0 & \mathsf{N} \end{bmatrix} \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \end{bmatrix} - \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \end{bmatrix} \mathsf{L} = - \begin{bmatrix} \mathsf{A}_{13} \\ \mathsf{A}_{23} \end{bmatrix} \mathsf{R}. \tag{2.12}$$

For the right hand side in (2.12) we have

$$\left\| - \begin{bmatrix} \mathsf{A}_{13} \\ \mathsf{A}_{23} \end{bmatrix} \mathsf{R} \right\| \leq \left\| \begin{bmatrix} -\mathsf{A}_{13} \\ -\mathsf{A}_{23} \end{bmatrix} \right\| \|\mathsf{R}\| \equiv: \eta \|\mathsf{R}\| , \tag{2.13}$$

where η is defined implicitly. For the left hand side in (2.12) we have

$$\left\| \begin{bmatrix} \mathsf{M} & \mathsf{A}_{12} \\ 0 & \mathsf{N} \end{bmatrix} \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \end{bmatrix} - \begin{bmatrix} \mathsf{P} \\ \mathsf{Q} \end{bmatrix} \mathsf{L} \right\| \geq \|\mathsf{NQ} - \mathsf{QL}\| . \tag{2.14}$$

Next, let $\hat{\mathsf{Q}} = (1/\|\mathsf{Q}\|) \cdot \mathsf{Q}$, then

$$\begin{aligned} \|\mathsf{NQ} - \mathsf{QL}\| &= \left\| \mathsf{N}\hat{\mathsf{Q}} - \hat{\mathsf{Q}}\mathsf{L} \right\| \|\mathsf{Q}\| \\ &\geq \inf_{\|\mathsf{Z}\|=1} \|\mathsf{NZ} - \mathsf{ZL}\| \, \|\mathsf{Q}\| \\ &=: \operatorname{sep}(\mathsf{N}, \mathsf{L}) \, \|\mathsf{Q}\| . \end{aligned}$$

The matrix Z in the definition of sep is of size $m \times l$, where m denotes the size of N, i. e., the dimension of $\mathcal{V}$ and ℓ denotes the size of L, i. e., the dimension of $\mathcal{U}$ and $\mathcal{X}$. Since we required $\mathcal{U} \neq \mathcal{K}$, we have $\ell, m > 0$. Using (2.13)–(2.14) together yields

$$\|\mathsf{Q}\| \leq \eta \frac{\|\mathsf{R}\|}{\operatorname{sep}(\mathsf{N}, \mathsf{L})}.$$

By using the fact that

$$\left\| \begin{bmatrix} \mathsf{Q} \\ \mathsf{R} \end{bmatrix} \right\|^2 \leq \|\mathsf{Q}\|^2 + \|\mathsf{R}\|^2$$

and taking square roots on both sides of the inequality we obtain

$$\left\| \begin{bmatrix} \mathsf{Q} \\ \mathsf{R} \end{bmatrix} \right\| \leq \|\mathsf{R}\| \sqrt{1 + \frac{\eta^2}{\operatorname{sep}(\mathsf{N}, \mathsf{L})^2}} .$$

Now, using angles instead of norms, it follows

$$\sin \angle_{\mathsf{B}}(\mathcal{U}, \mathcal{X}) \leq \sin \angle_{\mathsf{B}}(\mathcal{K}, \mathcal{X}) \sqrt{1 + \frac{\eta^2}{\operatorname{sep}(\mathsf{N}, \mathsf{L})^2}} .$$

Next, define

$$\begin{aligned}\operatorname{sep}(\mathcal{V},\mathcal{X}) &= \operatorname{sep}(\mathsf{V}^\star\mathsf{AV},\, \mathsf{X}^\star\mathsf{AX}) \\ &= \inf_{\|\mathsf{Z}\|=1} \|(\mathsf{V}^\star\mathsf{AV})\mathsf{Z} - \mathsf{Z}(\mathsf{X}^\star\mathsf{AX})\| \,.\end{aligned}$$

The definition makes sense since it does not depend on the choice of the bases of $\mathcal{V},\mathcal{X}$ as long as they are B-orthonormal. This is confirmed by the following Lemma.

Lemma 2.14
$\operatorname{sep}(\mathcal{V},\mathcal{X})$ *does not depend on the bases for* $\mathcal{V}$ *and* $\mathcal{X}$ *as long as they are* B*-orthonormal.*

Proof. Let $\mathsf{V}_1, \mathsf{V}_2$ be bases for $\mathcal{V}$ that are B-orthonormal. Then the matrix $\mathsf{Q} := \mathsf{V}_2^\star\mathsf{B}\mathsf{V}_1$ fulfills $\mathsf{V}_1 = \mathsf{V}_2\mathsf{Q}$. Next, it can be seen that Q is unitary since we have $\mathsf{Q}^\star\mathsf{Q} = \mathsf{Q}^\star\mathsf{V}_2^\star\mathsf{B}\mathsf{V}_2\mathsf{Q} = \mathsf{V}_1^\star\mathsf{B}\mathsf{V}_1 = \mathsf{I}$. The same analysis holds for the basis of $\mathcal{X}$. The change of B-orthonormal bases of $\mathcal{V},\mathcal{X}$ reduces to a unitary transformation of any of the two matrices $\mathsf{V}^\star\mathsf{AV},\ \mathsf{X}^\star\mathsf{AX}$. Such transformations leave sep unchanged, see [98]. □

The final result can now be stated as follows.

Theorem 2.15
Let $\mathcal{K}$ *be some subspace in* $\mathbb{C}^n$ *and let* $\mathcal{U} \subset \mathcal{K}$, $\mathcal{U} \neq \mathcal{K}$ *be a Ritz space (i. e., a subspace used in the Rayleigh–Ritz procedure). Let* $\mathcal{X}$ *be some eigenspace of the matrix pair* $(\mathsf{A},\, \mathsf{B})$ *with the same dimension as* $\mathcal{U}$*. Let* $\mathcal{V}$ *be the* B*-orthogonal complement of* $\mathcal{U}$ *in* $\mathcal{K}$ *(*$\mathcal{K} = \mathcal{V} \oplus \mathcal{U},\ \mathcal{V} \perp_{\mathsf{B}} \mathcal{U}$*). Then*

$$\sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{U}) \leq \sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{K})\sqrt{1 + \frac{\eta^2}{\operatorname{sep}(\mathcal{V},\mathcal{X})^2}} \tag{2.15}$$

with η *and* sep *defined above.*

Remark 2.16
Similarly to Theorem 2.11, the last theorem can be proven by taking the standard version [99], using the standardized equation $\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2}\mathsf{y} = \mathsf{y}\lambda$ and the relation between angles in the B-geometry and the standard ones. ◊

Remarks on sep and discussion of Theorem 2.15

The quantity sep (*separation*) of a matrix N and a matrix L is a measure for the distance between the spectra of the two matrices. It appears in many texts about perturbation of subspaces, for instance in [99]. To the best of our knowledge it was introduced by Stewart [98]. For more information, see also [110]. Interestingly, we only have to consider two single matrices instead of two matrix pairs; this is due to the fact that we can choose all bases B-orthonormal. In (2.15) the matrix

B appears only implicitly. Such a simple bound is not possible for general matrix pairs, in that case more complicated measures than sep have to be considered [98]. The quantity sep has the property [98]

$$\operatorname{sep}(\mathsf{N},\mathsf{L}) \leq \min|\operatorname{spec}(\mathsf{N}) - \operatorname{spec}(\mathsf{L})|\,. \tag{2.16}$$

In (2.16) we define $|\operatorname{spec}(\mathsf{N}) - \operatorname{spec}(\mathsf{L})| := \{|\nu - \mu| : \nu \in \operatorname{spec}(\mathsf{N}),\, \mu \in \operatorname{spec}(\mathsf{L})\}$. Unfortunately we only have the "$\leq$"-relation for sep, but the quantity appears in the denominator of (2.15). In the literature [103, p. 234] simple examples can be found where the two quantities in (2.16) differ significantly, at least in the relative sense (N, L can be constructed such that $\operatorname{sep}(\mathsf{N},\mathsf{L})/\min|\operatorname{spec}(\mathsf{N}) - \operatorname{spec}(\mathsf{L})|$ is arbitrarily small).

In the Hermitian case, when the pair (A, B) is definite as supposed, we can express the number sep in terms of the eigenvalues of N and L. For the Frobenius norm ($\|\mathsf{A}\|_F = (\sum_{ij} |\mathsf{A}(i,j)|^2)^{1/2}$) we have [98, Thm. 4.8]

$$\operatorname{sep}_F(\mathsf{N},\mathsf{L}) = \inf_{\|\mathsf{Z}\|_F=1} \|\mathsf{NZ} - \mathsf{ZL}\|_F = \min|\operatorname{spec}(\mathsf{N}) - \operatorname{spec}(\mathsf{L})|\,. \tag{2.17}$$

For the Frobenius norm of any matrix M we have

$$\|\mathsf{M}\|_F \geq \|\mathsf{M}\|_2\,.$$

However, it is not clear at first sight why this inequality should also hold for sep, since the infima in the definition are taken over different sets (for the Frobenius norm over the matrices Z with $\|\mathsf{Z}\|_F = 1$, for the 2-norm over the matrices Z with $\|\mathsf{Z}\|_2 = 1$). The following inequality, which we formulate as lemma, can be proven [98, p. 745].

Lemma 2.17
Let $\mathsf{N} \in \mathbb{C}^{m\times m}$, $\mathsf{L} \in \mathbb{C}^{\ell\times\ell}$. *Then the inequality*

$$\operatorname{sep}(\mathsf{N},\mathsf{L}) \geq \frac{\operatorname{sep}_F(\mathsf{N},\mathsf{L})}{\sqrt{\min\{m,\ell\}}}$$

holds.

Together with (2.17) we obtain from Lemma 2.17 the inequality

$$\operatorname{sep}(\mathsf{N},\mathsf{L}) \geq \frac{\min|\operatorname{spec}(\mathsf{N}) - \operatorname{spec}(\mathsf{L})|}{\sqrt{\min\{m,\ell\}}}, \tag{2.18}$$

for the case of N, L being Hermitian. Now, we can plug (2.18) into Formula (2.15) to obtain the bound

$$\sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{U}) \leq \sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{K})\sqrt{1 + \frac{\eta^2}{\min|\operatorname{spec}(\mathsf{N}) - \operatorname{spec}(\mathsf{L})|^2}\cdot\min\{m,\ell\}}\,. \tag{2.19}$$

Discussion. Theorem 2.15 gives a bound on the largest canonical angle between the exact eigenspace $\mathcal{X}$ of (A, B) and the Ritz space $\mathcal{U}$. The bound (in the form of (2.19)) depends on the quantities

- $\sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{K})$, the largest canonical angle between $\mathcal{X}$ and some space $\mathcal{K}$ enveloping $\mathcal{U}$. This angle fulfills $\sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{K}) \leq \sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{U})$ since $\mathcal{U} \subset \mathcal{K}$.
- A number η which is defined as the norm of certain blocks of the matrix A, transformed congruently via $[\mathsf{U},\mathsf{V},\mathsf{W}]$.
- The size of N, L.
- The separation of the spectra of N, L.

The computation of sep requires the computing the full spectra of N and L. This is possible if $m+\ell = \dim(\mathcal{K})$ is much smaller than n and if N,L are available.

The number η requires the computation of the norm of a matrix of size $(m+\ell)\times(n-(m+\ell))$, where $(n-(m+\ell))$ can be large. Also, $\sin\angle_{\mathsf{B}}(\mathcal{X},\mathcal{K})$ is not known because we aim at *computing* $\mathcal{X}$ and hence do not know this space. The use of the theorem is consequently of more theoretical nature. Stewart [99] mentions that the idea is to prove convergence of the Rayleigh–Ritz method if one has a sequence of subspaces $\mathcal{K}_j$ such that $\lim_j \sin\angle(\mathcal{X},\mathcal{K}_j) = 0$ (in case of the standard equation). He also writes that this does not suffice to prove the convergence of the method, since the Ritz pair U,M is not unique and $\min|\mathrm{spec}(\mathsf{N}) - \mathrm{spec}(\mathsf{L})|$ might become zero. Stewart's conclusion is that these problems do not have a very strong effect in practice, while pointing to [53].

2.1.5 Residual based bounds

In an actual computation, one needs reliable stopping criteria for the Rayleigh–Ritz process. The quantity that can usually be measured is the residual. Assume we have chosen a subspace U and extracted an eigenpair $(\widetilde{\mathsf{X}},\widetilde{\Lambda})$ with eigenvector matrix $\widetilde{\mathsf{X}} = [\widetilde{\mathsf{x}}_1,\ldots,\widetilde{\mathsf{x}}_m]$ and $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\alpha}_1,\ldots,\widetilde{\alpha}_m)$ from it. With $\widetilde{\mathsf{X}}$ and any matrix $\mathsf{H}\in\mathbb{C}^{m\times m}$ we have the *residual*

$$\mathsf{R} = \mathsf{R}(\mathsf{H}) = \mathsf{A}\widetilde{\mathsf{X}} - \mathsf{B}\widetilde{\mathsf{X}}\mathsf{H}. \tag{2.20}$$

In particular, we can compute $\mathsf{R}(\widetilde{\Lambda})$, the columns of which are

$$\mathsf{r}_j = \mathsf{A}\widetilde{\mathsf{x}}_j - \mathsf{B}\mathsf{x}_j\widetilde{\alpha}_j.$$

Residuals are easy to compute from the algorithmic point of view, the cost might be non-negligible. In the following, we state some error bounds for eigenvalues and eigenvectors based mainly on the residual norm. The arising inequalities are transferred to the definite generalized eigenproblem.

Most results are in terms of the residual $\mathsf{R}(\mathsf{A}_\mathsf{U})$ of the Rayleigh quotient $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star\mathsf{A}\mathsf{U}$ for some matrix U. The Rayleigh quotient minimizes the norm of the residual $\mathsf{R}(\mathsf{H})$, as is shown by the following lemma. It hence is bounded by all other residuals (see, e. g., [100, p. 254], [80, Thm. 11.4.2]).

Lemma 2.18
Let $[\mathsf{U}, \mathsf{U}_\perp]$ *be a unitary matrix with* $\mathsf{U} \in \mathbb{C}^{n\times m}$. *Then for the residual* $\mathsf{R} = \mathsf{A}\mathsf{U} - \mathsf{U}\mathsf{A}_\mathsf{U}$ *it holds*

$$\|\mathsf{R}\| = \min_{\mathsf{H}\in\mathbb{C}^{m\times m}} \|\mathsf{A}\mathsf{U} - \mathsf{U}\mathsf{H}\| .$$

In this case, we have $\|\mathsf{R}\| = \|\mathsf{U}_\perp^\star\mathsf{A}\mathsf{U}\|$.

For the generalized equation, let U be B-orthonormal; for the residual R from (2.20) we have

$$\begin{aligned}\hat{\mathsf{R}}(\mathsf{H}) &:= (\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2})\mathsf{B}^{1/2}\mathsf{U} - \mathsf{B}^{1/2}\mathsf{U}\mathsf{H}\\ &= \mathsf{B}^{-1/2}(\mathsf{A}\mathsf{U} - \mathsf{B}\mathsf{U}\mathsf{H})\\ &= \mathsf{B}^{-1/2}\mathsf{R}(\mathsf{H}).\end{aligned}$$

The norm of this matrix is minimized by the matrix

$$\mathsf{H} = (\mathsf{B}^{1/2}\mathsf{U})^\star(\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2})(\mathsf{B}^{1/2}\mathsf{U}) = \mathsf{U}^\star\mathsf{A}\mathsf{U} = \mathsf{A}_\mathsf{U},$$

the Rayleigh quotient of A with respect to the B-orthonormal basis U. Consequently, we have that

$$\left\|\hat{\mathsf{R}}(\mathsf{H})\right\| = \left\|\mathsf{B}^{-1/2}\mathsf{R}(\mathsf{H})\right\| = \|\mathsf{R}(\mathsf{H})\|_{\mathsf{B}^{-1},2}$$

is minimized by $\mathsf{H} = \mathsf{A}_\mathsf{U}$, where U is B-orthonormal.

Approximation error in eigenvalues

Let us investigate the approximation error in the Ritz values. To begin with, let A be Hermitian, y any unit vector and $\alpha \in \mathbb{C}$ any number. We then have the well-known result, which can, e. g., be found (with proof) in [80].

Theorem 2.19
With y, α *as above, there is an eigenvalue* λ *of* A *such that*

$$|\lambda - \alpha| \leq \|\mathsf{A}\mathsf{y} - \mathsf{y}\alpha\| .$$

Proof. For $\lambda = \alpha$ there is nothing to show, if $\lambda \neq \alpha$ for all $\lambda \in \operatorname{spec}(\mathsf{A})$, we may write $\mathsf{y} = (\mathsf{A} - \alpha\mathsf{I})^{-1}(\mathsf{A} - \alpha\mathsf{I})\mathsf{y}$. We then have

$$\begin{aligned}1 = \|\mathsf{y}\| &\leq \left\|(\mathsf{A} - \alpha\mathsf{I})^{-1}\right\| \|(\mathsf{A} - \alpha\mathsf{I})\mathsf{y}\|\\ &= (1/\min_j(|\lambda_j(\mathsf{A}) - \alpha|)) \|(\mathsf{A} - \alpha\mathsf{I})\mathsf{y}\| .\end{aligned}$$

The last equation follows since $\|(\mathsf{A}-\alpha\mathsf{I})^{-1}\| = \rho((\mathsf{A}-\alpha\mathsf{I})^{-1}) = 1/\min_j(|\lambda_j(\mathsf{A})-\alpha|)$. Next choose j^* such that $\min_j(|\lambda_j(\mathsf{A})-\alpha|) = |\lambda_{j^*}(\mathsf{A})-\alpha|$. Setting $\lambda = \lambda_{j^*}(\mathsf{A})$ completes the proof. □

Applying the theorem to the Rayleigh–Ritz approximation, we have that at least one eigenvalue of A resides in each interval $[\alpha_j - \|\mathsf{r}_j\|, \alpha_j + \|\mathsf{r}_j\|]$. If two (or more) of those intervals overlap, we do not know how to pair the Ritz values and the eigenvalues [80, Sec. 11.5]. The remedy is to bound the approximation errors in a whole cluster of eigenvalues [80].

Theorem 2.20 ([80, Thm. 11.5.1])
Let $\mathsf{U} \in \mathbb{C}^{n\times m}$ be an orthonormal matrix and let $\mathsf{R} = \mathsf{AU} - \mathsf{UH}$, where H is any Hermitian matrix. Then there are m pairs (i,j), where $1 \le i \le m$, $1 \le j \le n$ such that

$$|\alpha_i - \lambda_j| \le \|\mathsf{R}\| .$$

Next, let us adapt Theorem 2.20 to the case of a generalized problem. We can invoke the theorem with the matrices $\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2}$ and $\mathsf{B}^{1/2}\mathsf{U}$ to obtain a bound for the Ritz values of (A, B) in terms of the residual $\mathsf{B}^{-1/2}\mathsf{R}$ with R from (2.20). This yields the following result.

Theorem 2.21
Let $\mathsf{U} \in \mathbb{C}^{n\times m}$ be a B-orthonormal matrix and let $\mathsf{R} = \mathsf{AU} - \mathsf{BUH}$, where H is any Hermitian matrix. Then there are m pairs (i,j), $1 \le i \le m$, $1 \le j \le n$, such that

$$|\alpha_i - \lambda_j| \le \left\|\mathsf{B}^{-1/2}\mathsf{R}\right\| = \|\mathsf{R}\|_{\mathsf{B}^{-1},2} ,$$

where α_i are the Ritz values belonging to the space $\operatorname{span}(\mathsf{U})$.

Note that $\left\|\mathsf{B}^{-1/2}\mathsf{R}\right\|$ can be further bounded by $\lambda_{\min}(\mathsf{B})^{-1/2}\,\|\mathsf{R}\|$. Theorem 2.19 can be generalized in the same way.

Angles between subspaces ($\sin\theta$ theorem)

In this section, we shall derive bounds for the angle between the computed subspace and a certain exact eigenspace in terms of the residual (2.20) and certain *gaps* in the spectrum. As we have seen before, a reasonable way to measure angles is in the B-geometry.

The essential work on this topic was published by Davis and Kahan in 1970 [18]. The main results are known as $\sin\theta$ theorem and $\tan\theta$ theorem, where θ denotes the angle between the subspaces under consideration. Davis and Kahan gave a short overview in [17]. In [18] also bounds on $\sin 2\theta$ and $\tan 2\theta$ can be found. Note that the results in this reference are much more general than stated here. Recently, Nakatsukasa [75] showed that certain prerequisites can be relaxed.

For introduction, let us start by stating the $\sin\theta$, $\tan\theta$ theorems in their primary form, for Hermitian matrices. To this end, let A be Hermitian and let

$\mathsf{X} = [\mathsf{X}_1, \mathsf{X}_2]$ with $\mathsf{X}_1 \in \mathbb{C}^{n\times m}$ be a square matrix of eigenvectors, such that $\mathsf{X}^\star\mathsf{A}\mathsf{X} = \Lambda$ is the diagonal matrix of eigenvalues. Split $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ according to X. Suppose any orthonormal matrix U is given and let $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star\mathsf{A}\mathsf{U}$ be its Rayleigh quotient. The crucial quantity is the gap between the spectra of A_U and Λ_2. It can easily be explained in terms of intervals. Suppose there is a compact interval $[a, b] \supset \mathrm{spec}(\mathsf{A}_\mathsf{U})$ and a number $\delta > 0$ such that $\mathrm{spec}(\Lambda_2) \subset (-\infty, a - \delta]$ entirely *or* $\mathrm{spec}(\Lambda_2) \subset [b + \delta, \infty)$ entirely. Then, the following is true

1.
$$\|\mathbf{tan}\Theta\| \leq \frac{\|\mathsf{A}\mathsf{U} - \mathsf{U}\mathsf{A}_\mathsf{U}\|}{\delta}, \tag{2.21}$$

2.
$$\|\mathbf{sin}\Theta\| \leq \frac{\|\mathsf{A}\mathsf{U} - \mathsf{U}\mathsf{A}_\mathsf{U}\|}{\delta}. \tag{2.22}$$

Here, Θ denotes the diagonal matrix consisting of the m canonical angles ($< \pi/2$) between U and X_1 and **sin** and **tan** denote their element-wise sines and tangents, respectively. In [75], $\mathbf{tan}\Theta$ and $\mathbf{sin}\Theta$ are replaced by matrices whose singular values are the tangents and sines of the respective angles. The matrices $\mathbf{tan}\Theta$, $\mathbf{sin}\Theta$ have exactly these singular values. Since sine and tangent are monotone on $[0, \pi/2)$, we can safely replace the left hand sides of (2.21) and (2.22) by $\tan\angle(\mathsf{U}, \mathsf{X}_1)$ and $\sin\angle(\mathsf{U}, \mathsf{X}_1)$, respectively. This is due to the fact that $\angle(\mathsf{U}, \mathsf{X}_1)$ is the largest among the canonical angles from Θ. Note that the inequality involving the tangent is sharper than that one involving the sine, since $\tan\theta \geq \sin\theta$ for $0 \leq \theta < \pi/2$, see [75]. In the $\sin\theta$ theorem, the prerequisites can be relaxed. Requiring

- $\mathrm{spec}(\Lambda_2) \subset [a, b]$ and $\mathrm{spec}(\mathsf{A}_\mathsf{U}) \subset (-\infty, a - \delta] \cup [b + \delta, \infty)$ or
- $\mathrm{spec}(\mathsf{A}_\mathsf{U}) \subset [a, b]$ and $\mathrm{spec}(\Lambda_2) \subset (-\infty, a - \delta] \cup [b + \delta, \infty)$

is enough [75]. Nakatsukasa [75] showed that the first case is allowed as requirement in the $\tan\theta$ theorem. Verbally, $\mathrm{spec}(\mathsf{A}_\mathsf{U})$ is allowed to lie on both sides of $[a, b]$, where the distance at either side has to be at least δ, and not just entirely below or above this interval. We obtain the following generalization of the $\tan\theta$ and $\sin\theta$ theorems.

Theorem 2.22 ($\tan\theta$ and $\sin\theta$ for generalized eigenproblems)
Let (A, B) *be a definite matrix pair and let* $\mathsf{X} = [\mathsf{X}_1, \mathsf{X}_2]$ *with* $\mathsf{X}_1 \in \mathbb{C}^{n\times m}$ *be a full matrix of eigenvectors with* $\mathsf{X}^\star\mathsf{B}\mathsf{X} = \mathsf{I}$*, such that* $\mathsf{X}^\star\mathsf{A}\mathsf{X} = \Lambda$ *is the diagonal matrix of eigenvalues of* (A, B)*. Split* $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ *according to* X*. Suppose any* B*-orthonormal matrix* U *is given and let* $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star\mathsf{A}\mathsf{U}$ *be its Rayleigh quotient. Suppose further that there is* $\delta > 0$ *such that*

$$\mathrm{spec}(\Lambda_2) \subset [a, b] \text{ and } \mathrm{spec}(\mathsf{A}_\mathsf{U}) \subset (-\infty, a - \delta] \cup [b + \delta, \infty).$$

We then have

- $\sin \angle_{\mathsf{B}}(\mathsf{U}, \mathsf{X}_1) \leq \dfrac{\|\mathsf{AU} - \mathsf{BUH}\|_{\mathsf{B}^{-1},2}}{\delta}$ *and*
- $\tan \angle_{\mathsf{B}}(\mathsf{U}, \mathsf{X}_1) \leq \dfrac{\|\mathsf{AU} - \mathsf{BUH}\|_{\mathsf{B}^{-1},2}}{\delta}$

for any Hermitian H*, in particular for* A_U.

Proof. The statement on the sine follows from applying the classical $\sin\theta$ theorem [18, 75] to the standardized equation

$$(\mathsf{B}^{-1/2}\mathsf{A}\mathsf{B}^{-1/2})\mathsf{B}^{1/2}\mathsf{X} = \mathsf{B}^{1/2}\mathsf{X}\Lambda$$

to obtain

$$\sin \angle(\mathsf{B}^{1/2}\mathsf{U}, \mathsf{B}^{1/2}\mathsf{X}_1) \leq \frac{\left\|\mathsf{B}^{-1/2}(\mathsf{AU} - \mathsf{BUA}_\mathsf{U})\right\|}{\delta}.$$

Rewriting both sides yields

$$\sin \angle_{\mathsf{B}}(\mathsf{U}, \mathsf{X}_1) \leq \frac{\|\mathsf{AU} - \mathsf{BUA}_\mathsf{U}\|_{\mathsf{B}^{-1},2}}{\delta}.$$

The Rayleigh quotient A_U on the right hand side of the inequality can be replaced by any Hermitian matrix H since it minimizes the residual. Everything applies likewise to the statement on the tangent by using Nakatsukasa's relaxed conditions for the $\tan\theta$ theorem [75]. □

A lower bound on δ can be described in terms of indices as follows. Let a set of m Ritz values α_j of (A, B), which fulfill the inequality from Theorem 2.21, be given; let I denote the set of indices of all other eigenvalues. Let

$$\text{gap} := \min\{|\alpha_j - \lambda_i| : 1 \leq j \leq m;\, i \in I\}.$$

In words, gap is the minimum distance between all Ritz values α_i to the eigenvalues of (A, B) that cannot (necessarily) be paired according to Theorem 2.21. We then have gap $\leq \delta$. A similar number as gap appeared in [119]

Note that gap is potentially smaller than sep from Section 2.1.4, since it measures the minimum distance to *all* other eigenvalues that do not belong to certain Ritz values. The number sep captures distances of Ritz values to just a subset of the eigenvalues.

2.1.6 Harmonic Rayleigh–Ritz

It is noted in the literature, e. g., [72, 100] that the convergence of Ritz pairs with Ritz value in the interior of the spectrum is expected to be not as fast as convergence of exterior Ritz pairs. When considering the transformation $\mathsf{A} \mapsto (\mathsf{A} - \sigma\mathsf{I})^{-1}$, we see that the (interior) eigenvalues of A near σ are mapped to

eigenvalues with large modulus, i. e., at the extremes of $\operatorname{spec}((\mathsf{A}-\sigma\mathsf{I})^{-1})$. The eigenvectors stay the same. In formulas, we have

$$\lambda \in \operatorname{spec}(\mathsf{A}) \Longrightarrow (\lambda-\sigma)^{-1} \in \operatorname{spec}((\mathsf{A}-\sigma\mathsf{I})^{-1}).$$

Using the Rayleigh–Ritz method with subspace U and the matrix $(\mathsf{A}-\sigma\mathsf{I})^{-1}$ is not a cheap operation since it requires the solution of linear systems with system matrix $\mathsf{A}-\sigma\mathsf{I}$. We hence can revert to the space spanned by $(\mathsf{A}-\sigma\mathsf{I})\mathsf{U}$. This choice of matrix and subspace leads to the harmonic Rayleigh–Ritz procedure as described in [100]. It was first introduced by Morgan [72] without using the term "harmonic". We obtain the equation

$$\mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{I})^\star(\mathsf{A}-\sigma\mathsf{I})^{-1}(\mathsf{A}-\sigma\mathsf{I})\mathsf{U}\mathsf{w} = \mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{I})^\star(\mathsf{A}-\sigma\mathsf{I})\mathsf{U}\mathsf{w}\frac{1}{\rho},$$

where $\rho = \mu - \sigma$ for a Ritz value μ of A. Of course, it is supposed that $\rho \neq 0$. Rewriting yields

$$\mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{I})^\star(\mathsf{A}-\sigma\mathsf{I})\mathsf{U}\mathsf{w} = \mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{I})\mathsf{U}\mathsf{w}\rho. \tag{2.23}$$

The harmonic Rayleigh–Ritz procedure arises when using a blockwise version of (2.23) instead of the equation $\mathsf{A}_\mathsf{U}\mathsf{W} = \mathsf{B}_\mathsf{U}\mathsf{W}\Lambda$ inside the Rayleigh–Ritz procedure. It can easily be adopted to the generalized equation, resulting in

$$\mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})^\star(\mathsf{A}-\sigma\mathsf{B})\mathsf{U}\mathsf{w} = \mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})\mathsf{B}\mathsf{U}\mathsf{w}\rho, \tag{2.24}$$

see [100, p. 299], [45]. Note that the matrices $(\mathsf{A}-\sigma\mathsf{I}), (\mathsf{A}-\sigma\mathsf{B})$ are Hermitian if $(\mathsf{A},\,\mathsf{B})$ is a definite pair and if σ is real, in this case (2.24) becomes

$$\mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})^2\mathsf{U}\mathsf{w} = \mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})\mathsf{B}\mathsf{U}\mathsf{w}\rho. \tag{2.25}$$

In Hochstenbach [45], the *harmonic Ritz pair* $(\mathsf{U}\mathsf{w}, \mu)$, $\mu = \rho + \sigma$, is defined, where (w, ρ) is any solution of (2.24). The scalar μ is then taken as approximation to some eigenvalue of $(\mathsf{A},\,\mathsf{B})$ near σ. In this reference, Hochstenbach also describes how residual bounds can be obtained. For that purpose, let $(\mathsf{U}\mathsf{w}, \rho + \sigma)$ be a harmonic Ritz pair and left-multiply (2.24) by $\mathsf{w}^\star$. We obtain

$$\|(\mathsf{A}-\sigma\mathsf{B})\mathsf{u}\|^2 \leq |\rho| \cdot \|(\mathsf{A}-\sigma\mathsf{B})\mathsf{u}\| \cdot \|\mathsf{B}\mathsf{u}\|\,, \tag{2.26}$$

where $\mathsf{u} = \mathsf{U}\mathsf{w}$. This yields (divide (2.26) by $\|(\mathsf{A}-\sigma\mathsf{B})\mathsf{u}\|$)

$$\|(\mathsf{A}-\sigma\mathsf{B})\mathsf{u}\| \leq |\rho| \cdot \|\mathsf{B}\mathsf{u}\| = |\rho| \cdot \|\mathsf{u}\|_{\mathsf{B}^2}\,. \tag{2.27}$$

Here, u can be chosen such that the right hand side of (2.27) is $|\rho|$, i. e., normalized with respect to the B^2-norm. Bound (2.27) is true for any harmonic Ritz pair $(\mathsf{U}\mathsf{w}, \mu)$, where $|\mu - \sigma| \leq |\rho|$ [100]. Such residual bounds are not available for standard Rayleigh–Ritz [45].

The convergence of the harmonic Rayleigh–Ritz method was first analyzed in [14] and [52]; in [45] an analysis for the generalized version can be found. Those publications deal with general, non-Hermitian eigenproblems. In Morgan's article [72] error bounds for Hermitian matrices can be found. The name "harmonic" was, to the best of our knowledge, first used in [79], while that publication is dealing with Krylov subspaces and not general subspaces as above. The error bounds for harmonic Ritz vectors in [52] are similar to those stated earlier in this work, but they include the norm of $(\mathsf{U}(\mathsf{A}-\sigma\mathsf{I})^\star\mathsf{U})^{-1}$.

Harmonic Rayleigh–Ritz is not at the core of this work, however, all described subspace methods can in principle be implemented with harmonic Rayleigh–Ritz instead of standard Rayleigh–Ritz.

Practical aspects

Let us conclude this section with some practical considerations and a numerical example concerning harmonic Rayleigh–Ritz. To anticipate, the quintessence will be that the use of harmonic Rayleigh–Ritz in the context of the FEAST algorithm shows no improvements compared to standard Rayleigh–Ritz. It even leads to difficulties that do not occur in the standard case. The FEAST algorithm is introduced in Chapter 3, it aims at computing eigenpairs with eigenvalue in a given interval I_λ.

In harmonic Rayleigh–Ritz, (2.25) instead of the standard Rayleigh–Ritz equation

$$\mathsf{U}^\star\mathsf{A}\mathsf{U}\mathsf{w} = \mathsf{U}^\star\mathsf{B}\mathsf{U}\mathsf{w}\lambda$$

can be used if the matrices A, B are Hermitian. The harmonic Ritz pair $(\mathsf{U}\mathsf{w}, \rho+\sigma)$ is taken to approximate an eigenpair of (A, B) with eigenvalue near σ. The first difficulty that is seen when considering (2.25) is that the matrix $\mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})\mathsf{B}\mathsf{U}$ is not positive definite in general. This problem can still be circumvented by employing the equation

$$\mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})\mathsf{B}\mathsf{U}\mathsf{w} = \mathsf{U}^\star(\mathsf{A}-\sigma\mathsf{B})^2\mathsf{U}\mathsf{w}\frac{1}{\rho}$$

instead, where the right hand side involves a positive definite matrix. Harmonic Rayleigh–Ritz in this form can also be found in Morgan [72]. Unfortunately, two harmonic Ritz vectors $\mathsf{U}\mathsf{v}$, $\mathsf{U}\mathsf{w}$ belonging to different harmonic Ritz values will not be B-orthogonal. Furthermore, it is not clear how the shift σ should be chosen, since in the context of the FEAST algorithm we seek eigenvalues in an interval and not in the vicinity of a given target value. It seems obvious that it should be chosen somewhere inside the interval I_λ. In the following experiment, we applied the FEAST algorithm to a standard eigenvalue problem while using harmonic Rayleigh–Ritz. The reader should not care about the details of FEAST, which are discussed in Chapter 3.

α in $\sigma(\alpha) = \underline{\lambda} + \alpha \cdot d$	0.0	0.25	0.5	0.75	1.0
# Iterations(#E.values)	—	4(199)	4(198)	4(199)	10(31)

Table 2.2: Iteration counts for FEAST with harmonic Rayleigh–Ritz. The symbol "—" means that the algorithm failed to converge towards meaningful results.

Experiment 2.23
We chose a matrix A from electronic structure calculations of size 1629. We sought for the 200 eigenpairs with eigenvalues $\lambda_{n-400}, \ldots, \lambda_{n-200-1}$ and chose the interval boundaries slightly below and above these values, respectively. The search space size $\widetilde{m}$ was chosen as $\widetilde{m} = 300$. We used Gauß–Legendre integration with 8 integration points.

The FEAST algorithm with standard Rayleigh–Ritz took 4 iterations for all 200 eigenpairs to converge. Then, we set $d = \overline{\lambda} - \underline{\lambda}$ and $\sigma(\alpha) = \underline{\lambda} + \alpha d$, where $\alpha = 0.0, 0.25, 0.5, 0.75, 1.0$. The resulting iteration counts are shown in Table 2.2, together with the actual number of converged eigenpairs in at most 10 iterations of FEAST. The numerical quality (in terms of residuals) of the results was comparable to that one obtained with standard Rayleigh–Ritz. The computed eigenvectors were only orthogonal and not orthonormal. When changing $\widetilde{m}$, the results did not differ significantly, as long as $\widetilde{m}$ was not chosen too small (cf. Section 3.2).
$\Diamond$

The experiment shows exemplarily that it does not make sense to use harmonic Rayleigh–Ritz without further modification. From a conceptual point of view it is clear that it is also not necessary to use a different way of eigenpair extraction for inner eigenvalues, since the integration based eigensolver is designed for just this kind of problem, see the detailed discussion in Section 2.4. If harmonic Rayleigh–Ritz is used for some reason, the shift should be chosen in the center of the interval, at least somewhere inside the interval and not on one of the boundaries.

2.2 A few facts from complex analysis

In this section, we give formal definitions for complex contour integration and state some of the results from complex analysis that we will need later. Most of the material is adapted from [2] and [42] but it can be found in most textbooks on basic complex analysis.

Definition 2.24. Let $\Omega \subset \mathbb{C}$ be an open subset and $f : \Omega \longrightarrow \mathbb{C}$ some function. Then f is said to be *holomorphic* (or *analytic*) if for every point $z \in \Omega$ the derivative

$$f'(z) := \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

exists in $\mathbb{C}$. We will denote the set of analytic functions $f : \Omega \longrightarrow \mathbb{C}$ by $H(\Omega, \mathbb{C})$.

The set of analytic functions mapping into some set Y is denoted analogously by $H(\Omega, Y)$. ◊

We do not simply say that the function f is differentiable since the holomorphicity of a complex function is a much stronger condition than the differentiability of a real-valued function of a real variable. Holomorphicity implies that the function has a local power series expansion and a local primitive.

Definition 2.25. A *contour* (also known as *curve* or *path*) is a continuous function φ defined on a real interval,

$$\varphi : [\alpha, \beta] \longrightarrow \mathbb{C}. \tag{2.28}$$

The contour is called *continuously differentiable* (for short: differentiable) if its real and imaginary part are differentiable in (α, β), respectively, and if the derivative $\varphi'(t) = (\mathrm{Re}(\varphi(t)))' + \mathbf{i}(\mathrm{Im}(\varphi(t)))'$ is continuous on $[\alpha, \beta]$. It is called *piecewise differentiable* if there is a finite number of subdivision points of $[\alpha, \beta]$, $\alpha = \tau_0 < \tau_1 < \cdots < \tau_k = \beta$, such that it is differentiable in each subinterval $(\tau_j, \tau_{j+1}),\ j = 0, \ldots, k-1$. ◊

By the symbol $\mathcal{C}$ we will denote the contour (2.28) as well as the set it maps onto, $\mathcal{C} = \{\varphi(t) : t \in [\alpha, \beta]\}$. In the rest of this text by "curve"/"contour" we mean a curve that is at least piecewise differentiable.

Definition 2.26. Let $\mathcal{C}$ be a curve in some set $\Omega \subset \mathbb{C}$, parametrized by φ. If the function $f : \Omega \longrightarrow \mathbb{C}$ is continuous on $\mathcal{C}$, we define

$$\int_{\mathcal{C}} f(z)\mathrm{d}z = \int_{\alpha}^{\beta} f(\varphi(t))\varphi'(t)\mathrm{d}t, \tag{2.29}$$

where the integration interval on the right hand side of (2.29) has to be subdivided into subintervals such that φ is differentiable in each of those subintervals. ◊

Definition 2.27. A contour $\mathcal{C}$, parametrized by φ, is called *closed* if $\varphi(\alpha) = \varphi(\beta)$. The number

$$W_\varphi(z) := \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} \frac{1}{\zeta - z}\mathrm{d}\zeta,$$

defined for a closed curve $\mathcal{C}$ and $z \notin \mathcal{C}$, is called *winding number* of z with respect to φ. The set $\mathrm{Int}(\mathcal{C}) := \{z \notin \mathcal{C} : W_\varphi(z) \neq 0\}$ is called the *interior* of $\mathcal{C}$. The set $\mathrm{Ext}(\mathcal{C}) := \mathbb{C} \setminus (\mathrm{Int}(\mathcal{C}) \cup \mathcal{C})$ is called the *exterior* of $\mathcal{C}$.

A contour is called *simply closed* if its interior is nonempty and if $W_\varphi \equiv 1$ in the interior. ◊

Note that $W_\varphi(z) \in \mathbb{Z}$ for all $z \in \mathbb{C} \setminus \mathcal{C}$. In the following, we only consider simply closed curves. A very simple example of such a curve is the circle $\varphi(t) = r\exp(\mathbf{i}t)$, $0 \leq t \leq 2\pi$, $r > 0$. The following well-known theorem is in particular true for $f \equiv 1$ as a trivial case.

Theorem 2.28 (Cauchy's integral formula)
Let $\Omega \subset \mathbb{C}$, $\mathcal{C}$ be a closed curve in Ω with $\mathrm{Int}(\mathcal{C}) \subset \Omega$, parametrized by φ and let $f \in H(\Omega, \mathbb{C})$. Then it holds

$$\frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} \frac{f(\zeta)}{\zeta - z} \mathrm{d}\zeta = W_\varphi(z) f(z)$$

for all $z \notin \mathcal{C}$.

This is the theorem that will later help to explain why the integral based eigenvalue solvers work. Note that in particular the integral does *not* depend on the specific choice of $\mathcal{C}$, the curve only has to be simply closed and fulfill $z \notin \mathcal{C}$.

We will need some special series expansion of certain functions, the well-known Laurent expansion. It can be found, for instance, in Ahlfors [2, Ch. 7].

Theorem 2.29 (Laurent expansion)
Let

$$A = A(a_-, a_+) := \{\zeta \in \mathbb{C} : a_- < |\zeta| < a_+\},\ 0 \leq a_- < a_+$$

be an annulus and $f \in H(A, \mathbb{C})$. Then, f can be expanded into a Laurent series

$$f(z) = \sum_{k=-\infty}^{\infty} a_k z^k,$$

where the coefficients are uniquely defined. They are given by

$$a_k = \frac{1}{2\pi\mathbf{i}} \int_{|\zeta|=r} \frac{f(\zeta)}{\zeta^{k+1}} \mathrm{d}\zeta$$

for $a_- < r < a_+$.

Of course, ζ may be shifted to any $\zeta - c$ for $c \in \mathbb{C}$. We will use some other notions from complex analysis from time to time; the reader may find them in one of the textbooks cited above.

2.3 Numerical integration

In a numerical algorithm, integrals also have to be evaluated numerically, as long as the primitive of the integrand is not known. Here, we concisely survey numerical integration (also known as numerical quadrature). As an introductory text, [20] can be used. Of course there are many other textbooks such as [62, 63]. We follow these three books closely and also borrow most of our notation from there.

2.3.1 Basics

Numerical integration is needed whenever a definite integral of a function f has to be computed and the primitive of f is not available or the values of the function are only known at discrete points [62]. Also, it might be cheaper to apply a quadrature rule than to evaluate the primitive. For instance, the primitive of $f(t) = 1/(1+t^2)$ is $\arctan(t)$ which is not trivial to evaluate [73].

In the following, we discuss the numerical integration of some real valued continuous function $f : [\,\alpha,\, \beta\,] \longrightarrow \mathbb{R}$ and set

$$Q(f) := \int_{\alpha}^{\beta} f(t)\mathrm{d}t.$$

For complex valued functions the real and imaginary part are integrated separately. All integration schemes discussed take the form

$$(\omega_j, t_j)_{j=0,\ldots,p}, \tag{2.30}$$

where ω_j are the *integration weights* and $t_j \in [\,\alpha,\, \beta\,]$ are the *integration points*. The number p is called the *order* of the integration scheme (note that there are $p+1$ integration points). Applying the integration scheme (2.30) is done by forming

$$Q(f) \approx Q_p(f) := \sum_{j=0}^{p} \omega_j f(t_j).$$

Here, we introduce two basic kinds of quadrature, in a very condensed manner. *Interpolatory quadrature*, also known as Newton–Cotes quadrature is based on the evaluation of the integrand on equidistant points. These methods can be derived by integrating the corresponding interpolating polynomial.

The other type of quadrature rules we discuss are so called *Gaussian* integration rules, which employ values of the integrand at non-equidistant points. This often leads to better accuracy with equal computational effort. However, Newton–Cotes formulas are very useful when integrating periodic functions. For an overview, see [114]. We will see that Gauß type formulas are also of interpolatory type, but in the following by "interpolatory type quadrature rule" we will refer to a formula with an equidistant subdivision of the integration interval.

2.3.2 Interpolatory quadrature

Let us review interpolatory quadrature. Suppose, the integration points are fixed, e. g., an equidistant spacing of the interval $[\,\alpha,\, \beta\,]$ is given. We then can only choose the $p+1$ weights ω_j, so the best we can expect is to get a formula that is exact for $p+1$ linearly independent functions from $C\,[\,\alpha,\, \beta\,]$. Now, we wish for

$Q_0 f = (\beta - \alpha) f(\alpha)$	Rectangular rule
$Q_1 f = \frac{\beta-\alpha}{2}(f(\alpha) + f(\beta))$	Trapezoidal rule
$Q_2 f = \frac{\beta-\alpha}{6}(f(\alpha) + 4f(\frac{\alpha+\beta}{2}) + f(\beta))$	Simpson's rule

Table 2.3: 3 closed Newton–Cotes formulas

an exact integration rule for all polynomials with degree of at most p (the set of those will be denoted by $\mathbb{P}_p$). The ω_j's are the solutions of the system of linear equations

$$\int_\alpha^\beta t^k \mathrm{d}t \;=\; \sum_{j=0}^{p} \omega_j t_j^k,\; k = 0, \ldots, p. \tag{2.31}$$

The system matrix $(t_j^k)_{j,k=0,\ldots,p}$ of (2.31) is the so called Vandermonde matrix, which is nonsingular as long as the integration points are distinct [62, p. 186]. It can be shown that this choice of the weights ω_j is equivalent to the following procedure [20, p. 74]. First, f is interpolated at the points $t_0, \ldots, t_p$ by a polynomial $P \in \mathbb{P}_p$. For the basics on polynomial interpolation see, e. g., [62, Ch. 8] (or some other basic text on numerical analysis). The resulting polynomial P is integrated and the integral is expressed as

$$\int_\alpha^\beta P(t)\mathrm{d}t = \sum_{j=0}^{p} \omega_j P(t_j) = \sum_{j=0}^{p} \omega_j f(t_j).$$

This choice is well defined since the interpolating polynomial for distinct points is unique.

This motivates the name "interpolatory quadrature". If we now let $t_j = \alpha + hj,\; h = (\beta - \alpha)/p$, we obtain so called *closed Newton–Cotes* formulas. For $p = 0$ we get the *rectangular* rule with $\omega_0 = 1,\; t_0 = \alpha$. This rule computes the area of the rectangle with the corners $(\alpha, 0)$, $(\beta, 0)$, $(\alpha, f(\alpha))$, $(\beta, f(\alpha))$. For $p = 1$ we get the trapezoidal rule with $\omega_0 = \omega_1 = 1/2 \cdot (\beta - \alpha)$. This rule simply captures the area of the trapezoid that is defined by the linear function that interpolates f in α and β. The first three formulas are summarized in Table 2.3. If we let

$$t_j = \alpha + \frac{j+1}{p+2} \cdot (\beta - \alpha),$$

we obtain the so called *open Newton–Cotes* formulas, where the integration points are centered between two points of the corresponding closed formula. The probably most prominent member of this class is also the most simple, for $p = 0$ we obtain the *midpoint* rule

$$Q_0(f) := (\beta - \alpha) f\left(\frac{\alpha + \beta}{2}\right).$$

Compound formulas

Only limited accuracy can be achieved by Newton–Cotes formulas of modest order and those formulas become more and more unstable with growing p [63, p. 131]. Hence it is a good idea to apply the formulas Q_p to subintervals. Most of the following is taken from [63].

Let the interval $[\,\alpha,\,\beta\,]$ be equidistantly divided,

$$\alpha = \bar{t}_0 < \bar{t}_1 < \cdots < \bar{t}_k = \beta, \quad \bar{t}_j = \alpha + j \cdot \frac{\beta - \alpha}{k}, \quad j = 0, \ldots, k.$$

Then, the p-point Newton–Cotes formula is applied to each of the intervals defined by the points $\bar{t}_j$. To this end let $t_j,\ j = 0, \ldots, p$ denote the integration points of the formula Q_p applied to the interval $[-1,\,1]$ and set

$$t_{ij} := \bar{t}_{j-1} + \frac{\beta - \alpha}{2k}(1 + t_i), \quad i = 1, \ldots, p, \quad j = 1, \ldots, k.$$

The resulting *compound* formula $k \times Q_p$ then can be written as

$$(k \times Q_p)f = \frac{\beta - \alpha}{2k} \sum_{j=1}^{k} \sum_{i=0}^{p} \omega_i f(t_{ij}).$$

One of the most important compound formulas is the compound trapezoidal rule

$$T_k f := \frac{\beta - \alpha}{k} \left[\frac{1}{2} f(\alpha) + f(\alpha + h) + \cdots + f(\alpha + (k-1)h) + \frac{1}{2} f(\beta) \right] \tag{2.32}$$

with $h = (\beta - \alpha)/k$. This formula is easily geometrically understood, it is nothing else but computing the areas of the trapezoidals that are defined by f and the interval subdivision, see Figure 2.1. In formula (2.32) the function f has to be evaluated $k + 1$ times. In order to keep notation simple we will use p instead of k in the following and say that T_p is a *trapezoidal rule of order p* (note that it has $p + 1$ points).

2.3.3 Gauß quadrature

In this subsection, we discuss the so called *Gaussian quadrature* rules. These are actually rules for integrating the product $w(t)f(t)$ for some *weight function* w. We assume that w is a positive continuous function on $(\alpha,\,\beta)$ and that its integral over $[\,\alpha,\,\beta\,]$ exists. In all of our applications we will have $w \equiv 1$. The resulting Gauß rule is called *Gauß–Legendre* rule.

If we let the integration rule define the integration weights *and* the integration points, we have $2p + 2$ degrees of freedom and hence can hope for a rule that is exact for all polynomials from the $(2p + 2)$-dimensional space $\mathbb{P}_{2p+1}$. This is exactly what a Gaussian rule achieves and it is also our definition.

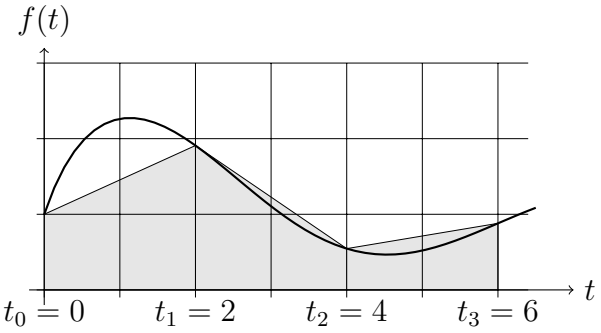

Figure 2.1: Example for trapezoidal rule $3 \times Q_2$ for $f(t) = 3\frac{\sin(t)}{1+t} + 1$ on the interval $[0, 6]$.

Definition 2.30 ([**62, Def. 9.12**])**.** A quadrature formula

$$\int_{\alpha}^{\beta} w(t) f(t) \mathrm{d}t \approx \sum_{j=0}^{p} \omega_j f(t_j)$$

with $p+1$ distinct quadrature points is called a *Gaussian quadrature formula* if it is exact for all polynomials from $\mathbb{P}_{2p+1}$, i. e.,

$$\int_{\alpha}^{\beta} w(t) P(t) \mathrm{d}t = \sum_{j=0}^{p} \omega_j P(t_j)$$

for all $P \in \mathbb{P}_{2p+1}$. ◊

There is a lot to say about Gaussian rules. The function w defines a scalar product on $C[\alpha, \beta]$ via

$$\langle f, g \rangle_w := \int_{\alpha}^{\beta} w(t) f(t) g(t) \mathrm{d}t.$$

It can be shown that there is a unique sequence $(q_p)_{p \geq 0}$ of polynomials with $q_0 \equiv 1$ and leading coefficient 1 for $p > 0$ such that q_p has degree p and such that the functions q_p, $p = 0, 1, \ldots$ form an orthogonal basis of the space of all polynomials. More precisely, we have

$$\langle q_p, q_r \rangle_w = 0 \text{ for } \qquad p \neq r,$$

and

$$\mathbb{P}_p = \operatorname{span}\{q_0, q_1, \ldots, q_p\} \text{ for } p \in \mathbb{Z}_{\geq 0},$$

see [62, Lem. 9.15]. Further, it can be shown that each of the polynomials q_p has exactly p distinct zeros in (α, β) ([62, Lem. 9.15]). We then have the following theorem.

Theorem 2.31 ([62, Thms. 9.17, 9.18])
For each $p = 0, 1, \ldots$ there is a unique Gaussian quadrature formula of order p. The corresponding integration points are the zeros of q_{p+1}. The corresponding weights are all positive.

So far it is not clear how to actually compute the weights and integration points of a Gauß rule. To see how this missing part can be accomplished, we study an article of Golub and Welsch [38]. In this article it was shown that a sequence of orthogonal polynomials $(q_p)_p$ fulfills a three term recurrence relationship,

$$q_p(x) = (a_p x + b_p) q_{p-1}(x) - c_p q_{p-2}(x), \quad p = 1, 2, \ldots \tag{2.33}$$

$$a_p, c_p \neq 0, \quad q_{-1} := 0, \quad q_0 \equiv 1. \tag{2.34}$$

By (2.33) it is clear that q_p must have degree p. The equations (2.33), (2.34) can be re-written in matrix notation [38, Sec. 2],

$$x \begin{bmatrix} q_0(x) \\ q_1(x) \\ \vdots \\ q_p(x) \end{bmatrix} = \begin{bmatrix} -b_1/a_1 & 1/a_1 & & & & \\ c_2/a_2 & -b_2/a_2 & 1/a_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & 1/a_{p-1} & \\ & & & c_{p+1}/a_{p+1} & -b_{p+1}/a_{p+1} & \end{bmatrix} \cdot \begin{bmatrix} q_0(x) \\ q_1(x) \\ \vdots \\ q_p(x) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ q_{p+1}(x)/a_{p+1} \end{bmatrix}. \tag{2.35}$$

If we set $\mathbf{q}(x) := [q_0(x), \ldots, q_p(x)]^\mathsf{T}$ and call the tridiagonal matrix in (2.35) T, we see that $q_{p+1}(t) = 0$ if and only if the eigenvector equation

$$\mathsf{T}\mathbf{q}(t) = \mathbf{q}(t)t$$

is fulfilled. In other words, the integration points t_j of the Gauß rule are the eigenvalues of T. It can be shown that the matrix T is real and symmetric when, as supposed, the polynomials form an orthonormal system. This is shown in [117, p. 54]. Next, let $\mathsf{q}_j = \mathbf{q}(t_j)$, $j = 0, \ldots, p$ denote the eigenvectors of T corresponding to eigenvalue (i. e., integration point) t_j. Suppose the vectors q_j form an orthonormal system of vectors. Then, Golub and Welsch [38] show that

$$\omega_j = \mathsf{q}_j(1)^2, \; j = 0, \ldots, p.$$

for the weight function $w \equiv 1$. Another explicit formula for the weights ω_j can also be given [20, p. 97]. The results are summarized in the following theorem.

Theorem 2.32
The integration points t_j of a Gauß rule of order p are the zeros of the polynomial q_{p+1}. Those correspond to the eigenvalues of the matrix T from (2.35). They are all distinct and reside in (α, β). The weights ω_j are given by

$$\omega_j = -\frac{k_{p+2}}{k_{p+1}} \cdot \frac{1}{q_{p+2}(t_j) \cdot q'_{p+1}(t_j)}$$

where k_{p+1}, k_{p+2} are the leading coefficients of q_{p+1}, q_{p+2}, respectively. Another formulation is

$$\omega_j = \mathsf{q}_j(1)^2, \ j = 0, \ldots, p,$$

where

$$\mathsf{q}_j = \mathbf{q}(t_j), \ j = 0, \ldots, p$$

denote the orthonormal eigenvectors of T.

For the computation of eigenvalues we refer to the discussion in Chapter 1.

The following theorem establishes the relation between Gaussian integration rules and interpolatory integration rules.

Theorem 2.33 ([62, Lem. 9.14])
Let $t_0, \ldots, t_p$ be $p+1$ distinct integration points chosen as the zeros of q_{p+1}. Then the corresponding interpolatory rule coincides with the Gauß rule that is given by these points and weights (and hence is exact for all polynomials from $\mathbb{P}_{2p+1}$).

It can be shown that Gauß type integration schemes are optimal in the sense that there is no formula of order p that is exact for all polynomials of degree $2p + 2$ [63, p. 138].

Let us finish this section with a remark concerning the integration interval.

Remark 2.34
In the literature Gauß–Legendre rules sometimes are defined for the interval $[-1, 1]$ (see, e. g., [19]). In this case, the points and weights have to be translated to the interval $[\alpha, \beta]$. The points can be transformed via a simple transformation

$$[-1, 1] \longrightarrow [\alpha, \beta], \ t \mapsto \frac{\beta - \alpha}{2} t + \frac{\alpha + \beta}{2}.$$

The weights then have to be multiplied by $(\beta - \alpha)/2$ to get the weights corresponding to the interval $[\alpha, \beta]$. The weights sum up to

$$\sum_{j=1}^{p} \omega_j = \beta - \alpha$$

in this case. This can be seen by integrating the constant function $f \equiv 1$. $\Diamond$

2.3.4 Error statements

In this section, we state in condensed form some error estimates for integration rules that we will make use of later.

Interpolatory rules

For some integration methods and sufficiently smooth functions very simple (but potentially rough) error estimates can be derived. For the closed (non-compound) Newton–Cotes formulas, they are, e. g.,

$$Q_2 f - Q(f) = \frac{(\beta-\alpha)^3}{12} f''(\xi_2),$$
$$Q_3 f - Q(f) = \frac{(\beta-\alpha)^5}{2880} f^{(4)}(\xi_3),$$

where $\xi_p \in [\,\alpha, \beta\,]$, see [63, p. 132]. Here, it is supposed that the respective derivative of f is continuous on $[\,\alpha, \beta\,]$. Upper bounds for the errors can be derived by taking supremum norms of the derivatives. Weideman [114] gives examples where these simple bounds overestimate the actual error by several orders of magnitude. Similar bounds can be derived for the open Newton–Cotes formulas.

From the simple bounds mentioned above, error estimates for the compound formulas can be derived. For instance, the compound trapezoidal formula $T_p f$ (2.32) takes the asymptotic error [63, p. 146]

$$T_p f - Q(f) \approx \frac{h^2}{12}(f'(\beta) - f'(\alpha)),\ h = \frac{\beta-\alpha}{p}.$$

Note that neither an interior point of $[\,\alpha, \beta\,]$ nor a higher derivative appears in this formula. The right hand side of the formula is zero for a periodic integrand. It can be shown that in this case the error only depends on the derivative of order $2m+1$ if $f \in C^{2m+1}(\mathbb{R})$, see Theorem 2.36 below.

Gauß rules

For Gauß type rules, similar error bounds as those for interpolatory rules can be proven. They also depend on the value of some derivative at an intermediate value. The following result is compiled from [20, p. 98]. First, recall the quantities k_p from Theorem 2.32.

Theorem 2.35
Let $w(t)$ be a weight function and let $(t_j, \omega_j)_{j=0,\dots,p}$ define a Gauß rule. Further, let $f \in C^{2p+2}[\alpha, \beta]$ (i. e., $f^{(2p+2)} \in C[\alpha, \beta]$). Then

$$\begin{aligned} E_{G_{p+1}}(f) &:= \int_\alpha^\beta w(t) f(t) \mathrm{d}t - \sum_{j=0}^{p} \omega_j f(t_j) \\ &= \frac{f^{(2p+2)}(\xi)}{(2p+2)! k_{p+1}^2} \end{aligned}$$

for some $\alpha < \xi < \beta$. In the case $w \equiv 1$ (i. e., when using Gauß–Legendre integration) we obtain

$$E_{G_{p+1}}(f) = \frac{(\beta-\alpha)^{2p+3}((p+1)!)^4}{(2p+3)((2p+2)!)^3} f^{(2p+2)}(\xi) \tag{2.36}$$

for some $\alpha < \xi < \beta$.

The factor in front of the derivative in (2.36) looks not very meaningful at first glance, but it is already of order 10^{-18} for $p = 7$ (i. e., 8 integration points), when the interval length is 2. Nonetheless, $\max_\xi \left| f^{(2p+2)}(\xi) \right|$ might grow very fast with p.

2.3.5 Integration of periodic functions

So far, everything was about integrating general, probably smooth, continuous functions. In particular, the error statements from Section 2.3.4 are valid for all functions that fulfill the respective requirements. The application of integration rules to a smaller class of functions should potentially deliver better results; this is true for interpolatory type rules applied to the class of periodic functions. In particular, let us consider the 2π-periodic functions on $\mathbb{R}$, i. e., the functions defined on $\mathbb{R}$ with $f(t) = f(t + 2\pi)$ for all $t \in \mathbb{R}$. It is intuitively understood that the trapezoidal rule works quite well for periodic functions:

> " When the function is periodic and one integrates over one full period, there are about as many sections of the graph that are concave up as concave down, so the errors cancel. This leaves one with a much better approximation than would have been the case had the function been monotonic."

This is how Weideman [114] would explain the phenomenon to a student.

Besides this understanding rigid error formulas that are much stricter than those for general functions can be derived. We will present two approaches in the following, one is based on the so called Euler-Maclaurin expansion and the other one on the theory of analytic functions. The fundamental difference is that the second one is derivative-free. We will make use of these error estimates later in the text, when coming to integration based eigensolvers.

Euler-Maclaurin-based error

The Euler-Maclaurin expansion [20, p. 136] is a formula that leads to an explicit expression for the error in the (compound) trapezoidal rule (2.32).

Theorem 2.36 ([62, Cor. 9.27])
Let $m, p \in \mathbb{Z}_{>0}$, $f \in C^{2m+1}(\mathbb{R})$ be 2π-periodic and T_p be the p-point compound trapezoidal rule (2.32)*. Then we have*

$$\begin{aligned} \left|E_{T_p}(f)\right| &:= \left| \int_0^{2\pi} f(t)\mathrm{d}t - T_p f \right| \\ &\leq \frac{2\zeta(2m+1)}{p^{2m+1}} \int_0^{2\pi} \left|f^{(2m+1)}(t)\right| \mathrm{d}t \qquad (2.37) \\ &\leq \frac{4\pi\zeta(2m+1)}{p^{2m+1}} \sup_t \left|f^{(2m+1)}(t)\right|, \qquad (2.38) \end{aligned}$$

where

$$\zeta(2m+1) = \sum_{k=1}^{\infty} \frac{1}{k^{2m+1}}$$

is the Riemann zeta function.

Note that (2.38) is a standard estimate applied to (2.37); this is also the form the theorem appears in, e. g., [20]. The theorem essentially says that the error of the p-point trapezoidal rule is small when f has derivatives that have a uniform bound. Note that the formula is valid for all values of m if $f \in C^\infty(\mathbb{R})$. It is not hard to see that the values of $\zeta(2m+1)$ are bounded for $m \to \infty$. Obviously $\zeta(2m+1) \geq 0$ for $m > 0$. Furthermore it can be shown that ζ can be extended to a function holomorphic on $\mathbb{C} \setminus \{1\}$ with negative derivative on $(1, \infty)$ [3, Ch. 5.4]. These facts imply that $\zeta(2m+1) \leq \zeta(2) = \pi^2/6$.

Derivative free error

In the preceding paragraph, we required the function to be sufficiently smooth and to be defined on the real numbers. If the integrand is even analytic on a strip in $\mathbb{C}$ containing the real axis, error bounds without derivatives in the estimate can be derived. Here, we state such a result in a very simple form. The proof is based on techniques from complex analysis, as the residue theorem, the Schwarz reflection principle and Cauchy's integral theorem [2]. It is not repeated here.

Theorem 2.37 ([62, Thm. 9.28])
Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be real analytic (i. e., it has a power series expansion) and of period 2π. Then there is a number $a > 0$ such that f can be extended to a bounded

2π-periodic analytic function in $D := \{z \in \mathbb{C} : -a < \operatorname{Im}(z) < a\} \subset \mathbb{C}$. *The error* $E_{T_p}(f)$ *of the trapezoidal rule can be estimated by*

$$\left|E_{T_p}(f)\right| \leq \frac{4\pi M}{\exp(pa) - 1}.$$

The constant M *can be chosen as* $\sup_{z \in D} |f(z)|$.

The theorem reveals that the error of the trapezoidal rule decays exponentially with the order p of the rule. It also depends on the size of the strip of analyticity D in the same manner (of course, for growing a, the number M will typically also grow). Several other estimates are available, which might be sharper depending on the function f, see [61]. Weideman [114] notes that in many cases the midpoint rule is as efficient as the trapezoidal rule. He states several examples for this fact, without quantifying it for the general case.

2.4 Eigensolvers based on integration

So far we did not actually discuss how the subspace $\mathcal{U}$ at the heart of the Rayleigh–Ritz-method is being computed. This—of course—is the crucial point of the method. The rest is just basic operations and the application of a standard software to solve the small scale eigenvalue problems.

In [60] we analyzed a method that is known as FEAST method and was proposed by Polizzi [85]. It has a plain Rayleigh–Ritz procedure at its core, as was pointed out in [60]. The subspace in discussion is computed by (numerical) contour integration as mentioned at the beginning of this chapter. Several other methods based on integration are described in the literature, while FEAST seems to be the most simple one, regarding presentation and implementation, and also quite powerful as we shall see later.

2.4.1 Literature review

Let us shortly review the available literature on eigensolvers based on numerical integration. To the best of our knowledge, one of the first methods of that kind was that of Sakurai and Sugiura published in 2003 [92]. An extension was published in 2009 by Sakurai and others [7]. Ikegami et al. further enhanced the method [49]. The Sakurai-Sugiura method and its descendants seem not as well-suited for high performance computations as FEAST.

Then, in 2009 Polizzi published his FEAST algorithm [85]. Recently, Laux [65] published a study concerning the application of FEAST to a problem from physics. We also published three papers concerning FEAST so far, see [33,34,60]. Recently, Tang and Polizzi published an analysis of the FEAST method [105].

Bertrand and Philippe [10] published an integration based method for counting eigenvalues over a decade ago. Beyn's integral method [11] is suited for the solution of nonlinear eigenvalue problems.

2.4.2 Spectral projectors and resolvent

Let us refine our knowledge on projectors, cf. page 5. The notion of a *spectral projector* plays an important role in eigenvalue computations. The spectral projector associated with an eigenvalue λ is the B-orthogonal projector P_λ that maps onto λ's eigenspace.

Let the pair (A, B) have the eigenvalues $\lambda_1, \ldots, \lambda_n$ together with B-orthonormal eigenvectors $\mathsf{x}_1, \ldots, \mathsf{x}_n$. We then have, for $i = 1, \ldots, n$,

$$\mathsf{A}\mathsf{x}_i = \mathsf{B}\mathsf{x}_i\lambda_i = \mathsf{B}\mathsf{x}_i\mathsf{x}_i^\star\mathsf{B}\mathsf{x}_i\lambda_i$$

since $\mathsf{x}_i^\star\mathsf{B}\mathsf{x}_i = 1$. This expression can further be extended to

$$\mathsf{A}\mathsf{x}_i = \sum_{j=1}^{n} \mathsf{B}(\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B})\mathsf{x}_i\lambda_j, \tag{2.39}$$

since all summands are zero but that one with $j = i$. Equation (2.39) is true for all eigenvectors x_i, $i = 1, \ldots, n$, which span the whole space $\mathbb{C}^n$, and hence for all vectors from that space. It follows

$$\mathsf{A} = \sum_{j=1}^{n} \lambda_j \mathsf{B}(\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}).$$

This can be simplified by grouping multiple eigenvalues. Let $\mu_1, \ldots, \mu_k$ be the distinct eigenvalues of (A, B) and X_{μ_j} be the matrix that collects the corresponding eigenvectors x_i; we then have

$$\mathsf{A} = \sum_{j=1}^{k} \mu_j \mathsf{B}(\mathsf{X}_{\mu_j}\mathsf{X}_{\mu_j}^\star\mathsf{B}).$$

This is, up to the choice of the basis matrices X_{μ_j}, a unique decomposition of (A, B) when multiple eigenvalues occur [80, p. 8], in contrast to the standard spectral Theorem 1.3. In that theorem, e. g., signs of eigenvectors and the order of eigenvalues are not uniquely determined. Due to the uniqueness of spectral projectors it is clear that the matrix $\mathsf{P}_{\mu_j} := \mathsf{X}_{\mu_j}\mathsf{X}_{\mu_j}^\star\mathsf{B}$ *is* the B-orthogonal spectral projector onto $\operatorname{span}(\mathsf{X}_{\mu_j})$. We then obtain the more abstract decomposition of (A, B),

$$\mathsf{A} = \sum_{j=1}^{k} \mu_j \mathsf{B}\mathsf{P}_{\mu_j},$$

which does not explicitly make use of eigenvectors and is unique. As we shall see below, the knowledge of eigenvectors is not necessary in order to compute spectral projectors, nonetheless both objects translate into each other. In this context, note once again that the specific representation of P_{μ_j} is not unique. Note also that the sum of spectral projectors maps onto the direct sum of their images. Using this fact, we can construct projectors belonging to eigenspaces that correspond to whole subsets of eigenvalues.

The use of spectral projectors in eigenvalue computations is quite obvious. Having a spectral projector P_λ for some eigenvalue λ at hand, we can multiply it to some test vectors Y to obtain a matrix $\mathsf{U} = \mathsf{P}_\lambda\mathsf{Y}$. We then can expect that U spans the eigenspace $\mathcal{U}_\lambda$. This indeed is true if Y has full rank, i. e., its rank equals the multiplicity of λ and if in addition no components of it are B-orthogonal to $\mathcal{U}_\lambda$. Recall Theorem 1.4 which states that a basis of an eigenspace is sufficient to compute the corresponding eigenvalues. It follows that our matrix U is the optimal candidate for the Rayleigh–Ritz process (in fact, no iterative "process" is necessary when an exact basis is available).

Integrating the resolvent

In the following, we essentially repeat our analysis from [60] to derive another representation of spectral projectors based on integration of the so called *resolvents*.

We begin by considering one eigenpair $(\mathsf{x}_k, \lambda_k)$ of (A, B). Let $z \in \mathbb{C}$ be any number that is no eigenvalue. We then have

$$(z\mathsf{B} - \mathsf{A})\mathsf{x}_k = (z - \lambda_k)\mathsf{B}\mathsf{x}_k \iff \mathsf{B}^{-1}(z\mathsf{B} - \mathsf{A})\mathsf{x}_k = (z - \lambda_k)\mathsf{x}_k,$$

in other words, $z - \lambda_k$ is an eigenvalue of $\mathsf{B}^{-1}(z\mathsf{B} - \mathsf{A})$. By inverting this matrix we obtain

$$(z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{x}_k = (z - \lambda_k)^{-1}\mathsf{x}_k. \tag{2.40}$$

In the following, let $G(z) := (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}$ be the so called *resolvent*. Note that $(z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B} = (z\mathsf{I} - \mathsf{B}^{-1}\mathsf{A})^{-1}$ and hence $G(z)$ coincides with the usual definition of a resolvent of the linear operator $\mathsf{B}^{-1}\mathsf{A}$, see [56]. Now, let $\mathcal{C}_k$ be a curve in $\mathbb{C}$ surrounding eigenvalue λ_k and no other (recall Definition 2.25 and that we suppose all curves to be piecewise differentiable). Integrating $(z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}$ along $\mathcal{C}_k$ yields the integral

$$\mathsf{Q} := \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}_k} (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathrm{d}z. \tag{2.41}$$

We shall see that this is the projector onto the eigenspace belonging to λ_k, i. e., to $\mathrm{null}(\lambda_k\mathsf{B} - \mathsf{A})$. First, let us analyze the function $z \mapsto G(z) = (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}$ further. It is obviously defined on $\mathbb{C} \setminus \mathrm{spec}(\mathsf{A}, \mathsf{B})$; in the eigenvalues it has singularities. Saad [91] gives a short analysis, once more only for the standard case $\mathsf{B} = \mathsf{I}$; but

as explained before everything also applies to $(z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}$. Let us follow his lines here.

To that end, let z_0 be any point that is no eigenvalue, then

$$\begin{aligned} G(z) &= (z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B} \\ &= ((z_0\mathsf{B}-\mathsf{A})-(z_0-z)\mathsf{B})^{-1}\mathsf{B} \\ &= (\mathsf{B}^{-1}(z_0\mathsf{B}-\mathsf{A})-(z_0-z)\mathsf{I})^{-1} \\ &= G(z_0)(\mathsf{I}-(z_0-z)G(z_0))^{-1}. \end{aligned}$$

Hence, due to the Neumann series expansion [56], the function $G(z)$ can be expanded into a Taylor series in the open disk with center z_0 and radius $1/\rho(G(z_0))$. This disc of analyticity then has an eigenvalue of $(\mathsf{A},\,\mathsf{B})$ on its boundary. Consequently, Cauchy's theorem is applicable in the region of analyticity.

Using Cramer's rule, it can be seen that

$$G(z) = \frac{1}{\det(z\mathsf{B}-\mathsf{A})\det(\mathsf{B}^{-1})}\mathrm{adj}(\mathsf{B}^{-1}(z\mathsf{B}-\mathsf{A})),$$

where $\det(z\mathsf{B}-\mathsf{A})$ is a polynomial in z of degree n. The adjugate matrix $\mathrm{adj}(\mathsf{B}^{-1}(z\mathsf{B}-\mathsf{A}))$ (sometimes simply called adjoint, which could lead to confusion with the Hermitian adjoint matrix) is a matrix of the same size as its argument, with determinants of certain submatrices as entries. This means that the entries are polynomials in z. In consequence, $G(z)$ is a matrix whose elements are rational functions in z. For more information on adj and on Cramer's rule, see [112]. To sum up, $G(z)$ is a function defined on $\mathbb{C}\setminus\mathrm{spec}(\mathsf{A},\,\mathsf{B})$ with non-essential singularities in the eigenvalues of $(\mathsf{A},\,\mathsf{B})$.

Let us return to the integral (2.41) and apply it to the eigenvector x_k. By using eq. (2.40) we obtain

$$\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}_k}(z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}\mathsf{x}_k\mathrm{d}z = \frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}_k}\frac{1}{z-\lambda_k}\mathsf{x}_k\mathrm{d}z = \frac{1}{2\pi\mathbf{i}}2\pi\mathbf{i}\mathsf{x}_k = \mathsf{x}_k,$$

where the last equation is due to Cauchy's Theorem 2.28. Taking the curve around any other eigenvalue will deliver an integral that is zero. This shows that $\mathrm{span}(\mathsf{x}_k)\subseteq\mathrm{range}(\mathsf{Q})$ and that $\mathsf{Q}^2=\mathsf{Q}$ on $\mathrm{span}(\mathsf{x}_k)$. It can easily be shown that Q itself is a projector [91]. Note that the value of the integrals does not depend on the actual choice of the curve $\mathcal{C}$, as long as it fulfills the respective requirements. One might hence choose a very simple curve, i. e., a circle of appropriate radius and placement.

Next, we consider a bunch $\{\lambda_k : k\in I\}$ of eigenvalues for some index set I. Let $\mathcal{C}$ be a curve surrounding this subset and no other eigenvalue and let $\mathcal{C}_k$ be

a curve surrounding only eigenvalue λ_k for each $k \in I$. We then obtain for an eigenvector x_j [60]

$$\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} G(z)\mathrm{d}z\mathsf{x}_j = \sum_{k\in I}\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}_k} G(z)\mathrm{d}z\mathsf{x}_j = \sum_{k\in I}\delta_{k,j}\mathsf{x}_j = \begin{cases}\mathsf{x}_j, & \text{if } j\in I\\ 0, & \text{otherwise}\end{cases}. \quad (2.42)$$

On the other hand, if we collect the vectors x_k, $k \in I$, in a B-orthogonal $n \times |I|$-matrix X and form the corresponding orthogonal projector $\mathsf{X}\mathsf{X}^\star\mathsf{B}$, we obtain the B-orthogonal projector onto span(X) according to the beginning of Section 2.4.2. If we compare the effect of multiplying $\mathsf{X}\mathsf{X}^\star\mathsf{B}$ with any of the basis vectors $\mathsf{x}_k \in$ span(X) with that one described by equation (2.42) we see that it is the same. Hence the integral on the left hand side of (2.42) and the spectral projector $\mathsf{X}\mathsf{X}^\star\mathsf{B}$ coincide.

We now again follow Saad [91, Sec. 3.1.4], to show—without the detour over the eigenvector representation—that the operator Q from (2.41) maps *into* the eigenspace belonging to eigenvalue λ_k. This means that we will see range(Q) $\subseteq$ null($\mathsf{B}^{-1}\mathsf{A} - \lambda_k\mathsf{I}$). By the foregoing analysis it will then be clear that range(Q) is exactly the eigenspace belonging to λ_k, since we already know that span(x_k) $\subseteq$ range(Q) for every eigenvector x_k belonging to λ_k. Now, let us drop the subscript k for λ and $\mathcal{C}$; we then have for any $z \notin$ spec(A, B)

$$(z-\lambda)\mathsf{I} = \mathsf{B}^{-1}(z\mathsf{B}-\mathsf{A}) - \mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A}).$$

By right-multiplying with $G(z)$ we obtain

$$(z-\lambda)G(z) = \mathsf{I} - \mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A})G(z). \quad (2.43)$$

Next, the integral over $\mathcal{C}$ is applied to both sides of the equation, the outcome of which is

$$\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}}(z-\lambda)G(z)\mathrm{d}z = -\mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A})\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} G(z)\mathrm{d}z \quad (2.44)$$

$$= -\mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A})\mathsf{Q} \quad (2.45)$$

since integrating I over a closed curve yields zero.

By multiplying (2.44)–(2.45) with $\mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A})$ from left and using (2.43) it can inductively be seen that

$$\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}}(z-\lambda)^m G(z)\mathrm{d}z = -(\mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A}))^m\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} G(z)\mathrm{d}z$$

$$= -(\mathsf{B}^{-1}(\lambda\mathsf{B}-\mathsf{A}))^m\mathsf{Q} \quad (2.46)$$

for all integers $m > 0$. The leftmost term of (2.46) is exactly the coefficient with index $-(m+1)$ of the Laurent expansion of $G(z)$ around λ. Since λ is a

non-essential singularity of the function G, see above, there is a number m^* such that

$$(\mathsf{B}^{-1}(\mathsf{A}-\lambda\mathsf{B}))^m\mathsf{Q}\mathsf{x} = (\mathsf{B}^{-1}\mathsf{A}-\lambda\mathsf{I})^m\mathsf{Q}\mathsf{x} = \mathsf{o}$$

for all $m \geq m^*$ and $\mathsf{x} = \mathsf{Q}\mathsf{x} \in \text{range}(\mathsf{Q})$. This means nothing else but that Q maps into the eigenspace belonging to λ.

We now have two representations of the spectral projector, in addition to the abstract one, at hand; the one based on eigenvectors and the one based on integration. The latter one is even more general since it does not require the matrix A or B to be Hermitian. Then, still a projector is obtained, but an oblique one. The notion of B-orthogonality then is not even more properly defined.

Remark 2.38 (Singularities of $G(z)$)
It can easily be seen, e. g., by considering the eigenvector expansion

$$G(z) = \sum_{j=1}^{n} \frac{1}{z-\lambda_j}\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}$$

that $G(z)$ has poles of order 1 in the eigenvalues λ_j of (A, B). ◊

Remark 2.39 (Eigenvectors of Q)
It is worth noting that by interpreting (2.42) we see that the eigenvectors of Q are just the eigenvectors of (A, B). Eigenvector x_j corresponds to eigenvalue 1 of Q if for the corresponding eigenvalue λ_j of (A, B) we have $j \in I$. All other eigenvectors x_j correspond to eigenvalue 0 of Q. ◊

2.4.3 Computing an eigenspace

Suppose we have the integral

$$\mathsf{Q} := \frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} (z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}\mathrm{d}z \tag{2.47}$$

at hand, where $\mathcal{C}$ is surrounding an interval I_λ which contains some eigenvalues $\Theta = \{\lambda_1, \ldots, \lambda_k\}$. We do not care for the numbering at the moment, the eigenvalues are not supposed to be ordered according to their size. Let $\mathsf{x}_1, \ldots, \mathsf{x}_k$ be the corresponding eigenvectors, let $\mathcal{X}$ be their span, i. e., the corresponding eigenspace and let X be the matrix consisting of the eigenvectors.

We saw that $\text{range}(\mathsf{Q}) = \mathcal{X}$, so we have to apply Q to "enough" (and suitable) vectors in order to obtain a basis of $\mathcal{X}$. To this end, let Y be a full rank $n \times k$-matrix and compute $\mathsf{U} := \mathsf{Q}\mathsf{Y} = \mathsf{X}\mathsf{X}^\star\mathsf{B}\mathsf{Y}$. We see that U is a linear combination of the columns of X; to be precise

$$\mathsf{U} = \mathsf{X}\cdot(\mathsf{X}^\star\mathsf{B}\mathsf{Y}).$$

Hence if Y was chosen carefully—of size k with full rank and no components B-orthogonal to $\mathcal{X}$—we have that $\mathcal{U} := \operatorname{span}(\mathsf{U}) = \mathcal{X}$. In other words, we found an *exact* eigenspace, a complicated object it is usually hard to find a nice formula for. The eigenvalues of $(\mathsf{U}^\star\mathsf{AU}, \mathsf{U}^\star\mathsf{BU})$ are the elements of Θ.

We now could also append some columns to Y and would still obtain a space $\mathcal{U}$ that contains $\mathcal{X}$. Note that the requirements of the Rayleigh–Ritz Theorem 2.1 are also met exactly, hence an approximation to U will be a perfect candidate for a basis in the Rayleigh–Ritz method. However, the computation of U is no easy task, neither with respect to numerical effort nor to other issues such as exactness. At least two kinds of error will be introduced. The first one is that of the linear systems involving the matrix $z\mathsf{B}-\mathsf{A}$ and the other one is the approximation error of the integration method in use.

The algorithm that arises when first applying projector (2.47) to some matrix Y and then performing a Rayleigh–Ritz process with the so obtained basis U is nothing but Polizzi's FEAST algorithm [85]. This dissection of FEAST was performed previously in [60]. Further, the repetition of the two steps mentioned above is nothing but subspace iteration with the matrix Q, see Section 2.1.2 and [105, 111].

2.5 Error analysis of integration based eigensolvers

2.5.1 Introduction

In the following, we will further analyze the errors that occur when solving eigenvalue problems with eigensolvers based on numerical integration. While in Section 2.1 we focused on general subspace based eigensolvers and different kinds of errors, we now discuss the errors that arise when computing the eigenspace. The eigenspace in the integration based solver is computed as a contour integral of the resolvent of (A, B), hence the error in the subspace U is the sum of the errors from numerical integration and the solution of linear systems involving the matrix $z\mathsf{B} - \mathsf{A}$.

We start by giving convergence proofs for the trapezoidal as well as for the Gauß–Legendre rule. This means $\|\mathsf{U} - \widetilde{\mathsf{U}}_p\| \to 0$ for growing integration order p, where $\widetilde{\mathsf{U}}_p$ denotes the approximation of U obtained by numerical integration of order p with one of the schemes noted above. In Section 2.5.2, the error of the trapezoidal rule is analyzed, the error bound does not contain any derivatives. Similarly, in Section 2.5.3 a derivative free error bound for the Gauß–Legendre rule is obtained. Recall, e. g., Theorem 2.35 which gives an error bound for the Gauß–Legendre rule depending on the derivatives, and does therefore not ensure convergence unless enough information about the derivatives is available. The results obtained are more formal proofs of convergence than practical error bounds, since they are very pessimistic. In practice, often much better results are

achieved. Further, the computation of U is not at the heart of our interest, it is only an intermediate step to the solution of a certain eigenproblem. However, we considered it important to give a convergence proof. Together with the results from Section 2.1 it is the theoretical justification for the use of eigensolvers based on integration.

In Section 2.5.4 the impact of using different integration contours is discussed. Finally, in Section 2.5.5, we explain the effect of the errors that occur in the solution of linear systems.

Setting

In the following, suppose that all occurring resolvents $(z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}$ are computed exactly. We have to make a concrete choice for the curve $\mathcal{C}$, as mentioned before. Suppose we are seeking the real eigenvalues of the pair (A, B) in the interval $I_\lambda = \left[\underline{\lambda}, \overline{\lambda}\right]$. Let $c := (\underline{\lambda}+\overline{\lambda})/2$ denote the center of the interval and $r := (\overline{\lambda}-\underline{\lambda})/2$ the radius. At the moment, the reader may also think of r being slightly larger than the actual radius of I_λ. The curve $\mathcal{C}$ can be chosen as a circle with radius r and center c. This curve can be parametrized by the function

$$\varphi : [0, 2\pi] \longrightarrow \mathbb{C},\ \varphi(t) = c + r\exp(\mathbf{i}t),$$

Note that φ is 2π-periodic. Clearly, other curves that admit 2π-periodic parametrizations are possible. Of course, 2π is no magic number in this setting, other periods are possible as well but for simplicity we restrict ourselves to this period. All other periods can easily be scaled to 2π.

Integrating $(z\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY}$ over $\mathcal{C}$ then results in

$$\begin{aligned}\mathsf{U} &= \frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} (z\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY}\mathrm{d}z \\ &= \frac{1}{2\pi\mathbf{i}}\int_0^{2\pi} \mathbf{i}r\exp(\mathbf{i}t)((c+r\exp(\mathbf{i}t))\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY}\mathrm{d}t, \end{aligned} \tag{2.48}$$

recall Definition 2.26. Note that $\mathbf{i}$ cancels. Applying any numerical integration rule Q_p to the last integral yields an approximation $Q_p h = \widetilde{\mathsf{U}}_p \approx \mathsf{U}$, where h denotes the integrand of (2.48). In the following, $E_{T_p}(h)$, $E_{G_p}(h)$ denote the errors $\mathsf{U} - Q_p h$ for the trapezoidal and Gauß–Legendre rules, respectively.

2.5.2 Error in the integration—Trapezoidal rule

We start by analyzing the error that arises when applying techniques from numerical integration to compute the subspace U. The projector Q has to be applied

to some starting basis Y in order to obtain

$$\mathsf{U} = \mathsf{QY} = \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{BY}\mathrm{d}z.$$

Note that U now denotes the exact space.

Euler-Maclaurin Error

As a first approach, one can employ the trapezoidal rule from equation (2.32) to approximate the integral (2.48). Since the integrand

$$h(t) := \frac{1}{2\pi} r \exp(\mathbf{i}t)((c + r\exp(\mathbf{i}t))\mathsf{B} - \mathsf{A})^{-1}\mathsf{BY} \tag{2.49}$$

is periodic, we can invoke Theorem 2.36 concerning the error in the trapezoidal rule. The function h is differentiable and has k-th derivative [16]

$$h^{(k)}(t) = \frac{\mathbf{i}^k}{2\pi} \sum_{j=0}^{k} (\varphi(t) - c)^{j+1} c_{jk} (\varphi(t)\mathsf{I}_n - \mathsf{B}^{-1}\mathsf{A})^{-(j+1)}\mathsf{Y},$$

where c_{jk} are some constants. The formula in [16] is more complicated since it also involves a certain function f and its derivatives. For a general curve φ the formula would also involve different derivatives of φ and their powers. For any k and any order p of the integration method we have

$$\|\mathsf{U} - T_p h\| \leq \frac{4\pi\zeta(2k+1)}{p^{2k+1}} \sup_{0\leq t\leq 2\pi} \left\|h^{(2k+1)}(t)\right\|. \tag{2.50}$$

As mentioned before, the problem is the bound that depends on the derivatives of h. Davies and Higham [16] point out the problems.

- We have that $|(\varphi(t) - c)^{j+1}| = r^{j+1}$, the bound (2.50) is hence asymptotically proportional to $r^{2k+2}/p^{2k+1} = r(r/p)^{2k+1}$. Therefore, r should be (much) smaller than p to achieve a small error for a moderate number k.
- The norms of $(\varphi(t)\mathsf{I}_n - \mathsf{B}^{-1}\mathsf{A})^{-(1+j)}$ tend to be large when φ passes the eigenvalues of (A, B) too close.
- The constants c_{jk} increase quickly with k, see [16]. The error is also proportional to the norm of Y.

These effects show that, first, the circle defined by c and r should be chosen small and, second, such that it does not come too close to eigenvalues. The matrix Y should be chosen with small norm, e. g., with orthonormal columns.

Derivative free error

Now, we develop error bounds for the trapezoidal rule applied to the integral U that do *not* depend on the derivatives of h, the quantities that cause the large terms in the error bounds stated above. The result is similar to Theorem 2.37.

If f is a 2π-periodic function, the compound trapezoidal rule of order p (i. e., with $p+1$ points) (2.32) reads for f being defined on $[0, 2\pi]$

$$\frac{1}{2\pi}\int_0^{2\pi} f(t)\mathrm{d}t \approx \frac{1}{p}\sum_{j=0}^{p-1} f\left(\frac{2\pi j}{p}\right), \tag{2.51}$$

since $f(0) = f(2\pi)$. In particular, the summation index only ranges up to $p-1$. Note that (2.51) is the trapezoidal rule multiplied by $1/(2\pi)$.

Subsequently, we follow a recent analysis by Beyn [11]. He showed that the trapezoidal rule converges when applying it to the function h (2.49) under certain conditions. To get a better understanding *why* the trapezoidal rule for periodic functions works so well we present large parts of Beyn's analysis, while adapting some of his statements to our problems. We mainly repeat his analysis concerning the convergence of the trapezoidal rule applied to scalar valued periodic functions and add some additional explanation where necessary. The reader will gain some insight why the considered functions have to be holomorphic and periodic. Beyn's work [11] is about the solution of nonlinear eigenvalue problems, it hence includes a much more complicated theory concerning the eigenvalue problem.

Let us start with a theorem that is a generalization of Theorem 2.37. The strip of analyticity is allowed to be unsymmetric and the function can take complex values for real arguments.

Theorem 2.40 ([11, Theorem 4.1])
Let $f \in H(S(s_-, s_+), \mathbb{C})$ be 2π-periodic on the strip

$$S = S(s_-, s_+) := \{z \in \mathbb{C} : s_- < \mathrm{Im}(z) < s_+\},\ s_- < 0 < s_+$$

containing the real line. Then the error of the trapezoidal sum

$$E_{T_p}(f) := \frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{d}x - \frac{1}{p}\sum_{j=0}^{p-1} f\left(\frac{2\pi j}{p}\right) \tag{2.52}$$

satisfies for all $0 > \sigma_- > s_-$ and $0 < \sigma_+ < s_+$:

$$\left|E_{T_p}(f)\right| \le \max_{\mathrm{Im}(z)=\sigma_+} |f(z)|F(\exp(-p\sigma_+)) + \max_{\mathrm{Im}(z)=\sigma_-} |f(z)|F(\exp(p\sigma_-)),$$

where $F(t) = \frac{t}{1-t}$ for $t \neq 1$.

To prove this theorem, we first state a lemma concerning the error of the trapezoidal rule applied to a holomorphic function on an annulus. It was basically already given in [20, Sec. 4.6.5], while Beyn [11] generalizes it slightly.

Lemma 2.41 (Beyn, [11, Thm. 4.3])
Define the annulus

$$A = A_R(a_-, a_+) := \{\zeta \in \mathbb{C} : a_- R < |\zeta| < a_+ R\},$$
$$a_- < 1 < a_+,\ R > 0.$$

Let $g \in H(A_R(a_-, a_+), \mathbb{C})$ and let $a_- < \alpha_- < 1 < \alpha_+ < a_+$. Suppose g is being integrated over the circle $|\zeta| = R$. Then, the error $E_{T_p}(g)$ of the trapezoidal rule on this circle fulfills

$$\left|E_{T_p}(g)\right| \leq \max_{|\zeta|=\alpha_+ R} |g(\zeta)|\, F(\alpha_+^{-p}) + \max_{|\zeta|=\alpha_- R} |g(\zeta)|\, F(\alpha_-^{p}) \tag{2.53}$$

with $F(t) = \frac{t}{1-t}$.

Proof. (See [11].) The function g has a Laurent expansion on the annulus A (see Theorem 2.29), i. e., we may write

$$g(\zeta) = \sum_{k=-\infty}^{\infty} a_k \zeta^k \tag{2.54}$$

for some coefficients a_k. The coefficients have the form

$$a_k = \frac{1}{2\pi \mathbf{i}} \int_{\mathcal{C}} \frac{g(\zeta)}{\zeta^{k+1}} \mathrm{d}\zeta, \tag{2.55}$$

where $\mathcal{C}$ is a circle in A with radius $R > 0$ around zero. The error $E_{T_p}(g)$ takes the form

$$E_{T_p}(g) := \frac{1}{2\pi \mathbf{i}} \int_{|\zeta|=R} g(\zeta)\mathrm{d}\zeta - \frac{R}{p} \sum_{j=0}^{p-1} g\left(R \exp\left(\frac{2\pi \mathbf{i}}{p}\right)^j\right) \exp\left(\frac{2\pi \mathbf{i}}{p}\right)^j \tag{2.56}$$

for the parametrization $t \mapsto R\exp(\mathbf{i}t)$. Next, let us compute the error for the monomial functions $u_k(\zeta) = \zeta^k,\ k \in \mathbb{Z}$. The exact integral in (2.56) is zero for u_k for all $k \neq -1$. For the trapezoidal formula over the circle $|\zeta| = R$ applied to u_k we obtain

$$\frac{R}{p} \sum_{j=0}^{p-1} \left(R \exp\left(\frac{2\pi \mathbf{i}}{p}\right)^j\right)^k \exp\left(\frac{2\pi \mathbf{i}}{p}\right)^j = \frac{R^{k+1}}{p} \sum_{j=0}^{p-1} \exp\left(\frac{2\pi \mathbf{i} j(k+1)}{p}\right).$$

The expression $\exp(2\pi \mathbf{i}(k+1)/p)$ is a p-th root of unity, hence its j-th powers sum up to p if $k+1$ is a multiple of p or to 0 if this is not the case. We obtain for the error

$$E_{T_p}(u_k) = \begin{cases} -R^{\ell p}, & k+1 = \ell p,\ 0 \neq \ell \in \mathbb{Z} \\ 0, & \text{otherwise} \end{cases}.$$

The error $E_{T_p}(u_{-1})$ is zero since

$$\frac{1}{2\pi\mathbf{i}}\int_{|\zeta|=R}\frac{1}{\zeta}\mathrm{d}\zeta = 1$$

and this is also the value the trapezoidal rule delivers if applied to this function. Now, plugging the Laurent expansion of g, (2.54)–(2.55), into the error operator E_{T_p} yields

$$E_{T_p}(g) = -\sum_{\ell=1}^{\infty}(a_{\ell p}R^{\ell p} + a_{-\ell p}R^{-\ell p}). \tag{2.57}$$

Every single term of (2.57) with positive index can be estimated as follows,

$$\begin{aligned}
\left|a_{\ell p}R^{\ell p}\right| &= \left|\frac{R^{\ell p}}{2\pi\mathbf{i}}\int_{|\zeta|=R} g(\zeta)\zeta^{-\ell p-1}\mathrm{d}\zeta\right| \\
&= \left|\frac{R^{\ell p}}{2\pi\mathbf{i}}\int_{|\zeta|=\alpha_+ R} g(\zeta)\zeta^{-\ell p-1}\mathrm{d}\zeta\right| \\
&\leq \frac{R^{\ell p}}{2\pi}\cdot 2\pi\alpha_+ R\cdot \max_{|\zeta|=\alpha_+ R}|g(\zeta)|\,(\alpha_+ R)^{-\ell p-1} \\
&= \max_{|\zeta|=\alpha_+ R}|g(\zeta)|\,(\alpha_+^{-p})^{\ell}.
\end{aligned} \tag{2.58}$$

The integral can be taken over the slightly larger circle in (2.58) due to Cauchy's Theorem 2.28 and the structure of A. Similarly, for the coefficients with negative index, we have

$$\begin{aligned}
\left|a_{-\ell p}R^{-\ell p}\right| &= \left|\frac{R^{-\ell p}}{2\pi\mathbf{i}}\int_{|\zeta|=R} g(\zeta)\zeta^{-(-\ell p)-1}\mathrm{d}\zeta\right| \\
&= \left|\frac{R^{-\ell p}}{2\pi\mathbf{i}}\int_{|\zeta|=\alpha_- R} g(\zeta)\zeta^{\ell p-1}\mathrm{d}\zeta\right| \\
&\leq \frac{R^{-\ell p}}{2\pi}\cdot 2\pi\alpha_- R\cdot \max_{|\zeta|=\alpha_- R}|g(\zeta)|\,(\alpha_- R)^{\ell p-1} \\
&= \max_{|\zeta|=\alpha_- R}|g(\zeta)|\,(\alpha_-^{p})^{\ell}.
\end{aligned}$$

Observing that the term with $\ell = 0$ in (2.57) is missing, we obtain the desired result (2.53) because

$$\sum_{\ell=1}^{\infty}(\alpha_+^{-p})^{\ell} = \frac{\alpha_+^{-p}}{1-\alpha_+^{-p}} = F(\alpha_+^{-p})$$

and likewise for the terms of (2.57) with negative index. □

Next, we proceed to the proof of Theorem 2.40 (see [11]).

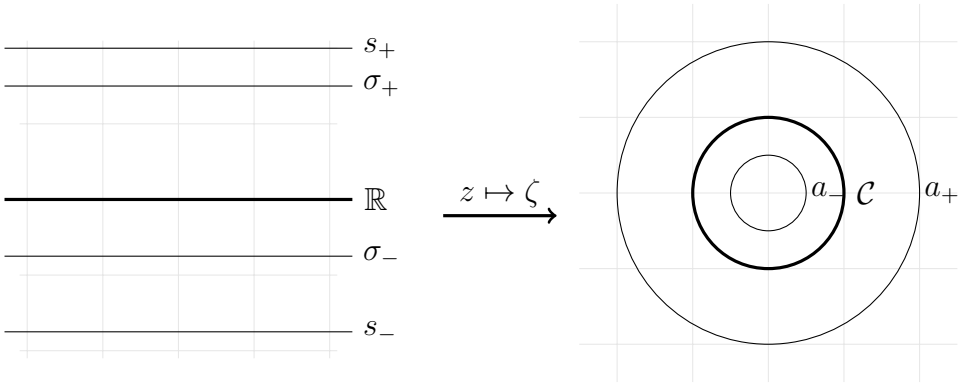

Figure 2.2: The strip S and the annulus A, transformed via the map $z \mapsto \zeta = \exp(\mathbf{i}z)$. The upper boundary of S is mapped to the inner circle with radius a_-. The lower boundary of S is mapped to the outer circle with radius a_+. We have $a_- = \exp(-s_+)$, $a_+ = \exp(-s_-)$. The real line is mapped to the circle $\mathcal{C}$ with radius 1.

Proof of Theorem 2.40. The strip $S(s_-, s_+)$ is mapped bijectively (modulo periodicity) onto the annulus

$$\begin{aligned} A = A_1(a_-, a_+) &= \{\zeta \in \mathbb{C} : a_- < |\zeta| < a_+\}, \\ a_- &= \exp(-s_+), \\ a_+ &= \exp(-s_-) \end{aligned}$$

via the map $z \mapsto \zeta := \exp(\mathbf{i}z)$. Note that this map transforms lines that are parallel to the real axis to circles. For a complex number z we have $|\exp(\mathbf{i}z)| = \exp(-\operatorname{Im}(z))$. It follows that a number z with $\operatorname{Im}(z) < 0$ is mapped to the exterior of the unit circle, a number with $\operatorname{Im}(z) > 0$ to the interior. The regions S and A and some transformed quantities can be seen in Figure 2.2.

Now, if f is analytic on the strip S defined in Theorem 2.40, it has a Fourier expansion [2, Ch. 7]

$$f(z) = \sum_{k=-\infty}^{\infty} a_k \exp(\mathbf{i}kz).$$

This is a special version of the Laurent expansion. By changing variables

$$z = \mathbf{i}^{-1} \log \zeta \iff \zeta = \exp(\mathbf{i}z),$$

we find that $g(\zeta) := f(z) = f(\mathbf{i}^{-1} \log \zeta)$ is defined on A, where it also is analytic. Note that the log function can be well defined on A due to the periodicity of the exponential. In particular g is a well defined function due to the periodicity of f. It follows that $g(\zeta)$ has a Laurent expansion on the annulus A with the same coefficients as the Fourier expansion of f. Transforming the integral from the

strip to the annulus yields

$$\frac{1}{2\pi}\int_0^{2\pi} f(z)\mathrm{d}z = \frac{1}{2\pi}\int_{\mathcal{C}} \mathbf{i}^{-1}(\frac{\mathrm{d}}{\mathrm{d}\zeta}\log(\zeta))g(\zeta)\mathrm{d}\zeta$$
$$= \frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} \zeta^{-1}g(\zeta)\mathrm{d}\zeta. \tag{2.59}$$

Hence we have to apply formula (2.53) to the function $\tilde{g}(\zeta) := \zeta^{-1}g(\zeta)$ which is holomorphic in A_1. The error bound from Lemma 2.41 becomes

$$\left|E_{T_p}(\tilde{g})\right| \leq \max_{|\zeta|=\alpha_+ R} \frac{1}{R}\left|g(\zeta)\right| F(\alpha_+^{-p}) + \max_{|\zeta|=\alpha_- R} \frac{1}{R}\left|g(\zeta)\right| F(\alpha_-^{p}),$$

where $R = 1$, $\alpha_- = \exp(-\sigma_+)$, $\alpha_+ = \exp(-\sigma_-)$. Note that the circles defined by the radii α_-, α_+ reside in the annulus A. Next, note that the application of the trapezoidal rule to the integral (2.59) in the annulus yields—by construction—the same result as the application of the trapezoidal rule to f in the interval $[0, 2\pi] \subset S$. Therefore, we obtain for the error (2.52)

$$\left|E_{T_p}(f)\right| \leq \max_{|\zeta|=\alpha_+} \left|g(\zeta)\right| F(\alpha_+^{-p}) + \max_{|\zeta|=\alpha_-} \left|g(\zeta)\right| F(\alpha_-^{p}).$$

Via the variable transformation $\zeta \mapsto z$, the circles $|\zeta| = \alpha_-$, $|\zeta| = \alpha_+$ map back to lines $\mathrm{Im}(z) = \sigma_-$, $\mathrm{Im}(z) = \sigma_+$, respectively. We have $\alpha_-^p = \exp(-\sigma_+)^p = \exp(-p\sigma_+)$ and $\alpha_+^{-p} = \exp(-\sigma_-)^{-p} = \exp(p\sigma_-)$. By using $g(\zeta) = f(z)$ we obtain the desired result. □

In contrast to Theorem 2.37, the strip S in Theorem 2.40 may be unsymmetric with respect to the real axis. Furthermore, the maxima are not taken over the whole strip but only only over a line parallel to the real numbers that can be chosen arbitrarily.

Subsequently, we will—in foresight to the integration of the resolvent $G(z) = (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}$ from section 2.4.2—investigate the functions

$$r_\lambda(z) := (z - \lambda)^{-1}.$$

Note that $G(z)$ has eigenvalues $(z - \lambda)^{-1}$ for any eigenvalue $\lambda \neq z$ of the pair (A, B). Consequently it has the eigenvector expansion $G(z) = \sum_{k=1}^{n} r_{\lambda_k}(z)\mathsf{x}_k\mathsf{x}_k^\star\mathsf{B}$. For the Laurent coefficients of r_λ around 0,

$$a_k = \frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}} \frac{r_\lambda(\zeta)}{\zeta^{k+1}}\mathrm{d}\zeta,\ 0 \in \mathrm{Int}(\mathcal{C}),$$

we have that $a_k = 0,\ k \leq -1$ for $\lambda \in \mathrm{Ext}(\mathcal{C})$ and $a_k = 0,\ k \geq 0$ for $\lambda \in \mathrm{Int}(\mathcal{C})$, see [11]. Hence, in formula (2.53), only one of the two summands has to be considered. Nonetheless, we obtain independently of the position of λ

$$\begin{aligned} E_{T_p}(r_\lambda) &\leq \max_{|\zeta|=\alpha_+ R} |r_\lambda(\zeta)|\, F(\alpha_+^{-p}) + \max_{|\zeta|=\alpha_- R} |r_\lambda(\zeta)|\, F(\alpha_-^{p}) \\ &= \frac{1}{\min_{|\zeta|=\alpha_+ R} |\zeta - \lambda|} F(\alpha_+^{-p}) + \frac{1}{\min_{|\zeta|=\alpha_- R} |\zeta - \lambda|} F(\alpha_-^{p}). \end{aligned}$$

Note that the min terms boil down to $\min\{|\alpha_\pm R - \lambda|\,, |-\alpha_\pm R - \lambda|\}$ if λ is real as supposed in this whole work.

In a numerical setting we will however not want to integrate functions r_λ around a circle $\mathcal{C}$ directly, but apply the trapezoidal rule to the function $r_\lambda \circ \varphi$ on a real interval, see the introduction on page 67. The function φ is, due to periodicity, not only defined on the interval $[0, 2\pi]$ but on all real numbers. Suppose further that φ can be extended analytically to the strip $S(s_-, s_+)$ from Theorem 2.40. This is, for instance, the case for $\varphi(t) = c + r\exp(\mathbf{i}t)$.

Let s_-, s_+ be chosen such that φ does not hit λ (for the conditions, see Section 2.5.3), we then have

$$\frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} r_\lambda(z)\mathrm{d}z = \frac{1}{2\pi\mathbf{i}} \int_0^{2\pi} r_\lambda(\varphi(t))\varphi'(t)\mathrm{d}t.$$

Consequently, for the error it holds [11]

$$\begin{aligned} E_{T_p}(r_\lambda) &= \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} r_\lambda(z)\mathrm{d}z - \frac{1}{\mathbf{i}p} \sum_{j=0}^{p-1} r_\lambda\left(\varphi\left(\frac{2\pi j}{p}\right)\right) \varphi'\left(\frac{2\pi j}{p}\right) \\ &= \frac{1}{2\pi\mathbf{i}} \int_0^{2\pi} r_\lambda(\varphi(t))\varphi'(t)\mathrm{d}t - \frac{1}{\mathbf{i}p} \sum_{j=0}^{p-1} r_\lambda\left(\varphi\left(\frac{2\pi j}{p}\right)\right) \varphi'\left(\frac{2\pi j}{p}\right), \end{aligned} \tag{2.60}$$

cf. equation (2.56). The function $r_\lambda \circ \varphi$ is 2π-periodic, hence we may apply Theorem 2.40 to this function and obtain for the error (2.60)

$$\begin{aligned} \left|E_{T_p}(r_\lambda)\right| \leq \max_{\mathrm{Im}(z)=\sigma_+} |\varphi'(z)|\, |r_\lambda(\varphi(z))|\, F(\exp(-p\sigma_+)) + \\ \max_{\mathrm{Im}(z)=\sigma_-} |\varphi'(z)|\, |r_\lambda(\varphi(z))|\, F(\exp(-p\sigma_-)), \end{aligned} \tag{2.61}$$

for certain $0 > \sigma_- > s_-,\ 0 < \sigma_+ < s_+$. The following lemma substantiates (2.61). It is a special case of [11, Lemma 4.6], where it is stated for functions $(z-\lambda)^{-j}$.

Lemma 2.42 (Beyn, [11, Lemma 4.6])
Let φ be defined on S and 2π-periodic, further let $\varphi(z) \in \mathrm{Int}(\mathcal{C})$ for $\mathrm{Im}(z) > 0$

and $\varphi(z) \in \mathrm{Ext}(\mathcal{C})$ *for* $\mathrm{Im}(z) < 0$. *Let* $\mathrm{dist}(\lambda, \mathcal{C}) = \min_{z\in\mathcal{C}} |\lambda - z|$. *Then there are* $C_1,\ C_2,\ C_3 > 0$ *such that*

$$\left|E_{T_p}(r_\lambda)\right| \leq C_1 \mathrm{dist}(\lambda, \mathcal{C})^{-1} \exp(-C_2\, p\, \mathrm{dist}(\lambda, \mathcal{C})) \tag{2.62}$$

for $\mathrm{dist}(\lambda, \mathcal{C}) \leq C_3$. *The constants are independent of* λ *and* p.

Since the inequality in (2.62) holds with the same constants for all λ, we can estimate this further by [11]

$$C_1 \mathrm{dist}(\lambda, \mathcal{C})^{-1} \exp(-C_2\, p\, \mathrm{dist}(\lambda, \mathcal{C})) \leq C_1 d(\mathcal{C})^{-1} \exp(-C_2\, p\, d(\mathcal{C})),$$
$$d(\mathcal{C}) = \min_{\lambda\in\mathrm{spec}(\mathsf{A},\,\mathsf{B})} \mathrm{dist}(\lambda, \mathcal{C}).$$

Next, recall that we have for the integral that we actually want to compute,

$$\begin{aligned}
\mathsf{U} &= \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{Y}\mathrm{d}z \\
&= \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} \sum_{j=1}^{k} r_{\lambda_j}(z)\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\mathsf{Y}\mathrm{d}z \\
&= \frac{1}{2\pi\mathbf{i}} \sum_{j=1}^{k} \int_{\mathcal{C}} r_{\lambda_j}(z)\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\mathsf{Y}\mathrm{d}z,
\end{aligned} \tag{2.63}$$

where k is the number of eigenvalues inside $\mathcal{C}$. For the integrands in (2.63), let us call them g_j, we have

$$\left\|E_{T_p}(g_j)\right\| = \left|E_{T_p}(r_{\lambda_j}(z))\right| \cdot \left\|\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\mathsf{Y}\right\|.$$

Then, the error estimator from Lemma 2.42 can be applied to every term of (2.63) because $E_{T_p}(\cdot)$ is linear in its argument. This results in

$$\left\|\mathsf{U} - \widetilde{\mathsf{U}}_p\right\| \leq k \cdot C_1 d(\mathcal{C})^{-1} \exp(-C_2\, p\, d(\mathcal{C})) \max_{j=1,\ldots,k} \left\|\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\right\| \|\mathsf{Y}\|.$$

To get the right notion of geometry, we should measure the norms of the error in the computation of U in the B2-norm. To this end, we multiply (2.63) by $\mathsf{B}^{1/2}$ and obtain

$$\mathsf{B}^{1/2}\mathsf{U} = \frac{1}{2\pi\mathbf{i}} \sum_{j=1}^{k} \int_{\mathcal{C}} r_{\lambda_j}(z)\mathsf{B}^{1/2}\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\mathsf{Y}\mathrm{d}z.$$

In this case we obtain, again denoting the integrands g_j,

$$\begin{aligned}
\left\|E_{T_p}(g_j)\right\| &= \left|E_{T_p}(r_{\lambda_j}(z))\right| \cdot \left\|\mathsf{B}^{1/2}\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\mathsf{Y}\right\| \\
&\leq \left|E_{T_p}(r_{\lambda_j}(z))\right| \cdot \left\|\mathsf{B}^{1/2}\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}^{1/2}\right\| \cdot \left\|\mathsf{B}^{1/2}\mathsf{Y}\right\| \\
&= \left|E_{T_p}(r_{\lambda_j}(z))\right| \cdot \left\|\mathsf{B}^{1/2}\mathsf{Y}\right\| \\
&\leq \left|E_{T_p}(r_{\lambda_j}(z))\right| \cdot \left\|\mathsf{B}^{1/2}\right\| \|\mathsf{Y}\|.
\end{aligned}$$

We obtain the following theorem, once again taken from [11]. We adapt it to the linear eigenvalue problem.

Theorem 2.43 (Beyn [11, Thm. 4.7])
Let the curve φ be the parametrization of $\mathcal{C}$ and fulfill the prerequisites of Lemma 2.42. *Then there are constants C_1, C_2 as defined above such that*

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\| \leq k \cdot C_1 d(\mathcal{C})^{-1} \exp(-C_2\, p\, d(\mathcal{C})) \max_{j=1,\ldots,k} \left\| \mathsf{x}_j \mathsf{x}_j^\star \mathsf{B} \right\| \, \|\mathsf{Y}\|$$

with $d(\mathcal{C}) = \min_{\lambda \in \operatorname{spec}(\mathsf{A},\,\mathsf{B})} \operatorname{dist}(\lambda, \mathcal{C})$. If $\varphi(t) = c + r\exp(\mathbf{i}t)$ we have

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\| \leq k \cdot C_1 (\alpha_-^p + \alpha_+^p) \max_{j=1,\ldots,k} \left\| \mathsf{x}_j \mathsf{x}_j^\star \mathsf{B} \right\| \, \|\mathsf{Y}\|$$

with

$$\alpha_- = \max_{\lambda \in \operatorname{spec}(\mathsf{A},\,\mathsf{B}),\, |\lambda - c| < r} \frac{|\lambda - c|}{r}, \quad \alpha_+ = \max_{\lambda \in \operatorname{spec}(\mathsf{A},\,\mathsf{B}),\, |\lambda - c| > r} \frac{r}{|\lambda - c|}.$$

With the same notation, we have that

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\|_{\mathsf{B}2} \leq k \cdot C_1 d(\mathcal{C})^{-1} \exp(-C_2\, p\, d(\mathcal{C})) \left\| \mathsf{B}^{1/2} \right\| \, \|\mathsf{Y}\|$$

and

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\|_{\mathsf{B}2} \leq k \cdot C_1 (\alpha_-^p + \alpha_+^p) \left\| \mathsf{B}^{1/2} \right\| \, \|\mathsf{Y}\| \, .$$

The theorem finally shows that the subspace we are computing converges exponentially to the desired subspace we are seeking for in our integration based eigenvalue solver. The estimates from the theorem then can be plugged into Theorems 2.5 and 2.13 of Argentati and Knyazev [59]. Moreover, the estimates can be used in the perturbation analysis of Rayleigh quotients in Section 2.1.3.

In the adaption of the theorem we made use of the fact that each eigenvalue of a definite matrix pair is a pole of order one of the resolvent, see Remark 2.38. In the original version [11], also the order of the poles at the eigenvalues has to be considered.

Remark 2.44 (Rounding errors)
Rabinowitz [86] emphasizes that the obtained error bounds for the trapezoidal rule are more of theoretical interest. If the error coefficients $F(\exp(-ps)$ become smaller than machine precision, they have no more practical relevance. ◊

Summary

We stated a simple error bound that basically depends on the derivative of the function $h(t) = \varphi'(t)(\varphi(t)\mathsf{B} - \mathsf{A})^{-1}\mathsf{BY}$. Then, we performed a rather technical analysis based on a work of Beyn that shows an exponential decay in the integration error with respect to the number of integration points under rather mild

assumptions. These assumptions can typically be fulfilled. They include that the curve $\mathcal{C}$ has a periodic parametrization, as the unit circle parametrized by the exponential. The other assumption is that the curve $\mathcal{C}$ may not pass through any eigenvalue. This is assumed anyway because if it was the case the integral was not defined. The obtained exponential decay is independent of the derivatives of the function h. The error bounds include the reciprocal of the distance between the curve and the closest eigenvalue.

The convergence of eigenvalues follows immediately. The convergence of eigenvectors is, as often, much more complicated to show and is treated in section 2.1.4.

2.5.3 Error in the integration—Gauß–Legendre

Results similar to the Euler-Maclaurin based error formula (Theorem 2.35) for the trapezoidal rule can be deduced for using Gauß–Legendre integration. The error can also be bounded by some constant multiplied by a derivative of h. Recall Theorem 2.35, declaring the error to be

$$E_{G_{p+1}}(h) = \frac{(2\pi)^{2p+3}((p+1)!)^4}{(2p+3)((2p+2)!)^3} h^{(2p+2)}(\xi), \quad 0 < \xi < 2\pi. \tag{2.64}$$

Note that here, in contrast to the error for the trapezoidal rule, the order of the derivative depends on p. The norm of $h^{(2p+2)}$ can grow fast with p as stated before (cf. Sec. 2.3.4) and hence (2.64) can be a large number even though the constant decays very fast with p. Accordingly, convergence *cannot* be guaranteed if the derivative grows faster than the constant decays with p.

Derivative free Gauß–Legendre error

In the following, we will derive a derivative free error bound for $\widetilde{\mathsf{U}}_p$ when it is computed using the Gauß–Legendre rule. Before we start let us note some facts on the numerical integration of periodic functions using the Gauß–Legendre rule.

Recall that the trapezoidal rule is particularly well suited for periodic functions as we saw above. Weideman [114] gives several examples to illustrate the power of simple trapezoidal rules applied to periodic functions, but he also notes "One should not conclude from this, however, that the midpoint or the trapezoidal rule beat all-comers hands down when the integrand is smooth and periodic. For $f_4(x) = 1/(a - \cos(x))$, with $a = 1 + \varepsilon$ and $0 < \varepsilon \ll 1$, the powerful Gauss–Legendre rule is superior, although this superiority disappears as a increases". Weideman borrowed the example f_4 from Davis [19].

Interestingly, Davis [19] explained in 1958 *why*, and under which conditions, the Gauß–Legendre rule might be superior when applied to periodic functions. For the errors $E_{G_p}(f)$ of the p-point Gauß–Legendre rule and $E_{T_p}(f)$ of the p-point trapezoidal rule he shows

$$\frac{E_{G_{p_k}}(f)}{E_{T_{p_k}}(f)} = \mathcal{O}(\mu^{p_k}) \tag{2.65}$$

for a subsequence $(p_k)_k$ of the positive integers and some value μ with $0 < \mu < 1$.

In other words, under certain conditions there is a value of p such that $\left|E_{G_p}(f)\right| \leq C \left|E_{T_p}(f)\right|$ for some constant $C < 1$. Davis' results are only true for even functions (i. e., functions f with $f(t) = f(-t)$, $t \in \mathbb{R}$). He also mentions a class of periodic functions for which the error of the Gauß–Legendre rule is zero beyond a certain integration order. For the same functions, the error of the trapezoidal rule is nonzero. These functions are based on Bernoulli polynomials, see [20, p. 135].

The following lemma helps to explain under which conditions the Gauß–Legendre rule applied to the integral from the FEAST algorithm converges well. We state it as a special case of the lemma in [19, p. 51], with interval $[0, 2\pi]$, while in the reference it is given for an arbitrary compact interval. It is the central part for the proof of the upper bound of the numerator in (2.65). To prove (2.65), we would also need a lower bound on E_{T_p} for general periodic functions. This is impossible since E_{T_p} can be zero for odd functions [19].

Lemma 2.45 ([19])
Let the (scalar valued) function f be analytic on $[0, 2\pi]$ and continuable analytically throughout the interior of an ellipse whose foci are at 0 and 2π and whose sum of semi-axis is γ. Then, for every $\varepsilon > 0$ there is an integer p_ε such that for all integers $p > p_\varepsilon$ we have

$$\left|E_{G_p}(f)\right| \leq 4\pi \left(\frac{\pi}{\gamma} + \varepsilon\right)^{2p+1}. \tag{2.66}$$

For completeness and in order to motivate the number $2p + 1$ we add the proof.

Proof. For sufficiently large p, we can find a polynomial P_{2p+1} of degree $2p + 1$ such that

$$|f(t) - P_{2p+1}(t)| \leq \left(\frac{\pi}{\gamma} + \varepsilon\right)^{2p+1}, \ 0 \leq t \leq 2\pi.$$

This is due to a result of Bernstein, see [77, p. 194][2]. Next, recall that Gauß–Legendre of order p integrates polynomials of degree $2p + 1$ exactly and that the error E_{G_p} is a linear operator. Let $(t_j, \omega_j)_{j=0,\ldots,p}$ denote the Gauß–Legendre

[2]Davis used the older German translation [76, p. 172].

integration scheme on the interval $[0, 2\pi]$. We then have

$$\begin{aligned}
\left|E_{G_p}(f)\right| &= \left|E_{G_p}(f) - E_{G_p}(P_{2p+1})\right| \\
&= \left|E_{G_p}(f - P_{2p+1})\right| \\
&\leq \int_0^{2\pi} |f(t) - P_{2p+1}(t)| \,\mathrm{d}t + \sum_{j=0}^{p} \omega_j \, |f(t_j) - P_{2p+1}(t_j)| \\
&\leq 2\pi \left(\frac{\pi}{\gamma} + \varepsilon\right)^{2p+1} + 2\pi \left(\frac{\pi}{\gamma} + \varepsilon\right)^{2p+1} \qquad (2.67) \\
&= 4\pi \left(\frac{\pi}{\gamma} + \varepsilon\right)^{2p+1}.
\end{aligned}$$

Inequality (2.67) holds because $\sum_j \omega_j = 2\pi$ and all weights are positive, cf. Remark 2.34.

□

The lemma declares that the convergence of Gauß–Legendre is the faster the larger the region of analyticity of f is. Note that it neither makes use of the derivatives of f nor presumes that f is periodic.

Discussion

Next, let us come back to the eigenvalue problem and the function

$$h(t) = \varphi'(t)(\varphi(t)\mathsf{B} - \mathsf{A})^{-1} \qquad (2.68)$$

that we want to integrate. Let φ once again be defined as

$$\varphi(t) = c + r\exp(\mathbf{i}t).$$

The individual entries of the matrix h fulfill the prerequisites of Lemma 2.45. Since h itself is a matrix valued function, a normwise error bound for the Gauß–Legendre rule applied to h will be more complicated than that in the lemma. The error bound for h will be derived below. The function h, although originally defined on the real numbers, can naturally be extended to $\mathbb{C}$, up to its singularities, which are poles. Those poles are well known, they are the values $z \in \mathbb{C}$ for which $\varphi(z)$ hits an eigenvalue λ of (A, B). The equation

$$\lambda = c + r\exp(\mathbf{i}z)$$

can easily be solved for z, one obtains the values

$$z = \varphi^{-1}(\lambda) = \begin{cases} \mathbf{i}^{-1}\log(\frac{\lambda - c}{r}) & , \lambda > c, \\ -\mathbf{i}^{-1}\log(\frac{r}{c-\lambda}) + \pi & , \lambda < c. \end{cases}$$

If c itself is an eigenvalue, it has no preimage under φ since exp does not map to zero. Of course, those are not the only preimages of λ under φ due to periodicity. Note that $\mathbf{i}^{-1} = -\mathbf{i}$ and that $\mathbf{i}^{-1}\log(\frac{\lambda-c}{r}) + 2\pi$ is also a preimage of λ under φ for $\lambda > c$. Eigenvalues inside the circle defined by c and r are mapped to values $\varphi^{-1}(\lambda)$ with positive imaginary part, those residing outside the circle to values with negative imaginary part.

For $\lambda \to c + r$ or $\lambda \to c - r$, we have $\mathrm{Im}(z) \to 0$. The ellipse of analyticity of h degenerates and the number γ of the lemma decreases to π. We have $\gamma > \pi$ if and only if no eigenvalue of (A, B) is on the boundary of $[c - r, c + r]$. In that case, we would run into trouble anyway since the matrix $(\varphi(t)\mathsf{B} - \mathsf{A})$ then is singular. For every $\varepsilon > 0$ we then have $(\pi/\gamma + \varepsilon) \geq 1$ and the bound (2.66) is meaningless. We hence can expect good convergence of the Gauß–Legendre rule as long as the contour φ is well separated from all eigenvalues of (A, B). This is a result of the same quality as that in Theorem 2.43, where the number $d(\mathcal{C}) = \min_{\lambda \in \mathrm{spec}(\mathsf{A}, \mathsf{B})} \mathrm{dist}(\lambda, \mathcal{C})$ appears explicitly. Similarly, the distance of the interval boundary to the next eigenvalue plays an important role in the approximation point of view, see Section 3.2.3.

Next, we will compute the number γ depending on the locations of the eigenvalues λ of (A, B). Recall, γ is the sum of semi axis a and b of the ellipse. It is determined by the numbers

$$\eta_1 = \min\left\{\left|\mathrm{Im}\left(\varphi^{-1}(\lambda)\right)\right| : \lambda \in \mathrm{spec}(\mathsf{A}, \mathsf{B}),\ \lambda > c\right\},$$
$$\eta_2 = \min\left\{\left|\mathrm{Im}\left(\varphi^{-1}(\lambda)\right)\right| : \lambda \in \mathrm{spec}(\mathsf{A}, \mathsf{B}),\ \lambda < c\right\}.$$

An ellipse of analyticity for h according to Lemma 2.45, including the quantities a, b, η_1, η_2 is shown in Figure 2.3. It has foci 0 and 2π and the values of $\varphi^{-1}(\lambda)$ outside or at the utmost on the boundary. Each of the values η_1, η_2 in fact defines an ellipse with focal points 0 and 2π and semi axis a_1, b_1 and a_2, b_2, respectively. An ellipse of analyticity according to Lemma 2.45 is such that both $0 + \mathbf{i}\eta_1$ and $\pi + \mathbf{i}\eta_2$ are at the exterior of it. It hence can be chosen with semi axis a, b such that

$$a < \min(a_1, a_2), \tag{2.69}$$
$$b < \min(b_1, b_2). \tag{2.70}$$

The numbers a_1, a_2, b_1, b_2 can be computed by means of elementary geometry, see, e. g., [13, pp. 221–222]. For the ellipse defined by η_2 (the height of the ellipse over π) we have $\pi = \sqrt{a_2^2 - b_2^2}$, with $b_2 = \eta_2$, hence $a_2 = {}_+\sqrt{\pi^2 + \eta_2^2}$.

The numbers a_1, b_1 are a little harder to track, we can compute them from the equations

$$\eta_1 = \frac{b_1^2}{a_1}, \tag{2.71}$$
$$1 = \left(\frac{x - \pi}{a_1}\right)^2 + \left(\frac{y}{b_1}\right)^2, \tag{2.72}$$

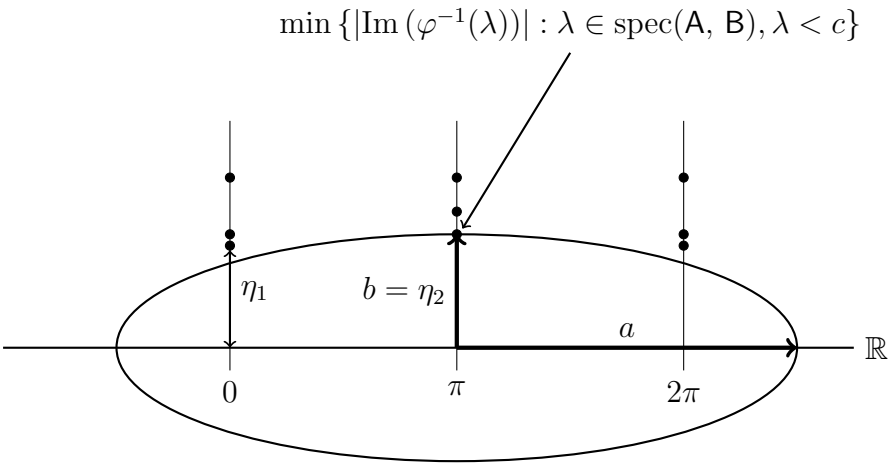

Figure 2.3: Location of the ellipse from Lemma 2.45. The semi axis are denoted a and b and marked by the arrows. The dots on the vertical lines denote the absolute values of $\varphi^{-1}(\lambda)$ for eigenvalues λ.

where now (x, y) denotes any point on the ellipse. Solving (2.71) for a_1 and inserting in (2.72) at $x = 0$ yields the equation of fourth order in b_1,

$$\frac{\pi^2}{(b_1^2/\eta_1)^2} + \frac{\eta_1^2}{b_1^2} = 1.$$

The positive solution of this equation is

$$b_1 = {}_+\sqrt{\frac{\eta_1^2}{2} + \sqrt{\left(\frac{\eta_1^2}{2}\right)^2 + \pi^2\eta_1^2}}.$$

The main semi axis a_1 then can be computed from (2.71).

Matrix valued function

Using Lemma 2.45, let us now derive an error bound for the integral

$$\mathsf{U} = \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{Y}\mathrm{d}z, \tag{2.73}$$

approximated via Gauß–Legendre integration. For simplicity we may suppose that $\mathcal{C}$ is parametrized by a circle $\varphi(t) = c + r\exp(\mathbf{i}t)$. Then, the function (2.68) (multiplied by Y) is to integrate. In the following theorem neither the special nature of $\mathcal{C}$ nor the periodicity of φ is being used. The use of a circle of course makes the computation of the ellipse much simpler, since preimages of the eigenvalues of (A, B) under φ can easily be computed. Subsequently, we assume that $\lambda_1, \ldots, \lambda_k$ are the only eigenvalues of (A, B) that reside in $\mathrm{Int}(\mathcal{C})$, i. e., in the search interval I_λ.

Theorem 2.46 (Error of Gauß–Legendre applied to $(z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}$)
Let (A, B) *be a definite matrix pair and let* $\mathsf{X} = [\mathsf{x}_1, \ldots, \mathsf{x}_n]$ *be its full eigenvector matrix, consisting of* B*-orthogonal eigenvectors. Let* a, b *be chosen according to* (2.69)–(2.70) *such that* $\gamma := a + b > \pi$*, defined by the curve* $\mathcal{C}$ *and the eigenvalues* $\lambda_1, \ldots, \lambda_n$ *of* (A, B)*. Suppose, only the eigenvalues* $\lambda_1, \ldots, \lambda_k$ *reside in* $\mathrm{Int}(\mathcal{C})$*. Then, for every* $\varepsilon > 0$ *there is a number* $p_\varepsilon \in \mathbb{Z}_{\geq 0}$ *such that for all* $p > p_\varepsilon$ *we have*

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\|_2 \leq 2\kappa(\mathsf{X}) \cdot \left(\frac{\pi}{\gamma} + \varepsilon \right)^{2p+1} \cdot \|\mathsf{Y}\|_2 \tag{2.74}$$

and

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\|_{\mathsf{B}2} \leq 2k \cdot \left(\frac{\pi}{\gamma} + \varepsilon \right)^{2p+1} \cdot \|\mathsf{Y}\|_{\mathsf{B}2}\,, \tag{2.75}$$

where $\widetilde{\mathsf{U}}_p$ *denotes the approximation of* U *via the Gauß–Legendre method of order* p*.*

Proof. To prove (2.74) we first write

$$\begin{aligned} h(t) &= \varphi'(t)(\varphi(t)\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{Y} \\ &= \varphi'(t)(\varphi(t)\mathsf{I} - \mathsf{B}^{-1}\mathsf{A})^{-1}\mathsf{Y} \\ &= \varphi'(t)\mathsf{X} \cdot \mathrm{diag}\left(r_{\lambda_1}(\varphi(t)), \ldots, r_{\lambda_n}(\varphi(t))\right) \cdot \mathsf{X}^{-1}\mathsf{Y} \\ &= \mathsf{X} \cdot \mathrm{diag}\left(\varphi'(t) r_{\lambda_1}(\varphi(t)), \ldots, \varphi'(t) r_{\lambda_n}(\varphi(t))\right) \cdot \mathsf{X}^{-1}\mathsf{Y}, \end{aligned}$$

where $r_\lambda(z) = (z-\lambda)^{-1}$. Define $g_j(t) := \varphi'(t) r_{\lambda_j}(\varphi(t))$, $j = 1, \ldots, n$. Then, for every j, the function g_j can analytically be continued to the interior of the ellipse. Hence, the prerequisites of Lemma 2.45 are fulfilled. For every $\varepsilon > 0$ and every j we can find a number $p_j(\varepsilon)$ such that

$$\left| E_{G_{p_j}}(g_j) \right| \leq 4\pi \left(\frac{\pi}{\gamma} + \varepsilon \right)^{2p_j+1},\ p_j \geq p_j(\varepsilon),\ j = 1, \ldots, n.$$

Set $p := \max_j(p_j(\varepsilon))$. Then, for every j we have

$$\left| E_{G_p}(g_j) \right| \leq 4\pi \cdot \left(\frac{\pi}{\gamma} + \varepsilon \right)^{2p+1}. \tag{2.76}$$

It follows (note the factor $1/(2\pi\mathbf{i})$ in the integral (2.73))

$$\begin{aligned}
\left\|\mathsf{U}-\widetilde{\mathsf{U}}_p\right\|_2 &= \left\|\frac{1}{2\pi\mathbf{i}}E_{G_p}(h)\right\|_2 \\
&= \frac{1}{2\pi}\left\|E_{G_p}\big(\varphi'(t)(\varphi(t)\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}\big)\mathsf{Y}\right\|_2 \\
&= \frac{1}{2\pi}\left\|\mathsf{X}\cdot\operatorname{diag}\left(E_{G_p}(\varphi'(t)r_{\lambda_1}(\varphi(t))),\ldots,E_{G_p}(\varphi'(t)r_{\lambda_n}(\varphi(t)))\right)\cdot\mathsf{X}^{-1}\mathsf{Y}\right\|_2 \\
&\le \frac{1}{2\pi}\kappa(\mathsf{X})\left\|\mathsf{Y}\right\|_2\cdot\left\|\operatorname{diag}\left(E_{G_p}(g_1),\ldots,E_{G_p}(g_n)\right)\right\|_2 \\
&= \frac{1}{2\pi}\kappa(\mathsf{X})\left\|\mathsf{Y}\right\|_2\cdot\max_j\left|E_{G_p}(g_j)\right| \\
&\le \frac{1}{2\pi}\cdot 4\pi\cdot\kappa(\mathsf{X})\cdot\left(\frac{\pi}{\gamma}+\varepsilon\right)^{2p+1}\cdot\left\|\mathsf{Y}\right\|_2 \\
&= 2\kappa(\mathsf{X})\cdot\left(\frac{\pi}{\gamma}+\varepsilon\right)^{2p+1}\cdot\left\|\mathsf{Y}\right\|_2,
\end{aligned}$$

where the last inequality is due to (2.76).

In order to prove the other inequality (2.75) we use the eigenvector expansion

$$(z\mathsf{B}-\mathsf{A})^{-1}\mathsf{B}\mathsf{Y}=\sum_{j=1}^{n}r_{\lambda_j}(z)\mathsf{x}_j\mathsf{x}_j^{\star}\mathsf{B}\mathsf{Y},$$

which induces

$$\begin{aligned}
\mathsf{U} &= \frac{1}{2\pi\mathbf{i}}\sum_{j=1}^{n}\int_{\mathcal{C}}r_{\lambda_j}(z)\mathrm{d}z\,\mathsf{x}_j\mathsf{x}_j^{\star}\mathsf{B}\mathsf{Y} \\
&= \frac{1}{2\pi\mathbf{i}}\sum_{j=1}^{k}\int_{0}^{2\pi}\varphi'(t)r_{\lambda_j}(\varphi(t))\mathrm{d}t\,\mathsf{x}_j\mathsf{x}_j^{\star}\mathsf{B}\mathsf{Y},
\end{aligned}$$

since $\lambda_{k+1},\ldots,\lambda_n\notin\operatorname{Int}(\mathcal{C})$. Consequently, because the error $E_{G_p}(\cdot)$ is a linear

operator, we have with $p = \max_j(p_j(\varepsilon))$

$$\begin{aligned}
\left\|\mathsf{U} - \widetilde{\mathsf{U}}_p\right\|_{\mathsf{B}2} &= \left\|\mathsf{B}^{1/2}(\mathsf{U} - \widetilde{\mathsf{U}}_p)\right\|_2 \\
&= \left\|\mathsf{B}^{1/2} E_{G_p}(h)\right\|_2 \\
&= \frac{1}{2\pi}\left\|\sum_{j=1}^{k}\left[E_{G_p}\left(\varphi'(t) r_{\lambda_j}(\varphi(t))\right)\right]\mathsf{B}^{1/2}\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}\mathsf{Y}\right\|_2 \\
&\leq \frac{1}{2\pi}\sum_{j=1}^{k}\left|E_{G_p}(g_j)\right| \cdot \left\|\mathsf{B}^{1/2}\mathsf{x}_j\mathsf{x}_j^\star\mathsf{B}^{1/2}\mathsf{B}^{1/2}\mathsf{Y}\right\|_2 \\
&\leq \frac{1}{2\pi}\cdot k \cdot 4\pi\left(\frac{\pi}{\gamma}+\varepsilon\right)^{2p+1} \cdot \max_j \left\|\mathsf{B}^{1/2}\mathsf{x}_j\right\|_2^2 \cdot \left\|\mathsf{B}^{1/2}\mathsf{Y}\right\|_2 \\
&= 2 \cdot k \cdot \left(\frac{\pi}{\gamma}+\varepsilon\right)^{2p+1} \cdot \|\mathsf{Y}\|_{\mathsf{B}2}\,.
\end{aligned}$$

The last equality follows by $\left\|\mathsf{B}^{1/2}\mathsf{x}_j\right\|_2 = \|\mathsf{x}_j\|_{\mathsf{B}} = 1$. □

The first inequality shown is more appropriate for the standard equation, since then $\kappa(\mathsf{X}) = 1$ is possible and $\|\mathsf{Y}\|_2 = 1$ if Y is chosen to have orthonormal columns. In the second inequality, Y can be chosen with B-orthonormal columns, yielding $\|\mathsf{Y}\|_{\mathsf{B}2} = 1$. Theorem 2.46 formally shows the convergence of Gauß–Legendre applied to the integral (2.73). We have $\gamma > \pi$ if and only if the contour $\mathcal{C}$ does not hit any eigenvalue. In this case, Theorem 2.46 ensures convergence with, e. g., $\varepsilon = (1 - \gamma/\pi)/2 > 0$. Note that the bounds obtained cannot be used in an algorithm as error indicator, they are far too pessimistic. For instance, for a circle with $r = 1$ and eigenvalues coming as close as 10^{-5} to the boundary on both sides of the interval, we already have $\gamma/\pi = 0.999997$. Furthermore, the important quantities γ and $\kappa(\mathsf{X})$ in the right hand sides of (2.74), (2.75) are typically not known at runtime. At least γ could be estimated from the computed Ritz values. In this discussion one should keep in mind that the normwise error that is shown to tend to zero is not the most important measure for accuracy. It is more important that the computed spaces point in the right direction, i. e., the angles between the computed and exact spaces are small. Of course, both measures are closely connected, see Section 2.1. See also the experiment in Section 3.6.5.

2.5.4 Choice of integration contour

At least for the analysis of the trapezoidal rule, see Section 2.5.2, we need a periodic parametrization φ. For the statements of the convergence of the Gauß–Legendre rule we need a function φ that can globally be defined as the restriction of an analytic function to a real interval. So far, we only used a circle $t \mapsto c + r\exp(\mathbf{i}t)$ as integration contour, where $c = (\underline{\lambda} + \overline{\lambda})/2$ and $r = (\overline{\lambda} -$

$\underline{\lambda})/2$ (or slightly larger), fulfilling these requirements. Another possible periodic parametrization is an ellipse with semi axis α and β,

$$\varphi(t) = \alpha\cos(t) + \mathbf{i}\beta\sin(t),\ t \in [0, 2\pi].$$

Here, $\alpha = (\overline{\lambda} - \underline{\lambda})/2$ (or slightly larger) and the center c is chosen as $(\underline{\lambda} + \overline{\lambda})/2$ again. The second semi axis β is at our disposal.

In all our numerical experiments we did not reach any improvements in the final quality of the results or in terms of runtime by using an ellipse instead of a circle offhand. However, for two reasons they might be used.

The first one affects the convergence rate and the size of the ellipse of analyticity from Theorem 2.46 and is hence only applying if the Gauß–Legendre rule is used. For simplicity, assume the ellipse is centered at zero, then for $0 \leq \lambda \leq \alpha$ and $\beta > \alpha$ the solutions of $\varphi(z) = \lambda$ fulfill[3]

$$z = \varphi^{-1}(\lambda) = \arccos\left(\frac{\alpha\lambda - \sqrt{-\alpha^2\beta^2 + \beta^4 + \beta^2\lambda^2}}{\alpha^2 - \beta^2}\right). \tag{2.77}$$

The arcus cosine of an argument $\notin [-1, 1]$ may be defined by means of the principal branch of the complex logarithm [2, p. 47], $\arccos(z) = -\mathbf{i}\log(z \pm \mathbf{i}\sqrt{z^2 - 1})$. The values of (2.77), where $\lambda \in \operatorname{spec}(\mathsf{A}, \mathsf{B})$, can be used to construct ellipses of analyticity according to Theorem 2.46. If $\beta > \alpha$, the resulting ellipse of analyticity is larger compared to the case where a circular contour was used (in that case $\alpha = \beta$).

The second reason for using ellipses is that the integration points are moved away farther from the real axis if $\beta > \alpha$ compared to the case where a circle is used. This will typically decrease the condition number of the system $(z\mathsf{B} - \mathsf{A})\mathsf{V} = \mathsf{BY}$ that is to solve, leading to better performance of iterative linear solvers. See Section 3.6.1.

In our experiments, we observed that the use of an ellipse with a moderate ratio β/α, e. g., $\beta/\alpha = 2$, sometimes can yield faster convergence in terms of FEAST iterations. Also, doubling β/α will roughly half the condition number of $(z_j\mathsf{B} - \mathsf{A})$, if $z_j = \varphi(t_j)$ for fixed integration points t_j. To reach this effect, the integral

$$\hat{\mathsf{U}} = \mathsf{B}^{-1}\mathsf{A}\frac{1}{2\pi\mathbf{i}}\int_C z^{-1}(z\mathsf{B} - \mathsf{A})^{-1}\mathsf{BY}\mathrm{d}z \tag{2.78}$$

has to be used instead of the usual one. The justification for the use of (2.78) is given in [41] in a slightly different context. In our case, it is not immediately clear why the use of (2.78) is allowed, i. e., why $\hat{\mathsf{U}} = \mathsf{U}$. The following lemma states that it is allowed, indeed. For the generalized case A has to be replaced by $\mathsf{B}^{-1}\mathsf{A}$ (besides some minor modifications in the proof).

[3] ask your computer algebra system

Lemma 2.47
For a Hermitian matrix A *and a contour* $\mathcal{C}$ *we have*

$$\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}}(z\mathsf{I}-\mathsf{A})^{-1}\mathrm{d}z=\frac{1}{2\pi\mathbf{i}}\mathsf{A}\int_{\mathcal{C}}z^{-1}(z\mathsf{I}-\mathsf{A})^{-1}\mathrm{d}z. \tag{2.79}$$

Proof. Suppose, $\mathcal{C}$ contains (after a possible renumbering) the eigenvalues with numbers $1,\ldots,k$ and let A have the eigenvector expansion

$$\mathsf{A}=\sum_{j=1}^{n}\lambda_j\mathsf{x}_j\mathsf{x}_j^\star$$

for orthonormal eigenvectors x_j, $j=1,\ldots,n$. We know that the left hand side of (2.79) equals $\sum_{j=1}^{k}\mathsf{x}_j\mathsf{x}_j^\star$. Let us show that the right hand side also equals this sum. Write

$$\begin{aligned}\frac{1}{2\pi\mathbf{i}}\mathsf{A}\int_{\mathcal{C}}z^{-1}(z\mathsf{I}-\mathsf{A})^{-1}\mathrm{d}z&=\frac{1}{2\pi\mathbf{i}}\sum_{j=1}^{k}\int_{\mathcal{C}}\frac{z^{-1}}{z-\lambda_j}\mathsf{A}\mathsf{x}_j\mathsf{x}_j^\star\mathrm{d}z.\\&=\frac{1}{2\pi\mathbf{i}}\sum_{j=1}^{k}\int_{\mathcal{C}}\frac{z^{-1}}{z-\lambda_j}\mathrm{d}z\mathsf{A}\mathsf{x}_j\mathsf{x}_j^\star.\end{aligned} \tag{2.80}$$

By Cauchy's theorem, the integrals $\frac{1}{2\pi\mathbf{i}}\int_{\mathcal{C}}\frac{z^{-1}}{z-\lambda_j}\mathrm{d}z$ have the value $1/\lambda_j$; hence, for (2.80) it holds

$$\begin{aligned}\frac{1}{2\pi\mathbf{i}}\sum_{j=1}^{k}\int_{\mathcal{C}}\frac{z^{-1}}{z-\lambda_j}\mathrm{d}z\mathsf{A}\mathsf{x}_j\mathsf{x}_j^\star&=\sum_{j=1}^{k}\lambda_j^{-1}\mathsf{A}\mathsf{x}_j\mathsf{x}_j^\star\\&=\sum_{j=1}^{k}\lambda_j^{-1}\lambda_j\mathsf{x}_j\mathsf{x}_j^\star\\&=\sum_{j=1}^{k}\mathsf{x}_j\mathsf{x}_j^\star.\end{aligned}$$

□

Note that when speaking about the convergence of integration rules, the shape of the contour will have no effect on the convergence of the trapezoidal rule, at least in the sense as the convergence was analyzed here. Here, the quantity that has to be considered is the distance from the curve $\mathcal{C}$ to spec(A, B), cf. Theorem 2.43.

2.5.5 Influence of error in linear systems

In the foregoing analysis we assumed that the matrices $z_j\mathsf{B}-\mathsf{A}$ for certain values of $z_j = \varphi(t_j)$ are inverted exactly. In practice however, there is an error in the solution $\mathsf{U}_z := (z\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY}$ (let $\widetilde{\mathsf{U}}_z$ denote the computed counterpart) that depends on the chosen method for the solution and the computer architecture in use. When the matrix pair (A, B) is large and sparse, direct factorization methods are not practical and an iterative process as GMRES [89, 90] is more appropriate. Such methods can solve linear systems to a prescribed accuracy.

We aim at approximating the integral (2.48) by a finite sum

$$\sum_{j=0}^{p} \omega_j(\varphi(t_j)\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY},$$

where the numbers ω_j now also contain all appearing scalars such as the derivative of φ. This sum is effectively approximated by a matrix $\widetilde{\Sigma}$ with the property [20, Ch. 4]

$$\widetilde{\Sigma} = \sum_{j=0}^{p} \omega_j(\varphi(t_j)\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY} + \mathsf{R},$$

where R is some error matrix. We hence obtain for the total error

$$\left\| \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B}-\mathsf{A})^{-1}\mathsf{BY}\mathrm{d}z - \widetilde{\Sigma} \right\| \leq \|\mathsf{E}_p\| + \|\mathsf{R}\|,$$

where E_p denotes the approximation error of the integration scheme in use, see [20, Ch. 4].

A normwise bound for R is available as soon as we can bound the errors $\mathsf{U}_{\varphi(t_j)} - \widetilde{\mathsf{U}}_{\varphi(t_j)}$ in the solution of the linear systems for all j. This amounts to error estimation for the solution of linear systems which, unfortunately, is hardly possible. For the error of the GMRES methods, for instance, no simple bound is known that is not too coarse. Of course, if $\mathsf{Mx} = \mathsf{b}$ is to be solved and $\widetilde{\mathsf{x}}$ is the obtained solution, we have with $\mathsf{r} = \mathsf{b} - \mathsf{M}\widetilde{\mathsf{x}}$ that $\|\mathsf{x}-\widetilde{\mathsf{x}}\| \leq \|\mathsf{M}^{-1}\|\,\|\mathsf{r}\|$. Hence, $\|\mathsf{r}\|$ is proportional to the error, while the constant $\|\mathsf{M}^{-1}\|$ is usually not known. However, estimates for the error norm of GMRES are available [69].

2.6 Conclusion

We presented a general, theoretical framework for the integration based solution of eigenvalue problems belonging to the equation $\mathsf{Ax} = \mathsf{Bx}\lambda$, where (A, B) is a definite matrix pair.

The general algorithm consists of a combination of a Rayleigh–Ritz process with numerical integration. In Section 2.1 we investigated subspace based eigensolvers in some detail. It is the author's impression that most literature on subspace eigensolvers deals only with the standard matrix equation, hence with a single matrix. It was shown that several results concerning the approximation error of eigenvalues, convergence of Ritz vectors and residual bounds can be extended to the generalized equation in a unified way. By using the geometry defined by B for norms and angles instead of the standard one, most of the results translate one-to-one.

The subspace in use is defined via a contour integral. It can be shown easily that in exact arithmetic, the desired subspace indeed is computed. This was discussed in Section 2.4. When integrating numerically, approximation errors are introduced. The error in the numerical integration can be estimated and was shown to decay exponentially for the trapezoidal and Gauß–Legendre rule, respectively, in Section 2.5. The bounds obtained can be used in the results from Section 2.1 to obtain bounds on angles and approximation errors in the eigenvalues.

Let us summarize the errors that occur in the process:

- The errors in the solution of linear systems depend heavily on the method used for the solution of those systems. They can be seen as an error that occurs when forming the sum for numerical integration.

- Errors in numerical integration can be analyzed theoretically. When using a circle (or some other periodic curve) as integration contour and the Gauß–Legendre or trapezoidal rule as integration scheme, error bounds can be derived. The error decays exponentially with the order of the integration scheme, while that obtained for the Gauß–Legendre scheme is typically valid only for high orders.

- Error bounds for the approximation error in the eigenvalues and the angle between the eigenvectors are available. Some of them are formulated in terms of the distance between the exact eigenspace and the computed eigenspace. This number can be computed or at least estimated by means of the error bounds explained before.

Using these error bounds in combination yields, under suitable conditions, convergence of the complete eigenvalue method.

Chapter 3

FEAST eigensolver

Synopsis

After having established the fundamental theoretical properties of integration based subspace solvers, this chapter is devoted to approaching their algorithmic and practical aspects. It was mentioned before that Polizzi's FEAST algorithm [85] is obtained when performing a certain contour integration of the resolvent $G(z)$. It is this algorithm we now want to analyze in detail, while we are going to present some new ideas. Most of our theoretical findings and proposed methods are confirmed by numerical experiments. A test based analysis of Polizzi's FEAST method has been presented in [60] and together with parts of the analysis, we presented some algorithmic improvements.

We start by introducing the basic FEAST algorithm in Section 3.1. In Section 3.2 we explain how eigenvalues inside a given contour can be counted reliably and how this information can be used in the FEAST algorithm. In Section 3.3, the numerical integration process is viewed in a different light, leading to an approximation process. One special approximation process, based on polynomials, is introduced and extensively tested in Section 3.4. Afterwards, in Section 3.5, we introduce a method that transforms the region of integration, leading to much better results in some cases. Several smaller topics are discussed in Section 3.6. Finally, in Section 3.7 we conclude this chapter.

Throughout this chapter, (A, B) will denote a definite matrix pair. The eigenvectors of (A, B) are denoted by $\mathsf{x}_1, \ldots, \mathsf{x}_n$ and are supposed to form a B-orthonormal system. The corresponding eigenvalues are $\lambda_1, \ldots, \lambda_n$, ordered ascendingly. The computed counterpart to a quantity will get a "$\tilde{}$"-symbol on top.

3.1 Basic algorithm

First, let us discuss the main steps of the algorithm. A very high-level pseudocode, similar to that one in [60] is shown in Algorithm 3.1. Again, as in the previous chapter, it is the integral

$$\mathsf{Q} = \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1} \mathsf{B} \mathrm{d}z, \tag{3.1}$$

applied to some matrix $\mathsf{Y} \in \mathbb{C}^{n \times \widetilde{m}}$ that is in our interest. With the matrix U computed in this way we aim to span a subspace that approximates the subspace $\mathcal{X} = \mathrm{span}(\mathsf{X})$ corresponding to the set $\mathrm{spec}(\mathsf{A}, \mathsf{B}) \cap I_\lambda$. A short discussion of the algorithm follows.

Algorithm 3.1 Skeleton of the FEAST algorithm

Input: An interval $I_\lambda = \left[\underline{\lambda}, \overline{\lambda}\right]$ and an estimate $\widetilde{m}$ of the number of eigenvalues in I_λ.

Output: $\hat{m} \leq \widetilde{m}$ eigenpairs with eigenvalue in I_λ.

1: Choose $\mathsf{Y} \in \mathbb{C}^{n \times \widetilde{m}}$ of rank $\widetilde{m}$ and compute

$$\mathsf{U} := \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1} \mathsf{B}\, \mathsf{Y} \mathrm{d}z.$$

2: Form the Rayleigh quotients $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star \mathsf{A} \mathsf{U}$, $\mathsf{B}_\mathsf{U} = \mathsf{U}^\star \mathsf{B} \mathsf{U}$.

3: Solve the size-$\widetilde{m}$ generalized eigenproblem $\mathsf{A}_\mathsf{U} \widetilde{\mathsf{W}} = \mathsf{B}_\mathsf{U} \widetilde{\mathsf{W}} \widetilde{\Lambda}$.

4: Compute the approximate Ritz pairs $(\widetilde{\mathsf{X}} := \mathsf{U} \cdot \widetilde{\mathsf{W}}, \widetilde{\Lambda})$.

5: If convergence is not reached then go to Step 1 with $\mathsf{Y} := \widetilde{\mathsf{X}}$.

Input Besides the trivial input—the matrix pair (A, B)—the interval I_λ and a number $\widetilde{m}$ are required. Being just a single integer, the choice of $\widetilde{m}$ is difficult and crucial for a robust behavior of the algorithm. The calculation of $\widetilde{m}$ is part of Section 3.2. Actually, the matrix Y from line 1 also belongs to the input, while it also can be chosen randomly. We will not comment further on the choice of Y, sometimes we only require it to be (B-)orthonormal. For a discussion, see [60].

Output A number $\hat{m} \leq \min\left\{m, \widetilde{m}\right\}$ of eigenpairs, where m denotes the actual number of eigenpairs in I_λ. The nature of the algorithm renders it impossible to find more than $\widetilde{m}$ pairs.

Line 1 The matrix U is computed via numerical integration, which was treated in Section 2.3 and analyzed in Section 2.5. In Section 3.4 we introduce a different way to compute U, based on approximation rather than on integration. This is an essential part of this chapter.

Line 2 consists only of basic operations, where typically sparse matrix routines have to be employed for the products AU, BU.

Line 3 includes the solution of a full small scale generalized eigenproblem. For this task, any suitable library, e. g., LAPACK [5] can be used.

Line 5 Stopping criteria will be discussed later, see Section 3.6.4.

Algorithm 3.1 can be seen (actually: *is*) nothing but projected subspace iteration with the matrix Q (3.1), while of course only an approximation of Q is being used. This was previously stated in [105, 111].

3.2 Counting eigenvalues and size of search space

In this section, we address the question of choosing the input parameter $\widetilde{m}$ of Algorithm 3.1.

The FEAST algorithm "as is" [85] and also the software FEAST 2.0 [83] need this parameter as input. Our goal was to redesign the algorithm in such a way that it accepts a probably rough overestimation of $\widetilde{m} \geq m$ as input and then decreases this number to a reasonable value. The techniques that arise can also be used as stand-alone methods for counting eigenvalues, we introduced them already in [34]. In [60] we studied the effects that occur when choosing $\widetilde{m}$ too small or too large.

3.2.1 Problems with wrongly chosen $\widetilde{m}$

Let us briefly discuss the cases $\widetilde{m} > m$ and $\widetilde{m} < m$, the discussion appeared previously in [60].

Case $\widetilde{m} > m$. In this case, the number $\widetilde{m}$ is larger than the actual number of eigenvalues in I_λ. Consequently, the matrix $\mathsf{U} = \mathsf{QY}$ does not have full rank (recall that Q has rank m). The matrix $\mathsf{B_U} = \mathsf{U}^\star\mathsf{BU}$ is consequently not positive definite anymore and the small scale eigenequation defined by $(\mathsf{A_U}, \mathsf{B_U})$ is not definite. For the consequences, see Example 1.2.

Case $\widetilde{m} < m$. The space spanned by U cannot contain the complete eigenspace associated with the eigenvalues in I_λ.

The following experiment shows the behavior for different choices of $\widetilde{m}$ and motivates the need for efficient estimators for m.

Experiment 3.1 (from [60])
We consider a size-1059 matrix A = `LAP_CIT_1059` [107] from modeling cross-citations in scientific publications and $\mathsf{B} = \mathsf{I}_n$. In this test, we search for $\widetilde{m} = 1, \ldots, 450$ eigenpairs with eigenvalues in an interval I_λ containing the

$m = 300$ lowest eigenvalues. The maximum number of iterations allowed for FEAST was 20.

The left panel of Figure 3.1 shows the number of iterations necessary for FEAST to calculate all eigenpairs within I_λ with sufficiently small residual $\|\mathsf{A}\widetilde{\mathsf{x}} - \mathsf{B}\widetilde{\mathsf{x}}\widetilde{\lambda}\| \leq \varepsilon \cdot n \cdot \max\{|\underline{\lambda}|, |\overline{\lambda}|\}$, as a function of $\widetilde{m}$. An iteration count of 20 typically implies that either none or not all eigenpairs converged within these 20 iterations. The right panel shows the residual span for all computed eigenpairs with eigenvalues in the interval after the respective number of iterations (20 or fewer, if convergence was reached beforehand). Again, these numbers are given as a function of $\widetilde{m}$. We see that, leaving aside the very small region around the exact eigenspace size, either all or none of the eigenpairs show a sufficiently small residual. While for $\widetilde{m} < m$ no eigenpairs converge and especially the minimum residuals are large, for $\widetilde{m} > m$ also the maximum residuals begin to drop significantly and typically all eigenpairs may converge if only enough iterations are performed. With $\widetilde{m}$ just slightly larger than m, all eigenpairs reach convergence within a few iterations.

For a better understanding of the evolution of the computed eigenspace, we monitored the largest canonical angle $\angle\left(\mathsf{X}^{(i)}, \mathsf{X}_{I_\lambda}\right)$ between the current approximate eigenspace $\mathsf{X}^{(i)}$ obtained from the Rayleigh–Ritz process and the exact eigenspace X_{I_λ}, as well as the angle $\angle\left(\mathsf{X}^{(i)}, \mathsf{X}^{(i-1)}\right)$ between the current and the previous iterate. Figure 3.2 provides these angles for three values of $\widetilde{m}$, $\widetilde{m} = 250$, $\widetilde{m} = 300$ and $\widetilde{m} = 350$. In this last case, after five iterations the computed eigenspace contains the exact one and does not vary anymore; these two facts imply convergence. By contrast, the curves for $\widetilde{m} = 250$ indicate that after more than 20 iterations the computed eigenspace is contained in the exact one. Nevertheless, it keeps varying and is not reaching convergence in a reasonable number of iterations. Interestingly, the worst convergence with respect to the exact eigenspace seems to occur for $\widetilde{m} = 300$. This can intuitively be understood by the fact that two subspaces of the same dimension need to be identical in order to have an angle of zero between each other. ◊

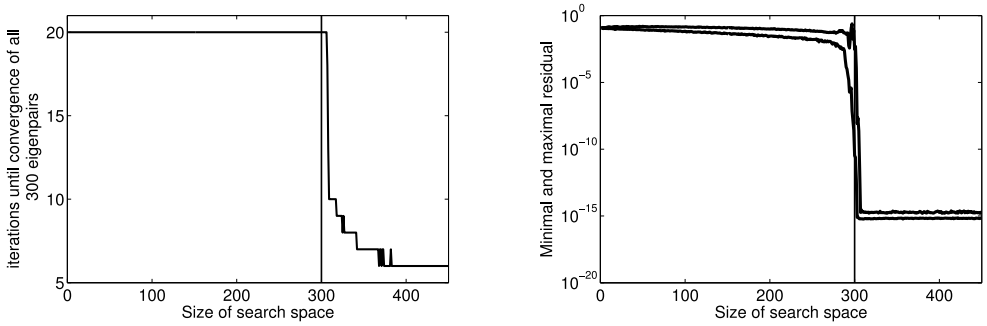

Figure 3.1: Left: Number of necessary iterations. Right: Minimal (lower line) and maximal (upper line) residual.

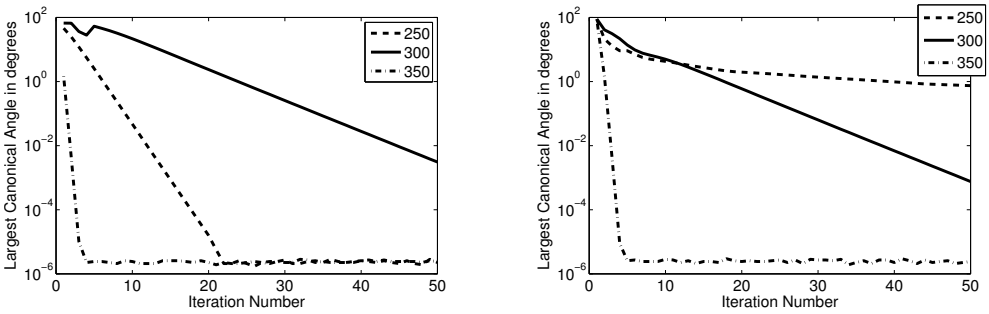

Figure 3.2: Canonical angles (left: between current iterate $\mathsf{X}^{(i)}$ and exact eigenspace X_{I_λ}; right: between current iterate $\mathsf{X}^{(i)}$ and previous iterate $\mathsf{X}^{(i-1)}$) in degrees for $\widetilde{m} = 250, 300, 350$.

The experiment motivates that the number $\widetilde{m}$ by no means should be chosen too *small*. The experiment also motivates that no convergence at all is a good indicator that $\widetilde{m}$ was chosen too small. This was validated by numerous other experiments. In [33] we proposed to increase $\widetilde{m}$ in this case, e. g., by a factor of 2. A number $\widetilde{m} > m$ can be detected by several indicators, as was shown in [33, 34, 60].

The rest of this section is devoted to the accurate determination of $\widetilde{m}$ at runtime.

3.2.2 The selection function

In order to approach the problem of finding an appropriate value for $\widetilde{m}$, we will review the integration process in the FEAST algorithm, not so much from the numerical point of view but rather in a conceptual way. We deduce a function S that will play an important role, where we follow [60, 65].

Recall that we are computing the integral

$$\mathsf{U} = \frac{1}{2\pi \mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1} \mathsf{B}\mathsf{Y} \mathrm{d}z.$$

For the resolvent $G(z) = (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}$ we have the eigenvector expansion

$$G(z) = \sum_{k=1}^{n} r_{\lambda_k}(z) \mathsf{x}_k \mathsf{x}_k^\star \mathsf{B}$$

with the functions $r_{\lambda_k}(z) = (z - \lambda_k)^{-1},\ k = 1, \ldots, n$. If we let

$$\mathsf{X} = [\mathsf{x}_1, \ldots, \mathsf{x}_n]$$

we also have $G(z) = \mathsf{X}\,\mathrm{diag}(r_{\lambda_1}(z), \ldots, r_{\lambda_n}(z))\mathsf{X}^\star\mathsf{B}$. Let for simplicity $\mathcal{C}$ encircle the eigenvalues $\lambda_1, \ldots \lambda_m$ and let I_λ denote a corresponding interval, i. e.,

$\lambda_1, \ldots, \lambda_m \in I_\lambda$, $\lambda_{m+1}, \ldots, \lambda_n \notin I_\lambda$. Integrating the functions r_{λ_k} around $\mathcal{C}$ yields

$$\frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} r_{\lambda_k}(z)\mathrm{d}z = \begin{cases} 1, & 1 \leq k \leq m \\ 0, & \text{otherwise} \end{cases},$$

or for short, the integral only depends on whether λ_k is located inside or outside of $\mathcal{C}$. Integrating $G(z)\mathsf{Y}$ around $\mathcal{C}$ consequently yields

$$\begin{aligned} \mathsf{U} &= \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} G(z)\mathsf{Y}\mathrm{d}z \\ &= \sum_{k=1}^{n} \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} r_{\lambda_k}(z)\mathsf{x}_k\mathsf{x}_k^\star\mathsf{B}\mathsf{Y}\mathrm{d}z \qquad (3.2) \\ &= \sum_{k=1}^{n} \chi_{I_\lambda}(\lambda_k)\mathsf{x}_k\mathsf{x}_k^\star\mathsf{B}\mathsf{Y}. \qquad (3.3) \end{aligned}$$

In (3.3), χ_{I_λ} denotes the function that is 1 inside I_λ and $\chi_{I_\lambda} \equiv 0$ outside I_λ (the *characteristic function* of I_λ). Note that the variable z only appears in the argument of r_{λ_k} in (3.2). For any $\lambda \notin \left\{\underline{\lambda}, \overline{\lambda}\right\}$ we have

$$\frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} r_\lambda(z)\mathrm{d}z = \chi_{I_\lambda}(\lambda), \qquad (3.4)$$

while the left hand side of (3.4) is not defined for $\lambda \in \left\{\underline{\lambda}, \overline{\lambda}\right\}$. Next, let us study the effect of numerical integration applied to $G(z)\mathsf{Y}$. Let $\mathcal{C}$ be parametrized by $\varphi : [0, 2\pi] \longrightarrow \mathbb{C}$ and let $(\omega_j, t_j)_{j=0,\ldots,p}$ denote an integration scheme. We then have

$$\begin{aligned} \mathsf{U} &= \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{Y}\mathrm{d}z \\ &= \frac{1}{2\pi\mathbf{i}} \int_{0}^{2\pi} \varphi'(t)(\varphi(t)\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\,\mathsf{Y}\mathrm{d}t \\ &\approx \frac{1}{2\pi\mathbf{i}} \sum_{j=0}^{p} \omega_j\, \varphi'(t_j)(\varphi(t_j)\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\;\mathsf{Y} \\ &= \sum_{j=0}^{p} \omega_j'\, (z_j\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\,\mathsf{Y} \qquad (3.5) \end{aligned}$$

with $z_j = \varphi(t_j)$ and $\omega_j' = \frac{1}{2\pi\mathbf{i}}\varphi'(t_j)\omega_j$. Plugging in the eigenvector expansion of $G(z)$ again, (3.5) becomes

$$\mathsf{U} \approx \sum_{k=1}^{n} \left[\sum_{j=0}^{p} \omega_j' r_{\lambda_k}(z_j)\right] \mathsf{x}_k\mathsf{x}_k^\star\mathsf{B}\,\mathsf{Y}. \qquad (3.6)$$

Comparing (3.3) with (3.6) shows that the integration scheme $(\omega_j, t_j)_j$ is exact if

$$\sum_{j=0}^{p} \omega_j' r_{\lambda_k}(z_j) = \chi_{I_\lambda}(\lambda_k) \tag{3.7}$$

for all λ_k. In other words, the problem of integrating the resolvent exactly is equivalent to solving the *approximation problem* (3.7) exactly.

Laux [65] interprets the sum $\sum_{j=0}^{p} \omega_j' r_{\lambda_k}(z_j)$ as a function of λ_k. Dropping the subscript k, we obtain a function of λ,

$$S(\lambda) = \sum_{j=0}^{p} \omega_j' r_\lambda(z_j) \approx \frac{1}{2\pi\mathbf{i}} \int_{\mathcal{C}} r_\lambda(z)\mathrm{d}z = \chi_{I_\lambda}(\lambda),\ \lambda \notin \left\{\underline{\lambda}, \overline{\lambda}\right\}. \tag{3.8}$$

This function is called *selection function* by Laux; we also will use this term. Note that S only depends on the curve $\mathcal{C}$ (via the z_j) and the integration scheme (via the ω_j'). The function S is a continuous function on $\mathbb{R}$ as long as we have $z_j \notin \mathbb{R}$ for $j = 0, \ldots, p$. It hence cannot be chosen such that it exactly coincides with χ_{I_λ} in that case, because χ_{I_λ} is not continuous. Later we will relax our definition of the selection function, it will just be some function approximating χ_{I_λ}, and does not necessarily have the form (3.8). In Figure 3.9, selection functions belonging to the trapezoidal and midpoint rule are displayed.

In the following, let us take a closer look on certain function values of S. Let a circle $\mathcal{C}$ be parametrized by $\varphi : [0, 2\pi] \longrightarrow \mathbb{C}$, $\varphi(t) = c + r\exp(\mathbf{i}t)$ with $c \in \mathbb{R}$ and $r > 0$. For simplicity, we may assume $c = 0$ and $r = 1$, all other cases can easily be transformed to this case, see [105]. In this reference, the authors show that $S(-1) = S(1) = 1/2$ and that $S(0) = 1$ if Gauß–Legendre quadrature is used. For the selection function belonging to a general search interval $I_\lambda = \left[\underline{\lambda}, \overline{\lambda}\right]$, we have $S(\underline{\lambda}) = S(\overline{\lambda}) = 1/2$ and $S((\underline{\lambda}+\overline{\lambda})/2) = 1$ if a Gauß–Legendre rule is used. In particular the values on the interval boundaries will play a prominent role in the sequel, see Section 3.2.4 below. For other integration rules, $S(\underline{\lambda}) = S(\overline{\lambda}) = 1/2$ is not true in all cases. The selection function belonging to the trapezoidal rule even has poles in the interval boundaries, see Lemma 3.6 in Section 3.3.1. The midpoint rule for p being even has a continuous selection function with value $1/2$ on the interval boundaries, see Remark 3.7. The function S can attain values slightly below 0, in particular for the Gauß–Legendre rule, see [105].

For the three integration schemes mentioned above, the selection function also fulfills $S(\lambda) \approx 0$ for λ outside I_λ. This fact is quantified for Gauß–Legendre in [105]. For trapezoidal and midpoint rule, see Section 3.3.1. In the rest of this work we suppose that $S(\underline{\lambda}) = S(\overline{\lambda}) = 1/2$, unless otherwise stated.

3.2.3 Convergence rate

In this subsection, we examine the speed of convergence of FEAST, depending on the function S. This will also explain why the convergence is slow if the subspace

size is chosen too small.

It was explained above that the selection function S takes values around 1 inside I_λ, around 0 outside and in some cases 1/2 on the boundary. In the following let us assume that S has value 1/2 on the interval boundaries (in particular it has no poles there). It then can be seen as a function approximating the parameter dependent integral (3.4).

Let (X, Λ) be a full eigen decomposition of (A, B) with B-orthonormal eigenvectors $\mathsf{X} = [\mathsf{x}_1, \ldots, \mathsf{x}_n]$. The following analysis is by Tang and Polizzi [105]. For brevity, let $\hat{\mathsf{A}} = \mathsf{B}^{-1}\mathsf{A}$. We then have $\mathsf{X}^{-1}\hat{\mathsf{A}}\mathsf{X} = \Lambda$, or equivalently $\hat{\mathsf{A}} = \mathsf{X}\Lambda\mathsf{X}^{-1}$. It follows that

$$S(\hat{\mathsf{A}}) = S(\mathsf{X}\Lambda\mathsf{X}^{-1}) = \mathsf{X}S(\Lambda)\mathsf{X}^{-1} = \mathsf{X}S(\Lambda)\mathsf{X}^{\star}\mathsf{B}. \tag{3.9}$$

The function S with a matrix argument is a so called *matrix function*. Such functions are beyond the scope of this work, an introduction can be found e. g., in [36, Ch. 11], [44]. For our purposes it suffices to know that for any rational function of the kind $f(z) = (\alpha - z)^{-1}$ with $\alpha \notin \operatorname{spec}(\mathsf{A})$ we can define $f(\mathsf{A}) := (\alpha\mathsf{I} - \mathsf{A})^{-1}$. Furthermore, for any invertible matrix X of appropriate size we have $f(\mathsf{X}^{-1}\mathsf{A}\mathsf{X}) = \mathsf{X}^{-1}f(\mathsf{A})\mathsf{X}$. The value of a matrix function of a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is defined element-wise as $f(\Lambda) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n))$.

Tang and Polizzi note that the diagonal entries $\gamma_1, \ldots, \gamma_n$ of $\Gamma := S(\Lambda)$ are the eigenvalues of $S(\hat{\mathsf{A}})$ corresponding to eigenvectors x_j, $j = 1, \ldots, n$ (this follows immediately from (3.9), cf. Remark 2.39). We hence may assume

$$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m \geq 1/2 > |\gamma_{m+1}| > \cdots > |\gamma_n|$$

(see [105], see also Section 3.2.4 below). In particular, the eigenvectors of $S(\hat{\mathsf{A}})$ and (A, B) coincide. Note that $S(\hat{\mathsf{A}})$ is nothing but a numerical approximation of the integral (3.1).

The convergence of subspace methods also relies on certain ratios of eigenvalues, see Section 2.1.2. For instance, the convergence of the simple power method [37, Sec. 8.2.1] depends on the ratio $|\lambda_2/\lambda_1|$. In particular, no convergence can be guaranteed if this fraction is 1. There are several generalizations to the convergence of subspaces iteration, see, e. g., [91, Thm. 5.2]. In his thesis [111], Viaud adapted the theorem from [91] to the FEAST algorithm. The key to the use of [91, Thm. 5.2] is that FEAST can be seen as subspace iteration (like presented in Section 2.1.2) with the matrix $S(\hat{\mathsf{A}})$. The adaption is straightforward, it basically requires some renaming. We state it here, without proof, and such that the notation matches ours.

Theorem 3.2 (Convergence rate, Viaud [111, Prop. 2.1])
Let $\mathcal{U}^{(0)}$ be the initial subspace used in the FEAST algorithm, spanned by $\mathsf{Y} = \mathsf{U}^{(0)} = [u_1^{(0)}, \ldots, u_m^{(0)}]$. Let $\mathcal{U}^{(k)}$ be the subspace in iteration k. Let P_k be the orthogonal projector onto that subspace and assume that the set

$$\left\{\mathsf{Q}\mathsf{u}_j^{(0)} : j = 1, \ldots, m\right\}$$

is linearly independent, where Q *denotes the projector* (3.1) *onto the desired subspace. Let* $\hat{\mathsf{Q}} = S(\hat{\mathsf{A}})$ *denote the approximate projector computed (e. g.,) by numerical integration. Then, for each eigenvector* x_j *of* $\hat{\mathsf{Q}}$, $j = 1, \ldots, m$ *with corresponding eigenvalue* γ_j, *there is a unique vector* $\mathsf{y}_j \in \mathsf{U}^{(0)}$ *such that* $\mathsf{Q}\mathsf{y}_j = \mathsf{x}_j$ *and*

$$\|\mathsf{x}_j - \mathsf{P}_k\mathsf{x}_j\| \leq \|\mathsf{x}_j - \mathsf{y}_j\| \left(\left| \frac{\gamma_{m+1}}{\gamma_m} \right| + \delta_k \right)^k \tag{3.10}$$

for a sequence $(\delta_k)_k$ *with* $\lim_{k \longrightarrow \infty} \delta_k = 0$.

Note that in the FEAST algorithm we also allow a larger subspace Y, containing $\widetilde{m} > m$ basis vectors. In this case, the inequality (3.10) remains true while the uniqueness property of the vector y_j can be violated. The theorem basically ensures convergence of the eigenvectors computed by the FEAST algorithm under the following conditions.

- $|\gamma_{m+1}/\gamma_m| < 1$.
- QY has full rank.

If both conditions are fulfilled, the convergence of the projections $\mathsf{P}_k\mathsf{x}_j$ towards x_j depends on the quality of the initial subspace (via the norm $\|\mathsf{x}_j - \mathsf{y}_j\|$ and the absolute value of the fraction γ_{m+1}/γ_m).

The theorem gives a bound on the normwise error of computed eigenvectors, a measure that we did not use before. However, it provides a good understanding under which conditions fast convergence can be expected. The function S should be able to divide the wanted from all other (unwanted) eigenvalues in a relative sense, making the ratio $|\gamma_{m+1}/\gamma_m|$ as small as possible. This is also the task of approximation techniques, see Section 3.3. In [111] a method is proposed that indeed minimizes this ratio. In [105] a similar bound as (3.10) can be found, specially tailored for the generalized eigenvalue problem.

Let us come back to Experiment 3.1. There, it turned out that the search space should have a higher dimension than the number of eigenvalues in the interval. If the search space dimension is smaller or equal to the number of eigenvalues inside I_λ, all eigenvalues λ are mapped to a value close to 1 by the function S. The ratio in (3.10) approaches 1. This explanation was given by Tang and Polizzi [105].

The ratio $|\gamma_{m+1}/\gamma_m|$ in (3.10) depends on the eigenvalues of the matrix pair (A, B). Viaud [111] notes that it has

$$\frac{\max_{\lambda \notin I_\lambda} |S(\lambda)|}{\min_{\lambda \in I_\lambda} |S(\lambda)|}$$

as an upper bound, an expression that does not depend on single eigenvalues. He then refines this expression to

$$\hat{u} = \frac{\max_{\lambda \in (-\infty, l_0^-) \cup (l_0^+, +\infty)} |S(\lambda)|}{\min_{\lambda \in (l_1^-, l_1^+)} |S(\lambda)|}, \tag{3.11}$$

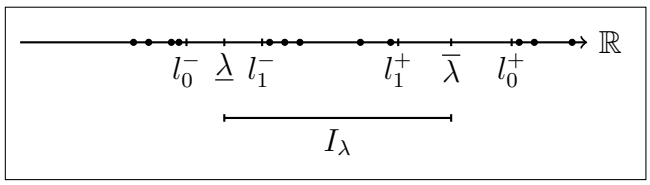

Figure 3.3: Illustration of the situation from (3.11). The dots represent the eigenvalues.

where $l_0^- < \underline{\lambda} < l_1^- < l_1^+ < \overline{\lambda} < l_0^+$ and no eigenvalue of (A, B) resides in (l_0^-, l_1^-) and (l_1^+, l_0^+). Actually, max and min should of course be replaced by sup and inf. The situation is illustrated in Figure 3.3. This refinement is necessary because the more simple expression is meaningless in the case that S is continuous at $\overline{\lambda}$ or $\underline{\lambda}$ [111].

The selection function S is independent of the eigenvalues of (A, B), hence it should be such that the interval (l_1^-, l_1^+) is as large as possible with the property $S \approx 1$ inside. Similarly, l_0^- and l_0^+ should be as close as possible to $\underline{\lambda}$ and $\overline{\lambda}$ respectively, with the property $S \approx 0$ on $(-\infty, l_0^-) \cup (l_0^+, +\infty)$. Speaking simply, the function S should be as "steep" as possible around $\underline{\lambda}$, $\overline{\lambda}$. The selection functions corresponding to two different integration schemes are plotted in Figure 3.9 on page 113. From these plots it can be realized that the intervals (l_0^-, l_1^-), (l_1^+, l_0^+) are not very small at least in the case of the midpoint rule.

3.2.4 Eigenvalues of B_U

Let us come to the use of the selection function. In [34], we evaluated several methods for counting eigenvalues in I_λ. It turned out that one of the most reliable ones was measuring the rank of U, which coincides, in exact arithmetic, with m. It is a well-known fact that the number of nonzero singular values of a matrix coincides with its rank, again in exact arithmetic. In a computational setting, one cannot just count nonzero singular values but has rather to define a tolerance $\delta > 0$ and to count only such singular values $> \delta$, see [36, Sec. 5.5.8]. The number δ must be chosen depending on machine precision and the uncertainty in the data. For our particular problem, it was shown in [34] that $\delta = 0.5$ is the correct choice, this was confirmed numerically. Being more precise, the number of singular values $> \delta = 0.5$ is the number of eigenvalues inside I_λ (actually, we had to count those $\geq \delta$).

Having a look at (3.5) and comparing that equation to the selection function

$$S(\lambda) = \sum_{j=0}^{p} \omega_j' r_\lambda(z_j), \tag{3.12}$$

we see that U is effectively approximated by the matrix function $S(\mathsf{B}^{-1}\mathsf{A})\mathsf{Y}$.

Applying the selection function S to $\mathsf{B}^{-1}\mathsf{A}$ yields (note that the argument of (3.12) is λ)

$$\begin{aligned} S(\mathsf{B}^{-1}\mathsf{A})\mathsf{Y} &= \sum_{j=0}^{p} \omega_j'(z_j\mathsf{I} - \mathsf{B}^{-1}\mathsf{A})^{-1}\mathsf{Y} \\ &= \sum_{j=0}^{p} \omega_j'(z_j\mathsf{B} - \mathsf{A})^{-1}\mathsf{B}\mathsf{Y}, \end{aligned}$$

which coincides with (3.5).

Coming back to eigensystems, let $(\mathsf{X}, \mathsf{\Lambda})$ be a full eigen decomposition of (A, B) with B-orthonormal eigenvectors $\mathsf{X} = [\mathsf{x}_1, \ldots, \mathsf{x}_n]$. Next, let us (again, see Section 3.2.3) follow Tang and Polizzi [105]. For brevity, let $\hat{\mathsf{A}} = \mathsf{B}^{-1}\mathsf{A}$, we then have $\mathsf{X}^{-1}\hat{\mathsf{A}}\mathsf{X} = \mathsf{\Lambda}$, or equivalently $\hat{\mathsf{A}} = \mathsf{X}\mathsf{\Lambda}\mathsf{X}^{-1}$. We already saw that

$$S(\hat{\mathsf{A}}) = S(\mathsf{X}\mathsf{\Lambda}\mathsf{X}^{-1}) = \mathsf{X}S(\mathsf{\Lambda})\mathsf{X}^{-1} = \mathsf{X}S(\mathsf{\Lambda})\mathsf{X}^{\star}\mathsf{B}.$$

Let $\gamma_1, \ldots, \gamma_n$ denote the diagonal entries of $\mathsf{\Gamma} := S(\mathsf{\Lambda})$. We see that

$$S(\hat{\mathsf{A}})\mathsf{x}_j = \mathsf{x}_j\gamma_j,\ j = 1, \ldots, n.$$

For short, the operator $S(\hat{\mathsf{A}})$ has the same eigenvectors as (A, B), cf. Remark 2.39.

Building the Rayleigh quotient B_U from $\mathsf{U} = S(\hat{\mathsf{A}})\mathsf{Y}$ yields

$$\begin{aligned} \mathsf{B}_\mathsf{U} &= \mathsf{U}^{\star}\mathsf{B}\mathsf{U} \\ &= \mathsf{Y}^{\star}S^{\star}(\hat{\mathsf{A}})\mathsf{B}S(\hat{\mathsf{A}})\mathsf{Y} \\ &= \mathsf{Y}^{\star}\mathsf{B}\mathsf{X}S^{\star}(\mathsf{\Lambda})\underbrace{\mathsf{X}^{\star}\mathsf{B}\mathsf{X}}_{=\mathsf{I}}S(\mathsf{\Lambda})\mathsf{X}^{\star}\mathsf{B}\mathsf{Y} \\ &= \mathsf{Y}^{\star}\mathsf{B}\mathsf{X}S^{\star}(\mathsf{\Lambda})S(\mathsf{\Lambda})\mathsf{X}^{\star}\mathsf{B}\mathsf{Y}. \end{aligned}$$

Suppose, Y was chosen B-orthonormal. We then have

$$(\mathsf{X}^{\star}\mathsf{B}\mathsf{Y})^{\star}(\mathsf{X}^{\star}\mathsf{B}\mathsf{Y}) = \mathsf{Y}^{\star}\mathsf{B}\mathsf{X}\mathsf{X}^{\star}\mathsf{B}\mathsf{Y} = \mathsf{Y}^{\star}\mathsf{B}\mathsf{Y} = \mathsf{I},$$

since $\mathsf{X}\mathsf{X}^{\star}\mathsf{B} = \mathsf{I}$. In other words, the matrix $\mathsf{X}^{\star}\mathsf{B}\mathsf{Y}$ has orthonormal columns. We see that the matrix B_U is also a Rayleigh quotient of the matrix $S^{\star}(\mathsf{\Lambda})S(\mathsf{\Lambda})$ belonging to the orthonormal basis $\mathsf{X}^{\star}\mathsf{B}\mathsf{Y}$. We have $S^{\star}(\mathsf{\Lambda})S(\mathsf{\Lambda}) = S^2(\mathsf{\Lambda})$, since the entries of $S(\mathsf{\Lambda})$ are real. It follows that B_U has eigenvalues $S(\lambda_j)^2,\ j = 1, \ldots, n$, if Y was chosen square and unitary.

Now it is intuitively understood that the eigenvalues of B_U approach the numbers $S(\lambda_j)^2$ as $\widetilde{m}$ approaches n, see Chapter 2. As a practical implication, we can count the eigenvalues of B_U which are greater or equal than $1/4 (= (1/2)^2)$ and take this number as an estimate for m. From Rayleigh–Ritz theory it is clear that the quality of this estimation is improved if more components of the eigenvectors belonging to I_λ are represented in U. These components are typically

amplified over the FEAST iterations, hence the estimation of m is improved. Tang and Polizzi [105] also give a quantitative analysis of how close the eigenvalues of B_U get to the numbers γ_j^2. The occurring bounds are not stated in terms of computable quantities that can be monitored since the eigenvectors of (A, B) are involved. Hence, they finally also give the advice to count eigenvalues greater or equal than $1/4$. In formulas, a number q is being computed with

$$q = \left|\left\{\gamma^2 \in \operatorname{spec}(\mathsf{B}_\mathsf{U}) : \gamma^2 \geq 1/4\right\}\right|. \tag{3.13}$$

Note that in our publication [34] we were talking about singular values of B_U in order to get a unified representation, see below. Obviously, singular values and eigenvalues of B_U coincide, since it is a positive (semi) definite, Hermitian matrix.

Standard case

If $\mathsf{B} = \mathsf{I}$, we have $\mathsf{B}_\mathsf{U} = \mathsf{U}^\star\mathsf{U}$. Letting $\mathsf{U} = \mathsf{V}\Sigma\mathsf{W}^\star$ be the thin SVD of U, we see that $\mathsf{B}_\mathsf{U} = \mathsf{W}\Sigma^2\mathsf{W}^\star$, i. e., the eigenvalues of B_U are the squared singular values of U. Neglecting the large amount of work necessary—U has n rows—we can compute the SVD of U and perform the same analysis as previously presented with the tolerance $1/2$ instead of $1/4$. The advantages are that the problem of computing the SVD of U is better conditioned then that one of $\mathsf{U}^\star\mathsf{U}$ and that columns of U belonging to relevant eigenvalues can easily be extracted.

Costs

The additional cost for determining q is the computation of the full spectrum of B_U. This requires $4\widetilde{m}^3/3$ operations for performing a tridiagonal factorization of B_U and lower order additional cost for computing the eigenvalues, see [36].

3.2.5 Efficient computation of a basis for the search space

Suppose $\mathsf{U} \in \mathbb{C}^{n\times\widetilde{m}}$ has been computed and a certain number $q \leq \widetilde{m}$ that estimates the actual number m of eigenvalues in I_λ has been calculated based on the rule (3.13). The most simple way to proceed would be to set $\widetilde{m} = q$ and restart Algorithm 3.1. This would cause a loss of information that has already been obtained. The method of choice is hence to extract useful columns from U and proceed only with those. In the standard case $\mathsf{B} = \mathsf{I}$ this can be done as described before, by computing the SVD $\mathsf{U} = \mathsf{V}\Sigma\mathsf{W}^\star$ and taking $\mathsf{U} := \mathsf{V}(:, 1:q)$. Here, it is sufficient to take the first q columns of V because the singular values of U, i. e., the diagonal entries of Σ, are ordered descendingly.

Fortunately, there is a way to avoid the SVD of the large matrix U. Let $\mathsf{B}_\mathsf{U} = \mathsf{U}^\star\mathsf{B}\mathsf{U} = \mathsf{V}\Sigma\mathsf{V}^\star$ be the singular value decomposition of B_U. The difference to an eigenvalue decomposition is that the values on the diagonal of Σ are sorted descendingly. Suppose the first q entries of Σ are $\geq 1/4$. Then, the columns

$1, \ldots, q$ of V belong to those singular values. The singular values are the entries of the $q \times q$-matrix

$$\Sigma(1:q, 1:q) = \mathsf{V}(:, 1:q)^\star \mathsf{B}_\mathsf{U} \mathsf{V}(:, 1:q) = \mathsf{V}(:, 1:q)^\star \mathsf{U}^\star \mathsf{B} \mathsf{U} \mathsf{V}(:, 1:q).$$

This shows that $\mathsf{U}_q := \mathsf{U} \cdot \mathsf{V}(:, 1:q)$ is a matrix with Rayleigh quotient $\mathsf{U}_q^\star \mathsf{B} \mathsf{U}_q$ that has all singular values $\geq 1/4$. Hence it is also a suitable matrix to transform (A, B) with. Before continuing the algorithm, one simply sets $\mathsf{U} := \mathsf{U}_q$ and forms the Rayleigh quotients with this matrix.

In order to avoid the effects that occur with a search space chosen too small, the number q should be replaced by some number slightly larger than the actual eigenvalue count obtained by any of the methods above. These effects were explained in Section 3.2.1. The choice $\tilde{q} = \min\{\alpha q, \widetilde{m}, n\}$ is reasonable for some $\alpha > 1$. In [85], $\alpha = 1.5$ is suggested, while often $\alpha = 1.1$ is enough according to our experience.

3.2.6 Preprocessing of FEAST

In [34] a technique for detecting "empty" intervals was proposed, i. e., intervals I_λ with $I_\lambda \cap \operatorname{spec}(\mathsf{A}, \mathsf{B}) = \emptyset$.

The previous analysis showed that if there are any eigenvalues of (A, B) in I_λ, the number of those is approximated by the number of eigenvalues of B_U greater or equal than $1/4$. Consequently, if there is no eigenvalue in I_λ at all, all eigenvalues of B_U are smaller than $1/4$. As a preprocessing step of the FEAST algorithm we can hence run one or two iterations with small search space size and check if any eigenvalue of B_U is $> 1/4$. If not, the interval can be neglected in the computation. Otherwise, the algorithm has to be restarted with a reasonably large size of search space $\widetilde{m}$. For this preprocessing step, $\widetilde{m} = 3$ proved to be sufficient, see [34].

If the algorithm was started for the full computation with reasonably large $\widetilde{m}$, in the second iteration the first estimation of the number of eigenvalues inside I_λ is performed via (3.13). If this check delivers $q = \widetilde{m}$, there might be $m > \widetilde{m}$ eigenvalues inside I_λ that cannot all be computed in that case. The algorithm should be restarted with a larger number $\widetilde{m}$, or a smaller interval has to be chosen. For details, see [33, 34].

3.2.7 Alternatives and further discussion

So far we presented a method for determining an estimate q for the number of eigenvalues in the search interval I_λ. This number can be used to set the size $\widetilde{m} := q$ of the search space used in the FEAST algorithm. The use of q is well understood theoretically, see the discussion above and [105]. In the case of the standard eigenvalue equation, also the singular values of U can be used for computing q.

Alternatives involving $\mathsf{B_U}$

Let us briefly discuss some alternatives to monitoring the singular values of $\mathsf{B_U}$ or U, respectively. Some of these alternatives have previously been presented in [34].

Counting Ritz values. A very simple but inaccurate method consists of just counting the Ritz values, i. e., those diagonal entries of $\widetilde{\mathsf{\Lambda}}$ in Line 4 of Algorithm 3.1 lying in I_λ. This technique just requires $\widetilde{m}$ comparisons.

Rank revealing QR. Another method that can in general be used to compute the rank of a matrix, is the so called rank revealing QR decomposition (rrQR), see [36, Sec. 5.4], or, e. g., [40] for more details. This decomposition consists of a factorization $\mathsf{X\Pi = QR}$, where $\mathsf{X} \in \{\mathsf{U}, \mathsf{B_U}\}$. The matrix $\mathsf{\Pi}$ is a suitable permutation such that Q is orthonormal and R is an upper triagonal matrix whose diagonal entries $r_{jj} = \mathsf{R}(j, j)$ are ordered descendingly according to their absolute value. Then the rank of X can be estimated by counting those values r_{jj} that are larger than some threshold that has to be supplied.

It is difficult to make a connection between the individual values of r_{jj} and the singular values of X. However, in the context of the FEAST algorithm, the matrix U is expected to have columns with a good level of orthogonality after a few iterations, meaning the scalar products $\mathsf{U}(:, i)^\star \mathsf{BU}(:, j)$, $j \neq i$, are small. The square matrix $\mathsf{B_U}$ is hence a diagonal matrix (or at least a matrix with small off-diagonal entries). When forming an rrQR of that matrix, $\mathsf{B_U\Pi = QR}$, we see that R is a diagonal matrix scaling Q's columns, Q is essentially the identity and hence $\mathsf{QR\Pi}^\star$ is an approximate SVD of $\mathsf{B_U}$. In the case of the standard equation, $\mathsf{B = I}$, an rrQR $\mathsf{U\Pi = QR}$ consists similarly of the matrix Q with reordered, normalized columns of U and of an upper triangular matrix R containing the corresponding scaling factors.

All these facts lead to the observation that counting singular values of $\mathsf{B_U}$ (U in the standard case, respectively) larger than $1/2$ or $1/4$, respectively, can be replaced by counting the corresponding diagonal entries of R. The reason for using this technique is that the costs are slightly lower than those for computing eigenvalues of $\mathsf{B_U}$, resulting in $3\widetilde{m}^2 r - 4r^2\widetilde{m} + 4r^3/3$ operations for the rrQR of $\mathsf{B_U}$ with rank r [36].

An alternative way to compute a QR decomposition of U in the case $\mathsf{B = I}$ is the so called CholQR method [97]. Let R now be the upper triangular Cholesky factor of $\mathsf{B_U}$, i. e., $\mathsf{B_U = R^\star R}$ and Q the orthonormal factor of a QR decomposition of U. Then Gander [35] states that

$$\mathsf{B_U = U^\star U = R^\star R = R^\star Q^\star QR}.$$

It follows that $\mathsf{U = QR}$ is a QR decomposition of U. Further, Q can be computed as $\mathsf{Q = UR}^{-1}$. This way for computing the QR factorization is not recommended

[35] due to numerical instability. The Cholesky algorithm becomes more and more unstable as the condition number of B_U grows (see, e. g., [97, 108]). The matrix B_U is expected to be singular in exact arithmetic, in practice it is typically non-singular such that the Cholesky decomposition exists. Besides all shortcomings, the CholQR method might suffice for our purposes and we expect B_U to have small condition due to the orthogonality of U, see Experiment 3.5.

Frobenius norm. In exact arithmetic and when no eigenvalues are on the boundary of I_λ, the matrices U and B_U only have singular values 0 and 1. Observing that for any matrix X with rank r and singular values σ_j we have $\|\mathsf{X}\|_F = (\sum_{j=1}^{r} \sigma_j)^{1/2}$, we can deduce $\|\mathsf{B}_\mathsf{U}\|_F = \sqrt{r}$. This gives at least a rough approximation after the first FEAST iteration and costs about $2\widetilde{m}^2$ operations for $\|\mathsf{B}_\mathsf{U}\|_F$.

If eigenvalues are on the boundary of I_λ, the matrix B_U has singular values $1/4$ and hence we cannot extrapolate the number of eigenvalues inside I_λ from a certain value of $\|\mathsf{B}_\mathsf{U}\|_F$.

Other techniques

Several techniques for counting eigenvalues in a certain domain are available that do not make use of techniques similar to those presented above, see, e. g., [80, Ch. 3].

That one coming closest to the techniques presented here was presented by Philippe and co-workers in [10], and recently [54]. It also makes use of contour integration and the fact that

$$N_\mathcal{C} = \frac{1}{2\pi\mathbf{i}} \int_\mathcal{C} \frac{f'(z)}{f(z)} \mathrm{d}z$$

is the number of zeros of the analytic function f in $\mathrm{Int}(\mathcal{C})$. Now setting $f(z) = \det(z\mathsf{B} - \mathsf{A})$ and computing the integral yields the number of eigenvalues of (A, B) inside $\mathcal{C}$. For computing $\det(z\mathsf{B} - \mathsf{A})$ (in the references above only $\mathsf{B} = \mathsf{I}$ is considered), the matrix $z\mathsf{B} - \mathsf{A}$ has to be factorized in its LU-factors for different points z. This also results in a considerably large amount of work. However, the method is tailored for general (i. e., non-Hermitian) eigenvalue problems.

Recently, a report [24] was published, investigating different approximations of the selection function for counting eigenvalues. One of those is the approximation by polynomials, a topic we treat in Section 3.4 extensively. The difference is that we use the approach to actually solve the eigenproblem, while in [24] it is only used for counting eigenvalues (which we also did in our experiments in Section 3.4). In [24], the authors also suggest to widen the interval $I_\lambda = \left[\underline{\lambda}, \overline{\lambda}\right]$ by (e. g.) $(\overline{\lambda} - \underline{\lambda})/4$ at both boundaries. We proposed a similar technique in [34] for also solving the eigenvalue problem. The simultaneous use of two intervals, a larger

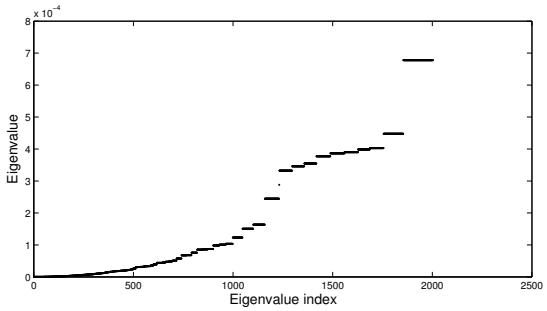

Figure 3.4: Eigenvalues of T_{2003}.

one for counting eigenvalues and a smaller one for actually searching eigenvalues is also thinkable, see [34].

3.2.8 Numerical experiments

In [34], different methods for counting eigenvalues were assessed. It turned out that the method based on the eigenvalues of B_U (singular values of U, resp.) from Section 3.2.4 is superior compared to the others. In particular, it is most robust when eigenvalues are located at the boundary of I_λ. Let us repeat parts of our experiments from [34] here, complemented by some others.

The setting of our experiments is as follows. The interval I_λ is moved over a part of the spectrum. This in fact yields a sequence

$$I_\lambda^{(1)} = \left[\underline{\lambda}_1, \overline{\lambda}_1\right], \ldots, I_\lambda^{(k)} = \left[\underline{\lambda}_k, \overline{\lambda}_k\right]$$

of different intervals, each possibly containing different eigenvalues. We will usually have a constant interval length $\overline{\lambda}_j - \underline{\lambda}_j$ for all j and a constant stepsize $\underline{\lambda}_j - \underline{\lambda}_{j-1}$ for $j = 2, \ldots, k$. Such a sequence of intervals was called *interval progression* in [34]. Then, at least two FEAST iterations are performed; in the second iteration the eigenvalue count is performed. In the first iteration, usually the eigenspace did not converge enough to deliver reliable results on eigenvalue counts. We tested the different methods on a matrix $\mathsf{A} := \mathsf{T}_{2003}$, which we chose due to its quite challenging spectrum. The eigenvalues are numbered ascendingly, $\lambda_1 < \lambda_2 < \ldots < \lambda_{2003}$. In particular it has eigenvalues of small absolute value ($\lambda_1 = \mathcal{O}(10^{-10})$) while the complete spectrum ranges up to 10^{-3}. Further, some eigenvalues are very close to each other both in relative and absolute sense. To get an understanding of A's spectrum, it is depicted in Fig. 3.4.

Experiment 3.3
To highlight the possible shortcomings of the Frobenius and Ritz methods from Section 3.2.7, we first repeat our Experiment 2.3 from [34]. Therein, an interval progression was performed, including an arbitrary starting point of the intervals, a fixed interval length and a fixed stepsize. All values were chosen arbitrarily.

In Figure 3.5, the interval progression is displayed, the results for the eigenvalue count can be seen in Figure 3.6. The SVD estimator was performed on B_U. The results for rrQR were exactly the same as those for the SVD, just as expected. For this reason they are not plotted. It can be seen that the SVD (and therefore the rrQR) estimator always delivers correct results. The Ritz count estimator sometimes overestimates, and the Frobenius norm estimator shows a behavior that cannot be predicted. The underestimations are from eigenvalues on the boundary of I_λ, while the overestimations might come from numerical inaccuracy, e. g., if some of the singular values of B_U are slightly larger than 1. These errors might sum up such that a too large number of eigenvalues is counted. This

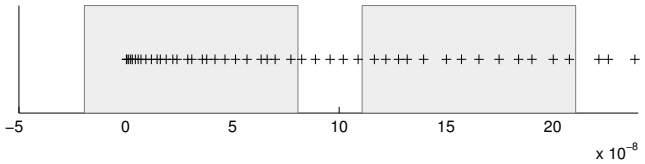

Figure 3.5: Interval progression for Experiment 3.3. A gray shaded area stands for one interval of the progression. The plot has previously been published in [34].

behavior is typical and could be observed in numerous other experiments, while with this choice of matrix the shortcomings of the Frobenius count might be particularly bad. ◊

The next experiment is stressing the estimators a little more.

Experiment 3.4
In this experiment, we apply the methods to the same matrix as in the previous Experiment 3.3, while the interval is chosen such that its boundary hits a cluster of eigenvalues. We let $\underline{\lambda} = \lambda_{1704}$, $\overline{\lambda} = \lambda_{2003}$, hence I_λ is containing exactly 300 eigenvalues. We choose $\widetilde{m} = 330$. The eigenvalues $1680, \ldots, 2003$ are shown in Figure 3.7. We see that the eigenvalues appear in three large clusters, one of them is on the upper boundary and one on the lower. The lower cluster contains some more eigenvalues below $\underline{\lambda}$, such that together 317 eigenvalues are in the three clusters. This special eigenvalue structure is quite demanding to eigenvalue algorithms.

We ran the FEAST algorithm and counted the eigenvalues in I_λ with the SVD and rrQR methods. First, we observed the evolution of the singular values of B_U over the iterations in the FEAST algorithm. The singular values of B_U in the first two iterations are shown in the top pictures of Figure 3.8. It can be seen that the singular values in the first iteration are completely meaningless in our context. This is due to the random starting basis. Only few components of such a starting base lie in the direction of the eigenspace. The singular values in the second iteration already show the desired behavior. They look very similar to those in the third iteration, given in the bottom picture of Figure 3.8, together

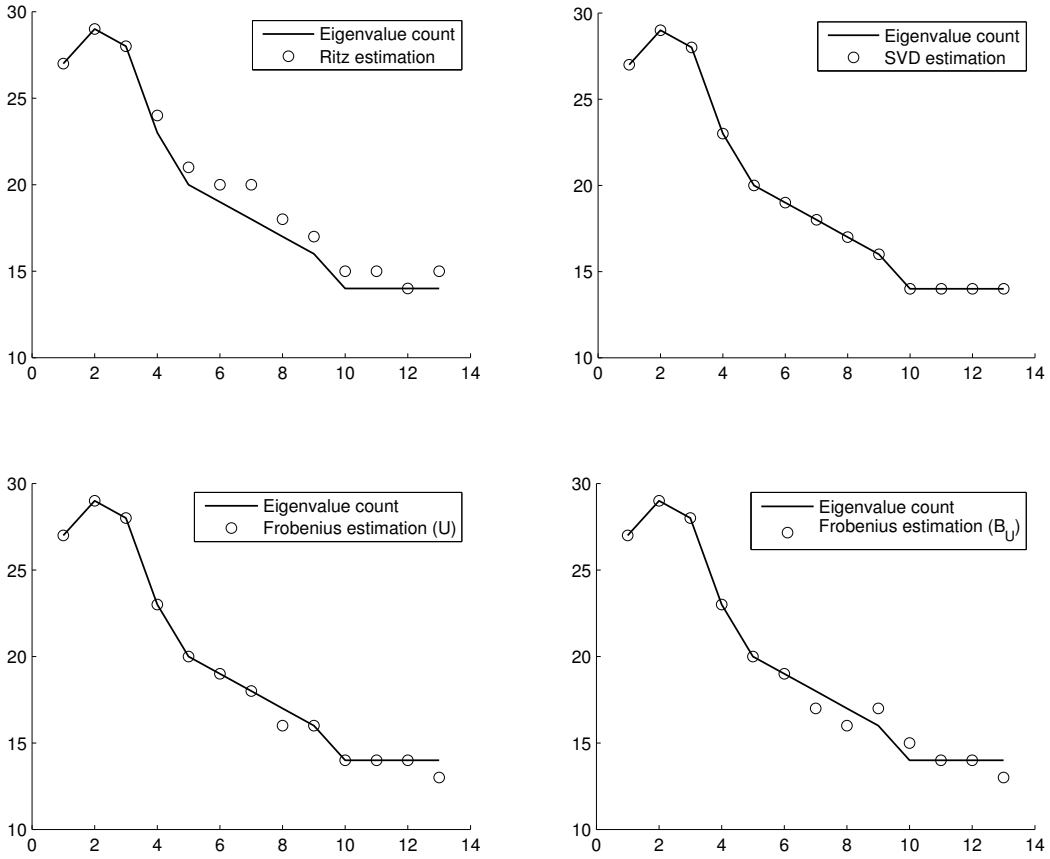

Figure 3.6: Results for Experiment 3.3. For each interval we plotted the exact number of eigenvalues in it (solid lines) and the estimated number of eigenvalues counted with the method named in the legend. The numbers on the abscissa indicate the number of the interval in the progression. The plots have previously been published in [34].

with the absolute values of the diagonal of the triangular factor of the rrQR of B_U. This plot reveals that all except one of the singular values of B_U coincide with the values obtained from the rrQR.

Next, we took a closer look on the computed (singular) values, in particular those on the 1/4-level. Some of them are slightly larger than 1/4, while some are slightly smaller (0.2499...). This suggests counting singular values larger or equal than 0.2499, giving an eigenvalue count of 317. Counting those values ≥ 0.25 gives a count of only 269. The method does not only count the eigenvalues inside the interval, but also all eigenvalues of the cluster.

In this example, the Frobenius norm fails completely, yielding an estimate of 113 eigenvalues. The Ritz count delivered a count of 294, being not too bad considering its cost. ◊

The experiment reveals that the singular values of B_U collapse into three groups, they either take values "around 0", "around 1/4" or "around 1" ("around" meaning slightly larger or smaller). The fact that these values are not hit exactly is clearly due to the different errors that were introduced and discussed in Chapter 2. However, since there are now singular values between, say, 0.2 and 0.25 that belong to the group of the 0.25-values, it is reasonable to count all values above, e. g., 0.2. In our example, even 0.2499 was small enough. The experiment also shows that one should wait until the second iteration before starting the eigenvalue counting process.

Finally, let us assess the abilities of the ostensibly unstable CholQR method from Section 3.2.7.

Experiment 3.5
The setting is the same as in Experiment 3.4. We computed an upper triangular matrix R such that $\mathsf{R}^\star\mathsf{R} = \mathsf{B}_\mathsf{U}$. The Cholesky factorization exists, since B_U has numerically full rank (even though it should have exact rank 317). Next, we solved the linear systems $\mathsf{QR} = \mathsf{U}$ for Q. We could measure $\kappa(\mathsf{Q}) = 1 + \varepsilon$, where ε is of order ε_M. Further, we could measure $\|\mathsf{Q}^\star\mathsf{Q} - \mathsf{I}_{330}\| = \mathcal{O}(10^{-15})$. Just as expected, these values are much better than those suspected in [97], because U itself is already well conditioned and close to orthonormality ($\kappa(\mathsf{U}) = 57.32$).

The eigenvalue count obtained from the diagonal entries of R was 317 just as in Experiment 3.4. ◊

3.3 Numerical integration revisited

In Section 3.2.2 we have seen that the numerical integration of the resolvent and hence the solution of the eigenvalue problem would be exact if the selection function S was exactly the function χ_{I_λ}. This motivates another way for evaluating

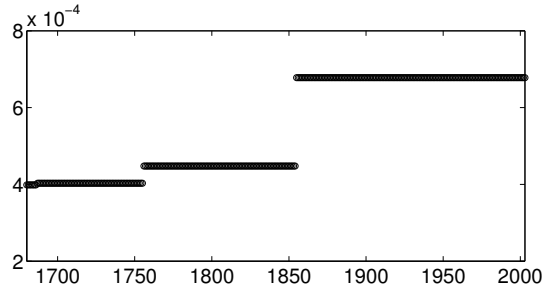

Figure 3.7: Eigenvalues $1680, \ldots, 2003$ of T_{2003}.

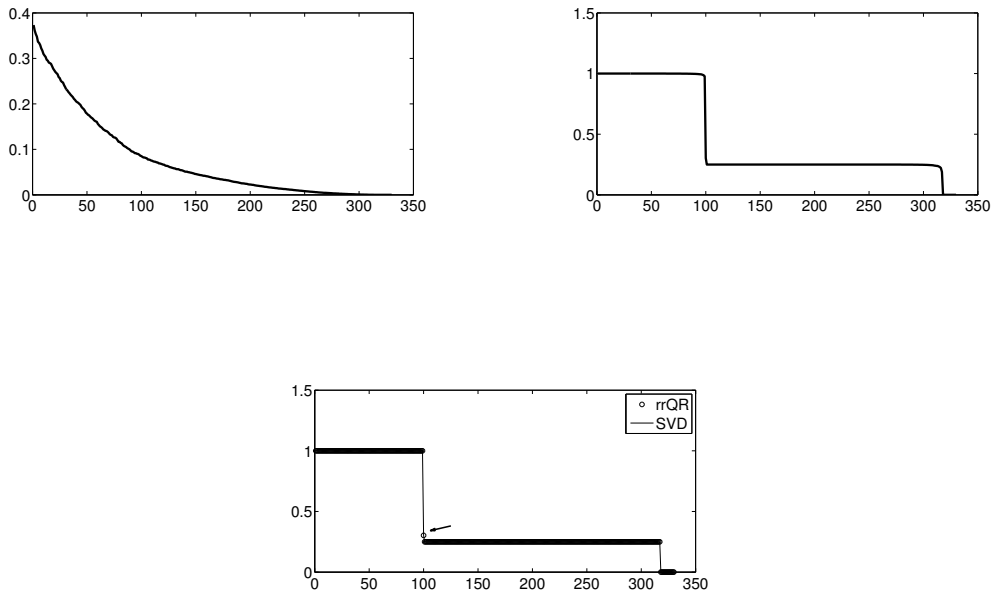

Figure 3.8: Results of Experiment 3.4. Top: Singular values of B_U in first and second iterations. Bottom: Singular values and absolute values of the diagonal entries of the triangular factor of B_U in the third iteration. The values coincide quite well, there is only one value of the rrQR that is larger than the corresponding singular value, marked by the arrow. All quantities are plotted against their index.

the integral

$$\mathsf{U} = \frac{1}{2\pi \mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1} \mathsf{B}\mathsf{Y} \mathrm{d}z. \tag{3.14}$$

The integration defined by weights and points $(\omega_j, t_j)_{j=0,\ldots,p}$ can be considered as the approximation problem

$$S(\lambda) = \sum_{j=0}^{p} \omega_j' r_\lambda(z_j) \approx \chi_{I_\lambda}(\lambda).$$

Three questions arise. First, given a certain integration scheme, how well does the corresponding selection function approximate χ_{I_λ}? Second, is there a way to find an integration method—or some other function S—that approximates χ_{I_λ} in the best way possible (the meaning of "best" is to be specified)?

Finally, we may ask how the approximation error of χ_{I_λ} translates to the integration error of the integral (3.14). This last question can be answered by going back to the matrix function approach of Section 3.2.4. Define for a bounded function f its norm as $\|f\|_\infty = \sup_t |f(t)|$ and suppose we have $\|S - \chi_{I_\lambda}\|_\infty \leq \varepsilon$. Next, recall that for the matrix $\hat{\mathsf{A}} := \mathsf{B}^{-1}\mathsf{A}$ and the eigenvectors X of (A, B) we have $S(\hat{\mathsf{A}}) = \mathsf{X}S(\Lambda)\mathsf{X}^{-1}$ and $\chi_{I_\lambda}(\hat{\mathsf{A}}) = \mathsf{X}\chi_{I_\lambda}(\Lambda)\mathsf{X}^{-1}$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues. Note that $\chi_{I_\lambda}(\hat{\mathsf{A}})$ is, up to the factor Y, the integral (3.14). We hence have for the integration error

$$\left\| \mathsf{X}(S(\Lambda) - \chi_{I_\lambda}(\Lambda))\mathsf{X}^{-1} \right\| \leq \kappa(\mathsf{X})\varepsilon.$$

We may even replace ε by $\max_j |S(\lambda_j) - \chi_{I_\lambda}(\lambda_j)|$. This representation can also be found in [44, Sec. 4.4]. By setting $\hat{\mathsf{A}} = \mathsf{K}^{-\star}\mathsf{A}\mathsf{K}^{-1}$ instead for some matrix K with $\mathsf{K}^\star\mathsf{K} = \mathsf{B}$, we obtain

$$S(\hat{\mathsf{A}}) = \mathsf{K}\mathsf{X}S(\Lambda)(\mathsf{K}\mathsf{X})^\star$$

and likewise for χ_{I_λ}. For the error we obtain in that case

$$\|\mathsf{K}\mathsf{X}(S(\Lambda) - \chi_{I_\lambda}(\Lambda))(\mathsf{K}\mathsf{X})^\star\| \leq \max_j |S(\lambda_j) - \chi_{I_\lambda}(\lambda_j)| \leq \varepsilon$$

since $\|\mathsf{K}\mathsf{X}\| = 1$.

The first two questions are approached in the following.

3.3.1 Approximation by integration methods

In this section, we show that the selection functions belonging to the trapezoidal and midpoint rules are simple, rational functions. We work out the details for the trapezoidal rule, the procedure for the midpoint rule is quite similar (see Remark 3.7).

The trapezoidal rule of order p on the interval $[0, 2\pi]$ reads $(\omega_j, t_j)_{j=0,\ldots,p}$ with $t_j = \frac{j2\pi}{p}$ and $\omega_j = 2\pi/p$, $j = 1, \ldots, p-1$, $\omega_0 = \omega_p = \pi/p$. For the contour being the unit circle we obtain the selection function

$$S(\lambda) = \frac{1}{2\pi\mathbf{i}} \sum_{j=0}^{p} \omega_j \mathbf{i} \exp(\mathbf{i}t_j) \frac{1}{\exp(\mathbf{i}t_j) - \lambda}.$$

Noticing that the first and last summand coincide and the corresponding weights have half the value of all the other weights, we have

$$S(\lambda) = \frac{1}{p} \sum_{j=0}^{p-1} \frac{\exp(\mathbf{i}t_j)}{\exp(\mathbf{i}t_j) - \lambda}. \tag{3.15}$$

Next, write for $|\lambda| < 1$,

$$\begin{aligned} \frac{\exp(\mathbf{i}t)}{\exp(\mathbf{i}t) - \lambda} &= \frac{1}{1 - \lambda \exp(-\mathbf{i}t)} \\ &= \sum_{n=0}^{\infty} (\lambda \exp(-\mathbf{i}t))^n. \end{aligned}$$

It follows for S from (3.15) that

$$S(\lambda) = \frac{1}{p} \sum_{j=0}^{p-1} \sum_{n=0}^{\infty} (\lambda \exp(-\mathbf{i}t_j))^n.$$

Due to absolute convergence we may reorder the double sum and obtain

$$S(\lambda) = \frac{1}{p} \sum_{n=0}^{\infty} \lambda^n \sum_{j=0}^{p-1} (\exp(-\mathbf{i}t_j))^n.$$

The terms $\exp(-\mathbf{i}t_j)^n,\ j = 0, \ldots, p-1$ sum up to p if $n = kp$ for $k \in \mathbb{Z}_{\geq 0}$ and to zero otherwise. Thus, we obtain

$$\begin{aligned} S(\lambda) &= \frac{1}{p} \sum_{n=0,\, p|n}^{\infty} \lambda^n \cdot p \\ &= \sum_{k=0}^{\infty} \lambda^{pk} \\ &= \frac{1}{1 - \lambda^p}. \end{aligned}$$

Next, for $|\lambda| > 1$, we may still write

$$\frac{\exp(\mathbf{i}t)}{\exp(\mathbf{i}t) - \lambda} = \frac{1}{1 - \lambda \exp(-\mathbf{i}t)}. \tag{3.16}$$

The geometric series

$$\sum_{n=0}^{\infty}(\lambda^{-1}\exp(\mathbf{i}t))^n$$

converges absolutely, the limit is

$$\begin{aligned}\sum_{n=0}^{\infty}(\lambda^{-1}\exp(\mathbf{i}t))^n &= \frac{1}{1-\lambda^{-1}\exp(\mathbf{i}t)} \\ &= \frac{\lambda}{\lambda-\exp(\mathbf{i}t)} \qquad (3.17)\\ &= \frac{\exp(-\mathbf{i}t)\lambda}{\exp(-\mathbf{i}t)\lambda-1} \qquad (3.18)\\ &= -\lambda\exp(-\mathbf{i}t)\frac{1}{1-\lambda\exp(-\mathbf{i}t)}, \qquad (3.19)\end{aligned}$$

where (3.17) and (3.18) are obtained by expanding the fraction with λ and $\exp(-\mathbf{i}t)$, respectively. Comparing (3.16) with (3.19) shows that

$$\begin{aligned}\frac{\exp(\mathbf{i}t)}{\exp(\mathbf{i}t)-\lambda} &= -\lambda^{-1}\exp(\mathbf{i}t)\sum_{n=0}^{\infty}(\lambda^{-1}\exp(\mathbf{i}t))^n \\ &= -\sum_{n=1}^{\infty}(\lambda^{-1}\exp(\mathbf{i}t))^n.\end{aligned}$$

For the selection function $S(\lambda)$ we obtain, by using the same arguments as for $|\lambda|<1$,

$$S(\lambda) = -\sum_{k=1}^{\infty}\lambda^{-pk} = -\frac{1}{1-\lambda^{-p}}+1.$$

A simple computation shows that

$$-\frac{1}{1-\lambda^{-p}}+1 = \frac{1}{1-\lambda^{p}}.$$

The following lemma sums up the result.

Lemma 3.6
For the interval $I_\lambda=[-1,1]$ and the p-point trapezoidal rule, the selection function is a rational function on $\mathbb{R}$ with poles in ±1, for even p, and in $+1$, for odd p. Outside the poles it is given by

$$\begin{aligned}S(\lambda) &= \frac{1}{1-\lambda^{p}} \qquad (3.20)\\ &= -\left(\prod_{j=0}^{p-1}(\lambda-\exp(\mathbf{i}t_j))\right)^{-1}. \qquad (3.21)\end{aligned}$$

It is immediately seen that $S(-1) = 1/2$ if p is odd. For general circles $\varphi(t) = c + r\exp(\mathbf{i}t)$ we have to replace λ by $(\lambda - c)/r$ in (3.20) and (3.21). By Lemma 3.6 we now have a simple formula for the rational approximation of χ_{I_λ} by the trapezoidal rule, it is given by

$$\chi_{I_\lambda}(\lambda) \approx \frac{1}{1-\lambda^p}.$$

The formulas (3.20)–(3.21) also suggest a different way to apply the trapezoidal rule to the matrix A or the matrix pair (A, B), respectively. Recall that we are actually interested in $S(\hat{\mathsf{A}})\mathsf{Y}$ with $\hat{\mathsf{A}} = \mathsf{B}^{-1}\mathsf{A}$, where Y denotes the starting basis. In the first formula, a single matrix A can be inserted for λ, resulting in $S(\mathsf{A})\mathsf{Y} = (\mathsf{I} - \mathsf{A}^p)^{-1}\mathsf{Y}$. For the second formula, first note that the product is well defined since shifted matrices commute. It is also applicable to a matrix pair, yielding

$$S(\hat{\mathsf{A}})\mathsf{Y} = -\left(\prod_{j=0}^{p-1} \mathsf{B}^{-1}(\mathsf{A} - \exp(\mathbf{i}t_j)\mathsf{B})\right)^{-1} \mathsf{Y}.$$

In both cases, the condition κ of the linear system to be solved grows exceedingly compared to the original system. Further, full size matrices need to be multiplied. However, formula (3.21) may be applied factor by factor to Y, i. e.,

$$S(\hat{\mathsf{A}})\mathsf{Y} = -\left(\prod_{j=0}^{p-1} (\mathsf{A} - \exp(\mathbf{i}t_j)\mathsf{B})^{-1}\mathsf{B}\right) \mathsf{Y}$$

is computed. This saves the additions needed in the classical trapezoidal rule.

The absolute value of the function S is plotted in the left plot of Figure 3.9. We see that it indeed approximates the function χ_{I_λ} well inside and outside of $[-1, 1]$. The large approximation error is coming from the singularities. Hence, from the approximation point of view no good convergence of the trapezoidal rule is to be expected, while looking directly at the integration rule shows the opposite, see the results in Chapter 2. One attempt to avoiding the singularities is to use the midpoint rule instead of the trapezoidal rule. The arising function has values inside $[0, 1]$ while having problems to approximate around the interval boundaries $-1, 1$. See the right plot in Figure 3.9. In theory, the trapezoidal rule will work well if no eigenvalues of (A, B) are close to the interval boundaries, see Section 3.2.3.

Remark 3.7
For the selection function belonging to the midpoint rule, a formula similar to (3.20) can be shown. The p-point midpoint rule possesses the selection function

$$S(\lambda) = \frac{1}{1+\lambda^p}.$$

This means that S is a continuous function on $\mathbb{R}$ with $S(1) = S(-1) = 1/2$ if p is even. For odd p, the function has a pole in -1. $\Diamond$

By the formulas for the trapezoidal and midpoint rule, the corresponding convergence rates according to Section 3.2.3 can easily be calculated. For instance, suppose we use the midpoint rule of order $p = 8$ on the interval $[-1, 1]$. Let the closest eigenvalue at each interval boundary have a distance of 10^{-6} to the boundary. We then obtain the convergence rate according to (3.11) as

$$\hat{u} = \frac{\frac{1}{1+(1+10^{-6})^p}}{\frac{1}{1+(1-10^{-6})^p}} = \frac{1+(1-10^{-6})^p}{1+(1+10^{-6})^p} = 0.999992 < 1.$$

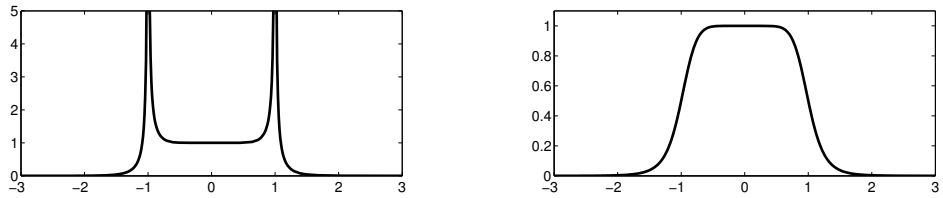

Figure 3.9: Left: Selection function for trapezoidal rule, truncated at function value 5. We plot $|S|$ instead of S. Right: Selection function for midpoint rule. Both rules with order $p = 8$.

3.3.2 Integration by approximation methods

This section is devoted to give an overview of the numerical approximation of the function χ_{I_λ} by functions that are easy to evaluate without further knowledge on the eigenvalues of (A, B).

Approximation by polynomials

Very simple functions for approximation that can cross someone's mind are polynomials. Approximation by a linear combination of Chebyshev polynomials turned out to be a powerful method, Section 3.4 is devoted to that topic.

Rational approximation

The points of discontinuity of χ_{I_λ} make it suitable for being approximated by a rational function. In [111] a technique for approximating χ_{I_λ} by a rational function was developed. The function can even be chosen in an optimal way, using results by Zolotarjov[1] [81]. One ends up with a computation of the subspace U similar to that one obtained by numerical integration. Still the inversion of $z\mathsf{B}-\mathsf{A}$ is necessary.

[1] This is only one possible transliteration from Cyrillic, as found in [81].

3.4 Polynomial approximation

3.4.1 Introduction

The most simple class of functions one could think of for approximating a given, nonlinear, function $f : I \longrightarrow \mathbb{R}$, $I \subset \mathbb{R}$, is the class of polynomials. To this end, one seeks a polynomial p, represented in a certain basis that fulfills $p \approx f$. Here, the sense of "$\approx$" has to be specified. Usually one will require uniform approximation, i. e., a small norm $\|f - p\|_I = \sup_{t \in I} |f(t) - p(t)|$. The existence of a polynomial that approximates χ_{I_λ} on a real interval containing I_λ can be justified, e. g., by the following theorem by Bernstein as found in [67, Thm. 1.1.1].

Theorem 3.8
Let $f : [0, 1] \longrightarrow \mathbb{R}$ be a function that is bounded on $[0, 1]$. Define

$$B_N^f(t) = \sum_{k=0}^{N} f\left(\frac{k}{N}\right) \binom{N}{k} t^k (1-t)^{N-k}.$$

We then have $\lim_{N\to\infty} B_N^f(t) = f(t)$ at each point of continuity of f. If f is continuous on $[0, 1]$, the limit is uniform.

The polynomials B_N^f are called *Bernstein polynomials*. The spectrum of a matrix A can simply be transformed such that it is contained in $[0, 1]$, then the theorem shows the existence of a pointwise polynomial approximation to χ_{I_λ}. It does however not suggest a practical way to construct such an approximation, see Section 3.4.9 below.

Let us come back to the FEAST algorithm. The characteristic function χ_{I_λ} of I_λ can be approximated by a polynomial or by the selection function belonging to a given integration scheme, see Section 3.3.1. We see that in the first case *no inversion* of the matrix is necessary in contrast to the latter one. Before we start our further discussion, let us list some obvious (dis)advantages ($\oplus$/$\ominus$) of polynomial approximation:

- $\oplus$ No matrix inversion is necessary.
- $\oplus$ It can be implemented by only using matrix-vector multiplication, since actually we are interested in $\chi_{I_\lambda}(\mathsf{A})\mathsf{Y}$.
- $\oplus$ Regarding the result as an approximate integral, all theory stays valid.
- $\oplus$ No complex arithmetic is necessary if the matrix is real.
- $\ominus$ The polynomial order can be high.
- $\ominus$ The polynomials are univariate, in other words, we can only insert *one* matrix. For the generalized problem involving (A, B) we have to revert to

$\hat{\mathsf{A}} \in \{\mathsf{K}^{-\star}\mathsf{A}\mathsf{K}^{-1}, \mathsf{B}^{-1}\mathsf{A}\}$. The second option is only of theoretical interest, and both involve the full and exact factorization of a full scale matrix.

There is another, more inherent reason for using polynomial approximation. The linear systems arising in the numerical integration employed in the FEAST algorithm typically have to be solved by an iterative linear solver, such as GMRES [90], cf. Section 3.6.1. Such solvers usually rely on Krylov subspaces, meaning they do nothing else than applying a polynomial in A to a certain starting vector. Hence, we end up in using polynomials again. We then can rather avoid the "formal" inversion of the matrix (actually, the inverse is approximated by polynomials) and use polynomials directly. See also the discussions in Sections 2.5.5 and 3.6.1.

3.4.2 Chebyshev approximation

A well known technique is the approximation by a so called Chebyshev[2] series. Here, a function f is approximated by a series

$$f(x) = \sum_{k=0}^{\infty} c_k T_k(x), \tag{3.22}$$

where the functions T_k are the *Chebyshev polynomials* and the numbers $c_k \in \mathbb{R}$ are certain coefficients. For an introduction to Chebyshev polynomials and approximation see, e. g., [115]. In practice, of course, the series (3.22) is truncated at a certain value $k = N$. This results in an approximation

In the following, we will use the term "Chebyshev approximation" for the approximation of a given function by a linear combination of Chebyshev polynomials.

It were—to the best of our knowledge—Druskin and Knizhnerman [27] who first introduced Chebyshev approximation for the evaluation of matrix functions. Since the approximation only works on $[-1, 1]$, the spectrum of the matrix has to be transformed to that interval via a simple linear transformation [27]. Having $\lambda_{\max}$ and $\lambda_{\min}$ at hand, we can transform the matrix A (and hence its spectrum) via the linear function

$$\lambda \mapsto \frac{2}{\lambda_{\max} - \lambda_{\min}}\lambda - \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}. \tag{3.23}$$

Here, the signs were changed compared to [27] in order to obtain an increasing function. Note that $\lambda_{\min}$, $\lambda_{\max}$ can also be substituted by lower and upper bounds of these values, respectively.

[2]Chebyshev was a Russian mathematician, there exist several transliterations of his name from Cyrillic to Latin letters

For the application of Chebyshev polynomials to matrices, the recurrence relation

$$T_0(x) = 1,\ T_1(x) = x,\ T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \tag{3.24}$$

rather than the explicit formulation

$$T_k(x) = \cos(k \arccos(x)),\ x \in [-1, 1]$$

can be used. In our case, where we want to approximate $\chi_{I_\lambda}(\mathsf{A})\mathsf{Y}$, Algorithm 3.2 can be used. Note that only multiplications of A with (probably narrow) matrices of size $n \times \widetilde{m}$ are employed, when $\mathsf{Y} \in \mathbb{C}^{n\times\widetilde{m}}$. Neither access to the individual entries of A nor the solution of linear systems with A is necessary. The coefficients

Algorithm 3.2 Application of Chebyshev polynomials

1: $\mathsf{T}^{(0)} = \mathsf{Y},\ \mathsf{T}^{(1)} = \mathsf{AY}$
2: $\widetilde{\mathsf{U}} = c_0\mathsf{T}^{(0)} + c_1\mathsf{T}^{(1)}$
3: **for** $k = 2, \ldots, N$ **do**
4: $\quad \mathsf{T}_{\text{new}} = 2\mathsf{AT}^{(1)} - \mathsf{T}^{(0)}$
5: $\quad \mathsf{T}^{(0)} = \mathsf{T}^{(1)},\ \mathsf{T}^{(1)} = \mathsf{T}_{\text{new}}$
6: $\quad \widetilde{\mathsf{U}} = \widetilde{\mathsf{U}} + c_k\mathsf{T}_{\text{new}}$

c_k for approximating a function f by a Chebyshev series (3.22) are given by [27]

$$c_k = \frac{\min(2, k+1)}{\pi} \int_{-1}^{1} f(t)T_k(t)(1-t^2)^{-1/2}\mathrm{d}t.$$

We have

$$\int T_k(t)(1-t^2)^{-1/2}\mathrm{d}t = \begin{cases} \arcsin(t) & k = 0 \\ \dfrac{-\sin(k\arccos(t))}{k} & k > 0 \end{cases}.$$

Noting $\arcsin(\underline{\lambda}) - \arcsin(\overline{\lambda}) = \arccos(\overline{\lambda}) - \arccos(\underline{\lambda})$ we obtain for the coefficients of the Chebyshev series of χ_{I_λ} with $-1 < \underline{\lambda} < \overline{\lambda} < 1$ (also [82]),

$$c_k = \begin{cases} \dfrac{\arccos(\underline{\lambda}) - \arccos(\overline{\lambda})}{\pi} & k = 0 \\ \frac{2}{k\pi}(\sin(k\arccos(\underline{\lambda})) - \sin(k\arccos(\overline{\lambda}))) & k > 0 \end{cases}. \tag{3.25}$$

It is known that Chebyshev approximation of functions with points of discontinuity leads to heavy oscillations (so called Gibbs oscillations) of the approximant near those points. The effect is described and illustrated with several examples in [115]. To avoid most of the oscillations in the approximant one can multiply the coefficients c_k by weights g_k, (Gibbs coefficients), leading to new coefficients $c'_k = c_k g_k$. The numbers g_k typically fulfill $|g_k| \leq 1$, see below. For the function

χ_{I_λ}, the so called Jackson kernel (cf. page 120) has proven to be most efficient, see [115]. The resulting weights g_k are given by (see [115])

$$g_k = \frac{1}{N+1}\left((N-k+1)\cos\left(\frac{\pi k}{N+1}\right) + \sin\left(\frac{\pi k}{N+1}\right)\cot\left(\frac{\pi}{N+1}\right)\right). \quad (3.26)$$

The emerging approximating polynomial of degree N will be denoted by Ψ_N in the sequel.

It should be noted that Chebyshev polynomials have been used in the solution of eigenvalue problems for a long time, see, e. g., Wilkinson's almost 50 year old monograph [118]. Their key feature is that they can be used to filter out the wanted parts of the spectrum. This has for instance been done in slightly different contexts in [93, 121, 122], also using the coefficients (3.26).

Here, the polynomial approximation is to be understood as an alternative to numerical integration in the context of the FEAST algorithm. In particular, the approximation can be used in an implementation of FEAST to replace the integration with very low programming effort.

3.4.3 Error estimation

The following is essentially Theorem 1 from [27], with the difference that Chebyshev polynomials are multiplied by matrices with $\widetilde{m}$ columns instead of single vectors.

Theorem 3.9
Assume that (3.22) *is absolutely convergent in* $[-1,1]$ *and let* g_k, $|g_k| \leq 1$ *denote certain Gibbs coefficients. Let* U *denote the exact integral and* $\|\mathsf{Y}\| = 1$, *we then have*

$$\left\|\mathsf{U} - \widetilde{\mathsf{U}}\right\| \leq \sum_{k=N+1}^{\infty} |c'_k| < +\infty.$$

Proof. (Essentially the proof of Theorem 1, [27]) Since we have $T_k(1) = 1$ for all k and (3.22) converges absolutely, we have

$$\sum_{k=0}^{\infty} |c_k| < +\infty.$$

Since $|g_k| \leq 1$, also the sum of the c'_k converges absolutely. Next, by using $\mathrm{spec}(\mathsf{A}) \subset [-1,1]$ we have $\|\mathsf{A}\| \leq 1$. Because $|T_k(x)| \leq 1$ for $x \in [-1,1]$ we obtain

$$\left\|\widetilde{\mathsf{U}} - \mathsf{U}\right\| \leq \sum_{k=N+1}^{\infty} |c'_k|\,\|T_k(\mathsf{A})\|\,\|\mathsf{Y}\| \leq \sum_{k=N+1}^{\infty} |c'_k|\,.$$

□

We cannot assume in general that the prerequisites of the theorem are met. For instance, with $\underline{\lambda} = \cos(\frac{\pi}{4}) < \overline{\lambda} = \cos(\frac{\pi}{8})$ we have for the coefficients (3.25), $k > 0$,

$$c_k = \frac{2}{k\pi}\left(\sin\left(k\frac{\pi}{4}\right) - \sin\left(k\frac{\pi}{8}\right)\right).$$

For $k = 4, 12, 20, 28\ldots$ we have $|c_k| = 2/k\pi$. Consequently,

$$\begin{aligned}\sum_{k=0}^{\infty} |c_k| &\geq \sum_{8|(k-4),\, k>0} |c_k| \\ &= \frac{2}{\pi} \sum_{8|(k-4),\, k>0} \frac{1}{k} \\ &= \frac{2}{\pi} \sum_{k=1}^{\infty} \frac{1}{8k-4}.\end{aligned}$$

The last sum is a partial series of the harmonic series, its divergence is easily seen.

The bound in the theorem is more a theoretical one. Using the approaches from Section 3.3—measuring uniform or pointwise errors in the approximation—we obtain for the pointwise error at point t

$$e(t) := \left|\chi_{I_\lambda}(t) - \sum_{k=0}^{N} c'_k T_k(t)\right| = \begin{cases} \left|\sum_{k=0}^{N} c'_k T_k(t)\right| & : t \notin I_\lambda \\ \left|1 - \sum_{k=0}^{N} c'_k T_k(t)\right| & : t \in I_\lambda \end{cases}. \tag{3.27}$$

These numbers can actually be computed. The points t that are of our interest are of course only the eigenvalues of the matrix A.

For a Hermitian matrix with eigenvalues λ_j and an orthonormal starting basis Y we then obtain

$$\left\|\mathsf{U} - \widetilde{\mathsf{U}}\right\| \leq \max_j e(\lambda_j). \tag{3.28}$$

In order to get an estimation of the error, we can use Ritz values that are already computed at a certain iteration of FEAST and apply formula (3.27) to those values, where the largest error usually occurs near the interval boundary. It hence might be sufficient to compute (3.27) only for some Ritz values near $\underline{\lambda}$, $\overline{\lambda}$ in order to get an estimation for the right hand side of (3.28). By using the formulation $T_k(t) = \cos(k \arccos(t))$, the evaluation of $e(t)$ requires N additions, $N + 1$ multiplications and $N + 1$ evaluations of cosine and arcus cosine. The numbers c'_k are available from the computation of the subspace itself and need to be computed only once.

In the following experiment we perform a thorough experimental analysis of the errors (3.27). The aim of the experiment is to give the reader an idea of how well χ_{I_λ} can be approximated by polynomials.

Experiment 3.10
For a fixed interval length we performed an interval progression (see Section 3.2) and let I_λ reside in different locations of $[-1, 1]$. At certain grid points we measured (3.27). Keep in mind that the given grid points should represent eigenvalues of some matrix. In steps 1.–4. we used the plain Chebyshev coefficients (3.25).

1. In the first experiment we let $I_\lambda^{(k)} = [k\cdot 0.2-1, (k+1)\cdot 0.2-1]$ for $k = 0, \ldots, 9$, i. e., we take 10 intervals each with length 0.2 whose union covers $[-1, 1]$. For each interval we compute Ψ_N for $N = 500$, 1000, 5000 and evaluate it at 500 equidistant points $t_j = -1 + 2(j-1)/499,\ j = 1, \ldots, 500$ from -1 to 1. To get a feeling for the shape of Ψ_N and the errors, we plotted $\Psi_{500}, \chi_{[0,0.2]}$ and the corresponding error in Figure 3.10.

   For each of the $I_\lambda^{(k)}$ and each value of N, we measured $\max_j e(t_j)$. The results are depicted in Figure 3.11. Every data point in this figure hence corresponds to the maximum of the solid lines of one plot as that one in Figure 3.10.

   When choosing an odd number of points t_j (or any other spacing such that the boundary of one interval I_λ hits one of the points t_j) the results differ. At the boundary of I_λ, the error is always approximately 0.5, see Section 3.4.4 below. Being more precise, we have $\Psi_N(\underline{\lambda}), \Psi_N(\overline{\lambda}) \to 0.5$ for $N \to \infty$.

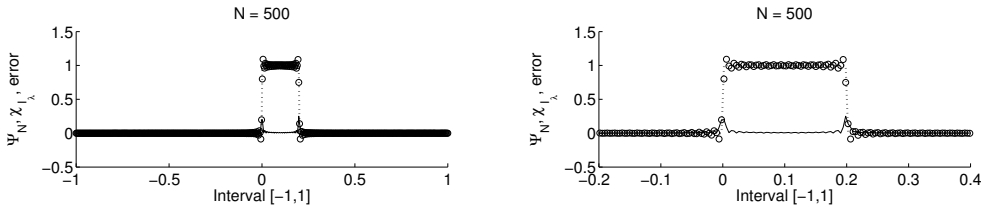

Figure 3.10: χ_{I_λ}, marked by dashed line, Ψ_{500}, marked by ∘ and corresponding error, marked by solid line. Left: complete interval $[-1, 1]$. Right: Magnification of $[-0.2, 0.4]$.

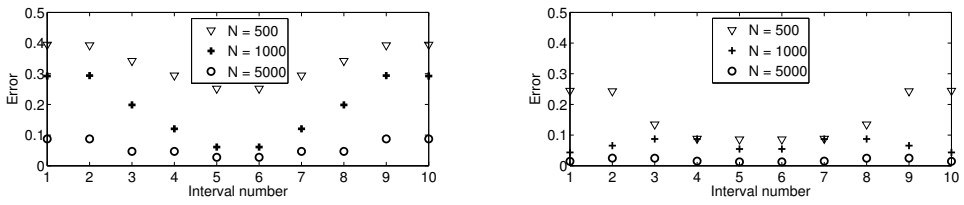

Figure 3.11: Results for Experiment 3.10, 1. (left) and 2. (right).

2. Next, let us repeat our experiment from 1. with 200 equidistant points $t_j = -1 + 2(j-1)/199$, $j = 1, \ldots, 200$ between -1 and 1. The results can also be seen in Figure 3.11. Comparison between 1. and 2. illustrates the behavior of $e(t)$ subject to the distance between t and any of the boundaries of the intervals $I_\lambda^{(k)}$. For the finer grid with 500 points we have larger errors. This is due to the fact that the approximation error is large near the boundary of $I_\lambda^{(k)}$ (see also 4.). For 500 points t_j, those points are closer to the interval boundaries than for 200 points.

3. In Figure 3.12, we see results of an experiment performed as in 1., but with grid points t_j chosen as uniformly and normal distributed pseudo random numbers, respectively. We formed the respective set of points by first creating 500 pseudo random numbers, then centering those numbers around zero (i. e., the smallest and largest number have the same absolute value) and finally scaling them to $[-1, 1]$. For each value of N we used the same set of grid points for all intervals $I_\lambda^{(k)}$.

   If $I_\lambda^{(k)}$ is located near the boundary of $[-1, 1]$, the error tends to be smaller than in the center of $[-1, 1]$.

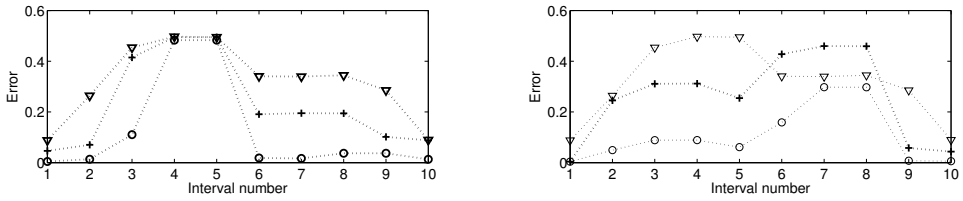

Figure 3.12: Results for Experiment 3.10, 3. The left plot was generated for uniform distributions, the right one for normal distributions. Once again, the $\triangledown$ symbol stands for $N = 500$, $+$ for $N = 1000$ and $\circ$ for $N = 5000$.

4. We now take a more detailed look on the errors $e(t)$ in order to see where in $[-1, 1]$ the largest errors occur. To see an interesting effect, we take the normal distributed grid points from 3. For these points t_j we plotted $\Psi_N(t_j)$, $\chi_{I_\lambda}(t_j)$ and $e(t_j)$ for $N = 500$, 1000, 5000. We fixed $I_\lambda = [0, 0.2]$, i. e., near the center of $[-1, 1]$, where we expect the largest errors. The results are shown in Figure 3.13. The plots reveal that the errors occur near the boundary of I_λ, while the approximation is quite good in the rest of the interval $[-1, 1]$.

   Note that there are only few oscillations near the boundary of I_λ. The oscillations are only relevant at discrete points. They would be damped if Gibbs coefficients such as those from equation (3.26) were used.

5. Now, let compare the effect of 5 different Gibbs coefficients which can be found in [115] and are compiled in the following list.

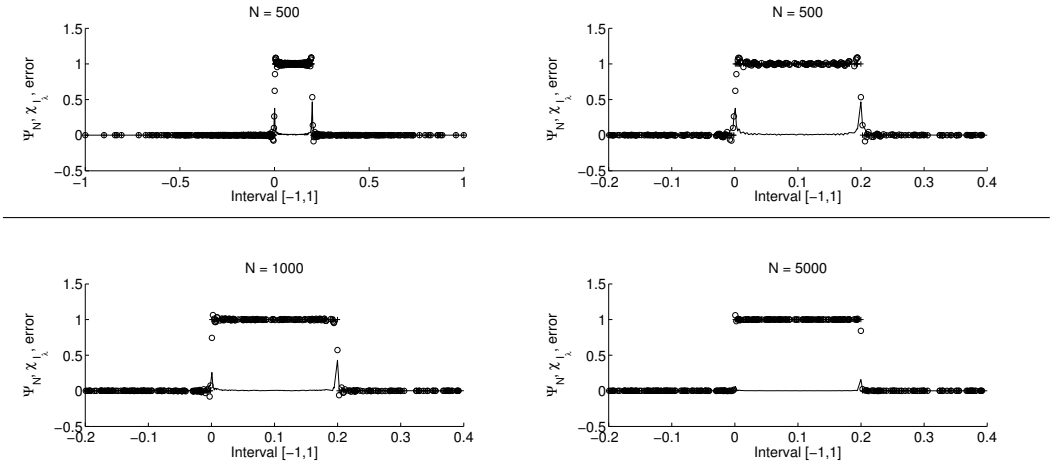

Figure 3.13: Results for Experiment 3.10, 4. The top left plot is the full picture for $N = 500$. What follows are magnifications of the region of interest. The values of $\Psi_N(t_j)$ are marked with $\circ$, the values of $\chi_{I_\lambda}(t_j)$ are marked by $+$ and the error is represented by the solid line.

- The Dirichlet kernel $g_k = 1$.
- The Jackson kernel (3.26).
- The Lorentz kernel $g_k = \sinh(\mu(1 - k/N))/\sinh(\mu),\ \mu \in \mathbb{R}$. Values $\mu \in [3, 5]$ are recommended. We used $\mu = 4$.
- The Fejér kernel $g_k = 1 - k/N$.
- The Lanczos kernel

$$g_k = \left(\frac{\sin(\pi k/N)}{\pi k/N}\right)^M,\ M \in \mathbb{Z}_{>0},\ k > 0,\ g_0 = 1.$$

  With $M = 3$, coefficients g_k similar to those given by the Jackson kernel are obtained [115].

The different kernels are shown in Figure 3.14 for $N = 100$. All kernels have absolute value ≤ 1. In Figure 3.15 we plotted similar results as in 4., with normal distributed grid points, the five different kernel types and $N = 500$. Once again we zoomed in on the region of interest. We see that indeed the oscillations decrease, while the errors near the boundary of I_λ are still present.

In our experiments the maximum errors measured in 1.–3. did not differ significantly for the different kernels. In the center of the interval $[-1, 1]$ we always observed errors around 0.5.

6. Finally, we repeated the experiment from 5. but measured average and median errors, since the maximum errors are a very local phenomenon near

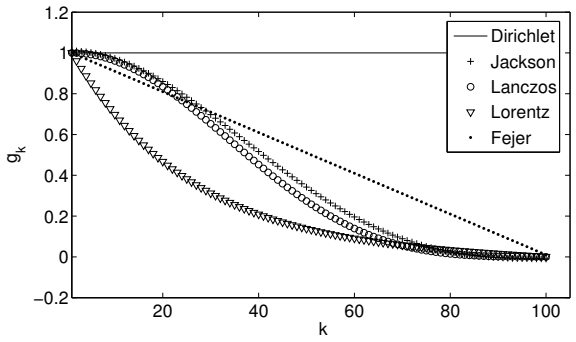

Figure 3.14: The 5 different kernels from Experiment 3.10, 5. for $N = 100$.

kernel	$N = 500$	$N = 1000$	$N = 5000$
Dirichlet	Avg.: 10^{-2} Med.: 10^{-3}	Avg.: 10^{-3} Med.: 10^{-3}	Avg.: 10^{-3} Med.: 10^{-4}
Jackson	Avg.: 10^{-2} Med.: 10^{-4}	Avg.: 10^{-3} Med.: 10^{-4}	Avg.: 10^{-3} Med.: 10^{-5}
Lorentz	Avg.: 10^{-2} Med.: 10^{-2}	Avg.: 10^{-2} Med.: 10^{-3}	Avg.: 10^{-2} Med.: 10^{-3}
Fejér	Avg.: 10^{-2} Med.: 10^{-3}	Avg.: 10^{-2} Med.: 10^{-3}	Avg.: 10^{-3} Med.: 10^{-4}
Lanczos	Avg.: 10^{-2} Med.: 10^{-5}	Avg.: 10^{-3} Med.: 10^{-5}	Avg.: 10^{-3} Med.: 10^{-7}

Table 3.1: Results for Experiment 3.10, 6. Average and median errors of approximation with Chebyshev polynomials using different kernels. The errors were measured at 500 random grid points. Only the order of magnitude is shown.

the interval boundary. The results are seen in Table 3.1. The Lanczos kernel shows a good behavior in the median error, but the arising polynomial is not very steep around the interval boundaries, see Figure 3.15. This effect is not desirable, cf. Section 3.4.6.

$\Diamond$

The experiments show that the Chebyshev approximation of the function χ_{I_λ} does not work too bad outside a narrow area around the boundaries of I_λ. If no eigenvalues are too close to $\underline{\lambda}$, $\overline{\lambda}$, the bound on the right hand side of (3.28) will be small. It is once more worth to mention that the normwise error in the subspace bases is not the most important measure, cf. Section 3.6.5.

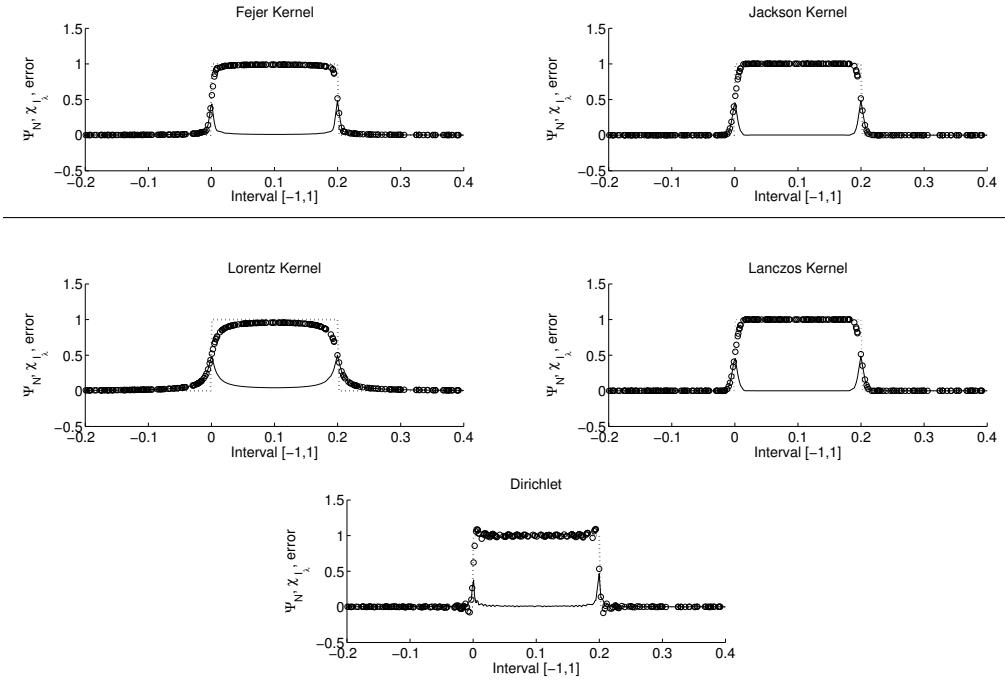

Figure 3.15: Results for Experiment 3.10, 5. Plots of χ_{I_λ}, Ψ_{500} and the error for five different kernels. The values of $\Psi_N(t_j)$ are marked with $\circ$, the values of $\chi_{I_\lambda}(t_j)$ are marked by the dashed line and the error is represented by the solid line.

3.4.4 Error at the boundary of I_λ

For the eigenvalue counting methods from Section 3.2 it is important to know which values the approximating polynomial attains at the boundary of I_λ. Surprisingly at first glance, it turned out that $\Psi_N(\underline{\lambda}), \Psi_N(\overline{\lambda})$ approaches 0.5 for $N \to \infty$. This is the same value that the selection functions belonging to certain integration methods attain at those points. The different eigenvalue counting methods from Section 3.2 can thus be used in the same way. The function Ψ_N might attain values slightly larger than 0.5 at $\underline{\lambda}$, $\overline{\lambda}$. The threshold consequently should be chosen slightly larger than 1/2 in practice, e. g., 0.55 to 0.6.

3.4.5 Experiments with Chebyshev-FEAST

Now, it is time to see the Chebyshev approximation in action within the FEAST algorithm. In this section, we will investigate the capability and reliability of the Chebyshev polynomial method in the context of the FEAST algorithm. To this end we will measure the needed iteration count of the outer FEAST iteration, the achieved numerical quality (see Sec. 1.5.1) and the interplay of N (the polynomial order) and n (the matrix size).

It should be noted that the focus is not so much on runtime of the method, since we do not use optimized code. One should rather think of a highly parallel

machine as target architecture, equipped with a very efficient and scaling matrix-vector product (or the product of a square with a narrow rectangular matrix).

Test design

In a first test run (Experiment 3.11) we will apply the—from now on called Chebyshev-FEAST —method to different matrices from small to modest size (modest means 10^6). We use different polynomial orders N, different search intervals I_λ and measure the iteration count that is necessary for all eigenpairs with eigenvalue in I_λ to converge. We further measure the blockwise relative residual $\mathsf{res} = \left\| \mathsf{A}\widetilde{\mathsf{X}} - \widetilde{\mathsf{X}}\widetilde{\Lambda} \right\| / \|\mathsf{A}\|$ and the achieved orthogonality $\mathsf{orth} = \max_{i \neq j, \widetilde{\lambda}_i, \widetilde{\lambda}_j \in I_\lambda} |\widetilde{\mathsf{x}}_i^\star \widetilde{\mathsf{x}}_j|$. The intervals I_λ are, except for Problem 5, chosen such that they include the indicated set of eigenvalues but no other eigenvalue. When setting one of the interval boundaries exactly to an eigenvalue (which is of course not known in practice), this eigenvalue might be lost in the computation. As convergence criterion we used a per-eigenpair norm criterion, see Section 3.6.4.

In a second run (Experiment 3.12) we fix one interval and let N range over a broader set of values.

Experiments

Experiment 3.11
We start with the first kind of test described above and tabulate the results. In the table for each problem only the order of magnitude of the values for res and orth is given. The test runs for Problem 1 were performed on a standard work station, the other test runs on a 12 core machine with 96 GB of main memory.

Problem 1. The first matrix is T_{2003} of size $n = 2003$ already used in Section 3.2.8, its spectrum is shown in Fig. 3.4. For this experiment, we forced the eigenvalues to be in $[-1, 1]$ by just multiplying the matrix with $\lambda_{\max}^{-1}$ (we have $|\lambda_{\max}| > |\lambda_{\min}|$).

The results are shown in Table 3.2. The difference between the second and the third interval seems to be marginal. While in the second case a cluster is cut, in the third case only complete clusters are included in the interval. We experimented with subspace sizes between $\widetilde{m} = 350$ and $\widetilde{m} = 450$, while larger subspaces lead to faster convergence. The eigenvalue count mechanism based on singular values from Section 3.2 automatically deletes as many vectors from the current basis as possible. As already mentioned in Experiment 3.4, one should wait at least until the second iteration. The reason is that in the first one the computed singular values do not yet carry enough information. It also turned out that the number of eigenvalues was estimated very accurately, but a search space slightly larger is desirable, cf. Experiment 3.1. See also [85], where $\widetilde{m} = 1.5 \cdot m$ was proposed. To get

I_λ	$N = 500$, **G**	$N = 500$,	$N = 1000$, **G**	$N = 1000$,
$[\lambda_{999}, \lambda_{1296}]$	4 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-13}$	9 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-13}$	4 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-14}$	7 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$
$[\lambda_{1704}, \lambda_{2003}]$	5 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-13}$	7 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-14}$	6 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-13}$	6 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-13}$
$[\lambda_{1687}, \lambda_{2003}]$	4 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-14}$	9 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-13}$	4 iterations $\mathsf{orth} = 10^{-15}$ $\mathsf{res} = 10^{-14}$	7 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$

Table 3.2: Results for Problem 1., Experiment 3.11. The letter **G** means that Gibbs coefficients defined by Jackson kernel were used.

comparable results, we used this estimate (of course, rounded to the next integer) to produce the results in Table 3.2, without shrinking the subspace. As starting basis we used a random (but fixed) orthonormal matrix.

It turned out that it sometimes is necessary to use a slightly larger interval than I_λ as stated in Table 3.2, in order to ensure convergence of the eigenvalues at the boundary. If doing so, some more eigenpairs are computed and finally those with eigenvalue outside I_λ are dropped.

Discussion of results of Problem 1. Looking at Table 3.2, we see that the best results for each interval are achieved when using a polynomial degree of 1000 and switching on the Jackson kernel. The differences from $N = 500$ to $N = 1000$ are marginal, while the number of necessary operations is roughly proportional to N. Hence, using $N = 500$ is preferable. The effect of switching the Gibbs coefficients on/off in contrast is remarkable and is reflected most in the iteration count.

We chose the matrix T_{2003} because of its challenging spectrum, consisting of many heavily clustered eigenvalues. The distances $\lambda_j - \lambda_{j-1},\ j = 2, \ldots, 2003$ are depicted in Figure 3.16. The distances belonging to eigenvalues inside a cluster are well distinguishable, we marked them with an ellipse. The distances inside the clusters are all of order about $10^{-12} - 10^{-15}$; the inverse of these numbers plays an important role in the theoretical assessment, cf. Chapter 2. Nonetheless, the method is performing well.

Problem 2. We performed the same kind of test as in Problem 1 with the matrix $\mathsf{A} = \texttt{LAP\_CIT\_6752}$ [107] of size $n = 6752$. This matrix originates from modeling citations in scientific publications. The spectrum of A, which is depicted in Figure 3.17, does not look very spectacular. Its difficulty is the cluster of zero eigenvalues of size 402. We hence focus on an interval

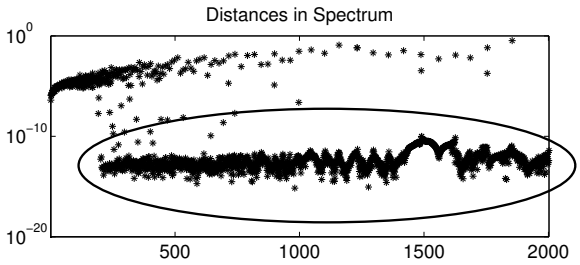

Figure 3.16: Distances in the spectrum of T_{2003}. The distances belonging to eigenvalues inside a cluster are marked with the ellipse.

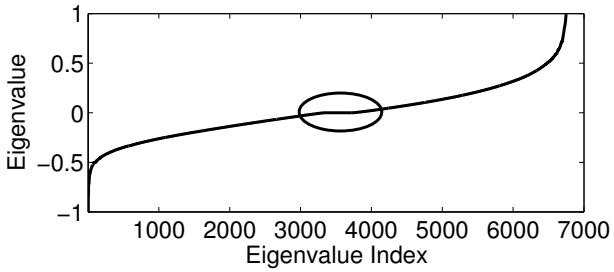

Figure 3.17: Eigenvalues of `LAP_CIT_6752`. The cluster of zero eigenvalues is marked with the ellipse.

containing the eigenvalues $\lambda_{3001}, \ldots, \lambda_{4000}$, therefore in particular the 402-fold zero eigenvalue. The results are shown in Table 3.3, in a similar fashion as the previous results. We chose $N = 250$ as smallest polynomial degree due to the experience from Experiment 3.12, where good convergence was achieved from that value on. The results confirm that this is the case here as well.

Problem 3. Next, we move on to a larger example. The matrix $\mathsf{A} =$ `Poly27069` [8] with $n = 27{,}069$ arises in electronic structure calculations. In those cal-

kernel	$N = 250$	$N = 500$	$N = 1000$
Jackson	4 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$	5 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$	4 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$
Dirichlet	10 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$	8 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$	8 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$

Table 3.3: Results for Problem 2., Experiment 3.11. Two runs were performed, one with Jackson kernel and one with Dirichlet kernel (i. e., $g_k = 1$, $k = 0, \ldots, N$).

kernel	$N = 250$	$N = 500$	$N = 1000$
Jackson	—	9 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-13}$	5 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-14}$
Dirichlet	10 iterations $\mathsf{orth} = 10^{-13}$ $(*)$ $\mathsf{res} = 10^{-13}$	10 iterations $\mathsf{orth} = 10^{-12}$ $(**)$ $\mathsf{res} = 10^{-13}$	9 iterations $\mathsf{orth} = 10^{-12}$ $\mathsf{res} = 10^{-13}$

Table 3.4: Results for Problem 3., Experiment 3.11. Two runs were performed, one with Jackson kernel and one with Dirichlet kernel (i. e., $g_k = 1$, $k = 0, \ldots, N$). In $(*)$ only 2579 eigenpairs converged, in $(**)$ the eigenpair with smallest eigenvalue was missed (i. e., 2999 eigenpairs converged). The symbol "—" means that no eigenpair converged to the desired accuracy.

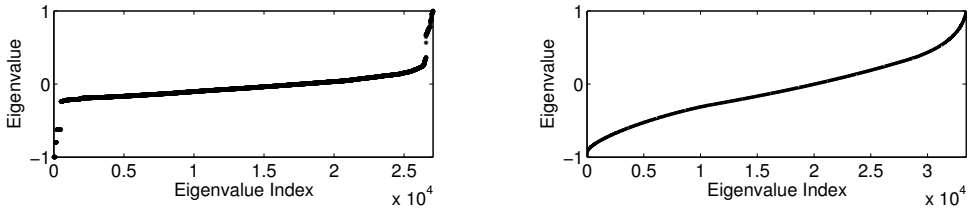

Figure 3.18: Eigenvalues of `Poly27069` (left) and `RAP_PARSEC_33401` (right).

culations, the eigenpairs with the lowest eigenvalues are desired. In this example, we tried to compute A's lowest 3000 eigenpairs with the Chebyshev-FEAST method as described. The goal is in particular to get a feeling for the connection between the matrix size and the necessary polynomial degree. The results are shown in Table 3.4. Here, the method failed to converge within 10 iterations with $N = 250$, using the Jackson kernel. A "fail" in this context means that not a single eigenpair converged to the desired accuracy. However, some eigenpairs reached a residual of order 10^{-10} after all. For $N = 500$, all desired 3000 eigenpairs converged when the Jackson kernel was used (we set $\mathsf{tol} = 10^{-12}$ here).

Interestingly, when using the Dirichlet kernel with $N = 250$, the method was able to compute 2579 eigenpairs to the desired accuracy in the given limit of 10 iterations. For the other polynomial degrees $N = 500$, $N = 1000$, we again observe the behavior seen before, e. g., in Problem 2. This includes higher reliability and/or lower iteration counts when using the Jackson kernel.

Note that the spectrum of A is challenging due to its large jumps in the lower eigenvalues, see the left plot in Figure 3.18. The polynomial degrees however could be chosen as in the smaller examples (Problems 1 and 2).

kernel	$N = 250$	$N = 500$	$N = 1000$
Jackson	5 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-15}$	5 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-14}$	4 iterations $\mathsf{orth} = 10^{-14}$ $\mathsf{res} = 10^{-15}$
Dirichlet	7 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-14}$	6 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-14}$	5 iterations $\mathsf{orth} = 10^{-13}$ $\mathsf{res} = 10^{-14}$

Table 3.5: Results for Problem 4., Experiment 3.11. Two runs were performed, one with Jackson kernel and one with Dirichlet kernel (i. e., $g_k = 1$, $k = 0, \ldots, N$).

Problem 4. We come to the matrix `RAP_PARSEC_33401` from network modeling [107] with $n = 33{,}401$, where we transformed the spectrum to $[-1, 1]$. It looks rather unspectacular, see the right plot in Figure 3.18. We sought the 3000 lowest eigenpairs. The difficulty is that $\lambda_2 - \lambda_1 \approx 0.05$, while all other eigenvalues are very close to each other in absolute sense. In average the distance of all neighbored eigenvalues is about 10^{-5} while some of the distances are much smaller. In a first attempt with $N = 250$, the eigenpairs with eigenvalues $\lambda_2, \ldots, \lambda_{3000}$ converged within 5 iterations to a blockwise relative residual norm of order 10^{-13}. The spectrum was transformed to $[-1, 1]$ via (3.23), meaning the smallest eigenvalue is exactly -1. This is not desirable because it is the smallest value that can be computed—in theory—by the Chebyshev method. In fact, the method sometimes failed to compute the lowest eigenvalue. We hence transformed the spectrum in a second attempt to $[-0.95, 0.95]$.

The results for the interval $[-0.95, 0.95]$ are shown in Table 3.5.

Problem 5. Finally, let us take a look on the behavior of the method for growing matrix size n, ranging over several orders of magnitude. The results are shown in Table 3.6. For that purpose we take matrices that occur in graphene modeling [28] which can easily be constructed in different sizes. Different matrix sizes represent different sizes of the underlying physical grid. Matrices of size 4,200, 84,000, 176,000, 840,000 and finally 1,050,000 were constructed with a tool by Andreas Pieper [82]. They are very sparse, having only about a fraction of 10^{-5} of the entries nonzero (i. e., $\mathsf{nnz}(\mathsf{A})/n^2 \approx 10^{-5}$). We did not know the spectrum of the larger matrices a priori, while that one of the matrices up to 176k could be computed with a direct method (we only computed the eigenvalues directly, not the eigenvectors). The full spectrum of the 176k-matrix is shown in Figure 3.19. The structure of the spectrum of this matrix was similar to that one of smaller matrices of the same class. In particular, it was contained in $[-3.5, 3]$. For

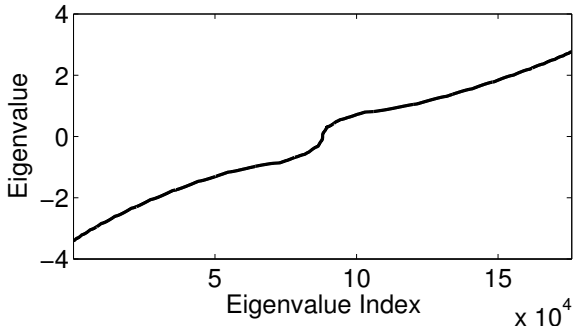

Figure 3.19: Eigenvalues of the 176k graphene matrix.

the first 3 problems, we conducted the following experiment. First, it was reasonable to force the spectrum into $[-1, 1]$ by multiplying the matrix with $1/3.5$, since its eigenvalues are contained in $[-3.5, 3]$. We then sought for eigenvalues in $I_\lambda = [-0.999, \overline{\lambda}]$. The upper bound $\overline{\lambda}$ was chosen such that some hundred eigenvalues should reside in I_λ. The size of the search space was controlled with the methods from Section 3.2. As long as possible, meaning for the 3 smaller problems, a polynomial of degree $N = 250$ was sufficient in order to bring all residuals below 10^{-12}.

For the larger problem with $n = 840{,}000$ we chose the parameters slightly different. A tolerance for the per-eigenpair residuals of 1.8×10^{-10} was computed from the input parameters, see Section 3.6.4. Here, a polynomial degree of 1500 was necessary in order to reach convergence within 10 FEAST iterations. Around the upper interval boundary $\overline{\lambda}$ there was a gap of order 10^{-6} to the closest eigenvalues of A, resulting in a low convergence rate. This explains the higher number of iterations. For $n = 1{,}050{,}000$, we considered 10^{-8} as small enough for the residual to flag an eigenpair as converged. With $N = 1500$, only 299 eigenpairs converged. The degree $N = 2000$ was high enough for 629 eigenpairs, which was also the estimated eigenvalue count in the $N = 1500$ case. The poor level of orthogonality for the largest example is caused by eigenpair locking. This means that converged eigenpairs are removed from the current computation. The rest of the eigenpairs then effectively is computed independently from those already converged, cf. Sections 3.6.3, 3.6.4. This poor level of orthogonality can be avoided by switching off eigenpair locking. A similar effect occurs in the Lanczos algorithm, see [80, Sec. 13.6]. ◊

Experiment 3.12
In this experiment we further study the interplay between the polynomial order N and the achievable accuracy, while measuring iteration counts. The aim of the experiment is to see from which polynomial degree on we can expect convergence of the eigenpairs. We took the same matrix T_{2003} as in Experiment 3.11/Problem 1,

n	4,200	84,000	176,000
Computed Eigenpairs	253	327	679
I_λ	$[-0.999, -0.8]$	$[-0.999, -0.963]$	$[-0.999, -0.963]$
Results	5 iterations orth $= 10^{-15}$ res $= 10^{-15}$	6 iterations orth $= 10^{-13}$ res $= 10^{-14}$	6 iterations orth $= 10^{-13}$ res $= 10^{-14}$

n	840,000	1,050,000
Computed Eigenpairs	499	629
I_λ	$[-0.972, -0.9712]$	$[-0.973, -0.9712]$
Results	9 iterations orth $= 10^{-12}$ res $= 10^{-12}$	10 iterations orth $= 10^{-8}$ res $= 10^{-10}$

Table 3.6: Results for Problem 5., Experiment 3.11. All test runs were performed using the Jackson kernel.

N	50	100	150	200	250	300	350	400	450	500
$I_\lambda = [\lambda_{999}, \lambda_{1269}]$	10	10	10	7	5	5	5	5	5	4
$I_\lambda = [\lambda_{1704}, \lambda_{2003}]$	10	10	10	10	6	5	5	5	5	5

Table 3.7: Iteration counts for Experiment 3.12. For every N the iteration count is specified for the two different intervals.

and switched on the Jackson kernel. We allowed 10 iterations to the FEAST algorithm. Then, we let N range from 50 to 500 with steps of 50 for different intervals. The results, showing the best achievable residual and the number of eigenpairs that have converged, are shown in Figure 3.20. The measured orthogonality was across the board of order $10^{-15} - 10^{-14}$. This is, to some extent, ensured by the design of the algorithm, cf. the discussion in Section 3.6.3. It is noteworthy that already for $N = 200$ to 250 all eigenpairs converged to small residuals (a fact that can not be seen from the figures). The number of FEAST iterations is shown in Table 3.7. It can be seen that the iteration count does not decrease significantly beyond $N = 250$. ◊

Discussion of Results

We applied the Chebyshev-FEAST method to different test problems, most of them from actual applications from science and engineering, where the matrix size n ranged from several thousand to 1,050,000. The method was able to deliver

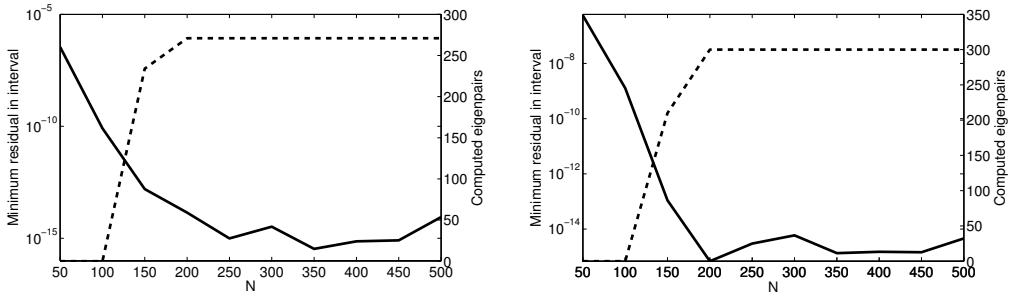

Figure 3.20: Results for Experiment 3.12. The dashed line represents the number of eigenpairs that were computed in at most 10 iterations, the solid line represents the minimum residual measured inside the interval. Left plot: $I_\lambda = [\lambda_{999}, \lambda_{1269}]$. Right plot: $I_\lambda = [\lambda_{1704}, \lambda_{2003}]$.

accurate results for almost all tested matrices. In one case (Problem 3) with $N = 250$ it failed, surprisingly with the Jackson kernel, which typically yielded better results. In this case, the failure can be explained by a lower rate of convergence for the Jackson kernel, cf. Section 3.4.6 below.

Ultimately, we tested larger problems from graphene modeling. Until $n = 176{,}000$ a polynomial degree of $N = 250$ was sufficient. For the largest example we needed $N = 2000$, which can be explained by the convergence rate again. The eigenvalues of such large matrices are typically very close to each other in absolute sense by the nature of the problem. Higher accuracy is also necessary when using another approximation method for χ_{I_λ}, as Gauß–Legendre in the standard FEAST method.

It is notable that all experiments were performed on rather small machines with an experimental MATLAB [106] code. However, the computations took only a few hours for the smaller matrices (up to 176k), in the 840k-example 66 hours and in the 1.05M-example about 100 hours. Using the numerical integration approach (or any other approach involving the solution of linear systems with the large matrix) would have taken much longer on the computing environment used, cf. the discussion in Section 3.6.1. Possibly, it even would have been impossible due to convergence or memory issues of the linear solver (e. g., if GMRES [89] was used).

In our experiments we used the techniques for eigenvalue counting from Section 3.2, which showed to be very reliable when used "in action".

3.4.6 Connection of polynomial degree and convergence rate

In this section, we analyze the connection between the polynomial order of the approximating polynomial used, the structure of the spectrum of the matrix and the convergence rate. First, recall Section 3.2.3. The convergence rate is determined

by the number

$$\hat{u} = \frac{\max_{\lambda\in(-\infty,l_0^-)\cup(l_0^+,+\infty)} |S(\lambda)|}{\min_{\lambda\in(l_1^-,l_1^+)} |S(\lambda)|}, \tag{3.29}$$

where $l_0^- < \underline{\lambda} < l_1^- < l_1^+ < \overline{\lambda} < l_0+$ and no eigenvalue of (A, B) resides in the exclusion intervals (l_0^-, l_1^-) and (l_1^+, l_0^+) [111]. The symbol S denotes the selection function, here it is of course the Chebyshev approximation polynomial, in the following denoted C, independently of the polynomial degree N.

What does C have to fulfill so that we can expect a good convergence behavior of the surrounding eigenvalue algorithm? It has to feature that l_0^-, l_1^- are as close as possible to $\underline{\lambda}$, the numbers l_0^+, l_1^+ are as close as possible to $\overline{\lambda}$ and $\hat{u}$ in (3.29) is still small, i. e., $\hat{u} \ll 1$. If we would choose a rational function for S, we could choose it in an optimal way, see [111]. Here, we are dealing with certain polynomials, hence what we can do is to determine the regions where C is "steep". The steeper the polynomial C is near $\underline{\lambda}$, $\overline{\lambda}$, the smaller the regions (l_0^-, l_1^-) and (l_1^+, l_0^+) will be. To identify these regions we start by analyzing the derivative of C.

Derivative of C

Let us compute the derivative of C in order to analyze its behavior around $\underline{\lambda}$, $\overline{\lambda}$. We have

$$C(\lambda) = \sum_{k=0}^{N} c_k T_k(\lambda)$$

and hence

$$C'(\lambda) = \sum_{k=0}^{N} c_k T_k'(\lambda).$$

The derivative of a Chebyshev polynomial T_k is

$$T_k'(\lambda) = \frac{k \sin(k \arccos(\lambda))}{\sqrt{1-\lambda^2}},\ \lambda \in \left(\underline{\lambda}, \overline{\lambda}\right),$$

we hence obtain (note that $T_0 \equiv 1$, $T_0' = 0$)

$$\begin{aligned} C'(\lambda) &= \sum_{k=0}^{N} c_k T_k'(\lambda) \\ &= \sum_{k=0}^{N} c_k \frac{k \sin(k \arccos(\lambda))}{\sqrt{1-\lambda^2}} \\ &= \sum_{k=1}^{N} c_k \frac{k \sin(k \arccos(\lambda))}{\sqrt{1-\lambda^2}}. \end{aligned}$$

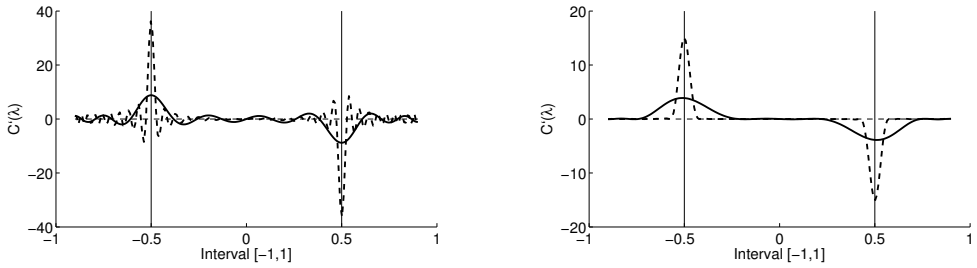

Figure 3.21: Derivatives C'. The left plot shows the derivatives obtained with the Dirichlet kernel, the right plot with the Jackson kernel. In each plot, the solid line stands for $N = 25$, the dashed line for $N = 100$.

The derivative C' is plotted in Figure 3.21 for different configurations. The plots for the approximation with the Dirichlet kernel $g_k = 1$ show oscillations, as expected. Another interesting fact is that C' as obtained with this kernel has a much higher magnitude than that one computed with the Jackson kernel. In our figure we show the derivatives only for $N = 25, 100$ because the general shape of the derivatives is better viewed for such small values of N. For $N = 500$ and using the Dirichlet kernel, C' attained an absolute value of about 183 at the interval boundaries.

Further, the plots suggest that C' attains a local maximum at $\underline{\lambda}$ and a local minimum at $\overline{\lambda}$, where $\underline{\lambda} = -0.5$, $\overline{\lambda} = 0.5$ in the plot. This maximum/minimum is also global on $[-1, 1]$. The second derivative of C is

$$C''(\lambda) = \sum_{k=2}^{N} c_k \left(\frac{k^2 \cos\left(k \arccos(\lambda)\right)}{\lambda^2 - 1} + \frac{\lambda k \sin\left(k \arccos(\lambda)\right)}{\left(1 - \lambda^2\right)^{3/2}} \right),$$

and indeed, by using MATHEMATICA [120] one can show that $C''(\underline{\lambda}) = C''(\overline{\lambda}) = 0$ for the case $\underline{\lambda} = -0.5$, $\overline{\lambda} = 0.5$. Further it can easily be seen using MATHEMATICA (or other computer algebra systems) that the third derivative of C is nonzero at $\underline{\lambda}$, $\overline{\lambda}$, showing that the derivative actually attains the extrema that can be seen in the figure. For values $\underline{\lambda} \neq -0.5$ or $\overline{\lambda} \neq 0.5$, the derivative C' in general does not attain extrema at $\underline{\lambda}$, $\overline{\lambda}$. However, the extrema are still reached near the interval boundaries, in other words, at $\underline{\lambda}$, $\overline{\lambda}$ the derivative C' is close to its maximum and minimum value, respectively. This behavior was observed in numerical experiments.

Together, these facts can be used to design a heuristic for the computation of exclusion intervals and convergence rates.

Computation of exclusion intervals

Approximate exclusion intervals (l_0^-, l_1^-) and (l_1^+, l_0^+) that are necessary to ensure a good convergence rate now can be computed by means of the derivative C'.

Let $\xi_1 = C'(\underline{\lambda})$, $\xi_2 = C'(\overline{\lambda})$ and let ℓ_1, ℓ_2 denote the tangents to $C(\lambda)$ at $\underline{\lambda}$ and $\overline{\lambda}$ respectively,

$$\begin{aligned} \ell_1(\lambda) &= \xi_1(\lambda - \underline{\lambda}) + C(\underline{\lambda}), \\ \ell_2(\lambda) &= \xi_2(\lambda - \overline{\lambda}) + C(\overline{\lambda}). \end{aligned}$$

Next, let us suppose for simplicity that C' takes extrema at $\underline{\lambda}$, $\overline{\lambda}$, i. e., that C has inflection points at these values, and that $C(\underline{\lambda}) = C(\overline{\lambda}) = 0.5$. Since ℓ_1, ℓ_2 are linearizations of C at the interval boundaries, we have the inequalities

$$\begin{cases} \ell_1(\lambda) \geq C(\lambda) & : \lambda \in I_\lambda \\ \ell_2(\lambda) \geq C(\lambda) & : \lambda \in I_\lambda \\ \ell_1(\lambda) \leq C(\lambda) & : \lambda \leq \underline{\lambda} \\ \ell_2(\lambda) \leq C(\lambda) & : \lambda \geq \overline{\lambda} \end{cases}. \tag{3.30}$$

In general, we can at least expect that the inequalities (3.30) are "approximately" true, since C' takes its extrema "near" $\underline{\lambda}$, $\overline{\lambda}$. The situation at $\underline{\lambda} = -0.5$ (for $\overline{\lambda} = 0.5$) is depicted in Figure 3.22. It is reasonable to set the exclusion intervals symmetrically around $\underline{\lambda}$, $\overline{\lambda}$, i. e.,

$$\begin{aligned} (l_0^-, l_1^-) &= (\underline{\lambda} - \delta_-, \underline{\lambda} + \delta_-), \\ (l_1^+, l_0^+) &= (\overline{\lambda} - \delta_+, \overline{\lambda} + \delta_+) \end{aligned}$$

for some numbers $\delta_-, \delta_+ > 0$. These numbers are computed from

$$\begin{aligned} 1 = 0.5 + \xi_1\delta_- &\iff \delta_- = 0.5/\xi_1, \\ 1 = 0.5 - \xi_2\delta_+ &\iff \delta_+ = -0.5/\xi_2, \end{aligned}$$

where the minus sign for δ_+ results from the fact that ξ_2 is negative. This computation yields numbers $l_0^-, l_1^-, l_1^+, l_0^+$ such that we have good reasons to expect

$$\begin{aligned} C(\lambda) &\approx 0 : \lambda \in (-\infty, l_0^-) \cup (l_0^+, +\infty), \\ C(\lambda) &\approx 1 : \lambda \in (l_1^-, l_1^+). \end{aligned}$$

The number δ_-, computed as above, is shown in the left plot of Figure 3.23 for different values of N and the Dirichlet as well as the Jackson kernel.

Analyzing the functional connection between δ and N can be done by plotting $\delta = \delta(N)$ on the double log scale in the first place, see Figure 3.24. The plot suggests that there is a $\delta = cN^\alpha$ connection. Then, by considering the equations

$$\begin{aligned} \delta_1(N_1) &= cN_1^\alpha \\ \delta_2(N_2) &= cN_2^\alpha \end{aligned}$$

for two distinct pairs (N_1, δ_1), (N_2, δ_2) we obtain

$$\alpha = \frac{\log\left(\frac{\delta_1}{\delta_2}\right)}{\log\left(\frac{N_1}{N_2}\right)}. \tag{3.31}$$

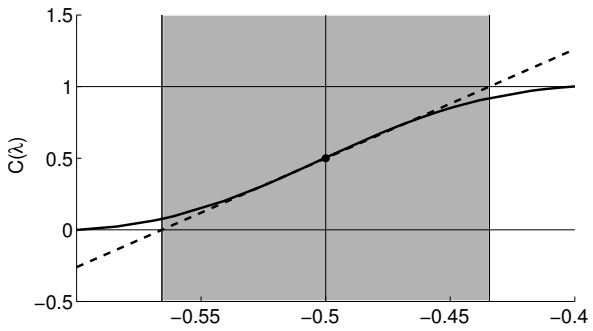

Figure 3.22: The approximating polynomial C for $N = 50$ with Jackson kernel. The dashed line is the tangent to $C(\lambda)$ at $\underline{\lambda} = -0.5$. The shaded region designates the interval (l_0^-, l_1^-).

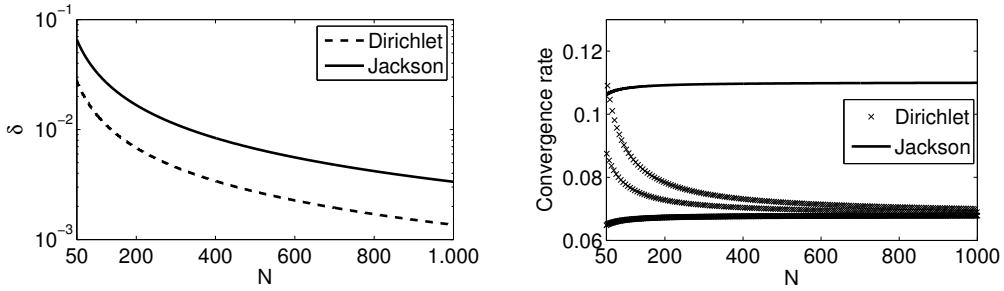

Figure 3.23: Left: Radius δ_- of (lower) exclusion interval $(l_0^-, l_1^-) = (\underline{\lambda} - \delta_-, \underline{\lambda} + \delta_-)$ for $\underline{\lambda} = -0.5$. Right: resulting convergence rate.

By evaluating (3.31) for different pairs (N_1, δ_1), (N_2, δ_2), we always obtained $\alpha \approx -1$, no matter if the Dirichlet or Jackson kernel was used and how $\underline{\lambda}$, $\overline{\lambda}$ were chosen. The constant c was always of order 1.

Now, we can approach another interesting question: how large does the polynomial degree N have to be chosen in order to ensure a low convergence rate in case of a known spectrum? If the distance δ from the interval boundaries to the next eigenvalue is known, we have $N = c/\delta$. This means, e. g., in the (academic) case of equidistant eigenvalues between -1 and 1 and matrix size n that we have $N = cn - c$ since $\delta = 1/(n-1)$. This is true in this case independently of how the interval boundaries are chosen, as long as they are in the middle of two eigenvalues. Note that this very high polynomial degree is only necessary in order to obtain a good (i. e., close to zero) convergence rate for a given eigenvalue distribution. This distribution is in general unknown, anyway. Note that a similar observation concerning the connection between polynomial degree and what we call exclusion intervals has been made in [93].

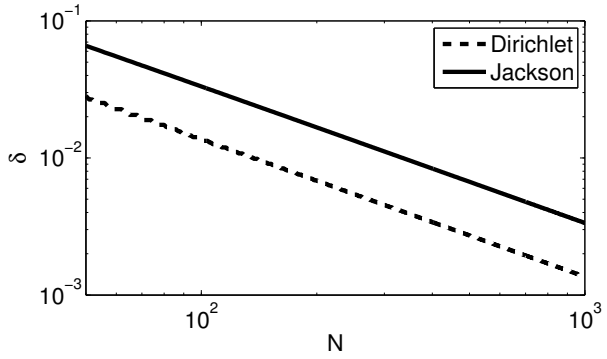

Figure 3.24: Radius of exclusion interval (see left plot in Figure 3.23) on the double log scale.

Convergence rate

The convergence rate $\hat{u}$, cf. (3.29) can be estimated by

$$\widetilde{u} := \max\left\{\left|C(l_0^-)/C(l_1^-)\right|, \left|C(l_0^+)/C(l_1^+)\right|\right\}, \tag{3.32}$$

where the exclusion intervals were computed via the linear approximations ℓ_1, ℓ_2 on C as described above. Formula (3.32) can be motivated by Figure 3.22. We have $\widetilde{u} \leq \hat{u}$.

The resulting approximate convergence rates $\widetilde{u}$, computed via (3.32) for different values of N are shown in the right plot of Figure 3.23. Note that in this figure, the "curve" for the Dirichlet kernel looks like three curves but actually we are dealing with one oscillating curve. The oscillations in the convergence rate clearly result from the different oscillations of the polynomial C for different polynomial degrees. This also leads to oscillations of the derivative C'. Taking a closer look on the right plot of Figure 3.23 reveals that the "higher level" convergence rates (starting at about 0.11 for $N = 50$) are attained for polynomial degrees $N = 53+6k$, $k > 0$. Those on the "medium level" (starting at about 0.08 for $N = 50$) are attained for polynomial degrees $N = 51 + 6k$, $k > 0$. Besides, it can be seen that the convergence rate approaches one single value for growing N.

For certain other radii $\delta = \delta_- = \delta_+$ of the exclusion intervals the resulting convergence rates are shown in Figure 3.25. Therein, δ was chosen from 0.05 to 0.45 for $I_\lambda = [-0.5, 0.5]$. Note that such values of δ are rather academic and that the "natural" choice of δ, determined by the slope of C, would be much lower.

In order to avoid too high polynomial degrees, an adaptive approach can be chosen, taking N small in the first place and increasing it later, see the next Section 3.4.7.

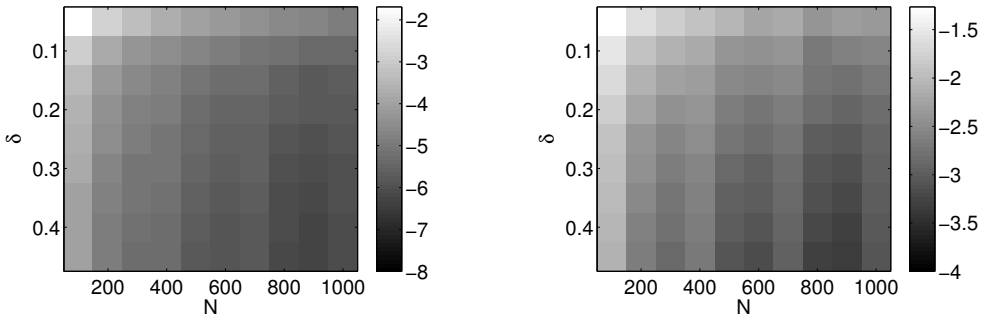

Figure 3.25: Left: Convergence rates $\widetilde{u}$ for the Jackson kernel. Right: Convergence rates $\widetilde{u}$ for the Dirichlet kernel. Both on the $\log_{10}$-scale.

Summary

The convergence rate for Chebyshev-FEAST can be determined theoretically by means of the derivative of the approximating polynomial. In particular, the analysis yields an interval of exclusion that should contain no eigenvalue. The existence of such exclusion intervals (i. e., of gaps in the spectrum) will typically yield a good convergence behavior.

In order to obtain "good" (i. e., close to zero) convergence rates it can be necessary to employ polynomials of extremely high degree. In case of an equidistant eigenvalue distribution we end up with a polynomial degree $N = \mathcal{O}(n)$, where n is the matrix size. The experiments in Section 3.4.5 showed that such high polynomial degrees are not necessary in practice.

3.4.7 Adaptive choice of polynomial degree

In the previous sections about numerical experiments (Section 3.4.5) and convergence rate (Section 3.4.6), we saw that the necessary polynomial degree depends on the distribution of eigenvalues. In many applications, not much is known about this distribution. One hence might require an adaptive control of the polynomial degree at runtime.

There are two possibilities to introduce adaptivity. The first one is straight forward; if the Chebyshev-FEAST method does not converge fast enough, the degree of the polynomial is increased *between* two iterations of the algorithm. The other possibility is to increase the polynomial degree *during* one iteration of the eigensolver. If it is observed after the Rayleigh–Ritz step of the eigenvalue algorithm that the achieved accuracy is too low, the degree of the polynomial is increased. This technique amounts to reusing the already computed polynomial. A new degree $N' > N$ is chosen and the remaining $N' - N$ summands are computed.

Increasing the degree of the approximating polynomial is possible if it is a

sum of Chebyshev polynomials weighted by coefficients that do not depend on the order N. This is basically the polynomial with Dirichlet kernel,

$$C_N(\lambda) = \sum_{k=0}^{N} c_k g_k T_k(\lambda), \tag{3.33}$$

where $g_k = 1$ and the subscript N denotes the degree. The Jackson kernel, which typically showed better behavior, is not applicable for this type of adaptivity since every summand of C_N depends on N. This dependency cannot be disentangled in a reasonable way, since all but two summands of (3.33) are not stored (being precise, none of the summands is stored at any time, but the result of its multiplication with a rectangular matrix Y).

For the Fejér kernel $g_k = 1 - k/N$, however, it is possible to increase the degree. We have

$$\begin{aligned} C_N^F(\lambda) &:= \sum_{k=0}^{N} c_k(1 - k/N)T_k(\lambda) \\ &= \sum_{k=0}^{N} c_k T_k(\lambda) - \frac{1}{N}\sum_{k=0}^{N} c_k k T_k(\lambda). \end{aligned}$$

If both

$$C_N^{F,1} := \sum_{k=0}^{N} c_k T_k(\lambda) \text{ and } C_N^{F,2} := \sum_{k=0}^{N} c_k k T_k(\lambda)$$

are stored, we can form $C_{N'}^F$ by updating $C_N^{F,1}$, $C_N^{F,2}$ in the usual way. The additional cost compared to using the Dirichlet kernel is essentially the storage for one $n \times \widetilde{m}$-matrix. Further, another addition of such matrices is necessary for $k = 0, \ldots, N$.

Adaptivity between iterations

Let us discuss the adaptivity that takes place between FEAST iterations in more detail. At a glance, it seems to be preferable, since the possible savings of the other type are ruined by the fact that the Jackson kernel cannot be used. This typically leads to higher iteration counts.

The typical effects that occur when controlling N adaptively are best illustrated with a numerical example. The operation count of the algorithm is roughly proportional to the number of matrix-vector multiplications. This number is hence a good measure to assess the effectivity of the adaptive choice of N. In formulas, it is given by the number

$$M = \sum_{j=1}^{\mathsf{maxit}} \widetilde{m}_j N_j,$$

N	32	50	100	250
M	134k	150k	256k	530k
Iter.	10	7	6	5
time/s	12.36	13.93	24.18	51.26

Table 3.8: Results for Experiment 3.13 without dynamics in N.

N_1	25	30	32	50
M	125k	118k	124k	172k
Iter.	9	7	7	6
time/s	11.78	10.93	11.68	16.49

Table 3.9: Results for Experiment 3.13 with dynamics in N.

where $\widetilde{m}_j$ denotes the size of the search space in iteration j and N_j denotes the polynomial degree in iteration j. The sum ranges up to maxit, the number of FEAST iterations allowed to be performed.

Experiment 3.13
We repeated the numerical test with the matrix of size $n = 4{,}200$ from graphene modeling, already used in Experiment 3.11. We were looking explicitly for the 300 eigenpairs with smallest eigenvalues, employing a search space with initial dimension $\widetilde{m}_1 = 500$, where dynamic choice of search space size was activated at iteration $j = 3$. The smallest value of N where all eigenpairs converged within 10 iterations was 32. In Table 3.8 we list some values of N and M (rounded to the nearest thousand), accompanied by the runtime of the Chebyshev routine listed in Algorithm 3.2 and the iteration count. The time as well as M do not grow linearly with N, since eigenpairs converge at different iterations. In particular, for larger values of N, the algorithm needs a lower total iteration count (while consuming more time).

Next, let us introduce dynamics in N. At a certain iteration j we have to decide whether residuals are small enough or not. This has do be done based on heuristics. In this test, we used the following procedure,

$$\begin{aligned} &\text{if} \quad j = 4 \wedge \max_{i=1,\dots,\widetilde{m}_j} \mathsf{res}_i > 10^{-6} \\ &\quad \text{set } N_5 = \lceil 1.5 \times N_4 \rceil \,, \end{aligned} \tag{3.34}$$

where $\mathsf{res}_i = \left\| \mathsf{A}\widetilde{\mathsf{x}}_i - \widetilde{\mathsf{x}}_i\widetilde{\lambda}_i \right\|$ denotes the residual with index i. Note that the procedure is applied only once, between the 4th and 5th iteration. In principle, the adaption of N is possible at every iteration. The results after applying procedure (3.34) are stated for some initial values N_1 in Table 3.9.

Interestingly, in the case of dynamically chosen N, convergence of all desired pairs could be reached even for $N_1 = 25$, while the overall number of matrix-vector multiplications compared to fixed N was decreased. If $N = 25$ was fixed over all iterations, only a few eigenvalues converged. For $N = 30$ (fixed), also not all eigenpairs converged. ♢

The experiment shows that a criterion/procedure as (3.34) can be used to improve convergence if the initial value of N was chosen too small. Using procedure (3.34) might even decrease the overall runtime. For instance, compare the

case $N = 32$ without dynamics to the case $N_1 = 25$ with dynamics. The values in (3.34) are based on experience, in particular the threshold 10^{-6} should be adapted for other problems. A threshold similar to that of the stopping criterion (cf. Section 3.6.4) can be used, in particular the matrix size has to be considered. This leads to a criterion

$$\max_{i=1,\ldots,\widetilde{m}} \mathsf{res}_i > \mathsf{tol} \cdot n \cdot \max\left\{ \left|\underline{\lambda}\right|, \left|\overline{\lambda}\right| \right\},$$

where tol is about 7 orders of magnitude higher than the value of tol used in Section 3.6.4. Note that this is only a heuristic. However, increasing the polynomial degree will rarely deliver less accurate results (we cannot report any case).

As mentioned before, the polynomial degree can also be increased between two arbitrary consecutive iterations of FEAST. For instance, the test (3.34) can be performed every second iteration. Further, the factor for increasing the degree can be chosen other than 1.5. It might be necessary to also adjust maxit, because the convergence could be too slow in the first FEAST iterations. Therefore not all desired eigenpairs might converge within maxit iterations.

3.4.8 Generalized problem

So far, the discussion about the FEAST algorithm with Chebyshev approximation only included one matrix A, i. e., a standard eigenvalue equation $\mathsf{Ax} = \mathsf{x}\lambda$. Since the generalized equation $\mathsf{Ax} = \mathsf{Bx}\lambda$ for Hermitian positive definite B is the actual topic of this work, we will briefly discuss its solution by Chebyshev approximation methods here.

At first glance there seems to be no way to treat the standard and the generalized equation in a unified way. This is possible for the solution of eigenvalue problems with numerical integration. However, there is some remedy to still use the polynomial approximation approach. The obvious way is to compute $\hat{\mathsf{A}} = \mathsf{B}^{-1}\mathsf{A}$ and then to proceed as before. This way was explained to be risky in Chapter 1, at least if $\kappa(\mathsf{B})$ is large. In case of Hermitian positive definite B with small condition number, one might attempt to use $\mathsf{B}^{-1}\mathsf{A}$ as follows. In the Chebyshev approximation, the matrix $\hat{\mathsf{A}}T_k(\hat{\mathsf{A}})\mathsf{Y}$ is needed for the computation of $T_{k+1}(\hat{\mathsf{A}})\mathsf{Y}$. Assuming $T_k(\hat{\mathsf{A}})\mathsf{Y}$ is available, first $\mathsf{S} := \mathsf{A}T_k(\hat{\mathsf{A}})\mathsf{Y}$ is formed. Then, the hpd system $\mathsf{BW} = \mathsf{S}$ is solved for $\mathsf{W} = \hat{\mathsf{A}}T_k(\hat{\mathsf{A}})\mathsf{Y}$. Since B is hpd, the first choice for a sparse linear solver is typically the conjugate gradients (cg) method [37,43]. It is known as a reliable and fast linear solver. Solving linear systems with B is a much simpler task than solving systems with the matrix $z\mathsf{B} - \mathsf{A}$, cf. Section 3.6.1.

If a factorization $\mathsf{K}^\star\mathsf{K} = \mathsf{B}$ is available, it can be used and the product $(\mathsf{K}^\star)^{-1}\mathsf{A}\mathsf{K}^{-1}$ can be applied factor by factor. The factorization has to be computed only once. This is in contrast to the integration based version of FEAST, where for each integration point a different system has to be solved.

A more sophisticated approach would be to use a bivariate polynomial to approximate the bivariate function $f(x, y) = \chi_{I_\lambda}(y^{-1}x)$. Bivariate Chebyshev

polynomials for approximating bivariate functions are well studied, see [9]. The approximating function is of the form

$$f(x,y) \approx \sum_{j=0}^{M} \sum_{k=0}^{N} c_{jk} T_j(x) T_k(y). \tag{3.35}$$

The coefficients c_{jk} can be computed by means of function evaluations at discrete points, hence avoiding to evaluate x/y at points with $y = 0$. The computation of the coefficients c_{jk} already takes $2(M+1)(N+1)$ evaluations of scalar Chebyshev polynomials. The final evaluation of (3.35) at $x = \mathsf{A}$ and $y = \mathsf{B}$ then requires NM multiplications of full size matrices with a block of vectors. This would result in enormous computational effort, already for moderately large N and M such as $N = M = 100$ we would need over $20,000$ multiplications (2 for each summand). In first tests, polynomial degrees of this order were necessary. For these reasons, the method is not feasible in this form in practice.

In [24], recently an approach was published for counting eigenvalues of a generalized equation in a given interval. It is based on Chebyshev approximation of high-pass filters. At first glance, there is no straightforward way to adapt the method to the actual solution of generalized eigenvalue problems. The authors of [24] also note that polynomials of very high degree can be necessary.

3.4.9 Why Chebyshev? (Other polynomials)

So far we only considered Chebyshev polynomials. The question arises, why we chose just this class of polynomials. For instance, one might come up with a *Bernstein polynomial* of order N, see the introduction in Section 3.4.1. For a function $f : [0,1] \to \mathbb{R}$ it is defined as [67, p. 3]

$$B_N^f(t) = \sum_{k=0}^{N} f\left(\frac{k}{N}\right) \binom{N}{k} t^k (1-t)^{N-k}. \tag{3.36}$$

The evaluation of (3.36) at t requires $\mathcal{O}(N^2)$ multiplications with t, it is therefore not applicable to matrices in practice. Another problem arises with the large values attained by the binomial coefficients. This can lead to numerical problems. Further, the convergence of (3.36) towards f is quite slow. In Figure 3.26 the Bernstein polynomial of order 150 belonging to $\chi_{[0.2,\,0.8]}$ is plotted.

The polynomials of choice are the so called *orthogonal polynomials*. Those are polynomials $(p_k)_{k\geq 0}$ with $\deg(p_k) = k$ that are pairwise orthogonal with respect to a certain scalar product on $C[a,b]$. The scalar product is defined via

$$\langle f, g \rangle_w = \int_a^b f(t) g(t) w(t) \mathrm{d}t, \tag{3.37}$$

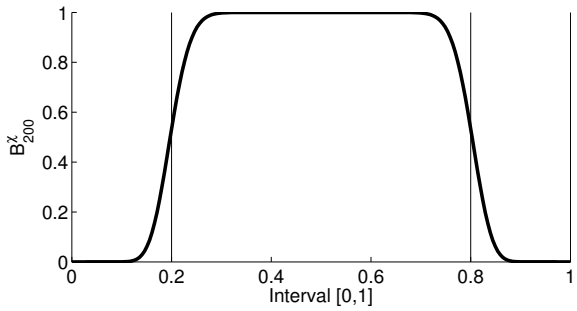

Figure 3.26: The Bernstein polynomial of order 150 belonging to $\chi_{[0.2,\,0.8]}$.

where w is a weight function with

$$0 < \int_a^b w(t)\mathrm{d}t < \infty.$$

The Chebyshev polynomials form an orthogonal system over $[-1,1]$ with respect to the scalar product (3.37) with $w(t) = (1-t^2)^{-1/2}$. For an introduction, see [30], see also Section 2.3.3. Orthogonal polynomials have in common that they are related via three term recurrences [30, p. 23], as the Chebyshev recurrence (3.24), hence evaluating a degree N polynomial requires $\mathcal{O}(N)$ multiplications with the argument.

The popularity of Chebyshev polynomials (in particular in linear algebra) is clearly due to their optimality properties. The Chebyshev polynomial T_N of degree N fulfills

$$\frac{1}{2^{N-1}}\left\|T_N\right\|_\infty = \min\left\{\|p\|_\infty : p \text{ is polynomial of degree} \le N,\, p(0)=1\right\},$$

where $\|\cdot\|_\infty$ denotes the maximum norm on $[-1,1]$ [30, Ch. 3]. However, let us discuss the *Legendre polynomials* exemplarily for other orthogonal polynomials. They are defined as the orthogonal polynomials $(P_k)_{k\ge 0}$ over $[-1,1]$ to the weight function $w \equiv 1$, scaled such that

$$\langle P_k, P_k\rangle_w = \frac{2}{2k+1}.$$

The corresponding recurrence relation is

$$P_k(t)\cdot k = (2k-1)\cdot t\cdot P_{k-1}(t) - (k-1)P_{k-2}(t),\ P_1(t)=t,\ P_0(t)=1,$$

see [1, Ch. 22]. Next, for any function $f:[-1,1]\to\mathbb{R}$ write f as the formal series (i. e., without saying anything about its convergence)

$$f(t) = \sum_{k=0}^{\infty} c_k P_k(t). \tag{3.38}$$

Multiplying both sides of (3.38) by $P_m(t)$ and integrating over $[-1, 1]$ yields the coefficient c_m,

$$\begin{aligned}\int_{-1}^{1} f(t)P_m(t)\mathrm{d}t &= \int_{-1}^{1} \sum_{k=0}^{\infty} c_k P_k(t)P_m(t)\mathrm{d}t \\ &= c_m \int_{-1}^{1} P_m(t)P_m(t)\mathrm{d}t \\ &= c_m \cdot \frac{2}{2m+1}.\end{aligned}$$

Thus,

$$c_m = \frac{2m+1}{2} \int_{-1}^{1} f(t)P_m(t)\mathrm{d}t.$$

This technique for the derivation of coefficients c_k can be used for other orthogonal polynomials $(p_k)_k$ in a similar fashion. The difference is that (3.38) has to be multiplied by $w(t)p_m(t)$, which we could skip since $w \equiv 1$ for Legendre polynomials. In the case of f being the characteristic function χ_{I_λ} of the search interval $I_\lambda \subset \mathbb{R}$, we obtain for the coefficients

$$\begin{aligned}c_k &= \frac{2k+1}{2} \int_{-1}^{1} \chi_{I_\lambda}(t)P_k(t)\mathrm{d}t \\ &= \frac{2k+1}{2} \int_{\underline{\lambda}}^{\overline{\lambda}} P_k(t)\mathrm{d}t \\ &= \frac{2k+1}{2} \cdot \frac{1}{2k+1} \cdot [P_{k+1}(t) - P_{k-1}(t)]_{t=\underline{\lambda}}^{\overline{\lambda}} \\ &= [P_{k+1}(t) - P_{k-1}(t)]_{t=\underline{\lambda}}^{\overline{\lambda}}. \end{aligned} \tag{3.39}$$

Here, we define $P_{-1} :\equiv 0$. For the primitive of P_k see [55, p. 500]. In the following, we will briefly study the FEAST method with Legendre approximation, i. e., we approximate the subspace U by

$$\mathsf{U} \approx L_N(\mathsf{A})\mathsf{Y} := \sum_{k=0}^{N} c_k P_k(\mathsf{A})\mathsf{Y}, \tag{3.40}$$

where c_k are the coefficients computed according to (3.39). In Figure 3.27 the function L_N is shown for several values of N and the interval $I_\lambda = [-0.5, 0.5]$. In Figure 3.28 we show the function L_{1000} on the region of interest around $\underline{\lambda}$. What follows is an experiment involving the subspace U computed according to (3.40).

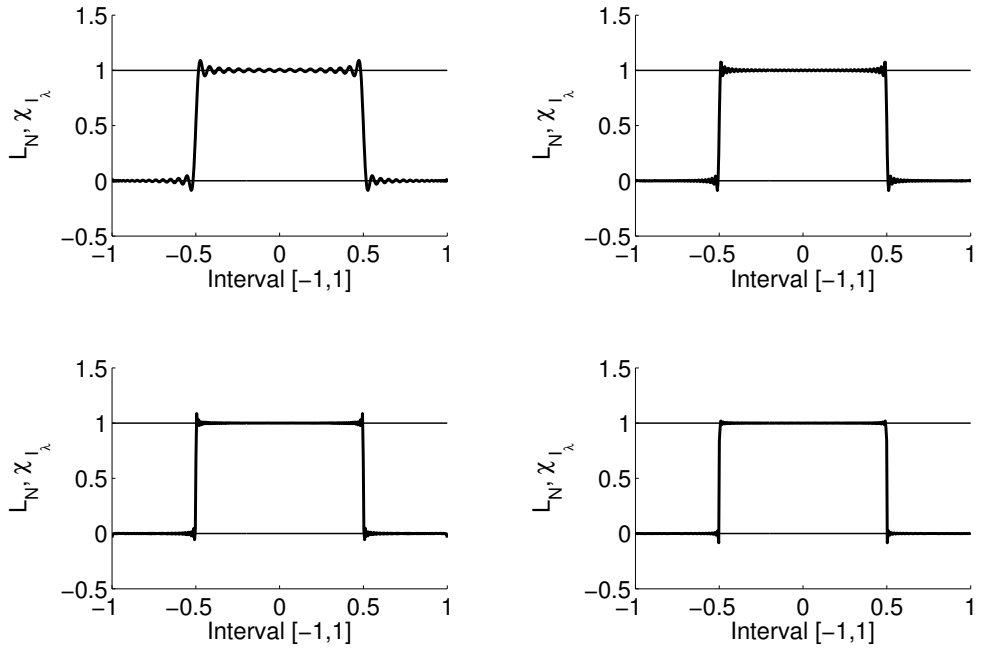

Figure 3.27: Legendre approximation of χ_{I_λ}. The polynomial degree is (top left, top right, bottom left, bottom right): 100, 250, 500, 1000.

Experiment 3.14
Let us repeat Experiment 3.11, Problem 1, the results of which are shown in Table 3.2. We perform exactly the same computation but with Legendre instead of Chebyshev polynomials. The results are shown in Table 3.10. In some cases the number of converged eigenpairs is higher than the number of eigenvalues in I_λ since the actual interval of computation was chosen slightly larger. Since in this experiment the same setting was used as in Experiment 3.11, Problem 1, we can compare the results to those obtained with Chebyshev polynomials and the Jackson kernel. We see that, in terms of iteration counts and convergence of all eigenpairs, the results are worse when Legendre polynomials are used. The numerical quality is only slightly lower. ◊

Conclusion

Other polynomials than Chebyshev can be used, although the Chebyshev polynomials enjoy optimality properties. Here, only orthogonal polynomials should be considered. For a class of polynomials $(p_k)_k$, orthogonal with respect to a function w, the coefficients of the respective series can be computed via

$$c_k = \langle p_k, p_k \rangle_w^{-1} \int_{\underline{\lambda}}^{\overline{\lambda}} p_k(t) w(t) \mathrm{d}t.$$

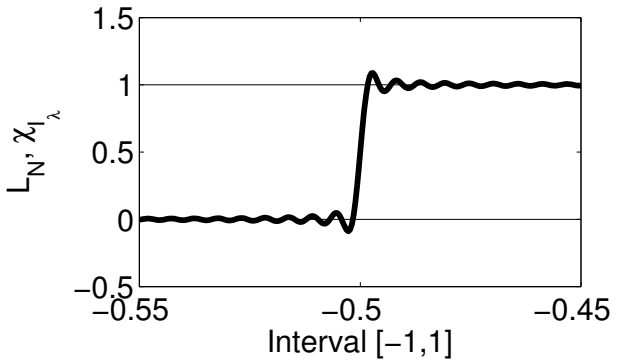

Figure 3.28: The function L_{1000} on the region around $\underline{\lambda}$.

I_λ # eigenpairs	$N = 500$	$N = 1000$	$N = 1500$
$[\lambda_{999}, \lambda_{1296}]$ 298 eigenpairs	9 iterations 298 eigenpairs orth $= 10^{-15}$ res $= 10^{-13}$	10 iterations 185 eigenpairs orth $= 10^{-15}$ res $= 10^{-14}$	8 iterations 298 eigenpairs orth $= 10^{-13}$ res $= 10^{-13}$
$[\lambda_{1704}, \lambda_{2003}]$ 300 eigenpairs	10 iterations 154 eigenpairs orth $= 10^{-13}$ res $= 10^{-13}$	7 iterations 303 eigenpairs orth $= 10^{-15}$ res $= 10^{-13}$	8 iterations 303 eigenpairs orth $= 10^{-13}$ res $= 10^{-13}$
$[\lambda_{1687}, \lambda_{2003}]$ 316 eigenpairs	10 iterations 168 eigenpairs orth $= 10^{-13}$ res $= 10^{-13}$	7 iterations 317 eigenpairs orth $= 10^{-14}$ res $= 10^{-13}$	8 iterations 317 eigenpairs orth $= 10^{-13}$ res $= 10^{-13}$

Table 3.10: Results for Legendre-FEAST, Experiment 3.14. We also give the number of eigenvalues inside I_λ as well as the number of converged eigenvalues.

For the special case of the Legendre polynomials we performed exactly the same test as with Chebyshev polynomials, resulting in much worse results. This suggests that other orthogonal polynomials will also show a poor performance and encourages us to stay with Chebyshev polynomials.

3.5 Transforming the integration region

So far, we considered the FEAST algorithm, based on the contour integral

$$\mathsf{U} = \frac{1}{2\pi \mathbf{i}} \int_{\mathcal{C}} (z\mathsf{B} - \mathsf{A})^{-1} \mathsf{B} \mathrm{d}z \mathsf{Y}, \tag{3.41}$$

where $\mathcal{C}$ is a curve encircling the eigenvalues in a certain search interval I_λ. All methods for computing (3.41)—whether based on integration or approximation—have in common that the numerical performance depends on the number $d(\mathcal{C}) := \min_{\lambda \in \mathrm{spec}(\mathsf{A},\,\mathsf{B})} \mathrm{dist}(\lambda, \mathcal{C})$.

3.5.1 Use of integral transformation

Is it possible to transform the region of integration such that the close passing of $\mathcal{C}$ to the spectrum of $(\mathsf{A},\, \mathsf{B})$ can be avoided, or that other improvements can be achieved? In the following, we will show that this is possible, however, so far only in very special cases. We adapt a method by Hale, Higham and Trefethen [41] for this purpose. Their method computes the values of analytic functions with matrix argument by means of contour integrals. To this purpose, the representation

$$f(\mathsf{A}) = \frac{1}{2\pi \mathbf{i}} \int_{\mathcal{C}} f(z)(z\mathsf{I} - \mathsf{A})^{-1} \mathrm{d}z \tag{3.42}$$

is used, where $\mathcal{C}$ is a simply closed curve around $\mathrm{spec}(\mathsf{A})$. This representation is a consequence of Cauchy's Theorem 2.28. In [41] it is supposed in the first place that A is real symmetric and positive definite. It is stated that $\mathcal{O}(\lambda_{\max}/\lambda_{\min})$ linear system solves are necessary to compute (3.42) if A is ill conditioned. In the method from [41], a conformal map from the annulus

$$A := \{z \in \mathbb{C} : r < |z| < R\}\,,\ r, R > 0,$$

to the doubly connected region

$$\Omega := \mathbb{C} \setminus ((-\infty, 0] \cup [\lambda_{\min}, \lambda_{\max}]) \tag{3.43}$$

is constructed, where $f \in H(\Omega, \mathbb{C})$. As usual, $\lambda_{\min}, \lambda_{\max}$ denote the smallest and largest eigenvalue of A, respectively. They may be replaced by lower and

upper bounds, as long as the lower bound is larger than zero. The curve $\mathcal{C}$ then can be chosen such that it entirely lies in Ω, encircling $[\lambda_{\min}, \lambda_{\max}]$. The transformation proposed in [41] is independent of f, it only relies on the structure of Ω. Consequently, we may adapt it to compute (3.41).

The requirement on Ω is that it consists of the slit plane $\mathbb{C} \setminus (-\infty, 0]$ where further an interval consisting of positive numbers is removed from as in (3.43). If the matrix pair (A, B) has the property that some eigenvalues, say, those contained in $I_\lambda = [\underline{\lambda}, \overline{\lambda}]$ are > 0, while the rest is less or equal than zero, we may use the method. This can be achieved by appropriate scaling and/or shifting of the matrix pair, if one is aiming to compute the eigenpairs with smallest or largest eigenvalues. Then, for $I_\lambda \subset (0, \infty)$ the region

$$\Omega = \Omega_\lambda := \mathbb{C} \setminus ((-\infty, 0] \cup I_\lambda)$$

is considered.

In [41] the integral

$$f(\mathsf{A}) = \frac{1}{2\pi\mathbf{i}} \mathsf{A} \int_{\mathcal{C}} z^{-1} f(z) (z\mathsf{I} - \mathsf{A})^{-1} \mathrm{d}z$$

is used instead of (3.42). The corresponding representation of U is also allowed in our case due to Lemma 2.47, but we will stay with (3.42) in the following.

3.5.2 Conformal transformation of integration region

Let us introduce the map from A to Ω_λ in three steps as in [41]. The introduction requires basic knowledge on elliptic functions and elliptic integrals, which will not be provided here but can be found in many books, e. g., [1, 2, 70]. The map constructed is in the first place a map

$$\{z \in A : \mathrm{Im}(z) \geq 0\} \longrightarrow \{z \in \Omega_\lambda : \mathrm{Im}(z) \geq 0\}.$$

Then, by using the Schwarz reflection principle [2, p. 170], it can be extended to a map from A to Ω_λ. The three steps of the map are as follows [41].

1. The function
$$s \mapsto t = \frac{2K\mathbf{i}}{\pi} \log(-\mathbf{i}s/r) \tag{3.44}$$
maps the upper half of the annulus A to the rectangle with vertices $-K, K, -K + \mathbf{i}K', K + \mathbf{i}K'$. The numbers K, K' are so called complete

elliptic integrals, defined by

$$K = K(m) = \int_0^{\pi/2} \frac{1}{\sqrt{1 - m\sin^2\theta}} \mathrm{d}\theta,$$

$$\mathbf{i}K' = \mathbf{i}K'(m_1) = \mathbf{i}\int_0^{\pi/2} \frac{1}{\sqrt{1 - m_1\sin^2\theta}} \mathrm{d}\theta,$$

where $m_1 = 1 - m$ and $m = k^2$ for a certain number k that is introduced below [70].

2. The Jacobi elliptic function

$$t \mapsto u = \mathrm{sn}(t) = \mathrm{sn}(t|k^2) \tag{3.45}$$

with

$$k = \frac{\sqrt{\overline{\lambda}/\underline{\lambda}} - 1}{\sqrt{\overline{\lambda}/\underline{\lambda}} + 1} < 1$$

moves the rectangle with corners $K, -K, -K + \mathbf{i}K', K + \mathbf{i}K'$ to the upper half plane, while the corners are mapped as

$$\begin{aligned} K &\mapsto 1, \\ K + \mathbf{i}K' &\mapsto k^{-1}, \\ -K &\mapsto -1, \\ -K + \mathbf{i}K' &\mapsto -k^{-1}. \end{aligned}$$

3. The last step maps the interval $[-k^{-1}, -1]$ to $[0, \underline{\lambda}]$ and $[1, k^{-1}]$ to $[\overline{\lambda}, \infty]$, while translating the upper half plane to itself. This is done by

$$u \mapsto z = \sqrt{\underline{\lambda}\overline{\lambda}} \cdot \frac{k^{-1} + u}{k^{-1} - u}.$$

An illustration of the map can be seen in Figure 3.29, see [41].

In [41], the annulus A itself is not used but rather the rectangle defined by (3.44), which is implicitly defined by A via the values of K and K'. Next, the definition of u (3.45) can be inserted into that one of z, yielding

$$z(t) = \sqrt{\underline{\lambda}\overline{\lambda}} \cdot \frac{k^{-1} + \mathrm{sn}(t)}{k^{-1} - \mathrm{sn}(t)},$$

which leads to

$$\frac{\mathrm{d}z}{\mathrm{d}t} = 2k^{-1}\sqrt{\underline{\lambda}\overline{\lambda}} \cdot \frac{\mathrm{cn}(t)\mathrm{dn}(t)}{(k^{-1} - \mathrm{sn}(t))^2}.$$

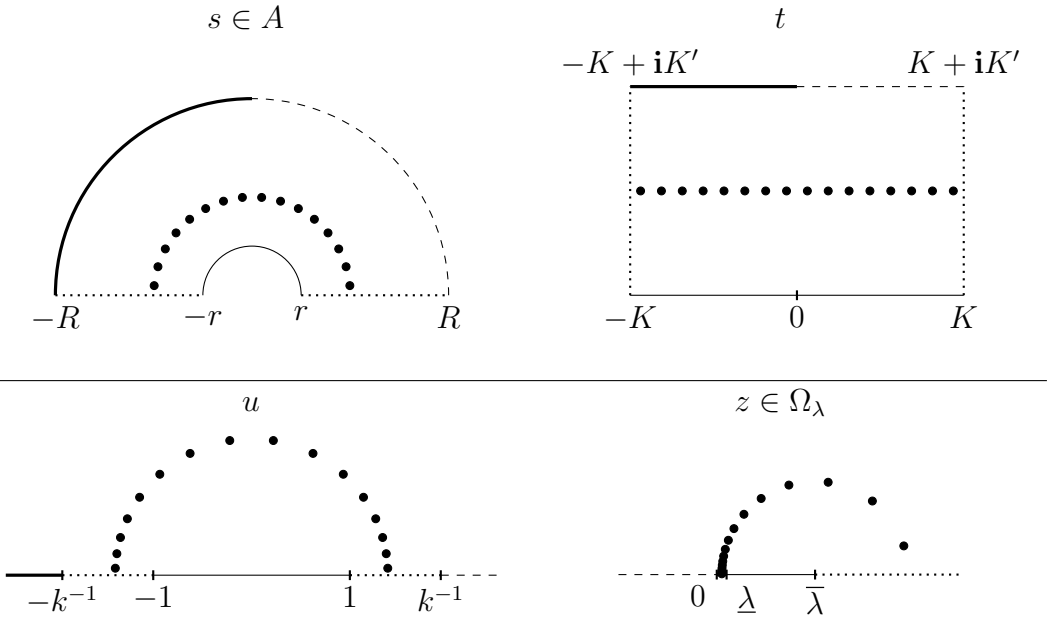

Figure 3.29: Illustration of the map $s \mapsto t \mapsto u \mapsto z$ from the annulus A (top left) to the region Ω_λ (bottom right). Here, it is shown for $\underline{\lambda} = 0.1$, $\overline{\lambda} = 1$. The dots represent the integration points. Parts that belong together, i. e., are translated by the map, are shown in the same line style. In the bottom right figure, the interval $(0, \underline{\lambda})$ belongs to the interval $(-\infty, -k^{-1})$ from the figure before. The figures have previously appeared in [41] in similar form, we only adapted the notation to ours.

Here, $\mathrm{sn}'(t) = \mathrm{cn}(t)\mathrm{dn}(t)$ and cn, dn are other types of Jacobi elliptic functions. Similar to the integral obtained in [41] we now can evaluate (3.41) as

$$\mathsf{U} = -\frac{\sqrt{\underline{\lambda}\overline{\lambda}}}{\pi \mathbf{i} k} \int_{-K+\mathbf{i}K'/2}^{3K+\mathbf{i}K'/2} \frac{\mathrm{cn}(t)\mathrm{dn}(t)}{(k^{-1} - \mathrm{sn}(t))^2} (z(t)\mathsf{B} - \mathsf{A})^{-1} \mathsf{B} \mathrm{d}t \mathsf{Y}. \tag{3.46}$$

Integrating over the complex interval $[-K + \mathbf{i}K'/2, 3K + \mathbf{i}K'/2]$ corresponds to integrating over a full circle inside A, which is traversed through in negative mathematical sense, this explains the minus sign in (3.46). Similar to the direct application of an integration rule [85], the integration over the lower part of the annulus A can be avoided for real symmetric matrices A and B. Then, the integration interval is restricted to $[-K + \mathbf{i}K'/2, K + \mathbf{i}K'/2]$, resulting in

$$\mathsf{U} = -\frac{2\sqrt{\underline{\lambda}\overline{\lambda}}}{\pi \mathbf{i} k} \,\mathrm{Im} \int_{-K+\mathbf{i}K'/2}^{K+\mathbf{i}K'/2} \frac{\mathrm{cn}(t)\mathrm{dn}(t)}{(k^{-1} - \mathrm{sn}(t))^2} (z(t)\mathsf{B} - \mathsf{A})^{-1} \mathsf{B} \mathrm{d}t \mathsf{Y}. \tag{3.47}$$

Ultimately, the authors of [41] suggest the use of a p-point midpoint rule $(t_j, \omega_j)_{j=1,\ldots,p}$ with

$$t_j = -K + \mathbf{i}K'/2 + \frac{(j - \frac{1}{2})K}{p},\ j = 1, \ldots, p$$

and $\omega_j = 2K/p,\ j = 1, \ldots, p$. This results in the formula for approximating (3.47),

$$\mathsf{U} \approx \mathsf{U}_p = \frac{-2\sqrt{\underline{\lambda}\overline{\lambda}}}{\pi k} \,\mathrm{Im} \sum_{j=1}^{p} \frac{2K}{p} \frac{\mathrm{cn}(t_j)\mathrm{dn}(t_j)}{(k^{-1} - \mathrm{sn}(t_j))^2} (z(t_j)\mathsf{B} - \mathsf{A})^{-1} \mathsf{B} \mathsf{Y}. \tag{3.48}$$

Of course, any other integration formula in principle can be applied by just replacing the integration points t_j and the numbers $\omega_j = 2K/p$ in (3.48) by the respective quantities.

Without proof, let us state the result from [41] which also applies in our case, since the integral is transformed in exactly the same way. All further constants, such as the norm of Y, are hidden in the $\mathcal{O}$-notation. We only adapted the notation to ours.

Theorem 3.15 ([41, Thm. 2.1])
Let A, B be real symmetric and let $I_\lambda = [\underline{\lambda}, \overline{\lambda}] \subset (0, +\infty)$. Then, the formula (3.48) *fulfills*

$$\|\mathsf{U} - \mathsf{U}_p\| = \mathcal{O}\left(\exp\left(\varepsilon - \frac{\pi K' p}{2K}\right)\right),$$

for any $\varepsilon > 0$ and $p \to \infty$. We have $\pi K'/(2K) \sim \pi^2/\log(\overline{\lambda}/\underline{\lambda})$ for $\overline{\lambda}/\underline{\lambda} \to \infty$. Further, we have for all $\underline{\lambda}, \overline{\lambda} > 0$

$$\|\mathsf{U} - \mathsf{U}_p\| = \mathcal{O}\left(\exp\left(-\pi^2 p/(\log(\overline{\lambda}/\underline{\lambda}) + 3)\right)\right).$$

3.5.3 Numerical experiments and discussion

To show the effectiveness of the method, let us conduct two simple experiments. For the implementation of the transformation method, we adapted the MATLAB [106] code printed in [41]. The functions cn, dn and sn are implemented in the Schwarz-Christoffel toolbox [25, 26].

Use within FEAST

Experiment 3.16
In this experiment, we calculate the eigenpairs belonging to the 300 largest eigenvalues of the size-1473 matrix pair $(\mathsf{A}, \mathsf{B}) = (\texttt{bcsstk11}, \texttt{bcsstm11})$ from structural engineering [68]. The results are shown in Table 3.11. The challenge of this pair is that the eigenvalues range widely, across 9 orders of magnitude. Via $\mathsf{A} \leftarrow \mathsf{A} - \lambda_{1173}\mathsf{B}$ we shifted (A, B) such that the 300 highest eigenvalues are positive while all others are non-positive. The sought eigenvalues ranged over 4 orders of magnitude. We allowed different versions of FEAST to run for 10 iterations, requiring a residual of 10^{-12}. The Gauß–Legendre, trapezoidal and midpoint rules were applied directly to the integral (3.41), where the contour was chosen to be a circle.

We counted the number of iterations, converged eigenpairs and linear systems solved. The solution of linear systems was counted per vector, i. e., the solution of $(z\mathsf{B} - \mathsf{A})\mathsf{V} = \mathsf{BY}$, where Y is an $n \times m$-matrix, was counted as m solves. The number of block solves is simply the iteration count times integration order. The first count is significant when using iterative solvers, the second count when using factorizations of $(z\mathsf{B} - \mathsf{A})$. In the latter case, the factorizations have to be computed only once per integration point.

Even for the highest order tested, i. e., $p = 16$, not all eigenpairs converged using the classical integration methods. The midpoint rule appears quite useless, which is not the whole truth. It is just not capable to reach the desired accuracy, while a per-eigenpair residual of 10^{-8} is quickly reached for some eigenpairs. ◊

The method with the transformed integration region showed superior performance in all three quantities measured, while the midpoint rule showed the worst performance. Interestingly, the underlying integration scheme in the transformation method *is* the midpoint rule. The other integration schemes mentioned showed worse performance in connection with the transformation method. To explain this effect, we may once again take a look at the selection functions, which are defined for an integration scheme $(t_j, \omega_j)_j$ according to (3.48) as

$$S(\lambda) = \frac{-2\sqrt{\underline{\lambda}\overline{\lambda}}}{\pi k} \operatorname{Im} \sum_{j=0}^{p} \omega_j \frac{\operatorname{cn}(t_j)\operatorname{dn}(t_j)}{(k^{-1} - \operatorname{sn}(t_j))^2} (z(t_j) - \lambda)^{-1}. \tag{3.49}$$

Note the slightly different numbering beginning at 0. The method is designed for computing upper eigenvalues; this behavior is also captured by the selection

Order (p)	Transformation method	Gauß–Legendre	Trapezoidal rule	Midpoint rule
4	300 eigenpairs 9 iterations 13k solves	0 eigenpairs 10 iterations 18k solves	65 eigenpairs 10 iterations 17k solves	0 eigenpairs 10 iterations 18k solves
8	300 eigenpairs 5 iterations 14k solves	0 eigenpairs 10 iterations 36k solves	68 eigenpairs 10 iterations 35k solves	0 eigenpairs 10 iterations 36k solves
16	300 eigenpairs 2 iterations 14k solves	294 eigenpairs 10 iterations 58k solves	82 eigenpairs 10 iterations 69k solves	0 eigenpairs 10 iterations 72k solves

Table 3.11: Comparison of integration methods for Experiment 3.16.

functions which do not decrease to zero immediately at the end of I_λ. For the interval I_λ used in Experiment 3.16, we plotted the selection function (3.49) for the Gauß–Legendre, trapezoidal and midpoint rule, each of order $p = 8$. They are shown in Figure 3.30. All three functions have in common that they make a "jump" at zero (actually they are only very steep at zero). This fact leads to a good convergence rate and small exclusion intervals, cf. Section 3.4.6. In the experiment we had $I_\lambda = [\underline{\lambda}, \overline{\lambda}] = [6.03 \times 10^{-5}, 0.94]$. The interval boundaries are marked by a vertical line. The best behavior inside the interval and below $\underline{\lambda}$ is seen for the midpoint rule (as suggested in [41]). It is also seen that the functions are not selection functions in the sense that they are approximating χ_{I_λ}. They rather are functions that damp the lower eigenvalues and amplify the upper eigenvalues, which is exactly the intended use in this application. The values of the selection functions (3.49) for $\lambda > \underline{\lambda}$ are not of interest in this context.

Normwise error

Similar to the experiment in [41, Sec. 2] we can measure the normwise errors in the subspace basis obtained by the transformation method. We will compute the numbers $\|\mathsf{U} - \mathsf{U}_p\|$ for different integration orders p. A similar experiment is also performed in Section 3.6.5 for other integration schemes. Let X be the "exact" eigenspace belonging to the eigenpairs sought for in Experiment 3.16. It was obtained by a direct method, MATLAB's `eig` [106], which internally calls LAPACK [5]. Then, for a fixed (but random) B-orthonormal starting basis Y with 300 columns, we compute $\mathsf{U} := \mathsf{X}\mathsf{X}^\star\mathsf{B}\mathsf{Y}$ (recall, $\mathsf{X}\mathsf{X}^\star\mathsf{B}$ is the B-orthogonal projector onto span(X)). The useful measure for the distance of U and U_p is the $\mathsf{B}2$-norm. For values $p = 5, 10, \ldots, 40$ we measured these norms. They can be seen in Figure 3.31, showing perfect exponential decay.

We also measured the error for the trapezoidal rule directly applied to the integral (3.41), the error was decaying for growing p, though it was of order 1

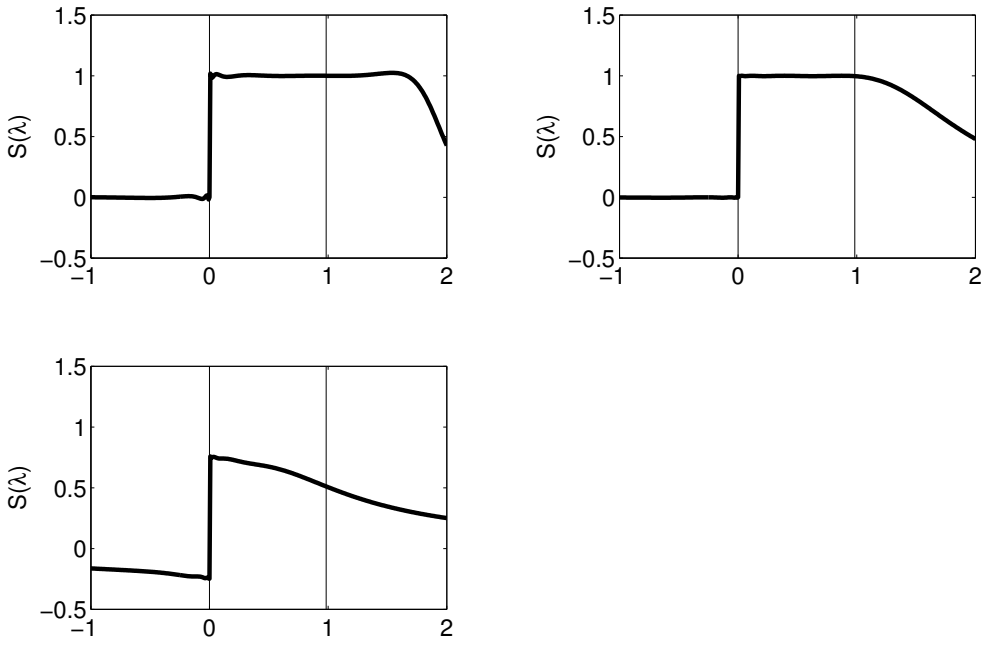

Figure 3.30: Selection functions for the transformed region. Gauß–Legendre (top, left), midpoint (top, right), trapezoidal (bottom).

for $p = 5, 10, \ldots, 40$. This rather large error seems to be surprising, but when recalling Theorem 2.43 it can be explained. The distance from the curve chosen to the next eigenvalue was in the best case $d = 3 \times 10^{-5}$ (the distance of $\underline{\lambda}$ to the next lower eigenvalue being 6×10^{-5}). In the theorem we have an error bound being a multiple of

$$C_1 k \cdot d^{-1} \exp(-C_2 p d), \tag{3.50}$$

where k denotes the number of eigenvalues sought for, i. e., 300 in our case. When thinking of $C_1 = C_2 = 1$, the number (3.50) becomes 9.92×10^6 for $p = 40$.

However, the normwise error is not the most important thing in the context of the FEAST algorithm, see Section 3.6.5.

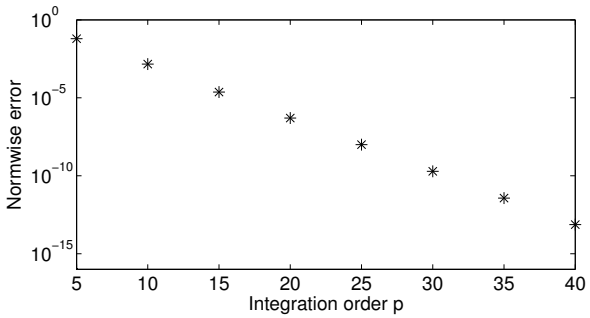

Figure 3.31: Normwise error in the basis, $\|\mathsf{U} - \mathsf{U}_p\|_{\mathsf{B2}}$.

Danger of too high accuracy

In some of our experiments with the transformation method within FEAST the obtained accuracy was in fact too *high*. When trying to repeat the test from Experiment 3.11, Problem 1, the computation of the 317 highest eigenpairs of the matrix T_{2003}, the following effect occurred. When using $p = 16$ in the transformation method, already in the first iteration of FEAST the matrix B_U had numerical rank 317, while being of size 450. This could rarely be observed when using the standard integration techniques, where the numerical rank of B_U typically was full, no matter how many eigenvalues were in I_λ. In this case here, the 317 vectors of interest can be extracted from the basis U via the approach from Section 3.2.5. With the obtained basis, the Rayleigh quotients $(\mathsf{A}_\mathsf{U}, \mathsf{B}_\mathsf{U})$ can be computed which then form a definite matrix pair.

Summary

The presented integration method is a slight modification of the method described in [41]. In some special cases, when the eigenpairs with largest or smallest eigenvalues are sought, it is applicable and superior compared to the classical integration methods. In some cases, it is even too exact in order to deliver a full rank matrix B_U. However, in practice this effect will rarely appear.

The method as presented can only be used for extremal eigenpairs. In the case of the standard eigenequation, a simple spectral transformation of the kind $\mathsf{A} \mapsto (\mathsf{A} - \sigma\mathsf{I})^2$ can be used in order to transform inner eigenvalues to extremal eigenvalues. The applicability of such transformations still has to be investigated.

3.6 Miscellaneous issues

In this section, we will give an overview of some shortcomings of the FEAST eigenvalue solver and test some parts of the algorithm that were not seen in action so far. Some of the material presented in Sections 3.6.1–3.6.4 has previously been discussed in [60] and is hence not presented in all details in this work.

3.6.1 Linear systems

In the standard version of FEAST, as proposed by Polizzi [85] and stated in Algorithm 3.1, for each integration point $z = z_j$ a block linear system of the form

$$(z\mathsf{B} - \mathsf{A})\mathsf{V} = \mathsf{BY} \tag{3.51}$$

has to be solved. The solution of (3.51) has already been discussed in [60], in particular the problems that can occur. As the matrix pair (A, B) is expected to be large and sparse, dense solvers, i. e., solvers that are based on Gaussian elimination, are not applicable. They require $\mathcal{O}(n^3)$ operations for each factorization

of the system matrix $\mathsf{M} := z\mathsf{B} - \mathsf{A}$, i. e., for each integration point z_j, $j = 0, \ldots, p$. Hence, iterative solvers are the methods to consider. Here, a method that is applicable to non-Hermitian matrices has to be used, since M is non-Hermitian in general. This can easily be explained by the fact that a Hermitian matrix has a real diagonal, but multiplying B with a complex number z yields a matrix $z\mathsf{B}$ with complex diagonal. The diagonal of $z\mathsf{B} - \mathsf{A}$ then is complex as well. One of the most widely used methods for general non-Hermitian matrices is GMRES, introduced by Saad and Schultz [89], see also Saad's monograph [90] for a comprehensive overview of iterative methods for linear systems.

In [60], we identified two more problems besides the choice of the method. One problem is that the condition number κ of the system matrix of (3.51) can be large. It is given by $\kappa(\mathsf{M}) = \|(z\mathsf{B} - \mathsf{A})^{-1}\| \cdot \|z\mathsf{B} - \mathsf{A}\|$, where in particular the first norm can become very large as z approaches $\operatorname{spec}(\mathsf{A}, \mathsf{B})$.

The second problem lies with the spectrum of the shifted matrix $\mathsf{M} = z\mathsf{B} - \mathsf{A}$. Even though the spectrum of (A, B) is real, the eigenvalues of $z\mathsf{B} - \mathsf{A}$ are typically scattered somewhere in the complex plane. Trefethen and Bau [108] explain very descriptively how the GMRES convergence depends on the structure of the spectrum. If M is diagonalizable, say $\mathsf{M} = \mathsf{W}\Xi\mathsf{W}^{-1}$ for a diagonal matrix $\Xi = \operatorname{diag}(\xi_1, \ldots, \xi_n)$, we have the following result that is not very hard to prove.

Theorem 3.17 ([108, Thm. 35.2])
Suppose, we want to solve $\mathsf{Mv} = \mathsf{b}$ *for a single vector* v. *Let* $\mathsf{r}_k = \mathsf{b} - \mathsf{Mv}_k$ *denote the residual for the k-th GMRES iterate* v_k. *We then have*

$$\frac{\|\mathsf{r}_k\|}{\|\mathsf{b}\|} \leq \inf_{p_k} \|p_k(\mathsf{M})\| \leq \kappa(\mathsf{W}) \inf_{p_k} \max_j |p_k(\xi_j)|\,, \tag{3.52}$$

where p_k *ranges over the set*

$$\{p : p \text{ is a polynomial, } \deg(p) \leq k,\ p(0) = 1\}\,.$$

In [108], an example can be found where the condition number of the matrix M is modest, while the eigenvalues are widely distributed, making it impossible to find a polynomial such that the upper bound in (3.52) becomes small. The authors also inform that this is typically the case if eigenvalues are located around the origin.

Preconditioners can be used to improve the convergence of GMRES. Tailored preconditioners for the matrices $z\mathsf{B} - \mathsf{A}$ probably will be part of future research. An overview of preconditioners can, e. g., be found in Saad's book [90].

Despite all problems with iterative linear solvers, they come with a feature that all iterative linear solvers have in common; they can be stopped at any iteration if one is satisfied with the reached accuracy. In the context of FEAST, this means that one might wish to compute the solution of the linear systems only to modest accuracy, of course expecting only modest accuracy in the eigenpairs. In order to investigate the connection between the two kinds of accuracy, we conducted the following experiment, it was first published in [60, Exp. 3.3].

Experiment 3.18 (Adapted from [60, Exp. 3.3])
We applied Algorithm 3.1 to the matrix pair (A, B), where A = `LAP_CIT_395` arises in the modeling of cross-citations in scientific publications [107], and B was chosen to be a diagonal matrix with random entries. We calculated the eigenpairs corresponding to the 10 largest eigenvalues. The linear systems (3.51) were solved column-by-column by running GMRES until $\|(z\mathsf{B} - \mathsf{A})\mathsf{v}_j - \mathsf{B}\mathsf{y}_j\| \,/\, \|\mathsf{r}_j^0\| \leq \varepsilon_{\text{lin}}$. Here, y_j denotes the j-th column of Y and r_j^0 is the starting residual corresponding to this column. Figure 3.32 reveals that the residual bounds required in the solution of the inner linear systems translated almost one-to-one into the residuals of the Ritz pairs. Even for a rather large bound such as $\varepsilon_{\text{lin}} = 10^{-6}$, the FEAST algorithm still converged (even though to a quite large residual). For the orthogonality of the computed eigenvectors x_j, the situation was different. After 20 FEAST iterations, an orthogonality level $\max_{i \neq j} |\widetilde{\mathsf{x}}_i^{\star} \mathsf{B} \widetilde{\mathsf{x}}_j|$ of order 10^{-15} could be reached for each of the bounds $\varepsilon_{\text{lin}} = 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}$ in the solution of the linear systems. Thus the achievable orthogonality seems not to be very sensitive to the accuracy of the linear solves. It also did not deteriorate significantly for a larger number of desired eigenpairs. ◊

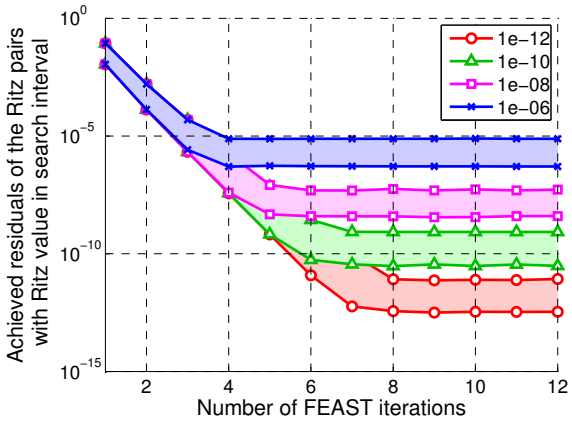

Figure 3.32: Range of all residuals among all Ritz pairs in I_λ for four different residual bounds ε_{lin} in the linear system solver [60].

When employing an iterative linear solver based on polynomial approximation, e. g., a Krylov subspace solver such as GMRES, one ends up with a matrix polynomial that should approximate the inverse of the matrix on a certain subspace. Recall that one could also think of approximating the integral in the FEAST algorithm directly by polynomials in this case, see Section 3.4.

3.6.2 Parallelism

Having very large, partial (maybe full) eigenproblems as key application for the FEAST algorithm in mind, it becomes clear that the algorithm *has* to be imple-

mented in parallel. Polizzi [85] already noted three different levels where parallelism can be introduced.

The first one is ostensibly the most simple one. By subdividing I_λ into K subintervals $I_\lambda^{(1)}, \ldots, I_\lambda^{(K)}$ such that

$$I_\lambda = I_\lambda^{(1)} \cup I_\lambda^{(2)} \cup \cdots \cup I_\lambda^{(K)}, \tag{3.53}$$

we also divide the required workload. Note that I_λ in (3.53) does not necessarily have to be an interval, it is rather a union of intervals. At first glance, the algorithm then can be run completely independently for each interval. Using this approach, problems can occur with the cross-interval orthogonality. The question is, for two computed eigenpairs $(\widetilde{\mathsf{x}}_1, \widetilde{\lambda}_1)$ and $(\widetilde{\mathsf{x}}_2, \widetilde{\lambda}_2)$ with $\widetilde{\lambda}_1, \widetilde{\lambda}_2$ residing in different intervals, can we guarantee or at least expect $|\widetilde{\mathsf{x}}_1^\star \mathsf{B} \widetilde{\mathsf{x}}_2|$ to be small? This is discussed in Section 3.6.3. On the other hand, the orthogonality can be used to check whether one eigenpair was doubly computed in different intervals (that probably have a slight overlap). In this case, one would check if $|\widetilde{\mathsf{x}}_1^\star \mathsf{B} \widetilde{\mathsf{x}}_2| \approx 1$ and discard one of the pairs. Another problem is load balancing. In order to achieve good utilization of a parallel environment, the subintervals should each contain a similar number of eigenvalues. In absence of detailed knowledge on the structure of the spectrum, this is of course not realizable a priori. A way out is implementing a master-slave approach and using a task queue. The intervals are inserted into this queue which is managed by one "master" process. This process distributes each task to a "slave" processes, as soon as one of those is free. A parallel C-code using MPI implementing this technique is being developed [32]. Recently, a numerical study has been conducted [4] that also takes load balancing into account. In this study, the problem of non-orthogonal eigenvectors was not addressed.

The second level of parallelism is the numerical integration step, where the solutions of different linear systems are summed up. For each integration point, the solution can be carried out separately, while the summation requires communication. Polizzi [85] also notes that the system matrices $z\mathsf{B} - \mathsf{A}$ only need to be factorized once for all FEAST iterations, if a linear solver based on factorization is employed. In this case, the factors of the system matrices have to be communicated.

The third level of parallelism is the solution of linear systems itself. It can be parallelized by solving the different columns of the system (3.51) independently. Further, the solver itself can be parallelized depending on its nature. If an iterative solver is used, its main computation time is typically consumed by simple matrix-vector products, which can efficiently be parallelized.

3.6.3 Orthogonality

It was already mentioned in Section 1.5.1 that we expect the computed eigenvectors of a Hermitian or definite generalized eigenproblem to be mutually or-

thogonal, at least to a certain degree. When subdividing I_λ according to (3.53), we can measure the orthogonality of eigenvectors belonging to eigenvalues in the same interval and of those belonging to eigenvalues in distinct intervals. In this context, it makes sense to introduce two different measures of orthogonality, the *global* orthogonality and the *local* orthogonality, restricted to the interval with number k. In formulas, we have the quantities [60]

$$\mathsf{orth}_{\text{global}} = \max_{i\neq j,\ \widetilde{\lambda}_i,\widetilde{\lambda}_j\in I_\lambda} |\widetilde{\mathsf{x}}_i^\star \mathsf{B}\widetilde{\mathsf{x}}_j|\,, \tag{3.54}$$

$$\mathsf{orth}_k = \max_{i\neq j,\ \widetilde{\lambda}_i,\widetilde{\lambda}_j\in I_\lambda^{(k)}} |\widetilde{\mathsf{x}}_i^\star \mathsf{B}\widetilde{\mathsf{x}}_j|\,. \tag{3.55}$$

In [33] it was pointed out that (3.55) might be small while (3.54) increases with the number of intervals K. Typically, when stepping from one to, say, 5 intervals, there is a jump of several orders of magnitude in $\mathsf{orth}_{\text{global}}$, while increasing K further only results in a moderate increase of $\mathsf{orth}_{\text{global}}$. In [60] we showed that in some cases, when splitting the interval I_λ in a quite unfortunate way, it is possible that (3.54) is larger than (3.55) by a factor of 10^{11} for $K = 2$. The corresponding Experiment is repeated below, including the figures from [60].

Experiment 3.19 (Adapted from [60, Exp. 4.2])
In this test, we consider a real unreduced tridiagonal matrix T_{2003} of size 2003 (from Experiment 3.11, Problem 1).

Its eigenvalues are simple, even though some are tightly clustered, see the top plots in Figure 3.33. The objective is to compute the 300 largest eigenpairs. To this end we initially split the interval $I_\lambda = [\lambda_{1704},\ \lambda_{2003}]$ into $I_\lambda^{(1)} = [\lambda_{1704}, \mu]$ and $I_\lambda^{(2)} = [\mu, \lambda_{2003}]$, with $\mu = \lambda_{1825} \approx 0.448 \times 10^{-3}$ chosen within a cluster of 99 eigenvalues. The relative gap between eigenvalue λ_{1825} and its neighbors is about 10^{-12} (i. e., agreement to roughly eleven leading decimal digits). A sketch of the spectrum with μ is given in the top left plot of Figure 3.33.

While FEAST attains very good local orthogonality for both subintervals ($\mathsf{orth}_1 = 4.4 \times 10^{-15}$ and $\mathsf{orth}_2 = 5.7 \times 10^{-14}$), it fails to deliver global orthogonality ($\mathsf{orth}_{\text{global}} = 4.7 \times 10^{-4}$). In the bottom left plot of Figure 3.33 we provide a pictorial description of $|\widetilde{\mathsf{x}}_i^\star \mathsf{B}\widetilde{\mathsf{x}}_j|$, $\widetilde{\lambda}_i$, $\widetilde{\lambda}_j \in I_\lambda$. The dark regions indicate that the loss of orthogonality emerges exclusively from eigenvectors belonging to the cluster of size 99. Next we divide the interval into 3 segments making sure not to break existing clusters (see top right of Figure 3.33). As illustrated in the bottom right plot, both the local and global orthogonality are satisfactory (10^{-13} or better). ◊

Thinking of the parallelization technique from Section 3.6.3 one immediately realizes that problems come up and a reorthogonalization step might be necessary; this is research in progress [32].

The quantity orth_k itself is basically ensured to be of low magnitude by design of the algorithm. This is easily explained, since in the Rayleigh–Ritz step of the

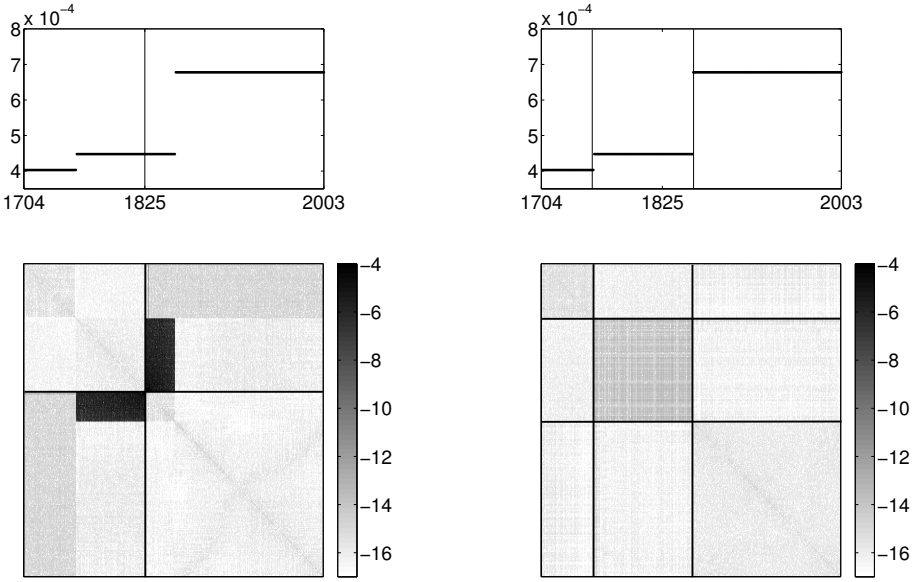

Figure 3.33: Results for Experiment 3.19. Computation of the eigenpairs corresponding to the 300 largest eigenvalues $\lambda_{1704}, \ldots, \lambda_{2003}$. The subdivision point $\mu = \lambda_{1825}$ in the left plots is taken from a group of very close eigenvalues. The top plots show the eigenvalues and the subdivision points (vertical lines), the bottom plots give a pictorial visualization of the orthogonality $|\widetilde{\mathsf{x}}_i^\star \mathsf{B} \widetilde{\mathsf{x}}_j|$, $i \neq j$.

FEAST algorithm, the small scale full eigen decomposition

$$\mathsf{A}_\mathsf{U} \widetilde{\mathsf{W}} = \mathsf{B}_\mathsf{U} \widetilde{\mathsf{W}} \widetilde{\Lambda} \tag{3.56}$$

is computed. This can be done such that $\widetilde{\mathsf{W}}$ is B_U-orthogonal, i. e., $\widetilde{\mathsf{W}}^\star \mathsf{B}_\mathsf{U} \widetilde{\mathsf{W}} = \widetilde{\mathsf{W}}^\star \mathsf{U}^\star \mathsf{B} \mathsf{U} \widetilde{\mathsf{W}} = \mathsf{I}$. The Ritz vectors then are computed as $\widetilde{\mathsf{X}} = \mathsf{U} \widetilde{\mathsf{W}}$, hence they are B-orthogonal. Of course, the obtained orthogonality depends on the achieved accuracy in the solution of (3.56).

Note that it is absolutely necessary that A_U, B_U are exactly Hermitian, in order to obtain B_U-orthogonal eigenvectors $\widetilde{\mathsf{W}}$. This can easily be achieved by updating $\mathsf{A}_\mathsf{U} \leftarrow 0.5(\mathsf{A}_\mathsf{U} + \mathsf{A}_\mathsf{U}^\star)$, $\mathsf{B}_\mathsf{U} \leftarrow 0.5(\mathsf{B}_\mathsf{U} + \mathsf{B}_\mathsf{U}^\star)$. Even an extremely small difference between A_U, $\mathsf{A}_\mathsf{U}^\star$ and B_U, $\mathsf{B}_\mathsf{U}^\star$, respectively, prevents some MATLAB [106] routines for computing (3.56) to treat it as a definite eigenproblem.

3.6.4 Stopping criteria and eigenpair locking

Another topic already addressed in [60] is the choice of reliable stopping criteria. They enter the picture in the last line of Algorithm 3.1. In Polizzi's first publication on FEAST [85], a criterion based on the *trace* i. e., of the sum of the computed Ritz values was used. The relative difference from iteration to iteration

is measured by the criterion

$$\left| \frac{\mathsf{trace}_k - \mathsf{trace}_{k-1}}{\mathsf{trace}_k} \right| < \mathsf{tol}. \tag{3.57}$$

Here, tol is a user specified tolerance and trace_k denotes the sum of the computed Ritz values in the k-th iteration, $k \geq 2$.

In [60] we pointed out three problems with criterion (3.57). The first one is a zero or almost zero denominator in (3.57), typically preventing the fraction from being small. The second problem arises if the numbers trace_{k-1}, trace_k are (almost) identical. This scenario is possible even if the individual Ritz values still are changing. In this case, criterion (3.57) is fulfilled and all eigenpairs are flagged as converged even though the residuals might still be large. The third problem mentioned in [60] is more general. In case of stagnation, any method will signal convergence if only the change of eigenvalues is taken into account.

Even if the Ritz values converge (and therefore the trace is doing so), the problem still lies with the Ritz vectors. It was worked out in Section 2.1 that the convergence of Ritz vectors relies on more complicated conditions than the convergence of Ritz values (e. g., the separation of eigenvalues). In particular, it is possible that Ritz values converge (i. e., (3.57) is fulfilled) while the corresponding Ritz vectors do not, cf. [100]. Hence, a per-eigenpair residual criterion was proposed in [60]. It takes the form

$$\left\| \mathsf{A}\widetilde{\mathsf{x}} - \mathsf{B}\widetilde{\mathsf{x}}\widetilde{\lambda} \right\| \leq \mathsf{tol} \cdot n \cdot \max\left\{ |\underline{\lambda}|, \left|\overline{\lambda}\right| \right\}, \tag{3.58}$$

where once more tol is the user specified tolerance, which should be at least as large as ε_M, the machine precision. The cost of computing the left hand side of (3.58) is not too high, as $\mathsf{B}\widetilde{\mathsf{x}}$ is one column of the matrix $\mathsf{B}\widetilde{\mathsf{X}}$ needed in the computation of the integral in Algorithm 3.1. The vector $\mathsf{A}\widetilde{\mathsf{x}}$ can be computed as $(\mathsf{AU})\mathsf{w}$, if w is the primitive Ritz vector of $\widetilde{\mathsf{x}}$. The matrix AU is available from forming the Rayleigh quotient $\mathsf{A}_\mathsf{U} = \mathsf{U}^\star\mathsf{AU}$. Without exploiting the sparsity, this computation costs $\mathcal{O}(\widetilde{m}^2 \cdot n)$ operations [60], because a product of the form $(\mathsf{AU})\mathsf{w}$ is computed $\widetilde{m}$ times (once for each Ritz vector). If $\max\left\{ |\underline{\lambda}|, \left|\overline{\lambda}\right| \right\}$ is very small, one should replace this quantity by a larger one, say σ, fulfilling

$$\max\left\{ |\underline{\lambda}|, \left|\overline{\lambda}\right| \right\} \leq \sigma \leq \max|\mathrm{spec}(\mathsf{A}, \mathsf{B})| = \left\| \mathsf{B}^{-1}\mathsf{A} \right\|.$$

A practical comparison between the convergence criteria based on the trace and on residual norms was performed in [60]. In there, examples can be found where the trace criterion signals convergence although the residuals are still large. Furthermore, examples where the residuals are already small but the trace criterion still does not signal convergence are given in [60]. Both scenarios are of course very undesirable.

Another benefit from using the per-eigenpair residual criterion is the possibility of *locking* single converged eigenpairs. By contrast, the trace criterion (3.57)

only allows to detect convergence of all eigenpairs with eigenvalue in the considered interval. Locking of eigenpairs was briefly discussed in [34]; it can easily be implemented as follows. Suppose, the computation is performed with a subspace of dimension $\widetilde{m}$, leading to a matrix $\widetilde{\mathsf{X}}$ consisting of $\widetilde{m}$ Ritz vectors and a diagonal matrix $\widetilde{\Lambda}$ of Ritz values, ordered accordingly. Suppose, one eigenpair, say with index k, $1 \leq k \leq \widetilde{m}$ has converged fulfilling criterion (3.58). Then, the next FEAST iteration is performed with a new starting basis Y consisting of all columns of $\widetilde{\mathsf{X}}$ with the k-th column removed. Of course, this can be done for more than one converged eigenpair. The converged eigenpair may stay in place in memory, it is just not further considered in the computation. This process leads to a decrease of the number of necessary operations in a single FEAST iteration in the same order as $\widetilde{m}$ is decreased. The reason is that the number of operations is basically linear with $\widetilde{m}$.

In Polizzi's FEAST 2.1 software [84], a per-eigenpair residual criterion similar to (3.58) was introduced. It is basically (3.58), checking for

$$\frac{\left\| \mathsf{A}\widetilde{\mathsf{x}} - \mathsf{B}\widetilde{\mathsf{x}}\widetilde{\lambda} \right\|_1}{\max\left\{ |\underline{\lambda}|, \left|\overline{\lambda}\right| \right\} \|\mathsf{B}\widetilde{\mathsf{x}}\|_1} \leq \mathsf{tol},$$

which is similar to (3.58) when requiring the Ritz vectors to be B-normalized (in the 2-norm) and then using the 1-norm instead.

3.6.5 Connection of integration error, eigenvalue convergence and subspace convergence

In this section, we will numerically investigate the connection between the normwise error in the subspace, the approximation error in the eigenvalues and the canonical angles between the computed subspaces. At first glance, some effects that occur seem to be contradictory, since the subspace convergence is often very slow while the eigenvalues converge. We will see that this effect matches the theory. Let U denote the exact integral U and $\widetilde{\mathsf{U}}_p$ the numerical approximation by an order-p integration scheme. Let us start with a small artificial example.

Experiment 3.20
We choose a symmetric matrix A of size $n = 100$ at random by setting $\check{\mathsf{A}} =$ `randn(n)`, $\mathsf{A} := \check{\mathsf{A}} + \check{\mathsf{A}}^\star$ in MATLAB [106] and perform essentially the steps of one FEAST iteration with a random orthonormal starting basis $\mathsf{Y} \in \mathbb{R}^{100\times 50}$. First, we measure the normwise errors $\|\mathsf{U} - \widetilde{\mathsf{U}}_p\|$. These errors are basically ensured to converge to zero by the theory in Sections 2.5.2 and 2.5.3 for the trapezoidal and Gauß–Legendre rule, respectively. In practice, this convergence will not necessarily take place since the actual subspace is chosen larger than the dimension of the space spanned by U.

We choose the curve $\mathcal{C}$ such that it encircles the first 50 eigenvalues of A. We have $\lambda_{51} - \lambda_{50} \approx 0.82$, hence we may choose the curve $\mathcal{C}$ such that $d := \operatorname{dist}(\mathcal{C}, \operatorname{spec}(\mathsf{A})) = 0.41$, which would be a fairly large number in practice. The errors $\|\mathsf{U} - \widetilde{\mathsf{U}}_p\|$ are shown in Figure 3.34. Note that the integration orders used are extremely large. They range up to $p = 2000$, while we used in the context of the FEAST algorithm, e. g., $p = 8$ or $p = 16$. For these comparatively small values from practice, the errors in our experiment were still of order 1. However, the computed subspaces were already able to deliver reasonable eigenvalue approximations. The first 50 exact eigenvalues of A as well as the Ritz values belonging to the subspaces computed by the Gauß–Legendre and trapezoidal rule, each of order $p = 16$, are shown in Figure 3.36. Of course, this figure is not very meaningful; the approximation error of each Ritz value is still about 0.9 on average, however, it can be seen that the Ritz values are at least of the correct order of magnitude. Note that the process described here corresponds to one single FEAST iteration with the search space size being exactly the dimension of the desired eigenspace, which was shown to be very problematic.

Some of the components of $\widetilde{\mathsf{U}}_p$ also move in the correct direction, the 50 canonical angles between U and $\widetilde{\mathsf{U}}_p$ are shown in Figure 3.35 for $p = 16$. For this value of p, the largest canonical angle still is very large i. e., close to π. For $p = 2000$, the largest canonical angles for both integration schemes were of order comparable to $\|\mathsf{U} - \widetilde{\mathsf{U}}_p\|$ as is stated by the theory, cf. Theorem. 2.13. ◊

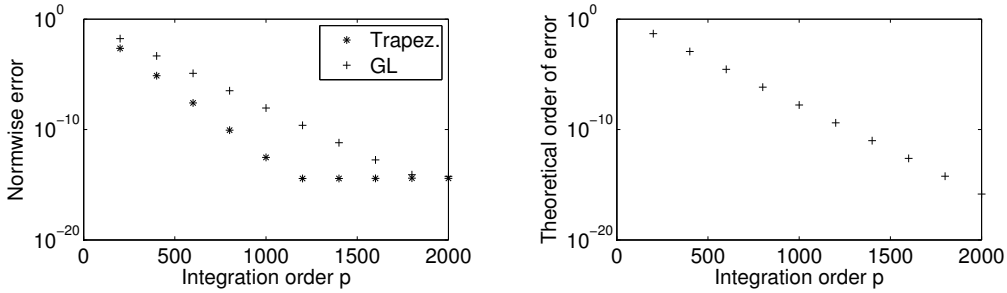

Figure 3.34: Left: Normwise integration error $\|\mathsf{U} - \widetilde{\mathsf{U}}_p\|$ for trapezoidal and Gauß–Legendre rule. Right: Estimated error of Gauß–Legendre rule.

The experiment has shown that a small normwise error in the subspace is not necessary for convergence of subspaces measured by canonical angles or for eigenvalue convergence.

The very slow convergence of the subspaces $\widetilde{\mathsf{U}}_p$ towards U can be justified theoretically. In case of the Gauß–Legendre rule we have, according to Theorem 2.46, the error bound

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\| \leq 2\kappa(\mathsf{X}) \left(\frac{\pi}{\gamma} + \varepsilon \right)^{2p+1} \cdot \|\mathsf{Y}\| \,, \; p > p_\varepsilon. \tag{3.59}$$

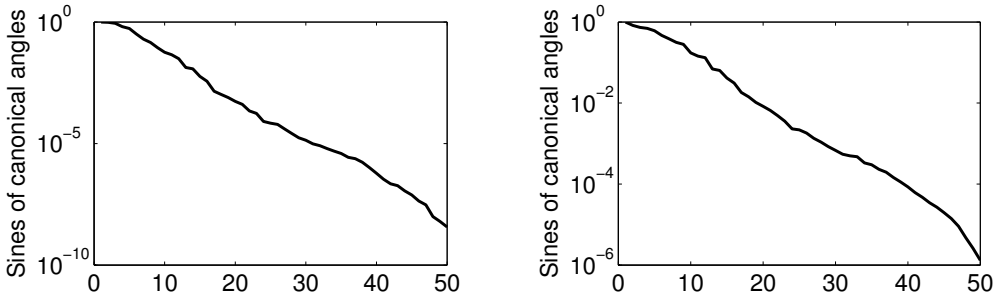

Figure 3.35: Sines of canonical angles between U and $\widetilde{\mathsf{U}}_p$ for $p = 16$. Left: Trapezoidal rule. Right: Gauß–Legendre rule.

Here, γ is a number that depends on the size of the region of analyticity of the resolvent and which is basically determined by the distance of the curve to the next eigenvalue. Both $\kappa(\mathsf{X})$ and $\|\mathsf{Y}\|$ have value 1, since X is the eigenvector matrix of the symmetric matrix A, hence orthonormal, and Y was chosen orthonormal. For the matrix and integration contour from Experiment 3.20 we found $\pi/\gamma > 0.99075$, hence being close to 1, even though the curve $\mathcal{C}$ has a comfortable distance of about 0.41 to the closest eigenvalue. The values $2(\pi/\gamma)^{2p+1}$ for $p = 200, 400, \ldots, 2000$ and $\pi/\gamma = 0.99075$ are shown in the right panel of Figure 3.34. We see that the theoretical prediction matches the measured value well. Note that we neglected $\varepsilon > 0$ from (3.59).

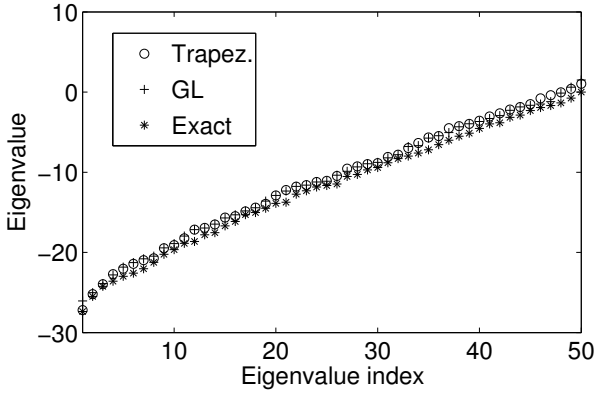

Figure 3.36: Approximation of eigenvalues computed by Gauß–Legendre and trapezoidal rule, respectively. We used $p = 16$ integration points in both cases.

For the trapezoidal rule things are slightly different. We have, according to Theorem 2.43

$$\left\| \mathsf{U} - \widetilde{\mathsf{U}}_p \right\| \leq k \cdot C_1 d^{-1} \exp(-C_2\, p\, d), \tag{3.60}$$

if the eigenvectors and Y are orthonormal. In this equation, C_1 and C_2 denote some positive constants and k is the number of eigenvalues inside $\mathcal{C}$. Again we

denote by d the distance from the curve $\mathcal{C}$ to the closest eigenvalue. With $d = 0.41$ as in Experiment 3.20, the right hand side of (3.60) converges extremely fast towards zero when thinking of C_1 and C_2 to be of order 1. Hence, the constant C_2 must be very small (but still positive).

For less artificial examples than in Experiment 3.20 the curve $\mathcal{C}$ typically passes the spectrum much more closely. For instance, if we have $d = 10^{-6}$ (which could be a value from practice), we would obtain $\pi/\gamma \approx 0.9999997$, a value whose positive powers converge extremely slow towards 0. The same holds for the error bound (3.60). However, the statements about the eigenvalue approximation and the canonical angles stay true, at least in a qualitative sense (the convergence of eigenvalues and canonical angles takes place much earlier than the normwise convergence of the subspaces).

In all our experiments with the FEAST algorithm we only used very modest integration orders, $p = 8, 16, 32$. In the literature however, in actual computations up to hundreds of thousands ($2^{18} \approx 262{,}000$) integration points were used in the context of matrix functions [16]. For nonlinear eigenvalue problems, Beyn used at least orders up to $p = 150$, see [11].

3.7 Conclusion

Chapter 3 was devoted to the practical aspects of the FEAST algorithm introduced by Polizzi [85]. After having introduced the basic algorithm we came to the important aspect of how to estimate the number of eigenvalues in an interval in Section 3.2. The presented techniques can also be used to compute the correct subspace needed.

In Section 3.3, the connection between numerical integration, approximation and matrix functions was made clear. It was shown that the selection functions belonging to the integration with the trapezoidal and midpoint rule, respectively, are simple rational functions if a circle as contour is used. The approximation of the characteristic function of the search interval by polynomials was discussed in Section 3.4. It was shown that using polynomials, eigenvalue problems up to a size of about one million can be solved on a rather small machine. This method is in particular applicable to sparse matrices, since only matrix-vector products are necessary. Next, we adapted techniques originating from the computation of matrix functions to the eigenvalue problem in Section 3.5. In some cases, this yielded much better results than in the standard algorithm. Section 3.6 was devoted to miscellaneous topics occurring in the implementation of the FEAST algorithm.

In this chapter we conducted numerous numerical experiments, shedding light on the aspects discussed. In Table 3.12, a list of the most important subjects treated numerically can be found.

Topic	Section	page	reference
Adaptive choice of polynomial degree N	3.4.7	139	
Chebyshev approximation	3.4.3	119	
Chebyshev-FEAST	3.4.5	124	
Error in Chebyshev approximation	3.4.3	119	
Eigenvalue counting: SVD, rrQR, Frobenius, Ritz	3.2.8	104, 105	[34]
Eigenvalue counting: CholQR	3.2.8	107	
Generalized problem	3.5.3	151	
Harmonic Rayleigh–Ritz	2.1.6	48	
Integration error/ eigenvalue and subspace convergence	3.6.5	161	
Legendre polynomials	3.4.9	144	
Linear solvers	3.6.1	156	[60]
Orthogonality	3.6.3	158	[60]
Selection function: midpoint and trapezoidal rule	3.3.1	113	
Stopping criterion	3.6.4	159	[60]
Transformation of problem	3.5.3	151	

Table 3.12: Selected numerical experiments concerning the FEAST algorithm.

Conclusion and outlook

The central topic of this work is the computation of eigenvalues, eigenvectors and invariant subspaces of a definite matrix pair (A, B) by using contour integrals. Methods using this technique rely on classical function theory, but have just recently been turned into algorithms, however with some shortcomings. We considered it necessary to list problems and provide suggestions to solve at least some of those.

In Chapter 1 we first introduced basic notions. An important part of this chapter is the theory of angles in a B-induced scalar product, whereof we compiled the most important facts.

We devoted Chapter 2 to the general theory of integration based eigensolvers. Here, it is central to separate the Rayleigh–Ritz part from the integration part. Several results concerning the Hermitian standard eigenvalue problem were generalized to eigenvalue problems involving definite matrix pairs. Next, a theoretical justification for the use of contour integrals was given. Finally, the convergence of the Gauß–Legendre and trapezoidal rule applied to the integral was proven.

Chapter 3 was all about algorithmics. We answered the important question of how to choose the correct subspace size. Next, the connection between numerical integration and approximation was made clear. This was followed by a longer section about polynomial approximation. The resulting polynomial based algorithm was extensively studied and tested. Using this technique it was possible to solve an eigenvalue problem larger than one million on a machine slightly larger than a desktop work station. A method from the computation of matrix functions was adapted to solve an eigenvalue problem. Finally, several minor issues were discussed.

To sum up, we analyzed—and enhanced to some extent—an eigensolver based on integration and approximation. The method is promising, while still having some shortcomings. At the moment it is mainly useful for expert users, since many parameters have to be set before starting the method. A real “black-box” method is not in sight at the moment.

List of contributions

The following is a list of what we consider the most important contributions in this work. Parts of the work have previously been published in [34, 60].

- Some error and perturbation bounds from the theory of the standard eigenproblem were adapted to the generalized eigenproblem in Section 2.1.
- We gave a mathematically precise introduction to eigensolvers based on numerical integration (Section 2.4).
- Convergence proofs of the Gauß–Legendre and trapezoidal rule applied to the resolvent were developed in Section 2.5.
- An overview of some methods for counting eigenvalues, including suggestions on how to compute the correct subspace is given in Section 3.2.
- In Section 3.3, we put the numerical integration part of the algorithm in a different light. The connection between numerical integration, approximation and matrix functions was made clear. Simple formulas for the integration with the trapezoidal and midpoint rules were given.
- The replacement of numerical integration by polynomial approximation was discussed thoroughly in Section 3.4.
- In Section 3.5 we adapted techniques originating from the computation of matrix functions to the eigenvalue problem. This resulted in some cases in much better results than if the standard algorithm was used.
- Finally, in Section 3.6 we performed some additional numerical experiments and listed some shortcomings of FEAST that can occur and still have to be addressed.

Outlook

As mentioned above, there are still many open questions, the following list summarizes what the author considers the most important ones.

- Many parameters still have to be tuned by hand (or chosen based on heuristics), e. g., the (initial) search space size, the integration order and the polynomial order in the approximation version. We should develop techniques for automatically choosing these parameters.
- Is there a way to limit the polynomial degree, independently of the system size?

- When using the integration based version, reliable iterative methods and preconditioners for the shifted block linear systems have to be developed.
- When splitting the search interval into parts, the global level of orthogonality increases. Hence, efficient re-orthogonalization schemes have to be developed. Even better, one should think about whether there is a possibility to ensure orthogonality without re-orthogonalization, similar to the MR^3 algorithm.
- An efficient, parallel implementation of all methods used has to be developed. This is already done at the moment [32].
- In principle, FEAST is also applicable to non-Hermitian eigenproblems. Furthermore similar methods also apply to non-linear eigenvalue problems [11]. The applicability of the discussed methods to non-Hermitian (and nonlinear) eigenproblems should be investigated, as well.

Finally, let us cite a statement by Christopher Paige [78], dating back to 1971, that still seems to be true:

> "Several methods are available for computing eigenvalues and eigenvectors of large sparse matrices, but as yet no outstandingly good algorithm is generally known."

Index

Summary of Notation

General

ε_M	Machine epsilon
ε	A "small" number in the current discussion
$\mathbf{i}$	Imaginary unit, $\mathbf{i}^2 = -1$
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$	Real and imaginary part of the complex number z, respectively
$\mathbb{Z}, \mathbb{Z}_{>0}, \mathbb{Z}_{\geq 0}$	Integers and positive and non-negative integers, respectively
$\widetilde{\star}$	Computed analogue to $\star$, where $\star$ can be replaced by any symbol
$\Diamond$	End of definition, remark, example, experiment

Linear Algebra

A, B	Square matrices of eigenvalue problem
K	Factor of hpd matrix B such that $\mathsf{B} = \mathsf{K}^\star\mathsf{K}$
n	Size of A and B
a, $\mathsf{b}, \ldots$	Vectors
$\mathrm{spec}(\mathsf{A}, \mathsf{B})$	Set of eigenvalues of (A, B)
$\rho(\mathsf{A})$	Spectral radius of A
$\mathrm{span}(\mathsf{M})$	Space spanned by M's columns
$\mathsf{nnz}(\mathsf{M})$	Number of nonzeros in M
o, $\mathsf{0}$	Zero vector and matrix, respectively
I_k	Identity matrix of size k
$\|\cdot\|$	Generic norm, 2-norm if not otherwise declared
$\|\cdot\|_2$	2-norm of matrix or vector
$\|\cdot\|_\mathsf{B}$	B-norm of matrix of vector
$\|\cdot\|_{\mathsf{B}2}$	B2-norm of a matrix

$\|\cdot\|_F$	Frobenius norm of a matrix or vector, respectively

Integration

p	Order of numerical integration
$\mathbb{P}_p$	Set of polynomials of degree p
$C[\alpha, \beta]$	Set of continuous functions on $[\alpha, \beta]$
$C^m[\alpha, \beta]$	Set of m-times continuously differentiable functions on $[\alpha, \beta]$
φ	Parametrization function
$\mathcal{C}$	Integration curve (= image of φ)
ω_j	Integration weights
t_j	Integration points
$E_{T_p}(\cdot)$	Error of trapezoidal rule
$E_{G_p}(\cdot)$	Error of Gauß–Legendre rule
$r_\lambda(z)$	Rational function $r_\lambda(z) = \frac{1}{z-\lambda}$
$\mathrm{sn}, \mathrm{cn}, \mathrm{dn}$	Jacobi elliptic functions

FEAST algorithm

I_λ	Eigenvalue interval
m	Actual number of eigenvalues in I_λ
$q, \tilde{q}$	Computed estimation of number of eigenvalues in I_λ
$\widetilde{m}$	(Current) chosen estimation of number of eigenvalues in I_λ
$\hat{u}$, $(\widetilde{u})$	Convergence rate (and estimation)
δ	Radius of exclusion interval

Approximation

χ_{I_λ}	Characteristic function of I_λ
N	Polynomial degree
C, C_N, Ψ_N	Approximating polynomial of degree N
$T_k(\cdot)$	Chebyshev polynomial of degree k
$L_k(\cdot)$	Legendre polynomial of degree k
c_k	Coefficients of Chebyshev or Legendre polynomial
g_k	Gibbs coefficients

List of Figures

List of Tables

List of Algorithms

Bibliography

[1] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, NY, 1970. 9th print.

[2] Lars V. Ahlfors. *Complex Analysis*. McGraw-Hill Book Company, New York, NY, 2nd edition, 1966.

[3] Lars V. Ahlfors. *Complex Analysis*. McGraw-Hill Book Company, New York, NY, 3rd edition, 1979.

[4] Hasan M. Aktulga, Lin Lin, Christopher Haine, Esmond G. Ng, and Chao Yang. Parallel eigenvalue calculation based on multiple shift-invert Lanczos and contour integral based spectral projection method. Preprint, 2012.

[5] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, James Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, Alan McKenney, and Danny Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, PA, third edition, 1999. Online available from http://www.netlib.org/lapack.

[6] Walter E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Qart. Appl. Math.*, 9:17–29, 1951.

[7] Junko Asakura, Tetsuya Sakurai, Hiroto Tadano, Tsutomu Ikegami, and Kinji Kimura. A numerical method for nonlinear eigenvalue problems using contour integrals. *JSIAM Letters*, 1:52–55, 2009.

[8] Thomas Auckenthaler, Volker Blum, Hans-Joachim Bungartz, Thomas Huckle, Rainer Johanni, Lukas Krämer, Bruno Lang, Hermann Lederer, and Paul R. Willems. Parallel solution of partial symmetric eigenvalue problems from electronic structure calculations. *Parallel Comput.*, 37(12):783–794, 2011.

[9] N. K. Basu. On double Chebyshev series approximation. *SIAM J. Numer. Anal.*, 10(3):496–505, 1973.

[10] Olivier Bertrand and Bernard Philippe. Counting the eigenvalues surrounded by a closed curve. *Sib. Zh. Ind. Mat.*, 4(2):73–94, 2001.

[11] Wolf-Jürgen Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra Appl.*, 436(10):3839–3863, 2012.

[12] Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Math. Comp.*, 27(123):579–594, 1973.

[13] Ilja N. Bronstein and Konstantin A. Semendjajew. *Taschenbuch der Mathematik*. Harri Deutsch, Frankfurt am Main, 21st edition, 1984.

[14] Guizhi Chen and Zhongxiao Jia. An analogue of the results of Saad and Stewart for harmonic Ritz vectors. *J. Comput. Appl. Math.*, 167(2):493–498, 2004.

[15] J. J. M. Cuppen. A divide and conquer method for the symmetric tridiagonal eigenproblem. *Numer. Math.*, 36:177–195, 1981.

[16] Philip I. Davies and Nicholas J. Higham. Computing $f(A)b$ for matrix functions f. In Artan Boriçi, Andreas Frommer, Bálint Joó, Anthony Kennedy, and Brian Pendleton, editors, *QCD and Numerical Analysis III*, volume 47 of *Lecture Notes in Computational Science and Engineering*, pages 15–24. Springer, Berlin, Heidelberg, 2005.

[17] Chandler Davis and William M. Kahan. Some new bounds on perturbation of subspaces. *Bull. Amer. Math. Soc.*, 75:863–868, 1969.

[18] Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

[19] Philip J. Davis. On the numerical integration of periodic analytic functions. In R. E. Langer, editor, *On numerical approximation*, Madison, WI, 1959. The University of Wisconsin Press.

[20] Philip J. Davis and P. Rabinowitz. *Methods of numerical integration*. Academic Press, Orlando, FL, second edition, 1984.

[21] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.

[22] James W. Demmel, Osni A. Marques, Beresford N. Parlett, and Christof Vömel. Performance and Accuracy of LAPACK's Symmetric Tridiagonal Eigensolvers. *SIAM J. Sci. Comput.*, 30(3):1508–1526, 2008.

[23] Inderjit S. Dhillon. *A new $O(n^2)$ algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem*. PhD thesis, University of California, Berkeley, 1997.

[24] Edoardo Di Napoli, Eric Polizzi, and Yousef Saad. Efficient estimation of eigenvalue counts in an interval. http://arxiv.org/abs/1308.4275, 2013.

[25] Tobin A. Driscoll. Schwarz-Christoffel toolbox user's guide, v2.3. Online available from http://www.math.udel.edu/~driscoll/SC/index.html. Checked 7th of November, 2013.

[26] Tobin A. Driscoll. Algorithm 843: Improvements to the Schwarz-Christoffel toolbox for MATLAB. *ACM Trans. Math. Softw.*, 31(2):239–251, 2005.

[27] Vladimir L. Druskin and Leonid A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *U.S.S.R. Comput. Math. Math. Phys.*, 29(6):112–121, 1989.

[28] Simon M.-M. Dubois, Zeila Zanolli, Xavier Declerck, and Jean-Christophe Charlier. Electronic properties and quantum transport in graphene-based nanostructures. *Eur. Phys. J. B*, 72(1):1–24, 2009.

[29] Ludwig Elsner. An optimal bound for the spectral variation of two matrices. *Linear Algebra Appl.*, 71:77–80, 1985.

[30] Bernd Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. Wiley-Teubner, New York, Leipzig, 1996.

[31] John G. F. Francis. The QR transformation: A unitary analogue to the LR transformation, parts I, II. *Computer J.*, 4:165–272, 332–345, 1961.

[32] Martin Galgon. Personal communication, 2010–2013.

[33] Martin Galgon, Lukas Krämer, and Bruno Lang. The FEAST algorithm for large eigenvalue problems. *Proc. Appl. Math. Mech.*, 11(1):747–748, 2011.

[34] Martin Galgon, Lukas Krämer, and Bruno Lang. Counting eigenvalues and improving the integration in the FEAST algorithm. Preprint BUW-IMACM 12/22, http://www.imacm.uni-wuppertal.de/imacm/research/preprints.html, 2012.

[35] Walter Gander. Algorithms for the QR decomposition. Technical report, Seminar für Angewandte Mathematik, ETH Zürich, 1980.

[36] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[37] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4th edition, 2013.

[38] Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Math. Comp.*, 23(106):221–230, 1969.

[39] Ming Gu and Stanley C. Eisenstat. A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM J. Matrix Anal. Appl.*, 16:172–191, 1995.

[40] Ming Gu and Stanley C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput.*, 17(4):848–869, 1996.

[41] Nicholas Hale, Nicholas J. Higham, and Lloyd N. Trefethen. Computing A^α, $\log(A)$ and related matrix functions by contour integrals. *SIAM J. Numer. Anal.*, 46(5):2505–2523, 2008.

[42] Peter Henrici. *Applied and Computational Complex Analysis*, volume I–Power Series—Integration—Conformal Mappings—Location of Zeros of *Pure & Applied Mathematics*. John Wiley & Sons, New York, NY, 1974.

[43] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, 49(6):409–436, 1952.

[44] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[45] Michiel E. Hochstenbach. Variations on harmonic Rayleigh–Ritz for standard and generalized eigenproblems. Technical report, TU Eindhoven, 2005. Available from http://www.win.tue.nl/~hochsten/.

[46] Olga Holtz and Michael Karow. Real and complex operator norms. Unpublished manuscript, available from http://arxiv.org/abs/math.FA/0512608, 2005.

[47] IEEE. *IEEE Standard 754-1985 for binary floating-point arithmetic*, 1985.

[48] IEEE. *IEEE Standard 754-2008 for floating-point arithmetic*, 2008.

[49] Tsutomu Ikegami, Tetsuya Sakurai, and Umpei Nagashima. A filter diagonalization for generalized eigenvalue problems based on the Sakurai-Sugiura projection method. *J. Comput. Appl. Math.*, 233(8):1927–1936, 2010.

[50] Carl G. J. Jacobi. Über eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen. *Astronom. Nachr.*, 1845:297–306, 1845.

[51] Carl G. J. Jacobi. Über ein leichtes Verfahren, die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *J. Reine und Angew. Math.*, 30:51–94, 1846.

[52] Zhongxiao Jia. The convergence of harmonic Ritz values, harmonic Ritz vectors, and refined harmonic Ritz vectors. *Math. Comp.*, 74(251):1441–1456, 2004.

[53] Zhongxiao Jia and Gilbert W. Stewart. An analysis of the Rayleigh–Ritz method for approximating eigenspaces. *Math. Comp.*, 70:637–647, 2001.

[54] Emmanuel Kamgnia and Bernard Philippe. Counting eigenvalues in domains of the complex field. *ETNA*, 40:1–16, 2013.

[55] Wilfred Kaplan. *Advanced Calculus*. Addison-Wesley, Reading, MA, 2nd edition, 1973.

[56] Tosio Kato. *Perturbation theory for linear operators*, volume 132 of *Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1966.

[57] Andrew V. Knyazev. *Computation of eigenvalues and eigenvectors for mesh problems: algorithms and error estimates*. Department of Numerical Mathematics, USSR Academy of sciences, Moscow, 1986. In Russian. Translation of title by author. Available online from http://math.ucdenver.edu/~aknyazev/research/papers/old/k.pdf, checked 12th of November, 2013.

[58] Andrew V. Knyazev and Merico E. Argentati. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM J. Sci. Comp.*, 23(6):2009–2041, 2002.

[59] Andrew V. Knyazev and Merico E. Argentati. On proximity of Rayleigh quotients for different vectors and Ritz values generated by different trial subspaces. *Linear Algebra Appl.*, 415(1):82–95, 2006.

[60] Lukas Krämer, Edoardo Di Napoli, Martin Galgon, Bruno Lang, and Paolo Bientinesi. Dissecting the FEAST algorithm for generalized eigenproblems. *J. Comput. Appl. Math.*, 244:1–9, 2013.

[61] Rainer Kress. On error norms of the trapezoidal rule. *SIAM J. Numer. Anal.*, 15(3):pp. 433–443, 1978.

[62] Rainer Kress. *Numerical Analysis*, volume 181 of *Graduate Texts in Mathematics*. Springer Netherlands, 1998.

[63] Arnold R. Krommer and Christoph W. Ueberhuber. *Computational Integration*. SIAM, Philadelphia, PA, 1998.

[64] Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, 45(4):255–282, 1950.

[65] Steven E. Laux. Solving complex band structure problems with the FEAST eigenvalue algorithm. *Phys. Rev. B*, 86:075103, 2012.

[66] Richard B. Lehoucq, Danny C. Sorensen, and Chao Yang. *ARPACK Users' Guide*. SIAM, Philadelphia, 1998.

[67] George G. Lorentz. *Bernstein Polynomials*. Chelsea Publishing Company, New York, NY, 1986.

[68] Matrix Market. http://math.nist.gov/MatrixMarket/. Checked at Nov. 4., 2013.

[69] Gérard Meurant. Estimates of the norm of the error in solving linear systems with FOM and GMRES. *SIAM J. Sci. Comp.*, 33(5):2686–2705, 2011.

[70] Louis M. Milne-Thomson. *Jacobian elliptic function tables*. Macmillan, London, Basingstoke, 1970.

[71] Cleve B. Moler and Gilbert W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.*, 10:241–256, 1973.

[72] Ronald B. Morgan. Computing interior eigenvalues of large matrices. *Linear Algebra Appl.*, 154–156:289–309, 1991.

[73] Jean-Michel Muller. *Elementary Functions—Algorithms and Implementation*. Birkhäuser, Boston–Basel–Berlin, 2nd edition, 2006.

[74] Yuji Nakatsukasa. Absolute and relative Weyl theorems for generalized eigenvalue problems. *Linear Algebra Appl.*, 432(1):242–248, 2010.

[75] Yuji Nakatsukasa. The $\tan\theta$ theorem with relaxed conditions. *Linear Algebra Appl.*, 436(5):1528–1534, 2012.

[76] Isidor P. Natanson. *Konstruktive Funktionentheorie*. Akademie-Verlag, Berlin, 1955. German translation by K. Bögel.

[77] Isidor P. Natanson. *Constructive Function Theory*, volume I: Uniform Approximation. Frederick Ungar Publishing, New York, NY, 1964. English translation by Alexis N. Obolensky.

[78] Christopher C. Paige. *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*. PhD thesis, University of London, England, 1971.

[79] Christopher C. Paige, Beresford N. Parlett, and Henk A. Van der Vorst. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numer. Linear Algebra Appl.*, 2(2):115–133, 1995.

[80] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*, volume 20 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, Classics edition, 1998.

[81] Penčo P. Petrushev and Vasil A. Popov. *Rational approximation of real functions*. Cambridge University Press, 1987.

[82] Andreas Pieper. Personal communication, 2013.

[83] Eric Polizzi. A high-performance numerical library for solving eigenvalue problems: Feast solver v2.0 user's guide. http://arxiv.org/abs/1203.4031v1 [cs.MS].

[84] Eric Polizzi. A high-performance numerical library for solving eigenvalue problems: Feast solver v2.1 user's guide. http://arxiv.org/abs/1203.4031v2 [cs.MS].

[85] Eric Polizzi. Density-matrix-based algorithm for solving eigenvalue problems. *Phys. Rev. B*, 79:115112, 2009.

[86] Philip Rabinowitz. Practical error coefficients in the integration of periodic analytic functions by the trapezoidal rule. *Comm. ACM*, 11:764–765, 1968.

[87] Lord Rayleigh. On the calculation of the frequency of vibration of a system in its gravest mode, with an example from hydrodynamics. *Philos. Mag. Series 5*, 47(289):566–572, 1899.

[88] Walter Ritz. Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. *J. Reine Angew. Math.*, 1909(135):1–61, 1909.

[89] Youcef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 7(3):856–869, 1986.

[90] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, 2nd edition, 2003.

[91] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, Philadelphia, PA, second edition, 2011.

[92] Tetsuya Sakurai and Hiroshi Sugiura. A projection method for generalized eigenvalue problems using numerical integration. *J. Comput. Appl. Math.*, 159:119–128, 2003.

[93] Grady Schofield, James R. Chelikowsky, and Yousef Saad. A spectrum slicing method for the Kohn–Sham problem. *Comput. Phys. Comm.*, 183(3):497–505, 2012.

[94] Gerard L. G. Sleijpen, Albert G. L. Booten, Diederik R. Fokkema, and Henk A. Van der Vorst. Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT*, 36(3):595–633, 1996.

[95] Gerard L. G. Sleijpen and Henk A. Van der Vorst. A Jacobi–Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17(2):401–425, 1996.

[96] Danny C. Sorensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM J. Matrix Anal. Appl.*, 13:357–385, 1992.

[97] Andreas Stathopoulos and Kesheng Wu. A block orthogonalization procedure with constant synchronization requirements. *SIAM J. Sci. Comput.*, 23(6):2165–2182, 2002.

[98] Gilbert W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.*, 15(4):727–764, 1973.

[99] Gilbert W. Stewart. A generalization of Saad's theorem on Rayleigh–Ritz approximations. *Linear Algebra Appl.*, 327(1–3):115–119, 2001.

[100] Gilbert W. Stewart. *Matrix Algorithms*, volume II, Eigensystems. SIAM, Philadelphia, PA, 2001.

[101] Gilbert W. Stewart. A Krylov–Schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.*, 23(3):601–614, 2002.

[102] Gilbert W. Stewart. An Elsner-like perturbation theorem for generalized eigenvalues. *Linear Algebra Appl.*, 390(0):1–5, 2004.

[103] Gilbert W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, San Diego, CA, 1990.

[104] Ji-guang Sun. Stability and accuracy—Perturbation analysis of algebraic eigenproblems. Technical Report UMINF 98.07, Umeå University, Departmend of Computer Science, 1998.

[105] Ping Tak Peter Tang and Eric Polizzi. Subspace iteration with approximate spectral projection. http://arxiv.org/abs/1302.0432 [math.NA], version 3, 2013.

[106] The MathWorks, Inc. Matlab R2013a, 1984–2013. Matlab is a registered trademark of The MathWorks, Inc.

[107] Mario Thüne. Personal communication. MPI MIS Leipzig, 2009.

[108] Lloyd N. Trefethen and David Bau, III. *Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.

[109] Charles F. Van Loan. A general matrix eigenvalue algorithm. *SIAM J. Numer. Anal.*, 12(6):819–834, 1975.

[110] James M. Varah. On the separation of two matrices. *SIAM J. Numer. Anal.*, 16(2):216–222, 1979.

[111] Gautier Viaud. The FEAST algorithm for generalised eigenvalue problems. Master's thesis, University of Oxford, 2012.

[112] Milan Vujičić. *Linear Algebra Thoroughly Explained*. Springer-Verlag, Berlin Heidelberg, 2008. Edited by Jeffrey Sanderson.

[113] David S. Watkins. The QR algorithm revisited. *SIAM Rev.*, 50(1):133–145, 2008.

[114] J. André C. Weideman. Numerical integration of periodic functions: A few examples. *Amer. Math. Monthly*, 109(1):21–36, 2002.

[115] Alexander Weiße, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. The kernel polynomial method. *Rev. Mod. Phys.*, 78:275–306, 2006.

[116] Hermann Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.*, 71:441–479, 1912.

[117] Herbert S. Wilf. *Mathematics for the Physical Sciences*. John Wiley and Sons, Inc., New York – London – Sydney, 1962.

[118] James H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, UK, 1965.

[119] Paul R. Willems and Bruno Lang. A framework for the MR^3 algorithm: Theory and implementation. *SIAM J. Sci. Comput.*, 35(2):A740–A766, 2013.

[120] Wolfram Research, Inc. Mathematica 5.2, 1988–2005. Mathematica is a registered trademark of Wolfram Research, Inc.

[121] Yunkai Zhou and Yousef Saad. A Chebyshev–Davidson algorithm for large symmetric eigenproblems. *SIAM J. Matrix Anal. Appl.*, 29(3):954–971, 2007.

[122] Yunkai Zhou, Yousef Saad, Murilo L. Tiago, and James R. Chelikowsky. Self-consistent-field calculations using Chebyshev-filtered subspace iteration. *J. Comp. Phys.*, 219(1):172–184, 2006.

[123] Peizhen Zhu and Andrew V. Knyazev. Principal angles between subspaces and their tangents. Technical Report 2012-058, Mitsubishi Electric Research Laboratories, 2012. http://www.merl.com/publications/TR2012-058.