Steffen Börm
Christian Mehl

Numerical Methods for Eigenvalue Problems

De Gruyter

Mathematics Subject Classification 2010: Primary: 15A18, 15A22, 15A23, 15A42, 65F15, 65F25; *Secondary:* 65N25, 15B57.

ISBN 978-3-11-025033-6
e-ISBN 978-3-11-025037-4

Library of Congress Cataloging-in-Publication Data

A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the internet at http://dnb.d-nb.de.

Typesetting: Da-TeX Gerd Blumenstein, Leipzig, www.da-tex.de
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
∞ Printed on acid-free paper

Printed in Germany

www.degruyter.com

Preface

Eigenvalues and eigenvectors of matrices and linear operators play an important role when solving problems from structural mechanics, and electrodynamics, e.g., by describing the resonance frequencies of systems, when investigating the long-term behaviour of stochastic processes, e.g., by describing invariant probability measures, and as a tool for solving more general mathematical problems, e.g., by diagonalizing ordinary differential equations or systems from control theory.

This book presents a number of the most important numerical methods for finding eigenvalues and eigenvectors of matrices. It is based on lecture notes of a short course for third-year students in mathematics, but it should also be accessible to students of physics or engineering sciences.

We discuss the central ideas underlying the different algorithms and introduce the theoretical concepts required to analyze their behaviour. Our goal is to present an easily accessible introduction to the field, including rigorous proofs of all important results, but not a complete overview of the vast body of research.

For an in-depth coverage of the theory of eigenvalue problems, we can recommend the following monographs:

- J. H. Wilkinson, "The Algebraic Eigenvalue Problem" [52]
- B. N. Parlett, "The Symmetric Eigenvalue Problem" [33]
- G. H. Golub and C. F. Van Loan, "Matrix Computations" [18]
- G. W. Stewart, "Matrix Algorithms" [43, 44]
- D. S. Watkins, "The matrix eigenvalue problem" [49]

We owe a great debt of gratitude to their authors, since this book would not exist without their work.

The book is intended as the basis for a short course (one semester or trimester) for third- or fourth-year undergraduate students. We have organized the material mostly in short sections that should each fit one session of a course. Some chapters and sections are marked by an asterisk *. These contain additional results that we consider optional, e.g., rather technical proofs of general results or algorithms for special problems. With one exception, the results of these optional sections are not required for the remainder of the book. The one exception is the optional Section 2.7 on non-unitary transformations, which lays the groundwork for the optional Section 4.8 on the convergence of the power iteration for general matrices.

In order to keep the presentation self-contained, a number of important results are proven only for special cases, e.g., for self-adjoint matrices or a spectrum consisting only of simple eigenvalues. For the general case, we would like to refer the reader to the monographs mentioned above.

Deviating from the practice of collecting fundamental results in a separate chapter, we introduce some of these results when they are required. An example is the Bauer–Fike theorem given as Proposition 3.11 in Section 3.4 on error estimates for the Jacobi iteration instead of in a separate chapter on perturbation theory. While this approach is certainly not adequate for a reference work, we hope that it improves the accessibility of lecture notes like this book that are intended to be taught in sequence.

We would like to thank Daniel Kressner for his valuable contributions to this book.

Kiel, December 2011
Steffen Börm

Berlin, December 2011
Christian Mehl

Contents

Chapter 1

Introduction

Eigenvalue problems play an important role in a number of fields of numerical mathematics: in structural mechanics and electrodynamics, eigenvalues correspond to resonance frequencies of systems, i.e., to frequencies to which these systems respond particularly well (or badly, depending on the context). When studying stochastic processes, invariant probability measures correspond to eigenvectors for the eigenvalue 1, and finding them yields a description of the long-term behaviour of the corresponding process.

This book gives an introduction to the basic theory of eigenvalue problems and focuses on important algorithms for finding eigenvalues and eigenvectors.

1.1 Example: Structural mechanics

Before we consider abstract eigenvalue problems, we turn our attention to some applications that lead naturally to eigenvalue problems.

The first application is the investigation of resonance frequencies. As an example, we consider the oscillations of a string of unit length. We represent the string as a function

$$u : \mathbb{R} \times [0,1] \to \mathbb{R}, \quad (t,x) \mapsto u(t,x),$$

where $u(t,x)$ denotes the deflection of the string at time t and position x (cf. Figure 1.1).

The oscillations are then described by the *wave equation*

$$\frac{\partial^2 u}{\partial t^2}(t,x) = c\frac{\partial^2 u}{\partial x^2}(t,x) \quad \text{for all } t \in \mathbb{R},\ x \in (0,1),$$

where $c > 0$ is a parameter describing the string's properties (e.g., its thickness). We assume that the string is fixed at both ends, i.e., that

$$u(t,0) = u(t,1) = 0 \quad \text{holds for all } t \in \mathbb{R}.$$

Since the differential equation is linear, we can separate the variables: we write u in the form

$$u(t,x) = u_0(x)\cos(\omega t) \quad \text{for all } t \in \mathbb{R},\ x \in [0,1]$$

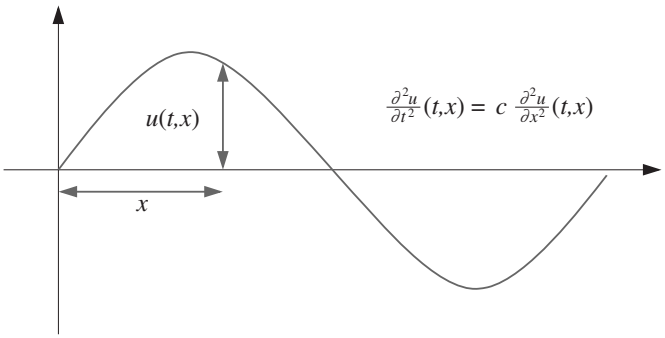

Figure 1.1. Mathematical model of a string.

with a *frequency parameter* $\omega \in \mathbb{R}_{\geq 0}$ and a function

$$u_0 : [0,1] \to \mathbb{R}, \quad x \mapsto u_0(x),$$

depending only on the location, but not on the time. The differential equation takes the form

$$\begin{aligned} -\omega^2 u_0(x)\cos(\omega t) &= \frac{\partial^2 u}{\partial t^2}(t,x) = c\frac{\partial^2 u}{\partial x^2}(t,x) \\ &= c u_0''(x)\cos(\omega t) \quad \text{for all } t \in \mathbb{R},\ x \in (0,1), \end{aligned}$$

and eliminating the time-dependent factor yields

$$-c u_0''(x) = \omega^2 u_0(x) \quad \text{for all } x \in (0,1). \tag{1.1}$$

We introduce $\lambda := \omega^2 \in \mathbb{R}_{\geq 0}$ and define the differential operator L by

$$L[u_0](x) := -c u_0''(x) \quad \text{for all } u_0 \in C^2(0,1),\ x \in (0,1)$$

in order to obtain

$$L[u_0] = \lambda u_0.$$

This is an eigenvalue problem in the infinite-dimensional space $C^2(0,1)$.

In order to be able to treat it by a numerical method, we have to *discretize* the problem. A simple approach is the *finite difference method*: Taylor expansion yields

$$\begin{aligned} u_0(x+h) &= u_0(x) + h u_0'(x) + \frac{h^2}{2}u_0''(x) + \frac{h^3}{6}u_0^{(3)}(x) + \frac{h^4}{24}u_0^{(4)}(\eta_+), \\ u_0(x-h) &= u_0(x) - h u_0'(x) + \frac{h^2}{2}u_0''(x) - \frac{h^3}{6}u_0^{(3)}(x) + \frac{h^4}{24}u_0^{(4)}(\eta_-) \end{aligned}$$

for $h \in \mathbb{R}_{>0}$ with $0 \leq x-h \leq x+h \leq 1$, where $\eta_+ \in [x, x+h]$ and $\eta_- \in [x-h, x]$. Adding both equations and using the intermediate value theorem yields

$$u_0(x-h) - 2u_0(x) + u_0(x+h) = h^2 u_0''(x) + \frac{h^4}{12} u_0^{(4)}(\eta)$$

with $\eta \in [x-h, x+h]$. Dividing by h^2 gives us an equation for the second derivative:

$$\frac{u_0(x-h) - 2u_0(x) + u_0(x+h)}{h^2} = u_0''(x) + \frac{h^2}{12} u_0^{(4)}(\eta).$$

We obtain a useful approximation by dropping the right-most term: fixing $n \in \mathbb{N}$ and setting

$$x_k := hk, \quad h := \frac{1}{n+1} \quad \text{for all } k \in \{0, \ldots, n+1\},$$

we find

$$\frac{u_0(x_{k-1}) - 2u_0(x_k) + u_0(x_{k+1})}{h^2} \approx u_0''(x_k) \quad \text{for all } k \in \{1, \ldots, n\},$$

and the term on the left-hand side requires only values of u_0 in the discrete points $x_0, \ldots, x_{n+1}$. We collect these values in a vector

$$e := \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_n) \end{pmatrix}, \quad e_0 = e_{n+1} = 0,$$

and replace $u_0''(x)$ in (1.1) by the approximation to get

$$c\frac{2e_k - e_{k-1} - e_{k+1}}{h^2} \approx \lambda e_k \quad \text{for all } k \in \{1, \ldots, n\}.$$

This system can be written as the algebraic eigenvalue problem

$$\frac{c}{h^2} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ \vdots \\ e_n \end{pmatrix} \approx \lambda \begin{pmatrix} e_1 \\ \vdots \\ \vdots \\ e_n \end{pmatrix}, \tag{1.2}$$

and solving the system yields approximations $u_0(x_k) \approx e_k$ of the values of u_0 in the points $x_1, \ldots, x_n$.

In order to reach a high accuracy, we have to ensure that h is small, so we have to be able to handle large values of n. We are typically only interested in computing a small number of the smallest eigenvalues, and this problem can be solved efficiently by specialized algorithms (e.g., the inverse iteration discussed in Chapter 4).

1.2 Example: Stochastic processes

The next example is not motivated by physics, but by computer science: we are interested in determining the "most important" pages of the world wide web. Let $n \in \mathbb{N}$ be the number of web pages, and let $L \in \mathbb{R}^{n\times n}$ represent the hyperlinks between these pages in the following way:

$$\ell_{ij} = \begin{cases} 1 & \text{if page } j \text{ contains a link to page } i, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \{1, \dots, n\}.$$

We follow the PageRank [31] approach: we consider a "random web user" that moves from page to page and compute the probability $p_j^{(m)} \in [0, 1]$ of him visiting a certain page j in his m-th step. We denote the number of links on page j by

$$\ell_j := \sum_{i=1}^{n} \ell_{ij} \quad \text{for all } j \in \{1, \dots, n\}$$

and assume that the random user chooses each of the links with equal probability $1/\ell_j$. In order to make this approach feasible, we have to assume $\ell_j \neq 0$ for all $j \in \{1, \dots, n\}$, i.e., we have to assume that each page contains at least one link.

This means that the probability of switching from page j to page i is given by

$$s_{ij} := \frac{\ell_{ij}}{\ell_j} = \begin{cases} 1/\ell_j & \text{if } \ell_{ij} = 1, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \{1, \dots, n\}.$$

The probability of visiting page i in step $m + 1$ is given by

$$p_i^{(m+1)} = \sum_{j=1}^{n} s_{ij}\, p_j^{(m)} \quad \text{for all } i \in \{1, \dots, n\},\ m \in \mathbb{N}_0.$$

Since this equation corresponds to a matrix-vector multiplication by $S = (s_{ij})_{i,j=1}^{n}$, we can write it in the compact form

$$p^{(m+1)} = Sp^{(m)} \quad \text{for all } m \in \mathbb{N}_0. \tag{1.3}$$

In order to ensure that the result does not depend on the arbitrarily chosen starting vector $p^{(0)}$, the PageRank algorithm uses the limit

$$p^* := \lim_{m\to\infty} p^{(m)} \tag{1.4}$$

to determine the "importance" of a web page: if p_j^* is large, the probability of a user visiting the j-th web page is high, therefore it is assumed to be important. Due to

$$p^* = \lim_{m\to\infty} p^{(m)} = \lim_{m\to\infty} p^{(m+1)} = \lim_{m\to\infty} Sp^{(m)} = S \lim_{m\to\infty} p^{(m)} = Sp^*,$$

the vector $p^* \in \mathbb{R}^n$ is an eigenvector of the matrix S for the eigenvalue one.

In this example, we not only reduce a problem related to stochastic processes (the "random walk" of the web user) to an algebraic eigenvalue problem, but we also find a simple algorithm for computing at least the eigenvector: due to (1.4), we can hope to approximate p^* by computing $p^{(m)}$ for a sufficiently large value of m. Due to (1.3), this requires only m matrix-vector multiplications, and since we can assume that each web page contains only a small number of links, these multiplications can be performed very efficiently.

In order to ensure convergence, the PageRank algorithm replaces the matrix S by the matrix $\widehat{S} = (\hat{s}_{ij})_{i,j=1}^n$ given by

$$\hat{s}_{ij} = (1-\alpha)s_{ij} + \alpha u_i \quad \text{for all } i, j \in \{1, \dots, n\},$$

where $\alpha \in (0, 1]$ (a typical value is 0.15) is a parameter controlling how close $\widehat{S}$ is to the original matrix S, while $u \in \mathbb{R}^n$ is a vector (the "teleportation vector") describing the event that a user switches to a different page without following a link: when visiting page j, the user either follows one of the links with a total probability of $1-\alpha$ or switches to another page with a total probability of α. In the latter case, u_i is the relative probability of switching to page i. For a suitable choice of u (e.g., a vector with strictly positive entries and $u_1 + \ldots + u_n = 1$), the *Perron–Frobenius theorem* [34] ensures convergence of the sequence $(p^{(m)})_{m=0}^{\infty}$ to p^*.

1.3 Example: Systems of linear differential equations

Another, more general example for eigenvalue problems are systems of linear differential equations that appear frequently in natural and engineering sciences. If $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $A \in \mathbb{F}^{n\times n}$ then

$$y' = Ay \tag{1.5}$$

is a *(homogeneous) system of linear differential equations* and a solution is defined to be a continuously differentiable function $y : \mathbb{R} \to \mathbb{F}^n$ satisfying $y'(t) = Ay(t)$ for all $t \in \mathbb{R}$. For some vector $y_0 \in \mathbb{F}^n \setminus \{0\}$ the ansatz $y(t) = e^{\lambda t} y_0$ yields the identity

$$\lambda e^{\lambda t} y_0 = y'(t) = Ay(t) = Ae^{\lambda t} y_0$$

which, after division on both sides by $e^{\lambda t}$, reduces to the characteristic equation

$$\lambda y_0 = Ay_0.$$

Thus, if $\lambda \in \mathbb{F}$ is an eigenvalue of A and $y_0 \in \mathbb{F}^n \setminus \{0\}$ is an associated eigenvector, then $y(t) = e^{\lambda t} y_0$ is a solution of the corresponding system of differential equations. It is well known from the theory of differential equations that if A is diagonalizable and if $v_1, \dots, v_n$ is a basis of $\mathbb{F}^n$ consisting of eigenvectors of A associated with

the eigenvalues $\lambda_1, \ldots, \lambda_n$, then any solution y of the system of differential equations (1.5) has the form

$$y(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t} v_i$$

for some coefficients $c_1, \ldots, c_n \in \mathbb{F}$. In the non-diagonalizable case, the general solution can be constructed from the so called *Jordan normal form*.

Instead of systems of linear differential equations of first order as in the form (1.5), one may also consider systems of linear differential equations of higher order having the general form

$$\sum_{k=0}^{\ell} A_k y^{(k)} = A_\ell y^{(\ell)} + A_{\ell-1} y^{(\ell-1)} + \cdots + A_2 y'' + A_1 y' + A_0 y = 0,$$

where $A_0, \ldots, A_\ell \in \mathbb{F}^{n\times n}$. In this case the ansatz $y(t) = e^{\lambda t} y_0$ for some nonzero vector $y_0 \in \mathbb{F}^n$ yields the identity

$$\sum_{k=0}^{\ell} A_k \lambda^k e^{\lambda t} y_0 = 0,$$

or equivalently, after division by $e^{\lambda t}$,

$$\left(\sum_{k=0}^{\ell} \lambda^k A_k \right) y_0 = 0. \tag{1.6}$$

The problem of solving (1.6) is called a *polynomial eigenvalue problem*.

A particular example in applications can be found in the theory of mechanical vibration. The equations of motion for a viscously damped linear system with n degrees of freedom are given by

$$My''(t) + Cy'(t) + Ky(t) = 0,$$

where M, C, K are $n \times n$ matrices called mass matrix, damping matrix, and stiffness matrix, respectively. The corresponding *quadratic eigenvalue problem* has the form

$$(\lambda^2 M + \lambda C + K)x = 0.$$

A simple example for the case $n = 1$ is the spring-mass system with damping by friction, see Figure 1.2.

By Hooke's law, the equation of motion for this system without friction is

$$my''(t) + ky(t) = 0,$$

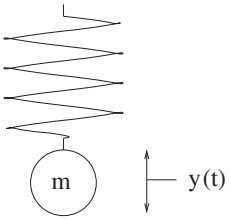

Figure 1.2. Mass-spring system.

where m is the mass attached to the spring and k is the spring constant. If friction is considered, it is usually modeled in such a way that the friction is assumed to be proportional to the velocity $y'(t)$ thus yielding the equation of motion

$$my''(t) + cy'(t) + ky(t) = 0,$$

for some constant c.

Chapter 2

Existence and properties of eigenvalues and eigenvectors

Summary

This chapter investigates existence and uniqueness of eigenvalues and eigenvectors for a given matrix. The key result is the *Schur decomposition* introduced in Theorem 2.46, a very useful tool for the investigation of eigenvalue problems. One of its most important consequences is the fact that a matrix can be diagonalized unitarily if and only if it is normal. The optional Section 2.7 presents a block-diagonalization result for general square matrices

Learning targets

- ✓ Recall the definition and some of the most important properties of eigenvalues, eigenvectors, similarity transformations and the characteristic polynomial corresponding to a matrix.
- ✓ Introduce a number of basic concepts of Hilbert space theory, e.g., the Cauchy–Schwarz inequality, self-adjoint, normal, isometric and unitary matrices.
- ✓ Prove the existence of the Schur decomposition.
- ✓ Use it to establish the existence of eigenvector bases for normal and self-adjoint matrices and of invariant subspaces in the general case.

2.1 Eigenvalues and eigenvectors

Let $n, m \in \mathbb{N}$, and let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ be the field of real or complex numbers.

We denote the space of matrices with n rows and m columns by $\mathbb{F}^{n\times m}$. The coefficients of a matrix $A \in \mathbb{F}^{n\times m}$ are given by

$$A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix}.$$

Zero coefficients in a matrix are frequently omitted in our notation, e.g., the n-dimensional identity matrix is usually represented by

$$I_n = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}.$$

If the dimension is clear from the context, we denote the identity matrix by I.

The product of a matrix by a vector $x \in \mathbb{F}^m$ is given by

$$Ax = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1m}x_m \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nm}x_m \end{pmatrix}$$

or, more precisely, by

$$(Ax)_i = \sum_{j=1}^{m} a_{ij}x_j \quad \text{for all } i \in \{1, \dots, n\}. \tag{2.1}$$

The mapping

$$\mathbb{F}^m \to \mathbb{F}^n, \quad x \mapsto Ax$$

is a linear operator mapping $\mathbb{F}^m$ to $\mathbb{F}^n$, and we use the corresponding notations:

Definition 2.1 (Null space and range). Let $A \in \mathbb{F}^{n\times m}$. The space

$$\mathcal{N}(A) := \{x \in \mathbb{F}^m \ : \ Ax = 0\}$$

is called the *null space* of A, and its dimension is called the *nullity* of A. The space

$$\mathcal{R}(A) := \{y \in \mathbb{F}^n \ : \ \text{there exists a vector } x \in \mathbb{F}^m \text{ with } Ax = y\}$$

is called the *range* of A, and its dimension is called the *rank* of A.

The matrix A is called *injective* if $\mathcal{N}(A) = \{0\}$ holds, and it is called *surjective* if $\mathcal{R}(A) = \mathbb{F}^n$ holds.

We recall the *rank-nullity theorem*: let $(y_i)_{i=1}^k$ be a basis of $\mathcal{R}(A)$. By definition, we can find $(\widehat{x}_i)_{i=1}^k$ in $\mathbb{F}^m$ such that

$$A\widehat{x}_i = y_i \quad \text{for all } i \in \{1, \dots, k\}.$$

Since the family $(y_i)_{i=1}^k$ is linearly independent, the same holds for $(\widehat{x}_i)_{i=1}^k$. This means that we can expand the family to a basis $(\widehat{x}_i)_{i=1}^m$ of $\mathbb{F}^m$. For each

$j \in \{k+1, \ldots, m\}$, we obviously have $A\widehat{x}_j \in \mathcal{R}(A)$ and can therefore find $z_j \in \operatorname{span}\{\widehat{x}_1, \ldots, \widehat{x}_k\}$ such that $A\widehat{x}_j = Az_j$ holds, i.e., $\widehat{x}_j - z_j \in \mathcal{N}(A)$. We define

$$x_i := \begin{cases} \widehat{x}_i & \text{if } i \leq k, \\ \widehat{x}_i - z_i & \text{otherwise} \end{cases} \qquad \text{for all } i \in \{1, \ldots, m\}$$

and see that $(x_i)_{i=1}^m$ is a basis of $\mathbb{F}^m$ such that $\operatorname{span}\{x_{k+1}, \ldots, x_m\} \subseteq \mathcal{N}(A)$ holds. This yields $\dim \mathcal{N}(A) = m - k$, i.e.,

$$\dim \mathcal{R}(A) + \dim \mathcal{N}(A) = m \qquad \text{for all } A \in \mathbb{F}^{n\times m}. \tag{2.2}$$

Definition 2.2 (Eigenvalue and Eigenvector). Let $A \in \mathbb{F}^{n\times n}$, and let $\lambda \in \mathbb{F}$. λ is called an *eigenvalue* of A, if there is a vector $x \in \mathbb{F}^n \setminus \{0\}$ such that

$$Ax = \lambda x \tag{2.3}$$

holds. Any such vector is called an *eigenvector* of A for the eigenvalue λ. A pair (λ, x) consisting of an eigenvalue and a corresponding eigenvector is called an *eigenpair*.

The set

$$\sigma(A) := \{\lambda \in \mathbb{F} \ : \ \lambda \text{ is an eigenvalue of } A\}$$

is called the *spectrum* of A.

Let $k \in \mathbb{N}$. The *product* AB of two matrices $A \in \mathbb{F}^{n\times k}$ and $B \in \mathbb{F}^{k\times m}$ is given by

$$(AB)_{ij} = \sum_{\ell=1}^{k} a_{i\ell} b_{\ell j} \qquad \text{for all } i \in \{1, \ldots, n\},\ j \in \{1, \ldots, m\}. \tag{2.4}$$

The definition ensures

$$ABx = A(Bx) \qquad \text{for all } x \in \mathbb{F}^m,$$

i.e., it is compatible with the matrix-vector multiplication.

Exercise 2.3 (Polynomials). Let $A \in \mathbb{F}^{n\times n}$. We define the *$\ell$-th power* of the matrix by

$$A^\ell := \begin{cases} I & \text{if } \ell = 0, \\ AA^{\ell-1} & \text{otherwise} \end{cases} \qquad \text{for all } \ell \in \mathbb{N}_0.$$

This definition allows us to apply polynomials to matrices: for each polynomial

$$p(t) = a_0 + a_1 t + \cdots + a_m t^m,$$

we define

$$p(A) := a_0 A^0 + a_1 A^1 + \cdots + a_m A^m.$$

Prove $p(\sigma(A)) \subseteq \sigma(p(A))$. Can you find A and p with $\sigma(p(A)) \neq p(\sigma(A))$? Hint: consider Exercise 2.6.

Exercise 2.4 (Projection). Let $P \in \mathbb{F}^{n\times n}$ be a *projection*, i.e., let it satisfy $P^2 = P$. Prove $\sigma(P) \subseteq \{0, 1\}$.

Is any matrix $A \in \mathbb{F}^{n\times n}$ satisfying $\sigma(A) \subseteq \{0, 1\}$ a projection?
Hint: consider $A \in \mathbb{F}^{2\times 2}$ with $a_{21} = 0$.

Exercise 2.5 (Nil-potent matrix). Let $N \in \mathbb{F}^{n\times n}$ be a *nil-potent* matrix, i.e., let there be a $k \in \mathbb{N}$ with $N^k = 0$. Prove $\sigma(N) \subseteq \{0\}$.

Exercise 2.6 (Empty spectrum). Let $\mathbb{F} = \mathbb{R}$. Consider the matrix

$$A := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

Prove $\sigma(A) = \emptyset$, i.e., show that A has no eigenvalues. Does the situation change if we let $\mathbb{F} = \mathbb{C}$?

If x is an eigenvector of a matrix $A \in \mathbb{F}^{n\times n}$, multiplying x by any non-zero number will again yield an eigenvector. Instead of dealing with non-unique eigenvectors, it is preferable to use an alternative characterization of eigenvalues:

Proposition 2.7 (Null space). *Let $A \in \mathbb{F}^{n\times n}$, and let $\lambda \in \mathbb{F}$. λ is an eigenvalue of A if and only if*

$$\mathcal{N}(\lambda I - A) \neq \{0\}$$

holds, i.e., if $\lambda I - A$ is not injective.

Proof. We first observe that for all $x \in \mathbb{F}^n$ and all $\lambda \in \mathbb{F}$, the following statements are equivalent:

$$\begin{aligned} \lambda x &= Ax, \\ \lambda x - Ax &= 0, \\ (\lambda I - A)x &= 0, \\ x &\in \mathcal{N}(\lambda I - A). \end{aligned}$$

If $\lambda \in \mathbb{F}$ is an eigenvalue, we can find a corresponding eigenvector $x \in \mathbb{F}^n \setminus \{0\}$, i.e., we have $Ax = \lambda x$, and therefore $x \in \mathcal{N}(\lambda I - A)$ and $\mathcal{N}(\lambda I - A) \supseteq \{0, x\} \neq \{0\}$.

If, on the other hand, $\mathcal{N}(\lambda I - A) \neq \{0\}$ holds, we can find $x \in \mathcal{N}(\lambda I - A) \setminus \{0\}$, and this vector x is an eigenvector. □

Since the null space of $\lambda I - A$ is uniquely determined by A and λ, working with it instead of individual eigenvectors offers significant advantages both for practical and theoretical investigations.

Definition 2.8 (Eigenspace). Let $A \in \mathbb{F}^{n\times n}$, and let $\lambda \in \sigma(A)$. Then

$$\mathcal{E}(A,\lambda) := \mathcal{N}(\lambda I - A)$$

is called the *eigenspace* of A for the eigenvalue λ.

Definition 2.9 (Geometric multiplicity). Let $A \in \mathbb{F}^{n\times n}$, and let $\lambda \in \sigma(A)$. The dimension of the eigenspace $\mathcal{E}(A,\lambda)$ is called the *geometric multiplicity* of λ and denoted by $\mu_g(A,\lambda)$.

Instead of looking for individual eigenvectors, we look for a basis of an eigenspace. This offers the advantage that we can change the basis during the course of our algorithms in order to preserve desirable properties like isometry or non-degeneracy.

Exercise 2.10 (Eigenspaces). Let $A \in \mathbb{F}^{n\times n}$, and let $\lambda, \mu \in \sigma(A)$ with $\lambda \neq \mu$. Prove

$$\mathcal{E}(A,\lambda) \cap \mathcal{E}(A,\mu) = \{0\}.$$

Exercise 2.11 (Geometric multiplicity). Let $A \in \mathbb{F}^{n\times n}$. Prove

$$\sum_{\lambda\in\sigma(A)} \mu_g(A,\lambda) \leq n.$$

2.2 Characteristic polynomials

By the rank-nullity theorem (2.2), λ is an eigenvalue of a matrix A if and only if $\lambda I - A$ is not invertible. Using this property, we can characterize the eigenvalues without explicitly constructing eigenvectors.

Let $n, m \in \mathbb{N}$, and let $j \in \{1,\dots,m\}$. The *j-th canonical unit vector* $\delta_j \in \mathbb{F}^m$ is given by

$$(\delta_j)_i := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i \in \{1,\dots,m\}, \tag{2.5}$$

and for a matrix $A \in \mathbb{F}^{n\times m}$, we denote the *$j$-th column vector* by

$$a_j := A\delta_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}.$$

The matrix A is injective if and only if its columns $a_1,\dots,a_m$ are linearly independent.

We can use the *determinant* to characterize tuples of linearly independent vectors. The determinant is a mapping

$$\det : (\mathbb{F}^n)^n \to \mathbb{F}$$

of n-tuples of n-dimensional vectors to scalar values that is *multilinear*, i.e., we have

$$\begin{aligned}&\det(x_1, \ldots, x_{j-1}, x_j + \alpha z, x_{j+1}, \ldots, x_n)\\ &\quad = \det(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n)\\ &\qquad + \alpha \det(x_1, \ldots, x_{j-1}, z, x_{j+1}, \ldots, x_n)\\ &\quad \text{for all } x_1, \ldots, x_n, z \in \mathbb{F}^n,\ \alpha \in \mathbb{F},\ j \in \{1, \ldots, n\}.\end{aligned}$$

The determinant is also *alternating*, i.e., we have

$$\begin{aligned}&\det(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n)\\ &\quad = -\det(x_1, \ldots, x_{i-1}, x_j, x_{i+1}, \ldots, x_{j-1}, x_i, x_{j+1}, \ldots, x_n)\\ &\qquad \text{for all } x_1, \ldots, x_n \in \mathbb{F}^n,\ i, j \in \{1, \ldots, n\} \text{ with } i < j\end{aligned}$$

and satisfies $\det(\delta_1, \ldots, \delta_n) = 1$.

Let $x_1, \ldots, x_n \in \mathbb{F}^n$. If there are $i, j \in \{1, \ldots, n\}$ with $i < j$ and $x_i = x_j$, the fact that the determinant is alternating implies

$$\begin{aligned}&\det(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n)\\ &\quad = -\det(x_1, \ldots, x_{i-1}, x_j, x_{i+1}, \ldots, x_{j-1}, x_i, x_{j+1}, \ldots, x_n)\\ &\quad = -\det(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n),\end{aligned}$$

and therefore $\det(x_1, \ldots, x_n) = 0$. Since the determinant is also multilinear, we can add any multiple of any argument to any other argument without changing the value of the determinant. In particular, it is possible to prove that the determinant is equal to zero if and only if its arguments are linearly dependent.

This means that a matrix $A \in \mathbb{F}^{n \times n}$ is invertible if and only if $\det(a_1, \ldots, a_n) \neq 0$, i.e., if the determinant of its column vectors vanishes. We extend the definition of the determinant to quadratic matrices by setting

$$\det : \mathbb{F}^{n \times n} \to \mathbb{F}, \quad A \mapsto \det(a_1, \ldots, a_n),$$

and have $\det(A) \neq 0$ if and only if A is invertible. Combining this property with Proposition 2.7, we obtain a characterization of eigenvalues that does not require eigenvectors:

Definition 2.12 (Characteristic polynomial). Let $A \in \mathbb{F}^{n \times n}$.

$$p_A : \mathbb{F} \to \mathbb{F}, \quad t \mapsto \det(tI - A),$$

is a polynomial of degree n. We call it the *characteristic polynomial* of A.

Proposition 2.13 (Zeros of p_A). *Let $A \in \mathbb{F}^{n\times n}$. $\lambda \in \mathbb{F}$ is an eigenvalue of A if and only if $p_A(\lambda) = 0$ holds.*

Proof. Let $\lambda \in \mathbb{F}$. If λ is an eigenvalue of A, Proposition 2.7 implies that $\lambda I - A$ is not injective, therefore this matrix has to be non-invertible, and we have $p_A(\lambda) = \det(\lambda I - A) = 0$.

If, on the other hand, we have $0 = p_A(\lambda) = \det(\lambda I - A)$, the matrix $\lambda I - A$ is non-invertible. Since it is a quadratic matrix, the rank-nullity theorem (2.2) implies that it cannot be injective, and Proposition 2.7 yields that λ has to be an eigenvalue. □

This result allows us to characterize the spectrum of a matrix $A \in \mathbb{F}^{n\times n}$ as the set of zeros of its characteristic polynomial:

$$\sigma(A) = \{\lambda \in \mathbb{F} \ : \ p_A(\lambda) = 0\}.$$

Given a polynomial p and a λ with $p(\lambda) = 0$, we can find a polynomial q such that

$$p(t) = (\lambda - t)q(t) \quad \text{for all } t \in \mathbb{F}.$$

If λ is a zero of q, we can apply this construction repeatedly to find the maximal power $k \in \mathbb{N}$ such that $(\lambda - t)^k$ is a divisor of p: the *multiplicity* of λ is the number $k \in \mathbb{N}$ uniquely defined by the property that there is a polynomial q satisfying

$$\begin{aligned} p(t) &= (\lambda - t)^k q(t) \quad \text{for all } t \in \mathbb{F}, \\ q(\lambda) &\neq 0. \end{aligned}$$

Definition 2.14 (Algebraic multiplicity). Let $A \in \mathbb{F}^{n\times n}$ and $\lambda \in \sigma(A)$. By Proposition 2.13, λ is a zero of p_A. We call its multiplicity the *algebraic multiplicity* of the eigenvalue λ and denote it by $\mu_a(A, \lambda)$.

If $\mu_a(A, \lambda) = 1$, λ is called a *simple eigenvalue*.

Exercise 2.15 (Algebraic multiplicity). Let $n \in \mathbb{N}$ and $A \in \mathbb{C}^{n\times n}$. Prove

$$\sum_{\lambda\in\sigma(A)} \mu_a(A, \lambda) = n. \tag{2.6}$$

Hint: the fundamental theorem of algebra states that every non-constant complex-valued polynomial has at least one root.

Exercise 2.16 (Companion matrix). Let $n \in \mathbb{N}_{\geq 2}$ and $c_0, c_1, \ldots, c_{n-1} \in \mathbb{F}$. Let

$$C := \begin{pmatrix} 0 & 1 & & & \\ 0 & 0 & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 0 & \ldots & 0 & 1 \\ -c_0 & -c_1 & \ldots & -c_{n-2} & -c_{n-1} \end{pmatrix}.$$

Prove

$$p_C(t) = c_0 + c_1 t + \cdots + c_{n-1} t^{n-1} + t^n \quad \text{for all } t \in \mathbb{F}.$$

This means that, given a monic polynomial p of order n, we can find a matrix $C \in \mathbb{F}^{n\times n}$ such that $p_C = p$ holds, i.e., the task of finding the zeros of a polynomial and the task of finding the eigenvalues of a matrix are equivalent.

The matrix C is called the *companion matrix* of the polynomial p.

(Hint: use Laplace's theorem and induction).

The Abel–Ruffini theorem states that there is no general closed-form algebraic solution to polynomial equations of degree five or higher. Since finding the roots of a polynomial is equivalent to finding the eigenvalues of the companion matrix, we cannot hope to find a general algorithm for computing the exact eigenvalues of matrices of dimension five or higher.

In fact, all practical algorithms compute arbitrarily accurate *approximations* of eigenvalues and eigenvectors.

2.3 Similarity transformations

Many popular algorithms for solving systems of linear equations are based on transforming the given matrix to a simple form (e.g., upper triangular) that can be handled efficiently. We are interested in following a similar approach.

Let $A \in \mathbb{F}^{n\times n}$, let $\lambda \in \sigma(A)$, and let $x \in \mathbb{F}^n \setminus \{0\}$ be an eigenvector of A for the eigenvalue λ. By definition (2.3), we have

$$Ax = \lambda x.$$

Let now $B \in \mathbb{F}^{n\times n}$ be an invertible matrix. Multiplying our equation on both sides with B^{-1} yields

$$B^{-1}Ax = \lambda B^{-1}x,$$

but since B^{-1} appears on the right-hand side, this equation is no longer related to eigenvalue problems. We fix this problem by introducing $\widehat{x} := B^{-1}x$ and find

$$B^{-1}AB\widehat{x} = \lambda\widehat{x}.$$

This is again an eigenvalue problem. Instead of looking for eigenvalues and eigenvectors of A, we can also look for eigenvalues and eigenvectors of the transformed matrix $\widehat{A} := B^{-1}AB$. Any eigenvalue of $\widehat{A}$ is also an eigenvalue of A, and any eigenvector of $\widehat{A}$ can be transformed to an eigenvector of A by multiplying by B.

Definition 2.17 (Similar matrices). Let $A, \widehat{A} \in \mathbb{F}^{n\times n}$. We call A and $\widehat{A}$ *similar*, if there is a invertible matrix $B \in \mathbb{F}^{n\times n}$ such that

$$\widehat{A} = B^{-1}AB.$$

The mapping $A \mapsto B^{-1}AB$ is called a *similarity transformation*. It can be interpreted as a change of basis: the linear mapping defined by A is represented in the basis given by the columns of B. As explained above, a change of basis cannot change the eigenvalues, and we can see that a number of other important properties are also left unchanged:

Proposition 2.18 (Similar matrices). *Let $A, \widehat{A} \in \mathbb{F}^{n\times n}$ be similar matrices, and let $B \in \mathbb{F}^{n\times n}$ be invertible with $\widehat{A} = B^{-1}AB$. Then we have*

- $\sigma(A) = \sigma(\widehat{A})$,
- $\mathcal{E}(A,\lambda) = B\mathcal{E}(\widehat{A},\lambda)$ *for all* $\lambda \in \sigma(A) = \sigma(\widehat{A})$,
- $p_A = p_{\widehat{A}}$ *and*
- $\mu_g(A,\lambda) = \mu_g(\widehat{A},\lambda)$ *and* $\mu_a(A,\lambda) = \mu_a(\widehat{A},\lambda)$ *for all* $\lambda \in \sigma(A) = \sigma(\widehat{A})$.

Proof. Let $\lambda \in \sigma(A)$, and let $x \in \mathbb{F}^n \setminus \{0\}$ be a corresponding eigenvector. We let $\widehat{x} := B^{-1}x$ and observe $\widehat{x} \neq 0$, $B\widehat{x} = x$. We find

$$\widehat{A}\widehat{x} = B^{-1}ABB^{-1}x = B^{-1}Ax = \lambda B^{-1}x = \lambda\widehat{x},$$

so $\widehat{x}$ is an eigenvector of $\widehat{A}$, and therefore $\lambda \in \sigma(\widehat{A})$. This implies

$$\sigma(A) \subseteq \sigma(\widehat{A}), \quad \mathcal{E}(A,\lambda) \subseteq B\mathcal{E}(\widehat{A},\lambda) \quad \text{for all } \lambda \in \sigma(A).$$

We also have $A = B\widehat{A}B^{-1}$, so we can exchange the roles of A and $\widehat{A}$ in order to obtain

$$\sigma(A) = \sigma(\widehat{A}), \quad \mathcal{E}(A,\lambda) = B\mathcal{E}(\widehat{A},\lambda), \quad \mu_g(A,\lambda) = \mu_g(\widehat{A},\lambda) \quad \text{for all } \lambda \in \sigma(A).$$

Since the determinant of the product of two matrices is the product of their determinants, we have

$$\begin{aligned} p_{\widehat{A}}(t) &= \det(tI - \widehat{A}) = \det(tB^{-1}B - B^{-1}AB) = \det(B^{-1}(tI - A)B) \\ &= \det(B^{-1})\det(tI - A)\det(B) = \det(B^{-1})\det(B)\det(tI - A) \\ &= \det(B^{-1}B)p_A(t) = \det(I)p_A(t) = p_A(t) \quad \text{for all } t \in \mathbb{F}, \end{aligned}$$

and this completes the proof. □

Similarity transformations are an important tool for the investigation of eigenvalues and eigenvectors. As an example of its many uses we consider the relationship between the geometric and the algebraic multiplicity.

Proposition 2.19 (Geometric and algebraic multiplicity). *Let $A \in \mathbb{F}^{n\times n}$, and let $\lambda \in \sigma(A)$. Then we have*

$$\mu_g(A,\lambda) \leq \mu_a(A,\lambda).$$

Proof. Let $k := \mu_g(A,\lambda) = \dim \mathcal{N}(\lambda I - A) \in \mathbb{N}$. By definition, we can find k linearly independent vectors $e_1,\ldots,e_k \in \mathbb{F}^n$ spanning $\mathcal{N}(\lambda I - A)$, i.e., satisfying

$$Ae_j = \lambda e_j \quad \text{for all } j \in \{1,\ldots,k\}.$$

We can extend $(e_j)_{j=1}^k$ to a basis $(e_j)_{j=1}^n$ of $\mathbb{F}^n$. Then the matrix $E \in \mathbb{F}^{n\times n}$ defined by using the basis vectors as columns, i.e., by

$$e_{ij} := (e_j)_i \quad \text{for all } i,j \in \{1,\ldots,n\}, \tag{2.7}$$

is invertible and satisfies $E\delta_j = e_j$. Let

$$\widehat{A} := E^{-1}AE.$$

For $j \in \{1,\ldots,k\}$, we obtain

$$\widehat{A}\delta_j = E^{-1}AE\delta_j = E^{-1}Ae_j = E^{-1}\lambda e_j = \lambda E^{-1}E\delta_j = \lambda\delta_j$$

and therefore

$$\widehat{A} := E^{-1}AE = \begin{pmatrix} \lambda I_k & B \\ & C \end{pmatrix}$$

with $B \in \mathbb{F}^{k\times(n-k)}$ and $C \in \mathbb{F}^{(n-k)\times(n-k)}$. Since A and $\widehat{A}$ are similar, we can apply Proposition 2.18 to get

$$\begin{aligned} p_A(t) = p_{\widehat{A}}(t) = \det(tI_n - \widehat{A}) &= \det\begin{pmatrix} tI_k - \lambda I_k & B \\ & tI_{n-k} - C \end{pmatrix} \\ &= \det((t-\lambda)I_k)\det(tI_{k-n} - C) \\ &= (t-\lambda)^k \det(tI_{k-n} - t) \quad \text{for all } t \in \mathbb{F}, \end{aligned}$$

i.e., λ is a zero of p_A of multiplicity at least $k = \mu_g(A,\lambda)$. □

Exercise 2.20 (Multiplicities). Consider the matrix

$$A := \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

Compute its eigenvalues and their algebraic and geometric multiplicities.

Of particular interest are similarity transformations turning a matrix into a diagonal matrix, since these are very useful both for theoretical investigations and practical applications.

Definition 2.21 (Diagonalizable matrix). Let $A \in \mathbb{F}^{n\times n}$. The matrix A is called *diagonalizable* if it is similar to a diagonal matrix, i.e., if there are an invertible matrix $B \in \mathbb{F}^{n\times n}$ and a diagonal matrix $D \in \mathbb{F}^{n\times n}$ satisfying

$$A = BDB^{-1}.$$

Proposition 2.22 (Diagonalizable matrix). *Let $A \in \mathbb{F}^{n\times n}$. The matrix A is diagonalizable if and only if there is a basis of $\mathbb{F}^n$ consisting only of its eigenvectors. This is equivalent to*

$$\sum_{\lambda\in\sigma(A)} \mu_g(A,\lambda) = n.$$

Proof. Assume that a basis $(e_j)_{j=1}^n$ of eigenvectors exists, and let $(\lambda_j)_{j=1}^n$ be the corresponding eigenvalues. Then the matrix $E \in \mathbb{F}^{n\times n}$ defined by (2.7) is invertible and satisfies

$$E^{-1}AE\delta_j = E^{-1}Ae_j = E^{-1}\lambda_j e_j = \lambda_j E^{-1}E\delta_j = \lambda_j\delta_j \quad \text{for all } j \in \{1,\dots,n\},$$

so $D := E^{-1}AE$ is a diagonal matrix, and therefore A is diagonalizable.

Assume now that A is diagonalizable, and let $E \in \mathbb{F}^{n\times n}$ be a invertible matrix and $D \in \mathbb{F}^{n\times n}$ a diagonal matrix satisfying $A = EDE^{-1}$. We define vectors $(e_j)_{j=1}^n$ by

$$e_j := E\delta_j \quad \text{for all } j \in \{1,\dots,n\}$$

and obtain

$$Ae_j = EDE^{-1}e_j = ED\delta_j = Ed_{jj}\delta_j = d_{jj}e_j,$$

i.e., e_j is an eigenvector corresponding to the eigenvalue d_{jj}. Since E is invertible, the vectors $(e_j)_{j=1}^n$ are linearly independent and therefore a basis of $\mathbb{F}^n$. □

Exercise 2.23 (Complex diagonalizability). Let $n \in \mathbb{N}$ and $A \in \mathbb{C}^{n\times n}$. Prove that A is diagonalizable if and only if

$$\mu_g(A,\lambda) = \mu_a(A,\lambda) \quad \text{holds for all } \lambda \in \sigma(A).$$

Hint: Proposition 2.19 and Exercise 2.15 could help.

Exercise 2.24 (Matrix exponential). Let $A \in \mathbb{C}^{n\times n}$ be diagonalizable. Prove that the *matrix exponential*

$$\exp(A) := \sum_{m=0}^{\infty} \frac{1}{m!}A^m$$

is well-defined and diagonalizable.

The matrix exponential is useful for solving linear ordinary differential equations: prove

$$\frac{\partial}{\partial t} \exp(tA) = A \exp(tA) \quad \text{for all } t \in \mathbb{R}.$$

Exercise 2.25 (Harmonic oscillator). Consider the ordinary differential equation

$$u'(t) = v(t), \quad v'(t) = -u(t) \quad \text{for all } t \in \mathbb{R}.$$

Use Exercise 2.24 to describe the space of solutions $t \mapsto (u(t), v(t))$ explicitly.

2.4 Some properties of Hilbert spaces

Our next goal is to find criteria for checking whether a given matrix is diagonalizable. $\mathbb{F}^n$ is a *Hilbert space* with the inner product given by

$$\langle x, y \rangle := \sum_{j=1}^{n} x_j \bar{y}_j \quad \text{for all } x, y \in \mathbb{F}^n$$

and the norm given by

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{j=1}^{n} |x_j|^2} \quad \text{for all } x \in \mathbb{F}^n. \tag{2.8}$$

One of the most important consequences of the close relationship between the inner product and the norm is the *Cauchy–Schwarz inequality*. Its proof serves to illustrate typical Hilbert space arguments:

Proposition 2.26 (Cauchy–Schwarz inequality). *Let $x, y \in \mathbb{F}^n$. We have*

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|. \tag{2.9}$$

Both sides are equal if and only if x and y are linearly dependent.

Proof. We assume $y \neq 0$ without loss of generality. Let $\mu \in \mathbb{F}$. We have

$$\begin{aligned} 0 \leq \|x - \mu y\|^2 &= \langle x - \mu y, x - \mu y \rangle = \langle x, x \rangle - \langle x, \mu y \rangle - \langle \mu y, x \rangle + \langle \mu y, \mu y \rangle \\ &= \|x\|^2 - \bar{\mu} \langle x, y \rangle - \mu \overline{\langle x, y \rangle} + |\mu|^2 \|y\|^2. \end{aligned}$$

In order to minimize the right-hand side, we choose

$$\mu := \frac{\langle x, y \rangle}{\|y\|^2}$$

and obtain

$$0 \leq \|x - \mu y\|^2 = \|x\|^2 - 2\frac{|\langle x, y\rangle|^2}{\|y\|^2} + \frac{|\langle x, y\rangle|^2}{\|y\|^4}\|y\|^2 = \|x\|^2 - \frac{|\langle x, y\rangle|^2}{\|y\|^2}, \quad (2.10)$$

and multiplying both sides by $\|y\|^2$ yields the estimate (2.9).

If we have equality, i.e., if $|\langle x, y\rangle| = \|x\| \, \|y\|$ holds, (2.10) implies $\|x - \mu y\| = 0$. □

In the following we consider a number of properties of Hilbert spaces that prove useful for the investigation of eigenvalue problems.

Definition 2.27 (Adjoint matrix). Let $n, m \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times m}$. The matrix $B \in \mathbb{F}^{m\times n}$ given by

$$b_{ij} := \bar{a}_{ji} \quad \text{for all } i \in \{1, \ldots, m\},\ j \in \{1, \ldots, n\}$$

is called the *adjoint* of A and denoted by A^*.

Lemma 2.28 (Adjoint matrix). *Let $n, m \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times m}$. We have*

$$\langle Ax, y\rangle = \langle x, A^*y\rangle \quad \textit{for all } x \in \mathbb{F}^m,\ y \in \mathbb{F}^n. \quad (2.11)$$

Proof. Let $B = A^*$. We have

$$\begin{aligned}\langle Ax, y\rangle &= \sum_{i=1}^{n} (Ax)_i \bar{y}_i = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} x_j \bar{y}_i = \sum_{j=1}^{m} x_j \sum_{i=1}^{n} \bar{b}_{ji}\bar{y}_i \\ &= \sum_{j=1}^{m} x_j (\overline{By})_j = \langle x, By\rangle = \langle x, A^*y\rangle \quad \text{for all } x \in \mathbb{F}^m,\ y \in \mathbb{F}^n.\end{aligned}$$ □

Applying (2.11) to $x = \delta_i$ and $y = \delta_j$ for $i \in \{1, \ldots, m\}$ and $y \in \{1, \ldots, n\}$ shows that this equation actually defines A^*. It can be used to generalize the concept of an adjoint to operators in general Hilbert spaces.

Lemma 2.29 (Null space and range). *Let $n, m \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times m}$. We have*

$$\langle x, y\rangle = 0 \quad \textit{for all } x \in \mathcal{R}(A),\ y \in \mathcal{N}(A^*). \quad (2.12)$$

In particular, $A^*x = 0$ implies $x = 0$ for all $x \in \mathcal{R}(A)$.

Proof. Let $x \in \mathcal{R}(A)$ and $y \in \mathcal{N}(A^*)$. By definition, we can find $z \in \mathbb{F}^m$ such that $x = Az$ and obtain

$$\langle x, y\rangle = \langle Az, y\rangle = \langle z, A^*y\rangle = \langle z, 0\rangle = 0,$$

since $y \in \mathcal{N}(A^*)$.

If now $A^*x = 0$ holds, we have $x \in \mathcal{N}(A^*)$ and find

$$\|x\|^2 = \langle x, x\rangle = 0$$

by applying (2.12) to $y = x$. □

Two vectors $x, y \in \mathbb{F}^n$ are called *perpendicular*, denoted by $x \perp y$, if $\langle x, y\rangle = 0$ holds. A vector $x \in \mathbb{F}^n$ is called perpendicular on a subspace $\mathcal{W} \subseteq \mathbb{F}^n$, denoted $x \perp \mathcal{W}$, if x and y are perpendicular for all $y \in \mathcal{W}$. Two subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathbb{F}^n$ are called perpendicular if all pairs $(x, y) \in \mathcal{V} \times \mathcal{W}$ are perpendicular. Lemma 2.29 states that the range of A and the null space of A^* are perpendicular.

Proposition 2.30 (Spectrum of adjoint matrix). *Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. We have*

$$\sigma(A^*) = \{\bar{\lambda} \,:\, \lambda \in \sigma(A)\}.$$

Proof. Let $\lambda \in \sigma(A)$, and let $x \in \mathbb{F}^n \setminus \{0\}$ be a corresponding eigenvector. By definition, we have $(\lambda I - A)x = 0$ and therefore

$$0 = \langle (\lambda I - A)x, y\rangle = \langle x, (\lambda I - A)^* y\rangle = \langle x, (A^* - \bar{\lambda} I)y\rangle \quad \text{for all } y \in \mathbb{F}^n.$$

Since x is perpendicular on $\mathcal{R}(A^* - \bar{\lambda} I)$, the rank-nullity theorem (2.2) implies $\mathcal{N}(A^* - \bar{\lambda} I) \neq \{0\}$, and any non-zero element of $\mathcal{N}(A^* - \bar{\lambda} I)$ is an eigenvector of A^* for $\bar{\lambda}$, so we have proven $\bar{\lambda} \in \sigma(A^*)$.

Due to $A^{**} = A$ and $\bar{\bar{\lambda}} = \lambda$, the proof is complete. □

Definition 2.31 (Self-adjoint matrix). Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. If $A = A^*$ holds, A is called *self-adjoint*.

Lemma 2.32 (Identity of self-adjoint matrices). *Let $n \in \mathbb{N}$, and let $A \in \mathbb{F}^{n\times n}$ be self-adjoint. If*

$$\langle Ax, x\rangle = 0 \quad \textit{holds for all } x \in \mathbb{F}^n,$$

we have $A = 0$.

Proof. Let $x, y \in \mathbb{F}^n$. We have

$$\begin{aligned} 0 &= \langle A(x + y), x + y\rangle = \langle Ax, x\rangle + \langle Ax, y\rangle + \langle Ay, x\rangle + \langle Ay, y\rangle \\ &= \langle Ax, y\rangle + \langle Ay, x\rangle = \langle Ax, y\rangle + \langle y, A^*x\rangle = \langle Ax, y\rangle + \langle y, Ax\rangle. \end{aligned}$$

Using $y := Ax$, we find

$$0 = 2\langle Ax, Ax\rangle = 2\|Ax\|^2 \quad \text{for all } x \in \mathbb{F}^n,$$

and this implies $A = 0$. □

Definition 2.33 (Normal matrix). Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. If $AA^* = A^*A$ holds, A is called *normal*.

Exercise 2.34. Prove that the following matrices are normal:

$$A_1 := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad A_2 := \begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix}, \quad A_3 := \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}.$$

Lemma 2.35 (Metric equivalence). *Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. A is normal if and only if*

$$\|Ax\| = \|A^*x\| \quad \text{holds for all } x \in \mathbb{F}^n. \tag{2.13}$$

Proof. Assume that A is normal. For $x \in \mathbb{F}^n$, we find

$$\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle = \langle AA^*x, x\rangle = \langle A^*x, A^*x\rangle = \|A^*x\|^2,$$

and this implies (2.13).

Assume now (2.13). We find

$$\begin{aligned} 0 &= \|Ax\|^2 - \|A^*x\|^2 = \langle Ax, Ax\rangle - \langle A^*x, A^*x\rangle = \langle A^*Ax, x\rangle - \langle AA^*x, x\rangle \\ &= \langle (A^*A - AA^*)x, x\rangle \quad \text{for all } x \in \mathbb{F}^n. \end{aligned}$$

Since $A^*A - AA^*$ is obviously self-adjoint, we can apply Lemma 2.32 to obtain $A^*A = AA^*$ and can conclude that A is normal. □

Proposition 2.36 (Eigenvectors). *Let $n \in \mathbb{N}$, and let $A \in \mathbb{F}^{n\times n}$ be a normal matrix. Let $\lambda \in \sigma(A)$ be an eigenvalue, and let $e \in \mathbb{F}^n \setminus \{0\}$ be a corresponding eigenvector. Then we have*

$$A^*e = \bar{\lambda}e,$$

i.e., e is an eigenvector of the adjoint A^* for the eigenvalue $\bar{\lambda}$.

In particular, A is diagonalizable if and only if A^* is, and both matrices can be diagonalized by the same similarity transformation.

Proof. Since e is an eigenvector, we have $Ae = \lambda e$, and therefore

$$0 = \|Ae - \lambda e\| = \|(A - \lambda I)e\|. \tag{2.14}$$

Since A is normal, we have

$$\begin{aligned} (A - \lambda I)(A - \lambda I)^* &= (A - \lambda I)(A^* - \bar{\lambda} I) = AA^* - \lambda A^* - \bar{\lambda}A + |\lambda|^2 I \\ &= A^*A - \bar{\lambda}A - \lambda A^* + |\lambda|^2 I = (A^* - \bar{\lambda} I)(A - \lambda I) \\ &= (A - \lambda I)^*(A - \lambda I), \end{aligned} \tag{2.15}$$

so $\lambda I - A$ is also a normal matrix. Applying Lemma 2.35 to (2.14), we obtain

$$0 = \|(A - \lambda I)e\| = \|(A - \lambda I)^* e\| = \|(A^* - \bar{\lambda} I)e\| = \|A^* e - \bar{\lambda} e\|$$

i.e., $A^* e = \bar{\lambda} e$. □

Applying general similarity transformations can cause numerical problems if the transformation is ill-conditioned, e.g., if rounding errors lead to mixed eigenspaces. We can avoid these problems by using unitary transformations, since these leave lengths and angles of vectors unchanged and therefore lead to very stable algorithms.

Definition 2.37 (Isometric matrix). Let $n, m \in \mathbb{N}$, and let $Q \in \mathbb{F}^{n \times m}$. If $Q^* Q = I$ holds, Q is called *isometric*. A square isometric matrix is called *unitary*.

A real unitary matrix is usually also called *orthogonal*. We always use the term "unitary" in the interest of consistency.

Lemma 2.38 (Isometry). *Let $n, m \in \mathbb{N}$, and let $Q \in \mathbb{F}^{n \times m}$. Q is isometric if and only if*

$$\|Qx\| = \|x\| \quad \text{holds for all } x \in \mathbb{F}^n. \tag{2.16}$$

Proof. Assume that Q is isometric. For $x \in \mathbb{F}^n$, we find

$$\|Qx\|^2 = \langle Qx, Qx \rangle = \langle Q^* Qx, x \rangle = \langle x, x \rangle = \|x\|^2,$$

and this implies (2.16).

Assume now (2.16). We find

$$\begin{aligned} 0 = \|Qx\|^2 - \|x\|^2 &= \langle Qx, Qx \rangle - \langle x, x \rangle = \langle Q^* Qx, x \rangle - \langle x, x \rangle \\ &= \langle (Q^* Q - I)x, x \rangle \quad \text{for all } x \in \mathbb{F}^n. \end{aligned}$$

Since $Q^* Q - I$ is self-adjoint, we can apply Lemma 2.32 to obtain $Q^* Q = I$ and conclude that Q is isometric. □

Proposition 2.39 (Unitary inverse). *Let $n \in \mathbb{N}$ and $Q \in \mathbb{F}^{n \times m}$ be isometric. Then we have*

$$QQ^* x = x \quad \text{for all } x \in \mathcal{R}(Q). \tag{2.17}$$

If Q is square, i.e., unitary, Q^* is also unitary and we have $Q^* = Q^{-1}$.

Proof. Let $x \in \mathcal{R}(Q)$. By definition, we can find $y \in \mathbb{F}^m$ such that $x = Qy$ and obtain

$$QQ^*x = QQ^*Qy = Qy = x.$$

Let now Q be square. Due to (2.16), Q is injective. Since Q is square, the rank-nullity theorem (2.2) implies that Q is also surjective, i.e., $\mathcal{R}(Q) = \mathbb{F}^n$. (2.17) yields $QQ^* = I = Q^*Q$. □

We frequently use unitary similarity transformations, and Proposition 2.39 allows us to express them using the readily available adjoint instead of the inverse.

2.5 Invariant subspaces

In general, we cannot diagonalize a matrix by unitary similarity transformations, but we can at least transform it to *triangular* form. This is the *Schur decomposition*, a useful tool that allows us to draw conclusions regarding the unitary diagonalizability of matrices.

Definition 2.40 (Triangular matrix). Let $n \in \mathbb{N}$ and $L, R \in \mathbb{F}^{n\times n}$. The matrix R is called *upper triangular* if

$$r_{ij} = 0 \quad \text{holds for all } i, j \in \{1, \ldots, n\} \text{ with } j < i$$

and the matrix L is called *lower triangular* if

$$l_{ij} = 0 \quad \text{holds for all } i, j \in \{1, \ldots, n\} \text{ with } i < j.$$

Consider an upper triangular matrix $R \in \mathbb{F}^{n\times n}$. For $j \in \{1, \ldots, n\}$, multiplying R by the canonical unit vector δ_j yields

$$(R\delta_j)_i = r_{ij} = \begin{cases} r_{ij} & \text{if } i \leq j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i \in \{1, \ldots, n\},$$

i.e., we have

$$R\delta_j = \sum_{i=1}^{j} r_{ij}\delta_i \in \operatorname{span}\{\delta_1, \ldots, \delta_j\}.$$

Introducing the subspace

$$\mathcal{X} := \operatorname{span}\{\delta_1, \ldots, \delta_j\},$$

we can write this property in the form

$$Rx \in \mathcal{X} \quad \text{for all } x \in \mathcal{X}. \tag{2.18}$$

Eigenspaces share the same structure: if $A \in \mathbb{F}^{n\times n}$ is a matrix and $\lambda \in \sigma(A)$ is one of its eigenvalues, we have

$$Ax = \lambda x \in \mathcal{E}(A,\lambda) \quad \text{for all } x \in \mathcal{E}(A,\lambda).$$

In this sense, subspaces with the property (2.18) are a generalization of eigenspaces.

Definition 2.41 (Invariant subspace). Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. A subspace $\mathcal{X} \subseteq \mathbb{F}^n$ is called *invariant* with respect to A if

$$Ax \in \mathcal{X} \quad \text{holds for all } x \in \mathcal{X}. \tag{2.19}$$

Exercise 2.42 (Invariant subspaces). Find at least five different invariant subspaces for the matrix

$$A := \begin{pmatrix} 2 & & & & \\ 4 & 3 & & & \\ & & 7 & 8 & \\ & & & 6 & \end{pmatrix} \in \mathbb{C}^{4\times 4}.$$

In practical algorithms, we can only work with a basis of a subspace $\mathcal{X}$ instead of the subspace itself. Combining the basis vectors as in (2.7), we can represent the basis by a matrix X such that $\mathcal{X} = \mathcal{R}(X)$.

Proposition 2.43 (Invariant subspace). *Let $n, m \in \mathbb{N}$, let $A \in \mathbb{F}^{n\times n}$ and $X \in \mathbb{F}^{n\times m}$. The range $\mathcal{X} := \mathcal{R}(X)$ of X is an invariant subspace with respect to A if and only if*

$$AX = X\Lambda \tag{2.20}$$

holds for a matrix $\Lambda \in \mathbb{F}^{m\times m}$.

Proof. Assume that $\mathcal{X}$ is invariant with respect to A. Let $j \in \{1,\ldots,m\}$. Due to $X\delta_j \in \mathcal{R}(X) = \mathcal{X}$, we have

$$AX\delta_j \in \mathcal{X} = \mathcal{R}(X),$$

so we can find $y_j \in \mathbb{F}^m$ with $AX\delta_j = Xy_j$. We define

$$\Lambda_{ij} := (y_j)_i \quad \text{for all } i, j \in \{1,\ldots,m\}$$

and conclude $\Lambda\delta_j = y_j$ and therefore

$$AX\delta_j = Xy_j = X\Lambda\delta_j \quad \text{for all } j \in \{1,\ldots,m\}.$$

This implies (2.20).

Assume now that (2.20) holds. For $x \in \mathcal{X} = \mathcal{R}(X)$, we can find $y \in \mathbb{F}^m$ with $x = Xy$ and obtain

$$Ax = AXy = X\Lambda y \in \mathcal{R}(X) = \mathcal{X},$$

i.e., $\mathcal{X}$ is invariant with respect to A. □

The equation (2.20) bears a close resemblance to the equation (2.3) defining eigenvalues and eigenvectors: the eigenvector x is replaced by the matrix X, and the eigenvalue λ is replaced by the matrix Λ.

Exercise 2.44 (Invariant subspaces and eigenvalues). Show that any eigenvalue of the matrix Λ in (2.20) is an eigenvalue of A if X is injective.

If λ is an eigenvalue of A for an eigenvector $x \in \mathbb{F}^n$ with $x \in \mathcal{R}(X)$, is it also an eigenvalue of Λ?

2.6 Schur decomposition

We can use equation (2.20) to find a similarity transformation that moves A "closer" to an upper triangular matrix: if $X \in \mathbb{F}^{n\times m}$ is injective and satisfies (2.20) with $\Lambda \in \mathbb{F}^{m\times m}$, we can use the basis extension theorem (e.g., as in the proof of Proposition 2.19) to find an invertible matrix $T \in \mathbb{F}^{n\times n}$ with

$$T = \begin{pmatrix} X & Y \end{pmatrix}$$

for a matrix $Y \in \mathbb{F}^{n\times(n-m)}$. (2.20) implies

$$T^{-1}AX = T^{-1}X\Lambda = \begin{pmatrix} I_m \\ 0 \end{pmatrix}\Lambda = \begin{pmatrix} \Lambda \\ 0 \end{pmatrix}, \quad T^{-1}AT = \begin{pmatrix} \Lambda & B \\ & C \end{pmatrix} \tag{2.21}$$

for some $B \in \mathbb{F}^{m\times(n-m)}$ and $C \in \mathbb{F}^{(n-m)\times(n-m)}$. If we can find suitable invariant subspaces for Λ and C, we can repeat the procedure until we reach upper triangular matrices.

We can even ensure that the similarity transformations are unitary. To this end, we introduce the *sign function*

$$\operatorname{sgn} : \mathbb{F} \to \mathbb{F}, \quad z \mapsto \begin{cases} z/|z| & \text{if } z \neq 0, \\ 1 & \text{otherwise.} \end{cases} \tag{2.22}$$

and define a unitary transformation mapping an arbitrary vector to a multiple of the first canonical unit vector:

Lemma 2.45 (Householder reflection). *Let $n \in \mathbb{N}$ and $q \in \mathbb{F}^n \setminus \{0\}$. The matrix $P \in \mathbb{F}^{n\times n}$ defined by*

$$w := q + \operatorname{sgn}(q_1)\|q\|\delta_1, \quad P := I - 2\frac{ww^*}{\|w\|^2}$$

(where we identify $\mathbb{F}$ with $\mathbb{F}^{1\times 1}$ and $\mathbb{F}^n$ with $\mathbb{F}^{n\times 1}$ in the obvious way) is called an elementary Householder reflection [21]. *It is unitary and self-adjoint and satisfies*

$$Pq = -\operatorname{sgn}(q_1)\|q\|\delta_1. \tag{2.23}$$

Proof. P is obviously self-adjoint. Due to $w^*w = \|w\|^2$ and

$$P^*P = \left(I - 2\frac{ww^*}{\|w\|^2}\right)\left(I - 2\frac{ww^*}{\|w\|^2}\right) = I - 4\frac{ww^*}{\|w\|^2} + 4\frac{ww^*ww^*}{\|w\|^4} = I,$$

it is also unitary. With $\sigma := \operatorname{sgn}(q_1)$, we have

$$\begin{aligned}
\langle q, w\rangle &= \langle q, q\rangle + \langle q, \sigma\|q\|\delta_1\rangle = \|q\|^2 + \|q\|\,|q_1|,\\
\|w\|^2 &= \langle q + \sigma\|q\|\delta_1, q + \sigma\|q\|\delta_1\rangle\\
&= \langle q, q\rangle + \bar\sigma\|q\|\langle q, \delta_1\rangle + \sigma\|q\|\langle\delta_1, q\rangle + \|q\|^2\langle\delta_1, \delta_1\rangle\\
&= 2(\|q\|^2 + \|q\|\,|q_1|),\\
Pq &= q - 2w\frac{\langle q, w\rangle}{\|w\|^2} = q - 2(q + \sigma\|q\|\delta_1)\frac{\|q\|^2 + \|q\|\,|q_1|}{2(\|q\|^2 + \|q\|\,|q_1|)}\\
&= q - q - \sigma\|q\|\delta_1 = -\sigma\|q\|\delta_1,
\end{aligned}$$

and this completes the proof. □

By the fundamental theorem of algebra, any complex polynomial has at least one zero. Applying this result to the characteristic polynomial yields that any complex matrix has at least one eigenvalue, and we can use a Householder reflection to transform the matrix to the "almost upper triangular" form (2.21). Repeating the process allows us to reach an upper triangular matrix using only unitary similarity transformations.

Theorem 2.46 (Schur decomposition). *Let $n \in \mathbb{N}$ and $A \in \mathbb{C}^{n\times n}$. There are a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ satisfying*

$$Q^*AQ = R. \tag{2.24}$$

Proof. By induction on $n \in \mathbb{N}$. The case $n = 1$ is trivial.

Let now $n \in \mathbb{N}$ and assume that any $A \in \mathbb{C}^{n\times n}$ is unitarily similar to an upper triangular matrix. Let $A \in \mathbb{C}^{(n+1)\times(n+1)}$. Due to the fundamental theorem of algebra, we can find $\lambda \in \mathbb{C}$ with $p_A(\lambda) = 0$. Since therefore $\lambda I - A$ is not injective, we can find a vector $q \in \mathcal{N}(\lambda I - A) \setminus \{0\}$ with $\|q\| = 1$. Obviously, q is an eigenvector for

the eigenvalue λ. Let $P \in \mathbb{C}^{(n+1)\times(n+1)}$ be the elementary Householder reflection (cf. Lemma 2.45) satisfying

$$Pq = -\operatorname{sgn}(q_1)\delta_1.$$

Due to $P = P^*$ and $1/\operatorname{sgn}(q_1) = \operatorname{sgn}(\bar{q}_1)$, we also have $P\delta_1 = -\operatorname{sgn}(\bar{q}_1)q$ and obtain

$$P^*AP\delta_1 = -\operatorname{sgn}(\bar{q}_1)P^*Aq = -\operatorname{sgn}(\bar{q}_1)\lambda P^*q = -\operatorname{sgn}(\bar{q}_1)\lambda Pq = \lambda\delta_1.$$

This means that we can find $\widehat{A} \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{1\times n}$ satisfying

$$P^*AP = \begin{pmatrix} \lambda & B \\ & \widehat{A} \end{pmatrix}.$$

Since $\widehat{A}$ is only an $n \times n$ matrix, we can apply the induction assumption and find a unitary matrix $\widehat{Q} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\widehat{R} \in \mathbb{C}^{n\times n}$ with

$$\widehat{Q}^*\widehat{A}\widehat{Q} = \widehat{R}.$$

Now we let

$$Q := P\begin{pmatrix} 1 & \\ & \widehat{Q} \end{pmatrix}, \quad R := \begin{pmatrix} \lambda & B\widehat{Q} \\ & \widehat{R} \end{pmatrix}$$

and observe

$$\begin{aligned} Q^*AQ &= \begin{pmatrix} 1 & \\ & \widehat{Q}^* \end{pmatrix} P^*AP \begin{pmatrix} 1 & \\ & \widehat{Q} \end{pmatrix} = \begin{pmatrix} 1 & \\ & \widehat{Q}^* \end{pmatrix} \begin{pmatrix} \lambda & B \\ & \widehat{A} \end{pmatrix} \begin{pmatrix} 1 & \\ & \widehat{Q} \end{pmatrix} \\ &= \begin{pmatrix} \lambda & B\widehat{Q} \\ & \widehat{Q}^*\widehat{A}\widehat{Q} \end{pmatrix} = \begin{pmatrix} \lambda & B\widehat{Q} \\ & \widehat{R} \end{pmatrix} = R. \end{aligned}$$

Since R is upper triangular, this completes the induction. □

Theorem 2.46 relies on the fundamental theorem of algebra that holds in $\mathbb{C}$, but not in $\mathbb{R}$. Since most of the following results are corollaries, they also hold only in the field of complex numbers.

Remark 2.47 (Order of eigenvalues). In the proof of Theorem 2.46, we can put *any* eigenvalue $\lambda \in \sigma(A)$ into the upper left entry of R, and by extension we can choose any order for the eigenvalues on the diagonal of R.

Exercise 2.48 (Trace). The *trace* of a matrix $A \in \mathbb{F}^{n\times n}$ is defined by

$$\operatorname{tr}(A) := \sum_{i=1}^{n} a_{ii},$$

i.e., by the sum of diagonal entries. Let $Q \in \mathbb{F}^{n\times n}$ be unitary. Prove

$$\operatorname{tr}(Q^* A Q) = \operatorname{tr}(A)$$

and conclude

$$\operatorname{tr}(A) = \sum_{\lambda\in\sigma(A)} \lambda \mu_a(A,\lambda),$$

i.e., that the trace is invariant under similarity transformations.

Hint: Start by proving $\operatorname{tr}(Q^* A Q) = \sum_{i,j,k=1}^{n} \bar{q}_{ji} q_{ki} a_{jk}$ and considering the first two factors.

While the upper triangular form is already useful, e.g., for finding all eigenvalues and the corresponding algebraic multiplicities, we are mainly interested in diagonalizing a matrix. The following Lemma suggests a possible approach:

Lemma 2.49 (Normal triangular matrix). *Let $R \in \mathbb{F}^{n\times n}$ be upper triangular. If R is normal, it is a diagonal matrix.*

Proof. By induction on $n \in \mathbb{N}$. The case $n = 1$ is trivial.

Let now $n \in \mathbb{N}$ and assume that any normal upper triangular matrix $R \in \mathbb{F}^{n\times n}$ is diagonal. Let $R \in \mathbb{F}^{(n+1)\times(n+1)}$ be normal and upper triangular. Due to Lemma 2.35, we have

$$|r_{11}|^2 = \|R\delta_1\|^2 = \|R^*\delta_1\|^2 = |r_{11}|^2 + \cdots + |r_{1,n+1}|^2$$

and conclude $|r_{12}| = \cdots = |r_{1,n+1}| = 0$. This means that we can find an upper triangular matrix $\widehat{R} \in \mathbb{F}^{n\times n}$ with

$$R = \begin{pmatrix} r_{11} & \\ & \widehat{R} \end{pmatrix}.$$

Due to

$$\begin{pmatrix} |r_{11}|^2 & \\ & \widehat{R}\widehat{R}^* \end{pmatrix} = RR^* = R^*R = \begin{pmatrix} |r_{11}|^2 & \\ & \widehat{R}^*\widehat{R} \end{pmatrix},$$

the matrix $\widehat{R}$ is normal, and we can use the induction assumption to complete the proof. □

Corollary 2.50 (Normal matrix). *Let $n \in \mathbb{N}$, and let $A \in \mathbb{C}^{n\times n}$. A is normal if and only if it is unitarily diagonalizable, i.e., if there are a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and a diagonal matrix $D \in \mathbb{C}^{n\times n}$ with*

$$Q^*AQ = D. \tag{2.25}$$

Proof. Assume that A is normal. Due to Theorem 2.46, we can find a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ with

$$A = QRQ^*.$$

Since A is a normal matrix, we have

$$RR^* = Q^*AQQ^*A^*Q = Q^*AA^*Q = Q^*A^*AQ = Q^*A^*QQ^*AQ = R^*R,$$

so R is also normal, and Lemma 2.49 yields that R is, in fact, diagonal.

Assume now that we can find a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and a diagonal matrix $D \in \mathbb{C}^{n\times n}$ satisfying (2.25). Then we have

$$AA^* = QDQ^*QD^*Q^* = QDD^*Q^* = QD^*DQ^* = QD^*Q^*QDQ^* = A^*A$$

and conclude that A is normal. □

Corollary 2.51 (Self-adjoint matrix). *Let $n \in \mathbb{N}$, and let $A \in \mathbb{C}^{n\times n}$. A is self-adjoint if and only if there are a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and a* real *diagonal matrix $D \in \mathbb{R}^{n\times n}$ with*

$$Q^*AQ = D. \tag{2.26}$$

Proof. Assume that A is self-adjoint. Then it is also normal, and Corollary 2.50 gives us a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and a diagonal matrix $D \in \mathbb{C}^{n\times n}$ with $A = QDQ^*$. Since A is self-adjoint, we obtain

$$D = Q^*AQ = Q^*A^*Q = (Q^*AQ)^* = D^*,$$

so all diagonal entries of D have to be real.

Assume now that we can find a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and a real diagonal matrix $D \in \mathbb{C}^{n\times n}$ satisfying (2.26). Then we have

$$A = QDQ^* = QD^*Q^* = A^*$$

and conclude that A is self-adjoint. □

Our proofs so far depend on the fundamental theorem of algebra and therefore apply only to complex matrices. Even if A is self-adjoint and real, Corollary 2.51 still may yield a *complex* matrix Q. By slightly modifying the proof of Theorem 2.46, we can improve the result:

Corollary 2.52 (Real self-adjoint matrix). *Let $n \in \mathbb{N}$ and $A \in \mathbb{R}^{n\times n}$. A is self-adjoint if and only if there are a unitary matrix $Q \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ with*

$$Q^* A Q = D. \tag{2.27}$$

Proof. Assume that we can find a unitary matrix $Q \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ satisfying (2.27). Then we have

$$A = QDQ^* = QD^*Q^* = A^*,$$

so A has to be self-adjoint.

Assume now that A is self-adjoint. According to Corollary 2.51, we can find an eigenvalue $\lambda \in \mathbb{R}$ of A, and due to $\mathcal{N}(\lambda I - A) \neq \{0\}$ also a *real* eigenvector $q \in \mathbb{R}^n$ with $\|q\| = 1$. Using this vector, we can proceed by induction as in the proof of Theorem 2.46. □

According to Theorem 2.46, any complex-valued matrix can be transformed to triangular form by a unitary transformation. We now investigate whether it is possible to get close to diagonal form by using further unitary transformations.

Definition 2.53 (Frobenius norm). Let $n, m \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times m}$. The *Frobenius norm* of A is given by

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

Proposition 2.54 (Frobenius norm). *Let $n, m \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times m}$. We have*

- $\|A\|_F = \|A^*\|_F$,
- $\|A\|_F = \|QA\|_F$ *for all unitary matrices* $Q \in \mathbb{F}^{n\times n}$,
- $\|A\|_F = \|AQ\|_F$ *for all unitary matrices* $Q \in \mathbb{F}^{m\times m}$.

Proof. Let $B := A^*$. Definition 2.53 implies

$$\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}|^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} |b_{ji}|^2 = \|B\|_F^2 = \|A^*\|_F^2.$$

Let $Q \in \mathbb{F}^{n\times n}$ be a unitary matrix. Using the canonical unit vectors $(\delta_j)_{j=1}^m$ in $\mathbb{F}^m$ defined by (2.5), we obtain

$$\begin{aligned}\|A\|_F^2 &= \sum_{j=1}^{m}\sum_{i=1}^{n}|a_{ij}|^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}|(A\delta_j)_i|^2 = \sum_{j=1}^{m}\|A\delta_j\|^2\\ &= \sum_{j=1}^{m}\|QA\delta_j\|^2 = \|QA\|_F^2\end{aligned}$$

by Lemma 2.38. Applying this result to A^* completes the proof. □

We measure the "distance" to diagonal form by taking the Frobenius norm of all entries except for the diagonal.

Definition 2.55 (Off-diagonal part). Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. We introduce the quantity

$$\operatorname{off}(A) := \left(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|^2\right)^{1/2}$$

describing the Frobenius norm of the off-diagonal part of A.

We have $\operatorname{off}(A) = 0$ if and only if A is a diagonal matrix.

We compare two unitary transformations of a matrix A to triangular form.

Proposition 2.56 (Invariant). *Let $n \in \mathbb{N}$ and $A \in \mathbb{C}^{n\times n}$. Let $Q_1, Q_2 \in \mathbb{C}^{n\times n}$ be unitary matrices and let $R_1, R_2 \in \mathbb{C}^{n\times n}$ be upper triangular matrices satisfying*

$$Q_1 R_1 Q_1^* = A = Q_2 R_2 Q_2^*.$$

Then we have $\operatorname{off}(R_1) = \operatorname{off}(R_2)$.

Proof. (cf. [18, Section 7.1]) Since A, R_1 and R_2 are similar matrices, Proposition 2.18 yields

$$\prod_{i=1}^{n}(r_{1,ii}-\lambda) = p_{R_1}(\lambda) = p_A(\lambda) = p_{R_2}(\lambda) = \prod_{j=1}^{n}(r_{2,jj}-\lambda) \quad \text{for all } \lambda \in \mathbb{C},$$

where $r_{1,ii}$ and $r_{2,jj}$ denote the i-th and j-th diagonal elements of R_1 and R_2, respectively. Two factorizations of p_A into linear factors can only differ by the ordering of the factors, and we obtain

$$\sum_{i=1}^{n}|r_{1,ii}|^2 = \sum_{j=1}^{n}|r_{2,jj}|^2.$$

Proposition 2.54 yields

$$\begin{aligned}\operatorname{off}(R_1)^2 &= \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} |r_{1,ij}|^2 = \|R_1\|_F^2 - \sum_{i=1}^{n} |r_{1,ii}|^2 = \|Q_1^* A Q_1\|_F^2 - \sum_{i=1}^{n} |r_{1,ii}|^2 \\ &= \|A\|_F^2 - \sum_{i=1}^{n} |r_{1,ii}|^2 = \|Q_2^* A Q_2\|_F^2 - \sum_{j=1}^{n} |r_{2,jj}|^2 \\ &= \|R_2\|_F^2 - \sum_{j=1}^{n} |r_{2,jj}|^2 = \operatorname{off}(R_2)^2,\end{aligned}$$

and the proof is complete. □

Let $A \in \mathbb{C}^{n\times n}$, and let $Q, R \in \mathbb{C}^{n\times n}$ be as in Theorem 2.46. Corollary 2.50 implies that A is normal if and only if R is diagonal, and this is the case if and only if

$$\Delta(A) := \operatorname{off}(R)$$

equals zero. In this sense, $\Delta(A)$ measures the *departure from normality* of A, and due to Proposition 2.56, it does not depend on the particular Schur decomposition.

2.7 Non-unitary transformations *

Let $A \in \mathbb{C}^{n\times n}$. Theorem 2.46 implies that we can find a unitary transformation $Q \in \mathbb{C}^{n\times n}$ such that

$$R = Q^* A Q$$

is upper triangular, and Proposition 2.56 implies that we cannot get "closer" to diagonal form using unitary transformations.

Therefore we now consider non-unitary transformations. We assume that an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ is given and split it into

$$R = \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix} \tag{2.28}$$

with $R_{11} \in \mathbb{C}^{k\times k}$, $R_{12} \in \mathbb{C}^{k\times \ell}$ and $R_{22} \in \mathbb{C}^{\ell\times\ell}$ with $n = k + \ell$. We consider the similarity transformation

$$T := \begin{pmatrix} I & X \\ & I \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} I & -X \\ & I \end{pmatrix} \tag{2.29}$$

for a matrix $X \in \mathbb{C}^{k\times\ell}$ and obtain

$$T^{-1} R T = \begin{pmatrix} I & -X \\ & I \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix} \begin{pmatrix} I & X \\ & I \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} + R_{11}X - XR_{22} \\ & R_{22} \end{pmatrix}.$$

In order to eliminate the upper-right block, we have to find $X \in \mathbb{C}^{k\times\ell}$ solving *Sylvester's equation*

$$R_{11}X - XR_{22} = -R_{12}.$$

This is a linear matrix equation.

Proposition 2.57 (Sylvester's equation). *Let $n, m \in \mathbb{N}$, $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$. Sylvester's equation*

$$AX - XB = C \tag{2.30}$$

has unique solutions $X \in \mathbb{C}^{n\times m}$ for all $C \in \mathbb{C}^{n\times m}$ if and only if $\sigma(A) \cap \sigma(B) = \emptyset$, i.e., if A and B have no eigenvalue in common.

Proof. We consider the linear mapping

$$S : \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}, \quad X \mapsto AX - XB.$$

If it is injective, the rank-nullity theorem implies that it is also bijective, and therefore equation (2.30) is uniquely solvable for all $C \in \mathbb{C}^{n\times m}$.

We will prove that $\sigma(A) \cap \sigma(B) \neq \emptyset$ is equivalent to S being not injective.

Let first $\sigma(A)\cap\sigma(B) \neq \emptyset$. We pick $\lambda \in \sigma(A)\cap\sigma(B)$ and eigenvectors $x \in \mathbb{C}^n\setminus\{0\}$ and $y \in \mathbb{C}^m \setminus \{0\}$ satisfying

$$Ax = \lambda x, \quad B^*y = \bar{\lambda} y.$$

The latter is possible due to Proposition 2.30. Now we let $X := xy^*$ and find

$$\begin{aligned} S[X] = AX - XB = Axy^* - xy^*B = (Ax)y^* - x(B^*y)^* \\ = (\lambda x)y^* - x(\bar{\lambda} y)^* = \lambda xy^* - \lambda xy^* = 0, \end{aligned}$$

i.e., the nullspace of S contains the non-zero matrix X, therefore S cannot be injective.

Let now S be not injective. Then we can find $X \in \mathbb{C}^{n\times m} \setminus \{0\}$ with $S[X] = 0$, i.e.,

$$AX = XB. \tag{2.31}$$

We first consider a special case: assume that B is upper triangular. Let $j \in \{1, \dots, m\}$ be the smallest index satisfying $x := X\delta_j \neq 0$, i.e., the number of the first non-zero column of X. Then we have

$$Ax = AX\delta_j = XB\delta_j = X\sum_{i=1}^{j} b_{ij}\delta_i = \sum_{i=1}^{j} b_{ij} X\delta_i.$$

Since $X\delta_i = 0$ for all $i < j$, this yields

$$Ax = b_{jj} X\delta_j = b_{jj}x,$$

i.e., x is an eigenvector of A for the eigenvalue b_{jj}. Since B is upper triangular, b_{jj} is also an eigenvalue of B, i.e., $b_{jj} \in \sigma(A) \cap \sigma(B)$.

Now we return to the general case. According to Theorem 2.46, we can find a Schur decomposition of B, i.e., a unitary matrix $Q \in \mathbb{C}^{m\times m}$ and an upper triangular matrix $\widehat{B}$ such that

$$B = Q\widehat{B}Q^*$$

holds. (2.31) yields

$$AX = XB = XQ\widehat{B}Q^*, \quad AXQ = XQ\widehat{B},$$

due to $Q^*Q = I$, and introducing $\widehat{X} := XQ$, we find

$$A\widehat{X} = \widehat{X}\widehat{B}.$$

We can apply the first step of the proof to show $\sigma(A) \cap \sigma(\widehat{B}) \neq \emptyset$, and since B and $\widehat{B}$ are similar, Proposition 2.18 yields $\sigma(A) \cap \sigma(B) \neq \emptyset$. □

If we can ensure $\sigma(R_{11}) \cap \sigma(R_{22}) = \emptyset$ in (2.28), Proposition 2.57 implies that the transformation T introduced in (2.29) can eliminate R_{12}. Repeating this procedure allows us to transform any square complex matrix to block-diagonal form:

Theorem 2.58 (Block-diagonalization). *Let $A \in \mathbb{C}^{n\times n}$. Let $\sigma(A) = \{\lambda_1, \ldots, \lambda_k\}$ with $k := \#\sigma(A)$. For each $\ell \in \{1, \ldots, k\}$ there are an integer $n_\ell \in \mathbb{N}$ and an upper triangular matrix $R_\ell \in \mathbb{R}^{n_\ell \times n_\ell}$ such that*

$$B^{-1}AB = \begin{pmatrix} R_1 & & \\ & \ddots & \\ & & R_k \end{pmatrix} \tag{2.32}$$

holds for an invertible matrix $B \in \mathbb{C}^{n\times n}$. For each $\ell \in \{1, \ldots, k\}$, we have $\sigma(R_\ell) = \{\lambda_\ell\}$ and $n_\ell = \mu_a(A, \lambda_\ell)$.

Proof. (cf. [18, Theorem 7.1.6]) By induction on the cardinality $\#\sigma(A)$ of the spectrum. If $\#\sigma(A) = 1$, the Schur decomposition of Theorem 2.46 yields the desired factorization with $B := Q$ and $R_1 := R$.

Assume now we have a $k \in \mathbb{N}$ such that the desired factorization exists for all $A \in \mathbb{C}^{n\times n}$ satisfying $\#\sigma(A) = k$.

Let $A \in \mathbb{C}^{n\times n}$ be given with $\#\sigma(A) = k + 1$ and $\sigma(A) = \{\lambda_1, \ldots, \lambda_{k+1}\}$. Due to Theorem 2.46, we can find a unitary matrix $Q \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ satisfying

$$Q^*AQ = R.$$

We define

$$n_1 := \#\{i \in \{1, \ldots, n\} \,:\, r_{ii} = \lambda_1\}, \quad \hat{n} := n - n_1.$$

According to Remark 2.47, we can ensure

$$\begin{aligned} r_{ii} &= \lambda_1 \quad \text{for all } i \leq n_1, \\ r_{ii} &\neq \lambda_1 \quad \text{for all } i > n_1. \end{aligned}$$

We split R into $R_1 \in \mathbb{R}^{n_1 \times n_1}$, $\widehat{R} \in \mathbb{R}^{\hat{n} \times \hat{n}}$ and $Y \in \mathbb{R}^{n_1 \times \hat{n}}$ such that

$$R = \begin{pmatrix} R_1 & Y \\ & \widehat{R} \end{pmatrix}.$$

By construction, we have $\sigma(R_1) = \{\lambda_1\}$ and $\lambda_1 \notin \sigma(\widehat{R})$. Due to Proposition 2.57, we can find $X \in \mathbb{R}^{n_1 \times (n-n_1)}$ satisfying Sylvester's equation

$$R_1 X - X\widehat{R} = -Y.$$

We define $T \in \mathbb{R}^{n \times n}$ as in (2.29) and obtain

$$T^{-1} R T = \begin{pmatrix} R_1 & \\ & \widehat{R} \end{pmatrix}.$$

Due to $\lambda_1 \notin \sigma(\widehat{R}) \subseteq \sigma(A)$, we have $\#\sigma(\widehat{R}) = k$ and can apply the induction assumption to find an invertible matrix $\widehat{B} \in \mathbb{C}^{\hat{n} \times \hat{n}}$ and upper triangular matrices $R_2, \ldots, R_{k+1}$ such that

$$\widehat{B}^{-1} \widehat{R} \widehat{B} = \begin{pmatrix} R_2 & & \\ & \ddots & \\ & & R_{k+1} \end{pmatrix}.$$

We let

$$B := QT \begin{pmatrix} I & \\ & \widehat{B} \end{pmatrix}$$

and obtain

$$\begin{aligned} B^{-1} A B &= \begin{pmatrix} I & \\ & \widehat{B}^{-1} \end{pmatrix} T^{-1} Q^* A Q T \begin{pmatrix} I & \\ & \widehat{B} \end{pmatrix} = \begin{pmatrix} I & \\ & \widehat{B}^{-1} \end{pmatrix} \begin{pmatrix} R_1 & \\ & \widehat{R} \end{pmatrix} \begin{pmatrix} I & \\ & \widehat{B} \end{pmatrix} \\ &= \begin{pmatrix} R_1 & \\ & \widehat{B}^{-1} \widehat{R} \widehat{B} \end{pmatrix} = \begin{pmatrix} R_1 & & & \\ & R_2 & & \\ & & \ddots & \\ & & & R_{k+1} \end{pmatrix}. \end{aligned}$$

This completes the induction.

Applying Proposition 2.18 to the representation (2.32) and taking advantage of the fact that the determinant of a block diagonal matrix is the product of the determinants of the diagonal blocks yields

$$p_A(\lambda) = \prod_{\ell=1}^{k} p_{R_\ell}(\lambda) = \prod_{\ell=1}^{k} (\lambda - \lambda_\ell)^{n_\ell} \quad \text{for all } \lambda \in \mathbb{C},$$

and therefore $\mu_a(A, \lambda_\ell) = n_\ell$. □

Even if a matrix is not diagonalizable, this result allows us to transform it to block-diagonal structure with one block corresponding to each eigenvalue. If we split the transformation B into

$$B = \begin{pmatrix} B_1 & \dots & B_k \end{pmatrix}, \quad B_\ell \in \mathbb{C}^{n \times n_\ell} \quad \text{for all } \ell \in \{1, \dots, k\},$$

the representation (2.32) takes the form

$$AB_\ell = B_\ell R_\ell \quad \text{for all } \ell \in \{1, \dots, k\}, \tag{2.33}$$

i.e., the columns of each B_ℓ are a basis of an invariant subspace.

Let $\ell \in \{1, \dots, k\}$. Since λ_ℓ is the only eigenvalue of the upper triangular matrix R_ℓ, all diagonal elements are equal to λ_ℓ, and therefore all diagonal elements of $\lambda_\ell I - R_\ell$ are equal to zero, i.e., the matrix is *strictly upper triangular*. We let $b_j := B_\ell \delta_j$ for all $j \in \{1, \dots, n_\ell\}$. Since $\lambda_\ell I - R_\ell$ is strictly upper triangular, we have

$$\begin{aligned}(\lambda_\ell I - A) b_j &= (\lambda_\ell I - A) B_\ell \delta_j = B_\ell (\lambda_\ell I - R_\ell) \delta_j \\ &= B_\ell \sum_{i=1}^{j-1} -r_{\ell,i,j} \delta_i = \sum_{i=1}^{j-1} -r_{\ell,i,j} b_i \quad \text{for all } j \in \{1, \dots, n_\ell\}.\end{aligned}$$

In particular, this implies

$$\begin{aligned}&(\lambda_\ell I - A) b_1 = 0, \\ &(\lambda_\ell I - A) b_j \in \operatorname{span}\{b_1, \dots, b_{j-1}\} \quad \text{for all } j \in \{2, \dots, n_\ell\}.\end{aligned}$$

A simple induction yields

$$(\lambda_\ell I - A)^j b_j = 0 \quad \text{for all } j \in \{1, \dots, n_\ell\},$$

so b_1 is an eigenvector, while b_j for $j \geq 2$ can be considered as a generalized eigenvector.

Definition 2.59 (Generalized eigenvector). Let $A \in \mathbb{C}^{n \times n}$, $\lambda \in \sigma(A)$ and $x \in \mathbb{C}^n \setminus \{0\}$. If there is a number $k \in \mathbb{N}$ with

$$(\lambda I - A)^k x = 0, \tag{2.34}$$

the vector x is called a *generalized eigenvector* of A for the eigenvalue λ. The smallest integer $k \in \mathbb{N}$ satisfying (2.34) is called the *order* of the generalized eigenvector.

By definition, if x is a generalized eigenvector of order $k > 1$ for a matrix A and an eigenvalue λ, $y := (\lambda I - A)x$ has to be a generalized eigenvector of order $k - 1$, since

$$(\lambda I - A)^{k-1} y = (\lambda I - A)^k x = 0$$

and $y \neq 0$ due to the minimality of k.

Theorem 2.58 states that for any complex matrix $A \in \mathbb{C}^{n \times n}$, we can find a basis of $\mathbb{C}^n$ consisting of generalized eigenvectors. This basis can be used to derive the even more specialized *Jordan normal form* of the matrix A. For our purposes, however, the block-diagonal decomposition of Theorem 2.58 is sufficient.

Chapter 3

Jacobi iteration

Summary

In this chapter, we consider the simple, but reliable, Jacobi iteration. Given a self-adjoint matrix $A \in \mathbb{F}^{n\times n}$, it performs a sequence of unitary similarity transformations that eliminate off-diagonal entries and thus moves the matrix closer to diagonal form. The Jacobi iteration may not be the fastest method, but it is guaranteed to converge for *any* self-adjoint matrix. The optional Section 3.5 proves that the Jacobi iteration will converge *quadratically* if a matrix is already close to diagonal form.

Learning targets

- ✓ Compute the two-dimensional Schur decomposition in a numerically stable way.
- ✓ Construct the elementary Jacobi step that eliminates a sub-diagonal entry of a matrix (and, due to symmetry, a corresponding super-diagonal entry as well).
- ✓ Use these elementary steps to construct a globally convergent iteration.
- ✓ Derive a posteriori bounds for the iteration error that allow us to judge the accuracy of each iteration step.

3.1 Iterated similarity transformations

The basic idea of the Jacobi iteration, and a number of other important iterative techniques for finding Schur decompositions, is to construct a sequence

$$A_0 = A, \quad A_{m+1} = Q_{m+1}^* A_m Q_{m+1} \quad \text{for all } m \in \mathbb{N}_0$$

of matrices $(A_m)_{m=0}^{\infty}$ by applying unitary similarity transformations $(Q_m)_{m=1}^{\infty}$. If these transformations are chosen appropriately, the sequence $A_0, A_1, A_2, \ldots$ will converge to a diagonal matrix.

Accumulating the transformations in a sequence

$$\widehat{Q}_0 = I, \quad \widehat{Q}_{m+1} = \widehat{Q}_m Q_{m+1} \quad \text{for all } m \in \mathbb{N}_0$$

yields

$$A_m = \widehat{Q}_m^* A \widehat{Q}_m \quad \text{for all } m \in \mathbb{N}_0,$$

and if A_m approximates a diagonal matrix D, we obtain

$$\widehat{Q}_m D \widehat{Q}_m^* \approx \widehat{Q}_m A_m \widehat{Q}_m^* = A,$$

i.e., an approximation of the Schur decomposition.

This approach has the significant advantage that the convergence behaviour of the sequence $A_0, A_1, A_2, \ldots$ can be controlled explicitly. Sophisticated iterations could, e.g., take advantage of submatrices that have already converged.

3.2 Two-dimensional Schur decomposition

Like many other numerical algorithms, the Jacobi iteration [22] replaces a complicated problem, in this case finding the Schur decomposition, by a sequence of simple problems. In the case of the Jacobi iteration, the simple problems consist of finding the Schur decomposition of 2×2 matrices. This decomposition can then be used to define a transformation of the original matrix.

In this section, we focus on the self-adjoint matrix

$$A = \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix} \in \mathbb{F}^{2\times 2}, \quad a, d \in \mathbb{R},\ b \in \mathbb{F}.$$

Due to Corollary 2.51, we know that there is a unitary transformation $Q \in \mathbb{F}^{2\times 2}$ diagonalizing the matrix A.

Following [18, Section 8.4.2], we consider unitary matrices of the form

$$Q = \begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix}, \quad c, s \in \mathbb{F}. \tag{3.1}$$

A matrix of this form is unitary if and only if

$$I = Q^*Q = \begin{pmatrix} \bar{c} & -\bar{s} \\ s & c \end{pmatrix} \begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix} = \begin{pmatrix} |c|^2 + |s|^2 & \bar{c}\bar{s} - \bar{s}\bar{c} \\ sc - cs & |s|^2 + |c|^2 \end{pmatrix}$$

holds, i.e., we have to use $c, s \in \mathbb{F}$ with $|c|^2 + |s|^2 = 1$. In order to diagonalize A, we have to find Q such that

$$\begin{aligned} B := Q^*AQ &= \begin{pmatrix} \bar{c} & -\bar{s} \\ s & c \end{pmatrix} \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix} \begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix} \\ &= \begin{pmatrix} \bar{c} & -\bar{s} \\ s & c \end{pmatrix} \begin{pmatrix} ac - bs & a\bar{s} + b\bar{c} \\ \bar{b}c - ds & \bar{b}\bar{s} + d\bar{c} \end{pmatrix} \\ &= \begin{pmatrix} \bar{c}(ac - bs) - \bar{s}(\bar{b}c - ds) & \bar{c}(a\bar{s} + b\bar{c}) - \bar{s}(\bar{b}\bar{s} + d\bar{c}) \\ s(ac - bs) + c(\bar{b}c - ds) & s(a\bar{s} + b\bar{c}) + c(\bar{b}\bar{s} + d\bar{c}) \end{pmatrix} \\ &= \begin{pmatrix} a|c|^2 + d|s|^2 - (b\bar{c}s + \bar{b}c\bar{s}) & (a-d)\bar{c}\bar{s} + b\bar{c}^2 - \bar{b}\bar{s}^2 \\ (a-d)cs - bs^2 + \bar{b}c^2 & a|s|^2 + d|c|^2 + (b\bar{c}s + \bar{b}c\bar{s}) \end{pmatrix} \end{aligned}$$

is a diagonal matrix, i.e., that

$$0 = b_{21} = (a-d)cs - bs^2 + \bar{b}c^2 \tag{3.2}$$

holds. If we have $b = 0$, we can let $c = 1$ and $s = 0$ and are done. Otherwise, we notice that, at least for $c, s \in \mathbb{R}$, the matrix Q defined in (3.1) can be considered as a clockwise rotation with the angle α given by $c = \cos\alpha$ and $s = \sin\alpha$. The tangent is given by

$$t := s/c = \frac{\sin\alpha}{\cos\alpha} = \tan\alpha$$

and can be used to simplify (3.2): with $s = tc$, we obtain

$$0 = (a-d)tc^2 - bt^2c^2 + \bar{b}c^2 = ((a-d)t - bt^2 + \bar{b})c^2$$

and can eliminate c^2 to get the quadratic equation

$$0 = bt^2 - (a-d)t - \bar{b}. \tag{3.3}$$

Multiplying by b, we obtain

$$0 = (bt)^2 - 2\frac{a-d}{2}(bt) - |b|^2 = \left(bt - \frac{a-d}{2}\right)^2 - \frac{(a-d)^2}{4} - |b|^2.$$

Taking the square root yields

$$bt = \frac{a-d}{2} \pm \sqrt{\frac{(a-d)^2}{4} + |b|^2}, \quad t = \frac{a-d}{2b} \pm \frac{|b|}{b}\sqrt{\left|\frac{a-d}{2b}\right|^2 + 1}.$$

We introduce

$$\tau := \frac{a-d}{2b},$$

recall (cf. 2.22) that the sign of a complex number is denoted by

$$\operatorname{sgn}(z) := \begin{cases} z/|z| & \text{if } z \neq 0, \\ 1 & \text{otherwise} \end{cases} \qquad \text{for all } z \in \mathbb{F},$$

and arrive at

$$t = \tau \pm \frac{1}{\operatorname{sgn}(b)}\sqrt{|\tau|^2 + 1}. \tag{3.4}$$

Using t, we can reconstruct c and s by

$$1 = |c|^2 + |s|^2 = |c|^2 + |tc|^2 = (1 + |t|^2)|c|^2, \quad c = \frac{1}{\sqrt{1+|t|^2}}, \quad s = tc. \tag{3.5}$$

We still have to choose the sign of the square root in (3.4). In order to obtain proper convergence of the iterative method, we have to ensure that the next iterate $B = Q^*AQ$ is close to A. For the rotation Q given in (3.1), this means that the sine s should be as small as possible, i.e., $|t|$ should be as small as possible: we have to choose the zero that is smaller in modulus. It is given by

$$t = \tau - \frac{\operatorname{sgn}(a-d)}{\operatorname{sgn}(b)}\sqrt{|\tau|^2+1} = \tau - \operatorname{sgn}(\tau)\sqrt{|\tau|^2+1},$$

but computing it this way can lead to problems when using floating-point arithmetic. Computing the zero

$$\hat{t} := \tau + \operatorname{sgn}(\tau)\sqrt{|\tau|^2+1}$$

that is larger in modulus, on the other hand, is stable. By construction, t and $\hat{t}$ are the zeros of the monic polynomial

$$p(z) := z^2 - 2\tau z - \bar{b}/b \quad \text{for all } z \in \mathbb{F},$$

i.e., we have

$$z^2 - 2\tau z - \bar{b}/b = p(z) = (z-t)(z-\hat{t}) = z^2 - (t+\hat{t})z + t\hat{t} \quad \text{for all } z \in \mathbb{F},$$

and comparing terms yields

$$t\hat{t} = \bar{b}/b, \quad t = \frac{\bar{b}/b}{\hat{t}}.$$

Using $\operatorname{sgn}(\bar{\tau}) = \operatorname{sgn}(a-d)/\operatorname{sgn}(\bar{b}) = \operatorname{sgn}(\tau)\operatorname{sgn}(b)/\operatorname{sgn}(\bar{b})$ allows us to find a more convenient expression:

$$\begin{aligned} t &= \frac{\operatorname{sgn}(\bar{b})/\operatorname{sgn}(b)}{\hat{t}} = \frac{1}{\operatorname{sgn}(b)(\tau + \operatorname{sgn}(\tau)\sqrt{|\tau|^2+1})/\operatorname{sgn}(\bar{b})} \\ &= \frac{1}{\bar{\tau} + \operatorname{sgn}(\bar{\tau})\sqrt{|\tau|^2+1}} = \frac{\tau}{\tau\bar{\tau} + \tau\operatorname{sgn}(\bar{\tau})\sqrt{|\tau|^2+1}} = \frac{\tau}{|\tau|^2 + |\tau|\sqrt{|\tau|^2+1}} \\ &= \frac{\operatorname{sgn}(\tau)}{|\tau| + \sqrt{|\tau|^2+1}}. \end{aligned} \tag{3.6}$$

The last term offers a numerically stable way of computing t.

Exercise 3.1 (Singular value decomposition). Let $A \in \mathbb{F}^{2\times 2}$ be an *arbitrary* matrix. Find unitary matrices $U, V \in \mathbb{F}^{2\times 2}$ such that $\Sigma := U^*AV$ is diagonal with real and non-negative diagonal entries.

If the diagonal entries of Σ are in descending order by magnitude, we call $A = U\Sigma V^*$ the singular value decomposition of A.

3.3 One step of the iteration

Let now $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix. In order to apply the result of the previous section, we choose $p, q \in \{1, \dots, n\}$ with $p < q$ and consider the 2×2 submatrix

$$\widehat{A} := \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix}.$$

We have seen that we can find $c, s \in \mathbb{F}$ such that

$$\widehat{Q} := \begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix}$$

is unitary and $\widehat{Q}^*\widehat{A}\widehat{Q}$ is diagonal. Applying the transformation $\widehat{Q}$ to the p-th and q-th row of an n-dimensional vector is a unitary transformation in $\mathbb{F}^{n\times n}$ corresponding to the matrix

$$Q := \begin{pmatrix} I_{p-1} & & & & \\ & c & & \bar{s} & \\ & & I_{q-p-1} & & \\ & -s & & \bar{c} & \\ & & & & I_{n-q} \end{pmatrix}, \tag{3.7}$$

where $I_k \in \mathbb{F}^{k\times k}$ again denotes the k-dimensional identity matrix. We compute the transformation $B := Q^*AQ$ in two steps: multiplication of A by Q yields the intermediate result

$$M := AQ = \begin{pmatrix} \dots & a_{1p}c - a_{1q}s & \dots & a_{1p}\bar{s} + a_{1q}\bar{c} & \dots \\ & \vdots & & \vdots & \\ \dots & a_{np}c - a_{nq}s & \dots & a_{np}\bar{s} + a_{nq}\bar{c} & \dots \end{pmatrix}, \tag{3.8}$$

where only the p-th and q-th column differ from the ones of the original matrix A. In the second step, we compute $B = Q^*M$. Multiplication by Q^* changes only the p-th and q-th row:

$$B = Q^*M = \begin{pmatrix} & \vdots & & \vdots & \\ \bar{c}m_{p1} - \bar{s}m_{q1} & & \dots & & \bar{c}m_{pn} - \bar{s}m_{qn} \\ & \vdots & & \vdots & \\ sm_{p1} + cm_{q1} & & \dots & & sm_{pn} + cm_{qn} \\ & \vdots & & \vdots & \end{pmatrix}. \tag{3.9}$$

We can easily verify

$$\widehat{M} := \begin{pmatrix} m_{pp} & m_{pq} \\ m_{qp} & m_{qq} \end{pmatrix} = \widehat{A}\widehat{Q}, \quad \widehat{B} := \begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} = \widehat{Q}^*\widehat{M} = \widehat{Q}^*\widehat{A}\widehat{Q}, \tag{3.10}$$

procedure jacobi_step(p, q, **var** A);
begin
  if $a_{pq} = 0$ **then**
    $t = 0$
  else begin
    $\tau \leftarrow (a_{pp} - a_{qq})/(2a_{pq})$;
    $t \leftarrow \operatorname{sgn}(\tau)/(|\tau| + \sqrt{|\tau|^2 + 1})$;
  end;
  $c \leftarrow 1/\sqrt{|t|^2 + 1}$; $s \leftarrow tc$;
  for $i \in \{1, \dots, n\}$ **do begin**
    $h \leftarrow a_{ip}$; $a_{ip} \leftarrow hc - a_{iq}s$; $a_{iq} \leftarrow h\bar{s} + a_{iq}\bar{c}$
  end;
  for $j \in \{1, \dots, n\}$ **do begin**
    $h \leftarrow a_{pj}$; $a_{pj} \leftarrow \bar{c}h - \bar{s}a_{qj}$; $a_{qj} \leftarrow sh + ca_{qj}$
  end
end

Figure 3.1. One step of the Jacobi iteration eliminates the entries a_{pq} and a_{qp}.

and since $\widehat{Q}^* \widehat{A} \widehat{Q}$ is diagonal by construction, we have $b_{qp} = b_{pq} = 0$. The resulting algorithm is summarized in Figure 3.1, A is overwritten with $B = Q^* A Q$.

Of course, it is not clear that eliminating one entry of A will get us closer to a diagonal matrix: it could just move entries around without actually reducing the off-diagonal part. Fortunately, the elimination step works perfectly:

Theorem 3.2 (Jacobi step). *Let $n \in \mathbb{N}$, and let $A \in \mathbb{F}^{n \times n}$ be self-adjoint. Let $p, q \in \{1, \dots, n\}$ with $p < q$. For the Jacobi matrix $Q \in \mathbb{F}^{n \times n}$ defined by* (3.7)*, we have*

$$\operatorname{off}(B)^2 = \operatorname{off}(A)^2 - 2|a_{pq}|^2, \quad B = Q^* A Q.$$

Proof. (cf. [18, eq. (8.4.2)]) The transformation Q can change only the p-th and q-th row and column. In particular, this implies

$$a_{ii} = b_{ii} \quad \text{for all } i \in \{1, \dots, n\},\ i \neq p,\ i \neq q.$$

Due to $\widehat{B} = \widehat{Q}^* \widehat{A} \widehat{Q}$ and Proposition 2.54, we have

$$|a_{pp}|^2 + 2|a_{pq}|^2 + |a_{qq}|^2 = \|\widehat{A}\|_F^2 = \|\widehat{B}\|_F^2 = |b_{pp}|^2 + |b_{qq}|^2,$$

since $b_{qp} = b_{pq} = 0$ and $a_{qp} = \bar{a}_{pq}$. Using Proposition 2.54 again, we obtain

$$\begin{aligned}\operatorname{off}(B)^2 &= \|B\|_F^2 - \sum_{i=1}^{n} |b_{ii}|^2 = \|B\|_F^2 - |b_{pp}|^2 - |b_{qq}|^2 - \sum_{\substack{i=1 \\ i \notin \{p,q\}}}^{n} |b_{ii}|^2 \\ &= \|A\|_F^2 - |a_{pp}|^2 - 2|a_{pq}|^2 - |a_{qq}|^2 - \sum_{\substack{i=1 \\ i \notin \{p,q\}}}^{n} |a_{ii}|^2 \\ &= \|A\|_F^2 - \sum_{i=1}^{n} |a_{ii}|^2 - 2|a_{pq}|^2 = \operatorname{off}(A)^2 - 2|a_{pq}|^2. \end{aligned}$$

□

Exercise 3.3 (Singular value decomposition). Let $A \in \mathbb{F}^{n\times n}$, and let $p, q \in \{1, \dots, n\}$ be given with $p < q$. Find unitary matrices $U, V \in \mathbb{F}^{n\times n}$ of similar form as (3.7) such that

$$\operatorname{off}(U^* A V)^2 = \operatorname{off}(A)^2 - |a_{pq}|^2 - |a_{qp}|^2.$$

As long as we choose non-zero entries a_{pq}, each Jacobi step reduces the norm of the off-diagonal part of the matrix. If we choose the maximal entries, we can obtain a simple linear convergence estimate:

Corollary 3.4 (Convergence). *Let $n \in \mathbb{N}$ and let $A \in \mathbb{F}^{n\times n}$ be self-adjoint. Let $p, q \in \{1, \dots, n\}$ with $p < q$ and*

$$|a_{ij}| \le |a_{pq}| \quad \text{for all } i, j \in \{1, \dots, n\}.$$

For the Jacobi matrix $Q \in \mathbb{F}^{n\times n}$ defined by (3.7)*, we have*

$$\operatorname{off}(B)^2 \le \left(1 - \frac{2}{n(n-1)}\right) \operatorname{off}(A)^2, \quad B = Q^* A Q.$$

Proof. (cf. [18, Section 8.4.3]) We have

$$\operatorname{off}(A)^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} |a_{ij}|^2 \le \sum_{\substack{i,j=1 \\ i \neq j}}^{n} |a_{pq}|^2 = |a_{pq}|^2 n(n-1),$$

so Theorem 3.2 yields

$$\begin{aligned}\operatorname{off}(B)^2 &\le \operatorname{off}(A)^2 - 2|a_{pq}|^2 \le \operatorname{off}(A)^2 - \frac{2}{n(n-1)} \operatorname{off}(A)^2 \\ &= \left(1 - \frac{2}{n(n-1)}\right) \operatorname{off}(A)^2. \end{aligned}$$

□

```
procedure jacobi_classical(ϵ, var A);
begin
  Choose 1 ≤ p < q ≤ n with |a_pq| = max{|a_ij| : i, j ∈ {1, ..., n}};
  while n(n − 1)|a_pq|^2 > ϵ^2 do begin
    jacobi_step(p, q, A);
    Choose 1 ≤ p < q ≤ n with |a_pq| = max{|a_ij| : i, j ∈ {1, ..., n}}
  end
end
```

Figure 3.2. Classical Jacobi iteration.

These results give rise to the *classical Jacobi iteration*: in each step, we choose the off-diagonal entry of maximal modulus and eliminate it, and we repeat until the off-diagonal entries are sufficiently small. Using $\operatorname{off}(A)^2 \leq n(n-1)|a_{pq}|^2$, we can compute an upper bound for $\operatorname{off}(A)$ and use it as a stopping criterion. The resulting algorithm is summarized in Figure 3.2.

An important advantage of the classical Jacobi iteration, as compared to other eigenvalue algorithms, is its guaranteed rate of convergence. Unfortunately, this advantage comes at a high cost: finding the maximal $|a_{pq}|$ directly requires checking all $n(n-1)/2$ superdiagonal entries, and even for moderately large values of n, this task dominates the computational cost. It is possible to reduce the complexity by using more sophisticated algorithms, but in practice it is more convenient to simply replace the search for the maximum by a strategy that cycles through all off-diagonal entries. An example is the *cyclic-by-row Jacobi iteration* given in Figure 3.3 that cycles through all rows and within each row through all columns. For a 5×5 matrix, the elements are eliminated in the following pattern:

$$\begin{pmatrix} \times & 1 & 2 & 3 & 4 \\ 1' & \times & 5 & 6 & 7 \\ 2' & 5' & \times & 8 & 9 \\ 3' & 6' & 8' & \times & 10 \\ 4' & 7' & 9' & 10' & \times \end{pmatrix}.$$

The entry $m \in \{1, \ldots, 10\}$ is eliminated in the m-th elementary step together with the entry m'. Diagonal elements marked with $\times$ are not eliminated.

The elementary Jacobi step given in Figure 3.1 requires 13 arithmetic operations to compute c and s and $12n$ operations to update A. The complexity of the overall Jacobi algorithm, both in the classical version given in Figure 3.2 and the cyclic version given in Figure 3.3 is usually measured in terms of one "sweep", i.e., a sequence of $n(n-1)/2$ elementary Jacobi steps, since exactly this number is required in the cyclic algorithm to eliminate all off-diagonal elements once. One sweep of the classical

```
procedure jacobi_cyclic(ϵ, var A);
begin
  γ ← 0;
  for i, j ∈ {1, ..., n}, i < j do γ ← γ + |a_ij|^2;
  while 2γ > ϵ^2 do begin
    for p = 1 to n − 1 do
      for q = p + 1 to n do
        jacobi_step(p, q, A);
    γ ← 0;
    for i, j ∈ {1, ..., n}, i < j do γ ← γ + |a_ij|^2
  end
end
```

Figure 3.3. Cyclic-by-row Jacobi iteration.

Jacobi algorithm requires

$$\frac{n(n-1)}{2}(12n+13) \approx 6n^3 + 13n^2/2$$

operations for the elementary steps and

$$\frac{n(n-1)}{2}\frac{n(n-1)}{2} \approx n^4/4$$

operations for finding the maximal elements in each step. One sweep of the cyclic Jacobi algorithm, including the computation of γ required for the stopping criterion, requires

$$\frac{n(n-1)}{2}(12n+13) + \frac{n(n-1)}{2}2 \approx 6n^3 + 15n^2/2$$

operations, i.e., it can be significantly more efficient than the classical algorithm.

3.4 Error estimates

We have seen that the matrices constructed by the Jacobi iteration converge to diagonal form, so we can expect the diagonal entries of these matrices to converge to eigenvalues. In this section, we provide a simple bound for the speed of convergence.

The Frobenius norm introduced in Definition 2.53 is not particularly well-suited for our purposes, so we introduce a second norm in the space of matrices:

Definition 3.5 (Spectral norm). Let $A \in \mathbb{F}^{n\times m}$.

$$\|A\| := \max\{\|Ax\| \,:\, x \in \mathbb{F}^m,\ \|x\| = 1\}$$

is called the *spectral norm* of A.

Although we use the same notation for the norm of a vector and the spectral norm of a matrix, the meaning is always clear, since the first only applies to vectors and the second only to matrices.

The spectral norm is *compatible* with the norm $\|\cdot\|$ in $\mathbb{F}^m$: for $x \in \mathbb{F}^m \setminus \{0\}$, the vector $x/\|x\|$ has unit norm, and we find

$$\|Ax\| = \left\| A \frac{x}{\|x\|} \right\| \|x\| \le \|A\| \, \|x\|.$$

This estimate obviously also holds for $x = 0$, so we have proven

$$\|Ax\| \le \|A\| \, \|x\| \quad \text{for all } x \in \mathbb{F}^m. \tag{3.11}$$

Exercise 3.6 (Spectral norm). Let $A \in \mathbb{F}^{n\times m}$. Prove

$$\|A\| = \max \left\{ \frac{|\langle Ax, y\rangle|}{\|x\| \, \|y\|} \ : \ x \in \mathbb{F}^m \setminus \{0\}, \ y \in \mathbb{F}^n \setminus \{0\} \right\}.$$

Show that this implies $\|A^*\| = \|A\|$ and $\|A^*A\| = \|A\|^2 = \|AA^*\|$.

The Frobenius norm and the spectral norm are related:

Proposition 3.7 (Norm estimates). *Let $A \in \mathbb{F}^{n\times m}$. We have*

$$\|A\| \le \|A\|_F \le \sqrt{m}\|A\|.$$

Proof. As in the proof of Proposition 2.54, we write the Frobenius norm as a sum of norms of the columns of A and apply (3.11):

$$\|A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2 = \sum_{j=1}^{m} \|A\delta_j\|^2 \le \sum_{j=1}^{m} \|A\|^2 \|\delta_j\|^2 = m\|A\|^2.$$

For the first estimate, we let $x \in \mathbb{F}^m$ with $\|x\| = 1$ and use the Cauchy–Schwarz inequality to obtain

$$\begin{aligned} \|Ax\|^2 &= \sum_{i=1}^{n} |(Ax)_i|^2 = \sum_{i=1}^{n} \Big| \sum_{j=1}^{m} a_{ij} x_j \Big|^2 \\ &\le \sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2 \sum_{k=1}^{m} |x_k|^2 = \|A\|_F^2 \|x\|^2 = \|A\|_F^2. \end{aligned}$$

Definition 3.5 yields $\|A\| \le \|A\|_F$. □

If the off-diagonal part of A has been sufficiently reduced, A is "almost" a diagonal matrix. Since it is very easy to find the eigenvalues and eigenvectors of a diagonal matrix, we would like to know how "close" they are to the eigenvalues and eigenvectors of A. For our proof of the corresponding perturbation estimate, we rely on the following simple version of the *Courant–Fischer–Weyl theorem* (cf. [6, §4.1]).

Lemma 3.8 (Minimization problem). *Let $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix, and let $\lambda \in \sigma(A)$ be its minimal eigenvalue. We have*

$$\lambda = \min\{\langle Ax, x\rangle \ : \ x \in \mathbb{F}^n, \ \|x\| = 1\}.$$

Proof. Due to Corollary 2.51, we find a unitary matrix $Q \in \mathbb{F}^{n\times n}$ and a real diagonal matrix $D \in \mathbb{R}^{n\times n}$ with

$$A = QDQ^*.$$

Since λ is the minimal eigenvalue, Proposition 2.18 yields

$$\lambda = \min\{d_{ii} \ : \ i \in \{1, \ldots, n\}\},$$

and we can find $j \in \{1, \ldots, n\}$ with $\lambda = d_{jj}$.

For all $x \in \mathbb{F}^n$ with $\|x\| = 1$, we have

$$\langle Ax, x\rangle = \langle QDQ^*x, x\rangle = \langle DQ^*x, Q^*x\rangle = \langle D\widehat{x}, \widehat{x}\rangle = \sum_{i=1}^{n} d_{ii}|\hat{x}_i|^2$$

for $\widehat{x} := Q^*x$. Lemma 2.38 yields $\|\widehat{x}\| = \|x\|$, and we obtain

$$\min\{\langle Ax, x\rangle \ : \ x \in \mathbb{F}^n, \ \|x\| = 1\} = \min\left\{\sum_{i=1}^{n} d_{ii}|\hat{x}_i|^2 \ : \ \widehat{x} \in \mathbb{F}^n, \|\widehat{x}\| = 1\right\}.$$

Due to

$$\lambda = \lambda\|\widehat{x}\|^2 = \sum_{i=1}^{n} d_{jj}|\hat{x}_i|^2 \leq \sum_{i=1}^{n} d_{ii}|\hat{x}_i|^2 \quad \text{for all } \widehat{x} \in \mathbb{F}^n, \ \|\widehat{x}\| = 1,$$

λ is a lower bound, and using $\widehat{x} = \delta_j$ demonstrates that it is the minimum. □

Exercise 3.9 (Maximization problem). Let $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix, and let $\lambda \in \sigma(A)$ be its *maximal* eigenvalue. Prove

$$\lambda = \max\{\langle Ax, x\rangle \ : \ x \in \mathbb{F}^n, \ \|x\| = 1\}.$$

Combine the result with Exercise 3.6 to show

$$\|A\| = \max\{|\lambda| \ : \ \lambda \in \sigma(A)\}.$$

The spectral norm owes its name to this property.

Exercise 3.10 (Eigenvectors). Let $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix, and let $\lambda \in \sigma(A)$ be its minimal eigenvalue. Let $x \in \mathbb{F}^n$ with $\|x\| = 1$. Prove that x is an eigenvector of A for the eigenvalue λ if and only if $\langle Ax, x\rangle = \lambda$ holds.

Based on Lemma 3.8, we can investigate the sensitivity of eigenvalues with respect to perturbations of the matrix. The result is the following simple version of the *Bauer–Fike theorem* [3]:

Proposition 3.11 (Perturbed eigenvalues). *Let $A, B \in \mathbb{F}^{n\times n}$ be self-adjoint matrices. For each eigenvalue μ of B, we can find an eigenvalue λ of A such that*

$$|\lambda - \mu| \le \|A - B\|.$$

Proof. Let $\mu \in \sigma(B)$. We choose $\lambda \in \sigma(A)$ with

$$|\lambda - \mu| = \min\{|\lambda' - \mu| \ : \ \lambda' \in \sigma(A)\}.$$

This choice implies that $\lambda - \mu$ is the smallest eigenvalue in modulus of $A - \mu I$. Since A and B are self-adjoint, we have $\lambda - \mu \in \mathbb{R}$, so $(\lambda - \mu)^2$ has to be the smallest eigenvalue of $(A - \mu I)^2$. Applying Lemma 3.8 to the latter matrix yields

$$(\lambda - \mu)^2 = \min\{\langle (A - \mu I)^2 x, x\rangle \ : \ x \in \mathbb{F}^n, \ \|x\| = 1\}.$$

Let $y \in \mathbb{F}^n \setminus \{0\}$ be an eigenvector of B for the eigenvalue μ. Since $y \neq 0$, we can assume $\|y\| = 1$ without loss of generality. Using (2.11) and (3.11), we find

$$\begin{aligned}
(\lambda - \mu)^2 &= \min\{\langle (A - \mu I)^2 x, x\rangle \ : \ x \in \mathbb{F}^n, \ \|x\| = 1\} \\
&\le \langle (A - \mu I)^2 y, y\rangle = \langle (B - \mu I + A - B)^2 y, y\rangle \\
&= \langle (B - \mu I)^2 y, y\rangle + \langle (A - B)(B - \mu I) y, y\rangle \\
&\quad + \langle (B - \mu I)(A - B) y, y\rangle + \langle (A - B)^2 y, y\rangle \\
&= \langle (B - \mu I) y, (A - B) y\rangle + \langle (A - B) y, (B - \mu I) y\rangle + \langle (A - B)^2 y, y\rangle \\
&= \langle (A - B) y, (A - B) y\rangle = \|(A - B) y\|^2 \le \|A - B\|^2 \|y\|^2 \\
&= \|A - B\|^2
\end{aligned}$$

due to $(B - \mu I) y = 0$, and since $\lambda - \mu$ is real, this implies $|\lambda - \mu| \le \|A - B\|$. □

Finding an estimate for the eigenvectors is slightly more challenging: if the "gap" between eigenvalues is small, perturbing the matrix may lead to large changes in the eigenvectors. As an example, we consider

$$A := \begin{pmatrix} 1 + 2\epsilon & \\ & 1 \end{pmatrix}, \quad B := \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 + \epsilon & \\ & 1 - \epsilon \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

for a small value $\epsilon > 0$. Using Exercise 3.9 and Proposition 2.13, we can compute the spectral norm

$$\|A - B\| = \left\| \begin{pmatrix} 2\epsilon & -\epsilon \\ -\epsilon & 0 \end{pmatrix} \right\| = \max\{|\lambda| \,:\, (\lambda - 2\epsilon)\lambda - \epsilon^2 = 0\}$$
$$= \max\{(1 - \sqrt{2})\epsilon, (1 + \sqrt{2})\epsilon\} = (1 + \sqrt{2})\epsilon.$$

Although the norm of the difference of A and B can be very small if ϵ is, the angles between each of the eigenvectors of both matrices always equal $\pi/2$. The reason is that the "gap" between the two eigenvalues of A equals only 2ϵ, and so a small perturbation of the eigenvalue can lead to a completely different eigenvector. In order to obtain a useful estimate for the approximation of eigenvectors, we can introduce a lower bound for this gap.

Definition 3.12 (Spectral gap). Let $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix. For all $\lambda \in \sigma(A)$, we define the *spectral gap* of λ and A by

$$\gamma_A(\lambda) := \inf\{|\mu - \lambda| \,:\, \mu \in \sigma(A) \setminus \{\lambda\}\}. \tag{3.12}$$

Using this quantity, we can formulate the following estimate for the eigenvectors of a perturbed matrix:

Proposition 3.13 (Perturbed eigenvectors). *Let $A, B \in \mathbb{F}^{n\times n}$ be self-adjoint matrices. For each eigenvector y of B for an eigenvalue μ we can find an eigenvalue λ of A and a vector $x \in \mathcal{E}(A, \lambda)$ in the corresponding eigenspace satisfying*

$$\|x - y\| \leq 2\frac{\|A - B\|}{\gamma_A(\lambda)}\|y\|.$$

Proof. Let $y \in \mathbb{F}^n \setminus \{0\}$ be an eigenvector for an eigenvalue μ of B. Due to Proposition 3.11, we can find an eigenvalue λ of A such that $|\lambda - \mu| \leq \|A - B\|$ holds.

According to Corollary 2.51, we can find a unitary matrix $Q \in \mathbb{F}^{n\times n}$ and a real diagonal matrix $D \in \mathbb{R}^{n\times n}$ satisfying

$$A = QDQ^*.$$

We let $\hat{y} := Q^* y$ and define $\hat{x} \in \mathbb{F}^n$ by

$$\hat{x}_j := \begin{cases} \hat{y}_j & \text{if } d_{jj} = \lambda, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } j \in \{1, \ldots, n\}.$$

For $x := Q\hat{x}$, we obtain

$$Ax = QDQ^*Q\hat{x} = QD\hat{x} = \lambda Q\hat{x} = \lambda x$$

since D is a diagonal matrix and $\hat{x}_j$ is only non-zero if $d_{jj} = \lambda$, and conclude $x \in \mathcal{E}(A, \lambda)$. Now we compute

$$\begin{aligned}\|(A-\lambda I)(x-y)\| &= \|(A-\lambda I)x - (A-\lambda I)y\| = \|(A-\lambda I)y\| \\ &= \|(B-\mu I)y + (A-B)y + (\mu-\lambda)y\| \\ &\leq \|(B-\mu I)y\| + \|(A-B)y\| + |\lambda-\mu|\,\|y\| \\ &\leq \|A-B\|\,\|y\| + \|A-B\|\,\|y\| = 2\|A-B\|\,\|y\|\end{aligned}$$

using the compatibility property (3.11) and the fact that y is an eigenvector of B. Now we only have to eliminate the matrix $A - \lambda I$ in this estimate:

$$\begin{aligned}\|(A-\lambda I)(x-y)\|^2 &= \|Q(D-\lambda I)Q^*(x-y)\|^2 = \|(D-\lambda I)(\widehat{x}-\widehat{y})\|^2 \\ &= \sum_{j=1}^{n} |d_{jj}-\lambda|^2 |\hat{x}_j - \hat{y}_j|^2 = \sum_{\substack{j=1 \\ d_{jj}\neq\lambda}}^{n} |d_{jj}-\lambda|^2 |\hat{x}_j - \hat{y}_j|^2 \\ &\geq \sum_{\substack{j=1 \\ d_{jj}\neq\lambda}}^{n} \gamma_A(\lambda)^2 |\hat{x}_j - \hat{y}_j|^2 = \gamma_A(\lambda)^2 \sum_{j=1}^{n} |\hat{x}_j - \hat{y}_j|^2 \\ &= \gamma_A(\lambda)^2 \|\widehat{x}-\widehat{y}\|^2 = \gamma_A(\lambda)^2 \|Q(\widehat{x}-\widehat{y})\|^2 \\ &= \gamma_A(\lambda)^2 \|x-y\|^2,\end{aligned}$$

where we have used our special choice of $\widehat{x}$ to avoid inconvenient terms in the sum and Lemma 2.38. □

The arguments used to eliminate $A - \lambda I$ in this proof are closely related to the concept of a *pseudo-inverse*: due to $\lambda \in \sigma(A)$, the matrix $A - \lambda I$ is not injective and therefore not invertible. Still, we can define an inverse in the *range* of $A-\lambda I$, and this is called a pseudo-inverse. In our case, the range of $A-\lambda I$ is the invariant subspace spanned by all eigenspaces except $\mathcal{E}(A, \lambda)$, and the vector x is chosen to ensure that $x - y$ is an element of this space.

Using this estimate, we can estimate the accuracy of the eigenvalue approximations:

Corollary 3.14 (Eigenvalue accuracy). *Let $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix. For each $i \in \{1, \ldots, n\}$, we can find $\lambda \in \sigma(A)$ such that*

$$|a_{ii} - \lambda| \leq \operatorname{off}(A),$$

and we can find a vector $x \in \mathcal{E}(A, \lambda)$ such that

$$\|\delta_i - x\| \leq 2\frac{\operatorname{off}(A)}{\gamma_A(\lambda)}.$$

Proof. We choose $B \in \mathbb{F}^{n\times n}$ to be the diagonal part of A, i.e.,

$$b_{ij} = \begin{cases} a_{ii} & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in \{1, \dots, n\}.$$

Due to Proposition 3.7, this implies

$$\|A - B\|^2 \leq \|A - B\|_F^2 = \operatorname{off}(A)^2.$$

Let $i \in \{1, \dots, n\}$. Since a_{ii} is an eigenvalue of the diagonal matrix B, Proposition 3.11 states that we can find an eigenvalue $\lambda \in \sigma(A)$ satisfying

$$|a_{ii} - \lambda| \leq \|B - A\| \leq \operatorname{off}(A).$$

Since δ_i is a corresponding eigenvector, we can apply Proposition 3.13 to find the required vector $x \in \mathcal{E}(A, \lambda)$. □

The estimate given here is suboptimal: combining Proposition 3.13 with the estimate provided by Theorem 4.6 for the *Rayleigh quotient* introduced in Section 4.2 yields that $|a_{ii} - \lambda|$ converges like $\operatorname{off}(A)^2$.

Under certain additional conditions, Jacobi's method offers the advantage that it can be used to compute eigenvalues to a very high *relative* accuracy, while most other algorithms typically reach only a high *absolute* accuracy [8].

3.5 Quadratic convergence *

Under certain conditions, the Jacobi iteration converges *quadratically*, i.e., it can be significantly faster if the matrix A is sufficiently close to diagonal form [19]. This section follows the outline of the proofs presented in [51] and [39], slightly adapted to fit our definition of the Jacobi iteration.

Obviously, the Jacobi iteration can only converge if the matrices do not change too much when applying the unitary transformation. This means that $|s|$ in (3.1), the sine of the rotation angle, has to be small. Due to (3.6), we have

$$|s| \leq |t|\,|c| \leq |t| \leq \frac{1}{|\tau| + \sqrt{|\tau|^2 + 1}} \leq \frac{1}{2|\tau|} = \frac{|a_{pq}|}{|a_{pp} - a_{qq}|} \tag{3.13}$$

for the elementary Jacobi step that eliminates a_{pq} with $1 \leq p < q \leq n$. This bound is only useful if we can ensure that the diagonal elements of the matrix are sufficiently far from each other.

Due to Corollary 3.4, we know that the diagonal elements converge to eigenvalues, so we can only expect well-separated diagonal elements if the eigenvalues are well-separated. For the remainder of this section, we assume that A has only simple eigenvalues, i.e., that

$$\#\sigma(A) = n \tag{3.14}$$

holds. The minimal distance of two eigenvalues is denoted by

$$\Delta := \min\{|\lambda - \mu| \,:\, \lambda, \mu \in \sigma(A),\ \lambda \neq \mu\}, \tag{3.15}$$

and we can use it in combination with Proposition 3.11 to obtain the following estimate for the diagonal entries:

Lemma 3.15 (Diagonal entries). *Let* $2\operatorname{off}(A) < \Delta$, *i.e.,* $\delta := \Delta - 2\operatorname{off}(A) > 0$. *Then we have*

$$|a_{pp} - a_{qq}| \geq \delta \quad \text{for all } p, q \in \{1, \ldots, n\},\ p \neq q.$$

Proof. (cf. [51, eq. (1.6)]) Let $D \in \mathbb{F}^{n\times n}$ be the diagonal part of A, i.e.,

$$D = \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{pmatrix},$$

We have $\|A - D\|_F = \operatorname{off}(A)$, and combining Proposition 3.11 with Proposition 3.7 yields that for each $\lambda \in \sigma(A)$ we can find $i_\lambda \in \{1, \ldots, n\}$ with

$$|\lambda - a_{i_\lambda i_\lambda}| \leq \operatorname{off}(A). \tag{3.16}$$

Let now $\mu \in \sigma(A)$ with

$$|\mu - a_{i_\lambda i_\lambda}| \leq \operatorname{off}(A).$$

By the triangle inequality, this implies

$$|\lambda - \mu| = |\lambda - a_{i_\lambda i_\lambda} + a_{i_\lambda i_\lambda} - \mu| \leq |\lambda - a_{i_\lambda i_\lambda}| + |\mu - a_{i_\lambda i_\lambda}| \leq 2\operatorname{off}(A) < \Delta,$$

i.e., $\lambda = \mu$ by our assumption (3.15). This means that the mapping $\lambda \mapsto i_\lambda$ defined by (3.16) is injective, and (3.14) yields that it also has to be bijective.

Let now $p, q \in \{1, \ldots, n\}$ be fixed with $p \neq q$. We have seen that there are $\lambda, \mu \in \sigma(A)$ with $\lambda \neq \mu$ such that $p = i_\lambda$ and $q = i_\mu$, and conclude

$$\begin{aligned} |a_{pp} - a_{qq}| &= |a_{i_\lambda i_\lambda} - a_{i_\mu i_\mu}| = |\lambda - \mu - (\lambda - a_{i_\lambda i_\lambda}) - (a_{i_\mu i_\mu} - \mu)| \\ &\geq |\lambda - \mu| - |\lambda - a_{i_\lambda i_\lambda}| - |\mu - a_{i_\mu i_\mu}| \geq \Delta - 2\operatorname{off}(A) = \delta \end{aligned}$$

by the triangle inequality. □

We consider one sweep of the algorithm, i.e., $N := n(n-1)/2$ elementary Jacobi steps. In the r-th step, we choose $1 \leq p_r < q_r \leq n$ and eliminate the entry in the p_r-th row and q_r-th column of the matrix. We denote the sine of the rotation angle (cf. (3.5)) by s_r and the cosine by c_r. The original matrix is written as $A^{(0)}$, and the matrix resulting from applying the r-th transformation to $A^{(r-1)}$ as $A^{(r)}$. Due to our construction the r-th transformation is chosen to eliminate $a_{p_r q_r}^{(r-1)}$. The key to the convergence analysis is the following bound for the Jacobi angles:

Lemma 3.16 (Jacobi angles). *Let* $\delta := \Delta - 2\operatorname{off}(A) > 0$. *We have*

$$\sum_{r=1}^{N} |s_r|^2 \leq \frac{\operatorname{off}(A^{(0)})^2}{2\delta^2}. \tag{3.17}$$

Proof. (cf. [51, eq. (2.2)] and [39, eq. (9)]) Theorem 3.2 yields

$$\operatorname{off}(A^{(r)})^2 = \operatorname{off}(A^{(r-1)})^2 - 2|a_{p_r q_r}^{(r-1)}|^2 \quad \text{for all } r \in \{1, \ldots, N\},$$

and a simple induction gives us

$$0 \leq \operatorname{off}(A^{(N)})^2 = \operatorname{off}(A^{(0)})^2 - 2\sum_{r=1}^{N} |a_{p_r q_r}^{(r-1)}|^2,$$

which implies

$$\sum_{r=1}^{N} |a_{p_r q_r}^{(r-1)}|^2 \leq \frac{\operatorname{off}(A^{(0)})^2}{2}. \tag{3.18}$$

Due to $\operatorname{off}(A^{(r-1)}) \leq \operatorname{off}(A^{(0)})$, we can apply Lemma 3.15 to $A^{(r-1)}$, and (3.13) yields

$$\sum_{r=1}^{N} |s_r|^2 \leq \sum_{r=1}^{N} \frac{|a_{p_r q_r}^{(r-1)}|^2}{|a_{p_r p_r}^{(r-1)} - a_{q_r q_r}^{(r-1)}|^2} \leq \sum_{r=1}^{N} \frac{|a_{p_r q_r}^{(r-1)}|^2}{\delta^2} \leq \frac{\operatorname{off}(A^{(0)})^2}{2\delta^2}. \qquad \square$$

In the r-th step of the sweep, the entry $a_{p_r q_r}^{(r-1)}$ is eliminated, we have $a_{p_r q_r}^{(r)} = 0$. Unfortunately, this entry will usually not stay eliminated: if there is an $\ell \in \{1, \ldots, N\}$ with $p_r = p_\ell$ and $q_r \neq q_\ell$, the p_r-th row will be changed and $a_{p_r q_r}^{(\ell)}$ may no longer be equal to zero. In fact, (3.9) yields

$$\begin{aligned} |a_{p_r q_r}^{(\ell)}| &= |\bar{c}_r a_{p_r q_r}^{(\ell-1)} - \bar{s}_r a_{q_\ell q_r}^{(\ell-1)}| \leq |\bar{c}_r|\,|a_{p_r q_r}^{(\ell-1)}| + |\bar{s}_r|\,|a_{q_\ell q_r}^{(\ell-1)}| \\ &\leq |a_{p_r q_r}^{(\ell-1)}| + |s_r|\,|a_{q_\ell q_r}^{(\ell-1)}|. \end{aligned} \tag{3.19a}$$

If we have $p_r = q_\ell$ and $q_r \neq p_\ell$, the p_r-th row will also be changed:

$$\begin{aligned} |a_{p_r q_r}^{(\ell)}| &= |s_r a_{p_\ell q_r}^{(\ell-1)} + c_r a_{p_r q_r}^{(\ell-1)}| \leq |s_r|\,|a_{p_\ell q_r}^{(\ell-1)}| + |c_r|\,|a_{p_r q_r}^{(\ell-1)}| \\ &\leq |a_{p_r q_r}^{(\ell-1)}| + |s_r|\,|a_{p_\ell q_r}^{(\ell-1)}|. \end{aligned} \tag{3.19b}$$

The same reasoning applies to the columns: if $q_r = p_\ell$ and $p_r \neq q_\ell$, (3.8) implies

$$\begin{aligned} |a^{(\ell)}_{p_r q_r}| &= |a^{(\ell-1)}_{p_r q_r} c_r - a^{(\ell-1)}_{p_r q_\ell} s_r| \leq |a^{(\ell-1)}_{p_r q_r}|\,|c_r| + |a^{(\ell-1)}_{p_r q_\ell}|\,|s_r| \\ &\leq |a^{(\ell-1)}_{p_r q_r}| + |s_r|\,|a^{(\ell-1)}_{p_r q_\ell}|, \end{aligned} \tag{3.19c}$$

while $q_r = q_\ell$ and $p_r \neq p_\ell$ leads to

$$\begin{aligned} |a^{(\ell)}_{p_r q_r}| &= |a^{(\ell-1)}_{p_r p_\ell} \bar{s}_r + a^{(\ell-1)}_{p_r q_r} \bar{c}_r| \leq |a^{(\ell-1)}_{p_r p_\ell}|\,|s_r| + |a^{(\ell-1)}_{p_r q_r}|\,|c_r| \\ &\leq |a^{(\ell-1)}_{p_r q_r}| + |s_r|\,|a^{(\ell-1)}_{p_r p_\ell}|. \end{aligned} \tag{3.19d}$$

For $p_r = p_\ell$ and $q_r = q_\ell$, $a^{(\ell-1)}_{p_r q_r}$ is eliminated again, and $p_r = q_\ell$ and $q_r = p_\ell$ is impossible due to $p_r < q_r$ and $p_\ell < q_\ell$. If $\{p_r, q_r\} \cap \{p_\ell, q_\ell\} = \emptyset$, $a^{(\ell-1)}_{p_r q_r}$ is not changed. We can summarize (3.19) as follows:

Lemma 3.17 (Bound for one Jacobi step). *Let $\ell \in \{1, \ldots, N\}$. Define the matrix $E^{(\ell)} \in \mathbb{F}^{n\times n}$ by*

$$e^{(\ell)}_{ij} = \begin{cases} a^{(\ell-1)}_{q_\ell j} & \text{if } i = p_\ell,\ j \neq q_\ell, \\ a^{(\ell-1)}_{p_\ell j} & \text{if } i = q_\ell,\ j \neq p_\ell, \\ a^{(\ell-1)}_{i q_\ell} & \text{if } j = p_\ell,\ i \neq q_\ell, \\ a^{(\ell-1)}_{i p_\ell} & \text{if } j = q_\ell,\ i \neq p_\ell, \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } i, j \in \{1, \ldots, n\}. \tag{3.20}$$

Then we have

$$|a^{(\ell)}_{ij}| \leq |a^{(\ell-1)}_{ij}| + |s_\ell|\,|e^{(\ell)}_{ij}| \quad \text{for all } i, j \in \{1, \ldots, n\}.$$

Proof. Combine the estimates (3.19) and take advantage of the fact that $A^* = A$ implies $(A^{(\ell-1)})^* = A^{(\ell-1)}$, which in turn implies $(E^{(\ell)})^* = E^{(\ell)}$. □

If we use a cyclic Jacobi method, i.e., if each off-diagonal entry is eliminated once during the sweep, we obtain the following convergence estimate:

Theorem 3.18 (Cyclic Jacobi method). *Let $\delta := \Delta - 2\,\mathrm{off}(A^{(0)}) > 0$. If for each $1 \leq p < q \leq n$ there is an $r \in \{1, \ldots, N\}$ with $p_r = p$ and $q_r = q$, i.e., if each off-diagonal entry is eliminated once during the sweep, we have*

$$\mathrm{off}(A^{(N)}) \leq \sqrt{\frac{N}{2\delta^2}}\,\mathrm{off}(A^{(0)})^2, \tag{3.21}$$

i.e., the Jacobi method converges quadratically.

Proof. (cf. [51, Section 2]) Let $1 \leq p < q \leq n$. By our assumption, we can find $r \in \{1, \ldots, N\}$ with $p_r = p$ and $q_r = q$. Since the Jacobi rotation eliminates $a_{p_r q_r}^{(r-1)}$, we have $a_{pq}^{(r)} = 0$. Lemma 3.17 and a simple induction yield

$$|a_{pq}^{(N)}| \leq \sum_{\ell=r+1}^{N} |s_\ell| \, |e_{pq}^{(\ell)}| \leq \sum_{\ell=1}^{N} |s_\ell| \, |e_{pq}^{(\ell)}|.$$

By the Cauchy–Schwarz inequality, we find

$$|a_{pq}^{(N)}|^2 \leq \left(\sum_{\ell=1}^{N} |s_\ell| \, |e_{pq}^{(\ell)}| \right)^2 \leq \left(\sum_{\ell=1}^{N} |s_\ell|^2 \right) \left(\sum_{\ell=1}^{N} |e_{pq}^{(\ell)}|^2 \right).$$

Using Lemma 3.16, we obtain

$$\operatorname{off}(A^{(N)})^2 \leq \left(\sum_{\ell=1}^{N} |s_\ell|^2 \right) \sum_{\ell=1}^{N} \operatorname{off}(E^{(\ell)})^2 \leq \frac{\operatorname{off}(A^{(0)})^2}{2\delta^2} \sum_{\ell=1}^{N} \operatorname{off}(E^{(\ell)})^2. \tag{3.22}$$

A look at (3.20) in combination with Theorem 3.2 reveals

$$\operatorname{off}(E^{(\ell)})^2 \leq \operatorname{off}(A^{(\ell)})^2 \leq \operatorname{off}(A^{(0)})^2,$$

so we have found

$$\operatorname{off}(A^{(N)})^2 \leq \frac{\operatorname{off}(A^{(0)})^2}{2\delta^2} \sum_{\ell=1}^{N} \operatorname{off}(E^{(\ell)})^2 \leq \frac{N}{2\delta^2} \operatorname{off}(A^{(0)})^4,$$

and this implies (3.21). □

Remark 3.19 (Improvements). This is obviously a very rough estimate, particularly the inequality $\operatorname{off}(E^{(\ell)})^2 \leq \operatorname{off}(A^{(0)})^2$ is very pessimistic, since most of the entries of $E^{(\ell)}$ are equal to zero.

For the cyclic-by-row Jacobi method, the convergence estimate can be improved significantly to obtain

$$\operatorname{off}(A^{(N)}) \leq \sqrt{\frac{1}{2\delta^2}} \operatorname{off}(A^{(0)})^2$$

by analyzing the entries of the intermediate matrices more carefully [51, Section 3].

Exercise 3.20 (Improved estimate). Take the structure of the matrices $(E^{(\ell)})_{\ell=1}^{N}$ into account to improve the result of Theorem 3.18 to

$$\operatorname{off}(A^{(N)}) \leq \sqrt{\frac{n}{\delta^2}} \operatorname{off}(A^{(0)})^2.$$

Hint: Theorem 3.2 implies $\operatorname{off}(A^{(\ell)}) \leq \operatorname{off}(A^{(0)})$ for all $\ell \in \{0, \dots, N\}$. How often is each of the entries in the matrix changed during a sweep? Can you improve the estimate even further, maybe using additional assumptions regarding the sequence of elementary steps?

Theorem 3.21 (Classical Jacobi method). *Let $\delta := \Delta - 2\operatorname{off}(A^{(0)}) > 0$, and let the indices $(p_r)_{r=1}^N$ and $(q_r)_{r=1}^N$ be chosen as in the classical Jacobi method, i.e., satisfying*

$$|a_{p_r q_r}^{(r-1)}| \geq |a_{ij}^{(r-1)}| \quad \text{for all } 1 \leq i < j \leq n,\ r \in \{1, \dots, N\}. \tag{3.23}$$

Then we have

$$\operatorname{off}(A^{(N)}) \leq \sqrt{\frac{n-2}{\delta^2}}\operatorname{off}(A^{(0)})^2, \tag{3.24}$$

i.e., the classical Jacobi method also converges quadratically.

Proof. (cf. [39, Satz 2]) Let $I_N := \{(i, j) : 1 \leq i < j \leq n\}$. We prove by induction that for each $r \in \{1, \dots, N\}$ we can find $I_r \subseteq I_N$ with

$$\#I_r = r, \qquad \sum_{(i,j)\in I_r} |a_{ij}^{(r)}|^2 \leq 2(n-2)\left(\sum_{\ell=1}^{r} |s_\ell|\,|a_{p_\ell q_\ell}^{(\ell-1)}|\right)^2. \tag{3.25}$$

We can see that for $r = N$, this estimate implies an upper bound for $\operatorname{off}(A^{(N)})^2/2$.

For $r = 1$, we choose $I_1 = \{(p_1, q_1)\}$. The Jacobi step guarantees $a_{p_1 q_1}^{(1)} = 0$, so (3.25) holds trivially.

Assume now that $r \in \{1, \dots, N-1\}$ is given such that a subset $I_r \subseteq I_N$ satisfying (3.25) exists. Let $J \subseteq I_N$ be a subset of cardinality $\#J = r$ such that

$$\sum_{(i,j)\in J} |a_{ij}^{(r)}|^2$$

is minimal among all those subsets. This implies

$$\sum_{(i,j)\in J} |a_{ij}^{(r)}|^2 \leq \sum_{(i,j)\in I_r} |a_{ij}^{(r)}|^2.$$

Since $r < N$ and (3.23) holds, we can assume $(p_r, q_r) \notin J$, since otherwise we could replace it by a different index pair without increasing the sum. Therefore the set

$$I_{r+1} := J \cup \{(p_r, q_r)\}$$

has cardinality $\#I_{r+1} = r + 1$. Due to Lemma 3.17, we have

$$|a_{ij}^{(r+1)}| \leq |a_{ij}^{(r)}| + |s_{r+1}|\,|e_{ij}^{(r+1)}| \quad \text{for all } (i, j) \in J.$$

A look at (3.20) reveals that only $4(n-2)$ entries of $E^{(\ell)}$, namely the p_ℓ-th and q_ℓ-th rows and columns with the exception of their intersections, are non-zero. To take advantage of this property, let

$$J_+ := \{(i,j) \in J \ : (i \in \{p_{r+1}, q_{r+1}\} \wedge j \notin \{p_{r+1}, q_{r+1}\}) \vee (j \in \{p_{r+1}, q_{r+1}\} \wedge i \notin \{p_{r+1}, q_{r+1}\})\}$$

denote the set of all coefficients in J changed during the update from $A^{(r)}$ to $A^{(r+1)}$. Its cardinality equals $\#J_+ = 2(n-2)$, since J contains only index pairs above the diagonal. We use (3.23) to bound $|e_{ij}^{(r+1)}|$ by $|a_{p_{r+1}q_{r+1}}^{(r)}|$ for all $(i,j) \in J_+$ and apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}
\sum_{(i,j)\in J} |a_{ij}^{(r+1)}|^2 &\le \sum_{(i,j)\in J} (|a_{ij}^{(r)}| + |s_{r+1}|\,|e_{ij}^{(r+1)}|)^2 \\
&= \sum_{(i,j)\in J} (|a_{ij}^{(r)}|^2 + 2|a_{ij}^{(r)}|\,|s_{r+1}|\,|e_{ij}^{(r+1)}| + |s_{r+1}|^2 |e_{ij}^{(r+1)}|^2) \\
&\le \sum_{(i,j)\in J} |a_{ij}^{(r)}|^2 + \sum_{(i,j)\in J_+} |s_{r+1}|^2 |a_{p_{r+1}q_{r+1}}^{(r)}|^2 \\
&\quad + 2|a_{p_{r+1}q_{r+1}}^{(r)}|\,|s_{r+1}| \sum_{(i,j)\in J_+} |a_{ij}^{(r)}| \\
&\le \sum_{(i,j)\in J} |a_{ij}^{(r)}|^2 + 2(n-2)|s_{r+1}|^2 |a_{p_{r+1}q_{r+1}}^{(r)}|^2 \\
&\quad + 2|a_{p_{r+1}q_{r+1}}^{(r)}|\,|s_{r+1}| \sqrt{2(n-2)} \sqrt{\sum_{(i,j)\in J_+} |a_{ij}^{(r)}|^2} \\
&= \left(\sqrt{\sum_{(i,j)\in J} |a_{ij}^{(r)}|^2} + \sqrt{2(n-2)}|s_{r+1}|\,|a_{p_{r+1}q_{r+1}}^{(r)}| \right)^2 .
\end{aligned}$$

Applying the induction assumption (3.25) to the first term yields

$$\begin{aligned}
\sum_{(i,j)\in J_+} |a_{ij}^{(r+1)}|^2 &\le \left(\sqrt{2(n-2)} \sum_{\ell=1}^{r} |s_\ell|\,|a_{p_\ell q_\ell}^{(\ell-1)}| + \sqrt{2(n-2)}|s_{r+1}|\,|a_{p_{r+1}q_{r+1}}^{(r)}| \right)^2 \\
&= 2(n-2) \left(\sum_{\ell=1}^{r+1} |s_\ell|\,|a_{p_\ell q_\ell}^{(\ell-1)}| \right)^2 .
\end{aligned}$$

This completes the induction.

We apply (3.25) to $r = N$ and use the Cauchy–Schwarz inequality, Lemma 3.16 and (3.18) to obtain

$$\begin{aligned}\operatorname{off}(A^{(N)})^2 &= 2 \sum_{(i,j)\in I_N} |a_{ij}^{(r+1)}|^2 \le 4(n-2) \left(\sum_{\ell=1}^{N} |s_\ell| \, |a_{p_\ell q_\ell}^{(\ell-1)}| \right)^2 \\ &\le 4(n-2) \left(\sum_{\ell=1}^{N} |s_\ell|^2 \right) \left(\sum_{\ell=1}^{N} |a_{p_\ell q_\ell}^{(\ell-1)}|^2 \right) \\ &\le 4(n-2) \frac{\operatorname{off}(A^{(0)})^2}{2\delta^2} \frac{\operatorname{off}(A^{(0)})^2}{2} = \frac{n-2}{\delta^2} \operatorname{off}(A^{(0)})^4. \end{aligned}$$

□

Remark 3.22 (Multiple eigenvalues). In order to prove quadratic convergence, we rely on the assumption that all eigenvalues of the matrix A are simple. This assumption can be avoided if we instead require all diagonal elements converging to the same multiple eigenvalue to be grouped together [46].

Applying generalized rotations to entire blocks of the matrix A, a variant of the cyclic Jacobi method can be derived that converges *globally* [13].

Chapter 4

Power methods

Summary

This chapter introduces the power iteration, a very simple method for computing eigenvectors that paves the way for a number of very important and efficient methods like the inverse iteration, the Rayleigh iteration, and the simultaneous iteration. We investigate the convergence of these iterations and provide simple estimates for the accuracy of the approximations of eigenvalues and eigenvectors that can be used to construct stopping criteria for the iteration.

Learning targets

- ✓ Introduce the power iteration and the (shifted) inverse iteration.
- ✓ Analyze the convergence of these methods.
- ✓ Introduce the Rayleigh iteration and establish its cubic convergence.
- ✓ Investigate the convergence to an invariant subspace instead of an eigenspace.
- ✓ Introduce the simultaneous iteration to compute a basis of this invariant subspace.

4.1 Power iteration

The power method (also known as the *Von Mises iteration* [47]) is a very simple and yet very flexible method for computing an eigenvector. The algorithm can be traced back to a publication [35] that appeared in 1921 and uses it (or more precisely a variant of the related inverse iteration) to solve a generalized eigenproblem arising in the context of structural mechanics. Our discussion of the method and its proof follows the theory already outlined in [35, 47].

Let $n \in \mathbb{N}$, and let $A \in \mathbb{F}^{n\times n}$ be a matrix. In order to keep the theoretical investigation simple, we assume that A is a normal matrix if $\mathbb{F} = \mathbb{C}$ or self-adjoint if $\mathbb{F} = \mathbb{R}$. Note that the theory can be extended to more general matrices without changing the algorithm (cf. Section 4.8).

The idea of the *power iteration* is the following: assume that a vector $x \in \mathbb{F}^n$ is given and that it can be represented as a linear combination

$$x = \alpha_1 e_1 + \cdots + \alpha_k e_k$$

of eigenvectors $e_1, \ldots, e_k \in \mathbb{F}^n$ corresponding to eigenvalues $\lambda_1, \ldots, \lambda_k$. Then we have

$$A^m x = \alpha_1 \lambda_1^m e_1 + \cdots + \alpha_k \lambda_k^m e_k \quad \text{for all } m \in \mathbb{N}_0.$$

If one of the eigenvalues is larger in modulus than the others, the corresponding term grows faster than all others for $m \to \infty$, and we can expect to obtain an approximation of an eigenvector.

In order to analyze the behaviour of the sequence $(x^{(m)})_{m=0}^{\infty}$ given by

$$x^{(m)} := A^m x^{(0)} = A x^{(m-1)} \quad \text{for all } m \in \mathbb{N} \tag{4.1}$$

more closely, we represent the vectors in a suitable basis consisting of eigenvectors. Since A is normal or self-adjoint, Corollaries 2.50 or 2.51, respectively, imply that we can find a unitary matrix $Q \in \mathbb{F}^{n\times n}$ and a diagonal matrix $D \in \mathbb{F}^{n\times n}$ such that the Schur decomposition takes the form

$$A = QDQ^*, \quad D = Q^* A Q. \tag{4.2}$$

The diagonal elements of D are its eigenvalues, and therefore also the eigenvalues of A. We denote these eigenvalues by $\lambda_1, \ldots, \lambda_n$ given by

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}. \tag{4.3}$$

We assume that A has a *dominant eigenvalue*, i.e., that one eigenvalue is simple and larger in modulus than all others. Since changing the order of the columns of Q and D will leave the Schur decomposition intact, we can assume

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|. \tag{4.4}$$

Let a starting vector $x^{(0)} \in \mathbb{F}^n \setminus \{0\}$ be given, and let the iteration vectors $x^{(1)}$, $x^{(2)}, \ldots$ be defined as in (4.1). We transform the vectors into the eigenvector basis, i.e., define

$$\widehat{x}^{(m)} := Q^* x^{(m)} \quad \text{for all } m \in \mathbb{N}_0,$$

and (4.1) takes the form

$$\begin{aligned} \widehat{x}^{(m+1)} &= Q^* x^{(m+1)} = Q^* A x^{(m)} = Q^* A Q \widehat{x}^{(m)} = D\widehat{x}^{(m)}, \\ \widehat{x}^{(m)} &= D^m \widehat{x}^{(0)} \quad \text{for all } m \in \mathbb{N}_0. \end{aligned}$$

Since D is a diagonal matrix, this equation allows us to compute the components of the iteration vectors directly:

$$\widehat{x}_j^{(m+1)} = \lambda_j \widehat{x}_j^{(m)} \quad \text{holds for all } m \in \mathbb{N}_0,\ j \in \{1, \ldots, n\}. \tag{4.5}$$

Due to (4.4), the first component of $\widehat{x}^{(m)}$ will grow more rapidly than any other component as $m \to \infty$, so we can expect the transformed iteration vectors to "converge" to multiples of the eigenvector δ_1 and the original vectors to multiples of the eigenvector $e_1 := Q\delta_1$.

Unfortunately, if $\lambda_1 \neq 1$, there can be no convergence in the usual sense, since the length of the iteration vectors is changed by applying the matrix. Instead, we look for convergence of the *angle* between the iteration vectors and the eigenvector.

Definition 4.1 (Angle between vectors). Let $x, y \in \mathbb{F}^n \setminus \{0\}$. We define

$$\cos \angle(x,y) := \frac{|\langle x, y\rangle|}{\|x\| \, \|y\|},$$
$$\sin \angle(x,y) := \sqrt{1 - \cos^2 \angle(x,y)},$$
$$\tan \angle(x,y) := \frac{\sin \angle(x,y)}{\cos \angle(x,y)}.$$

If $\cos \angle(x,y) = 0$ holds, we let $\tan \angle(x,y) = \infty$.

Since Q is unitary, Q^* is isometric, and Lemma 2.38 yields

$$\begin{aligned}\cos \angle(Q^*x, Q^*y) &= \frac{|\langle Q^*x, Q^*y\rangle|}{\|Q^*x\| \, \|Q^*y\|} = \frac{|\langle QQ^*x, y\rangle|}{\|x\| \, \|y\|} \\ &= \frac{|\langle x, y\rangle|}{\|x\| \, \|y\|} = \cos \angle(x,y) \quad \text{for all } x, y \in \mathbb{F}^n \setminus \{0\}. \end{aligned} \tag{4.6}$$

Due to our definition, the sine and the tangent also remain unchanged by unitary transformations. Now we can formulate a first convergence result: we let $e_1 := Q\delta_1$ and observe

$$Ae_1 = AQ\delta_1 = QQ^*AQ\delta_1 = QD\delta_1 = \lambda_1 Q\delta_1 = \lambda_1 e_1,$$

i.e., e_1 is a convenient eigenvector for the dominant eigenvalue λ_1. In the case of a dominant eigenvalue, i.e., if (4.4) holds, we can now establish the following convergence estimate:

Theorem 4.2 (Convergence). *Let* (4.2) *and* (4.3) *hold with*

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Let $e_1 := Q\delta_1$ *and* $\cos \angle(x^{(0)}, e_1) \neq 0$. *Let* $(x^{(m)})_{m=0}^{\infty}$ *be given by (4.1). Then we have*

$$\tan \angle(x^{(m+1)}, e_1) \leq \frac{|\lambda_2|}{|\lambda_1|} \tan \angle(x^{(m)}, e_1) \quad \textit{for all } m \in \mathbb{N}_0,$$

and due to $|\lambda_1| > |\lambda_2|$, *this means that the angle between the iteration vectors and the eigenvector* e_1 *converges to zero.*

Proof. Due to (4.6), we have

$$\begin{aligned}
\cos^2 \angle(x^{(m)}, e_1) &= \cos^2 \angle(\widehat{x}^{(m)}, \delta_1) = \frac{|\langle \widehat{x}^{(m)}, \delta_1 \rangle|^2}{\|\widehat{x}^{(m)}\|^2 \, \|\delta_1\|^2} = \frac{|\widehat{x}_1^{(m)}|^2}{\|\widehat{x}^{(m)}\|^2}, \\
\sin^2 \angle(x^{(m)}, e_1) &= 1 - \cos^2 \angle(x^{(m)}, e_1) = 1 - \frac{|\widehat{x}_1^{(m)}|^2}{\|\widehat{x}^{(m)}\|^2} = \frac{\|\widehat{x}^{(m)}\|^2 - |\widehat{x}_1^{(m)}| r}{\|\widehat{x}^{(m)}\|^2} \\
&= \frac{\sum_{j=1}^n |\widehat{x}_j^{(m)}|^2 - |\widehat{x}_1^{(m)}|^2}{\|\widehat{x}^{(m)}\|^2} = \frac{\sum_{j=2}^n |\widehat{x}_j^{(m)}|^2}{\|\widehat{x}^{(m)}\|^2}, \\
\tan^2 \angle(x^{(m)}, e_1) &= \frac{\sin^2 \angle(x^{(m)}, e_1)}{\cos^2 \angle(x^{(m)}, e_1)} = \frac{\sum_{j=2}^n |\widehat{x}_j^{(m)}|^2}{|\widehat{x}_1^{(m)}|^2} \qquad \text{for all } m \in \mathbb{N}_0.
\end{aligned}$$

Using (4.5) and (4.4) yields

$$\begin{aligned}
\tan^2 \angle(x^{(m+1)}, e_1) &= \frac{\sum_{j=2}^n |\widehat{x}_j^{(m+1)}|^2}{|\widehat{x}_1^{(m+1)}|^2} = \frac{\sum_{j=2}^n |\lambda_j|^2 |\widehat{x}_j^{(m)}|^2}{|\lambda_1|^2 |\widehat{x}_1^{(m)}|^2} \\
&\leq \frac{|\lambda_2|^2 \sum_{j=2}^n |\widehat{x}_j^{(m)}|^2}{|\lambda_1|^2 |\widehat{x}_1^{(m)}|^2} = \left(\frac{|\lambda_2|}{|\lambda_1|} \right)^2 \tan^2 \angle(x^{(m)}, e_1) \\
&\qquad \text{for all } m \in \mathbb{N}_0,
\end{aligned}$$

and taking the square root of both sides of the inequality yields the result. □

The condition $\cos \angle(x^{(0)}, e_1) \neq 0$ is required to ensure that the denominators in the proof are well-defined: it implies $\widehat{x}_1^{(0)} \neq 0$, and since we also have $|\lambda_1| > |\lambda_2| \geq 0$ due to (4.4), this implies

$$\widehat{x}_1^{(m)} = \lambda_1^m \widehat{x}_1^{(0)} \neq 0, \quad \|\widehat{x}^{(m)}\| \neq 0 \qquad \text{for all } m \in \mathbb{N}_0.$$

This requirement is also quite natural: if $\cos \angle(x^{(0)}, e_1) = 0$ holds, we have $\widehat{x}_1^{(0)} = 0$, so the starting vector contains no component in the direction of the desired eigenvector, and this will not change no matter how often we apply the iteration step.

The assumption that A is normal or self-adjoint is not necessary, it only serves to make the theoretical investigation of the convergence more elegant. If A is merely diagonalizable, i.e., if there is a regular matrix $T \in \mathbb{F}^{n\times n}$ such that

$$T^{-1}AT = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

holds, we can introduce transformed iteration vectors

$$\widehat{x}^{(m)} := T^{-1}x^{(m)} \quad \text{for all } m \in \mathbb{N}_0$$

and obtain

$$\widehat{x}^{(m+1)} = T^{-1}x^{(m+1)} = T^{-1}Ax^{(m)} = T^{-1}AT\widehat{x}^{(m)} = D\widehat{x}^{(m)} \quad \text{for all } m \in \mathbb{N}_0,$$

i.e., the transformed vectors result from a power iteration using the matrix D and the starting vector $\widehat{x}^{(0)}$. Applying Theorem 4.2 to the transformed vectors yields convergence estimates.

In a practical implementation, we have to replace the field $\mathbb{F}$ by machine numbers, and these numbers cannot exceed a given maximum. This can lead to problems if $|\lambda_1|$ is larger that one: the component $\widehat{x}_1^{(m)} = \lambda_1^m \widehat{x}_1^{(0)}$ can grow beyond the maximum, and this usually leads to an overflow. A similar effect occurs if $|\lambda_1|$ is smaller than one: $\widehat{x}_1^{(m)}$ tends to zero, and truncation will eliminate this component after a certain number of iterations. In order to avoid these undesirable effects, we *normalize* the iteration vectors, i.e., we ensure that they always have unit norm.

Using machine numbers also means that rounding effects will influence the convergence behaviour, cf. [52, Chapter 9].

```
procedure power_iteration(A, var x);
begin
  while „not sufficiently accurate“ do begin
    a ← Ax;
    x ← a/‖a‖
  end
end
```

Figure 4.1. Power iteration, still missing a stopping criterion.

Since the angle between vectors does not depend on their scaling, this leaves the convergence estimate of Theorem 4.2 intact and leads to the numerically stable algorithm summarized in Figure 4.1.

An important advantage of the power iteration is its flexibility: we only have to be able to evaluate matrix-vector products, we do not require the matrix A itself. In particular, we do not have to apply transformations to the matrix like in the case of the Jacobi iteration. This allows us to use the power iteration in situations where A is only given implicity: e.g., if A is the inverse of a matrix B, we can replace the matrix-vector multiplication $a = Ax = B^{-1}x$ by solving the system $Ba = x$, and this is usually easier and more stable.

Exercise 4.3 (Indefinite matrix). Consider the matrix

$$A := \begin{pmatrix} 2 & \\ & -2 \end{pmatrix}.$$

Investigate the convergence of the power iteration for the starting vectors

$$\widehat{x}^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \widehat{x}^{(0)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \widehat{x}^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Does the angle between $\widehat{x}^{(m)}$ and an eigenspace always converge? If not, explain why this is not a counterexample for Theorem 4.2.

4.2 Rayleigh quotient

Before we introduce refined variants of the power iteration, we investigate the question of a proper stopping criterion. The power iteration gives us a sequence $(x^{(m)})_{m=0}^{\infty}$ of vectors that we expect to converge to an eigenvector e_1 in the sense that the angles between the iterates and this eigenvector converge to zero, and we would like to know when the angle between $x^{(m)}$ and e_1 has become sufficiently small.

If we know the dominant eigenvalue λ_1, we have

$$Ae_1 = \lambda_1 e_1, \quad 0 = \lambda_1 e_1 - Ae_1,$$

so the norm

$$\|\lambda_1 x^{(m)} - Ax^{(m)}\|$$

should provide an estimate of how close we are to the eigenvector. In most applications, the eigenvalue is not known, so we have to approximate it as well. An elegant solution to this problem is provided by the *Rayleigh quotient*: assume that $e \neq 0$ is an eigenvector of A, but the corresponding eigenvalue λ is unknown. Our goal is to reconstruct λ using e. By our assumption, we have

$$Ae = \lambda e.$$

Taking the inner product with e on both sides yields

$$\langle Ae, e\rangle = \lambda \langle e, e\rangle,$$

and using

$$\langle e, e\rangle = \|e\|^2 > 0,$$

we can obtain

$$\lambda = \frac{\langle Ae, e\rangle}{\langle e, e\rangle}$$

and have found λ. This result gives rise to the following definition:

Definition 4.4 (Rayleigh quotient). Let $A \in \mathbb{F}^{n\times n}$. The mapping

$$\Lambda_A : \mathbb{F}^n \setminus \{0\} \to \mathbb{F}, \quad x \mapsto \frac{\langle Ax, x\rangle}{\langle x, x\rangle},$$

is called the *Rayleigh quotient* corresponding to the matrix A.

We have already encountered another important property of the Rayleigh quotient in Lemma 3.8: for a self-adjoint matrix A, its minimum and maximum are the smallest and largest eigenvalue, respectively.

For an *approximation* of an eigenvector, the Rayleigh quotient yields an approximation of an appropriate eigenvalue. In order to prove the corresponding estimate, we require the following alternate characterization of the sine of the angle between two vectors:

Lemma 4.5 (Sine). *Let $x, y \in \mathbb{F}^n \setminus \{0\}$. Then we have*

$$\sin \angle(x,y) = \min \left\{ \frac{\|x - \alpha y\|}{\|x\|} \;:\; \alpha \in \mathbb{F} \right\}. \tag{4.7}$$

Proof. We let $\alpha_0 := \langle x, y\rangle / \|y\|^2$ and observe

$$\langle x - \alpha_0 y, y\rangle = \langle x, y\rangle - \frac{\langle x, y\rangle}{\|y\|^2}\langle y, y\rangle = \langle x, y\rangle - \langle x, y\rangle = 0,$$

i.e., $x - \alpha_0 y$ and y are perpendicular vectors. For $\alpha \in \mathbb{F}$ and $\beta := \alpha - \alpha_0$, this implies

$$\begin{aligned}\|x - \alpha y\|^2 &= \|x - \alpha_0 y - \beta y\|^2 = \langle (x - \alpha_0 y) - \beta y, (x - \alpha_0 y) - \beta y\rangle \\ &= \|x - \alpha_0 y\|^2 - \langle \beta y, (x - \alpha_0 y)\rangle - \langle (x - \alpha_0 y), \beta y\rangle + |\beta|^2 \|y\|^2 \\ &= \|x - \alpha_0 y\|^2 + |\beta|^2 \|y\|^2,\end{aligned}$$

i.e., the right-hand side attains its minimum for $\alpha = \alpha_0$. Due to

$$\begin{aligned}\|x - \alpha_0 y\|^2 &= \|x\|^2 - \bar{\alpha}_0 \langle x, y\rangle - \alpha_0 \langle y, x\rangle + |\alpha_0|^2 \|y\|^2 \\ &= \|x\|^2 - \frac{|\langle x, y\rangle|^2}{\|y\|^2} - \frac{|\langle x, y\rangle|^2}{\|y\|^2} + \frac{|\langle x, y\rangle|^2}{\|y\|^2} = \|x\|^2 \left(1 - \frac{|\langle x, y\rangle|^2}{\|x\|^2 \|y\|^2}\right) \\ &= \|x\|^2 (1 - \cos^2 \angle(x,y)) = \|x\|^2 \sin^2 \angle(x,y),\end{aligned}$$

this minimum has to be $\sin \angle(x, y)$. □

Using this result, we can give the following elegant bound for the approximation error of the Rayleigh quotient:

Theorem 4.6 (Rayleigh quotient). *Let $A \in \mathbb{F}^{n\times n}$, and let $x \in \mathbb{F}^n \setminus \{0\}$. Let $e \in \mathbb{F}^n \setminus \{0\}$ be an eigenvector of A for the eigenvalue $\lambda \in \mathbb{F}$. Then we have*

$$|\Lambda_A(x) - \lambda| \le \|A - \lambda I\| \sin \angle(x,e) \le \|A - \lambda I\| \tan \angle(x,e).$$

If A is a normal matrix, we even have

$$|\Lambda_A(x) - \lambda| \le \|A - \lambda I\| \sin^2 \angle(x,e) \le \|A - \lambda I\| \tan^2 \angle(x,e).$$

Proof. Since e is an eigenvector, we have $(A - \lambda I)e = 0$, and we can use the Cauchy–Schwarz inequality (2.9) and the compatibility inequality (3.11) of the spectral norm to find

$$\begin{aligned}|\Lambda_A(x) - \lambda| &= \left| \frac{\langle Ax, x\rangle}{\langle x, x\rangle} - \frac{\langle \lambda x, x\rangle}{\langle x, x\rangle} \right| = \frac{|\langle (A - \lambda I)x, x\rangle|}{\langle x, x\rangle} \\ &= \frac{|\langle (A - \lambda I)(x - \alpha e), x\rangle|}{\langle x, x\rangle} \le \frac{\|(A - \lambda I)(x - \alpha e)\| \, \|x\|}{\|x\|^2} \\ &\le \frac{\|A - \lambda I\| \, \|x - \alpha e\| \, \|x\|}{\|x\|^2} = \|A - \lambda I\| \frac{\|x - \alpha e\|}{\|x\|} \quad \text{for all } \alpha \in \mathbb{F}.\end{aligned}$$

```
procedure power_iteration(A, var x);
begin
  x ← x/‖x‖;
  a ← Ax;
  λ ← ⟨a, x⟩;
  while ‖a − λx‖ > ϵ|λ| do begin
    x ← a/‖a‖;
    a ← Ax;
    λ ← ⟨a, x⟩
  end
end
```

Figure 4.2. Practical power iteration.

Due to Lemma 4.5, this yields

$$|\Lambda_A(x) - \lambda| \le \|A - \lambda I\| \sin \angle(x, e),$$

i.e., our first estimate. Let now A be a normal matrix. Due to Proposition 2.36, we now not only have $(A - \lambda I)e = 0$, but also $(A - \lambda I)^* e = 0$, and this allows us to improve the estimate as follows:

$$\begin{aligned} |\Lambda_A(x) - \lambda| &= \frac{|\langle (A - \lambda I)(x - \alpha e), x\rangle|}{\|x\|^2} = \frac{|\langle x - \alpha e, (A - \lambda I)^* x\rangle|}{\|x\|^2} \\ &= \frac{|\langle x - \alpha e, (A - \lambda I)^*(x - \alpha e)\rangle|}{\|x\|^2} = \frac{|\langle (A - \lambda I)(x - \alpha e), x - \alpha e\rangle|}{\|x\|^2} \\ &\le \frac{\|A - \lambda I\| \, \|x - \alpha e\|^2}{\|x\|^2} \quad \text{for all } \alpha \in \mathbb{F}, \end{aligned}$$

and using the optimal α of Lemma 4.5 yields

$$|\Lambda_A(x) - \lambda| \le \|A - \lambda I\| \sin^2 \angle(x, e).$$

Due to the Cauchy–Schwarz inequality (2.9), we have $\cos \angle(x, e) \le 1$ and therefore $\sin \angle(x, e) \le \tan \angle(x, e)$. □

Using the Rayleigh quotient, we can construct a stopping criterion for the power iteration: for the m-th iteration vector $x^{(m)}$, we compute an approximation $\lambda^{(m)} = \Lambda_A(x^{(m)})$ of the eigenvalue and check whether $\|Ax^{(m)} - \lambda^{(m)} x^{(m)}\|$ is sufficiently small. In our algorithm, $x^{(m)}$ is a unit vector, so the Rayleigh quotient takes the simple form $\Lambda_A(x^{(m)}) = \langle Ax^{(m)}, x^{(m)}\rangle$ and we obtain the practical algorithm given in Figure 4.2.

4.3 Residual-based error control

Since the error analysis in this section relies on Propositions 3.11 and 3.13, we require the matrix A to be self-adjoint. Since this is a particularly useful vector, it deserves to be given a name:

Definition 4.7 (Residual). Let $A \in \mathbb{F}^{n\times n}$, and let $x \in \mathbb{F}^n \setminus \{0\}$. The vector

$$r := \Lambda_A(x)x - Ax$$

is called the *residual* of x with respect to A.

The practical power iteration given in Figure 4.2 stops as soon as the norm of the residual drops below $\epsilon|\lambda^{(m)}|$, and we can use *backward error analysis* to derive error bounds for the eigenvalue and the corresponding eigenvector: given $x^{(m)}$ and $\lambda^{(m)} = \Lambda_A(x^{(m)})$, we construct a matrix B such that

$$Bx^{(m)} = \lambda^{(m)}x^{(m)}$$

holds, i.e., such that $x^{(m)}$ is an eigenvector of B for the eigenvalue $\lambda^{(m)}$. If we can bound $\|A - B\|$, the perturbation theory of Propositions 3.11 and 3.13 can be applied. In order to find a convenient bound for $\|A - B\|$, we require the following lemma:

Lemma 4.8 (Rank-2 matrix). *Let $a, b \in \mathbb{F}^n$ with $\langle a, b\rangle = 0$, and let $E := ab^* + ba^*$. Then we have $\|E\| = \|a\|\,\|b\|$.*

Proof. We first consider the special case $\|a\| = 1$, $\|b\| = 1$.

Let $x \in \mathbb{F}^n \setminus \{0\}$. We split x into components in the kernel of E and in the subspaces spanned by a and b:

$$x = y + \alpha a + \beta b, \quad \alpha := \langle x, a\rangle, \quad \beta := \langle x, b\rangle, \quad y := x - \alpha a - \beta b.$$

By assumption, we have $\langle a, b\rangle = 0$, and this implies

$$\begin{aligned}\langle y, a\rangle &= \langle x, a\rangle - \alpha\langle a, a\rangle - \beta\langle b, a\rangle = \langle x, a\rangle - \langle x, a\rangle = 0,\\ \langle y, b\rangle &= \langle x, b\rangle - \alpha\langle a, b\rangle - \beta\langle b, b\rangle = \langle x, b\rangle - \langle x, b\rangle = 0,\\ Ey &= ab^*y + ba^*y = a\langle y, b\rangle + b\langle y, a\rangle = 0,\end{aligned}$$

i.e., the vectors y, a and b are pairwise perpendicular and $y \in \mathcal{N}(E)$. We find

$$\begin{aligned}Ex &= E(y + \alpha a + \beta b) = Ey + \alpha Ea + \beta Eb\\ &= \alpha ab^*a + \alpha ba^*a + \beta ab^*b + \beta ba^*b = \alpha b + \beta a,\end{aligned}$$

and Pythagoras' theorem yields

$$\begin{aligned}\|Ex\|^2 &= |\alpha|^2\|b\|^2 + |\beta|^2\|a\|^2 = |\alpha|^2 + |\beta|^2 \\ &\le \|y\|^2 + |\alpha|^2 + |\beta|^2 = \|y\|^2 + \|\alpha a\|^2 + \|\beta b\|^2 = \|x\|^2,\end{aligned}$$

i.e., $\|E\| \le 1$. Due to

$$Ea = ab^*a + ba^*a = b, \quad \|Ea\| = \|b\| = 1,$$

we also have $\|E\| \ge 1$, and therefore $\|E\| = 1$.

Now we consider the general case. If $a = 0$ or $b = 0$, we have $E = 0$ and $\|E\| = 0 = \|a\|\,\|b\|$. Otherwise we let $\widehat{a} := a/\|a\|$ and $\widehat{b} := b/\|b\|$ and apply the special case to prove $\|\widehat{E}\| = 1$ for $\widehat{E} := \widehat{a}\widehat{b}^* + \widehat{b}\widehat{a}^*$. Observing $E = \|a\|\,\|b\|\,\widehat{E}$ completes the proof. □

Using this result, the backward error analysis for eigenvectors is straightforward:

Theorem 4.9 (Accuracy). *Let $y \in \mathbb{F}^n \setminus \{0\}$. Let $\mu := \Lambda_A(y)$, and let $r := \mu y - Ay$ denote the residual of y. Then there are an eigenvalue $\lambda \in \sigma(A)$ and a vector $x \in \mathcal{E}(A,\lambda)$ in the corresponding eigenspace satisfying*

$$|\lambda - \mu| \le \frac{\|r\|}{\|y\|}, \quad \|x - y\| \le 2\frac{\|r\|}{\gamma_A(\lambda)}.$$

Here $\gamma_A(\lambda)$ is the spectral gap introduced in (3.12).

Proof. We define the self-adjoint matrix

$$B := A + \frac{ry^*}{\|y\|^2} + \frac{yr^*}{\|y\|^2},$$

where we again interpret y and r as elements of $\mathbb{F}^{n\times 1}$ in the obvious way. Due to

$$r^*y = \langle y, r\rangle = \langle y, \mu y - Ay\rangle = \bar{\mu}\langle y, y\rangle - \langle y, Ay\rangle = \overline{\langle Ay, y\rangle} - \overline{\langle Ay, y\rangle} = 0,$$

our definition yields

$$By = Ay + r\frac{y^*y}{\|y\|^2} + y\frac{r^*y}{\|y\|^2} = Ay + r\frac{\|y\|^2}{\|y\|^2} = Ay + \mu y - Ay = \mu y,$$

i.e., y is an eigenvector of B for the eigenvalue μ. In order to be able to apply Propositions 3.11 and 3.13, we have to find a bound for the spectral norm of

$$B - A = \frac{ry^*}{\|y\|^2} + \frac{yr^*}{\|y\|^2} = \frac{1}{\|y\|^2}(ry^* + yr^*).$$

Due to $\langle r, y\rangle = 0$, we can apply Lemma 4.8 and obtain

$$\|A - B\| \le \frac{\|r\|\,\|y\|}{\|y\|^2} = \frac{\|r\|}{\|y\|}.$$

Applying Propositions 3.11 and 3.13 completes the proof. □

The practical power iteration in Figure 4.2 stops as soon as the norm of the residual $r^{(m)}$ defined by

$$r^{(m)} := \lambda^{(m)} x^{(m)} - A x^{(m)}, \quad \lambda^{(m)} := \langle A x^{(m)}, x^{(m)} \rangle \quad \text{for all } m \in \mathbb{N}_0 \tag{4.8}$$

drops below $\epsilon |\lambda^{(m)}|$. In this case, we can obtain the following estimates for the accuracy of the approximations of eigenvalues and eigenvectors:

Corollary 4.10 (Accuracy). *Let $\epsilon \in [0, 1)$ and $\|r^{(m)}\| \leq \epsilon |\lambda^{(m)}|$. Then we can find $\lambda \in \sigma(A)$ and $x \in \mathcal{E}(A, \lambda)$ such that*

$$|\lambda - \lambda^{(m)}| \leq \epsilon |\lambda^{(m)}| \leq \frac{\epsilon}{1-\epsilon} |\lambda|, \tag{4.9a}$$

$$\|x - x^{(m)}\| \leq 2\epsilon \frac{|\lambda^{(m)}|}{\gamma_A(\lambda)} \leq 2\epsilon \left(1 + \frac{\epsilon}{1-\epsilon}\right) \frac{|\lambda|}{\gamma_A(\lambda)}. \tag{4.9b}$$

Proof. By Theorem 4.9, $\|r^{(m)}\| \leq \epsilon |\lambda^{(m)}|$ implies that we can find an eigenvalue $\lambda \in \sigma(A)$ and a vector $x \in \mathcal{E}(A, \lambda)$ such that the bounds

$$|\lambda - \lambda^{(m)}| \leq \epsilon |\lambda^{(m)}|, \quad \|x - x^{(m)}\| \leq 2 \frac{\epsilon |\lambda^{(m)}|}{\gamma_A(\lambda)}$$

hold. In order to obtain an estimate for the relative error of the eigenvalue, we apply the triangle inequality to the first bound to get

$$\begin{aligned} |\lambda - \lambda^{(m)}| &\leq \epsilon |\lambda^{(m)}| \leq \epsilon |\lambda| + \epsilon |\lambda^{(m)} - \lambda|, \\ (1-\epsilon) |\lambda - \lambda^{(m)}| &\leq \epsilon |\lambda|, \\ |\lambda - \lambda^{(m)}| &\leq \frac{\epsilon}{1-\epsilon} |\lambda|, \end{aligned}$$

i.e., the relative error is bounded by $\epsilon/(1-\epsilon)$. Using this result, we can also derive a bound for the eigenvector:

$$\begin{aligned} \|x - x^{(m)}\| &\leq 2\epsilon \frac{|\lambda^{(m)}|}{\gamma_A(\lambda)} \leq 2\epsilon \frac{|\lambda| + |\lambda^{(m)} - \lambda|}{\gamma_A(\lambda)} \leq 2\epsilon \frac{|\lambda| + (\epsilon/(1-\epsilon))|\lambda|}{\gamma_A(\lambda)} \\ &= 2\epsilon \left(1 + \frac{\epsilon}{1-\epsilon}\right) \frac{|\lambda|}{\gamma_A(\lambda)}. \end{aligned}$$

□

We conclude that the stopping criterion guarantees a high relative accuracy of the eigenvalue. If the *relative* spectral gap $\gamma_A(\lambda)/|\lambda|$ is not too small, we can also expect an accurate approximation of an eigenvector.

4.4 Inverse iteration

The power iteration given in Figures 4.1 and 4.2 can only be used to compute the largest eigenvalues (in modulus). We now introduce a modification that allows us to find arbitrary eigenvalues: assume for the moment that A is invertible. Then λ is an eigenvalue of A if and only if $1/\lambda$ is an eigenvalue of A^{-1}, since

$$Ax = \lambda x \iff x = \lambda A^{-1}x \iff \frac{1}{\lambda}x = A^{-1}x \quad \text{for all } \lambda \in \mathbb{F} \setminus \{0\},\ x \in \mathbb{F}^n. \tag{4.10}$$

This means that applying the power iteration to A^{-1} instead of A should yield a sequence of vectors converging to the eigenspace corresponding to the *smallest* eigenvalue in modulus if

$$|\lambda_1| < |\lambda_2| \leq \cdots \leq |\lambda_n|,$$

and Theorem 4.2 yields a rate of $|\lambda_1|/|\lambda_2|$ for the convergence. The resulting algorithm is called the *inverse iteration*. It appears to have been developed first for applications in the field of structural mechanics [35].

We can refine the idea to obtain a method that can find eigenvectors for any eigenvalue, not only for the largest and smallest ones: we introduce a *shift parameter* $\mu \in \mathbb{F}$ and consider the matrix $A - \mu I$. Its eigenvalues are "shifted": we have

$$Ax = \lambda x \quad \iff \quad (A - \mu I)x = (\lambda - \mu)x \quad \text{for all } \lambda \in \mathbb{F},\ x \in \mathbb{F}^n,$$

and (4.10) yields

$$Ax = \lambda x \quad \iff \quad \frac{1}{\lambda - \mu}x = (A - \mu I)^{-1}x \quad \text{for all } \lambda \in \mathbb{F} \setminus \{\mu\},\ x \in \mathbb{F}^d \tag{4.11}$$

if $A - \mu I$ is invertible. In this case, we can apply the power iteration to the matrix $(A - \mu I)^{-1}$ and arrive at the *shifted inverse iteration*:

$$x^{(m)} := (A - \mu I)^{-1}x^{(m-1)} \quad \text{for all } m \in \mathbb{N}. \tag{4.12}$$

The convergence results of Theorem 4.2 carry over directly:

Corollary 4.11 (Convergence). *Let $\mu \in \mathbb{F}$, and let $A - \mu I$ be invertible. Let* (4.2) *and* (4.3) *hold with*

$$|\lambda_1 - \mu| < |\lambda_2 - \mu| \leq \cdots \leq |\lambda_n - \mu|. \tag{4.13}$$

Let $e_1 := Q\delta_1$ and $\cos \angle(x^{(0)}, e_1) \neq 0$, and let $(x^{(m)})_{m=0}^{\infty}$ be given by (4.12). Then we have

$$\tan \angle(x^{(m+1)}, e_1) \leq \frac{|\lambda_1 - \mu|}{|\lambda_2 - \mu|} \tan \angle(x^{(m)}, e_1) \quad \textit{for all } m \in \mathbb{N}_0,$$

and due to $|\lambda_1 - \mu| < |\lambda_2 - \mu|$, this means that the angle between the iteration vectors and the eigenvector e_1 converges to zero.

Proof. Let $\widehat{A} := (A - \mu I)^{-1}$ and

$$\hat{\lambda}_j := \frac{1}{\lambda_j - \mu} \quad \text{for all } j \in \{1, \ldots, n\}.$$

The equations (4.2) and (4.3) yield

$$Q^*(A - \mu I)Q = Q^*AQ - \mu Q^*Q = D - \mu I = \begin{pmatrix} \lambda_1 - \mu & & \\ & \ddots & \\ & & \lambda_n - \mu \end{pmatrix},$$

$$Q^*\widehat{A}Q = Q^*(A - \mu I)^{-1}Q = (D - \mu I)^{-1} = \begin{pmatrix} \hat{\lambda}_1 & & \\ & \ddots & \\ & & \hat{\lambda}_n \end{pmatrix}.$$

Using (4.13), we obtain

$$|\hat{\lambda}_1| > |\hat{\lambda}_2| \geq \cdots \geq |\hat{\lambda}_n|,$$

so we can apply Theorem 4.2 to $\widehat{A}$ in order to find

$$\begin{aligned} \tan \angle(x^{(m+1)}, e_1) &\leq \frac{|\hat{\lambda}_2|}{|\hat{\lambda}_1|} \tan \angle(x^{(m)}, e_1) \\ &= \frac{|\lambda_1 - \mu|}{|\lambda_2 - \mu|} \tan \angle(x^{(m)}, e_1) \quad \text{for all } m \in \mathbb{N}_0. \end{aligned}$$

□

If we can find a suitable shift parameter μ, the shifted inverse iteration converges rapidly and reliably: as a first example, we consider the matrix

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

Its eigenvalues are 1 and -1, and due to $|1| = |-1|$, we can expect no convergence for the power iteration. If we use a shifted inverse iteration with, e.g., $\mu = 1/2$, Corollary 4.11 predicts a convergence rate of

$$\frac{|1 - 1/2|}{|-1 - 1/2|} = \frac{|1/2|}{|-3/2|} = 1/3.$$

By choosing μ closer to 1, we can get even better rates.

The shift parameter can also help if the gap between eigenvectors is small: let $\epsilon > 0$ and consider

$$A = \begin{pmatrix} 1 + \epsilon & \\ & 1 \end{pmatrix}.$$

```
procedure inverse_iteration(A, μ, var x);
begin
  Prepare a solver for A − μI;
  x ← x/‖x‖;
  Solve (A − μI)a = x;
  λ̂ ← ⟨a, x⟩;   λ ← 1/λ̂ + μ;
  while ‖a − λ̂x‖ > ϵ|λ̂| do begin
    x ← a/‖a‖;
    Solve (A − μI)a = x;
    λ̂ ← ⟨a, x⟩;   λ ← 1/λ̂ + μ
  end
end
```

Figure 4.3. Inverse iteration.

The matrix A now has a dominant eigenvalue $1+\epsilon$ and a second eigenvalue 1, and Theorem 4.2 predicts a convergence rate of $1/(1+\epsilon)$. Using a shifted inverse iteration with $\mu = 1 + 2\epsilon/3$, we get a rate of

$$\frac{|1+\epsilon-(1+2\epsilon/3)|}{|1-(1+2\epsilon/3)|} = \frac{|\epsilon/3|}{|-2\epsilon/3|} = 1/2,$$

and again the rate can be improved by moving μ close to $1+\epsilon$. In both examples, we can also compute the second eigenvalue by choosing μ appropriately.

In a practical implementation of the inverse iteration, it is frequently more efficient to handle the inverse $(A-\mu I)^{-1}$ implicitly: instead of computing $a = (A-\mu I)^{-1}x$, which would require the inverse, we solve the linear system $(A-\mu I)a = x$ for the unknown vector a. Since each step of the iteration requires us to solve one of these systems for another right-hand side vector, it is usually a good idea to prepare the matrix $A-\mu I$ in advance, e.g., by computing a suitable factorization. The resulting algorithm is summarized in Figure 4.3.

The stopping criterion relies on the residual given by

$$\begin{aligned}\hat{\lambda}^{(m)} &:= \langle (A-\mu I)^{-1}x^{(m)}, x^{(m)}\rangle,\\ \widehat{r}^{(m)} &:= \hat{\lambda}^{(m)}x^{(m)} - (A-\mu I)^{-1}x^{(m)} \quad \text{for all } m \in \mathbb{N}_0\end{aligned}$$

instead of (4.8), since this vector can be computed without additional matrix-vector multiplications.

Corollary 4.12 (Accuracy). *Let $\epsilon \in [0,1)$ and $\|\widehat{r}^{(m)}\| \leq \epsilon|\hat{\lambda}^{(m)}|$. Then we can find an eigenvalue $\lambda \in \sigma(A)$ and $x \in \mathcal{E}(A,\lambda)$ in the corresponding eigenspace such that*

$$|\lambda - \lambda^{(m)}| \leq \epsilon|\lambda - \mu|,$$
$$\|x - x^{(m)}\| \leq 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\left(1 + \frac{|\lambda-\mu|}{\gamma_A(\lambda)}\right).$$

Proof. We apply Corollary 4.10 to $\widehat{A} := (A-\mu I)^{-1}$ in order to find $\hat{\lambda} \in \sigma(\widehat{A})$ and $x \in \mathcal{E}(\widehat{A},\hat{\lambda})$ such that

$$|\hat{\lambda} - \hat{\lambda}^{(m)}| \leq \epsilon|\hat{\lambda}^{(m)}|, \quad \|x - x^{(m)}\| \leq 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\frac{|\hat{\lambda}|}{\gamma_{\widehat{A}}(\hat{\lambda})}. \tag{4.14}$$

Due to (4.11), we can find $\lambda \in \sigma(A)$ satisfying $\hat{\lambda} = 1/(\lambda-\mu)$ and obtain

$$\begin{aligned}
|\hat{\lambda} - \hat{\lambda}^{(m)}| &= \left|\frac{1}{\lambda-\mu} - \frac{1}{\lambda^{(m)}-\mu}\right| = \frac{|(\lambda^{(m)}-\mu)-(\lambda-\mu)|}{|\lambda^{(m)}-\mu|\,|\lambda-\mu|} \\
&= \frac{|\lambda-\lambda^{(m)}|}{|\lambda^{(m)}-\mu|\,|\lambda-\mu|}, \\
|\hat{\lambda}^{(m)}| &= \left|\frac{1}{\lambda^{(m)}-\mu}\right| = \frac{|\lambda-\mu|}{|\lambda^{(m)}-\mu|\,|\lambda-\mu|}.
\end{aligned}$$

This already implies the first estimate:

$$|\lambda-\lambda^{(m)}| = |\hat{\lambda}-\hat{\lambda}^{(m)}|\,|\lambda^{(m)}-\mu|\,|\lambda-\mu| \leq \epsilon|\hat{\lambda}^{(m)}|\,|\lambda^{(m)}-\mu|\,|\lambda-\mu| = \epsilon|\lambda-\mu|.$$

An estimate for the eigenvector can be obtained by investigating the spectral gap for $\widehat{A}$: by definition, we find an eigenvalue $\hat{\lambda}' \in \sigma(\widehat{A}) \setminus \{\hat{\lambda}\}$ such that

$$\gamma_{\widehat{A}}(\hat{\lambda}) = |\hat{\lambda} - \hat{\lambda}'|.$$

Using (4.11) again, we find an eigenvalue $\lambda' \in \sigma(A) \setminus \{\lambda\}$ such that $\hat{\lambda}' = 1/(\lambda'-\mu)$. This means

$$\gamma_{\widehat{A}}(\hat{\lambda}) = |\hat{\lambda}-\hat{\lambda}'| = \left|\frac{1}{\lambda-\mu} - \frac{1}{\lambda'-\mu}\right| = \frac{|(\lambda'-\mu)-(\lambda-\mu)|}{|\lambda-\mu|\,|\lambda'-\mu|} = \frac{|\lambda'-\lambda|}{|\lambda-\mu|\,|\lambda'-\mu|},$$

and (4.14) yields

$$\begin{aligned}
\|x - x^{(m)}\| &\leq 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\frac{|\hat{\lambda}|}{\gamma_{\widehat{A}}(\hat{\lambda})} = 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\frac{|\lambda-\mu|\,|\lambda'-\mu|}{|\lambda-\mu|\,|\lambda'-\lambda|} \\
&= 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\frac{|\lambda'-\mu|}{|\lambda'-\lambda|} \leq 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\frac{|\lambda'-\lambda| + |\lambda-\mu|}{|\lambda'-\lambda|} \\
&\leq 2\epsilon\left(1 + \frac{\epsilon}{1-\epsilon}\right)\left(1 + \frac{|\lambda-\mu|}{\gamma_A(\lambda)}\right).
\end{aligned}$$

□

4.5 Rayleigh iteration

According to Corollary 4.11, the shifted inverse iteration converges rapidly if the shift parameter μ is close to an eigenvalue. Since Theorem 4.6 states that the Rayleigh quotient yields an approximation of an eigenvalue, using the Rayleigh quotient as shift should lead to a fast method. The resulting algorithm is called the *Rayleigh iteration*, and its iteration vectors are given by

$$\lambda^{(m)} := \frac{\langle Ax, x\rangle}{\langle x, x\rangle}, \quad x^{(m+1)} := (A - \lambda^{(m)} I)^{-1} x^{(m)} \quad \text{for all } m \in \mathbb{N}_0. \tag{4.15}$$

In practical implementations, we have to ensure that the entries of the iteration vectors do not become too large or too small by dividing $x^{(m+1)}$ by its norm. As in previous algorithms, this also allows us to avoid dividing by the norm when computing the Rayleigh quotient. Using the residual-based stopping criterion already employed for the power iteration, we obtain the algorithm given in Figure 4.4.

Compared to the shifted inverse iteration, the Rayleigh iteration typically requires more operations per step: the shift changes in each step, therefore we have to solve linear systems with *different* matrices that cannot be prepared in advance. On the other hand, the Rayleigh iteration can be significantly faster:

Proposition 4.13 (Convergence). *Let $A - \lambda^{(0)} I$ be invertible, let* (4.2) *and* (4.3) *hold with*

$$|\lambda_1 - \lambda^{(0)}| < |\lambda_2 - \lambda^{(0)}| \leq \cdots \leq |\lambda_n - \lambda^{(0)}|. \tag{4.16}$$

Let $e_1 := Q\delta_1$ and let $x^{(1)}$ be given by (4.15). *If there is a constant $c \in \mathbb{R}_{>0}$ such that*

$$\tan \angle(x^{(0)}, e_1) \leq c \leq \sqrt{\frac{|\lambda_1 - \lambda_2|}{2\|A - \lambda_1 I\|}}$$

holds, we have $\tan \angle(x^{(1)}, e_1) \leq c$ and

$$\tan \angle(x^{(1)}, e_1) \leq 2\frac{\|A - \lambda_1 I\|}{|\lambda_1 - \lambda_2|} \tan^3 \angle(x^{(0)}, e_1),$$

i.e., for a sufficiently good initial vector, the Rayleigh iteration is cubically convergent.

Proof. Since $\lambda^{(0)} = \Lambda_A(x^{(0)})$ by definition, Theorem 4.6 yields

$$|\lambda^{(0)} - \lambda_1| \leq \|A - \lambda_1 I\| \tan^2 \angle(x^{(0)}, e_1),$$

since we have assumed A to be normal (or even self-adjoint), and we also find

$$\begin{aligned} |\lambda^{(0)} - \lambda_2| &\geq |\lambda_1 - \lambda_2| - |\lambda^{(0)} - \lambda_1| \geq |\lambda_1 - \lambda_2| - \|A - \lambda_1 I\| \tan^2 \angle(x^{(0)}, e_1) \\ &\geq |\lambda_1 - \lambda_2| - \|A - \lambda_1 I\| c^2 \geq |\lambda_1 - \lambda_2| - |\lambda_1 - \lambda_2|/2 = |\lambda_1 - \lambda_2|/2. \end{aligned}$$

```
procedure rayleigh_iteration(A, var λ, x);
begin
  Solve (A − λI)y = x;
  x ← y/‖y‖;
  a ← Ax;
  λ ← ⟨a, x⟩;
  while ‖a − λx‖ > ϵ|λ| do begin
    Solve (A − λI)y = x;
    x ← y/‖y‖;
    a ← Ax;
    λ ← ⟨a, x⟩
  end
end
```

Figure 4.4. Rayleigh iteration.

Combining these estimates with Corollary 4.11 for $\mu = \lambda^{(0)}$ yields

$$\begin{aligned}
\tan \angle(x^{(1)}, e_1) &\le \frac{|\lambda_1 - \lambda^{(0)}|}{|\lambda_2 - \lambda^{(0)}|} \tan \angle(x^{(0)}, e_1) \\
&\le \frac{\|A - \lambda_1 I\| \tan^2 \angle(x^{(0)}, e_1)}{|\lambda_1 - \lambda_2|/2} \tan \angle(x^{(0)}, e_1) \\
&= 2\frac{\|A - \lambda_1 I\|}{|\lambda_1 - \lambda_2|} \tan^3 \angle(x^{(0)}, e_1) \\
&\le 2\frac{\|A - \lambda_1 I\|}{|\lambda_1 - \lambda_2|} \frac{|\lambda_1 - \lambda_2|}{2\|A - \lambda_1 I\|} \tan \angle(x^{(0)}, e_1) = \tan \angle(x^{(0)}, e_1) \le c.
\end{aligned}$$

These are the required estimates. □

Example 4.14 (One-dimensional model problem). The advantages of the Rayleigh iteration can be illustrated using the model problem (1.2). We let $c = 1$ and $n = 1000$, use the starting vector given by $x_j^{(0)} = j/n$, and compute the *second* eigenvalue by a shifted inverse iteration using the shift parameter $\mu = 30$ and a Rayleigh iteration starting with this initial shift. We have

$$\lambda_1 = 4h^{-2} \sin^2(\pi 2h/2) \approx 39.478, \quad \lambda_2 = 4h^{-2} \sin^2(\pi h/2) \approx 9.8696$$

and obtain the following results:

	Inverse it.		Rayleigh	
m	$\|r^{(m)}\|$	Ratio	$\|r^{(m)}\|$	Ratio
1	$1.72 \times 10^{+1}$		$1.72 \times 10^{+1}$	
2	$1.10 \times 10^{+1}$	0.64	$1.44 \times 10^{+1}$	0.84
3	$5.93 \times 10^{+0}$	0.54	$1.13 \times 10^{+1}$	0.78
4	$2.89 \times 10^{+0}$	0.49	$2.91 \times 10^{+0}$	0.26
5	$1.37 \times 10^{+0}$	0.47	2.89×10^{-2}	9.9×10^{-3}
6	6.45×10^{-1}	0.47	2.73×10^{-8}	9.5×10^{-7}
7	3.04×10^{-1}	0.47	conv.	
8	1.43×10^{-1}	0.47		

As we can see, the inverse iteration converges at a rate of approximately 0.47, which coincides with the expected value of $|\lambda_1 - \mu|/|\lambda_2 - \mu|$. The Rayleigh iteration shows inferior convergence rates during the first few steps, but once cubic convergence takes hold in the fourth step, machine accuracy is attained rapidly.

As with most non-linear iterations, the performance of the Rayleigh iteration depends crucially on the choice of the starting vector. If an unsuitable vector is chosen, the method may converge to an eigenspace corresponding to another eigenvalue, or it may fail to converge at all.

4.6 Convergence to invariant subspace

The convergence analysis presented so far relies on the dominance of one simple eigenvalue, the case of multiple eigenvalues with identical modulus has not been addressed. Since the investigation of the inverse iteration and the Rayleigh iteration relies on the convergence result provided by Theorem 4.2 for the power iteration, we focus on this simple case.

Instead of assuming that one eigenvalue is dominant, we assume that there is a $k \in \{1, \ldots, n-1\}$ such that

$$|\lambda_1| \geq \cdots \geq |\lambda_k| > |\lambda_{k+1}| \geq \cdots \geq |\lambda_n| \tag{4.17}$$

holds, i.e., we allow k eigenvalues to have the same norm. In this case, we cannot expect convergence to the eigenspace corresponding to λ_1, but we can still prove convergence to an *invariant subspace*.

Definition 4.15 (Angle between a vector and a subspace). Let $x \in \mathbb{F}^n \setminus \{0\}$, and let $\mathcal{Y} \subseteq \mathbb{F}^n$ be a non-trivial subspace. We define

$$\begin{aligned}\cos \angle(x, \mathcal{Y}) &:= \max\{\cos \angle(x, y) \ : \ y \in \mathcal{Y} \setminus \{0\}\},\\ \sin \angle(x, \mathcal{Y}) &:= \min\{\sin \angle(x, y) \ : \ y \in \mathcal{Y} \setminus \{0\}\},\\ \tan \angle(x, \mathcal{Y}) &:= \frac{\sin \angle(x, \mathcal{Y})}{\cos \angle(x, \mathcal{Y})}.\end{aligned}$$

If $\cos \angle(x, \mathcal{Y}) = 0$ holds, we let $\tan \angle(x, \mathcal{Y}) = \infty$.

In order to compute the angle between a vector and a subspace, the minimization property of Lemma 4.5 is useful: we have

$$\begin{aligned}\sin \angle(x, \mathcal{Y}) &= \min\{\sin \angle(x, y) \ : \ y \in \mathcal{Y}\}\\ &= \min\left\{\min\left\{\frac{\|x - \alpha y\|}{\|x\|} \ : \ \alpha \in \mathbb{F}\right\} \ : \ y \in \mathcal{Y}\right\}\\ &= \min\left\{\frac{\|x - y\|}{\|x\|} \ : \ y \in \mathcal{Y}\right\} \quad \text{for all } x \in \mathbb{F}^n \setminus \{0\},\end{aligned}$$

since $\mathcal{Y}$ is a subspace and therefore invariant under scaling. If we can find the best approximation of x in $\mathcal{Y}$, we can compute the angle. Such best approximations can be conveniently described by appropriate projection operators:

Definition 4.16 (Orthogonal projection). Let $P \in \mathbb{F}^{n\times n}$. If $P^2 = P$ holds, P is called a *projection*. If P is a selfadjoint projection, i.e., if $P^2 = P = P^*$ holds, we call it an *orthogonal projection*. If $\mathcal{Y} \subseteq \mathbb{F}^n$ is a subspace such that $\mathcal{Y} = \mathcal{R}(P)$, we call P a *projection onto* $\mathcal{Y}$.

Exercise 4.17 (Factorized projection). Let $m \in \mathbb{N}$, and let $V \in \mathbb{F}^{n\times m}$ be an isometric matrix (cf. Definition 2.37). Prove that $P := VV^*$ is an orthogonal projection onto $\mathcal{R}(V)$.

Proposition 4.18 (Orthogonal projection). *Let $\mathcal{Y} \subseteq \mathbb{F}^n$ be a subspace, and let $P \in \mathbb{F}^{n\times n}$. P is an orthogonal projection onto $\mathcal{Y}$ if and only if $\mathcal{R}(P) \subseteq \mathcal{Y}$ and*

$$\langle x, y\rangle = \langle Px, y\rangle \quad \text{for all } x \in \mathbb{F}^n,\ y \in \mathcal{Y}. \tag{4.18}$$

In this case, we have

$$\|x - Px\| \le \|x - y\| \quad \text{for all } y \in \mathcal{Y}, \tag{4.19a}$$

$$\|x - Px\|^2 = \|x\|^2 - \|Px\|^2 \tag{4.19b}$$

i.e., $Px \in \mathcal{Y}$ is the best approximation of x in $\mathcal{Y}$.

Proof. Let P be an orthogonal projection onto $\mathcal{Y}$. Let $x \in \mathbb{F}^n$ and $y \in \mathcal{Y}$. By definition, we have $y \in \mathcal{R}(P)$, i.e., we can find $z \in \mathbb{F}^n$ such that $y = Pz$. This implies

$$\langle x, y\rangle = \langle x, Pz\rangle = \langle x, P^2 z\rangle = \langle P^* x, Pz\rangle = \langle Px, Pz\rangle = \langle Px, y\rangle.$$

Let now $P \in \mathbb{F}^{n\times n}$ be a matrix satisfying (4.18). We first consider $y \in \mathcal{Y}$ and find

$$\begin{aligned}\|y - Py\|^2 &= \langle y - Py, y - Py\rangle = \langle y, y\rangle - \langle Py, y\rangle - \langle y, Py\rangle + \langle Py, Py\rangle \\ &= (\langle y, y\rangle - \langle Py, y\rangle) - (\langle y, Py\rangle - \langle Py, Py\rangle) = 0\end{aligned}$$

due to $Py \in \mathcal{R}(P) \subseteq \mathcal{Y}$. This implies $Py = y$ for all $\mathcal{Y}$ and therefore $\mathcal{R}(P) = \mathcal{Y}$.

Let $x \in \mathbb{F}^n$. Due to the previous equation, we have $y := Px \in \mathcal{R}(P) = \mathcal{Y}$ and $P^2 x = Py = y = Px$, and therefore $P^2 = P$.

Let $x, y \in \mathbb{F}^n$. Due to $Px, Py \in \mathcal{R}(P) = \mathcal{Y}$, we find

$$\langle x, Py\rangle = \langle Px, Py\rangle = \overline{\langle Py, Px\rangle} = \overline{\langle y, Px\rangle} = \langle Px, y\rangle,$$

and this implies $P = P^*$ by (2.11). We have proven that P is an orthogonal projection.

Let now $x \in \mathbb{F}^n$ and $y \in \mathcal{Y}$. We have $z := Px - y \in \mathcal{Y}$, and (4.18) yields

$$\begin{aligned}\|x - y\|^2 &= \|x - Px + Px - y\|^2 = \langle (x - Px) + z, (x - Px) + z\rangle \\ &= \langle x - Px, x - Px\rangle + \langle x - Px, z\rangle + \langle z, x - Px\rangle + \langle z, z\rangle \\ &= \|x - Px\|^2 + \langle x - Px, z\rangle + \overline{\langle x - Px, z\rangle} + \|z\|^2 \\ &= \|x - Px\|^2 + \|z\|^2,\end{aligned}$$

and this implies (4.19a) due to $\|z\|^2 \geq 0$. Applying the equation to the special choice $y := 0$ yields (4.19b). □

If we have an orthogonal projection at our disposal, we can find explicit equations for the angle between a vector and the range of the projection.

Proposition 4.19 (Angle characterized by projection). *Let $x \in \mathbb{F}^n \setminus \{0\}$, let $\mathcal{Y} \subseteq \mathbb{F}^n$ be a non-trivial subspace, and let $P \in \mathbb{F}^{n\times n}$ be an orthogonal projection onto this subspace. Then we have (assuming $Px \neq 0$ for the third equation)*

$$\cos\angle(x, \mathcal{Y}) = \frac{\|Px\|}{\|x\|}, \quad \sin\angle(x, \mathcal{Y}) = \frac{\|x - Px\|}{\|x\|}, \quad \tan\angle(x, \mathcal{Y}) = \frac{\|x - Px\|}{\|Px\|}.$$

Proof. According to Definitions 4.1 and 4.15 and Lemma 4.5, we have

$$\sin\angle(x, \mathcal{Y}) = \min\{\sin\angle(x, y) \ : \ y \in \mathcal{Y} \setminus \{0\}\} = \min\left\{\frac{\|x - y\|}{\|x\|} \ : \ y \in \mathcal{Y} \setminus \{0\}\right\}.$$

Since P is an orthogonal projection onto $\mathcal{Y}$, Proposition 4.18 can be applied, and (4.19a) yields

$$\sin \angle(x, \mathcal{Y}) = \min \left\{ \frac{\|x - y\|}{\|x\|} \ : \ y \in \mathcal{Y} \setminus \{0\} \right\} = \frac{\|x - Px\|}{\|x\|}.$$

Using $\sin^2 \angle(x, y) + \cos^2 \angle(x, y) = 1$ and (4.19b), we obtain

$$\begin{aligned} \cos^2 \angle(x, \mathcal{Y}) &= 1 - \sin^2 \angle(x, \mathcal{Y}) = 1 - \frac{\|x - Px\|^2}{\|x\|^2} = \frac{\|x\|^2 - \|x - Px\|^2}{\|x\|^2} \\ &= \frac{\|x\|^2 - (\|x\|^2 - \|Px\|^2)}{\|x\|^2} = \frac{\|Px\|^2}{\|x\|^2}, \end{aligned}$$

and taking the square root gives us the desired equation for the cosine. Applying the definition of the tangent completes the proof. □

Theorem 4.20 (Convergence). *Let* (4.2) *and* (4.3) *hold with*

$$|\lambda_1| \geq \cdots \geq |\lambda_k| > |\lambda_{k+1}| \geq \cdots \geq |\lambda_n|. \tag{4.20}$$

Let $\mathcal{E}_k := \operatorname{span}\{Q\delta_j \ : \ j \in \{1, \ldots, k\}\}$ *denote the invariant subspace spanned by the first* k *columns of* Q*, and let* $\cos \angle(x^{(0)}, \mathcal{E}_k) \neq 0$*. Let* $(x^{(m)})_{m=0}^{\infty}$ *be given by* (4.1)*. Then we have*

$$\tan \angle(x^{(m+1)}, \mathcal{E}_k) \leq \frac{|\lambda_{k+1}|}{|\lambda_k|} \tan \angle(x^{(m)}, \mathcal{E}_k) \quad \text{for all } m \in \mathbb{N}_0,$$

i.e., the angle between the iteration vectors and the subspace $\mathcal{E}_k$ *converges to zero.*

Proof. We define the matrix $\widehat{P} \in \mathbb{F}^{n \times n}$ by

$$\widehat{P} = \begin{pmatrix} I_k & \\ & 0 \end{pmatrix},$$

where $I_k \in \mathbb{F}^{k \times k}$ denotes the k-dimensional identity matrix, and let

$$P := Q \widehat{P} Q^*. \tag{4.21}$$

Due to

$$\begin{aligned} P^2 &= Q \widehat{P} Q^* Q \widehat{P} Q^* = Q \widehat{P} \widehat{P} Q^* = Q \widehat{P} Q^* = P, \\ P^* &= (Q \widehat{P} Q^*)^* = Q \widehat{P}^* Q^* = Q \widehat{P} Q^* = P, \end{aligned}$$

the matrix P is an orthogonal projection, and $\mathcal{R}(\widehat{P}) = \widehat{\mathcal{E}}_k := \operatorname{span}\{\delta_j \ : \ j \in \{1, \ldots, k\}\}$ implies $\mathcal{R}(P) = \mathcal{E}_k$. We conclude that P is an orthogonal projection into $\mathcal{E}_k$, so we can use Proposition 4.19 to handle the angle.

We can proceed as in the proof of Theorem 4.2: we let

$$\widehat{x}^{(m)} := Q^* x^{(m)} \quad \text{for all } m \in \mathbb{N}_0.$$

Due to $\cos \angle(x^{(0)}, \mathcal{E}_k) \neq 0$, we have $Px^{(0)} \neq 0$ and therefore $\widehat{P}\widehat{x}^{(0)} \neq 0$ and observe

$$\begin{aligned}\tan \angle(x^{(m)}, \mathcal{E}_k) &= \frac{\|x^{(m)} - Px^{(m)}\|}{\|Px^{(m)}\|} = \frac{\|Q\widehat{x}^{(m)} - Q\widehat{P}Q^*Q\widehat{x}^{(m)}\|}{\|Q\widehat{P}Q^*Q\widehat{x}^{(m)}\|} \\ &= \frac{\|\widehat{x}^{(m)} - \widehat{P}\widehat{x}^{(m)}\|}{\|\widehat{P}\widehat{x}^{(m)}\|} = \tan \angle(\widehat{x}^{(m)}, \widehat{\mathcal{E}}_k) \quad \text{for all } m \in \mathbb{N}_0 \quad (4.22)\end{aligned}$$

due to (2.16). The tangent of the transformed vector and subspace can be computed directly: Proposition 4.19 combined with the definition of the norm yields

$$\tan^2 \angle(\widehat{y}, \widehat{\mathcal{E}}_k) = \frac{\|\widehat{y} - \widehat{P}\widehat{y}\|^2}{\|\widehat{P}\widehat{y}\|^2} = \frac{\sum_{j=k+1}^n |\widehat{y}_j|^2}{\sum_{j=1}^k |\widehat{y}_j|^2} \quad \text{for all } \widehat{y} \in \mathbb{F}^n \text{ with } \widehat{P}\widehat{y} \neq 0,$$

and this results in

$$\begin{aligned}\tan^2 \angle(\widehat{x}^{(m+1)}, \widehat{\mathcal{E}}_k) = \tan^2 \angle(D\widehat{x}^{(m)}, \widehat{\mathcal{E}}_k) &= \frac{\sum_{j=k+1}^n |\lambda_j|^2 |\widehat{x}_j^{(m)}|^2}{\sum_{j=1}^k |\lambda_j|^2 |\widehat{x}_j^{(m)}|^2} \\ &\leq \frac{|\lambda_{k+1}|^2 \sum_{j=k+1}^n |\widehat{x}_j^{(m)}|^2}{|\lambda_k|^2 \sum_{j=1}^k |\widehat{x}_j^{(m)}|^2} = \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^2 \frac{\sum_{j=k+1}^n |\widehat{x}_j^{(m)}|^2}{\sum_{j=1}^k |\widehat{x}_j^{(m)}|^2} \\ &= \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^2 \tan^2 \angle(\widehat{x}^{(m)}, \widehat{\mathcal{E}}_k) \quad \text{for all } m \in \mathbb{N}_0.\end{aligned}$$

Using (4.22) gives us the final result. □

If λ_1 is a multiple eigenvalue, i.e., if $\lambda_1 = \lambda_2 = \cdots = \lambda_k$ holds, we have $\mathcal{E}_k = \mathcal{E}(A, \lambda_1)$, so the power iteration will compute an eigenvector for λ_1.

4.7 Simultaneous iteration

The result of Theorem 4.20 is not immediately useful, since it yields only convergence of the iteration vectors to the k-dimensional invariant subspace. In order to describe the entire subspace, we need a *basis*, i.e., k linearly independent vectors. A simple approach would be to pick k linearly independent starting vectors $x_1^{(0)}, \ldots, x_k^{(0)} \in \mathbb{F}^n$ and compute the corresponding iteration vectors

$$x_j^{(m+1)} := Ax_j^{(m)} \quad \text{for all } m \in \mathbb{N}_0,\ j \in \{1, \ldots, k\}$$

in parallel. We avoid the additional index by introducing matrices $(X^{(m)})_{m=0}^{\infty}$ in $\mathbb{F}^{n\times k}$ such that $x_j^{(m)}$ is the j-th column of $X^{(m)}$, i.e., such that

$$x_j^{(m)} = X^{(m)}\delta_j \quad \text{for all } m \in \mathbb{N}_0,\ j \in \{1,\dots,k\},$$

similar to the approach used in the proof of Proposition 2.19. The resulting iteration takes the form

$$X^{(m+1)} = AX^{(m)} \quad \text{for all } m \in \mathbb{N}_0. \tag{4.23}$$

According to Theorem 4.20, the columns of the matrices $X^{(m)}$ will converge to the invariant subspace $\mathcal{E}_k$, but they may stop being linearly independent, e.g., if one of the columns lies in the nullspace of A.

Since we are interested in finding a basis, we have to avoid this problem: we impose suitable starting conditions.

Proposition 4.21 (Linear independence). *Let A, Q, D, $\lambda_1,\dots,\lambda_n$ be as in Theorem* 4.20*, and let $P \in \mathbb{F}^{n\times n}$ be the projection onto $\mathcal{E}_k$ defined by* (4.21)*.*

If $PX^{(0)}$ is injective, i.e., if its columns are linearly independent, the same holds for all $PX^{(m)}$ with $m \in \mathbb{N}_0$. In particular, the columns of $X^{(m)}$ are also linearly independent, i.e., they span a k-dimensional subspace.

Proof. Let $PX^{(0)}$ be injective. Following the approach of Theorem 4.2, we define transformed matrices

$$\widehat{X}^{(m)} := Q^*X^{(m)} \quad \text{for all } m \in \mathbb{N}_0.$$

Since $\widehat{P} = Q^*PQ$ projects into the subspace $\mathbb{F}^k \times \{0\}$ of $\mathbb{F}^n$, it makes sense to consider the restriction of our matrices to the first k rows and columns: we define

$$\widehat{D} := D|_{k\times k} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix},$$

$$\widehat{Y}^{(m)} := \widehat{X}^{(m)}|_{k\times k}, \quad \widehat{P}\widehat{X}^{(m)} = \begin{pmatrix} \widehat{Y}^{(m)} \\ 0 \end{pmatrix} \quad \text{for all } m \in \mathbb{N}_0.$$

Due to (4.17), we have $|\lambda_1| \geq \cdots \geq |\lambda_k| > |\lambda_{k+1}| \geq 0$, so $\widehat{D}$ is invertible. Since $PX^{(0)}$ is injective, the matrix

$$\begin{pmatrix} \widehat{Y}^{(0)} \\ 0 \end{pmatrix} = \widehat{P}\widehat{X}^{(0)} = Q^*PQQ^*X^{(0)} = Q^*PX^{(0)}$$

is also injective, so $\widehat{Y}^{(0)}$ has to be injective.

Let $m \in \mathbb{N}_0$. Since $\widehat{P}$ and D are diagonal matrices, we find

$$\begin{aligned} PX^{(m)} &= Q\widehat{P}Q^*X^{(m)} = Q\widehat{P}\widehat{X}^{(m)} = Q\widehat{P}D^m\widehat{X}^{(0)} \\ &= QD^m\widehat{P}\widehat{X}^{(0)} = QD^m\begin{pmatrix}\widehat{Y}^{(0)}\\0\end{pmatrix} = Q\begin{pmatrix}\widehat{D}^m\widehat{Y}^{(0)}\\0\end{pmatrix}. \end{aligned}$$

We have already seen that $\widehat{D}$ is invertible, so the same holds for $\widehat{D}^m$. Since $\widehat{Y}^{(0)}$ is injective, the matrix $\widehat{D}^m\widehat{Y}^{(0)}$ also has to be injective, and thus the matrix $PX^{(m)}$. □

Although the matrices $(X^{(m)})_{m=0}^{\infty}$ will remain injective if $PX^{(0)}$ is, their columns may become "numerically linearly dependent", e.g., they could all converge to $\mathcal{E}(A,\lambda_1)$ if λ_1 is dominant. In order to avoid numerical instability, we guarantee linear independence by orthogonalization.

Definition 4.22 (QR factorization). Let $A \in \mathbb{F}^{n\times m}$ and $k := \min\{n,m\}$. If an isometric matrix $Q \in \mathbb{F}^{n\times k}$ and an upper triangular matrix $R \in \mathbb{F}^{k\times m}$ satisfy

$$A = QR,$$

we call the representation of A as the product QR a *(skinny) QR factorization*.

Proposition 4.23 (QR factorization). *For every matrix $A \in \mathbb{F}^{n\times m}$, there is a QR factorization $A = QR$.*

Proof. (cf. [43, Section 4.1]) By induction on $n \in \mathbb{N}$. The case $n = 1$ is trivial, since A is already upper triangular.

Let now $n \in \mathbb{N}$ and assume that for every matrix $A \in \mathbb{F}^{n\times m}$ with any $m \in \mathbb{N}_0$, there is a QR factorization. Let $m \in \mathbb{N}_0$, and let $A \in \mathbb{F}^{(n+1)\times m}$. If $m = 0$, we are done.

Otherwise we can use Lemma 2.45 to find a unitary matrix $Q_1 \in \mathbb{F}^{(n+1)\times(n+1)}$ with $Q_1^* = Q_1$ mapping the first column of A to a multiple of the first canonical unit vector. If $m = 1$, this yields

$$Q_1A = \begin{pmatrix}r_{11}\\0\end{pmatrix}, \quad A = Q_1\begin{pmatrix}1\\0\end{pmatrix}r_{11} = QR$$

with the matrices

$$Q := Q_1\begin{pmatrix}1\\0\end{pmatrix}, \quad R := \begin{pmatrix}r_{11}\end{pmatrix}.$$

For $m > 1$, on the other hand, we have

$$Q_1A = \begin{pmatrix}r_{11} & R_{1*}\\ & \widehat{A}\end{pmatrix}$$

with $r_{11} \in \mathbb{F}$, $R_{1*} \in \mathbb{F}^{1\times(m-1)}$ and $\widehat{A} \in \mathbb{F}^{n\times(m-1)}$. By the induction assumption, we find an isometric matrix $\widehat{Q} \in \mathbb{F}^{n\times\hat{k}}$ and an upper triangular matrix $\widehat{R} \in \mathbb{F}^{\hat{k}\times(m-1)}$ with $\hat{k} = \min\{n, m-1\}$ and $\widehat{A} = \widehat{Q}\widehat{R}$. We obtain

$$Q_1 A = \begin{pmatrix} r_{11} & R_{1*} \\ & \widehat{A} \end{pmatrix} = \begin{pmatrix} r_{11} & R_{1*} \\ & \widehat{Q}\widehat{R} \end{pmatrix} = \begin{pmatrix} 1 & \\ & \widehat{Q} \end{pmatrix} \begin{pmatrix} r_{11} & R_{1*} \\ & \widehat{R} \end{pmatrix}$$

and conclude

$$A = Q_1 \begin{pmatrix} 1 & \\ & \widehat{Q} \end{pmatrix} \begin{pmatrix} r_{11} & R_{1*} \\ & \widehat{R} \end{pmatrix} = QR$$

with the matrices

$$Q := Q_1 \begin{pmatrix} 1 & \\ & \widehat{Q} \end{pmatrix} \in \mathbb{F}^{(n+1)\times(\hat{k}+1)}, \quad R := \begin{pmatrix} r_{11} & R_{1*} \\ & \widehat{R} \end{pmatrix} \in \mathbb{F}^{(\hat{k}+1)\times m}.$$

Observing $k = \min\{n+1, m\} = \min\{n, m-1\}+1 = \hat{k}+1$ completes the proof. □

This implies that we can find isometric matrices $(Q^{(m)})_{m=0}^{\infty}$ in $\mathbb{F}^{n\times k}$ and upper triangular matrices $(R^{(m)})_{m=0}^{\infty}$ in $\mathbb{F}^{k\times k}$ such that

$$X^{(m)} = Q^{(m)} R^{(m)} \quad \text{for all } m \in \mathbb{N}_0. \tag{4.24}$$

In order to avoid numerical instabilities, we do not work directly with $X^{(m)}$: in a first step, we compute $Q^{(0)}$ and $R^{(0)}$ satisfying (4.24) directly. If $Q^{(m)}$ and $R^{(m)}$ have been computed, we have

$$X^{(m+1)} = AX^{(m)} = AQ^{(m)} R^{(m)}.$$

We compute a skinny QR factorization of

$$Y^{(m+1)} := AQ^{(m)},$$

i.e., we find an isometric $Q^{(m+1)} \in \mathbb{F}^{n\times k}$ and an upper triangular $\widehat{R}^{(m+1)} \in \mathbb{F}^{k\times k}$ such that

$$Y^{(m+1)} = Q^{(m+1)} \widehat{R}^{(m+1)}.$$

This implies

$$X^{(m+1)} = AQ^{(m)} R^{(m)} = Y^{(m+1)} R^{(m)} = Q^{(m+1)} \widehat{R}^{(m+1)} R^{(m)},$$

and by choosing

$$R^{(m+1)} := \widehat{R}^{(m+1)} R^{(m)},$$

we can obtain the required factorization without computing $X^{(m+1)}$ explicitly. The resulting algorithm is called the *simultaneous iteration* (sometimes also *subspace iteration* or *orthogonal iteration*) and summarized in Figure 4.5.

```
procedure simultaneous_iteration(A, var Q);
begin
  while „not sufficiently accurate“ do begin
    Y ← AQ;
    Compute skinny QR factorization Y = Q R̂
  end
end
```

Figure 4.5. Simultaneous iteration, still missing a stopping criterion.

For $k = 1$, the skinny QR decomposition coincides with dividing a one-column matrix by the norm of this column, so the simultaneous iteration is a generalization of the simple power iteration to the case of multiple vectors.

We can generalize the stopping criterion used for the power iteration: instead of computing the Rayleigh quotient

$$\lambda^{(m)} := \langle Ax^{(m)}, x^{(m)} \rangle \quad \text{for all } m \in \mathbb{N}_0,$$

(recall that we have $\|x^{(m)}\| = 1$ in the practical power iteration), we compute the $k \times k$ matrix

$$\Lambda^{(m)} := (Q^{(m)})^* A Q^{(m)} \quad \text{for all } m \in \mathbb{N}_0, \tag{4.25}$$

and instead of considering the norm of the residual vector $r^{(m)} = Ax^{(m)} - \lambda^{(m)} x^{(m)}$, we use the norm of the residual matrix

$$R^{(m)} := Q^{(m)} \Lambda^{(m)} - A Q^{(m)} \quad \text{for all } m \in \mathbb{N}_0.$$

Note that this is *not* the same matrix as the one used in (4.24), we only use the same letter, for obvious reasons. The resulting practical simultaneous iteration is given in Figure 4.6. The error analysis of Section 4.3 can be extended to this case:

Exercise 4.24 (Rank-$2k$ matrix). Let $C, D \in \mathbb{F}^{n \times k}$ with $D^*C = 0$, and let $X := CD^* + DC^*$. Prove $\|X\|_F^2 \leq 2\|C\|_F^2 \|D\|_F^2$ and $\|X\| \leq \|C\| \, \|D\|$. (Hint: Consider the Cauchy–Schwarz inequality for the first estimate and the proof of Lemma 4.8 for the second).

Exercise 4.25 (Eigenvalues). Let $m \in \mathbb{N}_0$ and $\mu \in \sigma(\Lambda^{(m)})$. Prove that there is a $\lambda \in \sigma(A)$ such that

$$|\mu - \lambda| \leq \|R^{(m)}\|.$$

(Hint: Consider the proof of Theorem 4.9).

```
procedure simultaneous_iteration(A, var Q);
begin
  Y ← AQ;
  Λ ← Q*Y;
  while ‖Y − QΛ‖ > ϵ do begin
    Compute skinny QR factorization Y = Q R̂;
    Y ← AQ;
    Λ ← Q*Y
  end
end
```

Figure 4.6. Practical simultaneous iteration.

Due to Theorem 4.20, we know that the angles between the columns of the matrices $X^{(m)}$ and the invariant subspace $\mathcal{E}_k$ spanned by the eigenvectors corresponding to the first k eigenvalues will converge to zero. Proposition 4.21 ensures that the columns of $X^{(m)}$ span the same subspace as those of $Q^{(m)}$, so the convergence result holds for the latter as well.

In view of Proposition 2.43, we can characterize "convergence to an invariant subspace" also in terms of equation (2.20) and obtain the following result.

Theorem 4.26 (Convergence to invariant subspace). *Let* (4.2) *and* (4.3) *hold with*

$$|\lambda_1| \geq \cdots \geq |\lambda_k| > |\lambda_{k+1}| \geq \cdots \geq |\lambda_n|.$$

Let $P \in \mathbb{F}^{n\times n}$ *be the projection onto* $\mathcal{E}_k$ *defined by* (4.21).

Let $X^{(0)} \in \mathbb{F}^{n\times k}$, *and assume that* $PX^{(0)}$ *is injective. We can find a matrix* $\Lambda \in \mathbb{F}^{k\times k}$ *such that*

$$\frac{\|(AX^{(m)} - X^{(m)}\Lambda)y\|}{\|PX^{(m)}y\|} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(AX^{(0)} - X^{(0)}\Lambda)y\|}{\|PX^{(0)}y\|}$$
$$\text{for all } y \in \mathbb{F}^k \setminus \{0\} \quad \text{and all } m \in \mathbb{N}_0.$$

Proof. As in the previous proofs, we introduce transformed matrices

$$\widehat{X}^{(m)} := Q^* X^{(m)} \quad \text{for all } m \in \mathbb{N}_0$$

and observe

$$\widehat{X}^{(m+1)} = Q^* X^{(m+1)} = Q^* A X^{(m)} = Q^* A Q \widehat{X}^{(m)} = D\widehat{X}^{(m)} \quad \text{for all } m \in \mathbb{N}_0.$$

We define $\widehat{X}_k^{(m)} \in \mathbb{F}^{k\times k}$ and $\widehat{X}_\perp^{(m)} \in \mathbb{F}^{(n-k)\times k}$ by

$$\widehat{X}^{(m)} = \begin{pmatrix} \widehat{X}_k^{(m)} \\ \widehat{X}_\perp^{(m)} \end{pmatrix} \quad \text{for all } m \in \mathbb{N}_0.$$

Due to (4.21) we have

$$Q^*PX^{(0)} = Q^*PQ\widehat{X}^{(0)} = \begin{pmatrix} I_k & \\ & 0 \end{pmatrix}\widehat{X}^{(0)} = \begin{pmatrix} I_k & \\ & 0 \end{pmatrix}\begin{pmatrix} \widehat{X}_k^{(0)} \\ \widehat{X}_\perp^{(0)} \end{pmatrix} = \begin{pmatrix} \widehat{X}_k^{(0)} \\ 0 \end{pmatrix},$$

and since $PX^{(0)}$ is injective, the matrix $\widehat{X}_k^{(0)}$ has to be invertible.

We split the diagonal matrix D into

$$D_k := \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix}, \quad D_\perp := \begin{pmatrix} \lambda_{k+1} & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

and obtain

$$\begin{aligned} AX^{(m)} - X^{(m)}\Lambda &= QDQ^*X^{(m)} - QQ^*X^{(m)}\Lambda = Q(D\widehat{X}^{(m)} - \widehat{X}^{(m)}\Lambda) \\ &= Q(D^{m+1}\widehat{X}^{(0)} - D^m\widehat{X}^{(0)}\Lambda) \\ &= Q\begin{pmatrix} D_k^{m+1}\widehat{X}_k^{(0)} - D_k^m\widehat{X}_k^{(0)}\Lambda \\ D_\perp^{m+1}\widehat{X}_\perp^{(0)} - D_\perp^m\widehat{X}_\perp^{(0)}\Lambda \end{pmatrix} \quad \text{for all } m \in \mathbb{N}_0,\ \Lambda \in \mathbb{F}^{k\times k}. \end{aligned}$$

Since $\widehat{X}_k^{(0)}$ is invertible, we can eliminate the first row by using

$$\Lambda := (\widehat{X}_k^{(0)})^{-1} D_k \widehat{X}_k^{(0)}$$

and use (2.16) and (3.11) to conclude

$$\begin{aligned} \|(AX^{(m)} - X^{(m)}\Lambda)y\| &= \left\| \begin{pmatrix} 0 \\ D_\perp^m(D_\perp\widehat{X}_\perp^{(0)} - \widehat{X}_\perp^{(0)}\Lambda)y \end{pmatrix} \right\| \\ &\le \|D_\perp\|^m \|(D_\perp\widehat{X}_\perp^{(0)} - \widehat{X}_\perp^{(0)}\Lambda)y\| \\ &= |\lambda_{k+1}|^m \|(D_\perp\widehat{X}_\perp^{(0)} - \widehat{X}_\perp^{(0)}\Lambda)y\| \\ &\quad \text{for all } m \in \mathbb{N}_0,\ y \in \mathbb{F}^k. \end{aligned}$$

This yields the numerator of the required estimate. For the denominator, we use

$$\begin{aligned} \|PX^{(m)}y\| &= \|Q\widehat{P}Q^*X^{(m)}y\| = \|Q\widehat{P}\widehat{X}^{(m)}y\| = \|Q\widehat{P}D^m\widehat{X}^{(0)}y\| \\ &= \left\| \begin{pmatrix} I_k & \\ & 0 \end{pmatrix} \begin{pmatrix} D_k^m & \\ & D_\perp^m \end{pmatrix} \begin{pmatrix} \widehat{X}_k^{(0)} \\ \widehat{X}_\perp^{(0)} \end{pmatrix} y \right\| \\ &= \|D_k^m\widehat{X}_k^{(0)}y\| \ge |\lambda_k|^m \|\widehat{X}_k^{(0)}y\| \quad \text{for all } y \in \mathbb{F}^k,\ m \in \mathbb{N}_0, \end{aligned}$$

where we have used

$$\|D_kz\| = \left(\sum_{i=1}^k |\lambda_i|^2|z_i|^2\right)^{1/2} \ge \left(\sum_{i=1}^k |\lambda_k|^2|z_i|^2\right)^{1/2} = |\lambda_k|\,\|z\| \quad \text{for all } z \in \mathbb{F}^k$$

in the last step. Using the definitions of P and $\widehat{X}^{(0)}$, we have

$$\|\widehat{X}_k^{(0)} y\| = \|\widehat{P}\widehat{X}^{(0)} y\| = \|PX^{(0)} y\| \quad \text{for all } m \in \mathbb{N}_0,\ y \in \mathbb{F}^k,$$

and this yields the required estimate for the denominator. □

During the simultaneous iteration, we do not compute $X^{(m)}$ but the isometric matrices $Q^{(m)}$. For these matrices, the convergence result takes a particularly simple form.

Corollary 4.27 (Invariant subspace). *Let the conditions of Theorem* 4.26 *hold, let* $(Q^{(m)})_{m=0}^{\infty}$ *and* $(R^{(m)})_{m=0}^{\infty}$ *be as in* (4.24)*, and let* $(\Lambda^{(m)})_{m=0}^{\infty}$ *be as in* (4.25)*. Then we have*

$$\|AQ^{(m)} - Q^{(m)}\Lambda^{(m)}\| \leq C_0 \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \quad \text{for all } m \in \mathbb{N}_0,$$

where the constant is given by

$$C_0 := \max\left\{\frac{\|(AX^{(0)} - X^{(0)}\Lambda)y\|}{\|PX^{(0)} y\|} \ :\ y \in \mathbb{F}^k \setminus \{0\}\right\}.$$

Proof. Let $m \in \mathbb{N}_0$. By definition, we have

$$AX^{(m)} - X^{(m)}\Lambda = AQ^{(m)}R^{(m)} - Q^{(m)}R^{(m)}\Lambda,$$

and Proposition 4.21 guarantees that $R^{(m)}$ is invertible, so we obtain

$$\begin{aligned} AX^{(m)} - X^{(m)}\Lambda &= AQ^{(m)}R^{(m)} - Q^{(m)}R^{(m)}\Lambda(R^{(m)})^{-1}R^{(m)} \\ &= (AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)})R^{(m)} \end{aligned}$$

with the matrix

$$\widetilde{\Lambda}^{(m)} := R^{(m)}\Lambda(R^{(m)})^{-1}.$$

Let $z \in \mathbb{F}^k \setminus \{0\}$, and let $y := (R^{(m)})^{-1}z$. Since (4.19b) and (2.16) imply $\|PQ^{(m)}z\| \leq \|Q^{(m)}z\| = \|z\|$, we find

$$\begin{aligned} \frac{\|(AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)})z\|}{\|z\|} &\leq \frac{\|(AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)})z\|}{\|PQ^{(m)}z\|} \\ &= \frac{\|(AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)})R^{(m)}y\|}{\|PQ^{(m)}R^{(m)}y\|} \\ &= \frac{\|(AQ^{(m)}R^{(m)} - Q^{(m)}R^{(m)}\Lambda(R^{(m)})^{-1}R^{(m)})y\|}{\|PQ^{(m)}R^{(m)}y\|} \\ &= \frac{\|(AX^{(m)} - X^{(m)}\Lambda)y\|}{\|PX^{(m)}y\|}. \end{aligned}$$

Applying Theorem 4.26 and the Definition 3.5 yields

$$\|AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)}\| \leq C_0 \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m. \tag{4.26}$$

Now we only have to replace $\widetilde{\Lambda}^{(m)}$ by $\Lambda^{(m)}$. To this end let $\Pi := Q^{(m)}(Q^{(m)})^*$. Since $Q^{(m)}$ is isometric, we have $\Pi^2 = Q^{(m)}(Q^{(m)})^* Q^{(m)}(Q^{(m)})^* = Q^{(m)}(Q^{(m)})^* = \Pi$ and $\Pi^* = \Pi$, so Π is an orthogonal projection onto the range $\mathcal{R}(Q^{(m)})$ of $Q^{(m)}$. Let $z \in \mathbb{F}^k$. Due to $Q^{(m)}\widetilde{\Lambda}^{(m)}z \in \mathcal{R}(Q^{(m)})$, we can use the best-approximation property (4.19a) to obtain

$$\begin{aligned}
\|(AQ^{(m)} - Q^{(m)}\Lambda^{(m)})z\| &= \|AQ^{(m)}z - Q^{(m)}(Q^{(m)})^* AQ^{(m)}z\| \\
&= \|AQ^{(m)}z - \Pi AQ^{(m)}z\| \leq \|AQ^{(m)}z - Q^{(m)}\widetilde{\Lambda}^{(m)}z\| \\
&= \|(AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)})z\| \\
&\leq \|AQ^{(m)} - Q^{(m)}\widetilde{\Lambda}^{(m)}\|\,\|z\|
\end{aligned}$$

and using the intermediate result (4.26) in combination with the Definition 3.5 of the spectral norm completes the proof. □

4.8 Convergence for general matrices *

The convergence analysis presented so far relies on the diagonalizability of the matrix A. In this section, we consider the convergence of the power iteration for non-diagonalizable matrices. In practice, a non-diagonalizable matrix can be represented as a limit of diagonalizable matrices, since we can always add a small perturbation that ensures that all eigenvalues are single. Since this approach would require a substantial generalization of the perturbation theory presented so far, we prefer to handle the non-diagonalizable case directly using the results of Section 2.7 to introduce a block-diagonal representation of A.

Let $A \in \mathbb{C}^{n\times n}$. According to Theorem 2.58, we can find an invertible matrix $B \in \mathbb{C}^{n\times n}$ and upper triangular matrices $R_1 \in \mathbb{C}^{n_1\times n_1}, \ldots, R_k \in \mathbb{C}^{n_k\times n_k}$ such that

$$A = B \begin{pmatrix} R_1 & & \\ & \ddots & \\ & & R_k \end{pmatrix} B^{-1} \tag{4.27}$$

and $\sigma(R_\ell) = \{\lambda_\ell\}$ for all $\ell \in \{1, \ldots, k\}$. This decomposition can take the place of the diagonalization (4.2) we relied on in the analysis of the power iteration in the diagonalizable case: the matrix is *block-diagonalizable*.

We have to investigate the matrices R_ℓ more closely. Let $\ell \in \{1, \ldots, k\}$. Since R_ℓ is upper triangular, Proposition 2.13 implies that the diagonal elements of R_ℓ are its eigenvalues. Due to $\sigma(R_\ell) = \{\lambda_\ell\}$, all diagonal elements have to equal λ_ℓ. This

suggests a decomposition of R_ℓ into the diagonal part and the remainder, and the remainder will be a special kind of triangular matrix.

Definition 4.28 (α-triangular matrix). Let $N \in \mathbb{F}^{n\times n}$ and $\alpha \in \mathbb{N}_0$. We call N *α-upper triangular*, if

$$n_{ij} = 0 \qquad \text{holds for all } i, j \in \{1, \ldots, n\} \text{ with } j < i + \alpha. \tag{4.28}$$

A 1-upper triangular matrix is also called *strictly upper triangular*.

We split the matrix R_ℓ into the diagonal part $\lambda_\ell I$ and a strictly upper triangular matrix $N_\ell \in \mathbb{C}^{n_\ell \times n_\ell}$:

$$R_\ell = \lambda_\ell I + N_\ell.$$

The vectors computed during the power iteration result from applying powers of the matrix to the starting vector, therefore we have to investigate the powers of R_ℓ. Since I and N_ℓ commute, these powers are given by

$$R_\ell^m = (\lambda_\ell I + N_\ell)^m = \sum_{j=0}^{m} \binom{m}{j} \lambda_\ell^{m-j} N_\ell^j \qquad \text{for all } m \in \mathbb{N}_0.$$

We can simplify this equations significantly by taking advantage of the properties of triangular matrices:

Lemma 4.29 (Products of triangular matrices). *Let $N, M \in \mathbb{F}^{n\times n}$ be α- and β-upper triangular, respectively. Then the product NM is $(\alpha + \beta)$-upper triangular.*

Proof. We use contraposition: let $i, j \in \{1, \ldots, n\}$ be given such that $(NM)_{ij} \neq 0$. Due to

$$0 \neq (NM)_{ij} = \sum_{k=1}^{n} n_{ik} m_{kj},$$

there has to be at least one index $k \in \{1, \ldots, n\}$ with $n_{ik} \neq 0$ and $m_{kj} \neq 0$. Since N is α-upper triangular and M is β-upper triangular, this implies $k \geq i + \alpha$ and $j \geq k + \beta$ due to (4.28). Combining both estimates gives us $j \geq k + \beta \geq i + \alpha + \beta$. We conclude that $(NM)_{ij} \neq 0$ implies $j \geq i + \alpha + \beta$, and contraposition yields that $j < i + \alpha + \beta$ implies $(NM)_{ij} = 0$, so NM has to be $(\alpha + \beta)$-upper triangular. □

Since N_ℓ is strictly upper triangular, Lemma 4.29 and a simple induction yield that $N_\ell^{n_\ell}$ is n_ℓ-upper triangular, and since the matrix is of size $n_\ell \times n_\ell$, this implies $N_\ell^{n_\ell} = 0$. Therefore the m-th power of R_ℓ takes the form

$$R_\ell^m = \sum_{j=0}^{n_\ell - 1} \binom{m}{j} \lambda_\ell^{m-j} N_\ell^j \qquad \text{for all } m \in \mathbb{N}_{\geq n_\ell - 1}.$$

In order to obtain a generalized version of Theorem 4.2, we have to find lower and upper bounds for matrices of this type.

Lemma 4.30 (Bounds for iteration vectors). *Let $\lambda \in \mathbb{F}$, let $N \in \mathbb{F}^{n\times n}$ be a strictly upper triangular matrix, and let $R = \lambda I + N$.*

Let $x \in \mathbb{F}^n$. There are a constant $c \in \mathbb{R}_{>0}$ and a polynomial p of degree not higher than $n-1$ such that

$$c|\lambda|^m\|x\| \leq \|R^m x\| \leq p(m)|\lambda|^{m-(n-1)}\|x\| \quad \text{for all } m \in \mathbb{N}_{\geq n-1}.$$

Proof. Since the result is trivial for $x = 0$, we consider only the case $x \neq 0$. Let $m \in \mathbb{N}_{\geq n}$, $P := R^m$ and $y := Px = R^m x$. Due to Lemma 4.29, the product of upper triangular matrices is upper triangular, and a simple induction yields that $P := R^m$ is also upper triangular.

We first prove the lower bound. Since $x \neq 0$, we can find an index k such that $x_k \neq 0$. The largest such index is given by

$$\ell := \max\{k \in \{1,\ldots,n\} \,:\, x_k \neq 0\}.$$

Since P is upper triangular and $x_k = 0$ for all $k > \ell$, we have

$$y_\ell = (Px)_\ell = \sum_{j=1}^{n} p_{\ell j}x_j = \sum_{j=\ell}^{n} p_{\ell j}x_j = p_{\ell\ell}x_\ell.$$

Since I and N commute, the binomial theorem yields

$$P = R^m = (\lambda I + N)^m = \sum_{j=0}^{m} \binom{m}{j} \lambda^{m-j} N^j. \tag{4.29}$$

N is strictly upper triangular, and the same holds for N^j with $j > 0$ by Lemma 4.29. Therefore the only term in (4.29) contributing to the diagonal of P is the term for $j = 0$, i.e., $\lambda^m I$. We conclude that all diagonal elements of P equal λ^m, and in particular

$$y_\ell = p_{\ell\ell}x_\ell = \lambda^m x_\ell.$$

Since $x_\ell \neq 0$, the constant

$$c := \frac{|x_\ell|}{\|x\|} > 0$$

is well-defined, and we obtain

$$\|R^m x\| = \|y\| \geq |y_\ell| = |\lambda|^m |x_\ell| = c|\lambda|^m\|x\|.$$

Now we consider the upper bound. Applying the triangle inequality to (4.29) yields

$$\|y\| = \|Px\| = \left\| \sum_{j=0}^{m} \binom{m}{j} \lambda^{m-j} N^j x \right\| \leq \sum_{j=0}^{m} \binom{m}{j} |\lambda|^{m-j} \|N^j x\|.$$

Since N is strictly upper triangular, Lemma 4.29 yields that N^n is n-upper triangular, i.e., $N^n = 0$, and therefore

$$\|y\| \leq \sum_{j=0}^{n-1} \binom{m}{j} |\lambda|^{m-j} \|N^j x\|.$$

Using the compatibility (3.11) of the norms, we conclude

$$\|y\| \leq \sum_{j=0}^{n-1} \binom{m}{j} |\lambda|^{m-j} \|N\|^j \|x\| = |\lambda|^{m-(n-1)} \sum_{j=0}^{n-1} \binom{m}{j} |\lambda|^{n-1-j} \|N\|^j \|x\|. \tag{4.30}$$

Now we only have to prove that the sum can be expressed by a polynomial in m. Due to

$$\binom{m}{j} = \frac{m!}{(m-j)!\,j!} = \frac{\prod_{k=m-j+1}^{m} k}{\prod_{k=1}^{j} k} = \prod_{k=1}^{j} \frac{k+m-j}{k} \quad \text{for all } j \in \{0, \ldots, n-1\},$$

the polynomials

$$p_j : \mathbb{F} \to \mathbb{F}, \quad z \mapsto \prod_{k=1}^{j} \frac{k+z-j}{k},$$

satisfy $p_j(m) = \binom{m}{j}$ for all $m \in \mathbb{N}_{\geq n}$ and $j \in \{0, \ldots, n-1\}$.

Since each p_j is the product of j linear factors, its order is bounded by j, and

$$p : \mathbb{F} \to \mathbb{F}, \quad z \mapsto \sum_{j=0}^{n-1} p_j(z) |\lambda|^{n-1-j} \|N\|^j,$$

is a polynomial of degree not larger than $n-1$ satisfying

$$p(m) = \sum_{j=0}^{n-1} \binom{m}{j} |\lambda|^{n-1-j} \|N\|^j$$

for all m. Combining this equation with (4.30) yields

$$\|y\| \leq |\lambda|^{m-(n-1)} \sum_{j=0}^{n-1} \binom{m}{j} |\lambda|^{n-1-j} \|N\|^j \|x\| = |\lambda|^{m-(n-1)} p(m) \|x\|,$$

and this is the required upper bound. □

Note that the constant c of Lemma 4.30 depends on the vector x, particularly on the ratio between its last non-zero entry and its Euclidean norm.

We can apply this Lemma to the decomposition (4.27) if we assume that one eigenvalue is dominant.

Theorem 4.31 (Convergence). *Let $A \in \mathbb{C}^{n\times n}$, and let $B, R_1, \ldots, R_k, n_1, \ldots, n_k$ and $\lambda_1, \ldots, \lambda_k$ be as in* (4.27)*. Let*

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_k|, \tag{4.31}$$

and let

$$\widehat{\mathcal{G}}_1 := \mathbb{C}^{n_1} \times \{0\} \subseteq \mathbb{C}^n, \quad \mathcal{G}_1 := B\widehat{\mathcal{G}}_1.$$

$\mathcal{G}_1$ is the invariant subspace of generalized *eigenvectors (cf. Definition* 2.59*) for the dominant eigenvalue λ_1. Let $x^{(0)} \in \mathbb{C}^n \setminus \{0\}$ satisfy $\cos\angle(B^{-1}x^{(0)}, \widehat{\mathcal{G}}_1) \neq 0$. Let $(x^{(m)})_{m=0}^{\infty}$ be the iteration vectors of the usual power iteration given by* (4.1)*. Then there is a polynomial p of degree not higher than $\hat{n} := \max\{n_2 - 1, \ldots, n_k - 1\}$ such that*

$$\sin\angle(x^{(m)}, \mathcal{G}_1) \leq \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{m-\hat{n}} p(m) \quad \text{for all } m \in \mathbb{N}_{\geq\hat{n}},$$

i.e., the angle between the iteration vectors and the subspace $\mathcal{G}_1$ converges to zero.

Proof. We follow the approach already used in the proof of Theorem 4.20: we define

$$\widehat{x}^{(m)} := B^{-1}x^{(m)} \quad \text{for all } m \in \mathbb{N}_0$$

and observe

$$\widehat{x}^{(m+1)} = \begin{pmatrix} R_1 & & \\ & \ddots & \\ & & R_k \end{pmatrix} \widehat{x}^{(m)} \quad \text{for all } m \in \mathbb{N}_0. \tag{4.32}$$

We define $\widehat{x}_\ell^{(m)} \in \mathbb{C}^{n_\ell}$ for all $m \in \mathbb{N}_0$ and $\ell \in \{1, \ldots, k\}$ by

$$\widehat{x}^{(m)} = \begin{pmatrix} \widehat{x}_1^{(m)} \\ \vdots \\ \widehat{x}_k^{(m)} \end{pmatrix} \quad \text{for all } m \in \mathbb{N}_0,$$

and (4.32) yields

$$\widehat{x}_\ell^{(m+1)} = R_\ell \widehat{x}_\ell^{(m)} \quad \text{for all } m \in \mathbb{N}_0,\ \ell \in \{1, \dots, k\}.$$

Straightforward inductions gives us

$$\widehat{x}_\ell^{(m)} = R_\ell^m \widehat{x}_\ell^{(0)} \quad \text{for all } m \in \mathbb{N}_0,\ \ell \in \{1, \dots, k\}, \tag{4.33}$$

so we can apply Lemma 4.30 to each of the subvectors.

We are interested in an upper bound of the angle between the transformed iteration vectors $\widehat{x}^{(m)}$ and the subspace $\widehat{\mathcal{G}}_1$ corresponding to the first eigenvalue λ_1. In order to be able to apply Proposition 4.19, we introduce the orthogonal projection

$$\widehat{P} := \begin{pmatrix} I_{n_1} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \in \mathbb{C}^{n \times n}$$

into the relevant subspace $\mathbb{C}^{n_1} \times \{0\} \subseteq \mathbb{C}^n$.

We first apply Lemma 4.30 to the block R_1. The Lemma yields a constant $c_1 \in \mathbb{R}_{>0}$, independent of m, such that

$$c_1 |\lambda_1|^m \|\widehat{x}_1^{(0)}\| \le \|R_1^m \widehat{x}_1^{(0)}\| \quad \text{for all } m \in \mathbb{N}_{\ge n_1}.$$

Due to the definition (2.8) of the Euclidean norm and (4.33), this implies

$$\begin{aligned} \|\widehat{P}\widehat{x}^{(m)}\|^2 = \|\widehat{x}_1^{(m)}\|^2 = \|R_1^m \widehat{x}_1^{(0)}\|^2 &\ge c_1^2 |\lambda_1|^{2m} \|\widehat{x}_1^{(0)}\|^2 \\ &= c_1^2 |\lambda_1|^{2m} \|\widehat{P}\widehat{x}^{(0)}\|^2 \quad \text{for all } m \in \mathbb{N}_{\ge \hat{n}}. \end{aligned} \tag{4.34}$$

By assumption and Proposition 4.19, we have $\|\widehat{P}\widehat{x}^{(0)}\| = \cos \angle(\widehat{x}^{(0)}, \widehat{\mathcal{G}}_1) \neq 0$, so both sides of (4.34) are strictly positive. Now we apply Lemma 4.30 to the remaining blocks $R_2, \dots, R_k$. We find polynomials $p_2, \dots, p_k$, independent of m, such that

$$\|R_\ell^m \widehat{x}_\ell^{(0)}\| \le p_\ell(m) |\lambda_\ell|^{m-(n_\ell-1)} \|\widehat{x}_\ell^{(0)}\| \quad \text{for all } m \in \mathbb{N}_{\ge n_\ell - 1},\ \ell \in \{2, \dots, k\}.$$

Using again (2.8) and (4.33), we obtain

$$\begin{aligned} \|(I - \widehat{P})\widehat{x}^{(m)}\|^2 = \sum_{\ell=2}^{k} \|\widehat{x}_\ell^{(m)}\|^2 &= \sum_{\ell=2}^{k} \|R_\ell^m \widehat{x}_\ell^{(0)}\|^2 \\ &\le \sum_{\ell=2}^{k} p_\ell(m)^2 |\lambda_\ell|^{2(m-\hat{n})} \|\widehat{x}_\ell^{(0)}\|^2 \end{aligned}$$

$$\leq \sum_{\ell=2}^{k} p_\ell(m)^2 |\lambda_2|^{2(m-\hat{n})} \|\widehat{x}_\ell^{(0)}\|^2$$

$$\leq \max\{p_\ell(m)^2 \,:\, \ell \in \{2,\dots,k\}\} |\lambda_2|^{2(m-\hat{n})} \sum_{\ell=2}^{k} \|\widehat{x}_\ell^{(0)}\|^2$$

for all $m \in \mathbb{N}_{\geq \hat{n}}$.

By taking the maximum of the absolute values of the coefficients of the polynomials $p_2, \dots, p_k$, we can construct a polynomial $p_{\max}$ of degree not larger than $\hat{n}$ such that

$$|p_\ell(m)| \leq p_{\max}(m) \quad \text{for all } m \in \mathbb{N}_{\geq \hat{n}},\ \ell \in \{2,\dots,k\},$$

and conclude

$$\begin{aligned} \|(I - \widehat{P})\widehat{x}^{(m)}\|^2 &\leq p_{\max}(m)^2 |\lambda_2|^{2(m-\hat{n})} \sum_{\ell=2}^{k} \|\widehat{x}_\ell^{(0)}\|^2 \\ &= p_{\max}(m)^2 |\lambda_2|^{2(m-\hat{n})} \|(I - \widehat{P})\widehat{x}^{(0)}\|^2 \quad \text{for all } m \in \mathbb{N}_{\geq \hat{n}}. \end{aligned} \tag{4.35}$$

Now we have to translate the estimates back to the original iteration vectors $x^{(m)}$. We let $\widetilde{P} := B\widehat{P}B^{-1}$ and observe that it is a projection onto $\mathcal{G}_1$, although unfortunately not orthogonal in general. We denote the orthogonal projection onto $\mathcal{G}_1$ by $P \in \mathbb{C}^{n\times n}$.

The best-approximation property (4.19a) and (3.11) yield

$$\begin{aligned} \|(I - P)x^{(m)}\| &\leq \|(I - \widetilde{P})x^{(m)}\| = \|B(I - \widehat{P})B^{-1}x^{(m)}\| \\ &= \|B(I - \widehat{P})\widehat{x}^{(m)}\| \leq \|B\|\,\|(I - \widehat{P})\widehat{x}^{(m)}\| \quad \text{for all } m \in \mathbb{N}_0, \end{aligned}$$

so we can apply (4.35) to obtain

$$\|(I - P)x^{(m)}\| \leq \|B\| p_{\max}(m) |\lambda_2|^{m-\hat{n}} \|(I - \widehat{P})\widehat{x}^{(0)}\| \quad \text{for all } m \in \mathbb{N}_{\geq \hat{n}}. \tag{4.36}$$

Using (3.11) and (4.19b) yields

$$\|B^{-1}\|\,\|x^{(m)}\| \geq \|B^{-1}x^{(m)}\| = \|\widehat{x}^{(m)}\| \geq \|\widehat{P}\widehat{x}^{(m)}\| \quad \text{for all } m \in \mathbb{N}_0,$$

so we can apply (4.34) to obtain

$$\|x^{(m)}\| \geq \frac{\|\widehat{P}\widehat{x}^{(m)}\|}{\|B^{-1}\|} \geq \frac{c_1 |\lambda_1|^m \|\widehat{P}\widehat{x}^{(0)}\|}{\|B^{-1}\|} \quad \text{for all } m \in \mathbb{N}_0. \tag{4.37}$$

Combining (4.36) and (4.37) with Proposition 4.19, we find

$$\begin{aligned}\sin\angle(x^{(m)},\mathcal{G}_1) &= \frac{\|(I-P)x^{(m)}\|}{\|x^{(m)}\|}\\ &\leq \frac{\|B\|\,\|B^{-1}\|}{c_1|\lambda_1|^{\hat{n}}}p_{\max}(m)\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{m-\hat{n}}\frac{\|(I-\widehat{P})\widehat{x}^{(0)}\|}{\|\widehat{P}\widehat{x}^{(0)}\|}\\ &= \frac{\|B\|\,\|B^{-1}\|}{c_1|\lambda_1|^{\hat{n}}}p_{\max}(m)\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{m-\hat{n}}\tan\angle(\widehat{x}^{(0)},\widehat{\mathcal{G}}_1)\\ &\quad\text{for all } m\in\mathbb{N}_{\geq\hat{n}},\end{aligned}$$

so setting

$$p(m) := \frac{\|B\|\,\|B^{-1}\|}{c_1|\lambda_1|^{\hat{n}}}p_{\max}(m)\tan\angle(\widehat{x}^{(0)},\widehat{\mathcal{G}}_1)\qquad\text{for all } m\in\mathbb{N}$$

gives us the required estimate. □

Exercise 4.32 (Transformed vectors). Given the assumptions of Theorem 4.31, prove that there is a polynomial $\hat{p}$ of degree not above n such that

$$\tan\angle(\widehat{x}^{(m)},\widehat{\mathcal{G}}_1)\leq\hat{p}(m)\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{m-n}\tan\angle(\widehat{x}^{(0)},\widehat{\mathcal{G}}_1)\qquad\text{for all } m\in\mathbb{N}_{\geq n}.$$

Remark 4.33 (Linear convergence). Theorem 4.31 yields "almost" the same convergence behaviour as Theorem 4.2, at least in the long run: let $\varrho := |\lambda_2|/|\lambda_1| < 1$ denote the rate of convergence for the diagonalizable case, and let $\hat{\varrho}\in(\varrho,1)$. Then we have $q := \varrho/\hat{\varrho} < 1$ and find

$$\sin\angle(x^{(m)},\mathcal{G}_1)\leq\varrho^{m-n}p(m)=\hat{\varrho}^{m-n}q^{m-n}p(m)\qquad\text{for all } m\in\mathbb{N}_{\geq n}.$$

To simplify the expression, we introduce $\hat{m} := m-n$ and $\hat{p}(\hat{m}) := p(\hat{m}+n)$ and obtain

$$\sin\angle(x^{(\hat{m}+n)},\mathcal{G}_1)\leq\hat{\varrho}^{\hat{m}}q^{\hat{m}}\hat{p}(\hat{m})\qquad\text{for all } \hat{m}\in\mathbb{N}_0.$$

Due to $q<1$, we can find $\gamma\in\mathbb{R}_{>0}$ such that $\exp(\gamma)=1/q$, and therefore $\exp(-\gamma)=q$ and $q^{\hat{m}}=\exp(-\gamma\hat{m})$. Since $\hat{m}$ is non-negative, we have

$$\exp(\gamma\hat{m})\geq\sum_{\ell=0}^{n+1}\frac{(\gamma\hat{m})^{\ell}}{\ell!}\geq\frac{\gamma^{n+1}}{(n+1)!}\hat{m}^{n+1}\qquad\text{for all } \hat{m}\in\mathbb{N}_0$$

and conclude

$$\sin\angle(x^{(\hat{m}+n)},\mathcal{G}_1)\leq\hat{\varrho}^{\hat{m}}\frac{\hat{p}(\hat{m})}{\exp(\gamma\hat{m})}\leq\hat{\varrho}^{\hat{m}}\frac{(n+1)!\hat{p}(\hat{m})}{\gamma^{n+1}\hat{m}^{n+1}}\qquad\text{for all } \hat{m}\in\mathbb{N}_0.$$

Since the degree of $\hat{p}$ is bounded by n, we can find a constant

$$C_q := \max\left\{\frac{(n+1)!\,\hat{p}(\hat{m})}{\gamma^{n+1}\hat{m}^{n+1}} \;:\; \hat{m} \in \mathbb{N}_0\right\}$$

and obtain

$$\sin\angle(x^{(m)}, \mathcal{G}_1) = \sin\angle(x^{(\hat{m}+n)}, \mathcal{G}_1) \leq C_q\hat{\varrho}^{\hat{m}} = C_q\hat{\varrho}^{m-n} \quad \text{for all } m \in \mathbb{N}_{\geq n},$$

i.e., the rate of convergence will get arbitrarily close to ϱ.

Chapter 5

QR iteration

Summary

The most successful algorithm for finding all eigenvalues and eigenvectors of a given basis is the *QR iteration*. The basic formulation of the method can be derived from the simultaneous iteration (cf. Section 4.7). In order to obtain an efficient algorithm, several improvements are required: the matrix has to be reduced to Hessenberg form, an appropriate shift strategy has to be used, and partially converged intermediate results have to be handled. The resulting practical QR algorithm can be expected to show quadratic or even cubic convergence and compute all eigenvalues and eigenvectors of an $n \times n$ matrix in $\mathcal{O}(n^3)$ operations.

Learning targets

- ✓ Introduce the basic QR iteration.
- ✓ Derive an algorithm for reducing a matrix to Hessenberg or tridiagonal form.
- ✓ Rewrite the iteration as an implicit algorithm requiring only a small amount of auxiliary storage.
- ✓ Get to know shift strategies for improving the speed of convergence.

5.1 Basic QR step

Historically, the QR iteration can be traced back to the quotient-difference algorithm [36] that characterizes eigenvalues as singularities of a meromorphic function that can be approximated by a series of fractions. It turned out [37] that the algorithm can be interpreted as a sequence of simple LR factorizations, and this interpretation led to an efficient method [38] for computing eigenvalues.

Since LR factorizations may be numerically unstable or even fail to exist for certain matrices, they were later replaced by the far more stable QR factorizations [14, 15, 26], resulting in the QR iteration. In this introduction, we follow an elegant derivation [48] of the method based on the simultaneous iteration (cf. Section 4.7).

Let $A \in \mathbb{F}^{n\times n}$. We consider the simultaneous iteration applied to a *unitary* first matrix $Q^{(0)} \in \mathbb{F}^{n\times n}$, i.e., to a matrix with n columns. By construction, the iteration

matrices satisfy

$$Q^{(m+1)}R^{(m+1)} = AQ^{(m)} \quad \text{for all } m \in \mathbb{N}_0. \tag{5.1}$$

Let $k \in \{1, \ldots, n-1\}$, and split the matrices into submatrices

$$\begin{aligned} Q^{(m)} &= \begin{pmatrix} Q_k^{(m)} & Q_\perp^{(m)} \end{pmatrix}, \quad Q_k^{(m)} \in \mathbb{F}^{n\times k}, \\ R^{(m)} &= \begin{pmatrix} R_{kk}^{(m)} & R_{k\perp}^{(m)} \\ & R_{\perp\perp}^{(m)} \end{pmatrix}, \quad R_{kk}^{(m)} \in \mathbb{F}^{k\times k} \quad \text{for all } m \in \mathbb{N}_0. \end{aligned} \tag{5.2}$$

Using these submatrices, (5.1) can be rewritten as

$$\begin{pmatrix} Q_k^{(m+1)} & Q_\perp^{(m+1)} \end{pmatrix} \begin{pmatrix} R_{kk}^{(m+1)} & R_{k\perp}^{(m+1)} \\ & R_{\perp\perp}^{(m+1)} \end{pmatrix} = A \begin{pmatrix} Q_k^{(m)} & Q_\perp^{(m)} \end{pmatrix} \quad \text{for all } m \in \mathbb{N}_0,$$

and due to the triangular structure of the second factor $R^{(m+1)}$, taking the first block column of this equation yields

$$Q_k^{(m+1)} R_{kk}^{(m+1)} = AQ_k^{(m)} \quad \text{for all } m \in \mathbb{N}_0.$$

Obviously this equation shares the structure of (5.1): the matrices $(Q_k^{(m)})_{m=0}^\infty$ are the result of a simultaneous iteration starting with the isometric matrix $Q_k^{(0)}$.

We assume for the moment that A is self-adjoint (for $\mathbb{F} = \mathbb{R}$) or normal (for $\mathbb{F} = \mathbb{C}$), respectively, and that its eigenvalues satisfy the condition (4.20). Let $P \in \mathbb{F}^{n\times n}$ denote the orthogonal projection into the invariant subspace spanned by the eigenvectors corresponding to the first k eigenvalues. We also assume $PQ_k^{(0)}$ to be injective.

Corollary 4.27 states

$$\|AQ_k^{(m)} - Q_k^{(m)}\Lambda_{kk}^{(m)}\| \le C_0 \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \quad \text{for all } m \in \mathbb{N}_0 \tag{5.3}$$

for the matrices $(\Lambda_{kk}^{(m)})_{m=0}^\infty$ given by

$$\Lambda_{kk}^{(m)} := (Q_k^{(m)})^* AQ_k^{(m)} \quad \text{for all } m \in \mathbb{N}_0.$$

Since the matrices $(Q^{(m)})_{m=0}^\infty$ are unitary, the matrices $(A^{(m)})_{m=0}^\infty$ defined by

$$A^{(m)} := (Q^{(m)})^* AQ^{(m)} \quad \text{for all } m \in \mathbb{N}_0$$

are unitarily similar to A.

In order to investigate these matrices, we fix $m \in \mathbb{N}_0$. Using the splitting (5.2), we obtain

$$A^{(m)} = (Q^{(m)})^* A Q^{(m)} = \begin{pmatrix} (Q_k^{(m)})^* \\ (Q_\perp^{(m)})^* \end{pmatrix} A \begin{pmatrix} Q_k^{(m)} & Q_\perp^{(m)} \end{pmatrix} = \begin{pmatrix} (Q_k^{(m)})^* A Q_k^{(m)} & (Q_k^{(m)})^* A Q_\perp^{(m)} \\ (Q_\perp^{(m)})^* A Q_k^{(m)} & (Q_\perp^{(m)})^* A Q_\perp^{(m)} \end{pmatrix}. \tag{5.4}$$

We are interested in the lower left block. Due to the estimate (5.3), we know that $\|AQ_k^{(m)} - Q_k^{(m)} \Lambda_{kk}^{(m)}\|$ converges to zero, and since $Q^{(m)}$ is unitary, we have

$$\begin{aligned} \begin{pmatrix} (Q_k^{(m)})^* Q_k^{(m)} & (Q_k^{(m)})^* Q_\perp^{(m)} \\ (Q_\perp^{(m)})^* Q_k^{(m)} & (Q_\perp^{(m)})^* Q_\perp^{(m)} \end{pmatrix} &= \begin{pmatrix} (Q_k^{(m)})^* \\ (Q_\perp^{(m)})^* \end{pmatrix} \begin{pmatrix} Q_k^{(m)} & Q_\perp^{(m)} \end{pmatrix} \\ &= (Q^{(m)})^* Q^{(m)} = I = \begin{pmatrix} I & \\ & I \end{pmatrix} \end{aligned}$$

and therefore $(Q_\perp^{(m)})^* Q_k^{(m)} = 0$ in particular. For the lower left block of $A^{(m)}$, this means

$$\begin{aligned} \|(Q_\perp^{(m)})^* A Q_k^{(m)}\| &= \|(Q_\perp^{(m)})^* Q_k^{(m)} \Lambda_{kk}^{(m)} + (Q_\perp^{(m)})^* (AQ_k^{(m)} - Q_k^{(m)} \Lambda_{kk}^{(m)})\| \\ &= \|(Q_\perp^{(m)})^* (AQ_k^{(m)} - Q_k^{(m)} \Lambda_{kk}^{(m)})\| \\ &\leq \left\| \begin{pmatrix} (Q_k^{(m)})^* \\ (Q_\perp^{(m)})^* \end{pmatrix} (AQ_k^{(m)} - Q_k^{(m)} \Lambda_{kk}^{(m)}) \right\| \\ &= \|(Q^{(m)})^* (AQ_k^{(m)} - Q_k^{(m)} \Lambda_{kk}^{(m)})\| \\ &= \|AQ_k^{(m)} - Q_k^{(m)} \Lambda_{kk}^{(m)}\| \leq C_0 \left(\frac{|\lambda_{k+1}|}{|\lambda_k|} \right)^m, \end{aligned} \tag{5.5}$$

where we have used the definition of the norm (2.8) in the third step and Proposition 2.39 and (2.16) in the fifth step.

This estimate means that the lower left block of $A^{(m)}$ in (5.4) converges to zero, therefore $A^{(m)}$ converges to *block upper triangular form*. We can use this property to develop an algorithm that takes the matrices arbitrarily close to triangular form, i.e., we can approximate the Schur decomposition introduced in Theorem 2.46.

We can judge the convergence of the iteration by checking the entries of the matrices $A^{(m)}$, therefore we are interested in finding an efficient way for computing them. Fortunately, we can rewrite the simultaneous iteration in a way that allows us to compute $A^{(m+1)}$ directly from $A^{(m)}$: let $m \in \mathbb{N}_0$. By definition, we have

$$Q^{(m+1)} R^{(m+1)} = AQ^{(m)}, \tag{5.6}$$

```
procedure qr_iteration(var A, Q);
begin
  while „not sufficiently triangular“ do begin
    Compute QR factorization A = Q̂ R;
    A ← R Q̂;
    Q ← Q Q̂
  end
end
```

Figure 5.1. Simple QR iteration.

and multiplying both sides of the equation by $(Q^{(m)})^*$ yields

$$(Q^{(m)})^* Q^{(m+1)} R^{(m+1)} = (Q^{(m)})^* A Q^{(m)} = A^{(m)}.$$

We introduce the unitary matrix

$$\widehat{Q}^{(m+1)} := (Q^{(m)})^* Q^{(m+1)} \tag{5.7}$$

and obtain

$$\widehat{Q}^{(m+1)} R^{(m+1)} = A^{(m)}. \tag{5.8}$$

We can reverse this construction: we compute a QR factorization $\widehat{Q}^{(m+1)} R^{(m+1)} = A^{(m)}$, let $Q^{(m+1)} := Q^{(m)} \widehat{Q}^{(m+1)}$, and observe that the defining equation (5.6) holds. We are interested in finding

$$A^{(m+1)} = (Q^{(m+1)})^* A Q^{(m+1)}.$$

Due to (5.7), we have

$$Q^{(m+1)} = Q^{(m)} \widehat{Q}^{(m+1)} \tag{5.9}$$

and find

$$\begin{aligned} A^{(m+1)} &= (Q^{(m+1)})^* A Q^{(m+1)} = (\widehat{Q}^{(m+1)})^* (Q^{(m)})^* A Q^{(m)} \widehat{Q}^{(m+1)} \\ &= (\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)}. \end{aligned}$$

Using (5.8), we conclude

$$\begin{aligned} A^{(m+1)} &= (\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)} = (\widehat{Q}^{(m+1)})^* \widehat{Q}^{(m+1)} R^{(m+1)} \widehat{Q}^{(m+1)} \\ &= R^{(m+1)} \widehat{Q}^{(m+1)}, \end{aligned}$$

i.e., the matrix $A^{(m+1)}$ can be obtained by multiplying the factors of the QR decomposition of $A^{(m)}$ in reversed order. The resulting simple QR iteration is summarized in Figure 5.1. If we are not interested in the eigenvectors, we can neglect to update the matrix Q in each step.

5.2 Hessenberg form

The simple QR iteration given in Figure 5.1 can require a large number of operations per step: finding the QR factorization $\widehat{Q}^{(m+1)}R^{(m+1)} = A^{(m)}$ of an $n \times n$ matrix $A^{(m)}$ takes approximately $4n^3/3$ arithmetic operations, and multiplying $R^{(m+1)}$ and $\widehat{Q}^{(m+1)}$ to compute $A^{(m+1)}$ can be expected to require a similar amount of operations.

We can improve the efficiency of the algorithm dramatically if the matrices $A^{(m)}$ are of a special form that allows us to compute the factorization more efficiently. In this context the *Hessenberg form* [20] is very useful, a close relative of the upper triangular form:

Definition 5.1 (Hessenberg matrix). Let $n \in \mathbb{N}$ and $H \in \mathbb{F}^{n\times n}$. The matrix H is called *Hessenberg matrix* or *in Hessenberg form* if

$$h_{ij} = 0 \quad \text{for all } i, j \in \{1, \ldots, n\} \text{ with } j + 1 < i.$$

If H^* is also in Hessenberg form, H is called *tridiagonal*.

For a triangular matrix, all entries below the diagonal have to be equal to zero. For a Hessenberg matrix, the entries immediately below the diagonal (the *first subdiagonal*) are allowed to be non-zero. For a tridiagonal matrix, only the entries on the diagonal and immediately below and immediately above the diagonal are allowed to be non-zero.

The resulting patterns of non-zero entries (denoted by "×") for Hessenberg (left) and tridiagonal matrices (right) look as follows:

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix} \qquad \begin{pmatrix} \times & \times & & & \\ \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

The QR factorization of a Hessenberg matrix H can be computed efficiently using Givens rotations [18, Section 5.1.8]: for $x, y \in \mathbb{F}$ with $(x, y) \neq 0$, we have

$$\begin{pmatrix} \bar{c} & -\bar{s} \\ s & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sqrt{|x|^2 + |y|^2} \\ 0 \end{pmatrix}, \quad c := \frac{x}{\sqrt{|x|^2 + |y|^2}}, \quad s := \frac{-y}{\sqrt{|x|^2 + |y|^2}},$$

so we can apply the unitary transformation

$$G := \begin{pmatrix} \bar{c} & -\bar{s} \\ s & c \end{pmatrix}$$

```
procedure givens_setup(var x, y, c, s);
begin
  if y = 0 then begin
    c ← 1;   s ← 0
  end else begin
    ϱ ← √(|x|² + |y|²);
    c ← x/ϱ;   s ← −y/ϱ;
    x ← ϱ;   y ← 0
  end
end;
procedure givens_apply(c, s, var x, y);
begin
  ζ ← x;   x ← c̄ζ − s̄y;   y ← sζ + cy;
end;
procedure givens_conjapply(c, s, var x, y);
begin
  ζ ← x;   x ← cζ − sy;   y ← s̄ζ + c̄y;
end
```

Figure 5.2. Preparation and application of Givens rotations.

to eliminate entries in a two-dimensional vector. Algorithms for finding c and s and applying G and G^* to vectors are given in Figure 5.2. In order to handle n-dimensional vectors, we use the same approach as in (3.7): we choose $p, q \in \{1, \ldots, n\}$ with $p \neq q$ and embed G in a matrix $G_{pq} \in \mathbb{F}^{n \times n}$ such that

$$(G_{pq}x)_i = \begin{cases} \bar{c}x_p - \bar{s}x_q & \text{if } i = p, \\ sx_p + cx_q & \text{if } i = q, \\ x_i & \text{otherwise} \end{cases} \quad \text{for all } x \in \mathbb{F}^n,\ i \in \{1, \ldots, n\}.$$

By multiplying with this matrix, we can eliminate the q-th entry of a given vector changing only the p-th and the q-th component.

We consider a matrix of the form

$$H = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}$$

as an example. We apply a Givens rotation $G^{(1)}$ to the first two rows in order to eliminate the first sub-diagonal element h_{21}:

$$H^{(1)} := G^{(1)} H = \begin{pmatrix} \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix},$$

where "$\otimes$" denotes matrix entries changed by the transformation. Next, we eliminate the second sub-diagonal element h_{32} by applying a Givens rotation $G^{(2)}$ to the second and third row:

$$H^{(2)} := G^{(2)} H^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

Givens rotations $G^{(3)}$ and $G^{(4)}$ applied to the third and fourth and fourth and fifth rows, respectively, lead to the required upper triangular form:

$$H^{(3)} := G^{(3)} H^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \\ & & & \times & \times \end{pmatrix},$$

$$R := G^{(4)} H^{(3)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \end{pmatrix}.$$

In general, we require $n - 1$ Givens rotations, and each rotation works on only two rows of the matrix, so the QR factorization $H = \widehat{Q} R$ of a Hessenberg matrix can be constructed in $\mathcal{O}(n^2)$ operations. In the case of a tridiagonal matrix, we can avoid computing zeros in the upper right triangle and require only $\mathcal{O}(n)$ operations.

For the next step of the QR iteration, we have to compute $\widehat{H} := R\widehat{Q}$, i.e., we have to apply the adjoint Givens rotations to the columns of R. In our example, we have

$Q = (G^{(1)})^*(G^{(2)})^*(G^{(3)})^*(G^{(4)})^*$ and compute

$$\widehat{H}^{(1)} := R(G^{(1)})^* = \begin{pmatrix} \otimes & \otimes & \times & \times & \times \\ \otimes & \otimes & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix},$$

$$\widehat{H}^{(2)} := \widehat{H}^{(1)}(G^{(2)})^* = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix},$$

$$\widehat{H}^{(3)} := \widehat{H}^{(2)}(G^{(3)})^* = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & & & \times \end{pmatrix},$$

$$\widehat{H} := \widehat{H}^{(3)}(G^{(4)})^* = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \end{pmatrix}.$$

Since each step changes only two columns of the matrix, we require only $\mathcal{O}(n^2)$ operations to perform one step of the QR iteration. This is a significant improvement compared to the cubic complexity required by a general QR factorization.

We can see that one step of the QR iteration preserves the Hessenberg form: if $A^{(0)}$ is a Hessenberg matrix, the same holds for all matrices $A^{(m)}$. This property allows us to perform the entire iteration more efficiently. The resulting efficient algorithm for performing one step of the QR iteration is summarized in Figure 5.3.

Proposition 5.2 (Complexity). *The algorithm given in Figure* 5.3 *requires not more than*

$$12n^2 \quad \textit{operations}$$

to perform one step of the QR iteration.

Proof. For a given $k \in \{1, \ldots, n-1\}$, computing ϱ, c_k and s_k requires not more than 8 operations. The application of the corresponding Givens rotation to all columns takes not more than $6(n-k)$ operations, the application to all rows $6k+2$ operations,

```
procedure qr_hessenberg(var A, Q);
begin
  for k = 1 to n − 1 do        { Find factorization A = Q̂ R }
    givens_setup(a_kk, a_{k+1,k}, c_k, s_k);
    for j ∈ {k + 1, ..., n} do givens_apply(c_k, s_k, a_kj, a_{k+1,j})
  end;
  for k = 1 to n − 1 do begin        { Compute R Q̂ and Q Q̂ }
    for i ∈ {1, ..., k} do givens_conjapply(c_k, s_k, a_ik, a_{i,k+1});
    a_{k+1,k} ← −s_k a_{k+1,k+1};    a_{k+1,k+1} ← c̄_k a_{k+1,k+1};
    for i ∈ {1, ..., n} do givens_conjapply(c_k, s_k, q_ik, q_{i,k+1})
  end
end
```

Figure 5.3. One step of the Hessenberg QR iteration.

and the application to all rows of Q $6n$ operations. The total number of operations is bounded by

$$\begin{aligned}\sum_{k=1}^{n-1} (8 + 6(n-k) + (6k+2) + 6n) \\ &= \sum_{k=1}^{n-1} (10 + 12n) \\ &= 10(n-1) + 12n(n-1) \leq 12n^2. \end{aligned}$$ □

The Hessenberg form has the additional advantage that it is easy to monitor convergence to upper triangular form: we only have to check the $n-1$ sub-diagonal entries for convergence to zero.

If A is self-adjoint, the same holds for all matrices $A^{(m)}$, since they are unitarily similar. If $A^{(0)}$ is also a Hessenberg matrix, the same holds for all $A^{(m)}$, so all of these matrices are self-adjoint Hessenberg matrices and therefore tridiagonal. This means that for self-adjoint matrices, we can take advantage of the tridiagonal structure to reduce the number of operations.

Example 5.3 (QR iteration). We consider the QR iteration applied to the matrix

$$A = \begin{pmatrix} 4 & & & \\ 1 & 3 & & \\ & 1 & 2 & \\ & & 1 & 1 \end{pmatrix}.$$

Since it is lower triangular, we easily find $\sigma(A) = \{1, 2, 3, 4\}$. Given the results of Corollary 4.27, we expect convergence rates of $3/4$, $2/3$ and $1/2$ for the first, second and third subdiagonal entries, respectively. The program gives the following results:

m	$\lvert a_{21}^{(m)}\rvert$	Ratio	$\lvert a_{32}^{(m)}\rvert$	Ratio	$\lvert a_{43}^{(m)}\rvert$	Ratio
1	7.46×10^{-1}		6.95×10^{-1}		4.13×10^{-1}	
2	5.48×10^{-1}	0.73	4.62×10^{-1}	0.66	1.75×10^{-1}	0.42
3	3.98×10^{-1}	0.73	2.95×10^{-1}	0.64	8.24×10^{-2}	0.47
4	2.86×10^{-1}	0.72	1.87×10^{-1}	0.63	4.09×10^{-2}	0.50
5	2.06×10^{-1}	0.72	1.20×10^{-1}	0.64	2.07×10^{-2}	0.51
6	1.49×10^{-1}	0.72	7.81×10^{-2}	0.65	1.05×10^{-2}	0.51
7	1.08×10^{-1}	0.72	5.13×10^{-2}	0.66	5.33×10^{-3}	0.51
8	7.88×10^{-2}	0.73	3.39×10^{-2}	0.66	2.70×10^{-3}	0.51

The subdiagonal entries appear to converge at the rate predicted by our theory.

Exercise 5.4 (Self-adjoint QR iteration). Derive a version of the Hessenberg QR iteration given in Figure 5.3 for self-adjoint matrices that requires only $\mathcal{O}(n)$ operations for each iteration.

Apparently the Hessenberg form is useful, therefore we are looking for an algorithm that transforms an arbitrary matrix into this form. We consider a matrix of the form

$$A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}.$$

We first look for a unitary similarity transformation that eliminates the entries a_{31}, a_{41} and a_{51}. This can be done by Givens rotations, but we choose to use *Householder reflections* for the sake of efficiency: according to Lemma 2.45, we can find a unitary self-adjoint matrix P_1 such that

$$P_1 \begin{pmatrix} a_{21} \\ a_{31} \\ a_{41} \\ a_{51} \end{pmatrix} = \begin{pmatrix} \sigma_1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

holds for a $\sigma_1 \in \mathbb{F}$, and applying this transformation to the second to fifth row of A yields

$$\begin{pmatrix} 1 & \\ & P_1 \end{pmatrix} A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \end{pmatrix}.$$

Since we are looking for a similarity transformation, we have to apply the inverse of P_1 to the second to fifth column. Due to $P_1 = P_1^* = P_1^{-1}$, computing the inverse is not really a challenge, and we obtain

$$\widehat{A}^{(1)} := \begin{pmatrix} 1 & \\ & P_1 \end{pmatrix} A \begin{pmatrix} 1 & \\ & P_1 \end{pmatrix} = \begin{pmatrix} \times & \otimes & \otimes & \otimes & \otimes \\ \times & \otimes & \otimes & \otimes & \otimes \\ & \otimes & \otimes & \otimes & \otimes \\ & \otimes & \otimes & \otimes & \otimes \\ & \otimes & \otimes & \otimes & \otimes \end{pmatrix}.$$

Now we find a Householder reflection P_2 that eliminates $\hat{a}_{42}^{(1)}$ and $\hat{a}_{52}^{(1)}$, i.e., such that

$$P_2 \begin{pmatrix} \hat{a}_{32}^{(1)} \\ \hat{a}_{42}^{(1)} \\ \hat{a}_{52}^{(1)} \end{pmatrix} = \begin{pmatrix} \sigma_2 \\ 0 \\ 0 \end{pmatrix}$$

holds for a $\sigma_2 \in \mathbb{F}$. Applying it to the third to fifth rows and columns yields

$$\begin{pmatrix} I_2 & \\ & P_2 \end{pmatrix} \widehat{A}^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \end{pmatrix},$$

$$\widehat{A}^{(2)} := \begin{pmatrix} I_2 & \\ & P_2 \end{pmatrix} \widehat{A}^{(1)} \begin{pmatrix} I_2 & \\ & P_2 \end{pmatrix} = \begin{pmatrix} \times & \times & \otimes & \otimes & \otimes \\ \times & \times & \otimes & \otimes & \otimes \\ & \times & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \end{pmatrix}.$$

In the last step, we find a Householder reflection P_3 that eliminates $\hat{a}_{53}^{(2)}$, i.e., such that

$$P_3 \begin{pmatrix} \hat{a}_{43}^{(2)} \\ \hat{a}_{53}^{(2)} \end{pmatrix} = \begin{pmatrix} \sigma_3 \\ 0 \end{pmatrix}$$

holds for a $\sigma_3 \in \mathbb{F}$. Applying this reflection to the fourth and fifth rows and columns gives us the required Hessenberg form:

$$\begin{pmatrix} I_3 & \\ & P_3 \end{pmatrix} \widehat{A}^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \end{pmatrix},$$

$$\widehat{A}^{(3)} := \begin{pmatrix} I_3 & \\ & P_3 \end{pmatrix} \widehat{A}^{(2)} \begin{pmatrix} I_3 & \\ & P_3 \end{pmatrix} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \end{pmatrix}.$$

We can "hide" this procedure in the choice of the initial basis for the QR iteration: we choose the initial matrix as

$$Q^{(0)} = \begin{pmatrix} 1 & \\ & P_1 \end{pmatrix} \begin{pmatrix} I_2 & \\ & P_2 \end{pmatrix} \begin{pmatrix} I_3 & \\ & P_3 \end{pmatrix}$$

and obtain

$$A^{(0)} = (Q^{(0)})^* A Q^{(0)} = \widehat{A}^{(3)}.$$

The resulting algorithm is given in Figure 5.4.

Exercise 5.5 (Triangular form). Explain why this procedure cannot be used to reach triangular form, i.e., to compute the Schur decomposition directly.

Exercise 5.6 (Self-adjoint matrix). Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint. In this case, it is customary to store only its lower or upper triangular part. Modify the algorithm for transforming a matrix to Hessenberg form (cf. Figure 5.4) in such a way that it works only with the lower triangular part of A.

Proposition 5.7 (Complexity). *The algorithm given in Figure* 5.4 *requires not more than*

$$\frac{16}{3}n^3 \quad \textit{operations}$$

to transform A into a Hessenberg matrix.

Proof. For a given $k \in \{1, \dots, n-2\}$, computing α, β and the Householder vector requires

$$(n-k) + 6 \quad \text{operations.}$$

Applying the Householder reflection to the rows of A requires

$$4(n-k)^2 + (n-k) \quad \text{operations,}$$

while applying it to the columns of A and the columns of Q takes

$$8n(n-k) + 2(n-k) \quad \text{operations.}$$

```
procedure hessenberg(var A, Q);
begin
  for k = 1 to n − 2 do begin
    α ← 0;
    for i ∈ {k + 1, ..., n} do α ← α + |a_ik|^2;
    if α ≠ 0 do begin
      α ← √α;  β ← 1/(α(α + |a_{k+1,k}|));  a_{k+1,k} ← a_{k+1,k} + sgn(a_{k+1,k})α;
      for j ∈ {k + 1, ..., n} do begin        { Apply to rows of A }
        γ ← 0;
        for i ∈ {k + 1, ..., n} do γ ← γ + ā_ik a_ij;
        γ ← γβ;
        for i ∈ {k + 1, ..., n} do a_ij ← a_ij − γ a_ik
      end;
      for i ∈ {1, ..., n} do begin            { Apply to columns of A }
        γ ← 0;
        for j ∈ {k + 1, ..., n} do γ ← γ + a_ij a_jk;
        γ ← γβ;
        for j ∈ {k + 1, ..., n} do a_ij ← a_ij − ā_jk γ
      end;
      for i ∈ {1, ..., n} do begin            { Apply to columns of Q }
        γ ← 0;
        for j ∈ {k + 1, ..., n} do γ ← γ + q_ij a_jk;
        γ ← γβ;
        for j ∈ {k + 1, ..., n} do q_ij ← q_ij − ā_jk γ
      end
    end
  end
end
```

Figure 5.4. Transformation to Hessenberg form.

The total number of operations is given by

$$\begin{aligned}
&\sum_{k=1}^{n-2} (8n(n-k) + 4(n-k)^2 + 4(n-k) + 6) \\
&\qquad = (8n+4) \sum_{k=1}^{n-2} (n-k) + 4 \sum_{k=1}^{n-2} (n-k)^2 + 6(n-2) \\
&\qquad \le (8n+4) \left(\frac{n(n-1)}{2} - 1 \right) + \frac{4}{6} n(n-1)(2n-1) + 6(n-2)
\end{aligned}$$

$$\leq \left(4n + 2 + \frac{2}{3}(2n-1)\right) n(n-1) = \left(\frac{16}{3}n + \frac{4}{3}\right)(n^2 - n)$$
$$= \frac{16}{3}n^3 + \frac{4}{3}n^2 - \frac{16}{3}n^2 - \frac{4}{3}n \leq \frac{16}{3}n^3. \qquad \square$$

Exercise 5.8 (Improved computation of Q). Assume that we start the transformation to Hessenberg form in the algorithm of Figure 5.4 with $Q = I$. Prove that in this case the number of operations can be reduced by $2n^3/3$ without changing the result. (Hint: It may be necessary to apply the Householder reflections in reversed order)

5.3 Shifting

The rate of convergence observed in Example 5.3 is too slow for a practical method: since the computational effort for each step of the iteration is proportional to n^2, we would like the number of steps to be as small as possible. We have already seen in Example 4.14 that shifting the spectrum can significantly improve the convergence, so we are interested in introducing a shift strategy for the QR iteration.

Since the QR method works with a full basis of $\mathbb{F}^n$, we can use shift strategies without the need for computing an inverse: if we have found a shift parameter $\mu \in \mathbb{F}$ sufficiently close to an eigenvalue $\lambda_n \in \sigma(A)$ and sufficiently far from all other eigenvalues, i.e., if we have

$$|\lambda_1 - \mu| \geq |\lambda_2 - \mu| \geq \cdots \geq |\lambda_{n-1} - \mu| > |\lambda_n - \mu|,$$

applying (5.5) to $k = n-1$ and the matrix $A - \mu I$ yields that the off-diagonal elements in the n-th row will converge to zero at a rate of $|\lambda_n - \mu|/|\lambda_{n-1} - \mu|$.

We can easily introduce the shift to the QR iteration: let $m \in \mathbb{N}_0$. One step of the simultaneous iteration for $A - \mu I$ is given by

$$Q_\mu^{(m+1)} R_\mu^{(m+1)} = (A - \mu I) Q_\mu^{(m)},$$

and using again

$$Q_\mu^{(m+1)} = Q_\mu^{(m)} \widehat{Q}_\mu^{(m+1)}$$

we obtain

$$Q_\mu^{(m)} \widehat{Q}_\mu^{(m+1)} R_\mu^{(m+1)} = (A - \mu I) Q_\mu^{(m)}.$$

Multiplication by $(Q_\mu^{(m)})^*$ yields

$$\widehat{Q}_\mu^{(m+1)} R_\mu^{(m+1)} = (Q_\mu^{(m)})^* (A - \mu I) Q_\mu^{(m)} = (Q_\mu^{(m)})^* A Q_\mu^{(m)} - \mu I.$$

We denote the iteration matrices by

$$A_\mu^{(m)} := (Q_\mu^{(m)})^* A Q_\mu^{(m)} \quad \text{for all } m \in \mathbb{N}_0$$

and write the first half of the shifted QR step in the form

$$\widehat{Q}_\mu^{(m+1)} R_\mu^{(m+1)} = A_\mu^{(m)} - \mu I. \tag{5.10}$$

In order to compute the next iteration matrix $A_\mu^{(m+1)}$, we consider

$$\begin{aligned} A_\mu^{(m+1)} &= (Q_\mu^{(m+1)})^* A Q_\mu^{(m+1)} = (\widehat{Q}_\mu^{(m+1)})^* (Q_\mu^{(m)})^* A Q_\mu^{(m)} \widehat{Q}_\mu^{(m+1)} \\ &= (\widehat{Q}_\mu^{(m+1)})^* A_\mu^{(m)} \widehat{Q}_\mu^{(m+1)} = (\widehat{Q}_\mu^{(m+1)})^* (A_\mu^{(m)} - \mu I) \widehat{Q}_\mu^{(m+1)} + \mu I \\ &= (\widehat{Q}_\mu^{(m+1)})^* \widehat{Q}_\mu^{(m+1)} R_\mu^{(m+1)} \widehat{Q}_\mu^{(m+1)} + \mu I \\ &= R_\mu^{(m+1)} \widehat{Q}_\mu^{(m+1)} + \mu I. \end{aligned} \tag{5.11}$$

Combining (5.10) and (5.11) yields the definition of the shifted QR iteration:

$$\begin{aligned} \widehat{Q}_\mu^{(m+1)} R_\mu^{(m+1)} &= A_\mu^{(m)} - \mu I, \\ A_\mu^{(m+1)} &= R_\mu^{(m+1)} \widehat{Q}_\mu^{(m+1)} + \mu I \quad \text{for all } m \in \mathbb{N}_0. \end{aligned}$$

As we can see, the only difference between the shifted QR iteration and the original one consists of subtracting μ from the diagonal elements prior to computing the QR factorization and adding μ after the multiplication. Compared to the overall complexity of the QR iteration, the additional operations required for the shift are negligible.

As in the case of the Rayleigh iteration, we can choose a different shift parameter for each step of the iteration, and doing so promises quadratic or even cubic convergence of the resulting shifted QR iteration. In the following, we will write $A^{(m)}$ instead of $A_\mu^{(m)}$ in order to keep the notation simple if the choice of shift parameter is clear from the context in order to keep the notation simple.

It is customary to try to choose the shift parameter to speed up convergence of the last row of the iteration matrices. According to (5.5), this means that we should ensure that the quotient $|\lambda_n - \mu|/|\lambda_{n-1} - \mu|$ is as small as possible. If the subdiagonal elements of the row are sufficiently close to zero, the diagonal element $a_{nn}^{(m)}$ is close to the eigenvalue λ_n, so it is a natural choice for the shift parameter.

Definition 5.9 (Rayleigh shift). The *Rayleigh shift strategy* is given by choosing the shift parameter in the $(m+1)$-th step of the QR iteration as $\mu = a_{nn}^{(m)}$.

The name "Rayleigh shift" is motivated by the close connection to the Rayleigh quotient introduced in Definition 4.4: if we denote the last column of $Q^{(m)}$ by $q^{(m)}$, we have

$$a_{nn}^{(m)} = ((Q^{(m)})^* A Q^{(m)})_{nn} = \langle A q^{(m)}, q^{(m)} \rangle = \Lambda_A(q^{(m)})$$

due to the definition of $A^{(m)}$ and $\|q^{(m)}\| = 1$. This equation suggests that we can expect the same good convergence behaviour as for the Rayleigh iteration described in Section 4.5.

We are interested in choosing a shift parameter that is close to an eigenvalue. In the case of the Rayleigh quotient, we have used the eigenvalue of the lower right 1×1 principal submatrix of $A^{(m)}$ as an approximation, so it is straightforward to generalize the strategy to consider larger principal submatrices. The case of the lower right 2×2 submatrix is particularly useful: we are looking for the eigenvalues of

$$\widehat{A} = \begin{pmatrix} a_{n-1,n-1}^{(m)} & a_{n-1,n}^{(m)} \\ a_{n,n-1}^{(m)} & a_{nn}^{(m)} \end{pmatrix}.$$

They can be determined by computing the zeros of the characteristic polynomial

$$\begin{aligned} p_{\widehat{A}}(t) &= \det(tI - \widehat{A}) = (t - \hat{a}_{11})(t - \hat{a}_{22}) - \hat{a}_{12}\hat{a}_{21} \\ &= t^2 - (\hat{a}_{11} + \hat{a}_{22})t + \hat{a}_{11}\hat{a}_{22} - \hat{a}_{12}\hat{a}_{21} \\ &= t^2 - (\hat{a}_{11} + \hat{a}_{22})t + \left(\frac{\hat{a}_{11} + \hat{a}_{22}}{2}\right)^2 - \left(\frac{\hat{a}_{11} + \hat{a}_{22}}{2}\right)^2 + \hat{a}_{11}\hat{a}_{22} - \hat{a}_{12}\hat{a}_{21} \\ &= \left(t - \frac{\hat{a}_{11} + \hat{a}_{22}}{2}\right)^2 - \left(\frac{\hat{a}_{11} - \hat{a}_{22}}{2}\right)^2 - \hat{a}_{12}\hat{a}_{21}, \end{aligned}$$

which are given by

$$\sigma(\widehat{A}) = \left\{ \frac{\hat{a}_{11} + \hat{a}_{22}}{2} \pm \sqrt{\left(\frac{\hat{a}_{11} - \hat{a}_{22}}{2}\right)^2 + \hat{a}_{12}\hat{a}_{21}} \right\}.$$

Since we are interested in convergence of the last row, we should choose the eigenvalue in $\sigma(\widehat{A})$ that is closest to the n-th diagonal element.

Definition 5.10 (Francis single shift). The *Francis single shift strategy* is given by choosing the element closest to $a_{nn}^{(m)}$ in the set

$$\sigma(\widehat{A}) = \left\{ \frac{a_{n-1,n-1}^{(m)} + a_{nn}^{(m)}}{2} \pm \sqrt{\left(\frac{a_{n-1,n-1}^{(m)} - a_{nn}^{(m)}}{2}\right)^2 + a_{n-1,n}^{(m)} a_{n,n-1}^{(m)}} \right\}$$

as the shift parameter in the $(m+1)$-th step of the QR iteration.

If we are dealing with a real matrix, the Francis shift strategy may yield a complex-valued shift parameter, and this would make the next iteration matrix also complex-valued. Since eigenvalues of real matrices always appear in complex-conjugate pairs, it is possible to avoid working with complex values by first performing a QR step with the Francis shift μ and following it by a QR step with its complex conjugate $\bar{\mu}$. Due to

$$(A - \mu I)(A - \bar{\mu} I) = A^2 - (\mu + \bar{\mu})A + |\mu|^2 I,$$

the resulting iteration matrix is again real-valued. It is possible to arrange the computation in such a way that complex numbers are avoided entirely, this results in the *Francis double shift strategy* (cf. Example 5.17).

Example 5.11 (Failure to converge). As we have seen in the case of the Rayleigh iteration (cf. Proposition 4.13), shift strategies work best if the iteration vectors or matrices are sufficiently close to the result. The matrix

$$P := \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

taken from [32] illustrates this point: its characteristic polynomial is given by

$$p_P(t) = t^3 - 1,$$

so its eigenvalues are the third roots of unity. Since they are equal in modulus, we can expect no convergence unless a suitable shift parameter is chosen.

Unfortunately, both the Rayleigh and the Francis shift strategies choose the shift parameter $\mu = 0$.

5.4 Deflation

Due to (5.5), we can expect the matrices $A^{(m)}$ to converge to a matrix of the form

$$A^{(m)} \approx \begin{pmatrix} B & F \\ & C \end{pmatrix}, \quad B \in \mathbb{F}^{k\times k},\ D \in \mathbb{F}^{(n-k)\times(n-k)}. \tag{5.12}$$

Once $A^{(m)}$ is sufficiently close to this block-triangular form, we can stop performing iterations for the entire matrix and focus on handling the diagonal blocks B and C individually: if we can find unitary matrices $Q_B \in \mathbb{F}^{k\times k}$ and $Q_C \in \mathbb{F}^{(n-k)\times(n-k)}$ such that $Q_B^* B Q_B$ and $Q_C^* C Q_C$ are sufficiently close to upper triangular form, the same holds for

$$\begin{aligned} \begin{pmatrix} Q_B & \\ & Q_C \end{pmatrix}^* A_\mu^{(m)} \begin{pmatrix} Q_B & \\ & Q_C \end{pmatrix} &\approx \begin{pmatrix} Q_B & \\ & Q_C \end{pmatrix}^* \begin{pmatrix} B & F \\ & C \end{pmatrix} \begin{pmatrix} Q_B & \\ & Q_C \end{pmatrix} \\ &= \begin{pmatrix} Q_B^* B Q_B & Q_B^* F Q_C \\ & Q_C^* C Q_C \end{pmatrix}, \end{aligned}$$

and we have approximated the Schur decomposition.

This observation suggests a strategy known as *deflation*: we perform iterations until one of the subdiagonal blocks has become sufficiently small. Then we repeat the procedure for the individual diagonal blocks. If we can choose the shift parameters

correctly, we will apply the procedure to smaller and smaller diagonal blocks until we arrive at 1×1 blocks and have reached an approximation of the Schur decomposition.

Applying the deflation approach to Hessenberg matrices is particularly simple: the subdiagonal blocks contain only one non-zero entry, so we can easily check for convergence. In a practical deflation approach, we first look for the largest index $\alpha \in \{1, \ldots, n\}$ such that

$$a_{k+1,k}^{(m)} \approx 0 \quad \text{for all } k \in \{1, \ldots, \alpha - 1\}.$$

If $\alpha = n$, the matrix $A^{(m)}$ is sufficiently close to upper triangular form and we can stop the iteration. Otherwise, we have $a_{\alpha+1,\alpha}^{(m)} \not\approx 0$ due to the maximality of α, and we look for the largest $\beta \in \{\alpha, \ldots, n\}$ such that

$$a_{k+1,k}^{(m)} \not\approx 0 \quad \text{for all } k \in \{\alpha, \ldots, \beta - 1\}.$$

If $\beta < n$, we have $a_{\beta+1,\beta}^{(m)} \approx 0$ due to the maximality of β.

We split the matrix $A^{(m)}$ in the form

$$A^{(m)} \approx \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ & H_{22} & H_{23} \\ & & H_{33} \end{pmatrix}, \quad \begin{matrix} H_{11} \in \mathbb{F}^{(\alpha-1)\times(\alpha-1)}, \quad H_{22} \in \mathbb{F}^{(\beta-\alpha+1)\times(\beta-\alpha+1)}, \\ H_{33} \in \mathbb{F}^{(n-\beta)\times(n-\beta)}. \end{matrix}$$

Due to our choice of α, H_{11} is sufficiently close to triangular form and the block below H_{11} is sufficiently close to zero. Due to our choice of β, the block below H_{22} is also sufficiently close to zero. H_{22} is a Hessenberg matrix, and we can efficiently apply the QR iteration to try to eliminate its subdiagonal entries. Once we have succeeded, H_{11} and H_{22} are upper triangular and we can apply the same procedure to the remaining matrix H_{33}.

The resulting algorithm is given in Figure 5.5. In order to decide whether a condition of the form "$a_{\alpha+1,\alpha}^{(m)} \approx 0$" is satisfied in practice, a condition like

$$|a_{\alpha+1,\alpha}^{(m)}| \leq \epsilon(|a_{\alpha\alpha}^{(m)}| + |a_{\alpha+1,\alpha+1}^{(m)}|)$$

is frequently used, where $\epsilon > 0$ denotes a relative error tolerance. A typical choice for ϵ depends on the machine's floating point accuracy ϵ_{mach}. Common choices are $\epsilon = c\epsilon_{\text{mach}}$ for a small constant c or $\epsilon = \sqrt{n}\epsilon_{\text{mach}}$, taking into account the matrix dimension.

Example 5.12 (QR iteration). We apply the QR iteration with Francis and Rayleigh shift to the matrix A given in Example 5.3. The program gives the following results:

```
procedure qr_deflation(var A, Q);
begin
  α ← 1;   while α < n and a_{α+1,α} ≈ 0 do α ← α + 1;
  β ← α + 1;   while β < n and a_{β+1,β} ≉ 0 do β ← β + 1;
  Choose a shift parameter μ;
  for k = α to β do a_{kk} ← a_{kk} − μ;
  for k = α to β − 1 do        { Find factorization A − μI = Q̂R }
    givens_setup(a_{kk}, a_{k+1,k}, c_k, s_k);
    for j ∈ {k + 1, ..., n} do givens_apply(c_k, s_k, a_{kj}, a_{k+1,j})
  end;
  for k = α to β − 1 do begin        { Compute RQ̂ and QQ̂ }
    for i ∈ {1, ..., k} do givens_conjapply(c_k, s_k, a_{ik}, a_{i,k+1});
    a_{k+1,k} ← −s_k a_{k+1,k+1};   a_{k+1,k+1} ← c̄_k a_{k+1,k+1};
    for i ∈ {1, ..., n} do givens_conjapply(c_k, s_k, q_{ik}, q_{i,k+1})
  end;
  for k = α to β do a_{kk} ← a_{kk} + μ
end
```

Figure 5.5. One step of the Hessenberg QR iteration with shift and deflation.

	Rayleigh			Francis		
m	$\lvert a_{21}^{(m)}\rvert$	$\lvert a_{32}^{(m)}\rvert$	$\lvert a_{43}^{(m)}\rvert$	$\lvert a_{21}^{(m)}\rvert$	$\lvert a_{32}^{(m)}\rvert$	$\lvert a_{43}^{(m)}\rvert$
1	6.78×10^{-1}	6.23×10^{-1}	conv.	6.78×10^{-1}	6.23×10^{-1}	conv.
2	3.86×10^{-1}	7.14×10^{-2}		3.66×10^{-1}	conv.	
3	1.73×10^{-1}	4.51×10^{-3}		conv.		
4	7.68×10^{-2}	2.32×10^{-5}				
5	3.60×10^{-2}	5.96×10^{-10}				
6	1.74×10^{-2}	conv.				
7	2.73×10^{-4}					
8	7.27×10^{-8}					
9	conv.					

We can see that the Rayleigh shift strategy seems to lead to quadratic convergence as soon as the subdiagonal entries are sufficiently close to zero. With the Francis shift strategy, the iteration requires only three steps to converge. For practical problems, the difference between both strategies can be expected to be far less pronounced.

5.5 Implicit iteration

The QR iteration presented in Figure 5.5 is still slightly inelegant: we have to subtract and add the shift parameter explicitly, and we have to store the values c_k and s_k corre-

sponding to the Givens rotations used during the factorization process. In this section, we derive an alternative version of the QR iteration that avoids these inconveniences and facilitates the implementation of multiple shift strategies like the Francis double shift or even more sophisticated techniques that can lead to significant improvements of the basic algorithm [50, 4, 5].

We have seen that the QR iteration preserves the Hessenberg form of the iteration matrices $A^{(m)}$: if $A^{(m)}$ is a Hessenberg matrix, then so is the next iteration matrix $A^{(m+1)} = (\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)}$. Unitarily similar Hessenberg matrices like $A^{(m)}$ and $A^{(m+1)}$ have special properties that we can use to simplify the algorithm: if we have two unitarily similar Hessenberg matrices with non-zero subdiagonal entries and if the first column of the corresponding similarity transformation matrix is the first canonical unit vector, then both matrices can only differ in the signs of their entries.

Theorem 5.13 (Implicit Q). *Let $H, J \in \mathbb{F}^{n\times n}$ be Hessenberg matrices and let $Q \in \mathbb{F}^{n\times n}$ be unitary such that*

$$J = Q^* H Q. \tag{5.13}$$

Let $m \in \{0, \ldots, n-1\}$ satisfy

$$j_{k+1,k} \neq 0 \quad \text{for all } k \in \{1, \ldots, m\}.$$

If the first column of Q is the canonical unit vector δ_1, the first $m+1$ columns of Q are multiples of the first $m+1$ canonical unit vectors.

Proof. (cf. [18, Theorem 7.4.2]) We prove by induction on $k \in \{1, \ldots, n\}$ that the k-th column $q_k := Q\delta_k$ of Q is a multiple of the k-th canonical unit vector.

For $k = 1$, this is given. Let now $k \in \{1, \ldots, m\}$ be such that $q_1, \ldots, q_k$ are multiples of the corresponding unit vectors. This implies

$$q_\ell = \delta_\ell q_{\ell\ell} \quad \text{for all } \ell \in \{1, \ldots, k\}. \tag{5.14}$$

We write (5.13) in the form

$$QJ = HQ$$

and multiply by the k-th canonical unit vector to obtain

$$\sum_{\ell=1}^{k+1} q_\ell j_{\ell k} = Q \sum_{\ell=1}^{k+1} \delta_\ell j_{\ell k} = QJ\delta_k = HQ\delta_k = Hq_k = H\delta_k q_{kk} = \sum_{\ell=1}^{k+1} \delta_\ell h_{\ell k} q_{kk}.$$

We apply (5.14) to the left-hand side to obtain

$$q_{k+1} j_{k+1,k} + \sum_{\ell=1}^{k} \delta_\ell q_{\ell\ell} j_{\ell k} = \sum_{\ell=1}^{k+1} \delta_\ell h_{\ell k} q_{kk},$$

and rearranging the terms yields

$$q_{k+1} j_{k+1,k} = \delta_{k+1} h_{k+1,k} q_k + \sum_{\ell=1}^{k} \delta_\ell (h_{\ell k} q_{kk} - q_{\ell\ell} j_{\ell k}).$$

Due to $j_{k+1,k} \neq 0$, we find

$$q_{k+1} = \frac{1}{j_{k+1,k}} \left(\delta_{k+1} h_{k+1,k} q_k + \sum_{\ell=1}^{k} \delta_\ell (h_{\ell k} q_{kk} - q_{\ell\ell} j_{\ell k}) \right),$$

i.e., $q_{k+1} \in \operatorname{span}\{\delta_1, \ldots, \delta_{k+1}\}$ and therefore

$$(q_{k+1})_\ell = 0 \quad \text{for all } \ell \in \{k+2, \ldots, n\}. \tag{5.15}$$

Using (5.14) again yields

$$(q_{k+1})_\ell = \langle q_{k+1}, \delta_\ell \rangle = \langle q_{k+1}, q_\ell / q_{\ell\ell} \rangle = 0 \quad \text{for all } \ell \in \{1, \ldots, k\} \tag{5.16}$$

since Q is unitary. Combining (5.15) and (5.16) yields that q_{k+1} has to be a multiple of δ_{k+1}. □

We aim to use this result to simplify the QR iteration: the QR iteration takes a Hessenberg matrix $A^{(m)}$ and computes a unitarily similar Hessenberg matrix $A^{(m+1)}$. If we can find a new algorithm that also yields a unitarily similar Hessenberg matrix $\widetilde{A}^{(m+1)}$ and if the first column of both similarity transformations are equal, Theorem 5.13 yields that $A^{(m+1)}$ and $\widetilde{A}^{(m+1)}$ can differ only in the signs of the coefficients.

Corollary 5.14 (Hessenberg similarity). *Let $A, B, \widetilde{B} \in \mathbb{F}^{n\times n}$ be Hessenberg matrices. Let $Q, \widetilde{Q} \in \mathbb{F}^{n\times n}$ be unitary matrices such that*

$$B = Q^* A Q, \quad \widetilde{B} = \widetilde{Q}^* A \widetilde{Q}.$$

Let the first columns of Q and $\widetilde{Q}$ be identical and let $m \in \{0, \ldots, n-1\}$ satisfy

$$b_{k,k+1} \neq 0 \quad \text{for all } k \in \{1, \ldots, m\}. \tag{5.17}$$

Then there are a unitary diagonal matrix $D \in \mathbb{F}^{(m+1)\times(m+1)}$ and a unitary matrix $\widehat{P} \in \mathbb{F}^{(n-m-1)\times(n-m-1)}$ such that

$$\widetilde{B} = \begin{pmatrix} D & \\ & \widehat{P} \end{pmatrix} B \begin{pmatrix} D & \\ & \widehat{P} \end{pmatrix}^*.$$

Proof. In order to be able to apply Theorem 5.13, we define $P := \widetilde{Q}^* Q$ and observe

$$\begin{aligned} A &= \widetilde{Q}\widetilde{B}\widetilde{Q}^*, \\ B &= Q^* A Q = Q^*\widetilde{Q}\widetilde{B}\widetilde{Q}^* Q = P^*\widetilde{B}P. \end{aligned}$$

Since the first columns of Q and $\widetilde{Q}$ are identical, the first column of P has to be the first canonical unit vector.

Theorem 5.13 yields that the first $m+1$ columns of P are multiples of the first $m+1$ canonical unit vectors, so we can find a diagonal matrix $D \in \mathbb{F}^{(m+1)\times(m+1)}$ and matrices $\widehat{P} \in \mathbb{F}^{(n-m-1)\times(n-m-1)}$ and $R \in \mathbb{F}^{(m+1)\times(n-m-1)}$ such that

$$P = \begin{pmatrix} D & R \\ & \widehat{P} \end{pmatrix}.$$

Since P is the product of unitary matrices, we have

$$\begin{aligned} \begin{pmatrix} I_{m+1} & \\ & I_{n-m-1} \end{pmatrix} = I = P^* P &= \begin{pmatrix} D^* & \\ R^* & \widehat{P}^* \end{pmatrix} \begin{pmatrix} D & R \\ & \widehat{P} \end{pmatrix} \\ &= \begin{pmatrix} D^* D & D^* R \\ R^* D & R^* R + \widehat{P}^* \widehat{P} \end{pmatrix}, \end{aligned}$$

i.e., $D^* D = I$, $D^* R = 0$ and $R^* R + \widehat{P}^* \widehat{P} = I$. The first equation implies that D is unitary. Since unitary matrices are invertible, the second equation yields $R = 0$. Since R is equal to zero, the third equation gives us $\widehat{P}^* \widehat{P} = I$, i.e., $\widehat{P}$ is also unitary. □

The Hessenberg QR step discussed in Section 5.2 consists of first finding Givens rotations $G_1, \ldots, G_{n-1}$ that render $A^{(m)}$ upper triangular

$$G_{n-1} \ldots G_1 A^{(m)} = R^{(m+1)}, \quad Q^{(m+1)} := G_1^* \ldots G_{n-1}^*$$

and then multiplying by their adjoints in reversed order to compute the next iteration matrix

$$\begin{aligned} A^{(m+1)} &= R^{(m+1)} Q^{(m+1)} = (Q^{(m+1)})^* A^{(m)} Q^{(m+1)} \\ &= G_{n-1} \ldots G_1 A^{(m)} G_1^* \ldots G_{n-1}^*. \end{aligned}$$

In our construction, only the first rotation G_1 can change the first component of a vector, therefore the first column of $Q^{(m+1)}$ depends only on G_1. If we can construct an alternate sequence of unitary transformations $\widetilde{G}_2, \ldots, \widetilde{G}_{n-1}$ that do not change the first component, the first columns of

$$Q^{(m+1)} = G_1^* G_2^* \ldots G_{n-1}^* \quad \text{and} \quad \widetilde{Q}^{(m+1)} = G_1^* \widetilde{G}_2^* \ldots \widetilde{G}_{n-1}^*$$

have to be equal. If we can ensure that

$$\widetilde{A}^{(m+1)} := (\widetilde{Q}^{(m+1)})^* A^{(m)} \widetilde{Q}^{(m+1)}$$

is also a Hessenberg matrix, Corollary 5.14 applies and guarantees that $A^{(m+1)}$ and $\widetilde{A}^{(m+1)}$ differ only in the signs of the coefficients, while the convergence of the resulting iteration remains unchanged. The condition (5.17) can be taken care of by using the deflation approach introduced in Section 5.4: as soon as a subdiagonal entry vanishes, the matrix is split into submatrices with non-zero subdiagonal entries and the submatrices are processed independently.

Our goal is now to construct the matrices $\widetilde{G}_2, \dots, \widetilde{G}_{n-1}$. We illustrate the procedure using a Hessenberg matrix of the form

$$H := A^{(m)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

Applying the first Givens rotation G_1 to H yields matrices of the form

$$G_1 H = \begin{pmatrix} \otimes & \otimes & \otimes & \otimes & \otimes \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix},$$

$$H^{(1)} := G_1 H G_1^* = \begin{pmatrix} \otimes & \otimes & \times & \times & \times \\ \otimes & \otimes & \times & \times & \times \\ \boxtimes & \otimes & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

We can see that $H^{(1)}$ is "almost" a Hessenberg matrix, only the coefficient $h_{31}^{(1)}$ in the first column of the third row poses a problem. Fortunately, we can apply a Givens rotation $\widetilde{G}_2$ to the second and third row to eliminate this entry and obtain

$$\widetilde{G}_2 H^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix},$$

$$H^{(2)} := \widetilde{G}_2 H^{(1)} \widetilde{G}_2^* = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & \boxtimes & \otimes & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

The entry in the first column is gone, unfortunately we have introduced a non-zero entry $h_{42}^{(2)}$ in the second column of the fourth row. We can apply a Givens rotation $\widetilde{G}_3$ to the third and fourth row to get rid of this entry and find

$$\widetilde{G}_3 H^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & & & \times & \times \end{pmatrix},$$

$$H^{(3)} := \widetilde{G}_3 H^{(2)} \widetilde{G}_3^* = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & \boxtimes & \otimes & \times \end{pmatrix}.$$

Once more only one entry causes problems, now the entry $h_{53}^{(3)}$ in the third column of the fifth row. We already know what to do: a Givens rotation $\widetilde{G}_4$ for the fourth and fifth row eliminates the offending entry:

$$\widetilde{G}_4 H^{(3)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \end{pmatrix},$$

$$H^{(4)} := \widetilde{G}_4 H^{(3)} \widetilde{G}_4^* = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \end{pmatrix}.$$

We can see that we have achieved our goal:

$$\widetilde{A}^{(m+1)} := H^{(4)} = \widetilde{G}_{n-1} \dots \widetilde{G}_2 G_1 A^{(m)} G_1^* \widetilde{G}_2^* \dots \widetilde{G}_{n-1}^*$$

is a Hessenberg matrix. This technique is called "bulge chasing", since the matrices $H^{(1)}, \dots, H^{(n-2)}$ differ from Hessenberg form only by the "bulge" marked as $\boxtimes$ consisting of one non-zero entry and this bulge is "chased" down and to the right until it "drops out of the matrix" and we have reached Hessenberg form.

```
procedure qr_implicit(var A, Q);
begin
  α ← 1;   while α < n and a_{α+1,α} ≈ 0 do α ← α + 1;
  if α = n then return;
  β ← α + 1;   while β < n and a_{β+1,β} ≉ 0 do β ← β + 1;
  Choose a shift parameter μ;
  a'_{αα} ← a_{αα} − μ;   a'_{α+1,α} ← a_{α+1,α};
  givens_setup(a'_{αα}, a'_{α+1,α}, c, s);
  for j ∈ {α, ..., n} do givens_apply(c, s, a_{αj}, a_{α+1,j});
  for i ∈ {1, ..., α + 1} do givens_conjapply(c, s, a_{iα}, a_{i,α+1});
  for i ∈ {1, ..., n} do givens_conjapply(c, s, q_{iα}, q_{i,α+1});
  for k = α to β − 2 do begin
    b ← −s a_{k+2,k+1};   a_{k+2,k+1} ← c̄ a_{k+2,k+1};      { Update row k + 2 }
    givens_setup(a_{k+1,k}, b, c, s);      { Givens rotation to eliminate bulge }
    for j ∈ {k + 1, ..., n} do givens_apply(c, s, a_{k+1,j}, a_{k+2,j});
    for i ∈ {1, ..., k + 2} do givens_conjapply(c, s, a_{i,k+1}, a_{i,k+2});
    for i ∈ {1, ..., n} do givens_conjapply(c, s, q_{i,k+1}, q_{i,k+2})
  end
end
```

Figure 5.6. One step of the implicit QR iteration with shift and deflation.

Since the rotations $\widetilde{G}_2, \ldots, \widetilde{G}_{n-1}$ do not change the first row, the conditions of Corollary 5.14 are fulfilled and we have computed an iteration matrix $\widetilde{A}^{(m+1)}$ that can replace $A^{(m+1)}$ without changing the important properties of our algorithm (as long as deflation is used). The resulting algorithm is called the *implicit QR method* and offers several advantages, most importantly an elegant way of handling shifts: for the shifted QR iteration, the next iteration matrix is given by

$$\begin{aligned} \widehat{Q}^{(m+1)} R^{(m+1)} &= A^{(m)} - \mu I, \\ A^{(m+1)} &= R^{(m+1)} \widehat{Q}^{(m+1)} + \mu I = (\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)}, \end{aligned}$$

and since adding and subtracting the identity does not change the Hessenberg form, we can also apply our construction with a small change: we compute the first Givens rotation G_1 not for the matrix $A^{(m)}$, but for the shifted matrix $A^{(m)} - \mu I$. Then we apply it to $A^{(m)}$ and proceed as before to regain Hessenberg form. For the implicit approach, the only difference between the original and the shifted iteration lies in the choice of the first Givens rotation. Combining the shifted implicit iteration with deflation leads to the algorithm given in Figure 5.6 that performs one step of the practical implicit QR iteration.

Proposition 5.15 (Complexity). *The algorithm given in Figure* 5.6 *requires not more than*

$$12n^2 + 8n \quad \textit{operations}$$

to perform a step of the implicit QR iteration. The estimate does not include the operations required for finding an appropriate shift parameter.

Proof. We require 8 operations to compute the first Givens rotation and $6(n - \alpha + 1)$ operations to apply it to the rows as well as $6(\alpha + 1)$ operations to apply it to the columns, for a total of

$$8 + 6(n - \alpha + 1 + \alpha + 1) = 8 + 6(n + 2) = 20 + 6n$$

operations. Applying the rotation to the columns of Q takes additional

$$6n$$

operations. For each $k \in \{\alpha, \dots, \beta - 2\}$, updating the $(k + 2)$-th row takes 2 operations, determining the Givens rotation to eliminate the bulge takes 6 operations, applying this rotation to the rows requires $6(n - k)$ operations, and applying it to the column requires $6(k + 2)$, for a total of

$$8 + 6(n - k + k + 2) = 8 + 6(n + 2) = 20 + 6n$$

operations. Applying the rotation to the columns of Q again takes additional

$$6n$$

operations. In the worst case, all subdiagonal entries are non-zero and we have $\alpha = 1$ and $\beta = n$, so the entire algorithm requires not more than

$$\begin{aligned} 20 + 12n + \sum_{k=\alpha}^{\beta-2} (20 + 12n) &= (20 + 12n)(1 + \beta - 2 - \alpha + 1) \\ &= (20 + 12n)(\beta - \alpha) \\ &= (20 + 12n)(n - 1) \leq 20n + 12n^2 - 20 - 12n \\ &= 12n^2 + 8n - 20 < 12n^2 + 8n \end{aligned}$$

operations. □

Although the number of operations for the implicit algorithm is slightly higher than in the case of the original algorithm given in Figure 5.5, its practical performance is frequently superior, e.g., in the self-adjoint case, A is a tridiagonal matrix, and the bulge-chasing algorithm works its way from the upper left to the lower right along the diagonal. If the tridiagonal matrix is stored appropriately, this operation can benefit from the cache memory of modern processors.

Remark 5.16 (Structured eigenvalue problems). Frequently, the matrix under investigation exhibits additional structure, e.g., it may be Hamiltonian, Skew–Hamiltonian or the product of two matrices. In these cases, it is possible to derive special variants of the QR iteration that can take advantage of the structure [24, 49].

5.6 Multiple-shift strategies *

The most important feature of the implicit QR iteration is that this approach facilitates the use of sophisticated multiple-shift strategies: we consider two consecutive steps of the iteration with shift values of μ_{m+1} and μ_{m+2}, given by

$$\begin{aligned}\widehat{Q}^{(m+1)}R^{(m+1)} &= A^{(m)} - \mu_{m+1}I, & A^{(m+1)} &= (\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)},\\ \widehat{Q}^{(m+2)}R^{(m+2)} &= A^{(m+1)} - \mu_{m+2}I, & A^{(m+2)} &= (\widehat{Q}^{(m+2)})^* A^{(m+1)} \widehat{Q}^{(m+2)}.\end{aligned}$$

We are looking for a way to get from $A^{(m)}$ to $A^{(m+2)}$ without computing $A^{(m+1)}$. We have

$$\begin{aligned}A^{(m+2)} &= (\widehat{Q}^{(m+2)})^* A^{(m+1)} \widehat{Q}^{(m+2)}\\ &= (\widehat{Q}^{(m+2)})^* (\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)} \widehat{Q}^{(m+2)}\\ &= (\widehat{Q}^{(m+1)} \widehat{Q}^{(m+2)})^* A^{(m)} \widehat{Q}^{(m+1)} \widehat{Q}^{(m+2)} = (\widehat{Q}_{\mathrm{dbl}}^{(m+2)})^* A^{(m)} \widehat{Q}_{\mathrm{dbl}}^{(m+2)}\end{aligned}$$

with the auxiliary unitary matrix

$$\widehat{Q}_{\mathrm{dbl}}^{(m+2)} := \widehat{Q}^{(m+1)} \widehat{Q}^{(m+2)}.$$

In order to reach our goal, we have to find this matrix without computing $A^{(m+1)}$ or $Q^{(m+1)}$. We know that the QR iteration is closely related to the simultaneous iteration, so it is reasonable to consider the result of performing two steps of the simultaneous iteration at once, i.e., to investigate the product matrix

$$\begin{aligned}&(A^{(m)} - \mu_{m+2}I)(A^{(m)} - \mu_{m+1}I)\\ &\quad= \widehat{Q}^{(m+1)}(\widehat{Q}^{(m+1)})^*(A^{(m)} - \mu_{m+2}I)\widehat{Q}^{(m+1)}(\widehat{Q}^{(m+1)})^*(A^{(m)} - \mu_{m+1}I)\\ &\quad= \widehat{Q}^{(m+1)}((\widehat{Q}^{(m+1)})^* A^{(m)} \widehat{Q}^{(m+1)} - \mu_{m+2}I)(\widehat{Q}^{(m+1)})^*(A^{(m)} - \mu_{m+1}I)\\ &\quad= \widehat{Q}^{(m+1)}(A^{(m+1)} - \mu_{m+2}I)R^{(m+1)}\\ &\quad= \widehat{Q}^{(m+1)} \widehat{Q}^{(m+2)} R^{(m+2)} R^{(m+1)}\\ &\quad= \widehat{Q}_{\mathrm{dbl}}^{(m+2)} R_{\mathrm{dbl}}^{(m+2)},\end{aligned}$$

where we have introduced the matrix

$$R_{\mathrm{dbl}}^{(m+2)} := R^{(m+2)} R^{(m+1)}.$$

Since $R^{(m+2)}$ and $R^{(m+1)}$ are upper triangular, their product $R_{\text{dbl}}^{(m+2)}$ shares this property (cf. Lemma 4.29) and the matrices $\widehat{Q}_{\text{dbl}}^{(m+2)}$ and $R_{\text{dbl}}^{(m+2)}$ define a QR decomposition

$$\widehat{Q}_{\text{dbl}}^{(m+2)} R_{\text{dbl}}^{(m+2)} = (A^{(m)} - \mu_{m+2} I)(A^{(m)} - \mu_{m+1} I) \tag{5.18}$$

of the product matrix. As long as μ_{m+1} and μ_{m+2} are not eigenvalues, the right-hand side matrix is invertible and this equation can be used to construct a matrix that differs from $\widehat{Q}_{\text{dbl}}^{(m+2)}$ only by the signs of its columns. If one of the shift parameters happens to be an eigenvalue, a subdiagonal entry of $A^{(m+2)}$ vanishes and we can use deflation.

We can construct $\widehat{Q}_{\text{dbl}}^{(m+2)}$ using Givens rotations or Householder reflections: we consider again the case of 5×5 matrices. Since $A^{(m)} - \mu_{m+1} I$ and $A^{(m)} - \mu_{m+2} I$ are Hessenberg matrices, their product has the form

$$B^{(0)} := (A^{(m)} - \mu_{m+2} I)(A^{(m)} - \mu_{m+1}) = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \end{pmatrix}.$$

We choose a Householder reflection M_1 that changes the first three rows to eliminate the entries $b_{21}^{(0)}$ and $b_{31}^{(0)}$:

$$B^{(1)} := M_1 B^{(0)} = \begin{pmatrix} \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \times \\ & & \times & \times & \times \end{pmatrix}.$$

A second Householder reflection M_2 works on the rows two to four to eliminate $b_{32}^{(1)}$ and $b_{42}^{(1)}$:

$$B^{(2)} := M_2 B^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & & \times & \times & \times \end{pmatrix}.$$

In the same way, we can construct reflections M_3 and M_4 acting on the rows three to five and four to five, respectively, that complete the transformation to upper triangular

form:

$$B^{(3)} := M_3 B^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \\ & & 0 & \otimes & \otimes \end{pmatrix},$$

$$B^{(4)} := M_4 B^{(3)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \end{pmatrix}.$$

We let $\widehat{Q}_{\mathrm{dbl}}^{(m+2)} := M_1^* \dots M_{n-1}^*$ and $R_{\mathrm{dbl}} := B^{(n-1)}$ and have found a QR factorization.

In order to construct an implicit method, we can once again use Corollary 5.14: we already know that $A^{(m+2)} = (\widehat{Q}_{\mathrm{dbl}}^{(m+2)})^* A^{(m)} \widehat{Q}_{\mathrm{dbl}}^{(m+2)}$ and $A^{(m)}$ are Hessenberg matrices. If we can find a unitary matrix $\widetilde{Q}_{\mathrm{dbl}}^{(m+2)}$ such that $\widetilde{A}^{(m+2)} := (\widetilde{Q}_{\mathrm{dbl}}^{(m+2)})^* A^{(m)} \widetilde{Q}_{\mathrm{dbl}}^{(m+2)}$ is a Hessenberg matrix and the first columns of $\widetilde{Q}_{\mathrm{dbl}}^{(m+2)}$ and $\widehat{Q}_{\mathrm{dbl}}^{(m+2)}$ are identical, we can replace $A^{(m+2)}$ by $\widetilde{A}^{(m+2)}$ without harming the convergence of the QR iteration.

We again employ bulge chasing to construct the matrix $\widetilde{Q}_{\mathrm{dbl}}^{(m+2)}$: we aim for $\widetilde{Q}_{\mathrm{dbl}}^{(m+2)} = M_1^* \widetilde{M}_2^* \dots \widetilde{M}_{n-1}^*$, where the unitary transformations $\widetilde{M}_2, \dots, \widetilde{M}_{n-1}$ do not change the first row and ensure that $\widetilde{A}^{(m+2)}$ is a Hessenberg matrix. We start with the Hessenberg matrix

$$H := A^{(m)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}$$

and apply the first Householder reflection M_1. Since it works on the first three rows, we end up with a larger bulge:

$$M_1 H = \begin{pmatrix} \otimes & \otimes & \otimes & \otimes & \otimes \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix},$$

$$H^{(1)} := M_1 H M_1^* = \begin{pmatrix} \otimes & \otimes & \otimes & \times & \times \\ \otimes & \otimes & \otimes & \times & \times \\ \boxtimes & \otimes & \otimes & \times & \times \\ \boxtimes & \boxtimes & \otimes & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

We can eliminate the offending entries $h_{31}^{(1)}$ and $h_{41}^{(1)}$ by applying a Householder reflection $\widetilde{M}_2$ to the rows two to four and find

$$\widetilde{M}_2 H^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & & & \times & \times \end{pmatrix},$$

$$H^{(2)} := \widetilde{M}_2 H^{(1)} \widetilde{M}_2^* = \begin{pmatrix} \times & \otimes & \otimes & \otimes & \times \\ \times & \otimes & \otimes & \otimes & \times \\ & \otimes & \otimes & \otimes & \times \\ & \boxtimes & \otimes & \otimes & \times \\ & \boxtimes & \boxtimes & \otimes & \times \end{pmatrix}.$$

As in the case of the simple implicit QR iteration, we have managed to chase the bulge one step towards the right lower edge of the matrix. We apply a Householder reflection $\widetilde{M}_3$ to the rows three to five to eliminate $h_{42}^{(2)}$ and $h_{52}^{(2)}$:

$$\widetilde{M}_3 H^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \end{pmatrix},$$

$$H^{(3)} := \widetilde{M}_3 H^{(2)} \widetilde{M}_3^* = \begin{pmatrix} \times & \times & \otimes & \otimes & \otimes \\ \times & \times & \otimes & \otimes & \otimes \\ & \times & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \\ & & \boxtimes & \otimes & \otimes \end{pmatrix}.$$

Part of the bulge has already disappeared, now we only have to apply one last Householder reflection $\widetilde{M}_4$ to the fourth and fifth row to eliminate $h_{53}^{(3)}$ and get

$$\widetilde{M}_4 H^{(3)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \end{pmatrix},$$

$$H^{(4)} := \widetilde{M}_4 H^{(3)} \widetilde{M}_4^* = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \end{pmatrix}.$$

We have constructed

$$\widetilde{A}^{(m+2)} := H^{(n-1)} = \widetilde{M}_{n-1} \dots \widetilde{M}_2 M_1 A^{(m)} M_1^* \widetilde{M}_2^* \dots \widetilde{M}_{n-1}^*,$$

and this is indeed again a Hessenberg matrix.

Example 5.17 (Francis double shift). A double shift strategy like the one described here is particularly useful when dealing with real-valued matrices: although all matrix entries are real, the eigenvalues may be complex. In this case, they have to appear in conjugate pairs, so it makes sense to perform a double shift using both numbers, i.e., to choose $\mu_{m+2} = \bar{\mu}_{m+1}$. This approach is particularly elegant since the matrix

$$\begin{aligned} (A^{(m)} - \mu_{m+2} I)(A^{(m)} - \mu_{m+1} I) &= (A^{(m)})^2 - (\mu_{m+1} + \mu_{m+2}) A^{(m)} \\ &\quad + \mu_{m+2} \mu_{m+1} I \\ &= (A^{(m)})^2 - (\mu_{m+1} + \bar{\mu}_{m+1}) A^{(m)} \\ &\quad + \bar{\mu}_{m+1} \mu_{m+1} I \\ &= (A^{(m)})^2 - 2\Re(\mu_{m+1}) A^{(m)} + |\mu_{m+1}|^2 I \end{aligned}$$

defining the first Householder reflection M_1 is real-valued, allowing us to perform the entire computation without resorting to complex numbers.

Remark 5.18 (Multiple-shift strategies). Obviously, we do not have to stop with double shifts, we can also apply r shifts $\mu_1, \dots, \mu_r$ simultaneously by replacing the equation (5.18) by

$$\widehat{Q}_{\text{multi}}^{(m+r)} R_{\text{multi}}^{(m+r)} = (A - \mu_r I) \dots (A - \mu_1 I).$$

The first Householder reflection M_1 now has to eliminate r entries in the first column, therefore the implicit method has to chase a bulge with r rows out of the matrix. Sophisticated multiple-shift strategies [4] can be used to apply a large number of shifts simultaneously.

Remark 5.19 (Aggressive deflation). During a multiple-shift algorithm, some eigenvalues tend to converge more rapidly than others, but the standard stopping criteria fail to take advantage of this property. Aggressive deflation methods [5] can be used to detect converged eigenvalues and apply deflation steps to improve the efficiency significantly.

Chapter 6

Bisection methods *

Summary

If we are mainly interested in the eigenvalues of a matrix, we can look for roots of its characteristic polynomial (cf. Proposition 2.13). In general, evaluating the characteristic polynomials takes too long, but if the matrix is self-adjoint, Householder transformations can be used to make it tridiagonal (cf. Figure 5.4), and the characteristic polynomial of tridiagonal matrices can be evaluated efficiently. In this setting, a bisection algorithm can be used to compute the eigenvalues. Similar to all bisection methods, it is guaranteed to always converge at a rate of $1/2$. In the case of characteristic polynomials of self-adjoint tridiagonal matrices, it is even possible to refine the bisection method to choose which eigenvalues are computed.

Learning targets

- ✓ Introduce the bisection algorithm for eigenvalue problems.
- ✓ Modify the algorithm to compute specific eigenvalues.
- ✓ Use Gershgorin circles to obtain a reliable initial guess for the bisection algorithm.

Using the algorithm given in Figure 5.4, we can turn any matrix $A \in \mathbb{F}^{n\times n}$ into a Hessenberg matrix $H \in \mathbb{F}^{n\times n}$ by using unitary similarity transformations, i.e., we have

$$Q^* A Q = H$$

with a unitary matrix $Q \in \mathbb{F}^{n\times n}$. If A is self-adjoint, i.e., if we have $A = A^*$, we find

$$H^* = (Q^* A Q)^* = Q^* A^* Q^{**} = Q^* A Q = H,$$

i.e., the Hessenberg matrix is also self-adjoint, and therefore tridiagonal.

In this chapter, we consider only self-adjoint tridiagonal matrices

$$T = \begin{pmatrix} a_1 & \bar{b}_1 & & \\ b_1 & a_2 & \ddots & \\ & \ddots & \ddots & \bar{b}_{n-1} \\ & & b_{n-1} & a_n \end{pmatrix} \quad \text{with } a \in \mathbb{R}^n,\ b \in \mathbb{F}^{n-1}. \tag{6.1}$$

We are looking for eigenvalues, and due to Proposition 2.13, these coincide with the zeros of the characteristic polynomial p_T given by

$$p_T(t) = \det(tI - T) = \det \begin{pmatrix} t-a_1 & -\bar{b}_1 & & \\ -b_1 & \ddots & \ddots & \\ & \ddots & \ddots & -\bar{b}_{n-1} \\ & & -b_{n-1} & t-a_n \end{pmatrix} \quad \text{for all } t \in \mathbb{F}.$$

In order to derive an efficient algorithm for evaluating p_T, we consider the characteristic polynomials

$$p_m(t) := \det \begin{pmatrix} t-a_1 & -\bar{b}_1 & & \\ -b_1 & \ddots & \ddots & \\ & \ddots & \ddots & -\bar{b}_{m-1} \\ & & -b_{m-1} & t-a_m \end{pmatrix} \quad \text{for all } m \in \{1,\dots,n\},\ t \in \mathbb{F}$$

corresponding to the m-th principal submatrices of T. We are looking for a recurrence relation that allows us to compute $p_n = p_T$.

Computing p_1 and p_2 is straightforward, therefore we only have to consider the computation of p_m for $m \in \{3,\dots,n\}$. We can evaluate the characteristic polynomial p_m for a given $t \in \mathbb{F}$ by Laplace expansion in the last column and row:

$$\begin{aligned} p_m(t) &= \det \left(\begin{array}{ccccc|c} t-a_1 & -\bar{b}_1 & & & & \\ -b_1 & \ddots & \ddots & & & \\ & \ddots & \ddots & -\bar{b}_{m-3} & & \\ & & -b_{m-3} & t-a_{m-2} & -\bar{b}_{m-2} & \\ & & & -b_{m-2} & t-a_{m-1} & -\bar{b}_{m-1} \\ & & & & -b_{m-1} & t-a_m \end{array} \right) \\ &= (t-a_m) \det \begin{pmatrix} t-a_1 & -\bar{b}_1 & & & \\ -b_1 & \ddots & \ddots & & \\ & \ddots & \ddots & -\bar{b}_{m-3} & \\ & & -b_{m-3} & t-a_{m-2} & -\bar{b}_{m-2} \\ & & & -b_{m-2} & t-a_{m-1} \end{pmatrix} \\ &\quad - (-\bar{b}_{m-1}) \det \left(\begin{array}{ccccc} t-a_1 & -\bar{b}_1 & & & \\ -b_1 & \ddots & \ddots & & \\ & \ddots & \ddots & -\bar{b}_{m-3} & \\ & & -b_{m-3} & t-a_{m-2} & -\bar{b}_{m-2} \\ \hline & & & & -b_{m-1} \end{array} \right) \end{aligned}$$

$$
\begin{aligned}
&= (t-a_m)\det\begin{pmatrix} t-a_1 & -\bar{b}_1 & & \\ -b_1 & \ddots & \ddots & \\ & \ddots & \ddots & -\bar{b}_{m-2} \\ & & -b_{m-2} & t-a_{m-1} \end{pmatrix} \\
&\quad - |b_{m-1}|^2 \det\begin{pmatrix} t-a_1 & -\bar{b}_1 & & \\ -b_1 & \ddots & \ddots & \\ & \ddots & \ddots & -\bar{b}_{m-3} \\ & & -b_{m-3} & t-a_{m-2} \end{pmatrix} \\
&= (t-a_m)p_{m-1}(t) - |b_{m-1}|^2 p_{m-2}(t).
\end{aligned}
$$

This equation allows us to compute $p_1, p_2, \ldots, p_n$ in not more than $6n$ operations by the recurrence relation

$$p_1(t) = t - a_1, \tag{6.2a}$$

$$p_2(t) = (t-a_2)p_1(t) - |b_1|^2, \tag{6.2b}$$

$$p_m(t) = (t-a_m)p_{m-1}(t) - |b_{m-1}|^2 p_{m-2}(t) \quad \text{for all } m \in \{3,\ldots,n\},\ t \in \mathbb{F}. \tag{6.2c}$$

Provided that we know $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$ and $p_n(\alpha)p_n(\beta) \leq 0$, the intermediate value theorem states that there exists a root $\lambda \in [\alpha, \beta]$ of p_n. We can apply a bisection method to approximate this root: we let $\gamma := (\beta + \alpha)/2$ denote the midpoint of the interval and check whether $p_n(\alpha)p_n(\gamma) \leq 0$ holds. If this is the case, we know that a root exists in $[\alpha, \gamma]$. Otherwise, we observe $p(\gamma)p(\beta) \leq 0$ and conclude that a root exists in $[\gamma, \beta]$. Repeating this process leads to an arbitrarily small interval containing a root.

This approach has two major disadvantages: we can only use it to compute one root, and we require an initial guess for the interval $[\alpha, \beta]$ such that the signs of $p_n(\alpha)$ and $p_n(\beta)$ differ. The rest of this chapter is dedicated to fixing these problems and thus providing us with a reliable method for finding the eigenvalues.

6.1 Sturm chains

We are interested in computing the roots of the characteristic polynomial p_T. Assume that p_T has n simple roots $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, and consider the derivative p_T'. For each $i \in \{1, \ldots, n-1\}$, the mean value theorem states that $p_T(\lambda_i) = 0 = p_T(\lambda_{i+1})$ implies that there has to be a zero λ_i' of p_T' in $[\lambda_i, \lambda_{i+1}]$. Since λ_i and λ_{i+1} are simple, we have $\lambda_i < \lambda_i' < \lambda_{i+1}$. Since p_T' is a non-zero polynomial of degree $n-1$, it cannot have more than these $n-1$ roots, and all of the roots have to be simple.

A straightforward induction yields that for $m \in \{1, \ldots, n\}$, the m-th derivative $p_T^{(m)}$ has exactly $n-m$ simple roots and that these roots are situated between the

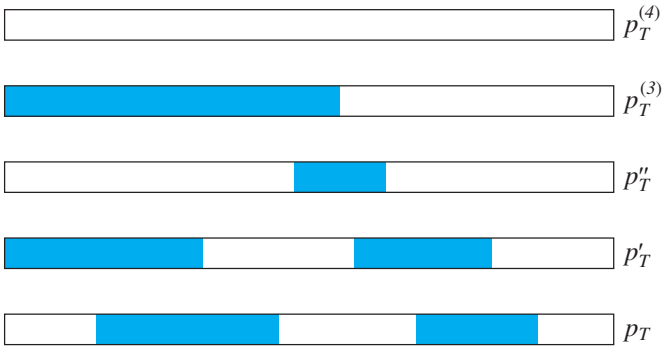

Figure 6.1. Sign pattern for p_T and its derivatives. The boxes represent the interval $[a, b]$, subintervals with negative values of the polynomials are marked in blue.

roots of $p_T^{(m-1)}$. We can take advantage of this property to determine how many roots of p_T are located within a given interval $[a, b]$: since all roots of $p_T^{(m)}$ are simple, the polynomial has to change its sign at each root. Since the roots of $p_T^{(m)}$ are interleaved with the roots of $p_T^{(m+1)}$, the signs change in a pattern like the one given in Figure 6.1.

We introduce a function s giving us the number of sign changes between p_T and its derivatives for any points $\lambda \in \mathbb{R}$

$$s(\lambda) := \#\{m \in \{0, \dots, n-1\} \ : \ p_T^{(m)}(\lambda) p_T^{(m+1)}(\lambda) < 0\}$$

and see that it decreases by one each time λ passes from left to right across a root of p_T and remains otherwise constant: it can be used to count the number of roots.

Computing all derivatives of p_T is possible, but not very elegant. Fortunately, we can replace $p_T^{(m)}$ by the characteristic polynomial $q_m := p_{n-m}$ of the $(n-m)$-th principal submatrix without losing the important sign-change properties. Since $p_1, p_2, \dots, p_n$ are computed anyway during our algorithm for evaluating $p_n = p_T$, checking the signs requires almost no additional effort and we obtain a very efficient algorithm.

Definition 6.1 (Sturm chain). Let $(q_0, q_1, \dots, q_n)$ be a tuple of polynomials. We call it a *Sturm chain* if the following conditions hold:

1. All zeros of q_0 are simple.
2. For $\lambda \in \mathbb{R}$ with $q_0(\lambda) = 0$, we have $q_0'(\lambda) q_1(\lambda) > 0$.
3. For $\lambda \in \mathbb{R}$, $m \in \{1, \dots, n-1\}$ with $q_m(\lambda) = 0$, we have $q_{m-1}(\lambda) q_{m+1}(\lambda) < 0$.
4. q_n has no zeros.

Theorem 6.2 (Number of zeros). *Let $(q_0, q_1, \dots, q_n)$ be a Sturm chain. We define*

$$S(\lambda) := \{m \in \{0, \dots, n-1\} \,:\, q_m(\lambda)q_{m+1}(\lambda) < 0 \text{ or } q_m(\lambda) = 0\},$$
$$s(\lambda) := \#S(\lambda) \quad \text{for all } \lambda \in \mathbb{R}.$$

Then we have

$$s(a) - s(b) = \#\{\lambda \in (a, b] \,:\, q_0(\lambda) = 0\} \quad \text{for all } a, b \in \mathbb{R},\ a < b,$$

i.e., the function s counts the number of zeros of q_0.

Proof. We are interested in the zeros of the polynomials $q_0, \dots, q_n$, i.e., in the set

$$Z(\lambda) := \{m \in \{0, \dots, n\} \,:\, q_m(\lambda) = 0\} \quad \text{for all } \lambda \in \mathbb{R}.$$

The function s changes it value only if the signs of the functions q_m change, i.e., in points $\lambda \in \mathbb{R}$ with $Z(\lambda) \neq \emptyset$.

Let $\lambda \in \mathbb{R}$ with $Z(\lambda) \neq \emptyset$. Since the set of zeros of polynomials cannot have limit points, there is an $\epsilon \in \mathbb{R}_{>0}$ such that $Z(\mu) = \emptyset$ and $q_0'(\mu) \neq 0$ for all $\mu \in (\lambda, \lambda + \epsilon]$.

Let $m \in Z(\lambda)$.

Case 1: Assume $m > 0$. Condition 4 yields $m \neq n$ and therefore $m \in \{1, \dots, n-1\}$. Condition 3 implies that $q_{m-1}(\lambda)$ and $q_{m+1}(\lambda)$ are not zero and have opposite signs, and due to the choice of ϵ, both q_{m-1} and q_{m+1} cannot change their signs in $[\lambda, \lambda+\epsilon]$.

If $m \in S(\lambda + \epsilon)$, we have $q_m(\lambda + \epsilon)q_{m+1}(\lambda + \epsilon) < 0$ and therefore $q_{m-1}(\lambda + \epsilon)q_m(\lambda + \epsilon) > 0$, i.e., $m - 1 \notin S(\lambda + \epsilon)$.

On the other hand if $m \notin S(\lambda + \epsilon)$, we have $q_m(\lambda + \epsilon)q_{m+1}(\lambda + \epsilon) > 0$ and therefore $q_{m-1}(\lambda + \epsilon)q_m(\lambda + \epsilon) < 0$, i.e., $m - 1 \in S(\lambda + \epsilon)$.

We let $\widehat{M}_m := \{m-1, m\}$ and $\widehat{S}_m(\mu) := S(\mu) \cap \{m-1, m\}$ for all $\mu \in [\lambda, \lambda + \epsilon]$ and summarize our findings as

$$\#\widehat{S}_m(\mu) = 1 \quad \text{for all } \mu \in (\lambda, \lambda + \epsilon].$$

Since $q_{m-1}(\lambda)$ is not zero, we also have $\widehat{S}_m(\lambda) = \{m\}$ and therefore $\#\widehat{S}_m(\mu) = 1$ for all $\mu \in [\lambda, \lambda + \epsilon]$.

Case 2: Assume now $m = 0$. Due to condition 2, we have $q_0'(\lambda)q_1(\lambda) > 0$, and due to our choice of ϵ, q_0' cannot change its sign in $(\lambda, \lambda + \epsilon]$.

By the fundamental theorem of calculus, we obtain

$$q_0(\lambda + \epsilon)q_1(\lambda) = \int_\lambda^{\lambda+\epsilon} q_0'(t)q_1(\lambda)dt > 0,$$

and since q_1 cannot change its sign in $[\lambda, \lambda + \epsilon]$, we conclude

$$q_0(\lambda + \epsilon)q_1(\lambda + \epsilon) > 0,$$

i.e., $0 \notin S(\lambda+\epsilon)$. By definition, we also have $0 \in S(\lambda)$. We let $\widehat{M}_0 := \{0\}$ and $\widehat{S}_0(\mu) := S(\mu) \cap \{0\}$ and summarize our result as $\#\widehat{S}_0(\lambda) = 1$ and $\#\widehat{S}_0(\lambda+\epsilon) = 0$.

Conclusion: Due to our choice of ϵ, the set

$$R := S(\mu) \setminus \bigcup_{m \in Z(\lambda)} \widehat{M}_m$$

does not depend on $\mu \in [\lambda, \lambda+\epsilon]$ and is therefore well-defined. The equation

$$S(\mu) = R \cup \bigcup_{m \in Z(\lambda)} \widehat{S}_m(\mu) \quad \text{for all } \mu \in [\lambda, \lambda+\epsilon]$$

defines unions of disjoint sets, so we have

$$s(\mu) = \#S(\mu) = \#R + \sum_{m \in Z(\lambda)} \#\widehat{S}_m(\mu) \quad \text{for all } \mu \in [\lambda, \lambda+\epsilon].$$

We have already proven that $\#\widehat{S}_m(\mu) = 1$ holds for all $\mu \in [\lambda, \lambda+\epsilon]$, $m \in Z(\lambda)\setminus\{0\}$. We also have $\#(S(\lambda)\cap\{0\}) = 1$ and $\#(S(\lambda+\epsilon)\cap\{0\}) = 0$ if $0 \in Z(\lambda)$ and conclude

$$s(\lambda+\epsilon) = \begin{cases} s(\lambda) & \text{if } 0 \notin Z(\lambda), \\ s(\lambda) - 1 & \text{otherwise} \end{cases},$$

i.e., s decreases by one each time we pass a zero of q_0. □

In order to apply this general result to the problem of computing eigenvalues of the tridiagonal matrix T, we have to check that the polynomials $q_0 := p_n, q_1 := p_{n-1}, \ldots, q_{n-1} := p_1, q_n := 1$ are a Sturm chain. This is not always the case: if we choose $T = I$, the polynomials q_m have a zero of multiplicity $n-m$ at $\lambda = 1$, so conditions 2 and 3 are violated. Fortunately, there is a simple criterion that helps us avoid this problem:

Definition 6.3 (Irreducible matrix). A Hessenberg matrix $H \in \mathbb{F}^{n\times n}$ is called *irreducible* (or *unreduced*) if

$$h_{i+1,i} \neq 0 \quad \text{for all } i \in \{1, \ldots, n-1\},$$

i.e., if all sub-diagonal elements are non-zero.

Checking whether a Hessenberg matrix is irreducible is straightforward, and if it is not, deflation (cf. Section 5.4) can be used to split it into irreducible submatrices.

Theorem 6.4 (Sturm chain). *Let $T \in \mathbb{F}^{n\times n}$ be an irreducible tridiagonal matrix of the form* (6.1)*, i.e., let*

$$b_i \neq 0 \quad \text{for all } i \in \{1,\dots,n-1\}.$$

Then the polynomials given by

$$q_n := 1, \quad q_m := p_{n-m} \quad \text{for all } m \in \{0,\dots,n-1\}$$

are a Sturm chain.

Proof. We are looking for eigenvectors and eigenvalues of the matrix T. Given a $t \in \mathbb{R}$, we look for a non-zero vector $e(t) \in \mathbb{F}^n$ such that the first $n-1$ components of $(tI - T)e(t)$ vanish. If the last component also vanishes, $e(t)$ is an eigenvector for the eigenvalue t.

In order to ensure $e(t) \neq 0$, we choose $e_1(t) = 1$. By considering the first row of $(tI - T)e(t)$, we obtain

$$\begin{aligned} 0 &= (t-a_1)e_1(t) - \bar{b}_1 e_2(t), \\ \bar{b}_1 e_2(t) &= (t-a_1)e_1(t) = p_1(t). \end{aligned}$$

For $m \in \{2,\dots,n-1\}$, we have

$$\begin{aligned} 0 &= -b_{m-1}e_{m-1}(t) + (t-a_m)e_m(t) - \bar{b}_m e_{m+1}(t), \\ \bar{b}_m e_{m+1}(t) &= (t-a_m)e_m(t) - b_{m-1}e_{m-1}(t). \end{aligned} \tag{6.3}$$

The right-hand side resembles (6.2c), and we can take advantage of this resemblance: we define $p_0 := 1$ and let

$$e_m(t) := \frac{p_{m-1}(t)}{\prod_{k=1}^{m-1}\bar{b}_k} \quad \text{for all } m \in \{1,\dots,n\},\ t \in \mathbb{R}.$$

By definition, we have $e_1(t) = 1$, and we also obtain

$$\begin{aligned} &-b_{m-1}e_{m-1}(t) + (t-a_m)e_m(t) - \bar{b}_m e_{m+1}(t) \\ &\quad = -b_{m-1}\frac{p_{m-2}(t)}{\prod_{k=1}^{m-2}\bar{b}_k} + (t-a_m)\frac{p_{m-1}(t)}{\prod_{k=1}^{m-1}\bar{b}_k} - \bar{b}_m\frac{p_m(t)}{\prod_{k=1}^{m}\bar{b}_k} \\ &\quad = \frac{-|b_{m-1}|^2 p_{m-2}(t) + (t-a_m)p_{m-1}(t) - p_m(t)}{\prod_{k=1}^{m-1}\bar{b}_k} \\ &\quad = \frac{p_m(t) - p_m(t)}{\prod_{k=1}^{m-1}\bar{b}_k} = 0 \quad \text{for all } m \in \{3,\dots,n-1\}. \end{aligned}$$

using (6.2c) in the last step. (6.3) implies that the vector $e(t)$ meets our requirements: the first $n-1$ components of $(tI-T)e(t)$ are equal to zero. Now we consider the last component given by

$$\begin{aligned}\gamma(t) &:= -b_{n-1}e_{n-1}(t) + (t-a_n)e_n(t) = -b_{n-1}\frac{p_{n-2}(t)}{\prod_{k=1}^{n-2}\bar{b}_k} + (t-a_n)\frac{p_{n-1}(t)}{\prod_{k=1}^{n-1}\bar{b}_k} \\ &= \frac{-|b_{n-1}|^2 p_{n-2}(t) + (t-a_n)p_{n-1}(t)}{\prod_{k=1}^{n-1}\bar{b}_k} = \frac{p_n(t)}{\prod_{k=1}^{n-1}\bar{b}_k} \quad \text{for all } t \in \mathbb{R}.\end{aligned}$$

The vector $e(t)$ is an eigenvector for an eigenvalue $t \in \mathbb{R}$ if and only if $\gamma(t) = 0$ holds, and this is equivalent to $p_n(t) = 0$. In general we have

$$(tI-T)e(t) = \begin{pmatrix} 0 \\ \gamma(t) \end{pmatrix} \quad \text{for all } t \in \mathbb{R},$$

and differentiating by t yields

$$e(t) + (tI-T)e'(t) = \begin{pmatrix} 0 \\ \gamma'(t) \end{pmatrix} \quad \text{for all } t \in \mathbb{R}. \tag{6.4}$$

Now we can verify the conditions of Definition 6.1. We begin with condition (2). Let $\xi \in \mathbb{R}$ be a root of $q_0 = p_n$, i.e., $\gamma(\xi) = 0$. Taking the inner product of (6.4) by the vector $e(\xi)$ yields

$$\begin{aligned}\bar{e}_n(\xi)\gamma'(\xi) &= \langle e(\xi) + (\xi I - T)e'(\xi), e(\xi)\rangle = \|e(\xi)\|^2 + \langle e'(\xi), (\xi I - T)e(\xi)\rangle \\ &= \|e(\xi)\|^2 > 0,\end{aligned}$$

and we conclude

$$0 < \bar{e}_n(\xi)\gamma'(\xi) = \frac{p_{n-1}(\xi)}{\prod_{k=1}^{n-1} b_k}\frac{p_n'(\xi)}{\prod_{k=1}^{n-1}\bar{b}_k} = \frac{p_{n-1}(\xi)p_n'(\xi)}{\prod_{k=1}^{n-1}|b_k|^2} = \frac{q_1(\xi)q_0'(\xi)}{\prod_{k=1}^{n-1}|b_k|^2},$$

so the condition (2) holds. It implies $q_0'(\xi) \neq 0$, therefore all zeros of q_0 are simple and (1) holds as well. $q_n = 1$ yields condition (4). This leaves only condition (3) to consider. Let $m \in \{1, \ldots, n-1\}$ and let $\lambda \in \mathbb{R}$ with $0 = q_m(\lambda) = p_{n-m}(\lambda)$. The recurrence relation (6.2c) yields

$$\begin{aligned}q_{m-1}(\lambda) &= p_{n-m+1}(\lambda) = (\lambda - a_{n-m+1})p_{n-m}(\lambda) - |b_{n-m}|^2 p_{n-m-1}(\lambda) \\ &= -|b_{n-m}|^2 q_{m+1}(\lambda),\end{aligned}$$

i.e., $q_{m+1}(\lambda)q_{m-1}(\lambda) = -|b_{n-m}|^2 q_{m+1}^2(\lambda) \leq 0$. To complete the proof, we only have to show $q_{m+1}(\lambda) \neq 0$. We can use (6.2c) to do this: $q_m(\lambda) = 0$ and $q_{m+1}(\lambda) = 0$ would imply $q_{m+2}(\lambda) = 0$, and a simple induction leads to $q_n(\lambda) = 0$. This would contradict $q_n = p_0 = 1$, so we can conclude $q_{m+1}(\lambda) \neq 0$. □

```
function sturm_evaluate(t, a, b);
  p_0 ← 1;   p_1 ← t − a_1;
  c ← 0;
  if p_0 p_1 < 0 or p_1 = 0 then c ← c + 1;
  for m = 2 to n do begin
    p_m ← (t − a_m) p_{m−1} − |b_{m−1}|^2 p_{m−2};
    if p_m p_{m−1} < 0 or p_m = 0 then c ← c + 1
  end;
  return c
end;

procedure sturm_bisection(k, a, b, var α, β);
begin
  while β − α > ε do begin
    γ ← (β + α)/2;   c ← sturm_evaluate(γ, a, b);
    if k ≤ n − c then
      β ← γ
    else
      α ← γ
  end
end
```

Figure 6.2. Compute an interval containing the k-th eigenvalue by a bisection method applied to Sturm chains.

We can use Theorem 6.2 to compute the k-th eigenvalue of the matrix T: we have

$$\lim_{t\to-\infty} p_m(t) = \begin{cases} \infty & \text{if } m \text{ is even,} \\ -\infty & \text{otherwise} \end{cases} \qquad \text{for all } m \in \{1,\dots,n\},$$

since the leading coefficients of the polynomials are always equal to one, and this implies

$$\lim_{t\to-\infty} s(t) = n.$$

Theorem 6.2 states that s decreases by one each time a zero of p_T is passed, so for any $t \in \mathbb{R}$, $n - s(t)$ gives us the number of zeros less than or equal to t. If we are looking for the k-th eigenvalue, we simply check whether $k \le n - s(t)$ holds: in this case we continue to search in $\mathbb{R}_{\le t}$, otherwise we consider $\mathbb{R}_{>t}$. The resulting algorithm is given in Figure 6.2.

The bisection method, particularly the evaluation of the function s, can react very sensitively to rounding errors, since even small perturbations of $p_m(t)$ and $p_{m-1}(t)$ may change the sign and therefore the result.

Remark 6.5 (Slicing the spectrum). The key ingredient of the method is the computation of the number of eigenvalues smaller and larger than a given value t. Instead of examining a Sturm sequence, we can also solve this problem by computing factorizations $\mu I - T = LDL^*$ of the shifted matrix and examining the number of negative and positive diagonal elements of the matrix D [33, page 15].

Remark 6.6 (Divide-and-conquer methods). The relatively simple structure of tridiagonal matrices can be used to construct alternative solution algorithms. One possibility are *divide-and-conquer methods* [7, 12] that split the tridiagonal matrix into submatrices, compute eigenvalues and eigenvectors of the submatrices, and then combine them to construct eigenvalues and eigenvectors of the original matrix. This class of algorithms is particularly well-suited for parallel implementations.

Remark 6.7 (MRRR). Based on sufficiently accurate approximations of the eigenvalues, the approach of *multiple relatively robust representations* (MRRR or MR^3) employs multiple factorizations $\mu I - T = LDL^*$ of shifted matrices in order to compute eigenvectors of T with high relative accuracy [10, 11].

6.2 Gershgorin discs

Let $A \in \mathbb{F}^{n\times n}$. In order to use the bisection algorithm, we have to find an interval $[\alpha, \beta]$ that is guaranteed to contain all eigenvalues of the matrix T. *Gershgorin discs* [16] provide us not only with this interval, but can also be used to determine bounds for the eigenvalues of general matrices.

We use the *maximum norm* defined by

$$\|x\|_\infty := \max\{|x_i| \ : \ i \in \{1, \ldots, n\}\} \quad \text{for all vectors } x \in \mathbb{F}^n$$

and by

$$\|A\|_\infty := \max\{\|Ax\|_\infty \ : \ x \in \mathbb{F}^n, \|x\|_\infty = 1\} \quad \text{for all matrices } A \in \mathbb{F}^{n\times n}.$$

These definitions ensure that we again have *compatible* norms, i.e.,

$$\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty \quad \text{for all } A \in \mathbb{F}^{n\times n},\ x \in \mathbb{F}^n. \tag{6.5}$$

Combining this property with the definition of the norm yields the estimate

$$\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty \quad \text{for all } A, B \in \mathbb{F}^{n\times n} \tag{6.6}$$

that leads to the following simple criterion for the invertibility of perturbed matrices.

Proposition 6.8 (Neumann series). *Let $X \in \mathbb{F}^{n\times n}$ be a matrix satisfying $\|X\|_\infty < 1$. Then $I - X$ is invertible.*

Proof. We have

$$(I - X)\sum_{k=0}^{m} X^k = \sum_{k=0}^{m} X^k - \sum_{k=1}^{m+1} X^k = I - X^{m+1} \quad \text{for all } m \in \mathbb{N}.$$

Due to $\|X\|_\infty < 1$ and (6.6), we have

$$\|\lim_{m\to\infty} X^m\|_\infty = \lim_{m\to\infty} \|X^m\|_\infty \leq \lim_{m\to\infty} \|X\|_\infty^m = 0,$$

i.e., $\lim_{m\to\infty} X^m = 0$ and therefore

$$(I - X)\sum_{k=0}^{\infty} X^k = I.$$

This implies that $I - X$ is invertible. □

In order to apply this result, it would be advantageous to be able to evaluate the norm $\|X\|_\infty$ for a matrix $X \in \mathbb{F}^{n\times n}$ explicitly. For the maximum norm, this task can be solved directly:

Proposition 6.9 (Maximum norm). *Let $A \in \mathbb{F}^{n\times n}$ be a matrix. Then we have*

$$\|A\|_\infty = \gamma := \max\left\{\sum_{j=1}^{n} |a_{ij}| \,:\, i \in \{1,\ldots,n\}\right\}. \tag{6.7}$$

Proof. The triangle inequality yields

$$\begin{aligned}
\|Ax\|_\infty &= \max\left\{\left|\sum_{j=1}^{n} a_{ij}x_j\right| \,:\, i \in \{1,\ldots,n\}\right\} \\
&\leq \max\left\{\sum_{j=1}^{n} |a_{ij}|\,|x_j| \,:\, i \in \{1,\ldots,n\}\right\} \\
&\leq \max\left\{\sum_{j=1}^{n} |a_{ij}| \,:\, i \in \{1,\ldots,n\}\right\} \|x\|_\infty = \gamma\|x\|_\infty \quad \text{for all } x \in \mathbb{F}^n,
\end{aligned}$$

and this implies $\|A\|_\infty \leq \gamma$.

Let $i \in \{1,\dots,n\}$ be the index for which the maximum in (6.7) is attained. We define a vector $x \in \mathbb{F}^n$ by

$$x_j := \operatorname{sgn}(\bar{a}_{ij}) \quad \text{for all } j \in \{1,\dots,n\}$$

and find $\|x\|_\infty = 1$ and

$$\|Ax\|_\infty \geq \left|\sum_{j=1}^n a_{ij}x_j\right| = \left|\sum_{j=1}^n a_{ij}\operatorname{sgn}(\bar{a}_{ij})\right| = \sum_{j=1}^n |a_{ij}| = \gamma$$

so we have also proven $\|A\|_\infty \geq \gamma$ and therefore $\|A\|_\infty = \gamma$. □

Theorem 6.10 (Gershgorin spheres). *Let $A \in \mathbb{F}^{n\times n}$. We define discs (or intervals)*

$$G_i := \left\{ \mu \in \mathbb{F} \,:\, |\mu - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^n |a_{ij}| \right\} \quad \textit{for all } i \in \{1,\dots,n\}.$$

Then we have

$$\sigma(A) \subseteq \bigcup_{i=1}^n G_i,$$

i.e., the spectrum of A is contained in discs (or intervals) centered at the diagonal elements.

Proof. (cf. [16, Satz II]) Let $\lambda \in \sigma(A)$, and let

$$D := \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{pmatrix}$$

be the diagonal part of A. If $\lambda I - D$ is not invertible, there has to be an index $i \in \{1,\dots,n\}$ such that $\lambda = a_{ii}$, therefore we have $\lambda \in G_i$.

Assume now that $\lambda I - D$ is invertible. Since λ is an eigenvalue, $\lambda I - A$ cannot be invertible, and the same holds for

$$(\lambda I - D)^{-1}(\lambda I - A) = (\lambda I - D)^{-1}(\lambda I - D - (A - D)) = I - (\lambda I - D)^{-1}(A - D).$$

Due to Propositions 6.8 and 6.9, this implies

$$1 \leq \|(\lambda I - D)^{-1}(A - D)\|_\infty = \max\left\{ \sum_{j=1}^n \frac{|a_{ij} - d_{ij}|}{|\lambda - a_{ii}|} \,:\, i \in \{1,\dots,n\} \right\}$$

$$= \max\left\{ \sum_{\substack{j=1 \\ j\neq i}}^n \frac{|a_{ij}|}{|\lambda - a_{ii}|} \,:\, i \in \{1,\dots,n\} \right\}.$$

We choose an index $i \in \{1, \dots, n\}$ for which the maximum is attained and conclude

$$1 \leq \frac{1}{|\lambda - a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad |\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,$$

i.e., $\lambda \in G_i$. □

In the context of Sturm chain bisection methods, we can use this result to compute an interval $[\alpha, \beta]$ given by

$$\alpha := \min\{a_1 - |b_1|, a_n - |b_{n-1}|, a_i - |b_i| - |b_{i-1}| \,:\, i \in \{2, \dots, n-1\}\},$$
$$\beta := \max\{a_1 + |b_1|, a_n + |b_{n-1}|, a_i + |b_i| + |b_{i-1}| \,:\, i \in \{2, \dots, n-1\}\},$$

that is guaranteed to include the entire spectrum of T (since T is self-adjoint, Corollary 2.51 yields that all eigenvalues are real, and Theorem 6.10 provides upper and lower bounds).

Chapter 7

Krylov subspace methods for large sparse eigenvalue problems

Summary

Large sparse eigenvalue problems typically have an underlying matrix that has a very large size, but also a special structure so that matrix-vector multiplications can efficiently be performed, but similarity transformations cannot, so different strategies for the computation of eigenvalues and eigenvectors are needed. The most common methods are Krylov subspace methods that can be used to compute some, but not all, eigenvalues of the matrix. Among those, the Arnoldi iteration and the symmetric Lanczos algorithm will be considered in detail in this chapter and also the convergence behaviour of these methods will be investigated with the help of an important tool: Chebyshev polynomials.

Learning targets

- ✓ Introduce projection methods for large sparse eigenvalue problems.
- ✓ Construct Krylov subspaces and investigate their basic properties.
- ✓ Introduce the Arnoldi iteration and the symmetric Lanczos algorithm as examples of Krylov subspace methods.
- ✓ Use Chebyshev polynomials to analyze the convergence behaviour of Krylov subspaces methods.

7.1 Sparse matrices and projection methods

Many applications in engineering sciences, physics, chemistry et cetera give rise to eigenvalue problems with matrices $A \in \mathbb{F}^{n\times n}$, where n is very large, e.g., $n \approx 10^8$ (or maybe even much larger). Typically, these matrices are *sparse*, i.e., most of the entries of the matrix are zero. An example of a sparse matrix is the matrix

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n} \tag{7.1}$$

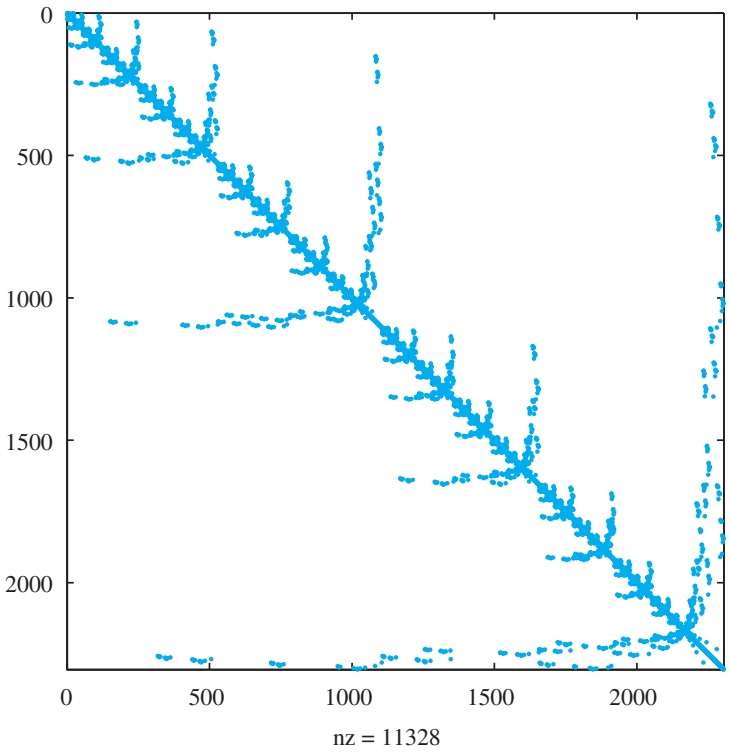

Figure 7.1. Depiction of a sparse matrix which represents the 5-point discrete Laplacian on a square 50-by-50 grid using a nested dissection ordering of the gridpoints. Nonzero entries of the matrix are plotted in blue. "nz" is the total number of nonzero entries.

from (1.2). We observe that regardless how large n is, there are at most three entries in each row that are distinct from zero. Another more complicated example is depicted in Figure 7.1. The matrix is of size 2304×2304, but only 11,328 of its 5,308,416 entries are nonzero.

Sparsity is a special matrix structure having two important properties that allow us to deal with matrices of very large size n. First, it is not necessary to store the matrix in the conventional way which would require to store all n^2 entries. Instead, we only need to store the nonzero entries and information on their positions within the matrix. For example, instead of storing the matrix

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 5 \end{pmatrix}$$

as a 4×4 array, we may store the three vectors

$$l = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \quad r = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 4 \\ 4 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3 \\ 4 \end{pmatrix},$$

instead, where l is the list of nonzero entries of A, and r and c are vectors containing the corresponding row and column indices of the entries, respectively. The matrix $A = (a_{ij})$ can be easily reconstructed from these three vectors by setting

$$a_{ij} = \begin{cases} l_k & \text{if } r_k = i \text{ and } c_k = j, \\ 0 & \text{else.} \end{cases} \tag{7.2}$$

Exercise 7.1 (Storing sparse matrices). Construct vectors l, r, c, so that the matrix generated by (7.2) corresponds to the matrix A in (7.1) for $n = 6$.

The second important property is that matrix-vector multiplications with sparse matrices can be performed considerably faster than compared to the conventional way. This is due to the fact that most entries of the matrix are zero. Multiplying a matrix $A \in \mathbb{F}^{n \times n}$ and a vector $x \in \mathbb{F}^n$, we have to multiply each of the n rows of A with x which adds up to $n \cdot n = n^2$ multiplications and $n \cdot (n-1) = n^2 - n$ additions, i.e., in total $2n^2 - n$ operations. On the other hand, if like in (7.1) only three entries per row are nonzero (except for the first and last row), then multiplying A with a vector $x \in \mathbb{F}^n$ costs $3n-2$ multiplications and $2n-2$ additions, i.e., $5n-4$ operations. Thus, the cost for computing a matrix-vector product is growing approximately linearly in n in this example as opposed to quadratically in the conventional case.

In some applications, it may actually be more convenient to not store the matrix A at all, but write a procedure instead that computes Ax for a given vector x. In this situation, we could think of our matrix A as being a *black box* that transforms a given vector x to the vector $y = Ax$.

$$x \longrightarrow \boxed{\text{A}} \longrightarrow Ax \tag{7.3}$$

For example, for our matrix A from (7.1) such a procedure that computes Ax from a given vector x is presented in Figure 7.2. In this situation, we cannot apply similarity transformations anymore, because A is not given explicitly. This also refers to the case that we actually do store the matrix, because then similarity transformations may (and do in general) destroy the sparsity of the matrix. We would then have to store the matrix in the conventional way and this may not be possible because of its large size. Therefore, the efficient QR iteration from Chapter 5 is no longer a method of choice as it heavily manipulates the given matrix. Therefore, we have to come up with

```
procedure sparse_matrix_mult(x, var y);
begin
  y_1 ← 2x_1 − x_2;
  for i ∈ {2, ..., n − 1} do
    y_i ← 2x_i − x_{i−1} − x_{i+1};
  y_n ← 2x_n − x_{n−1}
end
```

Figure 7.2. Sparse matrix multiplication with the matrix A from (7.1).

a different strategy which should be based on matrix-vector multiplications, because this is the only way to obtain information on our matrix A.

As mentioned in the beginning of this chapter, we are particularly interested in the case that n is very large. In most applications, not all eigenvalues and eigenvectors are needed, but just a few of them are of interest, for example, the eigenvalues that are largest or smallest in modulus, e.g., in Example 1.1 from the introduction, we are interested in a few of the smallest eigenvalues of A. However, one should keep in mind that for sizes as large as 10^8 "a few" may still mean "a few hundred" or "a few thousand".

A concept commonly used for the treatment of large sparse eigenvalue problems is the concept of *projection methods*. The idea is pretty simple: if $A \in \mathbb{F}^{n\times n}$ is a matrix and if $(\lambda, v) \in \mathbb{F} \times \mathbb{F}^n$ is an eigenpair of A, then $Av = \lambda v$, i.e., the vector $Av - \lambda v$ is the zero vector which is orthogonal to the space $\mathbb{F}^n$. Now let us replace the full space $\mathbb{F}^n$ by a subspace $\mathcal{X} \subseteq \mathbb{F}^n$ of smaller dimension, i.e., we will look for pairs $(\mu, x) \in \mathbb{F} \times \mathcal{X}$ so that $Ax - \mu x \perp \mathcal{X}$, i.e., $Ax - \mu x$ is orthogonal to $\mathcal{X}$. This condition is called *Ritz–Galerkin condition* as it is analogous to the Ritz–Galerkin conditions used in finite element methods.

Definition 7.2 (Ritz pair). Let $A \in \mathbb{F}^{n\times n}$ and let $\mathcal{X} \subseteq \mathbb{F}^n$ be a subspace. Then a pair $(\mu, x) \in \mathbb{F} \times \mathcal{X}$ is called a *Ritz pair* of A with respect to $\mathcal{X}$ if it satisfies the *Ritz–Galerkin condition*

$$Ax - \mu x \perp \mathcal{X}.$$

If (μ, x) is a Ritz pair, then μ is called a *Ritz value* and x is called the associated *Ritz vector*.

At this point, the following questions arise naturally:

1. How do we choose a suitable subspace $\mathcal{X}$?
2. How do we compute Ritz values and vectors?
3. Will the computed Ritz pairs be "good" approximations to eigenpairs?

We will turn to question 1. and 3. in the next section. Question 2., however, has a fairly immediate answer if we are given an *orthonormal basis* of $\mathcal{X}$, i.e., a basis with pairwise perpendicular vectors that all have norm one.

Theorem 7.3 (Detection of Ritz pairs). *Let $A \in \mathbb{F}^{n\times n}$ and let $\mathcal{X} \subseteq \mathbb{F}^n$ be a subspace. Furthermore let $Q \in \mathbb{F}^{n\times m}$ be an isometric matrix such that $\mathcal{X} = \mathcal{R}(Q)$, i.e., the columns of Q form a orthonormal basis of $\mathcal{X}$. Then the following two conditions are equivalent for $x = Qy \in \mathcal{X}$, where $y \in \mathbb{F}^m$,*

1. *$(\mu, Qy) \in \mathbb{F} \times \mathcal{X}$ is a Ritz pair of A with respect to $\mathcal{X}$;*
2. *$(\mu, y) \in \mathbb{F} \times \mathbb{F}^m$ is an eigenpair of the $m \times m$ matrix Q^*AQ.*

Proof. (μ, y) is an eigenpair of Q^*AQ if and only if

$$Q^*AQy = \mu y = \mu Q^*Qy$$

which itself is equivalent to

$$Q^*(AQy - \mu Qy) = 0.$$

The latter identity means that $AQy - \mu Qy$ is orthogonal to $\mathcal{R}(Q) = \mathcal{X}$, i.e., (μ, Qy) is a Ritz pair of A with respect to $\mathcal{X}$. □

7.2 Krylov subspaces

What kind of subspaces are suitable for computing Ritz pairs that are good approximations to eigenpairs of a given matrix $A \in \mathbb{F}^{n\times n}$? Recalling the power iteration from Section 4.8, we know from Theorem 4.2 that under some mild assumptions the angles between $A^m x$ and an eigenvector associated with the dominant eigenvalue converge to zero. We can therefore expect that for a given starting vector x the subspace spanned by $x, Ax, A^2x, \ldots, A^m x$ will contain at least good approximations to eigenvectors associated with the dominant eigenvalue of A. Given the fact that multiplying a given vector by A may be the only way to obtain information on the matrix A (cf. (7.3)), this kind of subspaces seems to be a natural candidate for the use of projection methods.

Definition 7.4 (Krylov subspace). Let $A \in \mathbb{F}^{n\times n}$ and $x \in \mathbb{F}^n$. Then

$$\mathcal{K}_m(A, x) := \operatorname{span}(x, Ax, \ldots, A^{m-1}x)$$

is called the *Krylov subspace* of order m generated by A and x.

The name refers to the Russian mathematician Alexei Krylov who used the sequence $(x, Ax, A^2x, \ldots)$ to find the coefficients of the characteristic polynomial of a given matrix A, see [25]. Obviously, we always have $\dim \mathcal{K}_m(A, x) \leq m$, but the dimension can be considerably smaller than m. For example, if $x \neq 0$ is an eigenvector of A, then we clearly have $\dim \mathcal{K}_m(A, x) = 1$ for all $m \in \mathbb{N}$.

Theorem 7.5 (Dimension of Krylov subspaces). *Let $A \in \mathbb{F}^{n\times n}$ and $x \in \mathbb{F}^n$. Then there exists a number $\ell \in \mathbb{N}_0$ with the following properties:*

1. $\dim \mathcal{K}_m(A,x) = m$ *for all* $m \le \ell$;
2. $\mathcal{K}_\ell(A,x) = \mathcal{K}_m(A,x)$ *and* $\dim \mathcal{K}_m(A,x) = \ell$ *for all* $\ell \le m$.

In particular, $\mathcal{K}_\ell(A,x)$ is an invariant subspace with respect to A.

Proof. 1. If $x = 0$, then there is nothing to show, and we have $\ell = 0$. If $x \neq 0$, then let $\ell \ge 1$ be the largest integer for which $x, Ax, \dots, A^{\ell-1}x$ are linearly independent. Then clearly $\ell \le n$ as $\dim \mathbb{F}^n = n$. As long as $m \le \ell$, it follows that the vectors $x, Ax, \dots, A^{m-1}x$ are also linearly independent, and we obtain $\dim \mathcal{K}_m(A,x) = m$.

2. Next let $\ell \le m$. By the construction of ℓ, it follows that $x, Ax, \dots, A^{\ell}x$ are linearly dependent, i.e., there exist scalars $\alpha_0, \dots, \alpha_\ell$, not all being zero, such that

$$\alpha_0 x + \alpha_1 Ax + \cdots + \alpha_\ell A^\ell x = 0.$$

Then $\alpha_\ell \neq 0$, because otherwise the linear independence of $x, Ax, \dots, A^{\ell-1}x$ would imply $\alpha_0 = \cdots = \alpha_{\ell-1} = 0$ in contradiction to the assumption that not all α_i, $i = 0, \dots, \ell$ are zero. But then, we have

$$A^\ell x = -\frac{\alpha_0}{\alpha_\ell} x - \cdots - \frac{\alpha_{\ell-1}}{\alpha_\ell} A^{\ell-1}x, \tag{7.4}$$

i.e., $A^\ell x$ is a linear combination of $x, Ax, \dots, A^{\ell-1}x$.

We now show by induction on $m \ge \ell$ that $\mathcal{K}_\ell(A,x) = \mathcal{K}_m(A,x)$ which by 1. immediately implies $\dim \mathcal{K}_m(A,x) = \ell$. The case $m = \ell$ is clear, so for the induction step assume $\ell < m$. By construction of the Krylov subspaces and $\ell < m$, it follows that $\mathcal{K}_\ell(A,x) \subseteq \mathcal{K}_m(A,x)$. By multiplying (7.4) on both sides with $A^{m-\ell}$, we obtain

$$A^m x = -\frac{\alpha_0}{\alpha_\ell} A^{m-\ell}x - \cdots - \frac{\alpha_{\ell-1}}{\alpha_\ell} A^{m-1}x$$

and thus $A^m x \in \operatorname{span}(A^{m-\ell}x, \dots, A^{m-1}x)$. By the induction hypothesis, we have $\mathcal{K}_\ell(A,x) = \mathcal{K}_{m-1}(A,x)$ which implies $x, Ax, \dots, A^{m-1}x \in \mathcal{K}_\ell(A,x)$. But then we also have $A^m x \in \mathcal{K}_\ell(A,x)$ which finally proves $\mathcal{K}_\ell(A,x) = \mathcal{K}_m(A,x)$. The proof that $\mathcal{K}_\ell(A,x)$ is an invariant subspace with respect to A is left as an exercise. □

Exercise 7.6 (Properties of Krylov subspaces). Let $A \in \mathbb{F}^{n\times n}$ and $x \in \mathbb{F}^n \setminus \{0\}$. Show that $\mathcal{K}_\ell(A,x)$ is the *smallest* invariant subspace (with respect to A) that contains x, that is, $\mathcal{K}_\ell(A,x)$ is an invariant subspace containing x, and any other invariant subspace $\mathcal{X}$ containing x satisfies $\mathcal{K}_\ell(A,x) \subseteq \mathcal{X}$.

For any polynomial $p(t) = \beta_\ell t^\ell + \cdots + \beta_1 t + \beta_0$ with coefficients in $\mathbb{F}$, we can define a polynomial map $p : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ by

$$p(A) := \beta_\ell A^\ell + \cdots + \beta_1 A + \beta_0 I_n,$$

where $A \in \mathbb{F}^{n\times n}$. The vector space of all polynomials over $\mathbb{F}$ will be denoted by $\mathbb{F}[t]$ and by $\mathbb{F}_m[t]$ we denote the subspace of all polynomials of degree not exceeding m. Then we have the following characterization of Krylov subspaces.

Lemma 7.7 (Structure of Krylov subspaces). *Let $A \in \mathbb{F}^{n\times n}$ and $x \in \mathbb{F}^n$. Then*

$$\mathcal{K}_m(A, x) = \{p(A)x \mid p \in \mathbb{F}_{m-1}[t]\}. \tag{7.5}$$

Proof. Let $y \in \mathcal{K}_m(A, x) = \operatorname{span}(x, Ax, \ldots, A^{m-1}x)$. Then there exist coefficients $\beta_0, \ldots, \beta_{m-1} \in \mathbb{F}$ such that

$$y = \beta_0 x + \beta_1 Ax + \cdots + \beta_{m-1} A^{m-1}x = p(A)x, \tag{7.6}$$

where $p(t) = \beta_{m-1}t^{m-1} + \cdots + \beta_1 t + \beta_0$ which proves "$\subseteq$". On the other hand, if we start with an arbitrary polynomial $p(t) = \beta_{m-1}t^{m-1} + \cdots + \beta_1 t + \beta_0$, then the identity in (7.6) implies $p(A)x \in \mathcal{K}_m(A, x)$. □

Let $A \in \mathbb{F}^{n\times n}$, $x \in \mathbb{F}^n \setminus \{0\}$, and let ℓ be as in Theorem 7.5. Then from (7.4) one finds that there exist coefficients $\beta_0, \ldots, \beta_{\ell-1} \in \mathbb{F}$ such that

$$A^\ell x = -\beta_0 x - \cdots - \beta_{\ell-1} A^{\ell-1}x,$$

where $\beta_i = \frac{\alpha_i}{\alpha_\ell}$. This equation can also be written as

$$p(A)x = 0, \quad \text{where } p(t) = t^\ell + \beta_{\ell-1}t^{\ell-1} + \cdots + \beta_1 t + \beta_0. \tag{7.7}$$

As $x, Ax, \ldots, A^{\ell-1}x$ are linearly independent, the coefficients $\beta_0, \ldots, \beta_{\ell-1} \in \mathbb{F}$ are uniquely determined and $q(A)x \neq 0$ for all polynomials $q \in \mathbb{F}[t]$ of degree less than ℓ. Thus, p in (7.7) is the unique *monic polynomial of minimal degree* satisfying $p(A)x = 0$ and it is called the *minimal polynomial* of x with respect to A. (A polynomial is called *monic* if the leading coefficient is equal to one.)

If $\ell = n$, then the minimal polynomial of x with respect to A is equal to the characteristic polynomial p_A of A, because the theorem of Cayley–Hamilton implies that $p_A(A) = 0$. This fact follows immediately from the uniqueness of the minimal polynomial and the characteristic polynomial being monic.

Exercise 7.8 (Minimal polynomials). Let $A \in \mathbb{F}^{n\times n}$ and $x \in \mathbb{F}^n$. Show that the minimal polynomial p of x with respect to A is a divisor of the characteristic polynomial p_A of A, i.e., there exists a polynomial $q \in \mathbb{F}[t]$ such that $p_A = p \cdot q$.

Hint: Use the remainder theorem, i.e., for nonzero polynomials p_1, p_2, there exist unique polynomials q, r such that $p_2 = p_1 \cdot q + r$ and $\deg r < \deg p_1$.

Exercise 7.9 (Dimensions of Krylov subspaces). Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & c \end{pmatrix},$$

where $c \in \mathbb{R}$. Show that there exists vectors x such that $\mathcal{K}_3(A, x) = \mathbb{R}^3$ if and only if $c \neq 1, 2$. For these values of c, determine all vectors y for which $\mathcal{K}_3(A, y) \neq \mathbb{R}^3$.

7.3 Gram–Schmidt process

In order to compute Ritz pairs with respect to a subspace $\mathcal{X}$ of $\mathbb{F}^n$, we need to compute an orthonormal basis for $\mathcal{X}$ by Theorem 7.3—a problem that we already encountered in Section 4.7, where we used Householder orthogonalization for the computation of orthonormal bases. In our particular situation, we are interested in computing an orthonormal basis of a Krylov subspace $\mathcal{K}_k(A, x) = \operatorname{span}(x, Ax, \ldots, A^{k-1}x)$ for many values of k. Suppose that $\dim \mathcal{K}_k(A, x) = k$ and that we have constructed an orthonormal basis $q_1, \ldots, q_k$ of $\mathcal{K}_k(A, x)$. If we then consider the next larger Krylov subspace $\mathcal{K}_{k+1}(A, x)$ (assuming $\dim \mathcal{K}_{k+1}(A, x) = k + 1$), we observe that

$$\mathcal{K}_{k+1}(A, x) = \operatorname{span}(x, Ax, \ldots, A^{k-1}x, A^k x) = \operatorname{span}(q_1, \ldots, q_k, A^k x).$$

Thus, in order to obtain an orthonormal basis of $\mathcal{K}_{k+1}(A, x)$, the only task is to orthonormalize the vector $A^k x$ against the previously constructed vectors $q_1, \ldots, q_k$. If we do this for $k = 1, \ldots, m$ (assuming $\dim \mathcal{K}_k(A, x) = k$ for $k = 1, \ldots, m$), we will finally obtain an orthonormal basis $q_1, \ldots, q_m$ of $\mathcal{K}_m(A, x)$ that satisfies

$$\mathcal{K}_k(A, x) = \operatorname{span}(x, Ax, \ldots, A^{k-1}x) = \operatorname{span}(q_1, \ldots, q_k), \quad k = 1, \ldots, m. \tag{7.8}$$

For this purpose, the so-called *Gram–Schmidt process* is more appropriate than Householder orthogonalization.

For deriving the corresponding algorithm in general, let us assume that we have m linearly independent vectors $v_1, \ldots, v_m \in \mathbb{F}^n$. Our task is to construct an orthonormal basis $q_1, \ldots, q_m$ of $\mathcal{V} = \operatorname{span}(v_1, \ldots, v_m)$ such that

$$\operatorname{span}(v_1, \ldots, v_k) = \operatorname{span}(q_1, \ldots, q_k), \quad k = 1, \ldots, m.$$

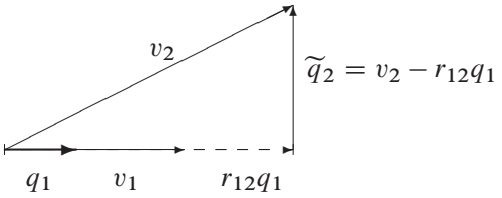

Figure 7.3. Second step in the Gram–Schmidt process interpreted geometrically.

The first step is immediate, we only have to normalize the vector v_1:

$$q_1 := \frac{1}{\|v_1\|} v_1.$$

For the second step, we look for a vector $\widetilde{q}_2 \in \operatorname{span}(v_1, v_2) = \operatorname{span}(q_1, v_2)$ that is orthogonal to q_1. Note that $\widetilde{q}_2$ must have a component in the direction v_2 (that is $\widetilde{q}_2 = \alpha_1 q_1 + \alpha_2 v_2$, where $\alpha_2 \neq 0$), because otherwise q_1 and $\widetilde{q}_2$ would be linearly dependent. As we have to normalize the vector $\widetilde{q}_2$ anyhow, we may assume without loss of generality that $\alpha_2 = 1$, i.e., $\widetilde{q}_2$ has the form

$$\widetilde{q}_2 := v_2 - r_{12}q_1, \tag{7.9}$$

for some $r_{12} \in \mathbb{F}$. (The minus sign in (7.9) has been introduced for notational convenience—this will become clear later.) Then the requirement that $\widetilde{q}_2$ be orthogonal to q_1 implies

$$0 = \langle \widetilde{q}_2, q_1 \rangle = \langle v_2 - r_{12}q_1, q_1 \rangle = \langle v_2, q_1 \rangle - r_{12},$$

because $\langle q_1, q_1 \rangle = 1$. Thus, we obtain the vector q_2 by

$$\begin{aligned} r_{12} &:= \langle v_2, q_1 \rangle && (7.10)\\ \widetilde{q}_2 &:= v_2 - \langle v_2, q_1 \rangle q_1, && (7.11)\\ r_{22} &:= \|\widetilde{q}_2\| \\ q_2 &:= \frac{1}{r_{22}} \widetilde{q}_2. \end{aligned}$$

The identity (7.11) can be interpreted geometrically in the following way: the vector $r_{12}q_1$ is the orthogonal projection of the vector v_2 onto the subspace generated by q_1. Thus, we obtain by $v_2 - r_{12}q_1$ a vector that is orthogonal to q_1, see Figure 7.3.

Assuming we have already computed orthonormal vectors $q_1, \ldots, q_{k-1}$ satisfying

$$\operatorname{span}(v_1, \ldots, v_j) = \operatorname{span}(q_1, \ldots, q_j), \quad j = 1, \ldots, k-1,$$

```
procedure gram_schmidt_procedure(v_1, ..., v_m, var R, q_1, ..., q_m);
begin
  for k ∈ {1, ..., m} do begin
    for i ∈ {1, ..., k − 1} do
      r_ik ← ⟨v_k, q_i⟩;
    q_k ← v_k;
    end
    for i ∈ {1, ..., k − 1} do
      q_k ← q_k − r_ik q_i;
    r_kk ← ‖q_k‖;
    end
    if r_kk = 0 then
      STOP
    else
      q_k ← (1/r_kk) q_k
  end
end
```

Figure 7.4. Gram–Schmidt procedure for vectors $v_1, \ldots, v_m \in \mathbb{F}^n$.

we now look for a vector $\widetilde{q}_k \in \operatorname{span}(v_1, \ldots, v_k) = \operatorname{span}(q_1, \ldots, q_{k-1}, v_k)$ orthogonal to $q_1, \ldots, q_{k-1}$. Analogous to the second step, we may assume that $\widetilde{q}_k$ has the form

$$\widetilde{q}_k = v_k - r_{1k} q_1 - \cdots - r_{k-1,k} q_{k-1}, \tag{7.12}$$

where $r_{1k}, \ldots, r_{k-1,k} \in \mathbb{F}$. From $\langle \widetilde{q}_k, q_j \rangle = 0$ for $j = 1, \ldots, k-1$, we obtain

$$r_{jk} = \langle v_k, q_j \rangle, \quad j = 1, \ldots, k-1,$$

using $\langle q_i, q_j \rangle = \delta_{ji}$. Here, the vector $r_{1k} q_1 + \cdots + r_{k-1,k} q_{k-1}$ can be interpreted as the orthogonal projection of v_k onto the subspace spanned by $q_1, \ldots, q_{k-1}$. We then obtain the vector q_k as follows.

$$\begin{aligned} r_{jk} &:= \langle v_k, q_j \rangle, \quad j = 1, \ldots, k-1 \\ \widetilde{q}_k &:= v_k - r_{1k} q_1 - \cdots - r_{k-1,k} q_{k-1}, \\ r_{kk} &:= \|\widetilde{q}_k\| \\ q_k &:= \frac{1}{r_{kk}} \widetilde{q}_k. \end{aligned}$$

The resulting algorithm is given in Figure 7.4.

The algorithm breaks down if $r_{kk} = 0$ in some step. But this would mean $\widetilde{q}_k = 0$ in (7.12) showing that $v_k \in \operatorname{span}(q_1, \ldots, q_{k-1}) = \operatorname{span}(v_1, \ldots, v_{k-1})$, i.e., $v_1, \ldots, v_k$

are linearly dependent. Thus, in exact arithmetic a breakdown will never occur as long as the vectors $v_1, \dots, v_m$ are linearly independent.

Remark 7.10 (Orthonormal bases for nested sets of subspaces). The application of the Gram–Schmidt process to the linearly independent vectors $v_1, \dots, v_m \in \mathbb{F}^n$ produces orthonormal vectors $q_1, \dots, q_m \in \mathbb{F}^n$ with the additional property

$$\operatorname{span}(v_1, \dots, v_k) = \operatorname{span}(q_1, \dots, q_k), \quad k = 1, \dots, m. \tag{7.13}$$

Remark 7.11 (Gram–Schmidt and QR factorization). Rewriting identity (7.12) we obtain

$$v_k = r_{1k}q_1 + \cdots + r_{k-1,k}q_{k-1} + \widetilde{q}_k = r_{1k}q_1 + \cdots + r_{kk}q_k$$

for $k = 1, \dots, m$. Setting $A = [v_1, \dots, v_m] \in \mathbb{F}^{n\times m}$, $Q = [q_1, \dots, q_m] \in \mathbb{F}^{n\times m}$ and

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ 0 & r_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & r_{mm} \end{pmatrix} \in \mathbb{F}^{m\times m},$$

we obtain $A = QR$ which is a skinny QR factorization of A as introduced in Section 4.7. Thus, the Gram–Schmidt process is an alternative method for computing a QR factorization of a matrix $A \in \mathbb{F}^{n\times m}$ if the columns of A are linearely independent.

Exercise 7.12 (Orthonormal bases for nested sets of subspaces). Let $v_1, \dots, v_m$ denote the columns of the matrix $A \in \mathbb{F}^{n\times m}$ and let $A = QR$ be a QR factorization of A, where we denote the columns of the isometric matrix $Q \in \mathbb{F}^{n\times m}$ by $q_1, \dots, q_m$. Show that if the columns of A are linearly independent then the identity (7.13) holds. What can be said if the columns of A are linearly dependent?

Unfortunately, the numerical behaviour of the Gram–Schmidt procedure is not too good: in finite precision arithmetic, the computed vectors $q_1, \dots, q_m$ typically show a significant loss in orthogonality due to roundoff errors. However, a slight modification of the algorithm already improves its performance considerably. Instead of orthogonally projecting v_k onto the subspace spanned by $q_1, \dots, q_k$, we first project v_k onto the subspace spanned by q_1 and then subtract the projection from v_k to obtain a vector $\widetilde{q}_k^{(1)}$ that is orthogonal with respect to q_1. Then $\widetilde{q}_k^{(1)}$ is projected onto the subspace spanned by q_2 and this projection is subtracted from $\widetilde{q}_k^{(1)}$ to obtain a vector $\widetilde{q}_k^{(2)}$ which is orthogonal to q_1, q_2. This procedure is then iterated, i.e., we compute

$$\widetilde{q}_k^{(j)} = \widetilde{q}_k^{(j-1)} - \langle \widetilde{q}_k^{(j-1)}, q_j \rangle q_j, \quad j = 1, \dots, k-1$$

```
procedure modified_gram_schmidt_procedure(v_1, ..., v_m, var R, q_1, ..., q_m);
begin
  for k ∈ {1, ..., m} do begin
    q_k ← v_k;
    for i ∈ {1, ..., k − 1} do begin
      r_ik ← ⟨q_k, q_i⟩;
      q_k ← q_k − r_ik q_i
    end;
    r_kk ← ‖q_k‖;
    if r_kk = 0 then
      STOP
    else
      q_k ← (1/r_kk) q_k
  end
end
```

Figure 7.5. Modified Gram–Schmidt procedure for vectors $v_1, \ldots, v_m \in \mathbb{F}^n$.

starting with $\widetilde{q}_k^{(0)} = v_k$. In exact arithmetic this procedure is equivalent to the Gram–Schmidt procedure as we can show by induction that

$$\widetilde{q}_k^{(j)} = v_k - \sum_{i=1}^{j} \langle v_k, q_j \rangle q_i, \qquad j = 0, \ldots, k-1.$$

Indeed, for $j = 0$ there is nothing to show and using the induction hypothesis for $j - 1$, we obtain that

$$\begin{aligned} \widetilde{q}_k^{(j)} &= \widetilde{q}_k^{(j-1)} - \langle \widetilde{q}_k^{(j-1)}, q_j \rangle q_j \\ &= v_k - \sum_{i=1}^{j-1} \langle v_k, q_j \rangle q_i - \left\langle v_k - \sum_{i=1}^{j-1} \langle v_k, q_j \rangle q_i, q_j \right\rangle q_j \\ &= v_k - \sum_{i=1}^{j} \langle v_k, q_j \rangle q_i, \end{aligned}$$

because $\langle q_i, q_j \rangle = \delta_{ij}$. Thus, after $k - 1$ steps we have $\widetilde{q}_k^{(k-1)} = \widetilde{q}_k$, where $\widetilde{q}_k$ is as in (7.12). The corresponding algorithm, the so-called *modified Gram–Schmidt process*, is given in Figure 7.5.

A heuristic explanation why the modified Gram–Schmidt process performs better than the classical Gram–Schmidt process is that in the computation of $\widetilde{q}_k^{(j)}$ the vector $\widetilde{q}_k^{(j)}$ is also orthogonalized to the errors made in the computation of $\widetilde{q}_k^{(j-1)}$.

```
procedure mgs_with_reorthogonalization(v_1, ..., v_m, var R, q_1, ..., q_m);
begin
  for k ∈ {1, ..., m} do begin
    q_k ← v_k;
    for i ∈ {1, ..., k − 1} do begin
      r_ik ← ⟨q_k, q_i⟩;
      q_k ← q_k − r_ik q_i
    end;
    for i ∈ {1, ..., k − 1} do begin
      s_i ← ⟨q_k, q_i⟩;
      q_k ← q_k − s_i q_i;
      r_ik ← r_ik + s_i
    end;
    r_kk ← ‖q_k‖;
    if r_kk = 0 then
      STOP
    else
      q_k ← (1/r_kk) q_k
  end
end
```

Figure 7.6. Modified Gram–Schmidt procedure with reorthogonalization.

Exercise 7.13 (Gram–Schmidt procedure). Use the Gram–Schmidt procedure and the modified Gram–Schmidt procedure to compute the QR factorization of the matrix

$$A = \begin{pmatrix} 1 & -5 & 4 \\ 2 & -4 & 8 \\ 2 & 2 & -1 \end{pmatrix}.$$

Unfortunately, even the modified Gram–Schmidt process still shows a significant loss in orthogonality in finite precision arithmetic, in particular, if the vectors $v_1, \ldots, v_m$ are nearly linearly dependent. Nevertheless, the computed basis vectors are less likely to be nearly linearly independent than the original vectors $v_1, \ldots, v_m$, so the key idea is *reorthogonalization*. Thus, after orthogonalizing v_k against $q_1, \ldots, q_{k-1}$, we orthogonalize the computed vector $\widetilde{q}_k$ again with respect to the vectors $q_1, \ldots, q_{k-1}$. Thus, using the classical Gram–Schmidt procedure, we compute

$$\widetilde{q}_k = v_k - r_{1k} q_1 - \cdots - r_{k-1,k} q_{k-1}, \tag{7.14}$$

$$\widetilde{q}_k^{(2)} = \widetilde{q}_k - s_{1k} q_1 - \cdots - s_{k-1,k} q_{k-1}, \tag{7.15}$$

	G.S.	modified G.S.	mod. G.S. with reorthogonaliz.
$n = 5$	5.5×10^{-08}	1.1×10^{-11}	2.8×10^{-16}
$n = 10$	3.0×10^{-00}	1.5×10^{-04}	4.4×10^{-16}
$n = 20$	12.7×10^{-00}	1.0×10^{-00}	5.9×10^{-16}

Table 7.1. Defect from orthogonality in some variants of the Gram–Schmidt process. The results were obtained in Matlab 7.11.0 with unit roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$.

with $r_{ik} = \langle v_k, q_i \rangle$ and $s_{ik} = \langle \widetilde{q}_k, q_i \rangle$. Note that combining (7.14) and (7.15) yields

$$\widetilde{q}_k^{(2)} = v_k - (r_{1k} + s_{1k})q_1 - \cdots - (r_{k-1,k} + s_{k-1,k})q_{k-1},$$

so the values r_{ij} have to be updated to $r_{ij} + s_{ij}$ in the algorithm to yield the corresponding QR factorization. In a similar way, reorthogonalization can be performed in the modified Gram–Schmidt process. The corresponding algorithm is given in Figure 7.6.

One may think that reorthogonalizing the computed vector $\widetilde{q}_k$ multiple times against $q_1, \ldots, q_{k-1}$ would produce an even better result in finite precision arithmetic. However, it has been observed empirically that in general it suffices to do the reorthogonalization once. We refer the reader to [49] and [17] for details.

Example 7.14 (Loss of orthogonality). Let $H_n \in \mathbb{R}^{n \times n}$ denote the $n \times n$ Hilbert matrix given by

$$H_n = \left[\frac{1}{i + j + 1} \right]_{i,j=1,\ldots,n} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{pmatrix}. \tag{7.16}$$

Applying the Gram–Schmidt procedure, the modified Gram–Schmidt procedure, and the modified Gram–Schmidt procedure with reorthogonalization to the columns of H_n and assembling the computed orthonormal basis vectors $q_1, \ldots, q_n$ produces a matrix $Q_n = [q_1, \ldots, q_n] \in \mathbb{R}^{n \times n}$ with defect from orthogonality as given in Table 7.1. As a measure of the defect from orthogonality, we computed $\| I_n - Q_n^* Q_n \|_2$.

The *Hilbert matrices* H_n from Example 7.14 are highly *ill-conditioned* in the sense that the columns of H_n are nearly linearly dependent. As a consequence, the inverse of H_n contains very large entries even for moderate values of n. Thus, Hilbert matrices represent a challenge to numerical algorithms and are often used as test matrices in numerical experiments.

Exercise 7.15 (Hilbert matrices). Compute the inverse of the Hilbert matrices H_2 and H_3 from (7.16).

7.4 Arnoldi iteration

Let $A \in \mathbb{F}^{n\times n}$ be a matrix and let us assume that for a suitable vector $x \in \mathbb{F}^n$ we have $\dim \mathcal{K}_m(A,x) = m$. In view of Theorem 7.3, we may then compute an orthonormal basis of the Krylov subspace $\mathcal{K}_m(A,x)$ in order to compute Ritz pairs of A with respect to $\mathcal{K}_m(A,x)$. A naive approach to compute such a basis would be to apply the Gram Schmidt process (with reorthogonalization) to the sequence $x, Ax, A^2x, \ldots, A^{m-1}x$. As we know from Remark 7.10 this will produce orthonormal vectors $q_1, \ldots, q_m$ satisfying

$$\mathcal{K}_k(A,x) = \operatorname{span}(x, Ax, \ldots, A^{k-1}x) = \operatorname{span}(q_1, \ldots, q_k), \quad k = 1, \ldots, m. \tag{7.17}$$

However, we know from Chapter 4 that the sequence $(A^k x)_{k=0}^{\infty}$ converges under appropriate assumptions to an eigenvector e_1 associated with the dominant eigenvalue, in the sense that the corresponding sequence of angles between the iterates and e_1 converges to zero. Thus, the set of vectors $x, Ax, A^2x, \ldots, A^{m-1}x$ may be in theory linearly independent, but numerically it may be linearly dependent, because the directions of the vectors are "not different enough".

Fortunately, a slight modification of our problem allows us to avoid this situation of nearly linearly independent vectors challenging our orthonormalization algorithm. To see this, let us assume that $q_1, \ldots, q_k$ in (7.17) have already been constructed. Thus, it remains to orthogonalize the vector $A^k x$ against the vectors $q_1, \ldots, q_k$ in order to obtain an orthonormal basis of $\mathcal{K}_{k+1}(A,x)$. Since

$$A^{k-1}x \in \mathcal{K}_k(A,x) = \operatorname{span}(q_1, \ldots, q_k),$$

there exist scalars $\alpha_1, \ldots, \alpha_k$ such that

$$A^{k-1}x = \alpha_1 q_1 + \cdots + \alpha_k q_k$$

which implies

$$A^k x = \alpha_1 A q_1 + \cdots + \alpha_k A q_k.$$

Note that $Aq_1, \dots, Aq_{k-1} \in \mathcal{K}_k(A, x)$, because $q_1, \dots, q_{k-1} \in \mathcal{K}_{k-1}(A, x)$. Thus, with $A^k x$ also Aq_k must be contained in $\mathcal{K}_{k+1}(A, x) \setminus \mathcal{K}_k(A, x)$ from which we obtain that

$$\mathcal{K}_{k+1}(A, x) = \operatorname{span}(q_1, \dots, q_k, A^k x) = \operatorname{span}(q_1, \dots, q_k, Aq_k).$$

Hence, instead of orthogonalizing $A^k x$ we can orthogonalize the vector Aq_k with respect to the vectors $q_1, \dots, q_k$. In exact arithmetic, there is no difference between these two strategies, but numerically, the vector Aq_k will be more "distinct" from the vectors $q_1, \dots, q_k$ in the sense explained above which may have a significant effect in finite precision arithmetic. Applying now one step of Gram Schmidt to Aq_k, we obtain

$$\widetilde{q}_{k+1} = Aq_k - \sum_{i=1}^{k} \langle Aq_k, q_i \rangle q_i = Aq_k - \sum_{i=1}^{k} h_{ik} q_i, \tag{7.18}$$

where we used the abbreviation $h_{ik} := \langle Aq_k, q_i \rangle$. Finally normalizing the vector $\widetilde{q}_{k+1}$ by

$$q_{k+1} = \frac{1}{h_{k+1,k}} \widetilde{q}_{k+1}, \quad \text{where } h_{k+1,k} = \|\widetilde{q}_{k+1}\| \tag{7.19}$$

we obtain an orthonormal basis $q_1, \dots, q_{k+1}$ of $\mathcal{K}_{k+1}(A, x)$. For the practical implementation, it is advisable to prefer the Gram Schmidt procedure with reorthogonalization. The corresponding algorithm is given in Figure 7.7 and is called *Arnoldi iteration* [1].

Let us denote by $Q_m \in \mathbb{F}^{n \times m}$ the isometric matrix with columns $q_1, \dots, q_m$. Then the eigenpairs of the matrix $H_m := Q_m^* A Q_m$ are the Ritz pairs of A with respect to $\mathcal{K}_m(A, x)$. It turns out that this matrix H_m has a special structure.

Theorem 7.16 (Krylov subspaces and Hessenberg matrices). *Let $A \in \mathbb{F}^{n \times n}$, $m \in \mathbb{N}$ and let $x \in \mathbb{F}^n$ be such $\dim \mathcal{K}_m(A, x) = m$. If $Q_m = [q_1, \dots, q_m] \in \mathbb{F}^{n \times m}$ is an isometric matrix such that (7.17) is satisfied, then $H_m = [h_{ij}] := Q_m^* A Q_m$ is an $m \times m$ Hessenberg matrix, i.e., $h_{ij} = 0$ for all $i > j + 1$, $j = 1, \dots, m - 1$.*

Proof. Let $j \in \{1, \dots, m-1\}$. Then by (7.17) we have $q_j \in \mathcal{K}_j(A, x)$ and thus

$$Aq_j \in \mathcal{K}_{j+1}(A, x) = \operatorname{span}(q_1, \dots, q_{j+1}).$$

Since the vectors $q_1, \dots, q_m$ are orthonormal, we obtain that $h_{ij} = \langle Aq_j, q_i \rangle = 0$ for $i = j + 2, \dots, m$. Thus, H_m is in Hessenberg form. □

Remark 7.17 (Possible breakdown of the Arnoldi iteration). Note that the Arnoldi iteration will break down if $h_{k+1,k} = 0$ for some $k < m$. By (7.18), we have $h_{k+1,k} \neq 0$ whenever $Aq_k \in \mathcal{K}_{k+1}(A, x) \setminus \mathcal{K}_k(A, x)$. If $\dim \mathcal{K}_m(A, x) = m$ then this is guaranteed for $k < m$. Under this hypothesis, the Arnoldi iteration will run without breakdown for at least $m - 1$ steps. For the mth step, however, there are two possibilities:

```
procedure arnoldi_iteration(A, x, var H, q_1, ..., q_m);
begin
  γ ← ‖x‖;
  q_1 ← (1/γ)x;
  for k ∈ {1, ..., m} do begin
    q_{k+1} ← Aq_k;
    for i ∈ {1, ..., k} do begin
      h_{i,k} ← ⟨q_{k+1}, q_i⟩;
      q_{k+1} ← q_{k+1} − h_{ik} q_i
    end;
    for i ∈ {1, ..., k} do begin
      s_i ← ⟨q_{k+1}, q_i⟩;
      q_{k+1} ← q_{k+1} − s_i q_i;
      h_{i,k} ← h_{i,k} + s_i
    end;
    h_{k+1,k} ← ‖q_{k+1}‖;
    if h_{k+1,k} = 0 then
      STOP
    else
      q_{k+1} ← (1/h_{k+1,k}) q_{k+1}
  end
end
```

Figure 7.7. Arnoldi iteration (using modified Gram–Schmidt with reorthogonalization).

1. If $h_{m+1,m} = 0$, then the algorithm breaks down and we have

$$Aq_k = \sum_{i=1}^{k+1} h_{ik} q_i, \quad \text{for } k = 1, \ldots, m-1,$$
$$Aq_m = \sum_{i=1}^{m} h_{im} q_i,$$

and thus $Aq_k \in \mathcal{K}_m(A, x) = \operatorname{span}(q_1, \ldots, q_k)$ for $k = 1, \ldots, m$. This implies that $\mathcal{K}_m(A, x)$ is an A-invariant subspace and hence, all eigenvalues of $H_m = Q_m^* A Q_m$, $Q = [q_1, \ldots, q_m]$ are also (exact) eigenvalues of A, see Exercise 2.44. Thus, a breakdown of the Arnoldi iteration is actually a preferable situation, because it gives us an orthonormal basis of an invariant subspace.

2. If $h_{m+1,m} \neq 0$, then we can continue to perform the mth step of the Arnoldi iteration to produce the next iterate q_{m+1}. Then we have

$$Aq_k = \sum_{i=1}^{k+1} h_{ik} q_i, \quad \text{for } k = 1, \ldots, m,$$

or, equivalently,

$$A\,[\,q_1, \ldots, q_m\,] = [\,q_1, \ldots, q_m, q_{m+1}\,] \left(\begin{array}{cccc} h_{11} & \ldots & \ldots & h_{1n} \\ h_{21} & h_{22} & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & h_{m,m-1} & h_{mm} \\ \hline 0 & \ldots & 0 & h_{m+1,m} \end{array}\right). \tag{7.20}$$

Note that the upper $m \times m$ block of the $(m+1) \times m$ matrix in (7.20) is just H_m and that the $(m+1)$st row is $h_{m+1,m}\delta_m^T$, where δ_m denotes the mth canonical unit vector of length m. Thus, we can write (7.20) in the shorter form

$$AQ_m = Q_m H_m + h_{m+1,m} q_{m+1} \delta_m^T \tag{7.21}$$

which is sometimes called an *Arnoldi configuration*.

Exercise 7.18 (Arnoldi configuration). Let $A \in \mathbb{F}^{n\times n}$, let $q_1, \ldots, q_m, q_{m+1} \in \mathbb{F}^n$ be linearly independent (but not necessarily orthonormal) such that (7.20) holds with $h_{k+1,k} \neq 0$ for $k = 1, \ldots, m+1$. Show that

$$\operatorname{span}(q_1, \ldots, q_k) = \mathcal{K}_k(A, q_1)$$

for $k = 1, \ldots, m+1$.

Exercise 7.19 (Krylov subspaces and Hessenberg matrices). Let $A \in \mathbb{F}^{n\times n}$ and let $Q = [q_1, \ldots, q_n] \in \mathbb{F}^{n\times n}$ be an invertible matrix such that $H = Q^{-1}AQ$ is in Hessenberg form. Show that

$$\operatorname{span}(q_1, \ldots, q_k) = \mathcal{K}_k(A, q_1)$$

for $k = 1, \ldots, n$ if and only if $h_{k+1,k} \neq 0$ for $k = 1, \ldots, n-1$.

Remark 7.20 (Computation of Ritz pairs). In both of the scenarios described in Remark 7.17, we are left with the task of computing the eigenvalues of the matrix $H_m = Q_m^* A Q_m$. Since H_m is already in Hessenberg form, those eigenvalues can be efficiently computed by the QR iteration discussed in Chapter 5. If μ is a computed eigenvalue, then we can compute a corresponding eigenvector $v \in \mathbb{F}^m$ by using inverse iteration with shift μ for an arbitrary starting vector, see Section 4.4. In practice, usually one or at most two iterations are sufficient.

Once we have successfully computed Ritz pairs, the question arises whether a given Ritz pair $(\lambda, v) \in \mathbb{F} \times \mathcal{K}_m(A, x)$ is a good approximation to an eigenpair. We answer this question by slightly extending the notion of *residual* from Definition 4.7.

Definition 7.21 (Residual). Let $A \in \mathbb{F}^{n\times n}$ and let $(\lambda, v) \in \mathbb{F} \times \mathbb{F}^n \setminus \{0\}$. Then the vector

$$r := \lambda v - Av$$

is called the *residual* of the pair (λ, v) with respect to A.

Theorem 7.22 (Backward error from the residual). *Let $A \in \mathbb{F}^{n\times n}$ and $(\lambda, v) \in \mathbb{F} \times \mathbb{F}^n$ be such that $\|v\| = 1$. Then there exists a matrix $E \in \mathbb{F}^{n\times n}$ such that*

$$(A + E)v = \lambda v \quad \text{and} \quad \|E\| \leq \|\lambda v - Av\|.$$

Proof. Set $E := (\lambda v - Av)v^*$. Then we have

$$(A + E)v = Av + (\lambda v - Av)\underbrace{v^*v}_{=1} = Av + \lambda v - Av = \lambda v$$

and $\|E\| = \|(\lambda v - Av)v^*\| \leq \|\lambda v - Av\|\|v^*\| = \|\lambda v - Av\|$. □

Thus, whenever the residual $r = \lambda v - Av$ of a Ritz pair (λ, v) is small, then we have a small *backward error* in the sense that (λ, v) is an exact eigenpair of a nearby matrix. A nice property of the Arnoldi iteration is the fact that the residual can be computed very cheaply.

Theorem 7.23 (Residual of Ritz pairs). *Let $A \in \mathbb{F}^{n\times n}$ be a matrix and $Q_m \in \mathbb{F}^{n\times m}$, $H_m \in \mathbb{F}^{m\times m}$, $q_{m+1} \in \mathbb{F}^n$, and $h_{m+1,m} \in \mathbb{F}$ be as in* (7.21)*. If $(\lambda, y) \in \mathbb{F} \times (\mathbb{F}^m \setminus \{0\})$ is an eigenpair of H_m, then (λ, v) with $v = Q_m y$ is a Ritz pair of A with respect to $\mathcal{K}_m(A, x)$ satisfying*

$$\|\lambda v - Av\| = |h_{m+1,m}| \cdot |y_m|,$$

where y_m denotes the last entry of the vector y.

Proof. Let $(\lambda, y) \in \mathbb{F} \times (\mathbb{F}^m \setminus \{0\})$ be an eigenpair of H_m. By Theorem 7.3, we know that $(\lambda, Q_m y)$ is a Ritz pair of A with respect to $\mathcal{K}_m(A, x) = \mathcal{R}(Q_m)$. Moreover, we have

$$\begin{aligned}\lambda v - Av &= \lambda Q_m y - AQ_m y \\ &\overset{(7.21)}{=} \lambda Q_m y - (Q_m H_m + h_{m+1,m} q_{m+1} \delta_m^T) y \\ &= Q_m \underbrace{(\lambda y - H_m y)}_{=0} - h_{m+1,m} q_{m+1} y_m,\end{aligned}$$

where we used that $y_m = \delta_m^T y$. But then we obtain $\|\lambda v - Av\| = |h_{m+1,m}| \cdot |y_m|$. □

Thus, the residual of the Ritz pair (λ, v) can be immediately computed from the two scalars $h_{m+1,m}$ and y_m without explicitly forming the Ritz vector v and without computing the matrix-vector product Av.

7.5 Symmetric Lanczos algorithm

In the special situation that our matrix $A^* = A \in \mathbb{F}^{n\times n}$ is self-adjoint, the Arnoldi iteration can be simplified significantly. After m steps of the Arnoldi iteration, we have computed a matrix $Q_m = [q_1, \ldots, q_m]$ with columns that form an orthonormal basis of the Krylov subspace $\mathcal{K}_m(A, x)$ for some initial vector $x \in \mathbb{F}^n \setminus \{0\}$. With A, also the matrix $H_m = Q_m^* A Q_m$ is self-adjoint and has the following form:

$$H_m = \begin{pmatrix} h_{11} & h_{12} & \ldots & \ldots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & h_{m-1,m} \\ 0 & \ldots & 0 & h_{m,m-1} & h_{mm} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \ldots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & \vdots \\ 0 & \beta_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{m-1} \\ 0 & \ldots & 0 & \beta_{m-1} & \alpha_m \end{pmatrix},$$

where $\alpha_k := h_{kk} \in \mathbb{R}$, because self-adjoint matrices have real diagonal elements, and $\beta_{k-1} := h_{k,k-1} = \|\widetilde{q}_k\| \in \mathbb{R}$ by (7.19) which implies $h_{k-1,k} = h_{k,k-1}^* = \beta_{k-1}$. Thus, even in the case that we started with a *complex* self-adjoint matrix A, the matrix $H_m = Q_m^* A Q_m$ will be tridiagonal and *real*. With this notation and starting with an arbitrary vector $x \in \mathbb{F}^n \setminus \{0\}$, the equations (7.18) and (7.19) simplify to

$$\begin{aligned} \widetilde{q}_{k+1} &= Aq_k - \alpha_k q_k - \beta_{k-1} q_{k-1} \\ \beta_k &= \|\widetilde{q}_{k+1}\| \\ q_{k+1} &= \frac{1}{\beta_k} \widetilde{q}_{k+1}, \quad k = 1, \ldots, m-1, \end{aligned} \tag{7.22}$$

where $q_0 := 0$, $\beta_0 := 0$, $q_1 := \frac{1}{\|x\|} x$. This special case of the Arnoldi iteration is called the *symmetric Lanczos iteration* [27]. The resulting algorithm is given in Figure 7.8

As in the Arnoldi iteration, the Lanczos iteration may break down in some step (say the mth step) yielding an invariant subspace given by the Krylov subspace $\mathcal{K}_m(A, x)$ which is spanned by the vectors $q_1, \ldots, q_m$. Otherwise, we may compute the eigenvalues of $H_m = Q_m A Q_m$ to obtain the Ritz values of A with respect to $\mathcal{K}_m(A, x)$.

An important feature of the symmetric Lanczos iteration is the so-called *three-term-recurrence* (7.22). In each iteration, only the vectors q_k and q_{k-1} from the two previous iterations are needed. However, the full set of vectors $q_1, \ldots, q_m$ is needed for the computation of Ritz vectors and for reorthogonalization in order to avoid a loss in orthogonality of the vectors $q_1, \ldots, q_m$.

```
procedure lanczos_iteration(A, x, var q_1, ..., q_m, α_1, ..., α_m, β_1, ..., β_{m-1});
begin
  q_0 ← 0;
  β_0 ← 0;
  γ ← ‖x‖;
  q_1 ← (1/γ) x;
  for k ∈ {1, ..., m − 1} do begin
    q_{k+1} ← A q_k;
    α_k ← ⟨q_{k+1}, q_k⟩;
    q_{k+1} ← q_{k+1} − α_k q_k − β_{k−1} q_{k−1};
    β_k ← ‖q_{k+1}‖;
    if β_k = 0 then
      STOP
    else
      q_{k+1} ← (1/β_k) q_{k+1}
  end
end
```

Figure 7.8. Lanczos iteration without reorthogonalization.

If only eigenvalues are needed, then one may use the Lanczos iteration without reorthogonalization and refrain from storing the vectors $q_1, \ldots, q_m$ except for those from the two previous iterations. As a consequence, the number of steps m can be taken much higher as with reorthogonalization, but on the other hand a loss in orthogonality of the vectors $q_1, \ldots, q_m$ will occur. This will typically lead to the occurrence of *ghost eigenvalues*. These are multiple "copies" of some eigenvalues of H_m, typically those that are well approximating the smallest and largest eigenvalues of A. Thus, the Ritz values give the wrong impression that the multiplicity of the corresponding eigenvalue of A is higher than it actually is. We refer the reader to the monograph [45] for an illustrative example and a heuristic explanation of this observation.

7.6 Chebyshev polynomials

In the previous sections we have discussed how to compute orthonormal bases of Krylov subspaces, how to compute the corresponding Ritz pairs, and how to check whether these Ritz pairs are good approximations to eigenpairs. However, a question that is still open is the following. Given a matrix $A \in \mathbb{F}^{n\times n}$ and a start vector $x \in \mathbb{F}^n$,

does $\mathcal{K}_m(A, x)$ *contain any good approximations to eigenpairs of* A? And if so, which eigenpairs are most likely to be detected first? We will see that this question can best be answered by the help of a new tool, the so-called *Chebyshev polynomials* that we will introduce in this section.

Example 7.24. Let $A = \operatorname{diag}(3, 2.99, 2.98, \ldots, 2.02, 2.01, 2, 1) \in \mathbb{R}^{102\times 102}$, i.e., the diagonal matrix with eigenvalues $\lambda_k = 3 - (k-1)/100$ for $k = 1, \ldots, 101$ and $\lambda_{102} = 1$. Applying the Arnoldi iteration (or Lanczos iteration) with the start vector $x \in \mathbb{R}^{102}$ with all entries equal to one, we obtain after 10 steps the 10×10 matrix H_{10} having the eigenvalues μ_i, $i = 1, \ldots, 10$ displayed in the following table rounded to five significant digits:

i	μ_i	$\min_j \lvert\mu_i - \lambda_j\rvert$
1	2.9840	6.2933×10^{-04}
2	2.9262	3.7955×10^{-03}
3	2.8192	7.5917×10^{-04}
4	2.6786	1.4213×10^{-03}
5	2.5180	1.9794×10^{-03}
6	2.3539	3.8811×10^{-03}
7	2.2038	3.7657×10^{-03}
8	2.0851	4.8982×10^{-03}
9	2.0130	2.9747×10^{-03}
10	1.0000	3.8769×10^{-12}

Apparently, the eigenvalue $\lambda_{102} = 1$ of A in Example 7.24 is well approximated by a Ritz value of A with respect to $\mathcal{K}_{10}(A, x)$. At first, this may be a surprise, because $\mathcal{K}_{10}(A, x)$ is spanned by $x, Ax, \ldots, A^9x$ and so by the results of Chapter 4 one may expect that we find a good approximation to eigenvectors associated with the dominant eigenvalue $\lambda_1 = 3$ in $\mathcal{K}_{10}(A, x)$. However, the gap between λ_1 and λ_2 is rather small, and so the convergence of the power method for A is rather slow which explains why we did not yet find a good approximation to an eigenvector associated with $\lambda_1 = 3$ after only ten iterations. On the other hand, we see that the smallest eigenvalue λ_{102} is approximated with an absolute error as small as 3.8769×10^{-12}. How can this be explained?

For the remainder of this section, we will assume that A is self-adjoint in order to keep the discussion simple. Thus, we may assume that $A \in \mathbb{F}^{n\times n}$ has the n eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ in descending order and a corresponding orthonormal basis of eigenvectors $e_1, \ldots, e_n$ associated with $\lambda_1, \ldots, \lambda_n$. The start vector $x \in \mathbb{F}^n \setminus \{0\}$ can then be represented as a linear combination of this basis vectors, i.e., there exists

coefficients $\alpha_1, \dots, \alpha_n \in \mathbb{F}$ such that

$$x = \sum_{i=1}^{n} \alpha_i e_i.$$

Recall that $\mathcal{K}_m(A, x) = \{p(A)x | p \in \mathbb{F}_{m-1}[t]\}$ by Lemma 7.7. For an arbitrary polynomial p the vector $p(A)x$ then has the form given in the following remark.

Remark 7.25 (Elements of Krylov subspaces). Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint and let $e_1, \dots, e_n$ be a basis of $\mathbb{F}^n$ of eigenvectors associated with the eigenvalues $\lambda_1, \dots, \lambda_n$. If $x = \sum_{i=1}^{n} \alpha_i e_i \in \mathbb{F}^n$, $\alpha_1, \dots, \alpha_n \in \mathbb{F}$, and if $p \in \mathbb{F}[t]$, then

$$p(A)x = \sum_{i=1}^{n} \alpha_i p(\lambda_i) e_i.$$

If we can now construct a polynomial p in such a way that

$$p(\lambda_j) = 1, \qquad \max_{i\neq j} |p(\lambda_i)| \leq \varepsilon,$$

then we obtain using $\langle e_i, e_j \rangle = \delta_{ij}$ that

$$\begin{aligned} \|\alpha_j e_j - p(A)x\|_2^2 &= \left\| \alpha_j e_j - \sum_{i=1}^{n} \alpha_i p(\lambda_i) e_i \right\|_2^2 = \left\| \sum_{i\neq j} \alpha_i p(\lambda_i) e_i \right\|_2^2 \\ &= \left\langle \sum_{i\neq j} \alpha_i p(\lambda_i) e_i, \sum_{i\neq j} \alpha_i p(\lambda_i) e_i \right\rangle = \sum_{i\neq j} |\alpha_i|^2 |p(\lambda_i)|^2 \\ &\leq \varepsilon^2 \sum_{i\neq j} |\alpha_i|^2. \end{aligned}$$

Assuming without loss of generality that $\|x\|_2 = 1$ (we can always scale the start vector without changing the corresponding Krylov subspace), we obtain that

$$\sum_{i\neq j} |\alpha_i|^2 = 1 - |\alpha_j|^2 = 1 - |\langle x, e_j \rangle| = 1 - \cos^2 \angle(x, e_j) = \sin^2 \angle(x, e_j).$$

Thus, we have obtained the following result.

Lemma 7.26 (Convergence of Krylov subspace methods). *Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and let $e_1, \dots, e_n$ be a corresponding orthonormal basis of eigenvectors, i.e., $Ae_i = \lambda_i e_i$ for $i = 1, \dots, n$. Furthermore let $m \in \mathbb{N}$, $\varepsilon > 0$ and*

$$x = \sum_{i=1}^{n} \alpha_i e_i \in \mathbb{F}^n, \quad \sum_{i=1}^{n} |\alpha_i|^2 = 1.$$

If there exists a polynomial $p \in \mathbb{F}_m[t]$ *with* $p(\lambda_j) = 1$ *and* $\max_{i \neq j} |p(\lambda_i)| \leq \varepsilon$, *then*

$$\|\alpha_j e_j - p(A)x\|_2 \leq \varepsilon \sin \angle(x, e_j) \leq \varepsilon.$$

Lemma 7.26 states that if ε is small and $\alpha_j \neq 0$, then the vector $p(A)x$ is a good approximation to the vector $\alpha_j e_j$ and hence to an eigenvector associated with the eigenvalue λ_j. Revisiting Example 7.24, we see that the minimal eigenvalue λ_{102} is well separated from the remaining eigenvalues in the interval $[2, 3]$. If we choose a polynomial

$$p = c \prod_{i=1}^{10} (t - \mu_i) \in \mathbb{F}_{10}[t], \quad c \in \mathbb{R}$$

having all zeros in the interval $[2, 3]$, and if we choose c such that $p(\lambda_{102}) = 1$, we can expect due to the continuity of polynomials that

$$\max_{i \neq 102} |p(\lambda_i)| \leq \varepsilon,$$

where ε is small. Thus, $\mathcal{K}_{10}(A, x)$ can be expected to contain good approximations to eigenvectors associated with λ_{102}. In order to obtain quantitative statements, it is beneficial to construct p with the help of *Chebyshev polynomials*.

Definition 7.27 (Chebyshev polynomials). The polynomials $T_n \in \mathbb{F}[t]$, $n \in \mathbb{N}$ that are recursively defined by the formula

$$\begin{aligned} T_0 &:= 1, \\ T_1 &:= t, \\ T_{k+2} &:= 2t\, T_{k+1} - T_k \quad \text{for all } k \in \mathbb{N}_0, \end{aligned}$$

are called the *Chebyshev polynomials of the first kind*.

It is easy to verify that T_n is a polynomial of degree n. In the following, we will only refer to them as *Chebyshev polynomials* thus dropping the extension "*of the first kind*".

Exercise 7.28 (Chebyshev polynomials). Prove by induction that T_n is a polynomial of degree n with leading coefficient 2^{n-1} for $n \geq 1$. Also show that

$$T_n(-x) = (-1)^n T_n(x)$$

for all $x \in \mathbb{F}$ and all $n \in \mathbb{N}$.

Example 7.29 (Low degree Chebyshev polynomials). The first six Chebyshev polynomials have the form

$$\begin{aligned} T_0 &= 1, & T_3 &= 4t^3 - 3t, \\ T_1 &= t, & T_4 &= 8t^4 - 8t^2 + 1, \\ T_2 &= 2t^2 - 1, & T_5 &= 16t^5 - 20t^3 + 5t. \end{aligned}$$

The Chebyshev polynomials have other representations that are quite useful for theoretical considerations.

Theorem 7.30 (Alternative representations of the Chebyshev polynomials). *The Chebyshev polynomials T_n, $n \in \mathbb{N}_0$ from Definition 7.27 satisfy the following conditions.*

1. $T_n(t) = \cos(n \arccos t)$ *for all* $t \in [-1, 1]$.
2. $T_n(t) = \dfrac{z^n + z^{-n}}{2}$, *where* $t = \dfrac{z + z^{-1}}{2}$, $z \in \mathbb{C} \setminus \{0\}$.

Proof. 1. We show that the functions $f_n : [-1, 1] \to \mathbb{R}, \quad t \mapsto \cos(n \arccos t)$ satisfy the recursive formula of Definition 7.27. This implies $T_n(t) = f_n(t)$ for all $t \in [-1, 1]$. Clearly, for all $t \in [-1, 1]$ we have

$$\begin{aligned} f_0(t) &= \cos(0 \cdot \arccos t) = \cos(0) = 1, \quad \text{and} \\ f_1(t) &= \cos(\arccos t) = t. \end{aligned}$$

Using the addition theorem $\cos(x + y) = \cos x \cos y - \sin x \sin y$ we then obtain for $k \in \mathbb{N}_0$ that

$$\begin{aligned} f_{k+2}(t) &= \cos((k+2) \arccos t) = \cos((k+1) \arccos t + \arccos t) \\ &= \cos((k+1) \arccos t) \cos(\arccos t) - \sin((k+1) \arccos t) \sin(\arccos t) \end{aligned}$$

and

$$\begin{aligned} f_k(t) &= \cos(k \arccos t) = \cos((k+1) \arccos t + (-\arccos t)) \\ &= \cos((k+1) \arccos t) \cos(\arccos t) + \sin((k+1) \arccos t) \sin(\arccos t), \end{aligned}$$

where we also used $\cos(x) = \cos(-x)$ and $\sin(x) = -\sin(-x)$. Adding both equations yields

$$f_{k+2}(t) + f_k(t) = 2 \cos((k+1) \arccos t) \cos(\arccos t) = 2t f_{k+1}(t)$$

which implies that the functions f_n satisfy the recursive formula of Definition 7.27.

2. For a fixed $t \in \mathbb{C}$, the equation

$$t = \frac{z + z^{-1}}{2} \tag{7.23}$$

has two solutions $z_{1/2} = t \pm \sqrt{t^2 - 1}$ (for some branch of the complex square root) that satisfy $z_2 = z_1^{-1}$, because

$$z_1 \cdot z_2 = (t + \sqrt{t^2 - 1})(t - \sqrt{t^2 - 1}) = t^2 - (t^2 - 1) = 1.$$

Let us now consider the functions

$$g_n : \mathbb{C} \to \mathbb{C}, \quad t \mapsto \frac{z^n + z^{-n}}{2}, \quad \text{where } t = \frac{z + z^{-1}}{2},$$

for $n \in \mathbb{N}$. Then the functions g_n, $n \in \mathbb{N}$ are well defined, because the expression $\frac{z^n + z^{-n}}{2}$ is invariant under replacing z with z^{-1}, so it does not matter which solution of (7.23) is used for the computation of $g_k(t)$. Now we have

$$g_0(t) = \frac{z^0 + z^{-0}}{2} = 1, \qquad g_1(t) = \frac{z^1 + z^{-1}}{2} = t, \quad \text{and}$$

$$\begin{aligned} 2\,t \cdot g_n(t) - g_{n-1}(t) &= 2 \cdot \frac{z + z^{-1}}{2} \cdot \frac{z^n + z^{-n}}{2} - \frac{z^{n-1} + z^{-(n-1)}}{2} \\ &= \frac{z^{n+1} + z^{-n+1} + z^{n-1} + z^{-n-1} - z^{n-1} - z^{-(n-1)}}{2} \\ &= \frac{z^{n+1} + z^{-(n+1)}}{2} = g_{n+1}(t), \end{aligned}$$

i.e., the functions g_n, $n \in \mathbb{N}$ satisfy the recursive formula of Definition 7.27 which implies $T_n(t) = g_n(t)$ for all $t \in \mathbb{C}$. □

Corollary 7.31 (Zeros and extrema of Chebyshev polynomials). *Let $n \in \mathbb{N}$. Then the following conditions are satisfied:*

1. *If $n \geq 1$, then T_n has n zeros, all being in $[-1, 1]$ and given by*

$$\cos\left(\frac{2k - 1}{2n}\pi\right), \quad k = 1, \ldots, n.$$

2. $\max_{t \in [-1,1]} |T_n(t)| = 1.$

3. *$T_n : [-1, 1] \to [-1, 1]$ has $n + 1$ local extrema given by*

$$s_k^{(n)} = \cos\left(\frac{k}{n}\pi\right), \quad T_n(s_k^{(n)}) = (-1)^k, \quad k = 0, \ldots, n.$$

Of those, $s_k^{(n)}$, $k = 1, \ldots, n - 1$ are also local extrema for $T_n : \mathbb{R} \to \mathbb{R}$, while $s_0^{(n)} = 1$ and $s_n^{(n)} = -1$ are not.

Proof. The proof follows immediately from Theorem 7.30. □

Exercise 7.32 (Properties of Chebyshev polynomials). Show that the Chebyshev polynomials satisfy the following conditions.

1. $T_n(T_m(t)) = T_{nm}(t)$ for all $t \in \mathbb{F}$ and all $n, m \in \mathbb{N}$.
2. The Chebyshev polynomials are orthogonal on the interval $(-1, 1)$ with respect to the weight $1/\sqrt{1-t^2}$, i.e.,
$$\int_{-1}^{1} T_n(t) T_m(t) \frac{1}{\sqrt{1-t^2}} dt = 0$$
for all $n, m \in \mathbb{N}$ with $m \neq n$.
3. For all $n \geq 1$, the Chebyshev polynomial $T_n : \mathbb{R} \to \mathbb{R}$ satisfies the differential equation
$$(1-t^2)T_n''(t) - t\, T_n'(t) + n^2 T_n(t) = 0.$$

What makes Chebyshev polynomials so useful is the fact that they satisfy an optimality condition. Recall that a polynomial is called *monic* if the leading coefficient is equal to one. Note that the Chebyshev polynomials are not monic for $n > 1$, but we can easily make them monic by normalizing them as $\widetilde{T}_n = 2^{1-n} T_n$, because by Exercise 7.28 the leading coefficient of T_n is 2^{n-1}.

Theorem 7.33 (Optimality of Chebyshev polynomials). *Let $n \geq 1$, $\gamma > 1$, $c \in \mathbb{R}$.*

1. *Among all monic polynomials $p \in \mathbb{R}_n[t]$ the polynomial $\widetilde{T}_n := 2^{1-n} T_n \in \mathbb{R}[t]$ has the smallest maximum norm on the interval $[-1, 1]$ which is given by 2^{1-n}, i.e.,*
$$2^{1-n} = \max_{t \in [-1,1]} |\widetilde{T}_n(t)| = \min_{\substack{p \in \mathbb{R}_n[t] \\ p \text{ monic}}} \max_{t \in [-1,1]} |p(t)|. \tag{7.24}$$
2. *Among all $p \in \mathbb{R}_n[t]$ satisfying $p(\gamma) = c$ the polynomial $\widehat{T}_n := \frac{c}{T_n(\gamma)} T_n \in \mathbb{R}[t]$ has the smallest maximum norm on the interval $[-1, 1]$ which is given by $|c|/T_n(\gamma)$, i.e.,*
$$\frac{|c|}{T_n(\gamma)} = \max_{t \in [-1,1]} |\widehat{T}_n(t)| = \min_{\substack{p \in \mathbb{R}_n[t] \\ p(\gamma) = c}} \max_{t \in [-1,1]} |p(t)|. \tag{7.25}$$

Proof. 1. By Corollary 7.31, we immediately obtain that
$$\max_{t \in [-1,1]} |\widetilde{T}_n(t)| = 2^{1-n}.$$

For the second equality in (7.24), the inequality "$\geq$" is trivial, so it remains to show the inequality "$\leq$". To this end, assume that $p \in \mathbb{F}_n[t]$ is a monic polynomial such that

$$|p(t)| < 2^{1-n} \quad \text{for all } t \in [-1, 1].$$

By Corollary 7.31 the polynomial $\widetilde{T}_n$ has $n+1$ local extrema in $[-1, 1]$ given by

$$\widetilde{T}_n(s_k^{(n)}) = 2^{1-n} T_n(s_k^{(n)}) = 2^{1-n}(-1)^k, \quad s_k^{(n)} = \cos\left(\frac{k}{n}\pi\right), \quad k = 0, \ldots, n.$$

Thus, we obtain that

$$p(s_k^{(n)}) - \widetilde{T}_n(s_k^{(n)}) \begin{cases} < 0 & \text{if } k \text{ is even,} \\ > 0 & \text{if } k \text{ is odd.} \end{cases}$$

for $k = 0, \ldots, n$, i.e., the function $p(t) - \widetilde{T}_n(t)$ changes sign n times in the interval $[-1, 1]$ and thus has n roots in $[-1, 1]$ by the intermediate value theorem. On the other hand, as p and $\widetilde{T}_n$ are both monic, the difference $p - \widetilde{T}_n$ is a polynomial of degree at most $n-1$ and hence it must be the zero polynomial because of having at least n pairwise distinct roots. This implies $p = \widetilde{T}_n$ in contradiction to the assumption.

2. The proof of 2. is quite similar to the proof of 1. and is left as an exercise. □

Exercise 7.34. Prove part 2. of Theorem 7.33.

7.7 Convergence of Krylov subspace methods

Let $A \in \mathbb{F}^{n\times n}$ be a self-adjoint matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. In the following, we will try to find approximations to eigenvectors associated with the extremal (largest and smallest) eigenvalues of A. We will restrict our attention to the largest eigenvalue λ_1, because we can always consider $-A$ instead of A so that the smallest eigenvalue λ_n of A becomes the largest eigenvalue of $-A$. By Lemma 7.26, we need to construct a polynomial of degree at most m that has small values in modulus in the interval $[\lambda_n, \lambda_2]$ in order to obtain a good approximation to an eigenvector associated with λ_1. Since the Chebyshev polynomials satisfy an optimality condition on the interval $[-1, 1]$ in the sense of Theorem 7.33, it is reasonable to transform the polynomials in such a way that this optimality is transferred to the interval $[\lambda_n, \lambda_2]$. Assuming $\lambda_2 > \lambda_n$, this can be done by help of the transformation

$$t \mapsto \frac{2t - (\lambda_2 + \lambda_n)}{\lambda_2 - \lambda_n}$$

which maps λ_2 to 1 and λ_n to -1 while λ_1 is mapped to

$$\gamma := \frac{2\lambda_1 - (\lambda_2 + \lambda_n)}{\lambda_2 - \lambda_n} = \frac{2(\lambda_1 - \lambda_2) + (\lambda_2 - \lambda_n)}{\lambda_2 - \lambda_n} = 2\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n} + 1. \tag{7.26}$$

Note that $\gamma > 1$ if $\lambda_1 > \lambda_2$. Using this transformation, we obtain the following optimality condition.

Lemma 7.35 (Optimality of transformed Chebyshev polynomials). *Let $\lambda_1 > \lambda_2 > \lambda_n$. Among all $p \in \mathbb{R}_m[t]$ satisfying $p(\lambda_1) = 1$, the polynomial $p_m \in \mathbb{R}_m[t]$ with*

$$p_m(t) := \frac{1}{T_m(\gamma)} T_m\left(\frac{2t - (\lambda_2 + \lambda_n)}{\lambda_2 - \lambda_n}\right), \quad \gamma := 2\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n} + 1$$

has the smallest maximum norm on the interval $[\lambda_n, \lambda_2]$. In particular, setting

$$\kappa := \frac{\lambda_1 - \lambda_n}{\lambda_1 - \lambda_2} \quad \text{and} \quad \varrho := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1},$$

we have

$$\min_{\substack{p \in \mathbb{R}_n[t] \\ p(\lambda_1)=1}} \max_{t \in [\lambda_n, \lambda_2]} |p(t)| = \max_{t \in [\lambda_n, \lambda_2]} |p_m(t)| = \frac{1}{T_m(\gamma)} = \frac{2\varrho^m}{1 + \varrho^{2m}}. \tag{7.27}$$

Proof. By applying part 2 of Theorem 7.24, we immediately obtain the first and the second identity in (7.27). For the third identity, we have to calculate the value of $1/T_m(\gamma)$. The trick is to use the second alternative representation of T_n from Theorem 7.30. Let us pick the smaller solution $z = \gamma - \sqrt{\gamma^2 - 1}$ of the two solutions of the equation

$$\gamma = \frac{z + z^{-1}}{2}.$$

(Note that both solutions are real, because of $\gamma > 1$.) Then we have

$$\begin{aligned}
\gamma^2 - 1 &= \left(2\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n} + 1\right)^2 - 1 = 4\frac{(\lambda_1 - \lambda_2)^2}{(\lambda_2 - \lambda_n)^2} + 4\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n} \\
&= 4\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_2)(\lambda_2 - \lambda_n)}{(\lambda_2 - \lambda_n)^2} = 4\frac{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_n)}{(\lambda_2 - \lambda_n)^2}.
\end{aligned}$$

This implies

$$\begin{aligned}
\gamma - \sqrt{\gamma^2 - 1} &= \frac{2(\lambda_1 - \lambda_2) + (\lambda_2 - \lambda_n) - 2\sqrt{\lambda_1 - \lambda_2}\sqrt{\lambda_1 - \lambda_n}}{\lambda_2 - \lambda_n} \\
&= \frac{(\lambda_1 - \lambda_2) - 2\sqrt{\lambda_1 - \lambda_2}\sqrt{\lambda_1 - \lambda_n} + (\lambda_1 - \lambda_n)}{\lambda_2 - \lambda_n} \\
&= \frac{(\sqrt{\lambda_1 - \lambda_2} - \sqrt{\lambda_1 - \lambda_n})^2}{(\sqrt{\lambda_2 - \lambda_n})^2}
\end{aligned}$$

$$= \left(\sqrt{\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}} - \sqrt{\frac{\lambda_1 - \lambda_n}{\lambda_2 - \lambda_n}} \cdot \sqrt{\frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2}} \right)^2$$
$$= \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n} \left(1 - \sqrt{\frac{\lambda_1 - \lambda_n}{\lambda_1 - \lambda_2}} \right)^2 .$$

Using

$$\frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n} = \frac{\lambda_1 - \lambda_2}{(\lambda_1 - \lambda_n) - (\lambda_1 - \lambda_2)} = \frac{1}{\kappa - 1},$$

we finally obtain

$$\gamma - \sqrt{\gamma^2 - 1} = \frac{(1 - \sqrt{\kappa})^2}{\kappa - 1} = \frac{(\sqrt{\kappa} - 1)^2}{(\sqrt{\kappa} + 1)(\sqrt{\kappa} - 1)} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = \varrho.$$

Then part 2. of Theorem 7.30 implies

$$\frac{1}{T_m(\gamma)} = \frac{2}{\varrho^m + \varrho^{-m}} = \frac{2\varrho^m}{\varrho^{2m} + 1}$$

which finishes the proof of the third identity in (7.27). □

Theorem 7.36 (Convergence rate for Krylov subspace methods I). *Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$, where $\lambda_2 > \lambda_n$ and let $e_1, \ldots, e_n$ be a corresponding orthonormal basis of eigenvectors such that $Ae_i = \lambda_i e_i$ for $i = 1, \ldots, n$. Furthermore let $m \in \mathbb{N}$ and*

$$x = \sum_{i=1}^{n} \alpha_i e_i \in \mathbb{F}^n, \qquad \sum_{i=1}^{n} |\alpha_i| = 1.$$

Then with the notation of Lemma 7.35 there exists a polynomial $p_m \in \mathbb{F}_m[t]$ with

$$\|\alpha_1 e_1 - p_m(A)x\|_2 \leq \frac{2\varrho_m}{1 + \varrho^{2m}} \sin \angle(x, e_1).$$

Proof. The proof follows immediately by combining Lemma 7.26 and Lemma 7.35. □

Note that Theorem 7.36 simply contains an existence statement: we know about the existence of the vector $p_m(A)x$ in the Krylov subspace $\mathcal{K}_m(A, x)$, but we do not know how to construct this vector. However, concerning eigenvalues, we can now give a relation between the largest eigenvalue λ_1 and the corresponding largest Ritz value with respect to $\mathcal{K}_m(A, x)$.

Theorem 7.37 (Convergence rate for Krylov subspace methods II). *Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$, where $\lambda_2 > \lambda_n$ and let $e_1, \ldots, e_n$ be a corresponding orthonormal basis of eigenvectors such that $Ae_i = \lambda_i e_i$ for $i = 1, \ldots, n$. Furthermore let $m \in \mathbb{N}$ and*

$$x = \sum_{i=1}^{n} \alpha_i e_i \in \mathbb{F}^n, \qquad \sum_{i=1}^{n} |\alpha_i| = 1,$$

where $\alpha_1 \neq 0$. If μ_1 is the largest Ritz value of A with respect to $\mathcal{K}_m(A, x)$, then with the notation of Lemma 7.35 we have

$$\lambda_1 \geq \mu_1 \geq \lambda_1 - (\lambda_1 - \lambda_n)\left(\frac{2\varrho^{m-1}}{1+\varrho^{2(m-1)}}\right)^2 \tan^2 \angle(x, e_1).$$

Proof. Without loss of generality, we may assume that $\dim \mathcal{K}_m(A, x) = m$. Otherwise, the result is trivial, because $\mathcal{K}_m(A, x)$ is an invariant subspace which must contain e_1 because of $\alpha_1 \neq 0$, see Exercise 7.38. Thus, let the columns of $Q_m \in \mathbb{F}^{m\times n}$ form an orthonormal basis of $\mathcal{K}_m(A, x)$. Then the largest Ritz value μ_1 of A with respect to $\mathcal{K}_m(A, x)$ is the largest eigenvalue of the matrix $Q_m^* A Q_m$ and thus, by Exercise 3.9 we have

$$\mu_1 = \max_{\substack{y\in\mathbb{F}\\ \|y\|=1}} \langle Q_m^* A Q_m y, y\rangle = \max_{y\in\mathbb{F}\setminus\{0\}} \frac{y^* Q_m^* A Q_m y}{y^* y}.$$

Recalling that

$$\mathcal{R}(Q_m) = \mathcal{K}_m(A, x) = \{p(A)x \mid p \in \mathbb{F}_{m-1}[t]\}$$

and letting p_{m-1} denote the transformed Chebyshev polynomial from Theorem 7.36 which satisfies $p_{m-1}(\lambda_1) = 1$, we obtain

$$\begin{aligned}
\mu_1 &= \max_{\substack{p\in\mathbb{F}_{m-1}[t]\\ p(A)x\neq 0}} \frac{x^* p(A)^* A p(A) x}{x^* p(A)^* p(A) x} \geq \frac{x^* p_{m-1}(A)^* A p_{m-1}(A) x}{x^* p_{m-1}(A)^* p_{m-1}(A) x} \\
&= \frac{\sum_{i=1}^n \lambda_i p_{m-1}(\lambda_i)^2 |\alpha_i|^2}{\sum_{i=1}^n p_{m-1}(\lambda_i)^2 |\alpha_i|^2} = \frac{\sum_{i=1}^n (\lambda_1 + \lambda_i - \lambda_1) p_{m-1}(\lambda_i)^2 |\alpha_i|^2}{\sum_{i=1}^n p_{m-1}(\lambda_i)^2 |\alpha_i|^2} \\
&= \lambda_1 - \frac{\sum_{i=1}^n (\lambda_1 - \lambda_i) p_{m-1}(\lambda_i)^2 |\alpha_i|^2}{\sum_{i=1}^n p_{m-1}(\lambda_i)^2 |\alpha_i|^2} \\
&\geq \lambda_1 - (\lambda_1 - \lambda_n) \frac{\sum_{i=2}^n p_{m-1}(\lambda_i)^2 |\alpha_i|^2}{|\alpha_1|^2 + \sum_{i=2}^n p_{m-1}(\lambda_i)^2 |\alpha_i|^2} \\
&\geq \lambda_1 - (\lambda_1 - \lambda_n) \frac{\max\{p_{m-1}(t)^2 \mid t \in [\lambda_n, \lambda_2]\} \sum_{i=2}^n |\alpha_i|^2}{|\alpha_1|^2} \\
&\geq \lambda_1 - (\lambda_1 - \lambda_n) \left(\frac{2\varrho^{m-1}}{1+\varrho^{2(m-1)}}\right)^2 \frac{1 - |\alpha_1|^2}{|\alpha_1|^2},
\end{aligned}$$

where we have used $\sum_{i=2}^{n} |\alpha_i|^2 = 1 - |\alpha_1|^2$. The result then follows from

$$\tan^2 \angle(x, e_1) = \frac{\sin^2 \angle(x, e_1)}{\cos^2 \angle(x, e_1)} = \frac{1 - \cos^2 \angle(x, e_1)}{\cos^2 \angle(x, e_1)}$$

and $\cos \angle(x, e_1) = |\langle e_1, x \rangle| = |\alpha_1|$. □

Exercise 7.38 (Invariant Krylov subspace). Let $A \in \mathbb{F}^{n \times n}$ be self-adjoint with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ and let $e_1, \ldots, e_n$ be a corresponding basis of eigenvectors. Furthermore let $m \in \mathbb{N}$ and

$$x = \sum_{i=1}^{n} \alpha_i e_i \in \mathbb{F}^n \setminus \{0\},$$

where $\alpha_1 \neq 0$. Show that if the Krylov subspace $\mathcal{K}_m(A, x)$ is invariant with respect to A then $e_1 \in \mathcal{K}_m(A, x)$.

Example 7.39 (Convergence of the largest Ritz value I). We illustrate the result of Theorem 7.37 for the diagonal 100×100 matrix A with diagonal entries $a_{kk} = 1 + k/100$ for $k = 1, \ldots, 99$ and $a_{100,100} = 2.2$, i.e.,

$$A = \operatorname{diag}(1.01,\ 1.02,\ 1.03,\ \ldots,\ 1.98,\ 1.99,\ 2.20) \in \mathbb{R}^{100 \times 100}.$$

Thus, there is a sufficiently large gap between the largest eigenvalue $\lambda_1 = 2.2$ and the remaining eigenvalues $\lambda_2, \ldots, \lambda_{100} \in [1, 1.99]$. Choosing the start vector

$$x = \frac{1}{\sqrt{100}} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{100},$$

we obtain the following error $\lambda_1 - \mu_1^{(m)}$ for the largest Ritz value $\mu_1^{(m)}$ with respect to $\mathcal{K}_m(A, x)$ as plotted in Figure 7.9.

As one can see, the convergence rate corresponds quite well to the convergence rate predicted by the bound from Theorem 7.37. In fact the error is a little bit smaller than predicted by the bound, but this effect is due to the simplifications we performed in the proof of Theorem 7.37. Numerical experiments show that the bound becomes much tighter if the gap between λ_1 and λ_2 is very large.

Example 7.40 (Convergence of the largest Ritz value II). For a second test, let B be the diagonal matrix with diagonal entries $b_{kk} = 1 + k/100$ for $k = 1, \ldots, 100$, i.e.,

$$B = \operatorname{diag}(1.01,\ 1.02,\ 1.03,\ \ldots,\ 1.98,\ 1.99,\ 2.00) \in \mathbb{R}^{100 \times 100}.$$

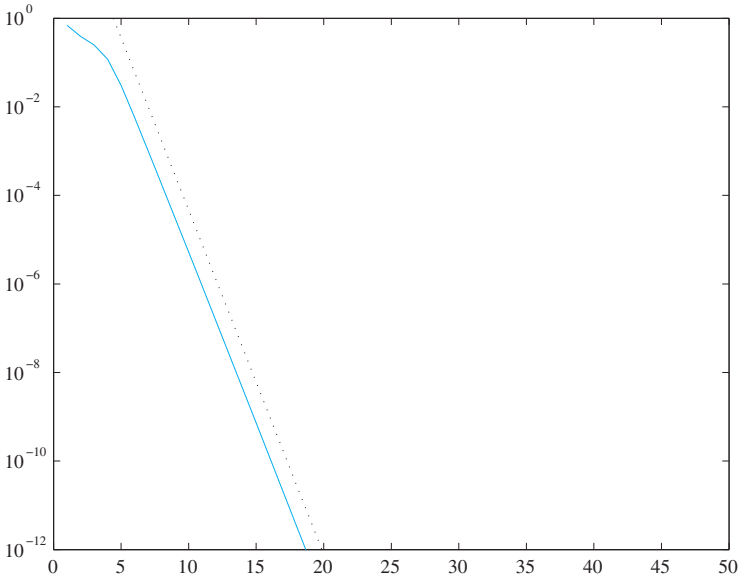

Figure 7.9. The error $\lambda_1 - \mu_1^{(m)}$ of the largest Ritz value $\mu_1^{(m)}$ of the matrix A in Example 7.39 with respect to $\mathcal{K}_m(A, x)$ for $m = 1, \ldots, 50$. The error is plotted as a solid line, the dotted line is the bound from Theorem 7.37.

In contrast to Example 7.39, the gap between the largest and second largest eigenvalues λ_1 and λ_2 is now much smaller. Choosing the same start vector x as in Example 7.39, we obtain the following error $\lambda_1 - \mu_1^{(m)}$ for the largest Ritz value $\mu_1^{(m)}$ with respect to $\mathcal{K}_m(A, x)$ as plotted in Figure 7.10.

In this case, the bound is quite an overestimate of the actual error. Moreover, the convergence seems to be superlinear at it seems that from step 25 onwards, the actual convergence rate becomes better than the linear convergence rate predicted by the bound from Theorem 7.37.

How can the observation in Example 7.40 be explained? Why does the convergence rate seem to accelerate for increasing values of m? The reason is that for computing the bound in Theorem 7.37, we used the transformed Chebyshev polynomial p_{m_1} which has the minimal maximum norm on the interval $[\lambda_n, \lambda_2]$ under the constraint $p_{m_1}(\lambda_1) = 1$. However, for our particular purpose, we do not need a polynomial that has minimal maximum norm on a whole interval, but it is sufficient if the polynomial is small on the eigenvalues $\lambda_n, \ldots, \lambda_2$. For example, we could use the polynomial

$$\widehat{p}_{m-2}(t) = \frac{t - \lambda_2}{\lambda_1 - \lambda_2} p_{m-2}(t),$$

where p_{m-2} is the transformed Chebyshev polynomial of degree $m - 2$ from Theorem 7.36 with $p_{m-2}(\lambda_1) = 1$ and minimal maximum norm on the interval $[\lambda_n, \lambda_3]$.

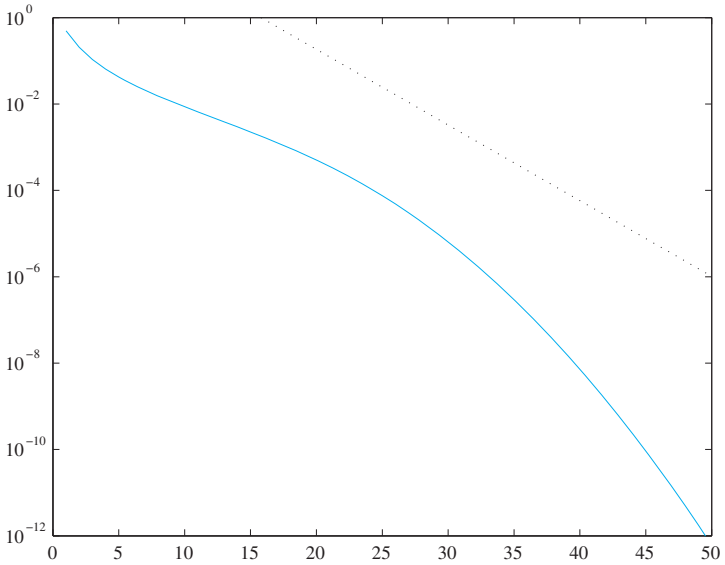

Figure 7.10. The error $\lambda_1 - \mu_1^{(m)}$ of the largest Ritz value $\mu_1^{(m)}$ of the matrix B in Example 7.40 with respect to $\mathcal{K}_m(A, x)$ for $m = 1, \ldots, 50$. The error is plotted as a solid line, the dotted line is the bound from Theorem 7.37.

Then $\widehat{p}_{m-2}$ is small on all eigenvalues $\lambda_n, \ldots, \lambda_2$, because λ_2 has been chosen to be a zero of $\widehat{p}_{m-2}$. We can use this polynomial to produce a different bound.

Theorem 7.41 (Convergence rate for Krylov subspace methods III). *Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint with eigenvalues $\lambda_1 \geq \lambda_2 > \lambda_3 \geq \cdots \geq \lambda_n$, where $\lambda_3 > \lambda_n$ and let $e_1, \ldots, e_n$ be a corresponding orthonormal basis of eigenvectors such that $Ae_i = \lambda_i e_i$ for $i = 1, \ldots, n$. Furthermore let $m \in \mathbb{N}$ and*

$$x = \sum_{i=1}^{n} \alpha_i e_i \in \mathbb{F}^n, \quad \sum_{i=1}^{n} |\alpha_i| = 1,$$

where $\alpha_1 \neq 0$. If

$$\widehat{\kappa} := \frac{\lambda_1 - \lambda_n}{\lambda_1 - \lambda_3} \quad \text{and} \quad \widehat{\varrho} := \frac{\sqrt{\widehat{\kappa}} - 1}{\sqrt{\widehat{\kappa}} + 1}$$

then

$$\lambda_1 \geq \mu_1 \geq \lambda_1 - (\lambda_1 - \lambda_n) \left(\frac{\lambda_n - \lambda_2}{\lambda_1 - \lambda_2} \right)^2 \left(\frac{2\widehat{\varrho}^{m-2}}{1 + \widehat{\varrho}^{2(m-2)}} \right)^2 \frac{\sum_{i=3}^{n} |\alpha_i|^2}{|\alpha_1|^2}.$$

where μ_1 is the largest Ritz value of A with respect to $\mathcal{K}_m(A, x)$.

Exercise 7.42 (Convergence rate for Krylov subspace methods).
Prove Theorem 7.41.
Hint: mimic the proof of Theorem 7.37 using

$$\widehat{p}_{m-2}(t) = \frac{t-\lambda_2}{\lambda_1-\lambda_2}\, p_{m-2}(t)$$

instead of p_{m-1}.

Note that if $\lambda_2 > \lambda_3$, then $\widehat{\varrho} < \varrho$. Thus, asymptotically the bound from Theorem 7.41 will be smaller than the one from Theorem 7.37. However, this may only be relevant for sufficiently large m as the new bound has a different constant now and the exponent has reduced from $m-1$ to $m-2$.

Example 7.43 (Convergence of the largest Ritz value III). Let $C \in \mathbb{R}^{100\times 100}$ be the diagonal matrix with diagonal entries

$$c_{kk} = 1 + k/100, \quad k = 1, \ldots, 98,$$

and $c_{99,99} = 2.19$, $c_{100,100} = 2.2$, i.e.,

$$C = \operatorname{diag}(1.01,\ 1.02,\ 1.03,\ \ldots,\ 1.97,\ 1.98,\ 2.19,\ 2.20) \in \mathbb{R}^{100\times 100}.$$

Thus, now have a nice gap between λ_2 and λ_3, whereas the gap between λ_1 and λ_2 is rather small. Choosing the start vector as in Example 7.39, we obtain the following error $\lambda_1 - \mu_1^{(m)}$ for the largest Ritz value $\mu_1^{(m)}$ with respect to $\mathcal{K}_m(A,x)$ as plotted in Figure 7.11.

This time, the convergence rate corresponds quite well to the convergence rate predicted by the bound from Theorem 7.41 while the bound from Theorem 7.36 is a huge overestimate due to the fact that the gap between λ_1 and λ_2 is rather small.

The observation from Example 7.43 can be generalized to explain also the effect observed in Example 7.40. If m is sufficiently large, then $\mathcal{K}_m(A,x)$ contains polynomials $\widehat{p}_{m-k}$ that have zeros in $\lambda_2, \ldots, \lambda_k$, that satisfy $\widehat{p}_{m-k}(\lambda_1) = 1$, and that have small norm in the interval $[\lambda_n, \lambda_{k+1}]$. For such polynomials, the relevant gap in eigenvalues is the gap between λ_{k+1} and λ_1 rather than λ_2 and λ_1. Thus, the convergence rate accelerates as m increases.

Exercise 7.44 (Convergence of the largest Ritz value). Under the hypothesis $\lambda_k > \lambda_{k+1} > \lambda_n$ construct polynomials $\widehat{p}_{m-k}$ as outlined above and obtain a result analogous to the one of Theorem 7.41 by improving the bound $\widehat{\varrho}$ to

$$\widetilde{\varrho} := \frac{\sqrt{\widetilde{\kappa}}-1}{\sqrt{\widetilde{\kappa}}+1} \quad \text{with} \quad \widetilde{\kappa} := \frac{\lambda_1-\lambda_n}{\lambda_1-\lambda_k}.$$

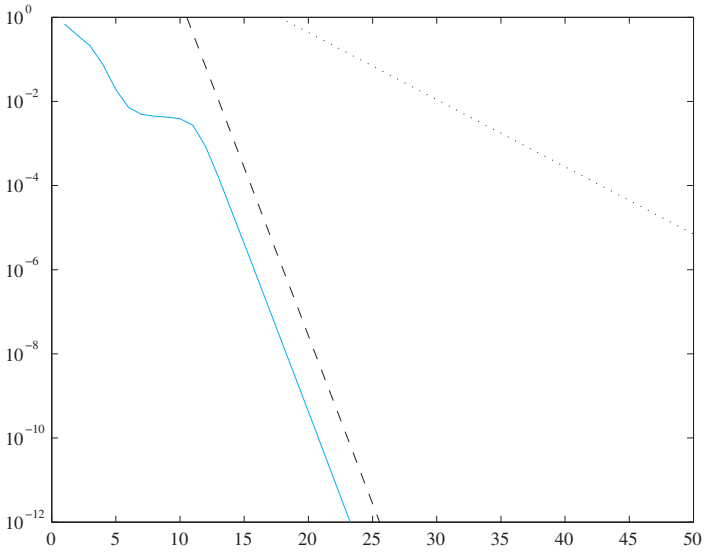

Figure 7.11. The error $\lambda_1 - \mu_1^{(m)}$ of the largest Ritz value $\mu_1^{(m)}$ of the matrix C in Example 7.43 with respect to $\mathcal{K}_m(A, x)$ for $m = 1, \ldots, 50$. The error is plotted as a solid line, the dotted line is the bound from Theorem 7.37, and the dashed line is the bound from Theorem 7.41.

Exercise 7.45 (Convergence of the second largest Ritz value). Let $A \in \mathbb{F}^{n\times n}$ be self-adjoint with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, where $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_n$. Derive a bound similar to the one in Theorem 7.37 that gives an estimate for the error $\lambda_2 - \mu_2$, where μ_2 is the second largest Ritz value of A with respect to $\mathcal{K}_m(A, x)$.

Hint: Construct a polynomial $\widetilde{p}_{m-2}$ with $\widetilde{p}_{m-2}(\lambda_1) = 0$ and $\widetilde{p}_{m-2}(\lambda_2) = 1$ that has small norm on the interval $[\lambda_n, \lambda_3]$.

By applying the results of this section to $-A$, we obtain corresponding bounds for the smallest Ritz values of the self-adjoint matrix A. It can be observed that among the Ritz values computed by Krylov subspace methods one typically finds first good approximations to the largest or smallest eigenvalues of A. Heuristically, this can be explained by the observation that for an eigenvalue $\widehat{\lambda}$ in the interior of the spectrum that is not well separated from the other eigenvalues, it is quite hard to find a polynomial $\widehat{p}$ of low degree such that $\widehat{p}(\widehat{\lambda}) = 1$ which takes values small in modulus at all remaining eigenvalues. This observation also generalizes to the case when A is not self-adjoint. Again, typically the eigenvalues near the boundary of the spectrum are located first.

Another important point in Krylov subspace method are restarts. At first, the start vector x in the Arnoldi or Lanczos iteration may be chosen randomly, but after a few iterations, we may actually find a vector $\widehat{x}$ in the computed Krylov subspace $\mathcal{K}_m(A, x)$ that would have been a better start vector. We could then restart our iteration with the new start vector $\widehat{x}$. In the Arnoldi iteration, this may be a huge advantage, because the summation in (7.18) becomes more and more expensive by each step, because in each step an additional summand occurs in the formula. A commonly used and rather effective methods is the so-called *implicitly restarted Arnoldi iteration*. For further reading we refer to [2, 49].

There are other algorithms for computing eigenvalues and eigenvectors of large sparse matrices. Two among those are the *Jacobi Davidson method* and *Preconditioned Inverse iteration* (PINVIT).

The Jacobi Davidson method introduced in [40] is a projection method, but uses other subspaces instead of Krylov subspaces. The basic idea is to enlarge the current subspace $\mathcal{R}(Q_m)$ in such a way that a particular eigenpair (λ, v) of A that is approximated by a Ritz pair $(\mu, Q_m y)$ will be approximated much better by a Ritz pair with respect to the enlarged subspace $\mathcal{R}(Q_{m+1})$. The resulting method is particularly suited for the computation of eigenvalues in the interior of the spectrum and leads to quadratic convergence of the selected Ritz pairs. However, typically only one Ritz pair converges at a time and once it has converged one focuses on the next Ritz pair. This is different to the case of Krylov subspaces where convergence of several Ritz pairs takes place simultaneously. For details, we refer the reader to [40] and [41].

The principal idea of PINVIT can be described as follows. If we consider inverse iteration and scale the next iteration vector by the Rayleigh quotient, the resulting iteration equation can be reformulated as

$$\begin{aligned} x_{i+1} &= \Lambda_A(x_i)A^{-1}x_i = x_i - A^{-1}Ax_i + \Lambda_A(x_i)A^{-1}x_i \\ &= x_i - A^{-1}(Ax_i - \Lambda_A(x_i)x_i). \end{aligned}$$

Then A^{-1} is replaced by an approximate inverse B^{-1} of A that can be easily calculated, a so-called *preconditioner*. For details, we refer the reader to [29, 30, 23].

Chapter 8

Generalized and polynomial eigenvalue problems *

Summary

Generalized eigenvalue problems are the natural extension of the standard eigenvalue problem. Their discussion will lead to the theory of matrix pencils. Concerning the solution of generalized eigenvalue problems, a generalization of the efficient QR iteration introduced in Chapter 5, the QZ algorithm, turns out to be an efficient method. It is based on a generalization of the Schur decomposition introduced in Chapter 2. An even further generalization of the standard eigenvalue problem are polynomial eigenvalue problems that can be reduced to generalized eigenvalue problems by the concept of linearization.

Learning targets

- ✓ Introduce matrix polynomials and matrix pencils.
- ✓ Introduce the eigenvalue ∞.
- ✓ Generalize the Schur decomposition to matrix pencils.
- ✓ Generalize the QR algorithm to the corresponding QZ algorithm for matrix pencils.

8.1 Polynomial eigenvalue problems and linearization

In the introduction, we have seen that systems of linear differential equations of higher order of the form

$$\sum_{k=0}^{\ell} A_k y^{(k)} = A_\ell y^{(\ell)} + A_{\ell-1} y^{(\ell-1)} + \cdots + A_2 y'' + A_1 y' + A_0 y = 0, \quad (8.1)$$

where $A_0, \ldots, A_\ell \in \mathbb{F}^{n\times n}$ lead to a polynomial eigenvalue problem of the form

$$\left(\sum_{k=0}^{\ell} \lambda^k A_k \right) y_0 = 0.$$

Definition 8.1 (Matrix polynomials). Let $\ell \in \mathbb{N} \setminus \{0\}$. The polynomial

$$P(t) = \sum_{k=0}^{\ell} t^k A_k \tag{8.2}$$

with matrix coefficients $A_0, \ldots, A_\ell \in \mathbb{F}^{n\times n}$, $A_\ell \neq 0$ is called an $n \times n$ *matrix polynomial* over $\mathbb{F}$ of degree ℓ.

Definition 8.2 (Polynomial and quadratic eigenvalue problems). Let P be an $n \times n$ matrix polynomial over $\mathbb{F}$ of degree ℓ. Then the problem of finding scalars $\lambda \in \mathbb{F}$ and nonzero vectors $x \in \mathbb{F}^n$ such that

$$P(\lambda)x = 0 \tag{8.3}$$

is called a *polynomial eigenvalue problem*. In this case, λ is called an eigenvalue of P and x is called an *eigenvector* of P for the eigenvalue λ. If $\ell = 2$, then the problem (8.3) is called a *quadratic eigenvalue problem*.

How can we solve the polynomial eigenvalue problem (8.3)? For a fixed $\lambda \in \mathbb{F}$, the equation $P(\lambda)x = 0$ has a nontrivial solution if the matrix $P(\lambda) \in \mathbb{F}^{n\times n}$ is singular, i.e., if $\det P(\lambda) = 0$.

Remark 8.3 (Determinant of matrix polynomials). Let P be an $n \times n$ matrix polynomial over $\mathbb{F}$. Then $\lambda \in \mathbb{F}$ is an eigenvalue of P if and only if $\det P(\lambda) = 0$.

A common method to tackle the polynomial eigenvalue problem (8.3) is the concept of *linearization*, i.e., of transforming the problem into an equivalent linear eigenvalue problem. A natural way to achieve this is based on the concept of order reduction in the theory of differential equations. Let us return to the system of differential equations (8.1). Introducing new variables $y_1 = y$, $y_2 = y'$, $\ldots$, $y_\ell = y^{(\ell-1)}$, we obtain the new system of linear differential equations

$$y_1' = y_2, \ \ldots, \ y_{\ell-1}' = y_\ell, \quad A_\ell y_\ell' + \sum_{k=0}^{\ell-1} A_k y_{k+1} = 0,$$

or, equivalently,

$$\begin{pmatrix} I_n & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & I_n & 0 \\ 0 & \ldots & 0 & A_\ell \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{pmatrix}' + \begin{pmatrix} 0 & -I_n & & 0 \\ \vdots & \ddots & \ddots & \\ 0 & \ldots & 0 & -I_n \\ A_0 & A_1 & \ldots & A_{\ell-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{pmatrix} = 0.$$

Similarly, introducing the new variables $x_1 = x$, $x_2 = \lambda x$, $\ldots$, $x_\ell = \lambda^{\ell-1}x$, we obtain from the polynomial eigenvalue problem

$$\left(\sum_{k=0}^{\ell} \lambda^k A_k\right) x = 0 \tag{8.4}$$

the new eigenvalue problem

$$\lambda \begin{pmatrix} I_n & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & I_n & 0 \\ 0 & \ldots & 0 & A_\ell \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{pmatrix} = \begin{pmatrix} 0 & I_n & & 0 \\ \vdots & \ddots & \ddots & \\ 0 & \ldots & 0 & I_n \\ -A_0 & -A_1 & \ldots & -A_{\ell-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{pmatrix}. \tag{8.5}$$

Note that the matrix on the right hand side of (8.5) has a block form analogous to the companion matrix of Exercise 2.16. This motivates the following definition.

Definition 8.4 (Companion polynomial). Let P be a matrix polynomial as in (8.2). Then the $n\ell \times n\ell$ matrix polynomial

$$C(t) = t \begin{pmatrix} I_n & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & I_n & 0 \\ 0 & \ldots & 0 & A_\ell \end{pmatrix} + \begin{pmatrix} 0 & -I_n & & 0 \\ \vdots & \ddots & \ddots & \\ 0 & \ldots & 0 & -I_n \\ A_0 & A_1 & \ldots & A_{\ell-1} \end{pmatrix} \tag{8.6}$$

is called the *companion polynomial* of P.

Exercise 8.5 (Companion polynomial). Show that $z \in \mathbb{F}^{n\ell} \setminus \{0\}$ is an eigenvector of C in (8.6) associated with the eigenvalue λ if and only if λ is an eigenvalue of P as in (8.2) with associated eigenvector x that equals the vector consisting of the last n components of z.

Note that the eigenvalue problem (8.5) is *linear* in the sense that is depends linearly on the eigenvalue parameter λ. If $A_\ell = I_n$, then this eigenvalue problem reduces to a standard eigenvalue problem with an $n\ell \times n\ell$ matrix, but if $A_\ell \neq I_n$, then it is a special case of a so-called *generalized eigenvalue problem*.

Definition 8.6 (Matrix pencil). Let $E, A \in \mathbb{F}^{n\times n}$. Then the matrix polynomial $L(t) = tE - A$ is called a *matrix pencil*. The corresponding polynomial eigenvalue problem $L(\lambda)x = 0$ is called a *generalized eigenvalue problem*.

The reason for introducing the minus sign in the term $tE - A$ is the fact that the corresponding generalized eigenvalue problem can then be written in the convenient form $\lambda Ex = Ax$.

8.2 Matrix pencils

In this section, we will investigate matrix pencils $L(t) = tE - A$, where $E, A \in \mathbb{F}^{n\times n}$ and the corresponding generalized eigenvalue problem $\lambda Ex = Ax$. From Remark 8.3 we know that $\lambda \in \mathbb{F}$ is an eigenvalue of L if and only if $\det L(\lambda) = 0$. It is also well known that any matrix $A \in \mathbb{F}^{n\times n}$ has exactly n eigenvalues in $\mathbb{C}$ counted with algebraic multiplicities. This may be different for matrix pencils which may also have infinitely many or no eigenvalues in $\mathbb{C}$.

Example 8.7 (Number of eigenvalues). Let $L_1(t) = tE_1 - A_1$ and $L_2(t) = tE_2 - A_2$, where

$$E_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then for all $\lambda \in \mathbb{C}$ we have

$$\det(\lambda E_1 - A_1) = \det \begin{pmatrix} 1 & \lambda \\ 0 & 0 \end{pmatrix} = 0 \quad \text{and} \quad \det(\lambda E_2 - A_2) = \det \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix} = 1,$$

thus $\lambda E_2 - A_2$ is nonsingular for all $\lambda \in \mathbb{C}$ and consequently $L_2(\lambda)$ does not have any eigenvalues according to Definition 8.2. On the other hand, $\lambda E_1 - A_1$ is singular for all $\lambda \in \mathbb{C}$ and for each $\lambda \in \mathbb{C}$, the vector

$$v_\lambda = \begin{pmatrix} -\lambda \\ 1 \end{pmatrix}$$

is an eigenvector of L_1 associated with the eigenvalue λ.

The pencil $L_1(t)$ in Example 8.7 is a pathologic case and we will therefore restrict ourselves to pencils for which the determinant is not identically zero.

Definition 8.8 (Regular pencils). An $n \times n$ matrix pencil $L(t)$ over $\mathbb{F}$ is called *regular* if $\det L(t)$ is not the zero polynomial. Otherwise, the pencil is called *singular*.

Exercise 8.9 (Singular pencils). Consider the singular matrix pencil

$$L(t) = tE - A = t \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Show that for any $\varepsilon > 0$ and any $\lambda_1, \lambda_2 \in \mathbb{R}$, there exists a regular pencil $t\widetilde{E} - \widetilde{A}$ having the eigenvalues λ_1 and λ_2 such that

$$\max(\|\widetilde{E} - E\|_2, \|\widetilde{A} - A\|_2) \leq \varepsilon.$$

Thus, singular matrix pencils are extremely ill-conditioned in the sense that arbitrarily small perturbations make them regular pencils with arbitrarily prescribed eigenvalues.

A key idea for solving the standard eigenvalue problem $Ax = \lambda x$, where $A \in \mathbb{F}^{n\times n}$, was the application of similarity transformations to A. There is a similar concept for matrix pencils. Let $E, A \in \mathbb{F}^{n\times n}$ and let $x \in \mathbb{F}^n \setminus \{0\}$ be an eigenvector of the pencil $tE - A$ associated with the eigenvalue $\lambda \in \mathbb{F}$, i.e., we have $\lambda Ex = Ax$. Multiplying this equation from the left by a nonsingular matrix $P \in \mathbb{F}^{n\times n}$ and introducing $\widehat{x} := Q^{-1}x$, where $Q \in \mathbb{F}^{n\times n}$ is nonsingular, we find

$$\lambda PEQ\widehat{x} = PAQ\widehat{x}.$$

Thus, any scalar λ is an eigenvalue of the pencil $t\widehat{E} - \widehat{A} := tPEQ - PAQ$ if and only if it is an eigenvalue of the pencil $tE - A$, and any associated eigenvector of $t\widehat{E} - \widehat{A}$ can be transformed to an eigenvector of $tE - A$ associated with λ by multiplication with Q.

Definition 8.10 (Equivalence of pencils). Two $n \times n$ matrix pencils $L_1(t) = tE_1 - A_1$ and $L_2(t) = tE_2 - A_2$ over $\mathbb{F}$ are called *equivalent* if there exist nonsingular matrices $P, Q \in \mathbb{F}^{n\times n}$ such that

$$E_2 = PE_1Q \quad \text{and} \quad A_2 = PA_1Q.$$

Remark 8.11 (Equivalence and similarity). Let $A \in \mathbb{F}^{n\times n}$. Then the standard eigenvalue problem $Ax = \lambda x$ can also be interpreted as a generalized eigenvalue problem of the form $\lambda I_n x = Ax$. Thus, any matrix A can be canonically identified with the matrix pencil $tI_n - A$. Two matrix pencils $L_1(t) = tI_n - A_1$ and $L_2(t) = tI_n - A_2$ with $A_1, A_2 \in \mathbb{F}^{n\times n}$ are equivalent if and only if there exist nonsingular matrices P, Q such that

$$PQ = I_n \quad \text{and} \quad A_2 = PA_1Q.$$

The first identity immediately implies $P = Q^{-1}$ and thus the pencils $L_1(t)$ and $L_2(t)$ are equivalent if and only if A_1 and A_2 are similar. Thus, Definition 8.10 is the natural extension of Definition 2.17 to matrix pencils.

Exercise 8.12 (Invariance of regularity under equivalence). Let the $n \times n$ pencils $L_1(t)$ and $L_2(t)$ over $\mathbb{F}$ be equivalent. Show that L_1 is regular if and only if L_2 is regular.

Exercise 8.13 (Pencils with commuting coefficient matrices). Let $E, A \in \mathbb{F}^{n\times n}$ such that the pencil $L(t) = tE - A$ is regular. Show that $L(t)$ is equivalent to a pencil with commuting coefficients as follows:

1. Let $B, C \in \mathbb{F}^{n\times n}$ be such that $B + C = I_n$. Show that B and C commute, i.e., $BC = CB$.

2. Let $\lambda \in \mathbb{F}$ be such that $\lambda E - A$ is nonsingular. Use the results of 1. to show that the matrices

$$\widetilde{E} := (\lambda E - A)^{-1} E \quad \text{and} \quad \widetilde{A} := (\lambda E - A)^{-1} A$$

commute.

Remark 8.14 (Reduction to standard eigenvalue problems). Let $E, A \in \mathbb{F}^{n\times n}$. If E is nonsingular, then the pencil $tE - A$ is equivalent to both the pencils $tI_n - E^{-1}A$ and $tI_n - AE^{-1}$, i.e., the generalized eigenvalue problem $\lambda Ex = Ax$ is equivalent to the standard eigenvalue problem with $E^{-1}A$ or AE^{-1}.

We have seen in Example 8.7 that a regular $n \times n$ pencil $tE - A$ may have less than n eigenvalues over $\mathbb{C}$ if the matrix E is singular. This comes from the fact that the determinant $\det(tE - A)$ is then a polynomial in t of degree less than n. This "defect" can be repaired by introducing the concept of *infinite eigenvalues*. We begin with the following observation. Let $x \in \mathbb{F}^n \setminus \{0\}$ be an eigenvector of the pencil $tA - E$ associated with the eigenvalue $\lambda \neq 0$, i.e., $\lambda Ax = Ex$. Then x is also an eigenvector of the pencil $tE - A$ now associated with the eigenvalue λ^{-1}, because

$$\lambda^{-1} Ex = \lambda^{-1}(\lambda Ax) = Ax.$$

Thus, the nonzero eigenvalues of the pencil $tE - A$ are exactly the reciprocals of the nonzero eigenvalues of $tA - E$. Extending this reciprocity to include the case $\lambda = 0$ with the convention that zero and ∞ are reciprocals motivates the following definition:

Definition 8.15 (Infinite eigenvalue). Let $E, A \in \mathbb{F}^{n\times n}$. If $x \in \mathbb{F}^n \setminus \{0\}$ is such that $Ex = 0$, then we say that x is an eigenvector of the pencil $tE - A$ associated with the eigenvalue ∞.

A more precise definition of the eigenvalue ∞ can be given by introducing *homogeneous parameters*. If $x \in \mathbb{F}^n \setminus \{0\}$ is an eigenvector of $tE - A$ associated with the eigenvalue $\lambda \in \mathbb{F}$ and if $(\alpha, \beta) \in \mathbb{F}^2$, $\beta \neq 0$ is a pair satisfying $\lambda = \frac{\alpha}{\beta}$, then

$$\lambda E x = Ax \iff \alpha E x = \beta A x.$$

Now let us introduce the equivalence relation $\sim$ on $\mathbb{F}^2 \setminus \{(0,0)\}$ given by

$$(\alpha_1, \beta_1) \sim (\alpha_2, \beta_2) \ :\Longleftrightarrow \ \alpha_1 \beta_2 = \alpha_2 \beta_1.$$

Then for $\alpha, \beta \in \mathbb{F}$ with $\beta \neq 0$ satisfying $\alpha E x = \beta A x$ for some nonzero vector $x \in \mathbb{F} \setminus \{0\}$, the equivalence class

$$[(\alpha, \beta)] := \{(\widetilde{\alpha}, \widetilde{\beta}) \,|\, (\widetilde{\alpha}, \widetilde{\beta}) \sim (\alpha, \beta)\}$$

can be identified with the eigenvalue $\lambda = \frac{\alpha}{\beta}$ of $tE - A$. Note that there is an additional equivalence class

$$\infty := \{(\alpha, 0) \,|\, \alpha \in \mathbb{F} \setminus \{0\}\}.$$

If $x \in \mathbb{F}^n \setminus \{0\}$ is such that $Ex = 0$, then

$$\alpha E x = 0$$

for all $\alpha \neq 0$. Thus, we interpret the equivalence class ∞ as the so-called infinite eigenvalue of $tE - A$. Associated eigenvectors are all nonzero vectors from the null space of E.

Definition 8.16 (Algebraic and geometric multiplicity). Let $E, A \in \mathbb{F}^{n \times n}$ such that $L(t) = tE - A$ is regular and let $\lambda \in \mathbb{F} \cup \{\infty\}$ be an eigenvalue of $L(t)$.

1. If $\lambda \neq \infty$ then the multiplicity $\mu_a(E, A, \lambda)$ of λ as a zero of $\det(tE - A)$ is called the *algebraic multiplicity* of λ.
2. If $\lambda = \infty$ and if $k \in \{0, 1, 2, \ldots, n\}$ is the degree of the polynomial $\det(tE - A)$, then $\mu_a(E, A, \infty) = n - k$ is called the *algebraic multiplicity* of ∞.
3. If $\lambda \neq \infty$ then $\mu_g(E, A, \lambda) := \dim \mathcal{N}(\lambda E - A)$ is called the *geometric multiplicity* of λ.
4. If $\lambda = \infty$ then $\mu_g(E, A, \infty) := \dim \mathcal{N}(E)$ is called the *geometric multiplicity* of ∞.

Remark 8.17 (Number of eigenvalues). By Definition 8.16, any regular $n \times n$ matrix pencil over $\mathbb{F}$ has exactly n eigenvalues over $\mathbb{C}$ counted with algebraic multiplicities.

Exercise 8.18 (Algebraic and geometric multiplicity). Let $E, A \in \mathbb{F}^{n\times n}$ such that $L(t) = tE - A$ is regular.

1. Prove that $\mu_g(E, A, \infty) \leq \mu_a(E, A, \infty)$. *Hint*: use that there exists $k \in \mathbb{N}$ and nonsingular matrices $P, Q \in \mathbb{F}^{n\times n}$ such that
$$PEQ = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}.$$
2. Find an example to show that $\mu_g(E, A, \infty) < \mu_a(E, A, \infty)$ is possible.

8.3 Deflating subspaces and the generalized Schur decomposition

In the theory of matrices, the concept of invariant subspaces is an important generalization of the concept of eigenvectors, see Section 2.5. In this section, we will generalize this concept to the concept of *deflating subspaces* that play the analogous role in the theory of matrix pencils. Recall that for $A \in \mathbb{F}^{n\times n}$ a subspace $\mathcal{X} \subseteq \mathbb{F}^n$ is called *invariant* with respect to A if $x \in \mathcal{X}$ implies $Ax \in \mathcal{X}$. In order to generalize this definition to the case of matrix pencils an alternative characterization of invariant subspaces is more suitable. If $\mathcal{X} \subseteq \mathbb{F}^n$ is a subspace of dimension k, we have

$$\mathcal{X} \text{ is invariant w.r.t. } A \iff \dim\{x + Ay \mid x, y \in \mathcal{X}\} = k.$$

Indeed, the subspace on the right hand side of the equivalence is contained in $\mathcal{X}$ if $\mathcal{X}$ is invariant with respect to A, and hence its dimension does not exceed k. On the other hand, the space also contains $\mathcal{X}$, and therefore it must be of dimension exactly k. We can now use this equivalence to generalize the concept of invariant subspaces to the case of matrix pencils.

Definition 8.19 (Deflating subspace). Let $tE - A$ be an $n \times n$ matrix pencil over $\mathbb{F}$. A subspace $\mathcal{X} \subseteq \mathbb{F}^n$ of dimension k is called a *deflating subspace* of $tE - A$ if

$$\dim\{Ex + Ay \mid x, y \in \mathcal{X}\} \leq k.$$

The inequality in Definition 8.19 is due to the fact that the pencil is allowed to be singular which may lead to a drop in the dimension.

Exercise 8.20 (Deflating subspaces of regular pencils). Let $tE - A$ be a regular $n \times n$ matrix pencil over $\mathbb{F}$ and let $\mathcal{X} \subseteq \mathbb{F}^n$ be a deflating subspace of $tE - A$. Show that

$$\dim\{Ex + Ay \mid x, y \in \mathcal{X}\} = k.$$

Theorem 8.21 (Alternative characterizations of deflating subspaces). *Let $tE - A$ be an $n \times n$ matrix pencil over $\mathbb{F}$ and $\mathcal{X} \subseteq \mathbb{F}^n$ be a subspace of dimension k. Then the following statements are equivalent:*

1. *$\mathcal{X}$ is a deflating subspace of $tE - A$.*
2. *There exists a subspace $\mathcal{Y} \subseteq \mathbb{F}^n$ with $\dim \mathcal{Y} \leq k$ and $A\mathcal{X}, E\mathcal{X} \subseteq \mathcal{Y}$.*
3. *There exist unitary matrices $Z = [z_1, \ldots, z_n]$, $Q \in \mathbb{F}^{n\times n}$ with $\mathrm{span}(z_1, \ldots, z_k) = \mathcal{X}$ such that*

$$tQ^*EZ - Q^*AZ = t\begin{pmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{pmatrix} - \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

where $E_{11}, A_{11} \in \mathbb{F}^{k\times k}$ and $E_{22}, A_{22} \in \mathbb{F}^{(n-k)\times(n-k)}$.

Proof. "1. $\Rightarrow$ 2.": Set $\mathcal{Y} := \{Ex + Ay \mid x, y \in \mathcal{X}\}$. Then $\dim \mathcal{Y} \leq k$ as $\mathcal{X}$ is a deflating subspace of $tE - A$. Clearly, we have $A\mathcal{X}, E\mathcal{X} \subseteq \mathcal{Y}$.

"2. $\Rightarrow$ 3.": Let $z_1, \ldots, z_k$ be an orthonormal basis of $\mathcal{X}$. This basis can be completed to an orthonormal basis $z_1, \ldots, z_n$ of $\mathbb{F}^n$. Similarly, let $q_{k+1}, \ldots, q_n \in \mathcal{Y}^\perp$ be a family of orthonormal vectors from the orthogonal complement of $\mathcal{Y}$ and let $q_1, \ldots, q_n$ be a completion to an orthonormal basis of $\mathbb{F}^n$. Set

$$Z = \begin{pmatrix} Z_1 & Z_2 \end{pmatrix} = [z_1, \ldots, z_n], \quad Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} = [q_1, \ldots, q_n],$$

where $Z_1 = [z_1, \ldots, z_k]$ and $Q_1 = [q_1, \ldots, q_k]$. Then $Q_2^*EZ_1 = 0 = Q_2^*AZ_1$ by construction which implies

$$\begin{aligned} tQ^*EZ - Q^*AZ &= t\begin{pmatrix} Q_1^*EZ_1 & Q_1^*EZ_2 \\ Q_2^*EZ_1 & Q_2^*EZ_2 \end{pmatrix} - \begin{pmatrix} Q_1^*AZ_1 & Q_1^*AZ_2 \\ Q_2^*AZ_1 & Q_2^*AZ_2 \end{pmatrix} \\ &= t\begin{pmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{pmatrix} - \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \end{aligned}$$

where $E_{ij} = Q_i^*EZ_j$ and $A_{ij} = Q_i^*AZ_j$ for $i, j = 1, 2$.

"3. $\Rightarrow$ 1.": Let $Z_1 = [z_1, \ldots, z_k]$. Then $\mathcal{R}(Z_1) = \mathcal{X}$ and by 3. we have

$$Q^*EZ_1 = \begin{pmatrix} E_{11} \\ 0 \end{pmatrix}, \quad Q^*AZ_1 = \begin{pmatrix} A_{11} \\ 0 \end{pmatrix}.$$

From this we obtain $Q^*\{Ex + Ay \mid x, y \in \mathcal{X}\} \subseteq \mathrm{span}(\delta_1, \ldots, \delta_k)$, where $\delta_1, \ldots, \delta_k$ are the first k canonical unit vectors. This implies $\dim\{Ex + Ay \mid x, y \in \mathcal{X}\} \leq k$ as Q is invertible and thus, $\mathcal{X}$ is a deflating subspace. □

Part 3 of Theorem 8.21 is the reason for the name *deflating subspace* that was introduced in [42], because the knowledge of a deflating subspace allows to deflate the eigenvalue problem $(\lambda E - A)x = 0$ into two smaller subproblems with the pencils $tE_{11} - A_{11}$ and $tE_{22} - A_{22}$.

Remark 8.22 (Left deflating subspaces). Sometimes, $\mathcal{Y}$ as in Theorem 8.21 is called a *left deflating subspace* associated with the right deflating subspace $\mathcal{X}$, so $\mathcal{X}, \mathcal{Y}$ together are called a *pair of right and left deflating subspaces*.

A key result in Section 2 was the Schur decomposition of a matrix, because it is a form that displays the eigenvalues of the matrix and can be achieved by applying unitary similarity transformations only. There is a corresponding decomposition of equal importance in the case of matrix pencils, called the *generalized Schur decomposition*.

Theorem 8.23 (Generalized Schur decomposition). *Let $tE - A$ be an $n \times n$ matrix pencil over $\mathbb{C}$. Then there exist unitary matrices $Q, Z \in \mathbb{C}^{n\times n}$ such that*

$$Q^*EZ = R := \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & r_{nn} \end{pmatrix}, \quad Q^*AZ = S := \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ 0 & s_{22} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & s_{nn} \end{pmatrix}.$$

If there exists $k \in \{1, \dots, n\}$ such that $r_{kk} = 0 = s_{kk}$ then $tE - A$ is singular. Otherwise, $tE - A$ has the n eigenvalues

$$\lambda_k = \frac{s_{kk}}{r_{kk}}, \quad k = 1, \dots, n,$$

where we interpret $\lambda_k = \infty$ if $r_{kk} = 0$.

Proof. The proof proceeds by induction on n. The case for $n = 1$ is trivial. Thus, let us assume $n > 1$ and that the assertion of the theorem holds for $n - 1$. Let $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n \setminus \{0\}$ be such that

$$\lambda Ex = Ax.$$

Such a choice is always possible. Indeed, if $tE - A$ is regular, then let λ be an eigenvalue and x be an associated eigenvector. If, on the other hand, $tE - A$ is singular, then $\det(tE - A) \equiv 0$ and for an arbitrary $\lambda \in \mathbb{C}$ choose a nonzero vector x from the kernel of $\lambda E - A$. Then $\operatorname{span}(x)$ is a deflating subspace of $tE - A$. Thus, by Theorem 8.21, there exist unitary matrices $Q, Z \in \mathbb{C}^{n\times n}$, where the first column z_1 of Z spans $\operatorname{span}(x)$ such that

$$tQ^*EZ - Q^*AZ = t \begin{pmatrix} r_{11} & E_{12} \\ 0 & E_{22} \end{pmatrix} - \begin{pmatrix} s_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

where $r_{11}, s_{11} \in \mathbb{C}$ and $E_{22}, A_{22} \in \mathbb{C}^{(n-1)\times(n-1)}$. By the induction hypothesis, there exist unitary matrices $Q_2, Z_2 \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $Q_2^* E_{22} Z_2$ and $Q_2^* A_{22} Z_2$ are both upper triangular. Setting then

$$Q = Q_1 \begin{pmatrix} 1 & 0 \\ 0 & Q_2 \end{pmatrix}, \quad Z = Z_1 \begin{pmatrix} 1 & 0 \\ 0 & Z_2 \end{pmatrix},$$

we obtain the desired decomposition. The remainder of the proof then follows from

$$\det(tE - A) = \det(Q)\det(tR - S)\det(Z^*) = \det(QZ^*) \prod_{k=1}^{n} (t r_{kk} - s_{kk}). \qquad \square$$

Exercise 8.24. Show that an $n \times n$ matrix pencil $tE - A$ over $\mathbb{C}$ has a diagonal generalized Schur decomposition (i.e., R and S in Theorem 8.23 are both diagonal) if and only if there exists an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of $tE - A$.

Exercise 8.25. Let $E, A \in \mathbb{C}^{n\times n}$ be Hermitian. Show that if the matrix pencil $tE - A$ has a diagonal generalized Schur decomposition, then E and A commute, i.e., $EA = AE$.

8.4 Hessenberg-triangular form

In the following, we aim to numerically compute the generalized Schur decomposition. How can this be achieved? Suppose that $L(t) := tE - A$ is an $n \times n$ matrix pencil over $\mathbb{C}$ and suppose that $Q, Z \in \mathbb{C}^{n\times n}$ are unitary such that

$$tQ^*EZ - Q^*AZ = tR - S \tag{8.7}$$

is the generalized Schur decomposition of L. Suppose for the moment, that E is invertible. Then we obtain from (8.7) that

$$Q^*(AE^{-1})Q = Q^*AZZ^*E^{-1}Q = SR^{-1}.$$

Since the inverse of an upper triangular matrix and the product of two upper triangular matrices are again upper triangular matrices, the latter identity means that $Q^*(AE^{-1})Q = SR^{-1}$ is a Schur decomposition of the matrix AE^{-1}. Thus, a possible idea is to first compute a unitary matrix Q that transforms AE^{-1} into Schur form and then to compute a unitary Z so that (8.7) holds. The standard method for computing Q is, of course, the QR iteration from Chapter 5. However, E may be ill-conditioned (or even singular), so we do not want (or are not able) to form the inverse E^{-1} and the product AE^{-1}. Instead, we will try to develop a method that implicitly performs the QR iteration for AE^{-1}, but works directly on the matrices A and E.

Before we do so, recall that the QR iteration requires only $\mathcal{O}(n^2)$ operations for Hessenberg matrices as opposed to $\mathcal{O}(n^3)$ operations for arbitrary matrices. Therefore, a related reduction for matrix pencils would be useful. Observe that if the pencil $tE - A$ is such that A is a Hessenberg matrix and E is upper triangular and invertible, then also AE^{-1} is a Hessenberg matrix.

Definition 8.26 (Hessenberg-triangular form). Let $E, A \in \mathbb{F}^{n\times n}$, where E is upper triangular and A is a Hessenberg matrix. Then we say that the pencil $tE - A$ is in *Hessenberg-triangular form*.

In the following, we will show constructively that for arbitrary $E, A \in \mathbb{F}^{n\times n}$ there exist unitary matrices $Q, Z \in \mathbb{F}^{n\times n}$ such that $tQ^*EZ - Q^*AZ$ is in Hessenberg-triangular form. We will depict the general procedure examplarily for $n = 5$. Thus, we start with matrices of the form

$$E = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}, \quad A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}.$$

In the first step, we compute the QR factorization of E (e.g., with the help of Householder reflections, see Section 4.7), i.e., we compute a unitary $Q_1 \in \mathbb{F}^{n\times n}$ such that Q_1^*E is upper triangular. Then we have

$$E^{(1)} := Q_1^*E = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A^{(1)} := Q_1^*A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix},$$

where blanks stand for zero entries.

In the following, we will transform $A^{(1)}$ step by step to Hessenberg form and simultaneously keep the triangular form of $E^{(1)}$. Alternatingly we will apply *elimination steps* and *restoration steps*. Thus, if k is odd, then after the kth step we eliminate an element of $A^{(k)}$ with the help of a Givens rotation (introduced in Section 5.2), and in the next step, we restore the triangular form of $E^{(k+1)}$, again with the help of a Givens rotation. We start by applying a Givens rotation $G^{(1)}$ that transforms the forth and fifth row in order to eliminate the $(5, 1)$-element of $A^{(1)}$:

$$E^{(2)} := G^{(1)}E^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & \boxtimes & \otimes \end{pmatrix}, \quad A^{(2)} := G^{(1)}A^{(1)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \end{pmatrix}.$$

Note that this transformation destroys the upper triangular form of $E^{(1)}$ as it introduces a fill-in in the $(5,4)$-position of the matrix. But we can restore the triangular form by multiplying $E^{(2)}$ with a Givens rotation $G^{(2)}$ which manipulates the forth and fifth columns of $E^{(2)}$. Application of this Givens rotation to $A^{(2)}$ will keep the zero entry in the $(5,1)$-position, because the first column remains unchanged.

$$E^{(3)} := E^{(2)}G^{(2)} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \end{pmatrix}, \quad A^{(3)} := A^{(2)}G^{(2)} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \end{pmatrix}.$$

Next, we perform an elimination step to eliminate the $(4,1)$-element of $A^{(3)}$ by applying a Givens rotation $G^{(3)}$ to the rows three and four (this time producing a fill-in in the $(4,3)$-position of $E^{(3)}$), followed by a restauration step to restore the triangular form of $E^{(3)}$ by applying a Givens rotation $G^{(4)}$ to the columns three and four:

$$E^{(4)} := G^{(3)}E^{(3)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & \boxtimes & \otimes & \otimes \\ & & & & \times \end{pmatrix}, \quad A^{(4)} := G^{(3)}A^{(3)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \times \end{pmatrix},$$

$$E^{(5)} := E^{(4)}G^{(4)} = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & 0 & \otimes & \times \\ & & & & \times \end{pmatrix}, \quad A^{(5)} := A^{(4)}G^{(4)} = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \end{pmatrix}.$$

We continue this procedure until we have annihilated all but the first two elements of the first column of A. In the case $n = 5$, we need one more pair of elimination-restauration-steps:

$$E^{(6)} := G^{(5)}E^{(5)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & \boxtimes & \otimes & \otimes & \otimes \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A^{(6)} := G^{(5)}A^{(5)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \end{pmatrix},$$

$$E^{(7)} := E^{(6)}G^{(6)} = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & 0 & \otimes & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A^{(7)} := A^{(6)}G^{(6)} = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \end{pmatrix}.$$

Observe that at this stage we cannot proceed with eliminating the element in the $(2,1)$-position of $A^{(7)}$, because this would produce a fill-in in the $(2,1)$-position of $E^{(7)}$, and the following restauration step would work on the first and second columns of our matrices, thereby destroying all the strenuously obtained zeros in $A^{(7)}$. Instead, we now continue with the second column of $A^{(7)}$, again starting from the bottom. Note that all our transformation will preserve the zero structure of the first column of $A^{(7)}$:

$$E^{(8)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & \boxtimes & \otimes \end{pmatrix}, \quad A^{(8)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes & \otimes \end{pmatrix},$$

$$E^{(9)} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \end{pmatrix}, \quad A^{(9)} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \times & \otimes & \otimes \end{pmatrix},$$

$$E^{(10)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & \boxtimes & \otimes & \otimes \\ & & & & \times \end{pmatrix}, \quad A^{(10)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & & \times & \times & \times \end{pmatrix},$$

$$E^{(11)} = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & 0 & \otimes & \times \\ & & & & \times \end{pmatrix}, \quad A^{(11)} = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \end{pmatrix}.$$

For the second column, we have to stop the procedure after eliminating all but the first three element, because proceeding by eliminating the $(3,2)$-entry of $A^{(11)}$ would lead to a fill-in in the $(3,1)$-position of $A^{(11)}$), thereby destroying the zero structure of the first column. We now continue this series of steps until A has reached Hessenberg form. In the case $n = 5$ only one more pair of steps is necessary:

$$E^{(12)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & \boxtimes & \otimes \end{pmatrix}, \quad A^{(12)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \end{pmatrix},$$

$$E^{(13)} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \end{pmatrix}, \quad A^{(13)} = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \end{pmatrix}.$$

Exercise 8.27. Write a procedure hessenberg_triangular(**var** E, A) that computes the Hessenberg-triangular form for an $n \times n$ matrix pencil $tE - A$ over $\mathbb{F}$.

Exercise 8.28. Calculate how many operations are necessary to compute the Hessenberg triangular form for an $n \times n$ matrix pencil $tE - A$ over $\mathbb{F}$. Also calculate the number of operations needed to explicitly compute the transformation matrices Q and Z.

8.5 Deflation

Deflation was an important strategy in the practical use of the QR iteration as noted in Section 5.4, and the same is true for generalized eigenvalue problems. Assume that our $n \times n$ matrix pencil $tE - A$ over $\mathbb{F}$ is regular and in Hessenberg-triangular form. If one or more of the subdiagonal elements of A are zero (or, sufficiently small in magnitude), then the corresponding generalized eigenvalue problem can be divided into two smaller generalized eigenvalue problems of smaller size. Let $k \in \{1, \ldots, n\}$ be an index such that $a_{k+1,k} \neq 0$. Then the pencil $tE - A$ has the form

$$tE - A = \begin{pmatrix} tE_{11} - A_{11} & tE_{12} - A_{12} \\ 0 & tE_{22} - A_{22} \end{pmatrix},$$

where $tE_{11} - A_{11}$ and $tE_{22} - A_{22}$ are two pencils of sizes $k \times k$ and $(n-k) \times (n-k)$, respectively, that are again in Hessenberg-triangular form. Thus, we can compute the eigenvalues of $tE - A$ by computing the eigenvalues of the pencils $tE_{11} - A_{11}$ and $tE_{22} - A_{22}$. In practice, we will deflate the problem as soon as there exists an index k for $a_{k+1,k} \approx 0$.

Next assume that all subdiagonal entries of A are nonzero and assume that the pencil $tE - A$ has the eigenvalue ∞. Then E is singular and necessarily must have a zero entry on the diagonal. Thus, the eigenvalue ∞ obviously plays an exceptional role, because it is automatically displayed after reduction to Hessenberg-triangular form. Due to this exceptional role, it may not be a complete surprise that we are also able to *deflate* it. The corresponding procedure is called *zero-chasing*, and we will explain it exemplarily on a 5×5 matrix pencil. Thus, let us assume that $e_{33} = 0$,

where $\gamma = 3$ is the smallest index for which $e_{kk} \neq 0$ for all $k = \gamma + 1, \ldots, n$. Thus, our pencil has the following pattern.

$$E = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

Our aim is to chase the 'zero' on the diagonal of E down to the lower right corner of E. In the first step, we apply a Givens rotation G_1 to E and A acting on the third and forth rows that will eliminate the $(4,4)$-entry of E.

$$E^{(1)} = G_1 E = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \\ & & & & \times \end{pmatrix}, \quad A^{(1)} = G_1 A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & \boxtimes & \otimes & \otimes & \otimes \\ & & & \times & \times \end{pmatrix}.$$

This procedure will generate a fill-in in the $(4,2)$-position of A thus creating a bulge in the Hessenberg form. However, applying a Givens rotation G_2 acting on the columns two and three, we can restore the Hessenberg form of $A^{(1)}$ without destroying the upper triangular form of $E^{(1)}$.

$$E^{(2)} = E^{(1)} G_2 = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & & & \times & \times \\ & & & & \times \\ & & & & \times \end{pmatrix}, \quad A^{(2)} = A^{(1)} G_2 = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & 0 & \otimes & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

Note that we chased the zero on the diagonal of E further down while the Hessenberg form of A has been restored. (There is an additional zero entry remaining in the $(3,3)$-position of E, but this one will fill up again later, so it is out of our focus.) We continue now to chase the zero down to the (n,n)-position.

$$E^{(3)} = G_3 E^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & & \times & \times \\ & & & & \otimes \\ & & & & 0 \end{pmatrix}, \quad A^{(3)} = G_3 A^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & \boxtimes & \otimes & \otimes \end{pmatrix},$$

$$E^{(4)} = E^{(3)} G_4 = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \boxtimes & \otimes & \times \\ & & & & \times \\ & & & & \end{pmatrix}, \quad A^{(4)} = A^{(3)} G_4 = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & 0 & \times & \times \end{pmatrix}.$$

Observe that the last step has finally filled up the $(3,3)$-entry of $E^{(3)}$. However, as $E^{(3)}$ still has upper triangular form, this is acceptable. The pattern of the matrices $E^{(4)}$ and $A^{(4)}$ is the typical pattern that we obtain after chasing down the zero on the diagonal of E to the (n,n)-position. In order to deflate the problem, we can apply a final Givens rotation G_5 acting on columns that eliminates the $(5,4)$-entry of $A^{(4)}$. This will create a fill-in in the $(4,4)$-position of $E^{(4)}$, but again this is acceptable as the upper triangular form is not destroyed.

$$E^{(5)} = E^{(4)}G_5 = \left(\begin{array}{ccccc} \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \boxtimes & \otimes \\ \hline & & & & \end{array}\right), \quad A^{(5)} = A^{(4)}G_5 = \left(\begin{array}{ccccc} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ \hline & & & 0 & \otimes \end{array}\right).$$

We can now deflate the eigenvalue ∞ and continue with the $(n-1)\times(n-1)$ matrix pencil from the upper left corner of $tE^{(5)} - A^{(5)}$ which is again in Hessenberg-triangular form.

Exercise 8.29. Write a procedure zero_chasing(**var** E, A) that deflates the eigenvalue ∞ (which may have algebraic multiplicity larger than one) for a given $n \times n$ matrix pencil $tE - A$ over $\mathbb{F}$ in Hessenberg-triangular form.

8.6 The QZ step

Let $tE - A$ be a regular $n \times n$ matrix pencil over $\mathbb{F}$ in Hessenberg-triangular form. By the previous section, we may also assume that E is invertible, because otherwise we could deflate the eigenvalue ∞ until we are left with a smaller pencil in Hessenberg-triangular form not having the eigenvalue ∞ any more. The idea is then to determine unitary matrices $Q, Z \in \mathbb{F}^{n\times n}$ such that $t\widetilde{E} - \widetilde{A} := Q^*(tE - A)Z$ is again in Hessenberg-triangular form and such that $\widetilde{A}\widetilde{E}^{-1}$ is the matrix that one would obtain after one step of the implicit QR iteration with multi-shift strategy outlined in Section 5.6. We will do this exemplary for the case $n = 5$ and the double shift strategy as in Example 5.17.

As in Chapter 5, the main idea is the application of Corollary 5.14. In the following let $H := AE^{-1}$ and let $\mu_1, \mu_2 \in \mathbb{C}$ be two shifts. In Section 5.6, we started by investigating the matrix $B = (H - \mu_1 I_n)(H - \mu_2 I_n)$ which has the form

$$B = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \end{pmatrix}.$$

We then computed a Householder reflection M_1 that eliminates the $(2,1)$- and $(3,1)$-entries of B, but instead of applying M_1 directly to B and computing the QR factorization of B, we applied M_1 directly to the matrix H and restored its Hessenberg form with *bulge chasing*. Corollary 5.14 then implied that the resulting Hessenberg matrix was essentially the matrix that we would have obtained after two consecutive steps of the QR iteration with shifts μ_1 and μ_2. Applying the same strategy here, we will have to compute a Householder reflection that reflects the first column b_1 of the matrix $B = (H - \mu_1 I_n)(H - \mu_2 I_n)$ to a multiple of the first canonical unit vector. It is important to highlight that the first column of B can be easily computed without explicitly forming E^{-1} and the product AE^{-1}, see Exercise 8.30.

Exercise 8.30 (First column of B). Let $E, A \in \mathbb{C}^{n\times n}$ such that E is nonsingular and $tE - A$ is in Hessenberg-triangular form. Show that only the first three entries of the first column

$$b_1 = (AE^{-1} - \mu_1 I_n)(AE^{-1} - \mu_2 I_n)\delta_1$$

of the matrix $B = (AE^{-1} - \mu_1 I_n)(AE^{-1} - \mu_2 I_n)$ are nonzero and provide a formula for its entries in dependence of μ_1, μ_2 and the entries of the first two columns of A and E.
Hint: Use that E^{-1} is upper triangular and that you will need the entries of the principal 2×2 submatrix only.

Once we have computed the first column b_1 of B, we can compute a Householder reflection M_1 such that $M_1 b_1$ is a multiple of the first canonical unit vector. By Exercise 8.30, only the first three entries of b_1 are nonzero, so M_1 will have the form

$$M_1 = \begin{pmatrix} P & 0 \\ 0 & I_{n-3} \end{pmatrix},$$

where P is a 3×3 Householder reflection. Thus M_1 only manipulates the first three rows of any matrix it will be applied to. Applying M_1 both to E and A, we therefore obtain matrices of the following form.

$$E^{(0)} = M_1 E = \begin{pmatrix} \times & \times & \times & \times & \times \\ \boxtimes & \times & \times & \times & \times \\ \boxtimes & \boxtimes & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A^{(0)} = M_1 A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \boxtimes & \times & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}.$$

Thus, we have created a triangular and a Hessenberg matrix each having a bulge. If we can now find unitary matrices $Q, Z \in \mathbb{C}^{n\times n}$ such that $tQ^*M_1EZ - Q^*M_1AZ$ is again in Hessenberg-triangular form and such that M_1^*Q and M_1^* have identical first columns, then by Corollary 5.14 the matrix

$$Q^*M_1AZ(Q^*M_1EZ)^{-1} = Q^*M_1AE^{-1}M_1^*Q$$

is essentially the matrix that we would have obtained after two steps of the QR iteration with shifts μ_1 and μ_2 applied to AE^{-1}. As in Chapter 5, we will achieve this by *bulge chasing*. We start by applying a Householder reflection Z_1 that changes the first three columns to eliminate $e_{31}^{(0)}$ and $e_{32}^{(0)}$. This will create fill-ins in the $(4,1)$- and $(4,2)$-positions of $A^{(0)}$.

$$E^{(1)} = E^{(0)}Z_1 = \begin{pmatrix} \otimes & \otimes & \otimes & \times & \times \\ \otimes & \otimes & \otimes & \times & \times \\ 0 & 0 & \otimes & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A^{(1)} = A^{(0)}Z_1 = \begin{pmatrix} \otimes & \otimes & \otimes & \times & \times \\ \otimes & \otimes & \otimes & \times & \times \\ \otimes & \otimes & \otimes & \times & \times \\ \boxtimes & \boxtimes & \otimes & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

Next, we apply a Householder reflection Z_2 that changes the first two columns to eliminate $e_{21}^{(0)}$. Note that this will not create any additional fill-ins in our matrices.

$$E^{(2)} = E^{(1)}Z_2 = \begin{pmatrix} \otimes & \otimes & \times & \times & \times \\ 0 & \otimes & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad A^{(2)} = A^{(1)}Z_2 = \begin{pmatrix} \otimes & \otimes & \times & \times & \times \\ \otimes & \otimes & \times & \times & \times \\ \otimes & \otimes & \times & \times & \times \\ \otimes & \otimes & \times & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

These two step have restored the triangular form of E. We now concentrate on the matrix $A^{(2)}$ and apply a Householder reflection Q_1^* changing the rows two to four to eliminate the entries $a_{31}^{(2)}$ and $a_{41}^{(2)}$.

$$E^{(3)} = Q_1^*E^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & \boxtimes & \otimes & \otimes & \otimes \\ & \boxtimes & \boxtimes & \otimes & \otimes \\ & & & & \times \end{pmatrix}, \quad A^{(3)} = Q_1^*A^{(2)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \otimes & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ 0 & \otimes & \otimes & \otimes & \otimes \\ & & & \times & \times \end{pmatrix}.$$

The fact that Q_1^* only changes the rows two to four, its conjugate transpose Q_1 will manipulate only the columns two to four if we multiply it from the right to M_1^*. Thus, M_1^* and $M_1^*Q_1$ will have identical first columns. Observe that as in Chapter 5, we have managed to chase the bulge one step towards the right lower edge of the matrix. We continue to chase the bulge further down and finally off the diagonal to restore the Hessenberg-triangular form of our pencil.

$$E^{(4)} = E^{(3)}Z_3 = \begin{pmatrix} \times & \otimes & \otimes & \otimes & \times \\ \times & \otimes & \otimes & \otimes & \times \\ & \otimes & \otimes & \otimes & \times \\ & 0 & 0 & \otimes & \times \\ & & & & \times \end{pmatrix}, \quad A^{(4)} = A^{(3)}Z_3 = \begin{pmatrix} \times & \otimes & \otimes & \otimes & \times \\ \times & \otimes & \otimes & \otimes & \times \\ & \otimes & \otimes & \otimes & \times \\ & \otimes & \otimes & \otimes & \times \\ & \boxtimes & \boxtimes & \otimes & \times \end{pmatrix},$$

$$E^{(5)} = E^{(4)}Z_4 = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ \times & \otimes & \otimes & \times & \times \\ 0 & \otimes & \times & \times & \\ & & \times & \times & \\ & & & \times & \end{pmatrix}, \quad A^{(5)} = A^{(4)}Z_4 = \begin{pmatrix} \times & \otimes & \otimes & \times & \times \\ \times & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \\ & \otimes & \otimes & \times & \times \end{pmatrix},$$

$$E^{(6)} = Q_2^* E^{(5)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & \boxtimes & \otimes & \otimes \\ & & \boxtimes & \boxtimes & \otimes \end{pmatrix}, \quad A^{(6)} = Q_2^* A^{(5)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \otimes & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \\ & 0 & \otimes & \otimes & \otimes \end{pmatrix},$$

$$E^{(7)} = E^{(6)}Z_5 = \begin{pmatrix} \times & \times & \otimes & \otimes & \otimes \\ & \times & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \\ & & 0 & 0 & \otimes \end{pmatrix}, \quad A^{(7)} = A^{(6)}Z_5 = \begin{pmatrix} \times & \times & \otimes & \otimes & \otimes \\ \times & \times & \otimes & \otimes & \otimes \\ & \times & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \\ & & \otimes & \otimes & \otimes \end{pmatrix},$$

$$E^{(8)} = E^{(7)}Z_6 = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & 0 & \otimes & \times \\ & & & & \times \end{pmatrix}, \quad A^{(8)} = A^{(7)}Z_6 = \begin{pmatrix} \times & \times & \otimes & \otimes & \times \\ \times & \times & \otimes & \otimes & \times \\ & \times & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \\ & & \otimes & \otimes & \times \end{pmatrix},$$

$$E^{(9)} = Q_3^* E^{(8)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \otimes & \otimes \\ & & & \boxtimes & \otimes \end{pmatrix}, \quad A^{(9)} = Q_3^* A^{(8)} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \otimes & \otimes & \otimes \\ & & 0 & \otimes & \otimes \end{pmatrix},$$

$$E^{(10)} = E^{(9)}Z_7 = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \\ & & & 0 & \otimes \end{pmatrix}, \quad A^{(10)} = A^{(9)}Z_7 = \begin{pmatrix} \times & \times & \times & \otimes & \otimes \\ \times & \times & \times & \otimes & \otimes \\ & \times & \times & \otimes & \otimes \\ & & \times & \otimes & \otimes \\ & & & \otimes & \otimes \end{pmatrix}.$$

In the general $n \times n$ case, this procedure yields matrices

$$Q = Q_0 \cdots Q_{n-2}, \quad Z = Z_1 \cdots Z_{2n-3}$$

such that the pencil $t\widetilde{E} - \widetilde{A} := Q^*(tE - A)Z$ is in Hessenberg-triangular form and such that $M_1^* Q$ and M_1^* have identical first columns. As explained above, the matrix $\widetilde{A}\widetilde{E}^{-1}$ is essentially the Hessenberg matrix that is obtained after two steps of the QR iteration with shifts μ_1 and μ_2 applied to the Hessenberg matrix AE^{-1}. We can now repeat the steps until one or more elements on the subdiagonal of AE^{-1} are

sufficiently small to be considered as zero. Clearly, the $(k+1,k)$-element in AE^{-1} is zero if and only if the $(k+1,k)$-element in A is zero, since with E also E^{-1} is upper triangular. In this case, we have deflation and can continue with two subproblems of smaller size in order to compute the eigenvalues of our matrix pencil.

The algorithm outlined above is the so-called *QZ-algorithm* and was originally proposed in [28]. It can be considered to be the standard algorithm for the solution of dense (i.e., not sparse) generalized eigenvalue problems. For further reading, we refer the reader to [28], [9], and [18].

Exercise 8.31. Write a procedure qz_step(**var** E, A) that performs a QZ step on an $n \times n$ matrix pencil $tE - A$ over $\mathbb{F}$ in Hessenberg-triangular form.

Exercise 8.32. Calculate how many operations are necessary to perform a QZ step on an $n \times n$ matrix pencil $tE - A$ over $\mathbb{F}$ in Hessenberg-triangular form. Also calculate the number of operations needed to explicitly compute the transformation matrices Q and Z.

Bibliography

[1] W. E. Arnoldi, The principle of minimized iterations in the solution of the matrix eigenvalue problem, *Quart. Appl. Math.* **9** (1951), 17–29.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems*, SIAM, Philadelphia, PA, 2000.

[3] F. L. Bauer and C. T. Fike, Norms and exclusion theorems, *Numer. Math.* **2** (1960), 137–141.

[4] K. Braman, R. Byers and R. Mathias, The Multishift QR Algorithm. Part I: Maintaining Well-Focused Shifts and Level 3 Performance, *SIAM Mat. Anal. Appl.* **23** (2002), 929–947.

[5] K. Braman, R. Byers and R. Mathias, The Multishift QR Algorithm. Part II: Aggressive Early Deflation, *SIAM Mat. Anal. Appl.* **23** (2002), 948–973.

[6] R. Courant and D. Hilbert, *Methoden der mathematischen Physik*, Springer, 1924.

[7] J. J. M. Cuppen, A Divide and Conquer Method for the Symmetric Tridiagonal Eigenproblem, *Numer. Math.* **36** (1981), 177–195.

[8] J. Demmel and K. Veselić, Jacobi's method is more accurate than QR, *SIAM J. Matrix Anal. Appl.* **13** (1992), 1204–1245.

[9] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, 1997.

[10] I. S. Dhillon, *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Ph. D. thesis, University of California, Berkeley, 1997.

[11] I. S. Dhillon and B. N. Parlett, Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices, *Lin. Alg. Appl.* **387** (2004), 1–28.

[12] J. Dongarra and D. Sorensen, A fully parallel algorithm for the symmetric eigenvalue problem, *SIAM J. Sci. Statist. Comput.* **8** (1987), 139–154.

[13] Z. Drmač, A Global Convergence Proof for Cyclic Jacobi Methods with Block Rotations, *SIAM J. Matrix Anal. Appl.* **31** (2009), 1329–1350.

[14] J. G. F. Francis, The QR Transformation. A Unitary Analogue to the LR Transformation — Part 1, *The Computer Journal* **4** (1961), 265–271.

[15] J. G. F. Francis, The QR Transformation — Part 2, *The Computer Journal* **4** (1962), 332–345.

[16] S. Gershgorin, Über die Abgrenzung der Eigenwerte einer Matrix, *Izv. Akad. Nauk SSSR Ser. Mat. 1* (1931), 749–754.

[17] L. Giraud, J. Langou, M. Rozložnik and J. van den Eshof, Rounding error analysis of the classical Gram-Schmidt orthogonalization process, *Numer. Math.* **101** (2005), 87–100.

[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, London, 1996.

[19] P. Henrici, On the Speed of Convergence of Cyclic and Quasicyclic Jacobi Methods for Computing Eigenvalues of Hermitian Matrices, *J. Soc. Ind. Appl. Math* **6** (1958), 144–162.

[20] K. Hessenberg, *Behandlung linearer Eigenwertaufgaben mit Hilfe der Hamilton-Cayleyschen Gleichung*, Institut für Praktische Mathematik, TH Darmstadt, Report no. 1, 1940, available at `http://www.hessenberg.de/numverf.pdf`.

[21] A. S. Householder, Unitary Triangularization of a Nonsymmetric Matrix, *Journal of the ACM* **5** (1958), 339–342.

[22] C. G. J. Jacobi, Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen, *Journal für die reine und angewandte Mathematik* **30** (1846), 51–94.

[23] A. Knyazev and K. Neymeyr, A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems, *Linear Algebra Appl.* **358** (2003), 95–114.

[24] D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems*, Springer, 2005.

[25] A. N. Krylov, On the numerical solution of the equation by which, in technical matters, frequencies of small oscillations of material systems are determined, *Izv. Akad. Nauk. S.S.S.R. Otdel. Mat. Estest.* **1** (1931), 491–539.

[26] B. H. Kublanovskaja, On some algorithms for the solution of the complete problem of proper values, *J. Comput. Math. and Math. Phys.* **1** (1961), 555–570.

[27] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Res. Nat. Bur. Stand.* **45** (1950), 255–282.

[28] C. B. Moler and G. W. Stewart, An algorithm for generalized matrix eigenvalue problems, *SIAM J. Numer. Anal.* **10** (1973), 241–256.

[29] K. Neymeyr, A geometric theory for preconditioned inverse iteration. I: Extrema of the Rayleigh quotient, *Linear Algebra Appl.* **322** (2001), 61–85.

[30] K. Neymeyr, A geometric theory for preconditioned inverse iteration. II: Convergence estimates, *Linear Algebra Appl.* **322** (2001), 87–104.

[31] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford InfoLab, Report, 1999.

[32] B. N. Parlett, Global Convergence of the Basic QR Algorithm on Hessenberg Matrices, *Math. Comp.* **22** (1968), 803–817.

[33] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, 1987.

[34] O. Perron, Zur Theorie der Matrices, *Mathematische Annalen* **64** (1907), 248–263.

[35] E. Pohlhausen, Berechnung der Eigenschwingungen statisch-bestimmter Fachwerke, *ZAMM — Zeitschrift für Angewandte Mathematik und Mechanik* **1** (1921), 28–42.

[36] H. Rutishauser, Der Quotienten-Differenzen-Algorithmus, *ZAMP — Zeitschrift für Angewandte Mathematik und Physik* **5** (1954), 233–251.

[37] H. Rutishauser, Une methode pour la determination des valeurs propres d'une matrice, *Comptes Rendues de l'Academie des Sciences Paris* **240** (1955), 34–36.

[38] H. Rutishauser, Solution of eigenvalue problems with the LR-transformation, *National Bureau of Standards Applied Mathematics Series* **49** (1958), 47–81.

[39] A. Schönhage, Zur Konvergenz des Jacobi-Verfahrens, *Numer. Math.* **3** (1961), 374–380.

[40] G. L. G. Sleijpen and H. A. A. van der Vorst, A Jacobi-Davidson iteration method for linear eigenvalue problems, *SIAM J. Matrix Anal. Appl.* **17** (1996), 401–425.

[41] G. L. G. Sleijpen and H. A. A. van der Vorst, A Jacobi-Davidson iteration method for linear eigenvalue problems, *SIAM Review* **42** (2000), 267–293.

[42] G. W. Stewart, On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$, *SIAM J. Numer. Anal.* **9** (1972), 669–686.

[43] G. W. Stewart, *Matrix Algorithms. Volume I: Basic Decompositions*, SIAM, 1998.

[44] G. W. Stewart, *Matrix Algorithms. Volume II: Eigensystems*, SIAM, 2001.

[45] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, 1997.

[46] H. P. M. van Kempen, On the Quadratic Convergence of the Special Cyclic Jacobi Method, *Numer. Math.* **9** (1966), 19–22.

[47] R. von Mises and H. Pollaczek-Geiringer, Praktische Verfahren der Gleichungsauflösung, *ZAMM — Zeitschrift für Angewandte Mathematik und Mechanik* **9** (1929), 152–164.

[48] D. S. Watkins, Understanding the QR algorithm, *SIAM Review* **24** (1982), 427–440.

[49] D. S. Watkins, *The matrix eigenvalue problem*, SIAM, 2007.

[50] D. S. Watkins, The QR algorithm revisited, *SIAM Rev,* **50** (2008), 133–145.

[51] J. H. Wilkinson, Note on the Quadratic Convergence of the Cyclic Jacobi Process, *Numer. Math.* **4** (1962), 296–300.

[52] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

Index