Lecture Notes for Inf-Mat 4350, 2010

Tom Lyche

August 28, 2010

Contents

Preface

These lecture notes contains the text for a course in matrix analysis and numerical linear algebra given at the beginning graduate level at the University of Oslo. In the appendix we give a review of basis linear algebra. Each of the chapters correspond approximately to one week of lectures.

Oslo, 12 August, 2010

Tom Lyche

List of Figures

List of Tables

List of Algorithms

List of Exercises

Chapter 1

Introduction

1.1 Notation

The following sets will be used throughout these notes.

1. The set of natural numbers, integers, rational numbers, real numbers, and complex numbers are denoted by $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, respectively.
2. We use the "colon equal" symbol $v := e$ to indicate that the symbol v is defined by the expression e.
3. $\mathbb{R}^n$ is the set of n-tuples of real numbers which we will represent as column vectors. Thus $\boldsymbol{x} \in \mathbb{R}^n$ means
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$
where $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Row vectors are normally identified using the transpose operation. Thus if $\boldsymbol{x} \in \mathbb{R}^n$ then $\boldsymbol{x}$ is a column vector and $\boldsymbol{x}^T$ is a row vector.
4. Addition and scalar multiplication are denoted, and defined by
$$\boldsymbol{x} + \boldsymbol{y} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}, \quad a\boldsymbol{x} = \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \quad a \in \mathbb{R}.$$
5. $\mathbb{R}^{m,n}$ is the set of $m \times n$ matrices with real elements represented as
$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The element in the ith row and jth column of a matrix $\boldsymbol{A}$ will be denoted by $a_{i,j}$, a_{ij}, $\boldsymbol{A}(i,j)$ or $(\boldsymbol{A})_{i,j}$. We use the notations

$$\boldsymbol{a}_{:j} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}, \quad \boldsymbol{a}_{i:}^T = [a_{i1}, a_{i2}, \ldots, a_{in}], \quad \boldsymbol{A} = [\boldsymbol{a}_{:1}, \boldsymbol{a}_{:2}, \ldots \boldsymbol{a}_{:n}] = \begin{bmatrix} \boldsymbol{a}_{1:}^T \\ \boldsymbol{a}_{2:}^T \\ \vdots \\ \boldsymbol{a}_{m:}^T \end{bmatrix}$$

for the columns $\boldsymbol{a}_{:j}$ and rows $\boldsymbol{a}_{i:}^T$ of $\boldsymbol{A}$. We often drop the colon and write $\boldsymbol{a}_j$ and $\boldsymbol{a}_i^T$ when no confusion can arise. If $m = 1$ then $\boldsymbol{A}$ is a row vector, if $n = 1$ then $\boldsymbol{A}$ is a column vector, while if $m = n$ then $\boldsymbol{A}$ is a square matrix. In this text we will denote matrices by boldface capital letters $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \cdots$ and vectors most often by boldface lower case letters $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \cdots$.

6. The imaginary unit $\sqrt{-1}$ is denoted by i. The complex conjugate and the modulus of a complex number z is denoted by $\overline{z}$ and $|z|$, respectively. Thus if $z = x + iy = re^{i\phi} = r(\cos\phi + i\sin\phi)$ is a complex number then $\overline{z} := x - iy = re^{-i\phi} = \cos\phi - i\sin\phi$ and $|z| := \sqrt{x^2 + y^2} = r$. $\mathrm{Re}(z) := x$ and $\mathrm{Im}(z) := y$ denote the real and imaginary part of the complex number z.
7. For matrices and vectors with complex elements we use the notation $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $\boldsymbol{x} \in \mathbb{C}^n$. We identify complex row vectors using either the transpose T or the conjugate transpose operation $\boldsymbol{x}^* := \overline{\boldsymbol{x}}^T = [\overline{x}_1, \ldots, \overline{x}_n]$.
8. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ and $a \in \mathbb{C}$ the operations of vector addition and scalar multiplication is defined by component operations as in the real case. (Cf. 4).
9. The arithmetic operations on rectangular matrices are

   - **matrix addition** $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$ if $c_{ij} = a_{ij} + b_{ij}$ for all i, j and $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are matrices of the same dimension.
   - **multiplication by a scalar** $\boldsymbol{C} = \alpha\boldsymbol{A}$, where $c_{ij} = \alpha a_{ij}$ for all i, j.
   - **multiplication by another matrix** $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$, $\boldsymbol{C} = \boldsymbol{A}\cdot\boldsymbol{B}$ or $\boldsymbol{C} = \boldsymbol{A}*\boldsymbol{B}$, where $\boldsymbol{A} \in \mathbb{C}^{m,p}$, $\boldsymbol{B} \in \mathbb{C}^{p,n}$, $\boldsymbol{C} \in \mathbb{C}^{m,n}$, and $c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$.
   - **element-by-element matrix operations** $\boldsymbol{C} = \boldsymbol{A} \times \boldsymbol{B}$ and $D = \boldsymbol{A}/\boldsymbol{B}$, and $E = \boldsymbol{A} \wedge r$ where all matrices are of the same dimension and $c_{ij} = a_{ij}b_{ij}$, $d_{ij} = a_{ij}/b_{ij}$ and $e_{ij} = a_{ij}^r$ for all i, j and suitable r. The element-by-element product $\boldsymbol{C} = \boldsymbol{A} \times \boldsymbol{B}$ is known as the **Schur product** and also the **Hadamard product**.

10. Let $\boldsymbol{A} \in \mathbb{R}^{m,n}$ or $\boldsymbol{A} \in \mathbb{C}^{m,n}$. The **transpose** $\boldsymbol{A}^T$, and **conjugate transpose** $\boldsymbol{A}^*$ are n, m matrices with elements $a_{ij}^T = a_{ji}$ and $a_{ij}^* = \overline{a}_{ji}$, respectively. If $\boldsymbol{B}$ is an n, p matrix then $(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T$ and $(\boldsymbol{A}\boldsymbol{B})^* = \boldsymbol{B}^*\boldsymbol{A}^*$.

11. The **unit vectors** in $\mathbb{R}^n$ and $\mathbb{C}^n$ are denoted by

$$\boldsymbol{e}_1 := \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \boldsymbol{e}_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \boldsymbol{e}_3 := \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \boldsymbol{e}_n := \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

while $\boldsymbol{I}_n = \boldsymbol{I} =: [\delta_{ij}]_{i,j=1}^n$, where

$$\delta_{ij} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

is the **identity matrix** of order n. Both the columns and the transpose of the rows of $\boldsymbol{I}$ are the unit vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.

12. We use the following notations for diagonal- and tridiagonal $n \times n$ matrices

$$\operatorname{diag}(d_i) = \operatorname{diag}(d_1, \ldots, d_n) := \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} = \begin{bmatrix} d_1 & & & \\ & \ddots & & \\ & & & d_n \end{bmatrix},$$

$$\boldsymbol{B} = \operatorname{tridiag}(a_i, d_i, c_i) = \operatorname{tridiag}(\boldsymbol{a}, \boldsymbol{d}, \boldsymbol{c}) := \begin{bmatrix} d_1 & c_1 & & & \\ a_2 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & d_{n-1} & c_{n-1} \\ & & & a_n & d_n \end{bmatrix}.$$

Here $b_{ii} = d_i$ for $i = 1, \ldots, n$, $b_{i+1,i} = a_{i+1}$, $b_{i,i+1} = c_i$ for $i = 1, \ldots, n-1$, and $b_{ij} = 0$ otherwise.

13. Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $1 \le i_1 < i_2 < \cdots < i_r \le m$, $1 \le j_1 < j_2 < \cdots < j_c \le n$. The matrix $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j}) \in \mathbb{C}^{r,c}$ is the submatrix of $\boldsymbol{A}$ consisting of rows $\boldsymbol{i} := [i_1, \ldots, i_r]$ and columns $\boldsymbol{j} := [j_1, \ldots, j_c]$

$$A(\boldsymbol{i}, \boldsymbol{j}) := \boldsymbol{A}\begin{pmatrix} i_1 & i_2 & \cdots & i_r \\ j_1 & j_2 & \cdots & j_c \end{pmatrix} = \begin{bmatrix} a_{i_1,j_1} & a_{i_1,j_2} & \cdots & a_{i_1,j_c} \\ a_{i_2,j_1} & a_{i_2,j_2} & \cdots & a_{i_2,j_c} \\ \vdots & \vdots & & \vdots \\ a_{i_r,j_1} & a_{i_r,j_2} & \cdots & a_{i_r,j_c} \end{bmatrix}.$$

For the special case of consecutive rows and columns we use the notation

$$A(r_1 : r_2, c_1 : c_2) := \begin{bmatrix} a_{r_1,c_1} & a_{r_1,c_1+1} & \cdots & a_{r_1,c_2} \\ a_{r_1+1,c_1} & a_{r_1+1,c_1+1} & \cdots & a_{r_1+1,c_2} \\ \vdots & \vdots & & \vdots \\ a_{r_2,c_1} & a_{r_2,c_1+1} & \cdots & a_{r_2,c_2} \end{bmatrix}.$$

Part I

Some Linear Systems with a Special Structure

Chapter 2

Examples of Linear Systems

Many problems in computational science involves linear systems where the coefficient matrix has a special structure. In this chapter we present two problems that lead to a linear system with a tridiagonal coefficient matrix. Such linear systems can be solved by a version of Gaussian elimination adapted to the special structure. We first consider block multiplication and some useful facts about triangular matrices.

2.1 Block Multiplication and Triangular Matrices

2.1.1 Block Multiplication

A rectangular matrix $\boldsymbol{A}$ can be partitioned into submatrices by drawing horizontal lines between selected rows and vertical lines between selected columns. For example, the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

can be partitioned as

$$(i)\ \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} = \left[\begin{array}{cc|c} 1 & 2 & 3 \\ 4 & 5 & 6 \\ \hline 7 & 8 & 9 \end{array}\right], \quad (ii)\ \begin{bmatrix} \boldsymbol{a}_{:1}, \boldsymbol{a}_{:2}, \boldsymbol{a}_{:3} \end{bmatrix} = \left[\begin{array}{c|c|c} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array}\right],$$

$$(iii)\ \begin{bmatrix} \boldsymbol{a}_{1:}^T \\ \boldsymbol{a}_{2:}^T \\ \boldsymbol{a}_{3:}^T \end{bmatrix} = \left[\begin{array}{ccc} 1 & 2 & 3 \\ \hline 4 & 5 & 6 \\ \hline 7 & 8 & 9 \end{array}\right], \quad (iv)\ \begin{bmatrix} \boldsymbol{A}_{11}, \boldsymbol{A}_{12} \end{bmatrix} = \left[\begin{array}{c|cc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array}\right].$$

In (i) the matrix $\boldsymbol{A}$ is divided into four submatrices

$$\boldsymbol{A}_{11} = \begin{bmatrix} 1 \end{bmatrix}, \ \boldsymbol{A}_{12} = \begin{bmatrix} 2, 3 \end{bmatrix}, \ \boldsymbol{A}_{21} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \text{ and } \boldsymbol{A}_{22} = \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix},$$

while in (ii) and (iii) $\boldsymbol{A}$ has been partitioned into columns and rows, respectively. The submatrices in a partition are often referred to as **blocks** and a partitioned matrix is sometimes called a **block matrix**.

In the following we assume that $\boldsymbol{A} \in \mathbb{C}^{m,p}$ and $\boldsymbol{B} \in \mathbb{C}^{p,n}$. Here are some rules and observations for block multiplication.

1. If $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{b}_{:1}, \ldots, \boldsymbol{b}_{:n}\end{bmatrix}$ is partitioned into columns then the partition of the product $\boldsymbol{AB}$ into columns is
$$\boldsymbol{AB} = \begin{bmatrix}\boldsymbol{Ab}_{:1}, \boldsymbol{Ab}_{:2}, \ldots, \boldsymbol{Ab}_{:n}\end{bmatrix}.$$
In particular, if $\boldsymbol{I}$ is the identity matrix of order p then
$$\boldsymbol{A} = \boldsymbol{AI} = \boldsymbol{A}\begin{bmatrix}\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p\end{bmatrix} = \begin{bmatrix}\boldsymbol{Ae}_1, \boldsymbol{Ae}_2, \ldots, \boldsymbol{Ae}_p\end{bmatrix}$$
and we see that column j of $\boldsymbol{A}$ can be written $\boldsymbol{Ae}_j$ for $j = 1, \ldots, p$.
2. Similarly, if $\boldsymbol{A}$ is partitioned into rows then
$$\boldsymbol{AB} = \begin{bmatrix} a_{1:}^T \\ a_{2:}^T \\ \vdots \\ a_{m:}^T \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} a_{1:}^T\boldsymbol{B} \\ a_{2:}^T\boldsymbol{B} \\ \vdots \\ a_{m:}^T\boldsymbol{B} \end{bmatrix}$$
and taking $\boldsymbol{A} = \boldsymbol{I}$ it follows that row i of $\boldsymbol{B}$ can be written $\boldsymbol{e}_i^T\boldsymbol{B}$ for $i = 1, \ldots, p$.
3. It is often useful to write the matrix-vector product $\boldsymbol{Ax}$ as a linear combination of the columns of $\boldsymbol{A}$
$$\boldsymbol{Ax} = x_1\boldsymbol{a}_{:1} + x_2\boldsymbol{a}_{:2} + \cdots + x_p\boldsymbol{a}_{:p}.$$
4. If $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{B}_1, \boldsymbol{B}_2\end{bmatrix}$, where $\boldsymbol{B}_1 \in \mathbb{C}^{p,r}$ and $\boldsymbol{B}_2 \in \mathbb{C}^{p,n-r}$ then
$$\boldsymbol{A}\begin{bmatrix}\boldsymbol{B}_1, \boldsymbol{B}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{AB}_1, \boldsymbol{AB}_2\end{bmatrix}.$$
This follows from Rule 1. by an appropriate grouping of columns.
5. If $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_1 \\ \boldsymbol{A}_2\end{bmatrix}$, where $\boldsymbol{A}_1 \in \mathbb{C}^{k,p}$ and $\boldsymbol{A}_2 \in \mathbb{C}^{m-k,p}$ then
$$\begin{bmatrix}\boldsymbol{A}_1 \\ \boldsymbol{A}_2\end{bmatrix}\boldsymbol{B} = \begin{bmatrix}\boldsymbol{A}_1\boldsymbol{B} \\ \boldsymbol{A}_2\boldsymbol{B}\end{bmatrix}.$$
This follows from Rule 2. by a grouping of rows.
6. If $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_1, \boldsymbol{A}_2\end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{B}_1 \\ \boldsymbol{B}_2\end{bmatrix}$, where $\boldsymbol{A}_1 \in \mathbb{C}^{m,s}$, $\boldsymbol{A}_2 \in \mathbb{C}^{m,p-s}$, $\boldsymbol{B}_1 \in \mathbb{C}^{s,n}$ and $\boldsymbol{B}_2 \in \mathbb{C}^{p-s,n}$ then
$$\begin{bmatrix}\boldsymbol{A}_1, \boldsymbol{A}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{B}_1 \\ \boldsymbol{B}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2\end{bmatrix}.$$
Indeed, $(\boldsymbol{AB})_{ij} = \sum_{j=1}^{p} a_{ik}b_{kj} = \sum_{j=1}^{s} a_{ik}b_{kj} + \sum_{j=s+1}^{p} a_{ik}b_{kj} = (\boldsymbol{A}_1\boldsymbol{B}_1)_{ij} + (\boldsymbol{A}_2\boldsymbol{B}_2)_{ij} = (\boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2)_{ij}$.

7. If $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$ then

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix},$$

provided the vertical partition in $\boldsymbol{A}$ matches the horizontal one in $\boldsymbol{B}$, i.e. the number of columns in $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{21}$ equals the number of rows in $\boldsymbol{B}_{11}$ and $\boldsymbol{B}_{12}$ and similar for the other blocks. To show this we use Rule 4. to obtain

$$\boldsymbol{AB} = \left[\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} \\ \boldsymbol{B}_{21} \end{bmatrix}, \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{12} \\ \boldsymbol{B}_{22} \end{bmatrix} \right].$$

We complete the proof using Rules 5. and 6.

8. For the general case see Section B.1.

Exercise 2.1 For any matrix $\boldsymbol{A}$ show that $a_{ij} = \boldsymbol{e}_i^T \boldsymbol{A} \boldsymbol{e}_j$ for all i, j.

Exercise 2.2 Let $\boldsymbol{B} = \boldsymbol{A}^T \boldsymbol{A}$. Explain why this product is defined for any matrix. Show that $b_{ij} = \langle \boldsymbol{a}_{:i}, \boldsymbol{a}_{:j} \rangle := \boldsymbol{a}_{:i}^T \boldsymbol{a}_{:j}$ for all i, j.

Exercise 2.3 For $\boldsymbol{A} \in \mathbb{R}^{m,n}$ and $\boldsymbol{B} \in \mathbb{R}^{p,n}$ show that

$$\boldsymbol{AB}^T = \boldsymbol{a}_{:1}\boldsymbol{b}_{:1}^T + \boldsymbol{a}_{:2}\boldsymbol{b}_{:2}^T + \cdots + \boldsymbol{a}_{:n}\boldsymbol{b}_{:n}^T.$$

This is called the **outer product expansion** of the columns of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Exercise 2.4 Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $\boldsymbol{B} \in \mathbb{R}^{m,p}$, and $\boldsymbol{X} \in \mathbb{R}^{n,p}$. Show that

$$\boldsymbol{AX} = \boldsymbol{B} \quad \Longleftrightarrow \quad \boldsymbol{A}\boldsymbol{x}_{:j} = \boldsymbol{b}_{:j}, \ j = 1, \ldots, p.$$

Exercise 2.5 Suppose $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1, \boldsymbol{A}_2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{0} \end{bmatrix}$. When is $\boldsymbol{AB} = \boldsymbol{A}_1\boldsymbol{B}_1$?

Exercise 2.6 Suppose $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n,n}$ are given in block form by

$$\boldsymbol{A} := \begin{bmatrix} \lambda & \boldsymbol{a}^T \\ \boldsymbol{0} & \boldsymbol{A}_1 \end{bmatrix}, \quad \boldsymbol{B} := \begin{bmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{B}_1 \end{bmatrix}, \quad \boldsymbol{C} := \begin{bmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{C}_1 \end{bmatrix},$$

where $\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1 \in \mathbb{R}^{n-1,n-1}$. Show that

$$\boldsymbol{CAB} = \begin{bmatrix} \lambda & \boldsymbol{a}^T\boldsymbol{B}_1 \\ \boldsymbol{0} & \boldsymbol{C}_1\boldsymbol{A}_1\boldsymbol{B}_1 \end{bmatrix}.$$

2.1.2 Triangular matrices

Recall that a matrix $\boldsymbol{R}$ is upper- or right triangular if $r_{ij} = 0$ for $i > j$, and a matrix $\boldsymbol{L}$ is lower- or left triangular if $l_{ij} = 0$ for $i < j$. If $\boldsymbol{R}$ is upper triangular then $\boldsymbol{R}^T$ is lower triangular.

We need some basic facts about triangular matrices and we start with

Lemma 2.7 *Suppose*

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ 0 & \boldsymbol{A}_{22} \end{bmatrix}$$

where $\boldsymbol{A}, \boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are square matrices. Then $\boldsymbol{A}$ is nonsingular if and only if both $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are nonsingular. In that case

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ 0 & \boldsymbol{A}_{22}^{-1} \end{bmatrix} \tag{2.1}$$

Proof. If $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{12}$ are nonsingular then

$$\begin{bmatrix} \boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ 0 & \boldsymbol{A}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ 0 & \boldsymbol{A}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix} = \boldsymbol{I}$$

and $\boldsymbol{A}$ is nonsingular with the indicated inverse. Conversely, let $\boldsymbol{B}$ be the inverse of the nonsingular matrix $\boldsymbol{A}$. We partition $\boldsymbol{B}$ conformally with $\boldsymbol{A}$ and have

$$\boldsymbol{B}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ 0 & \boldsymbol{A}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix} = \boldsymbol{I}$$

Using block-multiplication we find

$$\boldsymbol{B}_{11}\boldsymbol{A}_{11} = \boldsymbol{I}, \quad \boldsymbol{B}_{21}\boldsymbol{A}_{11} = \boldsymbol{0}, \quad \boldsymbol{B}_{21}\boldsymbol{A}_{12} + \boldsymbol{B}_{22}\boldsymbol{A}_{22} = \boldsymbol{I}.$$

The first equation implies that $\boldsymbol{A}_{11}$ is nonsingular, this in turn implies that $\boldsymbol{B}_{21} = \boldsymbol{0}$ in the second equation, and then the third equation simplifies to $\boldsymbol{B}_{22}\boldsymbol{A}_{22} = \boldsymbol{I}$. We conclude that also $\boldsymbol{A}_{22}$ is nonsingular. □

Consider now a triangular matrix.

Lemma 2.8 *An upper (lower) triangular matrix $\boldsymbol{A} = [a_{ij}] \in \mathbb{C}^{n,n}$ is nonsingular if and only if the diagonal elements a_{ii}, $i = 1, \ldots, n$ are nonzero. In that case the inverse is upper (lower) triangular with diagonal elements a_{ii}^{-1}, $i = 1, \ldots, n$.*

Proof. We use induction on n. The result holds for $n = 1$. The 1-by-1 matrix $\boldsymbol{A} = [a_{11}]$ is nonsingular if and only if $a_{11} \neq 0$ and in that case $\boldsymbol{A}^{-1} = [a_{11}^{-1}]$. Suppose the result holds for $n = k$ and let $\boldsymbol{A} \in \mathbb{C}^{k+1,k+1}$ be upper triangular. We partition $\boldsymbol{A}$ in the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_k & \boldsymbol{a}_k \\ 0 & a_{k+1,k+1} \end{bmatrix}$$

and note that $\boldsymbol{A}_k \in \mathbb{C}^{k,k}$ is upper triangular. By Lemma 1.1 $\boldsymbol{A}$ is nonsingular if and only if $\boldsymbol{A}_k$ and $(a_{k+1,k+1})$ are nonsingular and in that case

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_k^{-1} & -\boldsymbol{A}_k^{-1}\boldsymbol{a}_k a_{k+1,k+1}^{-1} \\ 0 & a_{k+1,k+1}^{-1} \end{bmatrix}.$$

By the induction hypothesis $\boldsymbol{A}_k$ is nonsingular if and only if the diagonal elements $a_{11}, \ldots, a_{kk}$ of $\boldsymbol{A}_k$ are nonzero and in that case $\boldsymbol{A}_k^{-1}$ is upper triangular with diagonal elements a_{ii}^{-1}, $i = 1, \ldots, k$. The result for $\boldsymbol{A}$ follows. □

Lemma 2.9 *The product $\boldsymbol{C} = \boldsymbol{AB} = (c_{ij})$ of two upper (lower) triangular matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ is upper (lower) triangular with diagonal elements $c_{ii} = a_{ii}b_{ii}$ for all i.*

Proof. Exercise. □

A matrix is **unit triangular** if it is triangular with 1's on the diagonal.

Lemma 2.10 *For a unit upper (lower) triangular matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$:*

1. *$\boldsymbol{A}$ is nonsingular and the inverse is unit upper(lower) triangular.*
2. *The product of two unit upper (lower) triangular matrices is unit upper (lower) triangular.*

Proof. 1. follows from Lemma 2.8, while Lemma 2.9 implies 2. □

2.2 The Second Derivative Matrix

Consider the simple two point boundary value problem

$$-u''(x) = f(x), \quad x \in [0,1], \quad u(0) = 0, \; u(1) = 0, \tag{2.2}$$

where f is a given continuous function on $[0,1]$. This problem is also known as the **one-dimensional (1D) Poisson problem** . In principle it is easy to solve (2.2) exactly. We just integrate f twice and determine the two integration constants so that the homogeneous boundary conditions $u(0) = u(1) = 0$ are satisfied. For example, if $f(x) = 1$ then $u(x) = x(x-1)/2$ is the solution. However, many functions f cannot be integrated exactly, and in such cases a numerical method can be used.

(2.2) can be solved approximately using the **finite difference method**. For this we choose a positive integer m, define the discretization parameter $h := 1/(m+1)$, and replace the interval $[0,1]$ by grid points $x_j := jh$ for $j = 0, 1, \ldots, m+1$. We then use the following finite difference approximation of the second derivative:

$$u''(x) \approx \frac{u(x-h) - 2u(x) + u(x+h)}{h^2}.$$

We obtain approximations v_j to the exact solution $u(x_j)$ for $j = 1, \ldots, m$ by replacing the differential equation by the difference equation

$$\frac{-v_{j-1} + 2v_j - v_{j+1}}{h^2} = f(jh), \quad j = 1, \ldots, m, \quad v_0 = v_{m+1} = 0.$$

Moving the h^2 factor to the right hand side this can be written as an $m \times m$ linear system

$$\boldsymbol{T}\boldsymbol{v} = \begin{bmatrix} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & & & \\ 0 & \ddots & \ddots & \ddots & & \\ & & & & & 0 \\ & & & -1 & 2 & -1 \\ & & & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{m-1} \\ v_m \end{bmatrix} = h^2 \begin{bmatrix} f(h) \\ f(2h) \\ \vdots \\ f((m-1)h) \\ f(mh) \end{bmatrix} =: \boldsymbol{b}. \tag{2.3}$$

The matrix $\boldsymbol{T}$ is called the **second derivative matrix** and will occur frequently in these notes.

2.3 LU Factorization of a Tridiagonal System

Consider the linear system $\boldsymbol{T}\boldsymbol{v} = \boldsymbol{b}$ given by (2.3). We show in Theorem 2.14 below that this system has a unique solution and that the algorithm we now describe is well defined.

The matrix $\boldsymbol{T}$ is an instance of a tridiagonal matrix $\boldsymbol{A} = \operatorname{tridiag}(a_i, d_i, c_i) \in \mathbb{C}^{n,n}$. An economical way to solve a tridiagonal system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is to construct, if possible, triangular matrices $\boldsymbol{L}$ and $\boldsymbol{R}$ such that the product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ has the form

$$\begin{bmatrix} d_1 & c_1 & & & \\ a_2 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & d_{n-1} & c_{n-1} \\ & & & a_n & d_n \end{bmatrix} = \begin{bmatrix} 1 & & & \\ l_2 & 1 & & \\ & \ddots & \ddots & \\ & & l_n & 1 \end{bmatrix} \begin{bmatrix} r_1 & c_1 & & \\ & \ddots & \ddots & \\ & & r_{n-1} & c_{n-1} \\ & & & r_n \end{bmatrix}. \tag{2.4}$$

Once $\boldsymbol{L}$ and $\boldsymbol{R}$ are determined we can find $\boldsymbol{x}$ by solving two simpler systems $\boldsymbol{L}\boldsymbol{y} = \boldsymbol{b}$ and $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$.

To find $\boldsymbol{L}$ and $\boldsymbol{R}$ we note that $\boldsymbol{L}$ and $\boldsymbol{R}$ are bidiagonal, $\boldsymbol{L}$ has ones on the diagonal, and that we have the same c_i elements on the super-diagonals of $\boldsymbol{A}$ and $\boldsymbol{R}$. By equating elements in (2.4) we find

$$d_1 = r_1, \quad a_k = l_k r_{k-1}, \quad d_k = l_k c_{k-1} + r_k, \quad k = 2, 3, \ldots, n.$$

Solving for l_k and r_k leads to

$$r_1 = d_1, \quad l_k = \frac{a_k}{r_{k-1}}, \quad r_k = d_k - l_k c_{k-1}, \quad k = 2, 3, \ldots, n. \tag{2.5}$$

We can then solve $\boldsymbol{Ly} = \boldsymbol{b}$ and $\boldsymbol{Rx} = \boldsymbol{y}$

$$\begin{array}{llll} y_1 = b_1, & y_k = b_k - l_k y_{k-1}, & k = 2, 3, \ldots, n, \\ x_n = y_n / r_n, & x_k = (y_k - c_k x_{k+1})/r_k, & k = n-1, \ldots, 2, 1. \end{array} \tag{2.6}$$

We formulate this as two algorithms.

Algorithm 2.11 (trifactor) Vectors $\boldsymbol{l}, \boldsymbol{r} \in \mathbb{C}^n$ are computed from $\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{d} \in \mathbb{C}^n$. This implements the LU factorization of a tridiagonal matrix. The first (dummy) component in $\boldsymbol{a}$ and last component of $\boldsymbol{c}$ are not used.

```
function [l,r]=trifactor(a,d,c)
r=d; l=d;
for k=2:length(d)
    l(k)=a(k)/r(k-1);
    r(k)=d(k)-l(k)*c(k-1);
end
```

Algorithm 2.12 (trisolve) The solution x of the tridiagonal system $\boldsymbol{LRx} = \boldsymbol{b}$ is found from (2.6). Here $\boldsymbol{l}, \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{b} \in \mathbb{C}^n$. The vectors $\boldsymbol{l}, \boldsymbol{r}$ are typically output from `trifactor`.

```
function x=trisolve(l,r,c,b)
x=b; n=length(b);
for k=2:n
    x(k)=b(k)-l(k)*x(k-1);
end
x(n)=x(n)/r(n);
for k=n-1:-1:1
    x(k)=(x(k)-c(k)*x(k+1))/r(k);
end
```

The number of floating point operations (flops) to compute the LU factorization of a tridiagonal matrix using Algorithm 2.11 is only $3n-3$, while the number of flops for Algorithm 2.12 is $5n-4$. This means that the number of flops (the complexity) to solve a tridiagonal system is $O(n)$, or more precisely $8n-7$, and this number only grows linearly with n. This should be compared to Gaussian elimination on a full $n \times n$ system which is an $O(n^3)$ process, i. e., it is proportional to n^3.

2.3.1 Diagonal Dominance

We show that Algorithms 2.11, 2.12 are well defined for a class of tridiagonal linear systems. Moreover, these linear systems have a unique solution.

Definition 2.13 *The matrix* $\boldsymbol{A} = [a_{ij}] \in \mathbb{C}^{n,n}$ *is* **diagonally dominant** *if*

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|, \ i = 1, \ldots, n. \tag{2.7}$$

It is **strictly diagonally dominant** *if strict inequality holds for* $i = 1, \ldots, n$.

Recall that a square matrix $\boldsymbol{A}$ is singular if $\boldsymbol{Ax} = \boldsymbol{0}$ for a nonzero vector $\boldsymbol{x}$. A diagonally dominant matrix can be singular. For example the matrices $\boldsymbol{A}_1 = \left[\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right]$ and $\boldsymbol{A}_2 = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right]$ are both diagonally dominant and singular. Indeed, $\boldsymbol{A}_2$ has a zero row, and for $\boldsymbol{A}_1$ column two is the sum of columns one and three. In the literature diagonal dominance is therefore most often defined by including some additional condition(s). Here we prove the following result.

Theorem 2.14 *Suppose* $\boldsymbol{A} = \mathit{tridiag}(a_i, d_i, c_i) \in \mathbb{C}^{n,n}$ *is tridiagonal and diagonally dominant. If* $|d_1| > |c_1|$ *and* $a_i \neq 0$ *for* $i = 2, \ldots, n-1$*, then* $\boldsymbol{A}$ *has a unique LU factorization* (2.4)*. If in addition* $d_n \neq 0$*, then* $\boldsymbol{A}$ *is nonsingular.*

Proof. The matrix $\boldsymbol{A}$ has an LU factorization if the r_k's in (2.5) are nonzero for $k = 1, \ldots, n-1$. We show by induction on k that $|r_k| > |c_k|$ for $k = 1, \ldots, n-1$. Clearly $|r_1| = |d_1| > |c_1|$. Suppose $|c_{k-1}|/|r_{k-1}| < 1$ for some $2 \leq k \leq n-1$. By (2.5)

$$|r_k| = |d_k - l_k c_{k-1}| = |d_k - \frac{a_k c_{k-1}}{r_{k-1}}| \geq |d_k| - \frac{|a_k||c_{k-1}|}{|r_{k-1}|}. \tag{2.8}$$

Since $a_k \neq 0$ and by diagonal dominance $|r_k| > |d_k| - |a_k| \geq |c_k|$. Thus $|r_k| > |c_k| \geq 0$, for $k = 1, \ldots, n-1$ and an LU factorization exists. It is unique since any LU factorization must satisfy (2.5). If $d_n \neq 0$ then by (2.8) $|r_n| > 0$ regarless of wether a_n is zero or nonzero, so both $\boldsymbol{L}$ and $\boldsymbol{R}$ have nonzero diagonal elements. By Lemma 2.8 the product $\boldsymbol{A} = \boldsymbol{LR}$ is nonsingular. □

Consider the system $\boldsymbol{Tv} = \boldsymbol{b}$. The matrix $\boldsymbol{T}$ is diagonally dominant and satisfies the additional conditions in Theorem 2.14. Thus it is nonsingular and we can solve the system in $O(n)$ arithmetic operations using Algorithms 2.11,2.12.

Alternatively, we could solve the system $\boldsymbol{Tv} = \boldsymbol{b}$ by using the inverse $\boldsymbol{T}^{-1}$ of $\boldsymbol{T}$ and simply compute the matrix vector product $\boldsymbol{v} = \boldsymbol{T}^{-1}\boldsymbol{b}$. However this is not a good idea. In fact, all elements in T^{-1} are nonzero and the calculation of $\boldsymbol{T}^{-1}\boldsymbol{b}$ requires $O(n^2)$ operations. See Exercise 2.16.

Exercise 2.15 *Show that* $\boldsymbol{T} = \boldsymbol{LR}$*, where*

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\frac{1}{2} & 1 & \ddots & & \vdots \\ 0 & -\frac{2}{3} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{m-1}{m} & 1 \end{bmatrix}, \boldsymbol{R} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ 0 & \frac{3}{2} & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \frac{m}{m-1} & -1 \\ 0 & \cdots & \cdots & 0 & \frac{m+1}{m} \end{bmatrix}. \tag{2.9}$$

Thus, $\boldsymbol{T} = \boldsymbol{LR}$ *is the LU factorization of* $\boldsymbol{T}$.

Exercise 2.16 *Let* $\boldsymbol{S} \in \mathbb{R}^{m,m}$ *have elements* s_{ij} *given by*

$$s_{i,j} = s_{j,i} = \frac{1}{m+1} j\big(m+1-i\big), \quad 1 \le j \le i \le m. \tag{2.10}$$

Show that $\boldsymbol{ST} = \boldsymbol{I}$ *and conclude that* $\boldsymbol{T}^{-1} = \boldsymbol{S}$.

Exercise 2.17 *We consider a finite difference method for the two point boundary value problem*

$$\begin{aligned} -u''(x) + r(x)u'(x) + q(x)u(x) = f(x), \ \text{for } x \in [a,b], \\ u(a) = g_0, \quad u(b) = g_1. \end{aligned} \tag{2.11}$$

We assume that the given functions f, q *and* r *are continuous on* $[a,b]$ *and that* $q(x) \ge 0$ *for* $x \in [a,b]$. *It can then be shown that* (2.11) *has a unique solution* u.

To solve (2.11) *numerically we choose* $m \in \mathbb{N}$, $h = (b-a)/(m+1)$, $x_j = a + jh$ *for* $j = 0, 1, \ldots, m+1$ *and solve the difference equation*

$$\frac{-v_{j-1} + 2v_j - v_{j+1}}{h^2} + r(x_j)\frac{v_{j+1} - v_{j-1}}{2h} + q(x_j)v_j = f(x_j), \quad j = 1, \ldots, m, \tag{2.12}$$

with $v_0 = g_0$ *and* $v_{m+1} = g_1$.

(a) *Show that* (2.12) *leads to a tridiagonal linear system* $\boldsymbol{Av} = \boldsymbol{b}$, *where* $\boldsymbol{A} = \mathit{tridiag}(a_j, d_j.c_j) \in \mathbb{R}^{m,m}$ *has elements*

$$a_j = -1 - \frac{h}{2}r(x_j), \ c_j = -1 + \frac{h}{2}r(x_j), \ d_j = 2 + h^2 q(x_j),$$

and

$$b_j = \begin{cases} h^2 f(x_1) - a_1 g_0, & \text{if } j = 1, \\ h^2 f(x_j), & \text{if } 2 \le j \le m-1, \\ h^2 f(x_m) - c_m g_1, & \text{if } j = m. \end{cases}$$

(b) *Show that the linear system satisfies the conditions in Theorem 2.14 if the spacing* h *is so small that* $\frac{h}{2}|r(x)| < 1$ *for all* $x \in [a,b]$.

(c) *Propose a method to find* $v_1, \ldots, v_m$.

Exercise 2.18 (a) *Consider the problem* (2.11) *with* $r = 0$, $f = q = 1$ *and boundary conditions* $u(0) = 1$, $u(1) = 0$. *The exact solution is* $u(x) = 1 - \sinh x / \sinh 1$. *Write a computer program to solve* (2.12) *for* $h = 0.1, 0.05, 0.025, 0.0125$, *and compute the "error"* $\max_{1 \le j \le m} |u(x_j) - v_j|$ *for each* h.

(b) *Make a combined plot of the solution* u *and the computed points* v_j, $j = 0, \ldots, m+1$ *for* $h = 0.1$.

(c) *One can show that the error is proportional to* h^p *for some integer* p. *Estimate* p *based on the error for* $h = 0.1, 0.05, 0.025, 0.0125$.

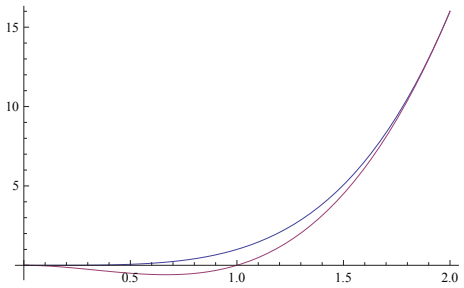


Figure 2.1. *The cubic Hermite interpolation polynomial interpolating* $f(x) = x^4$ *on* $[0, 2]$.

2.4 Cubic Spline Interpolation

We next consider another problem leading to a tridiagonal linear system. Given $n \geq 2$ interpolation sites $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ with $a := x_1 < \cdots < x_n =: b$, real y values $\boldsymbol{y} = [y_1, \ldots, y_n]^T$, and derivative values σ_a, σ_b. We seek a function $g : [a, b] \to \mathbb{R}$ such that

$$g(x_i) = y_i, \text{ for } i = 1, \ldots, n, \quad g'(a) = \sigma_a, \; g'(b) = \sigma_b. \tag{2.13}$$

The derivative conditions are known under various names as first derivative, clamped, or Hermite boundary conditions.

Since there are $n + 2$ interpolation conditions in (2.13) a natural choice for a function g is a polynomial with $n + 2$ coefficients, i. e., a polynomial of degree at most $n + 1$. As shown in many books on numerical methods such a g is uniquely defined and there are good algorithms for computing it. For example, when $n = 2$ the interpolant is known as the **cubic Hermite interpolation polynomial**.

Example 2.19 (Cubic Hermite interpolation) *Let* $f : [0, 2] \to \mathbb{R}$ *be given by* $f(x) = x^4$. *The cubic polynomial* g *given by* $g(x) = 4x^3 - 4x^2$ *satisfies*

$$g(0) = f(0), \quad g(2) = f(2), \quad g'(0) = f'(0), \quad g'(2) = f'(2).$$

We show g *and* f *in Figure 2.1*

The polynomial g of degree $\leq n + 1 = 15$ interpolating the function f given by $f(x) = \arctan(10x) + \pi/2$, $x \in [-1, 1]$ at the points $x_i = -1 + 2(i - 1)/(n - 1)$, $i = 1, \ldots, n$ with $g'(-1) = f'(-1)$, $g'(1) = f'(1)$ is shown in Figure 2.2. The interpolant has large oscillations near the end of the range.

2.4.1 C^2 Cubic Splines

When n is large a polynomial interpolant can have undesirable oscillations (cf. Figure 2.2 and in this section we consider an alternative which often leads to better results.

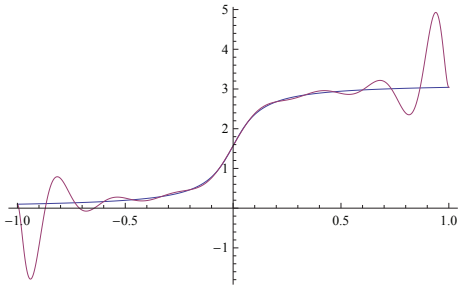


Figure 2.2. *The polynomial of degree 15 interpolating* $f(x) = \arctan(10x) + \pi/2$ *on* $[-1, 1]$. *See text*

Definition 2.20 *Let* $n \geq 3$, *and* $a = x_1 < x_2 < \ldots < x_n = b$. *A function* $g : [a, b] \to \mathbb{R}$ *of the form*

$$g(x) := \begin{cases} p_1(x), & \text{if } a \leq x < x_2, \\ p_2(x), & \text{if } x_2 \leq x < x_3, \\ \vdots & \\ p_{n-2}(x), & \text{if } x_{n-2} \leq x < x_{n-1}, \\ p_{n-1}(x), & \text{if } x_{n-1} \leq x \leq b, \end{cases} \tag{2.14}$$

is called a **cubic spline** *with* **knots** $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ *provided*

(i) *Each* p_i *is a polynomial of degree* ≤ 3.

(ii) $p_{i-1}(x_i) = p_i(x_i), \quad p'_{i-1}(x_i) = p'_i(x_i), \quad p''_{i-1}(x_i) = p''_i(x_i), \quad i = 2, \ldots, n-1.$

We note that.

1. By requirement **(ii)** the polynomials p_i are "glued" together with C^2 continuity, i. e., g, g' and g'' are continuous on $[a, b]$.

2. Each of the $n - 1$ cubic polynomials p_i has 4 coefficients and there are 3 continuity conditions at each of the $n - 2$ interior knots. Thus, a cubic spline appears to have $4(n - 1) - 3(n - 2) = n + 2$ degrees of freedom. This is the number of interpolation conditions in (2.13). We show in Theorem 2.22 below that there is a unique cubic spline of the form (2.14) satisfying (2.13). We call this function the **cubic spline interpolant**. A cubic spline interpolant to the function $f(x) = \arctan(10x) + \pi/2$ us shown in Figure 2.3 for $n = 14$ using the same uniform x_i's as in Figure 2.2. The spline interpolant is quite close to f. In fact it is hard to distinguish the two curves.

Example 2.21 *Show that* g *given by*

$$g(x) := \begin{cases} p_1(x) = -x^2 + 2x^3, & \text{if } 0 \leq x < 1, \\ p_2(x) = -4 + 12x - 13x^2 + 6x^3, & \text{if } 1 \leq x \leq 2, \end{cases} \tag{2.15}$$

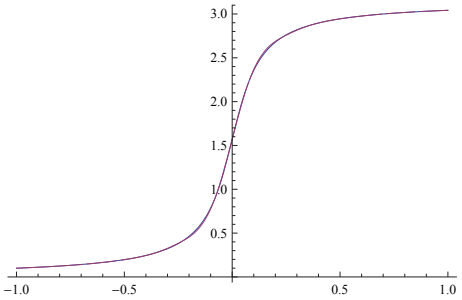


Figure 2.3. *A cubic spline interpolating* $f(x) = \arctan(10x) + \pi/2$ *on* $[-1, 1]$. *See text*

is a cubic spline interpolant to the data

$$\boldsymbol{x} = [0, 1, 2]^T, \quad \boldsymbol{y} = [0, 1, 4]^T, \quad \sigma_a = 0,\ \sigma_b = 32.$$

Discussion: Clearly g *is in the form* (2.14) *with knots* $\boldsymbol{x}$ *and* p_1, p_2 *are cubic polynomials. Since* $p_1(1) = 1 = p_2(1)$, $p_1'(1) = 4 = p_2'(1)$, $p_1''(1) = 32 = p_2''(1)$ *we see that* g *is a cubic spline. Moreover,* $g(x_1) = p_1(0) = 0 = y_1$, $g(x_2) = p_2(1) = 1 = y_2$, $g(x_3) = p_2(2) = 16 = y_3$, $g'(0) = p_1'(0) = 0 = \sigma_a$, $g'(2) = p_2'(2) = 32 = \sigma_b$, *and so* g *interpolates the data. The data is sampled from the function given by the rule* $f(x) = x^4$. *A plot of* f *and* g *is shown in Figure 2.4. It is hard to distinguish the two curves.*

2.4.2 Finding the Interpolant

For reasons of numerical accuracy it is convenient to use a representation called the **shifted power form** for each p_i.

$$p_i(x) = c_{1i} + c_{2i}(x - x_i) + c_{3i}(x - x_i)^2 + c_{4i}(x - x_i)^3. \tag{2.16}$$

A cubic spline is completely determined by the shift vector $\boldsymbol{x}_s \in \mathbb{R}^{n-1}$ and the coefficient matrix $\boldsymbol{C}$

$$\boldsymbol{x}_s := [x_1, \ldots, x_k]^T, \quad \boldsymbol{C} := \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1,k} \\ c_{21} & c_{22} & \cdots & c_{2,k} \\ c_{31} & c_{32} & \cdots & c_{3,k} \\ c_{41} & c_{42} & \cdots & c_{4,k} \end{bmatrix} \in \mathbb{R}^{4,k}, \quad k = n - 1. \tag{2.17}$$

We call (2.17) the **pp representation** of g (with respect to $\boldsymbol{x}_s$).

As an example, consider (2.15). With $x_2 = 1$ we obtain $p_2(x) = -4 + 12x - 13x^2 + 6x^3 = 1 + 4(x-1) + 5(x-1)^2 + 6(x-1)^3$ and since $x_1 = 0$ the shifted power form takes the form

$$g(x) := \begin{cases} p_1(x) = -x^2 + 2x^3, & \text{if } 0 \le x < 1, \\ p_2(x) = 1 + 4(x-1) + 5(x-1)^2 + 6(x-1)^3. & \text{if } 1 \le x \le 2. \end{cases}$$

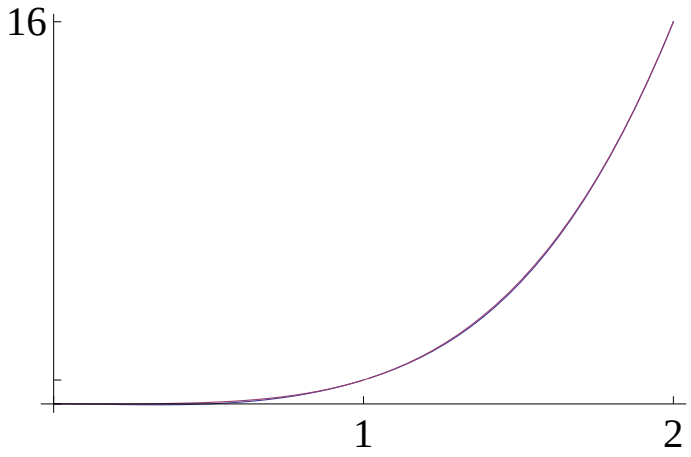


Figure 2.4. *A two piece cubic spline interpolant to* $f(x) = x^4$.

The pp representation is

$$\boldsymbol{x}_s = [0,1]^T, \quad \boldsymbol{C} = \begin{bmatrix} 0 & 1 \\ 0 & 4 \\ -1 & 5 \\ 2 & 6 \end{bmatrix}. \tag{2.18}$$

The following Theorem shows how to determine the pp form of the cubic spline interpolant.

Theorem 2.22 *Given* $a = x_1 < \ldots < x_n = b$, $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, *and* σ_a, σ_b. *Suppose the pp form* $\boldsymbol{x}_s$, $\boldsymbol{C} = [c_{ji}]$ *of a cubic spline* g *is given by* $\boldsymbol{x}_s^T = [x_1, \ldots, x_{n-1}]$ *and*

$$\begin{aligned} c_{1i} &:= y_i, \quad c_{2i} := s_i, \\ c_{3i} &:= (3\delta_i - 2s_i - s_{i+1})/h_i, \\ c_{4i} &:= (-2\delta_i + s_i + s_{i+1})/h_i^2, \end{aligned} \tag{2.19}$$

where $\boldsymbol{s} = [s_1, \ldots, s_n]^T$ *is the solution of the linear system*

$$\boldsymbol{N}\boldsymbol{s} = \begin{bmatrix} 1 & 0 & & & \\ \lambda_2 & 4 & \mu_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-1} & 4 & \mu_{n-1} \\ & & & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n-1} \\ s_n \end{bmatrix} = \begin{bmatrix} \sigma_a \\ \beta_2 \\ \vdots \\ \beta_{n-1} \\ \sigma_b \end{bmatrix} =: \boldsymbol{b}, \tag{2.20}$$

and where

$$\begin{aligned} h_i &= x_{i+1} - x_i, \quad \delta_i = \frac{y_{i+1} - y_i}{h_i}, \\ \lambda_i &= \frac{2h_i}{h_{i-1} + h_i}, \quad \mu_i = \frac{2h_{i-1}}{h_{i-1} + h_i}, \quad \beta_i = 3(\lambda_i \delta_{i-1} + \mu_i \delta_i), \end{aligned} \tag{2.21}$$

Then g is a cubic spline interpolant and $s_i := g'(x_i)$ for $i = 1, \ldots, n$. Moreover, (2.20) has a unique solution $\boldsymbol{s} = [s_1, \ldots, s_n]^T$, so the cubic spline interpolant is unique.

Proof. Taking derivatives in (2.16) give

$$\begin{aligned} p_i(x) &= c_{1i} + c_{2i}(x - x_i) + c_{3i}(x - x_i)^2 + c_{4i}(x - x_i)^3, \\ p_i'(x) &= c_{2i} + 2c_{3i}(x - x_i) + 3c_{4i}(x - x_i)^2, \\ p_i''(x) &= 2c_{3i} + 6c_{4i}(x - x_i). \end{aligned} \tag{2.22}$$

Suppose p_i is given by (2.16) where the $c_{j,i}$ are given by (2.19) for $i = 1, \ldots, n-1$, $s_1 = \sigma_a$ and $s_n = \sigma_b$, and $s_2, \ldots, s_{n-1}$ are at this point undetermined. Then for $i = 1, \ldots, n-1$

$$\begin{aligned} p_i(x_i) &= y_i, \quad p_i(x_{i+1}) = y_{i+1}, \\ p_i'(x_i) &= s_i, \quad p_i'(x_{i+1}) = s_{i+1}. \end{aligned} \tag{2.23}$$

Thus $g(x_i) = y_i$ and $g'(x_i) = s_i$ for $i = 1, \ldots, n$ and it follows that (2.13) holds. Moreover, by (2.23)

$$p_{i-1}(x_i) = p_i(x_i), \quad p_{i-1}'(x_i) = p_i'(x_i), \quad i = 2, \ldots, n-1.$$

Thus, g is a cubic spline interpolant if and only if

$$p_{i-1}''(x_i) = p_i''(x_i), \quad i = 2, \ldots, n-1. \tag{2.24}$$

By (2.22)

$$p_{i-1}''(x_i) = 2c_{3,i-1} + 6c_{4,i-1}h_{i-1}, \quad p_i''(x_i) = 2c_{3,i}.$$

Thus, (2.24) holds if and only if for $i = 2, \ldots, n-1$

$$\begin{aligned} 0 &= h_{i-1}h_i\big(\frac{1}{2}p_{i-1}''(x_i) - \frac{1}{2}p_i''(x_i)\big) \\ &= h_{i-1}h_i\big(c_{3,i-1} + 3c_{4,i-1}h_{i-1} - c_{3,i}\big) \\ &= h_i(3\delta_{i-1} - 2s_{i-1} - s_i) + 3h_i(-2\delta_{i-1} + s_{i-1} + s_i) - h_{i-1}(3\delta_i - 2s_i - s_{i+1}) \\ &= h_i s_{i-1} + 2(h_{i-1} + h_i)s_i + h_{i-1}s_{i+1} - 3(h_i\delta_{i-1} + h_{i-1}\delta_i). \end{aligned}$$

Multiplying the last expression by $2/(h_{i-1} + h_i)$ leads to the equations

$$\lambda_i s_{i-1} + 4s_i + \mu_i s_{i+1} = \beta_i, \quad i = 2, \ldots, n-1, \tag{2.25}$$

and (2.20) follows. The matrix $\boldsymbol{N}$ is nonsingular since $\lambda_i + \mu_i = 2 < 4$ and $\lambda_i \neq 0$, so that the conditions in Theorem 2.14 hold. For uniqueness suppose g_1 and g_2 are two cubic splines interpolating the same data (2.13). Then $g := g_1 - g_2$ is a cubic spline interpolating zero data. The unique solution of (2.20) is then $\boldsymbol{s} = \boldsymbol{0}$, and since (2.19) must hold for any cubic spline interpolant it easily follows that $\boldsymbol{C} = \boldsymbol{0}$. Thus $g = 0$ and $g_1 = g_2$. ∎

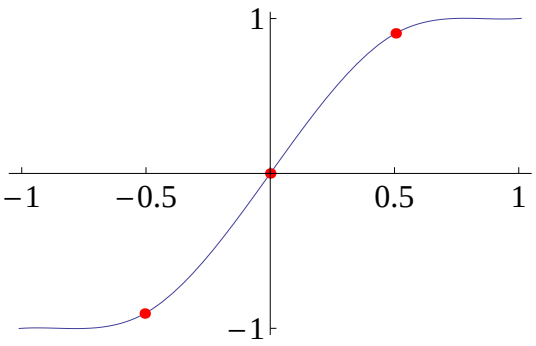


Figure 2.5. *Cubic spline interpolation to the data in Example 2.25. The breakpoints* (x_i, y_i), $i = 2, 3, 4$ *are marked with dots on the curve.*

2.4.3 Algorithms

To find the cubic spline interpolant we first solve (2.20) for $\boldsymbol{s}$ using Algorithms 2.11, 2.12. Then we determine the pp representation from (2.19).

To plot a cubic spline g we need to compute y values $q_j = g(r_j)$ at a number of x values $\boldsymbol{r} = [r_1, \ldots, r_m] \in \mathbb{R}^m$ for some reasonably large integer m. To determine $g(r_j)$ for some j we need to find an integer i_j so that $g(r_j) = p_{i_j}(r_j)$. The following Matlab function determines $\boldsymbol{i} = [i_1, \ldots, i_m]$. It uses the built in Matlab functions `length, min, sort, find`.

Algorithm 2.23 (findsubintervals) Given shifts $\boldsymbol{x}_s = [x_1, \ldots, x_k]$ and a real number r, an integer i is computed so that $i = 1$ if $r < x(2)$, $i = k$ if $r \geq x_k$, and $x_i \leq r < x_{i+1}$ otherwise. If $\boldsymbol{r}$ is a vector then a vector $\boldsymbol{i}$ is computed, such that the jth component of $\boldsymbol{i}$ gives the location of the jth component of $\boldsymbol{r}$.

```
function i=findsubintervals(xs,r)
    k=length(xs); m=length(r);
    xs(1)=min(r)-1;
    [sorted,j] = sort([xs(:)' r(:)']);
    i = find(j>k)-(1:m);
```

Here is the algorithm that was used to compute points for the plot in Figure 2.4. It uses Algorithm 2.23.

Algorithm 2.24 (cubppeval) Given a pp representation $(\boldsymbol{x}_s, \boldsymbol{C})$ of a cubic spline g together with x values $\boldsymbol{r} \in \mathbb{R}^m$. The vector $\boldsymbol{q} = g(\boldsymbol{r})$ is computed.

```
function q=cubppeval(xs,C,r)
i=findsubintervals(xs,r); q=r;
for j=1:length(r)
    k=i(j); t=r(j)-xs(k);
    q(j)=[1 t t^2 t^3]*C(:,k);
end
```

Example 2.25 *Consider the data* $\boldsymbol{x} = [-1, -1/2, 0, 1/2, 1]^T$, $\boldsymbol{y} = [-1, -0.9, 0, 0.9, 1]^T$ *and* $\sigma_a = \sigma_b = 1/10$. *By* (2.20)

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 1 \\ 60 \\ 108 \\ 60 \\ 1 \end{bmatrix}.$$

We find

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1/4 & 1 & 0 & 0 \\ 0 & 0 & 4/15 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 & 0 \\ 0 & 0 & 15/4 & 1 & 0 \\ 0 & 0 & 0 & 56/15 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$\boldsymbol{y} = [1, 59, 373/4, 527/15, 1]^T/10$, *and solution* $\boldsymbol{s} = [7, 64, 157, 64, 7]^T/70$. *A straightforward calculation gives the coefficients in the pp representation*

$$\boldsymbol{C} = \begin{bmatrix} -70 & -63 & 0 & 63 \\ 7 & 64 & 157 & 64 \\ -72 & 186 & 0 & -186 \\ 172 & -124 & -124 & 172 \end{bmatrix} /70.$$

The cubic spline interpolant is shown in Figure 2.5. Here Algorithm 2.24 was used with 200 uniform plot points.

The name spline is inherited from the "physical uncle", i. e., an elastic ruler that is used to draw smooth curves. Heavy weights, called **ducks**, are used to force the physical spline to pass through, or near given locations. (Cf. Figure 2.6).

Exercise 2.26 *In many cases the knots are uniformly spaced, i. e.,* $h_i = h$ *for all* i. *Show that* (2.20) *takes the form*

$$\begin{bmatrix} 1 & 0 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n-1} \\ s_n \end{bmatrix} = \begin{bmatrix} \sigma_a \\ 3(y_3 - y_1)/h \\ \vdots \\ 3(y_n - y_{n-2})/h \\ \sigma_b \end{bmatrix}. \tag{2.26}$$

Figure 2.6. *A physical spline with ducks.*

Exercise 2.27 *Show that*

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + \frac{h^2}{6} f^{(3)}(\eta), \quad x-h < \eta < x+h.$$

This is known as the central difference approximation to the first derivative.

Exercise 2.28 *Derive the pp representation of g in Example 2.25.*

Exercise 2.29 (Give me a Moment) *In this problem we determine the cubic spline interpolant with 1. derivative boundary conditions using the unknown second derivatives (sometimes called moments) $m_i := g''(x_i)$, $i = 1, \ldots, n$ as parameters in the shifted power representation* (2.16) *of g. Show the following theorem.*

Theorem 2.30 *Given $a = x_1 < \ldots < x_n = b$, $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, and σ_a, σ_b. Suppose the pp form $\boldsymbol{x}_s$, $\boldsymbol{C} = [c_{ji}]$ of a cubic spline g is given by $\boldsymbol{x}_s^T = [x_1, \ldots, x_{n-1}]$ and*

$$\begin{aligned} c_{1i} &:= y_i, \quad c_{2i} := \delta_i - \frac{m_i}{3} h_i - \frac{m_{i+1}}{6} h_i, \\ c_{3i} &:= m_i/2, \quad c_{4i} := \frac{m_{i+1} - m_i}{6h_i}, \\ h_i &:= x_{i+1} - x_i, \quad \delta_i := (y_{i+1} - y_i)/h_i, \end{aligned} \tag{2.27}$$

where $\boldsymbol{m} = [m_1, \ldots, m_n]^T$ *is the solution of the linear system*

$$\boldsymbol{N}_{21}\boldsymbol{m} = \begin{bmatrix} 2 & 1 & & & \\ \mu_2 & 4 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 4 & \lambda_{n-1} \\ & & & 1 & 2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-1} \\ m_n \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n-1} \\ \gamma_n \end{bmatrix} =: \boldsymbol{b}_{21}, \tag{2.28}$$

and where

$$\begin{aligned} &\lambda_i = \frac{2h_i}{h_{i-1}+h_i}, \quad \mu_i = \frac{2h_{i-1}}{h_{i-1}+h_i}, \\ &\gamma_i := 12\frac{\delta_i - \delta_{i-1}}{h_{i-1}+h_i}, \quad i = 2, \ldots, n-1, \\ &\gamma_1 := 6(\delta_1 - \sigma_a)/h_1, \quad \gamma_n := 6(\sigma_b - \delta_{n-1})/h_{n-1}. \end{aligned} \tag{2.29}$$

Then g *is the unique cubic spline interpolant to the data* (2.13) *and* $m_i := g''(x_i)$ *for* $i = 1, \ldots, n$*. Moreover,* (2.28) *has a unique solution.*

Hint: Show that $g(x_i) = y_i$*,* $g''(x_i) = m_i$ *for* $i = 1, \ldots, n$*, and* $g \in C^2$ *if* $p'_{i-1}(x_i) = p'_i(x_i)$ *for* $i = 2, \ldots, n-1$*. The requirements* $g'(a) = \sigma_a$*,* $g'(b) = \sigma_b$ *lead to the first and last equation in* (2.28)*.*

Exercise 2.31 Not-a-knot boundary condition. *Suppose* $n \geq 5$ *and consider finding a cubic spline* g *such that*

$$g(x_i) = y_i, \quad i = 1, \ldots, n, \quad p_1 = p_2, p_{n-2} = p_{n-1}.$$

Since x_2 *and* x_{n-1} *are no longer knots we refer to this as the not-a-knot condition. The spline* g *now consists of only* $n-3$ *pieces*

$$g(x) := \begin{cases} p_2(x), & \text{if } a \leq x < x_3, \\ p_3(x), & \text{if } x_3 \leq x < x_4, \\ \vdots & \\ p_{n-3}(x), & \text{if } x_{n-3} \leq x < x_{n-2}, \\ p_{n-2}(x), & \text{if } x_{n-2} \leq x \leq b. \end{cases} \tag{2.30}$$

With the shifted power form (2.16) *we obtain the pp representation*

$$\boldsymbol{x}_s = [x_2, \ldots, x_{n-2}], \quad \boldsymbol{C} := \begin{bmatrix} c_{12} & c_{13} & \cdots & c_{1,n-2} \\ c_{22} & c_{23} & \cdots & c_{2,n-2} \\ c_{32} & c_{33} & \cdots & c_{3,n-2} \\ c_{42} & c_{43} & \cdots & c_{4,n-2} \end{bmatrix} \in \mathbb{R}^{4,n-3}.$$

Show, using $p_2(a) = y_1$ *and* $p_{n-2}(b) = y_n$ *in addition to* (2.25) *for* $i = 3, \ldots, n-2$

that we obtain the linear system

$$\boldsymbol{N}_3\tilde{\boldsymbol{s}} := \begin{bmatrix} 2 & \mu_2 & & & \\ \lambda_3 & 4 & \mu_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-2} & 4 & \mu_{n-2} \\ & & & \lambda_{n-1} & 2 \end{bmatrix} \begin{bmatrix} s_2 \\ s_3 \\ \vdots \\ s_{n-2} \\ s_{n-1} \end{bmatrix} = \begin{bmatrix} \nu_2 \\ \beta_3 \\ \vdots \\ \beta_{n-2} \\ \nu_{n-1} \end{bmatrix}, \tag{2.31}$$

where $\lambda_i, \mu_i, \beta_i$ *is given by* (2.21) *and*

$$\begin{aligned} \nu_2 &:= \frac{1}{2}\lambda_2^2\delta_1 + \frac{1}{2}\mu_2^2(2 + 3h_2/h_1)\delta_2, \\ \nu_{n-1} &:= \frac{1}{2}\mu_{n-1}^2\delta_{n-1} + \frac{1}{2}\lambda_{n-1}^2(2 + 3h_{n-2}/h_{n-1})\delta_{n-2}. \end{aligned} \tag{2.32}$$

Explain why $\boldsymbol{N}_3$ *is nonsingular. Note that* s_{n-1} *is not needed for the pp representation of* g.

Chapter 3

LU Factorizations

In Chapter 2 we saw how an LU factorization of the coefficient matrix can be used to solve certain tridiagonal systems efficiently. In this chapter we consider the general theory of LU factorizations[1]. We consider some related factorizations called block LU, PLU, symmetric LU, and Cholesky.

3.1 The LU Factorization

We say that $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ is an **LU factorization** of $\boldsymbol{A} \in \mathbb{C}^{n,n}$ if $\boldsymbol{L} \in \mathbb{C}^{n,n}$ is lower triangular (**left triangular**)) and $\boldsymbol{R} \in \mathbb{C}^{n,n}$ is upper triangular (**right triangular**). In addition we will assume that $\boldsymbol{L}$ is unit triangular, i. e., it has ones on the diagonal. The LU factorization of the 2. derivative matrix $\boldsymbol{T}$ was given in (2.9). But not every nonsingular matrix has an LU factorization.

Example 3.1 An LU factorization of $\boldsymbol{A} = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ must satisfy the equations

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_1 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_3 \\ 0 & r_2 \end{bmatrix} = \begin{bmatrix} r_1 & r_3 \\ l_1 r_1 & l_1 r_3 + r_2 \end{bmatrix}$$

for the unknowns l_1 in $\boldsymbol{L}$ and r_1, r_2, r_3 in $\boldsymbol{R}$. Comparing $(1,1)$ elements we see that $r_1 = 0$, which makes it impossible to satisfy the condition $1 = l_1 r_1$ for the $(2,1)$ element. We conclude that $\boldsymbol{A}$ has no LU factorization.

We will make use of some special submatrices.

Definition 3.2 *For $k = 1, \ldots, n$ the matrices $\boldsymbol{A}_k \in \mathbb{C}^{k,k}$ given by*

$$\boldsymbol{A}_k := \boldsymbol{A}(1:k, 1:k) = \begin{bmatrix} a_{11} & \cdots & a_{k1} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}$$

[1] In the literature an upper triangular matrix is denoted by $\boldsymbol{U}$ in an LU factorization and $\boldsymbol{R}$ in a QR factorization. (see Chapter 12). We have chosen to use $\boldsymbol{R}$ to denote an upper triangular matrix both for LU and QR factorizations.

are called the **leading principal submatrices** *of* $\boldsymbol{A} = \boldsymbol{A}_n \in \mathbb{C}^{n,n}$. *More generally, a matrix* $\boldsymbol{B} \in \mathbb{C}^{k,k}$ *is called a* **principal submatrix** *of* $\boldsymbol{A}$ *if* $\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{r}, \boldsymbol{r})$, *where* $\boldsymbol{r} = [r_1, \ldots, r_k]$ *for some* $1 \le r_1 < \cdots < r_k \le n$. *Thus*

$$b_{i,j} = a_{r_i, r_j}, \quad i, j = 1, \ldots, k,$$

The determinant of a (leading) principal submatrix is called a **(leading) principal minor**.

A principal submatrix is leading if $r_j = j$ for $j = 1, \ldots, k$. Also a principal submatrix is special in that it uses the same rows and columns of $\boldsymbol{A}$. For example, for $k = 1$ the only principal submatrices are the diagonal elements of $\boldsymbol{A}$.

Example 3.3 *The principal submatrices of* $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{smallmatrix}\right]$ *are*

$$[1],\ [5],\ [9],\ \left[\begin{smallmatrix} 1 & 2 \\ 4 & 5 \end{smallmatrix}\right],\ \left[\begin{smallmatrix} 1 & 3 \\ 7 & 9 \end{smallmatrix}\right],\ \left[\begin{smallmatrix} 5 & 6 \\ 8 & 9 \end{smallmatrix}\right],\ \boldsymbol{A}.$$

The leading principal submatrices are

$$[1],\ \left[\begin{smallmatrix} 1 & 2 \\ 4 & 5 \end{smallmatrix}\right],\ \boldsymbol{A}.$$

Theorem 3.4 *Suppose the leading principal submatrices* $\boldsymbol{A}_k$ *of* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *are nonsingular for* $k = 1, \ldots, n-1$. *Then* $\boldsymbol{A}$ *has a unique LU factorization.*

Proof. We use induction on n to show that $\boldsymbol{A}$ has a unique LU factorization. The result is clearly true for $n = 1$, since the unique LU factorization of a 1-by-1 matrix is $[a_{11}] = [1][a_{11}]$. Suppose that $\boldsymbol{A}_{n-1}$ has a unique LU factorization $\boldsymbol{A}_{n-1} = \boldsymbol{L}_{n-1}\boldsymbol{R}_{n-1}$, and that $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n-1}$ are nonsingular. Since $\boldsymbol{A}_{n-1}$ is nonsingular it follows that $\boldsymbol{L}_{n-1}$ and $\boldsymbol{R}_{n-1}$ are nonsingular. But then

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{n-1} & \boldsymbol{b} \\ \boldsymbol{c}^T & a_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{n-1} & \boldsymbol{0} \\ \boldsymbol{c}^T\boldsymbol{R}_{n-1}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{n-1} & \boldsymbol{v} \\ 0 & a_{nn} - \boldsymbol{c}^T\boldsymbol{R}_{n-1}^{-1}\boldsymbol{v} \end{bmatrix} = \boldsymbol{L}\boldsymbol{R}, \tag{3.1}$$

where $\boldsymbol{v} = \boldsymbol{L}_{n-1}^{-1}\boldsymbol{b}$, and this is an LU factorization of $\boldsymbol{A}$. Since $\boldsymbol{L}_{n-1}$ and $\boldsymbol{R}_{n-1}$ are nonsingular the block (2,1) element in $\boldsymbol{L}$ and the block $(1,2)$ element in $\boldsymbol{R}$ are uniquely given in (3.1), and then r_{nn} is also determined uniquely. Thus the LU factorization is unique by construction. ∎

The following observation is useful.

Lemma 3.5 *Suppose* $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ *is an LU factorization of* $\boldsymbol{A} \in \mathbb{C}^{n,n}$. *For* $k = 1, \ldots, n$ *let* $\boldsymbol{A}_k, \boldsymbol{L}_k, \boldsymbol{R}_k$ *be the leading principal submatrices of* $\boldsymbol{A}, \boldsymbol{L}, \boldsymbol{R}$, *respectively. Then* $\boldsymbol{A}_k = \boldsymbol{L}_k\boldsymbol{R}_k$ *is an LU factorization of* $\boldsymbol{A}_k$ *for* $k = 1, \ldots, n$.

Proof. For $k = 1, \ldots, n-1$ we partition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ as follows:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_k & \boldsymbol{B}_k \\ \boldsymbol{C}_k & \boldsymbol{D}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_k & \boldsymbol{0} \\ \boldsymbol{M}_k & \boldsymbol{N}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_k & \boldsymbol{S}_k \\ 0 & \boldsymbol{T}_k \end{bmatrix} = \boldsymbol{L}\boldsymbol{R}, \tag{3.2}$$

where $\boldsymbol{D}_k, \boldsymbol{N}_k, \boldsymbol{T}_k \in \mathbb{C}^{n-k,n-k}$. Using block multiplication we find $\boldsymbol{A}_k = \boldsymbol{L}_k\boldsymbol{R}_k$. Since $\boldsymbol{L}_k$ is unit lower triangular and $\boldsymbol{R}_k$ is upper triangular we see that this gives an LU factorization of $\boldsymbol{A}_k$. ∎

There is a converse of Theorem 3.4.

Theorem 3.6 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ has an LU factorization. If $\boldsymbol{A}$ is nonsingular then the leading principal submatrices $\boldsymbol{A}_k$ are nonsingular for $k = 1, \ldots, n-1$.*

Proof. Suppose $\boldsymbol{A}$ is nonsingular with the LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$. Since $\boldsymbol{A}$ is nonsingular it follows that $\boldsymbol{L}$ and $\boldsymbol{R}$ are nonsingular. Let $1 \leq k \leq n$. By Lemma 3.5 it follows that $\boldsymbol{A}_k = \boldsymbol{L}_k\boldsymbol{R}_k$. Since $\boldsymbol{L}_k$ is unit lower triangular it is nonsingular. Moreover $\boldsymbol{R}_k$ is nonsingular since its diagonal elements are among the nonzero diagonal elements of $\boldsymbol{R}$. But then $\boldsymbol{A}_k$ is nonsingular.
∎

The following lemma shows that the LU factorization of a nonsingular matrix is unique.

Corollary 3.7 *The LU factorization of a nonsingular matrix is unique whenever it exists.*

Proof. By Theorem 3.6 the leading principal submatrices are nonsingular for $k = 1, \ldots, n-1$. But then by Theorem 3.4 the LU factorization is unique. ∎

Remark 3.8 *Theorem 3.6 is not true in general if $\boldsymbol{A}$ is singular. An LU factorization of an upper triangular matrix $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{I}\boldsymbol{A}$, and if $\boldsymbol{A}$ is singular it can have zeros anywhere on the diagonal. By Lemma 2.8, if some a_{kk} is zero then $\boldsymbol{A}_k$ is singular.*

Remark 3.9 *The LU factorization of a singular matrix need not be unique. In particular, for the zero matrix any unit lower triangular matrix can be used as $\boldsymbol{L}$ in an LU factorization.*

Remark 3.10 *We have shown that a nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ has an LU factorization if and only if the leading principle submatrices $\boldsymbol{A}_k$ are nonsingular for $k = 1, \ldots, n-1$. This condition seems fairly restrictive. However, for a nonsingular matrix $\boldsymbol{A}$ there always is a permutation of the rows so that the permuted matrix has an LU factorization. We obtain a factorization of the form $\boldsymbol{P}^T\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ or equivalently $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{R}$, where $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{L}$ is unit lower triangular, and $\boldsymbol{R}$ is upper triangular. We call this a* **PLU factorization** *of $\boldsymbol{A}$. (Cf. Section 3.7 and Appendix E.)*

Exercise 3.11 *Show that $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ has an LU factorization. Note that we have only interchanged rows in Example 3.1*

Exercise 3.12 *Find an LU factorization of the singular matrix* $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$. *Is it unique?*

Exercise 3.13 *Suppose* $\boldsymbol{A}$ *has an LU factorization* $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$. *Show that* $\det(\boldsymbol{A}_k) = r_{11}r_{22}\cdots r_{kk}$ *for* $k = 1, \ldots, n$.

Exercise 3.14 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *and* $\boldsymbol{A}_k$ *is nonsingular for* $k = 1, \ldots, n-1$. *Use Exercise 3.13 to show that the diagonal elements* r_{kk} *in the LU factorization are*

$$r_{11} = a_{11}, \quad r_{kk} = \frac{\det(\boldsymbol{A}_k)}{\det(\boldsymbol{A}_{k-1})}, \text{ for } k = 2, \ldots, n. \tag{3.3}$$

3.2 Block LU Factorization

Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is a block matrix of the form

$$\boldsymbol{A} := \begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1m} \\ \vdots & & \vdots \\ \boldsymbol{A}_{m1} & \cdots & \boldsymbol{A}_{mm} \end{bmatrix}, \tag{3.4}$$

where each (diagonal) block $\boldsymbol{A}_{ii}$ is square. We call the factorization

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I} & & & \\ \boldsymbol{L}_{21} & \boldsymbol{I} & & \\ \vdots & & \ddots & \\ \boldsymbol{L}_{m1} & \cdots & \boldsymbol{L}_{m,m-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{11} & & \cdots & \boldsymbol{R}_{1m} \\ & \boldsymbol{R}_{21} & \cdots & \boldsymbol{R}_{2m} \\ & & \ddots & \vdots \\ & & & \boldsymbol{R}_{mm} \end{bmatrix} \tag{3.5}$$

a **block LU factorization of** $\boldsymbol{A}$. Here the ith diagonal blocks $\boldsymbol{I}$ and $\boldsymbol{R}_{ii}$ in $\boldsymbol{L}$ and $\boldsymbol{R}$ have the same order as $\boldsymbol{A}_{ii}$, the ith diagonal block in $\boldsymbol{A}$.

The results for elementwise LU factorization carry over to block LU factorization as follows.

Theorem 3.15 *Suppose* $\boldsymbol{A} \in \mathbb{R}^{n,n}$ *is a block matrix of the form* (3.4), *and the* **leading principal block submatrices**

$$\boldsymbol{A}_k := \begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1k} \\ \vdots & & \vdots \\ \boldsymbol{A}_{k1} & \cdots & \boldsymbol{A}_{kk} \end{bmatrix}$$

are nonsingular for $k = 1, \ldots, m-1$. *Then* $\boldsymbol{A}$ *has a unique block LU factorization* (3.5). *Conversely, if* $\boldsymbol{A}$ *is nonsingular and has a block LU factorization then* $\boldsymbol{A}_k$ *is nonsingular for* $k = 1, \ldots, m-1$.

Proof. Suppose $\boldsymbol{A}_k$ is nonsingular for $k = 1, \ldots, m-1$. Following the proof in Theorem 3.4 suppose $\boldsymbol{A}_{m-1}$ has a unique LU factorization $\boldsymbol{A}_{m-1} = \boldsymbol{L}_{m-1}\boldsymbol{R}_{m-1}$,

and that $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{m-1}$ are nonsingular. Then $\boldsymbol{L}_{m-1}$ and $\boldsymbol{R}_{m-1}$ are nonsingular and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{m-1} & \boldsymbol{B} \\ \boldsymbol{C}^T & \boldsymbol{A}_{mm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{m-1} & \boldsymbol{0} \\ \boldsymbol{C}^T\boldsymbol{R}_{m-1}^{-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{m-1} & \boldsymbol{L}_{m-1}^{-1}\boldsymbol{B} \\ 0 & \boldsymbol{A}_{mm} - \boldsymbol{C}^T\boldsymbol{R}_{m-1}^{-1}\boldsymbol{L}_{m-1}^{-1}\boldsymbol{B} \end{bmatrix}, \tag{3.6}$$

is a block LU factorization of $\boldsymbol{A}$. It is unique by derivation. Conversely, suppose $\boldsymbol{A}$ is nonsingular and has a block LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$. Then as in Lemma 3.5 it is easily seen that $\boldsymbol{A}_k = \boldsymbol{L}_k\boldsymbol{R}_k$ is a block LU factorization of $\boldsymbol{A}_k$ for $k = 1, \ldots, m$. By Lemma 2.7 and induction a block triangular matrix is nonsingular if and only if the diagonal blocks are nonsingular and we see that $\boldsymbol{L}_k$ and $\boldsymbol{R}_k$ are nonsingular, and hence $\boldsymbol{A}_k$ is nonsingular for $k = 1, \ldots, m-1$. ∎

Remark 3.16 *The number of flops for the block LU factorization is the same as for the ordinary LU factorization. An advantage of the block method is that it combines many of the operations into matrix operations.*

Remark 3.17 *Note that* (3.5) *is not an LU factorization of* $\boldsymbol{A}$ *since the* $\boldsymbol{R}_{ii}$*'s are not upper triangular in general. To relate the block LU factorization to the usual LU factorization we assume that each* $\boldsymbol{R}_{ii}$ *has an LU factorization* $\boldsymbol{R}_{ii} = \tilde{\boldsymbol{L}}_{ii}\tilde{\boldsymbol{R}}_{ii}$*. Then* $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{R}}$*, where* $\hat{\boldsymbol{L}} := \boldsymbol{L}\,\mathrm{diag}(\tilde{\boldsymbol{L}}_{ii})$ *and* $\hat{\boldsymbol{R}} := \mathrm{diag}(\tilde{\boldsymbol{L}}_{ii}^{-1})\boldsymbol{R}$*, and this is an ordinary LU factorization of* $\boldsymbol{A}$*.*

Exercise 3.18 *Show that* $\hat{\boldsymbol{L}}$ *is unit lower triangular and* $\hat{\boldsymbol{R}}$ *is upper triangular.*

3.3 The Symmetric LU Factorization

We consider next LU factorization of a real symmetric matrix.

Definition 3.19 *Suppose* $\boldsymbol{A} \in \mathbb{R}^{n,n}$*. A factorization* $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$*, where* $\boldsymbol{L}$ *is unit lower triangular and* $\boldsymbol{D}$ *is diagonal is called a* **symmetric LU factorization** *of* $\boldsymbol{A}$*.*

A matrix which has a symmetric LU factorization must be symmetric since $\boldsymbol{A}^T = (\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T)^T = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T = \boldsymbol{A}$.

Theorem 3.20 *Suppose* $\boldsymbol{A} \in \mathbb{R}^{n,n}$ *is nonsingular. Then* $\boldsymbol{A}$ *has a symmetric LU factorization if and only if* $\boldsymbol{A} = \boldsymbol{A}^T$ *and* $\boldsymbol{A}_k$ *is nonsingular for* $k = 1, \ldots, n-1$*. The symmetric LU factorization is unique.*

Proof. If $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n-1}$ are nonsingular then Theorem 3.4 implies that $\boldsymbol{A}$ has a unique LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$. Since $\boldsymbol{A}$ is nonsingular it follows that $\boldsymbol{R}$ is nonsingular and since $\boldsymbol{R}$ is triangular the diagonal matrix $\boldsymbol{D} := \mathrm{diag}(r_{11}, \ldots, r_{nn})$ is nonsingular (cf. Lemma 2.8). But then $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{M}^T$, where $\boldsymbol{M}^T = \boldsymbol{D}^{-1}\boldsymbol{R}$ is

unit upper triangular. By symmetry $\boldsymbol{A} = \boldsymbol{L}(\boldsymbol{D}\boldsymbol{M}^T) = \boldsymbol{M}(\boldsymbol{D}\boldsymbol{L}^T) = \boldsymbol{A}^T$ are two LU factorizations of $\boldsymbol{A}$, and by uniqueness $\boldsymbol{M} = \boldsymbol{L}$. Thus $\boldsymbol{A}$ has a unique symmetric LU factorization.

Conversely, if $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ is the symmetric LU factorization of $\boldsymbol{A}$ then $\boldsymbol{A}$ is symmetric since $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ is symmetric, and $\boldsymbol{A}$ has an LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ with $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{L}^T$. By Theorem 3.6 we conclude that $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n-1}$ are nonsingular. □

3.4 Positive Definite- and Positive Semidefinite Matrices

Symmetric positive definite matrices occur often in scientific computing. For example, the second derivative matrix is symmetric positive definite, see Lemma 3.21 below. For symmetric positive definite and symmetric positive semidefinite matrices there is a special version of the symmetric LU factorization. Before considering this factorization we study some properties of positive (semi)definite matrices. We study only real matrices, but consider also the nonsymmetric case..

3.4.1 Definition and Examples

Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is a square matrix. The function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$

is called a **quadratic form**. We say that $\boldsymbol{A}$ is

(i) positive definite if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$.

(ii) positive semidefinite if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

(iii) negative (semi)definite if $-\boldsymbol{A}$ is positive (semi)definite.

(iv) symmetric positive (semi)definite if $\boldsymbol{A}$ is symmetric in addition to being positive (semi)definite.

(v) symmetric negative (semi)definite if $\boldsymbol{A}$ is symmetric in addition to being negative (semi)definite.

We observe the following.

- A matrix is positive definite if it is positive semidefinite and in addition

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0. \tag{3.7}$$

- The zero-matrix is symmetric positive semidefinite, while the unit matrix is symmetric positive definite.

- A positive definite matrix must be nonsingular. Indeed, if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ for some $\boldsymbol{x} \in \mathbb{R}^n$ then $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0$ which by (3.7) implies that $\boldsymbol{x} = \boldsymbol{0}$.

Lemma 3.21 *The second derivative matrix* $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{n,n}$ *is symmetric positive definite.*

Proof. Clearly $\boldsymbol{T}$ is symmetric. For any $\boldsymbol{x} \in \mathbb{R}^n$

$$\begin{aligned}\boldsymbol{x}^T\boldsymbol{T}\boldsymbol{x} &= 2\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n-1} x_i x_{i+1} - \sum_{i=2}^{n} x_{i-1}x_i \\ &= \sum_{i=1}^{n-1} x_i^2 - 2\sum_{i=1}^{n-1} x_i x_{i+1} + \sum_{i=1}^{n-1} x_{i+1}^2 + x_1^2 + x_n^2 \\ &= x_1^2 + x_n^2 + \sum_{i=1}^{n-1}(x_{i+1} - x_i)^2.\end{aligned}$$

Thus $\boldsymbol{x}^T\boldsymbol{T}\boldsymbol{x} \geq 0$ and if $\boldsymbol{x}^T\boldsymbol{T}\boldsymbol{x} = 0$ then $x_1 = x_n = 0$ and $x_i = x_{i+1}$ for $i = 1, \ldots, n-1$ which implies that $\boldsymbol{x} = 0$. Hence $\boldsymbol{T}$ is positive definite. ∎

Example 3.22 *Consider (cf.* (G.1)*) the gradient* ∇f *and hessian* $\nabla\nabla^T f$ *of a function* $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n, \quad \nabla\nabla^T f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n,n}.$$

We assume that f *has continuous first and second partial derivatives on* Ω.

Under suitable conditions on the domain Ω *it is shown in advanced calculus texts that if* $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ *and* $\nabla\nabla^T f(\boldsymbol{x})$ *is positive definite then* $\boldsymbol{x}$ *is a local minimum for* f. *This can be shown using the second-order Taylor expansion* (G.2). *Moreover,* $\boldsymbol{x}$ *is a local maximum if* $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ *and* $\nabla\nabla^T f(\boldsymbol{x})$ *is negative definite.*

3.4.2 Some Criteria for the Nonsymmetric Case

We treat the positive definite and positive semidefinite cases in parallel.

Theorem 3.23 *Let* m, n *be positive integers. If* $\boldsymbol{A} \in \mathbb{R}^{n,n}$ *is positive semidefinite and* $\boldsymbol{X} \in \mathbb{R}^{n,m}$ *then* $\boldsymbol{B} := \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \in \mathbb{R}^{m,m}$ *is positive semidefinite. If in addition* $\boldsymbol{A}$ *is positive definite and* $\boldsymbol{X}$ *has linearly independent columns then* $\boldsymbol{B}$ *is positive definite.*

Proof. Let $\boldsymbol{y} \in \mathbb{R}^m$ and set $\boldsymbol{x} := \boldsymbol{X}\boldsymbol{y}$. Then $\boldsymbol{y}^T\boldsymbol{B}\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$. If $\boldsymbol{A}$ is positive definite and $\boldsymbol{X}$ has linearly independent columns then $\boldsymbol{x}$ is nonzero if $\boldsymbol{y}$ is nonzero and $\boldsymbol{y}^T\boldsymbol{B}\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$. ∎

Taking $\boldsymbol{A} := \boldsymbol{I}$ and $\boldsymbol{X} := \boldsymbol{A}$ we obtain

Corollary 3.24 *Let m, n be positive integers. If $\boldsymbol{A} \in \mathbb{R}^{m,n}$ then $\boldsymbol{A}^T\boldsymbol{A}$ is positive semidefinite. If in addition $\boldsymbol{A}$ has linearly independent columns then $\boldsymbol{A}^T\boldsymbol{A}$ is positive definite.*

Theorem 3.25 *Any principal submatrix of a positive (semi)definite matrix is positive (semi)definite.*

Proof. Suppose the submatrix $\boldsymbol{B}$ is defined by the rows and columns $r_1, \ldots, r_k$ of $\boldsymbol{A}$. Then $\boldsymbol{B} := \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{X} = [\boldsymbol{e}_{r_1}, \ldots, \boldsymbol{e}_{r_k}] \in \mathbb{R}^{n,k}$, and $\boldsymbol{B}$ is positive (semi)definite by Theorem 3.23. □

If $\boldsymbol{A}$ is positive definite then the leading principal submatrices are nonsingular and we obtain:

Corollary 3.26 *A positive definite matrix has a unique LU factorization.*

Theorem 3.27 *A positive (semi)definite matrix $\boldsymbol{A}$ has positive (nonnegative) eigenvalues. Conversely, if $\boldsymbol{A}$ has positive (nonnegative) eigenvalues and orthonormal eigenvectors then it is positive (semi)definite.*

Proof. Consider the positive definite case. Suppose $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\boldsymbol{x} \neq \boldsymbol{0}$. Multiplying both sides by $\boldsymbol{x}^T$ and solving for λ we find $\lambda = \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} > 0$. Suppose conversely that $\boldsymbol{A} \in \mathbb{R}^{n,n}$ has eigenpairs $(\lambda_j, \boldsymbol{u}_j)$, $j = 1, \ldots, n$. Let $\boldsymbol{U} := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathbb{R}^{n,n}$ and $\boldsymbol{D} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Since $\boldsymbol{u}_i^T\boldsymbol{u}_j = \delta_{ij}$, $i, j = 1, \ldots, n$, it follows that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$. But then $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ as well. Now $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_j$ for $j = 1, \ldots, n$ implies $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{D}$ and therefore $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D} = \boldsymbol{D}$. Let $\boldsymbol{x} \in \mathbb{R}^n$ be nonzero and define $\boldsymbol{c} := \boldsymbol{U}^T\boldsymbol{x} = [c_1, \ldots, c_n]^T$. Then $\boldsymbol{U}\boldsymbol{c} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x} = \boldsymbol{x}$, and so

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{U}\boldsymbol{c})^T\boldsymbol{A}\boldsymbol{U}\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{D}\boldsymbol{c} = \sum_{j=1}^{n} \lambda_j c_j^2.$$

But $\boldsymbol{U}^T$ nonsingular implies $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{x} \neq 0$, and since $\lambda_j > 0$ for $j = 1, \ldots, n$ it follows that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$ so that $\boldsymbol{A}$ is positive definite. The positive semidefinite case is similar. □

Theorem 3.28 *If $\boldsymbol{A}$ is positive (semi)definite then* $\det(\boldsymbol{A}) > 0$ *(*$\det(\boldsymbol{A}) \geq 0$*).*

Proof. Since the determinant of a matrix is equal to the product of its eigenvalues this follows from Theorem 3.27. □

3.5 The Symmetric Case and Cholesky Factorization

For symmetric positive definite matrices there is an alternative to the symmetric LU factorization known as the Cholesky factorization. We consider also a closely related factorization of symmetric positive semidefinite matrices.

We need the following necessary conditions for symmetric positive semidefinite matrices.

Lemma 3.29 *If $\boldsymbol{A}$ is symmetric positive semidefinite then for all i, j*

1. $|a_{ij}| \leq (a_{ii} + a_{jj})/2$,
2. $|a_{ij}| \leq \sqrt{a_{ii}a_{jj}}$,
3. $\max_{i,j} |a_{ij}| = \max_i a_{ii}$,
4. $a_{ii} = 0 \Longrightarrow a_{ij} = a_{ji} = 0$, *fixed i, all j.*

Proof. 3. follows from 1. and 4. from 2. Now

$$0 \leq (\alpha \boldsymbol{e}_i + \beta \boldsymbol{e}_j)^T \boldsymbol{A}(\alpha \boldsymbol{e}_i + \beta \boldsymbol{e}_j) = \alpha^2 a_{ii} + \beta^2 a_{jj} + 2\alpha\beta a_{ij}, \text{ all } i, j, \ \alpha, \beta \in \mathbb{R}, \quad (3.8)$$

since $\boldsymbol{A}$ is symmetric positive semidefinite. Taking $\alpha = 1$, $\beta = \pm 1$ we obtain $a_{ii} + a_{jj} \pm 2a_{ij} \geq 0$ and this implies 1. 2. follows trivially from 1. if $a_{ii} = a_{jj} = 0$. Suppose one of them, say a_{ii} is nonzero. Note that $a_{ii} = \boldsymbol{e}_i^T \boldsymbol{A} \boldsymbol{e}_i > 0$. Taking $\alpha = -a_{ij}$, $\beta = a_{ii}$ in (3.8) we find

$$0 \leq a_{ij}^2 a_{ii} + a_{ii}^2 a_{jj} - 2a_{ij}^2 a_{ii} = a_{ii}(a_{ii}a_{jj} - a_{ij}^2).$$

But then $a_{ii}a_{jj} - a_{ij}^2 \geq 0$ and 2. follows. ∎

As an illustration consider the matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}, \quad \boldsymbol{A}_3 = \begin{bmatrix} -2 & 1 \\ 1 & 2 \end{bmatrix}.$$

None of them is positive semidefinite, since neither 1. nor 2. hold.

Definition 3.30 *A factorization $\boldsymbol{A} = \boldsymbol{R}^T \boldsymbol{R}$ where $\boldsymbol{R}$ is upper triangular with positive diagonal elements is called a* **Cholesky factorization**. *A factorization $\boldsymbol{A} = \boldsymbol{R}^T \boldsymbol{R}$ where $\boldsymbol{R}$ is upper triangular with nonnegative diagonal elements is called a* **semi-Cholesky factorization**.

Note that a semi-Cholesky factorization of a symmetric positive definite matrix is necessarily a Cholesky factorization. For if $\boldsymbol{A}$ is positive definite then it is nonsingular and then $\boldsymbol{R}$ must be nonsingular. Thus the diagonal elements of $\boldsymbol{R}$ cannot be zero.

Exercise 3.31 *Show that a symmetric matrix has a Cholesky factorization if and only if it has a symmetric LU factorization with positive diagonal elements in $\boldsymbol{D}$.*

Theorem 3.32 *A matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ has a Cholesky factorization $\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{R}$ if and only if it is symmetric positive definite.*

Proof. If $\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{R}$ is a Cholesky factorization then $\boldsymbol{A}$ is symmetric. Since $\boldsymbol{R}$ has positive diagonal elements it is nonsingular. Thus $\boldsymbol{A}$ is symmetric positive definite by Corollary 3.24. The proof of the converse will lead to an algortihm. We use induction on n. A positive definite matrix of order one has a Cholesky factorization since the one and only element in $\boldsymbol{A}$ is positive. Suppose any symmetric positive definite matrix of order $n-1$ has a Cholesky factorization and suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is symmetric positive definite. We partition $\boldsymbol{A}$ as follows

$$\boldsymbol{A} = \begin{bmatrix} \alpha & \boldsymbol{v}^T \\ \boldsymbol{v} & \boldsymbol{B} \end{bmatrix}, \quad \alpha \in \mathbb{R},\ \boldsymbol{v} \in \mathbb{R}^{n-1},\ \boldsymbol{B} \in \mathbb{R}^{n-1,n-1}. \tag{3.9}$$

Clearly $\alpha = \boldsymbol{e}_1^T\boldsymbol{A}\boldsymbol{e}_1 > 0$. We claim that $\boldsymbol{C} := \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T/\alpha$ is symmetric positive definite. $\boldsymbol{C}$ is symmetric. To show that $\boldsymbol{C}$ is positive definite we let $\boldsymbol{y} \in \mathbb{R}^{n-1}$ be nonzero and define $\boldsymbol{x}^T := [-\boldsymbol{v}^T\boldsymbol{y}/\alpha, \boldsymbol{y}^T] \in \mathbb{R}^n$. Then $\boldsymbol{x} \neq \boldsymbol{0}$ and

$$\begin{aligned} 0 < \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= [-\boldsymbol{v}^T\boldsymbol{y}/\alpha, \boldsymbol{y}^T] \begin{bmatrix} \alpha & \boldsymbol{v}^T \\ \boldsymbol{v} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} -\boldsymbol{v}^T\boldsymbol{y}/\alpha \\ \boldsymbol{y} \end{bmatrix} \\ &= [0, -(\boldsymbol{v}^T\boldsymbol{y})\boldsymbol{v}^T/\alpha + \boldsymbol{y}^T\boldsymbol{B}] \begin{bmatrix} -\boldsymbol{v}^T\boldsymbol{y}/\alpha \\ \boldsymbol{y} \end{bmatrix} \\ &= -(\boldsymbol{v}^T\boldsymbol{y})(\boldsymbol{v}^T\boldsymbol{y})/\alpha + \boldsymbol{y}^T\boldsymbol{B}\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}, \end{aligned} \tag{3.10}$$

since $(\boldsymbol{v}^T\boldsymbol{y})\boldsymbol{v}^T\boldsymbol{y} = (\boldsymbol{v}^T\boldsymbol{y})^T\boldsymbol{v}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{y}$. So $\boldsymbol{C} \in \mathbb{R}^{n-1,n-1}$ is symmetric positive definite and by the induction hypothesis it has a Cholesky factorization $\boldsymbol{C} = \boldsymbol{R}_1^T\boldsymbol{R}_1$. The matrix

$$\boldsymbol{R} := \begin{bmatrix} \beta & \boldsymbol{v}^T/\beta \\ \boldsymbol{0} & \boldsymbol{R}_1 \end{bmatrix}, \quad \beta := \sqrt{\alpha}, \tag{3.11}$$

is upper triangular with positive diagonal elements and

$$\boldsymbol{R}^T\boldsymbol{R} = \begin{bmatrix} \beta & \boldsymbol{0} \\ \boldsymbol{v}/\beta & \boldsymbol{R}_1^T \end{bmatrix} \begin{bmatrix} \beta & \boldsymbol{v}^T/\beta \\ \boldsymbol{0} & \boldsymbol{R}_1 \end{bmatrix} = \begin{bmatrix} \alpha & \boldsymbol{v}^T \\ \boldsymbol{v} & \boldsymbol{B} \end{bmatrix} = \boldsymbol{A}$$

is a Cholesky factorization of $\boldsymbol{A}$. □

We can now show

Theorem 3.33 *The following is equivalent for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$.*

1. *$\boldsymbol{A}$ is positive definite.*
2. *$\boldsymbol{A}$ has only positive eigenvalues.*
3. *All leading principal minors are positive.*
4. *$\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ for a nonsingular $\boldsymbol{B} \in \mathbb{R}^{n,n}$.*

Proof. A symmetric matrix has a set of eigenvectors that form an orthonormal basis for $\mathbb{R}^n$ (Cf. Theorem 6.5). Therefore, by Theorem 3.27 we know that $1 \Leftrightarrow 2$. We show that $1 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$.
1 $\Rightarrow$ **3**: By Theorem 3.25 the leading principal submatrix $\boldsymbol{A}_k$ of $\boldsymbol{A}$ is positive definite, and has a positive determinant by Theorem 3.28.
3 $\Rightarrow$ **4**: Since all principal minors of $\boldsymbol{A}$ are positive the principal submatrices $\boldsymbol{A}_k$ are nonsingular for all k and therefore $\boldsymbol{A}$ has a symmetric LU factorization. By Exercise 3.31 $\boldsymbol{A}$ has a Cholesky factorization and we can take $\boldsymbol{B} = \boldsymbol{R}$.
4 $\Rightarrow$ **1**: This follows from Corollary 3.24. □

Consider next the semi-Cholesky factorization.

Theorem 3.34 *A matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ has a semi-Cholesky factorization $\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{R}$ if and only if it is symmetric positive semidefinite.*

Proof. If $\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{R}$ is a semi-Cholesky factorization then $\boldsymbol{A}$ is symmetric and it is positive semidefinite by Corollary 3.24. Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is symmetric positive semidefinite. A symmetric positive semidefinite matrix of order one has a semi-Cholesky factorization since a_{11} is nonnegative. Suppose by induction on n that any symmetric positive semidefinite matrix $\boldsymbol{C}$ of order $n-1$ has a semi-Cholesky factorization. We partition $\boldsymbol{A}$ as in (3.9). There are two cases. If $\alpha > 0$ then we obtain a semi-Cholesky factorization of $\boldsymbol{A}$ as in the proof of Theorem 3.32 since $\boldsymbol{C}$ is symmetric positive semidefinite. This follows as in (3.10) since now $0 \leq \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}$. If $\alpha = 0$ then it follows from 4. in Lemma 3.29 that $\boldsymbol{v} = \boldsymbol{0}$. Moreover, $\boldsymbol{B} \in \mathbb{R}^{n-1,n-1}$ in (3.9) is positive semidefinite and therefore has a semi-Cholesky factorization $\boldsymbol{R}_1$. But then $\boldsymbol{R} = \begin{bmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{R}_1 \end{bmatrix}$ is a semi-Cholesky factorization of $\boldsymbol{A}$. Indeed, $\boldsymbol{R}$ is upper triangular and

$$\boldsymbol{R}^T\boldsymbol{R} = \begin{bmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{R}_1^T \end{bmatrix} \begin{bmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{R}_1 \end{bmatrix} = \begin{bmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix} = \boldsymbol{A}.$$

□

Theorem 3.35 *The following is equivalent for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$.*

1. *$\boldsymbol{A}$ is positive semidefinite.*
2. *$\boldsymbol{A}$ has only nonnegative eigenvalues.*
3. *$\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ for some $\boldsymbol{B} \in \mathbb{R}^{n,n}$.*
4. *All principal minors are nonnegative.*

Proof. The proof of 1. $\Leftrightarrow$ 2 follows as in the proof of Theorem 3.33. 1. $\Leftrightarrow$ 3. follows from Theorem 3.34 while 1. $\Rightarrow$ 4. is a consequence of Theorem 3.25. To prove 4. $\Rightarrow$ 1. one first shows that $\epsilon\boldsymbol{I} + \boldsymbol{A}$ is symmetric positive definite for all $\epsilon > 0$

(Cf. page 567 of [15]). But then $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \lim_{\epsilon\to 0} \boldsymbol{x}^T(\epsilon\boldsymbol{I} + \boldsymbol{A})\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$. ∎

In 4. of Theorem 3.35 we require nonnegativity of all principal minors, while only positivity of leading principal minors was required for positive definite matrices (cf. Theorem 3.33). To see that nonnegativity of the leading principal minors is not enough consider the matrix $\boldsymbol{A} := \left[\begin{smallmatrix} 0 & 0 \\ 0 & -1 \end{smallmatrix}\right]$. The leading principal minors are nonnegative, but $\boldsymbol{A}$ is not positive semidefinite.

3.6 An Algorithm for SemiCholesky Factorization of a Banded Matrix

Recall that a matrix $\boldsymbol{A}$ has bandwidth $d \geq 0$ if $a_{ij} = 0$ for $|i - j| > d$. A (semi)Cholesky factorization preserves bandwidth.

Theorem 3.36 *The Cholesky factor $\boldsymbol{R}$ given by* (3.11) *has the same bandwidth as $\boldsymbol{A}$.*

Proof. Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ has bandwidth $d \geq 0$. Then $\boldsymbol{v}^T = [\boldsymbol{u}^T, \boldsymbol{0}^T]$ in (3.9), where $\boldsymbol{u} \in \mathbb{R}^d$, and therefore $\boldsymbol{C} := \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T/\alpha$ differs from $\boldsymbol{B}$ only in the upper left $d \times d$ corner. It follows that $\boldsymbol{C}$ has the same bandwidth as $\boldsymbol{B}$ and $\boldsymbol{A}$. By induction on n, $\boldsymbol{C} = \boldsymbol{R}_1^T\boldsymbol{R}_1$, where $\boldsymbol{R}_1$ has the same bandwidth as $\boldsymbol{C}$. But then $\boldsymbol{R}$ in (3.11) has the same bandwidth as $\boldsymbol{A}$. ∎

Consider now implementing an algorithm based on the previous discussion. Since $\boldsymbol{A}$ is symmetric we only need to use the upper part of $\boldsymbol{A}$. The first row of $\boldsymbol{R}$ is $[\beta, \boldsymbol{v}^T/\beta]$ if $\alpha > 0$. If $\alpha = 0$ then by 4 in Lemma 3.29 the first row of $\boldsymbol{A}$ is zero and this is also the first row of $\boldsymbol{R}$.

Suppose we store the first row of $\boldsymbol{R}$ in the first row of $\boldsymbol{A}$ and the upper part of $\boldsymbol{C} = \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T/\alpha$ in the upper part of $A(2:n, 2:n)$. The first row of $\boldsymbol{R}$ and the upper part of $\boldsymbol{C}$ can be computed as follows.

```
if A(1,1) > 0
  A(1,1) = sqrt(A(1,1))
  A(1,2:n) = A(1,2:n)/A(1,1)
  for i = 2:n
    A(i,i:n) = A(i,i:n) - A(1,i) * A(1,i:n)
```

(3.12)

The code can be made more efficient when $\boldsymbol{A}$ is a band matrix. If the bandwidth is d we simply replace all occurrences of n by $\min(i + d, n)$.

Continuing the reduction we arrive at the following algorithm.

Algorithm 3.37 (bandcholesky) Suppose $\boldsymbol{A}$ is symmetric positive semidefinite. An upper triangular matrix $\boldsymbol{R}$ is computed so that $\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{R}$. This is the Cholesky factorization of $\boldsymbol{A}$ if $\boldsymbol{A}$ is symmetric positive definite and a semi-Cholesky factorization of $\boldsymbol{A}$ otherwise. The algorithm uses the Matlab command `triu`.

```
function R=bandcholesky(A,d)
n=length(A);
for k=1:n
   if A(k,k)>0
       kp=min(n,k+d);
      A(k,k)=sqrt(A(k,k));
      A(k,k+1:kp)=A(k,k+1:kp)/A(k,k);
       for i=k+1:kp
          A(i,i:kp)=A(i,i:kp)-A(k,i)*A(k,i:kp);
       end
   else
       A(k,k:kp)=zeros(1,kp-k+1);
   end
end
R=triu(A);
```

In the algorithm we overwrite the upper triangle of $\boldsymbol{A}$ with the elements of $\boldsymbol{R}$. Row k of $\boldsymbol{R}$ is zero for those k where $r_{kk} = 0$. We reduce round-off noise by forcing those rows to be zero. In the semidefinite case no update is necessary and we "do nothing".

There are many versions of Cholesky factorizations, see [3]. Algorithm 3.37 is based on outer products $\boldsymbol{v}\boldsymbol{v}^T$. An advantage of this formulation is that it can be extended to symmetric positive semidefinite matrices.

Consider next forward and backward substitution. Since $\boldsymbol{R}^T$ is lower triangular and banded the kth component of $\boldsymbol{R}^T\boldsymbol{y} = \boldsymbol{b}$ is $\sum_{j=\max(1,k-d)}^{k-1} r_{jk}y_j + r_{kk}y_k = b_k$, and solving for y_k

$$y_k = (b_k - \sum_{j=\max(1,k-d)}^{k-1} r_{jk}y_j)/r_{kk}, \text{ for } k = 1,\ldots,n, \tag{3.13}$$

Similarly the kth component of $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$ is $r_{kk}x_k + \sum_{i=k+1}^{\min(n,k+d)} r_{ki}x_i = y_k$, and solving for x_k

$$x_k = (y_k - \sum_{i=k+1}^{\min(n,k+d)} r_{ki}x_i)/r_{kk}, \text{ for } k = n, n-1,\ldots,1. \tag{3.14}$$

This give the following algorithms

Algorithm 3.38 (bandforwardsolve) Solves the lower triangular system $\boldsymbol{R}^T\boldsymbol{y} = \boldsymbol{b}$. $\boldsymbol{R}$ is upper triangular and banded with $r_{kj} = 0$ for $j - k > d$.

```
function y=bandforwardsolve(R,b,d)
 n=length(b); y=b(:);
for k=1:n
    km=max(1,k-d);
    y(k)=(y(k)-R(km:k-1,k)'*y(km:k-1))/R(k,k);
end
```

Algorithm 3.39 (bandbacksolve) Solves the upper triangular system $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$. $\boldsymbol{R}$ is upper triangular and banded with $r_{kj} = 0$ for $j - k > d$.

```
function x=bandbacksolve(R,y,d)
n=length(y); x=y(:);
for i=n:-1:1
    kp=min(n,k+d);
    x(k)=(x(k)-R(k,k+1:kp)*x(k+1:kp))/R(k,k);
end
```

For a full matrix ($d = n$) the number of flops needed for the Cholesky factorization including n square roots is given by

$$\sum_{k=1}^{n}\sum_{i=k+1}^{n}(1+\sum_{j=i}^{n}2)+n = \frac{1}{3}n(n+\frac{1}{2})(n+1) \approx n^3/3.$$

The number $n^3/3$ is half the number of flops needed for Gaussian elimination of an arbitrary matrix. We obtain this reduction since the Cholesky factorization takes advantage of the symmetry of $\boldsymbol{A}$.

The number of flops for the banded algorithms is given approximately by

$$\sum_{k=1}^{n}\sum_{i=k+1}^{k+d}(1+\sum_{j=i}^{k+d}2)+n = O(nd^2)$$

for Algorithm 3.37 and $O(2nd)$ for each of Algorithms 3.38 and 3.39. When d is small compared to n we see that these numbers are considerably smaller than the $O(n^3/3)$ and $O(2n^2)$ counts for the factorization of a full symmetric matrix.

There is also a banded version of the symmetric LU factorization which requires approximately the same number of flops as the Cholesky factorization. The choice between using a symmetric LU factorization or an R^TR factorization depends on several factors. Usually an LU or a symmetric LU factorization is preferred for matrices with small bandwidth (tridiagonal, pentadiagonal), while the R^TR factorization is restricted to symmetric positive semidefinite matrices and is often used when the bandwidth is larger.

3.7 The PLU Factorization

Suppose $\boldsymbol{A}$ is nonsingular. We show existence of a factorization $\boldsymbol{A} = \boldsymbol{PLR}$, where $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{L}$ is a unit lower triangular, and $\boldsymbol{R}$ is upper triangular. Recall that a **permutation matrix** is a matrix of the form

$$\boldsymbol{P} = [\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_n}] \in \mathbb{R}^{n,n},$$

where $\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_n}$ is a permutation of the unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n \in \mathbb{R}^n$. Since $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$ the inverse of $\boldsymbol{P}$ is equal to its transpose, $\boldsymbol{P}^{-1} = \boldsymbol{P}^T$ and $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}$ as well. A special case is an **(j,k)-Exchange Matrix** $\boldsymbol{I}_{jk}$ obtained by exchanging column j and k of the identity matrix. Since $\boldsymbol{I}_{jk} = \boldsymbol{I}_{kj}$, and we obtain the identity by applying $\boldsymbol{I}_{jk}$ twice, we see that $\boldsymbol{I}_{jk}^2 = \boldsymbol{I}$ and an exchange matrix is symmetric and equal to its own inverse. Pre-multiplying a matrix by an exchange matrix interchanges two rows of the matrix, while post-multiplication interchanges two columns.

Theorem 3.40 (The PLU theorem) *A nonsingular matrix $\boldsymbol{A}$ has a factorization $\boldsymbol{A} = \boldsymbol{PLR}$, where $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{L}$ is unit lower triangular, and $\boldsymbol{R}$ is upper triangular.*

Proof. We use induction on n. The result is obvious for $n = 1$. Suppose any nonsingular matrix of order $n-1$ has a PLU factorization and consider a nonsingular matrix $\boldsymbol{A}$ of order n. Since $\boldsymbol{A}$ is nonsingular one of the elements, say a_{r1}, in the first column of $\boldsymbol{A}$ must be nonzero. Let $\boldsymbol{B} := \boldsymbol{I}_{r1}\boldsymbol{A}$ and set

$$\boldsymbol{M}_1 := \boldsymbol{I} - \boldsymbol{m}\boldsymbol{e}_1^T, \quad \boldsymbol{m} = [0, \frac{b_{21}}{b_{11}}, \ldots, \frac{b_{n1}}{b_{11}}]^T.$$

Note that $\boldsymbol{M}_1$ is unit lower triangular and therefore nonsingular. We have $\boldsymbol{M}^{-1} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}_1^T$ since

$$(\boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}_1^T)(\boldsymbol{I} - \boldsymbol{m}\boldsymbol{e}_1^T) = \boldsymbol{I} - \boldsymbol{m}\boldsymbol{e}_1^T + \boldsymbol{m}\boldsymbol{e}_1^T - \boldsymbol{m}(\boldsymbol{e}_1^T\boldsymbol{m})\boldsymbol{e}_1^T = \boldsymbol{I}.$$

The first column of $\boldsymbol{M}_1\boldsymbol{B}$ is

$$\boldsymbol{M}_1\boldsymbol{B}\boldsymbol{e}_1 = \boldsymbol{B}\boldsymbol{e}_1 - \boldsymbol{m}\boldsymbol{e}_1^T\boldsymbol{B}\boldsymbol{e}_1 = \boldsymbol{B}\boldsymbol{e}_1 - b_{11}\boldsymbol{m} = [b_{11}, 0 \ldots, 0]^T$$

and we can write

$$\boldsymbol{M}_1\boldsymbol{B} = \boldsymbol{M}_1\boldsymbol{I}_{r1}\boldsymbol{A} = \begin{bmatrix} b_{11} & \boldsymbol{c}_2^T \\ 0 & \boldsymbol{D}_2 \end{bmatrix}, \quad \textit{with} \quad \boldsymbol{D}_2 \in \mathbb{R}^{n-1,n-1}. \tag{3.15}$$

The matrix $\boldsymbol{M}_1\boldsymbol{I}_{r1}\boldsymbol{A}$ is a product of nonsingular matrices and therefore nonsingular. By Lemma 2.7 the matrix $\boldsymbol{D}_2$ is nonsingular and by the induction hypothesis we have $\boldsymbol{D}_2 = \boldsymbol{P}_2\boldsymbol{L}_2\boldsymbol{R}_2$ or $\boldsymbol{P}_2^T\boldsymbol{D}_2 = \boldsymbol{L}_2\boldsymbol{R}_2$, where $\boldsymbol{P}_2 \in \mathbb{R}^{n-1,n-1}$ is a permutation matrix, $\boldsymbol{L}_2$ is unit lower triangular and $\boldsymbol{R}_2$ is upper triangular. Define matrices $\boldsymbol{Q}_2, \boldsymbol{M}_2, \boldsymbol{R}$ of order n by

$$\boldsymbol{Q}_2 = \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{P}_2 \end{bmatrix}, \quad \boldsymbol{M}_2 = \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{L}_2 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} b_{11} & \boldsymbol{c}_2^T \\ 0 & \boldsymbol{R}_2 \end{bmatrix}.$$

Then

$$\begin{aligned}\boldsymbol{Q}_2^T\boldsymbol{M}_1\boldsymbol{I}_{r1}\boldsymbol{A} &= \begin{bmatrix}1 & 0\\ 0 & \boldsymbol{P}_2^T\end{bmatrix}\begin{bmatrix}b_{11} & \boldsymbol{c}_2^T\\ 0 & \boldsymbol{D}_2\end{bmatrix} = \begin{bmatrix}b_{11} & \boldsymbol{c}_2^T\\ 0 & \boldsymbol{P}_2^T\boldsymbol{D}_2\end{bmatrix}\\ &= \begin{bmatrix}b_{11} & \boldsymbol{c}_2^T\\ 0 & \boldsymbol{L}_2\boldsymbol{R}_2\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & \boldsymbol{L}_2\end{bmatrix}\begin{bmatrix}b_{11} & \boldsymbol{c}_2^T\\ 0 & \boldsymbol{R}_2\end{bmatrix} = \boldsymbol{M}_2\boldsymbol{R},\end{aligned}$$

and hence

$$\boldsymbol{A} = \boldsymbol{I}_{r1}\boldsymbol{M}_1^{-1}\boldsymbol{Q}_2\boldsymbol{M}_2\boldsymbol{R} = \big(\boldsymbol{I}_{r1}\boldsymbol{Q}_2\big)\big(\boldsymbol{Q}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{Q}_2\big)\boldsymbol{M}_2\boldsymbol{R}.$$

Now

$$\begin{aligned}\boldsymbol{Q}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{Q}_2 &= \begin{bmatrix}1 & 0\\ 0 & \boldsymbol{P}_2^T\end{bmatrix}(\boldsymbol{I}+\boldsymbol{m}\boldsymbol{e}_1^T)\begin{bmatrix}1 & 0\\ 0 & \boldsymbol{P}_2\end{bmatrix} = \boldsymbol{I} + \begin{bmatrix}1 & 0\\ 0 & \boldsymbol{P}_2^T\end{bmatrix}\boldsymbol{m}\boldsymbol{e}_1^T\begin{bmatrix}1 & 0\\ 0 & \boldsymbol{P}_2\end{bmatrix}\\ &= \boldsymbol{I} + \begin{bmatrix}0\\ \boldsymbol{P}_2^T\boldsymbol{m}(2:n)\end{bmatrix}\boldsymbol{e}_1^T.\end{aligned}$$

Thus $\boldsymbol{Q}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{Q}_2$ is unit lower triangular and we have $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{R}$, where $\boldsymbol{P} = \boldsymbol{I}_{r1}\boldsymbol{Q}_2$ is a permutation matrix, $\boldsymbol{L} = \boldsymbol{Q}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{Q}_2\boldsymbol{M}_2$ is unit lower triangular, and $\boldsymbol{R}$ is upper triangular. □

To find the PLU factorization of a matrix we can use Gaussian elimination with row interchanges (pivoting). See Appendix E for details.

Chapter 4

The Kronecker Product

Matrices arising from 2D and 3D problems sometimes have a Kronecker product structure. Identifying a Kronecker structure can be very rewarding since it simplifies the study of such matrices.

4.1 Test Matrices

In this section we introduce some matrices which we will use to compare various algorithms in later chapters.

4.1.1 The 2D Poisson Problem

Consider the problem

$$-\nabla^2 u := -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \text{ on } \Omega := (0,1)^2 = \{(x,y) : 0 < x < 1,\ 0 < y < 1\}, \quad (4.1)$$

$$u := 0 \text{ on } \partial\Omega.$$

Here Ω is the open unit square while $\partial\Omega$ is the boundary of Ω. The function f is given and continuous on Ω and we seek a function $u = u(x,y)$ such that (4.1) holds and which is zero on $\partial\Omega$.

Let m be a positive integer. We solve the problem numerically by finding approximations $v_{j,k} \approx u(jh, kh)$ on a grid of points given by

$$\overline{\Omega}_h := \{(jh, kh) :\ j,k = 0, 1, \ldots, m+1\}, \quad \text{where} \quad h = 1/(m+1).$$

The points $\Omega_h := \{(jh, kh) :\ j,k = 1, \ldots, m\}$ are the interior points, while $\overline{\Omega}_h \setminus \Omega_h$ are the boundary points. The solution is zero at the boundary points. For an interior point we insert the difference approximations

$$\frac{\partial^2 u(jh,kh)}{\partial x^2} \approx \frac{v_{j-1,k} - 2v_{j,k} + v_{j+1,k}}{h^2}, \quad \frac{\partial^2 u(jh,kh)}{\partial y^2} \approx \frac{v_{j,k-1} - 2v_{j,k} + v_{j,k+1}}{h^2}$$

in (4.1) and multiply both sides by h^2 to obtain

$$(-v_{j-1,k} + 2v_{j,k} - v_{j+1,k}) + (-v_{j,k-1} + 2v_{j,k} - v_{j,k+1}) = h^2 f_{j,k} \tag{4.2}$$

or

$$4v_{j,k} - v_{j-1,k} - v_{j+1,k} - v_{j,k-1} - v_{j,k+1} = h^2 f_{jk} := h^2 f(jh, kh). \tag{4.3}$$

From the boundary conditions we have in addition

$$v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1} = 0, \quad j,k = 0, 1, \ldots, m+1. \tag{4.4}$$

The equations (4.3) and (4.4) define a linear set of equations for the unknowns $\boldsymbol{V} = [v_{jk}] \in \mathbb{R}^{m,m}$.

Observe that (4.2) can be written as a matrix equation in the form

$$\boldsymbol{TV} + \boldsymbol{VT} = h^2 \boldsymbol{F} \quad \text{with} \quad h = 1/(m+1), \tag{4.5}$$

where $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m,m}$ is the second derivative matrix given by (2.3), $\boldsymbol{V} = (v_{jk}) \in \mathbb{R}^{m,m}$, and $\boldsymbol{F} = (f_{jk}) = (f(jh, kh)) \in \mathbb{R}^{m,m}$. Indeed, the (j,k) element in $\boldsymbol{TV} + \boldsymbol{VT}$ is given by

$$\sum_{i=1}^{m} \boldsymbol{T}_{j,i} v_{i,k} + \sum_{i=1}^{m} v_{j,i} \boldsymbol{T}_{i,k},$$

and this is precisely the left hand side of (4.2).

To write (4.3) and (4.4) in standard form $\boldsymbol{Ax} = \boldsymbol{b}$ we need to order the unknowns $v_{j,k}$ in some way. The following operation of **vectorization** of a matrix gives one possible ordering.

Definition 4.1 *For any $\boldsymbol{B} \in \mathbb{R}^{m,n}$ we define the vector*

$$vec(\boldsymbol{B}) := [b_{11}, \ldots, b_{m1}, b_{12}, \ldots, b_{m2}, \ldots, b_{1n}, \ldots, b_{mn}]^T \in \mathbb{R}^{mn}$$

by stacking the columns of $\boldsymbol{B}$ on top of each other.

Let $n = m^2$ and $\boldsymbol{x} := vec(\boldsymbol{V}) \in \mathbb{R}^n$. Note that forming $\boldsymbol{x}$ by stacking the columns of $\boldsymbol{V}$ on top of each other means an ordering of the grid points which for $m = 3$ is illustrated in Figure 4.1. We call this the **natural ordering**. The location of the elements in (4.3) form a 5-point stencil, as shown in Figure 4.2.

To find the matrix $\boldsymbol{A}$ we note that for values of j, k where the 5-point stencil does not touch the boundary, (4.3) takes the form

$$4x_i - x_{i-1} - x_{i+1} - x_{i-m} - x_{i+m} = b_i,$$

where $x_i = v_{jk}$ and $b_i = h^2 f_{jk}$. This must be modified close to the boundary. We obtain the linear system

$$\boldsymbol{Ax} = \boldsymbol{b}, \quad \boldsymbol{A} \in \mathbb{R}^{n,n}, \quad b \in \mathbb{R}^n, \quad n = m^2, \tag{4.6}$$

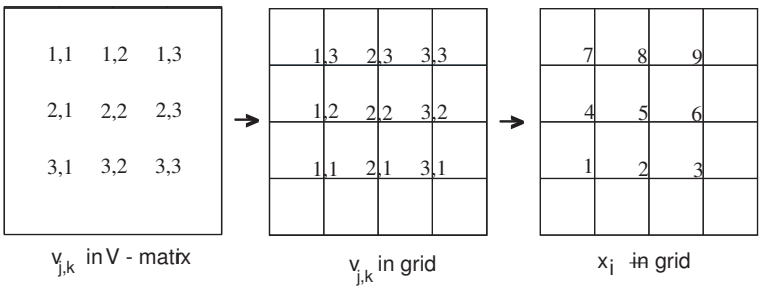


Figure 4.1. *Numbering of grid points*

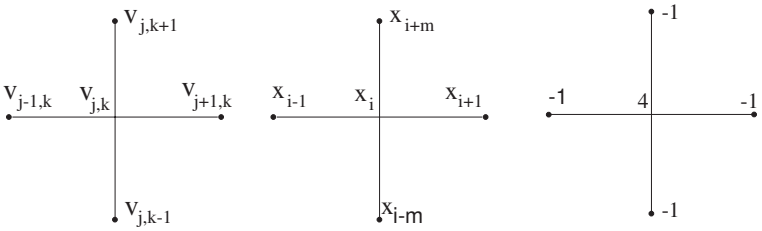


Figure 4.2. *The 5-point stencil*

where $\boldsymbol{x} = vec(\boldsymbol{V})$, $\quad \boldsymbol{b} = h^2 vec(\boldsymbol{F})$ with $\boldsymbol{F} = (f_{jk}) \in \mathbb{R}^{m,m}$ and $\boldsymbol{A}$ is the **Poisson matrix** given by

$$\begin{array}{rll} a_{ii} = 4, & i = 1, \ldots, n & \\ a_{i+1,i} = a_{i,i+1} = -1, & i = 1, \ldots, n-1, & i \neq m, 2m, \ldots, (m-1)m \\ a_{i+m,i} = a_{i,i+m} = -1, & i = 1, \ldots, n-m & \\ a_{ij} = 0, & \text{otherwise.} & \end{array} \tag{4.7}$$

For $m = 3$ we have the following matrix

$$\boldsymbol{A} = \begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix}.$$

Exercise 4.2 *Write down the Poisson matrix for $m = 2$ and show that it is strictly diagonally dominant.*

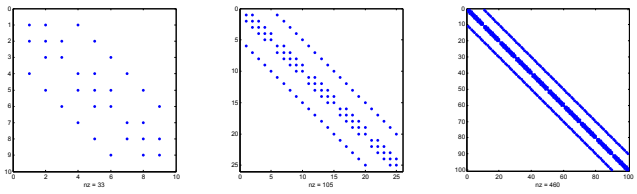


Figure 4.3. *Band structure of the 2D test matrix,* $n = 9$*,* $n = 25$*,* $n = 100$

4.1.2 The test Matrices

The second derivative matrix $\boldsymbol{T} = \text{tridiag}(-1, 2, -1)$ is a special case of the tridiagonal matrix

$$\boldsymbol{T}_1 := \begin{bmatrix} d & a & 0 & & & \\ a & d & a & & & \\ 0 & \ddots & \ddots & \ddots & & \\ & & & & & 0 \\ & & & a & d & a \\ & & & 0 & a & d \end{bmatrix}, \tag{4.8}$$

where $a, d \in \mathbb{R}$. We call this the **1D test matrix**. It is strictly diagonally dominant if $|d| > 2|a|$.

The (2 dimensional) Poisson matrix is a special case of the matrix $\boldsymbol{T}_2 = [a_{ij}] \in \mathbb{R}^{n,n}$ with elements

$$\begin{aligned} a_{i,i+1} = a_{i+1,i} &= a, \quad i = 1, \ldots, n-1, \quad i \neq m, 2m, \ldots, (m-1)m, \\ a_{i,i+m} = a_{i+m,i} &= a, \quad i = 1, \ldots, n-m, \\ a_{i,i} &= 2d, \quad i = 1, \ldots, n, \\ a_{i,j} &= 0, \quad \text{otherwise}, \end{aligned} \tag{4.9}$$

and where a, d are real numbers. We will refer to this matrix as simply the **2D test matrix**. The 2D test matrix is

- symmetric,
- a banded matrix with bandwidth $m = \sqrt{n}$, (Cf. Figure 4.3).
- strictly diagonally dominant if $|d| > 2|a|$,
- the Poisson matrix given by (4.7) when $a = -1$ and $d = 2$, This matrix is strictly diagonally dominant for $m = 2, n = 4$, but only diagonally dominant for $m > 2$.

- called the **averaging matrix** when $a = 1/9$ and $d = 5/18$. This matrix is strictly diagonally dominant for all n.

Properties of $\boldsymbol{T}_2$ can be derived from properties of $\boldsymbol{T}_1$ by using properties of the Kronecker product.

4.2 The Kronecker Product

Definition 4.3 *For any positive integers p, q, r, s we define the Kronecker product of two matrices $\boldsymbol{A} \in \mathbb{R}^{p,q}$ and $\boldsymbol{B} \in \mathbb{R}^{r,s}$ as a matrix $\boldsymbol{C} \in \mathbb{R}^{pr,qs}$ given in block form as*

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{A}b_{1,1} & \boldsymbol{A}b_{1,2} & \cdots & \boldsymbol{A}b_{1,s} \\ \boldsymbol{A}b_{2,1} & \boldsymbol{A}b_{2,2} & \cdots & \boldsymbol{A}b_{2,s} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{A}b_{r,1} & \boldsymbol{A}b_{r,2} & \cdots & \boldsymbol{A}b_{r,s} \end{bmatrix}.$$

We denote the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ by $\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B}$.

This definition of the Kronecker product is known more precisely as the *left Kronecker product*. In the literature one often finds the *right Kronecker product* which in our notation is given by $\boldsymbol{B} \otimes \boldsymbol{A}$.

As examples of Kronecker products which are relevant for our discussion, if

$$\boldsymbol{T}_1 = \begin{bmatrix} d & a \\ a & d \end{bmatrix} \quad \text{and} \quad \boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

then

$$\boldsymbol{T}_1 \otimes \boldsymbol{I} = \left[\begin{array}{cc|cc} d & a & 0 & 0 \\ a & d & 0 & 0 \\ \hline 0 & 0 & d & a \\ 0 & 0 & a & d \end{array}\right] \quad \text{and} \quad \boldsymbol{I} \otimes \boldsymbol{T}_1 = \left[\begin{array}{cc|cc} d & 0 & a & 0 \\ 0 & d & 0 & a \\ \hline a & 0 & d & 0 \\ 0 & a & 0 & d \end{array}\right].$$

Also note that the Kronecker product $\boldsymbol{u} \otimes \boldsymbol{v} = \left[\boldsymbol{u}^T v_1, \ldots, \boldsymbol{u}^T v_r\right]^T$ of two column vectors $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^r$ is a column vector of length $p \times r$.

The 2D test matrix $\boldsymbol{T}_2$ can be written as a sum of two Kronecker products. We see that

$$\boldsymbol{T}_2 = \begin{bmatrix} \boldsymbol{T}_1 & & & & \\ & \boldsymbol{T}_1 & & & \\ & & \ddots & & \\ & & & \boldsymbol{T}_1 & \\ & & & & \boldsymbol{T}_1 \end{bmatrix} + \begin{bmatrix} d\boldsymbol{I} & a\boldsymbol{I} & & & \\ a\boldsymbol{I} & d\boldsymbol{I} & a\boldsymbol{I} & & \\ & \ddots & \ddots & \ddots & \\ & & a\boldsymbol{I} & d\boldsymbol{I} & a\boldsymbol{I} \\ & & & a\boldsymbol{I} & d\boldsymbol{I} \end{bmatrix} = \boldsymbol{T}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}_1.$$

Definition 4.4 *Let for positive integers r, s, k, $\boldsymbol{A} \in \mathbb{R}^{r,r}$, $\boldsymbol{B} \in \mathbb{R}^{s,s}$ and $\boldsymbol{I}_k$ be the identity matrix of order k. The sum $\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B}$ is known as the* **Kronecker sum** *of $\boldsymbol{A}$ and $\boldsymbol{B}$.*

In other words, the 2D test matrix is the Kronecker sum of two identical 1D test matrices.

The following simple arithmetic rules hold for Kronecker products. For scalars λ, μ and matrices $\boldsymbol{A}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{C}$ of dimensions such that the operations are defined we have

$$\begin{aligned} \big(\lambda \boldsymbol{A}\big) \otimes \big(\mu \boldsymbol{B}\big) &= \lambda\mu\big(\boldsymbol{A} \otimes \boldsymbol{B}\big), \\ \big(\boldsymbol{A}_1 + \boldsymbol{A}_2\big) \otimes \boldsymbol{B} &= \boldsymbol{A}_1 \otimes \boldsymbol{B} + \boldsymbol{A}_2 \otimes \boldsymbol{B}, \\ \boldsymbol{A} \otimes \big(\boldsymbol{B}_1 + \boldsymbol{B}_2\big) &= \boldsymbol{A} \otimes \boldsymbol{B}_1 + \boldsymbol{A} \otimes \boldsymbol{B}_2, \\ (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes C &= \boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}), \\ (\boldsymbol{A} \otimes \boldsymbol{B})^T &= \boldsymbol{A}^T \otimes \boldsymbol{B}^T. \end{aligned} \tag{4.10}$$

Note however that in general we have $\boldsymbol{A} \otimes \boldsymbol{B} \neq \boldsymbol{B} \otimes \boldsymbol{A}$, but it can be shown that there are permutation matrices $\boldsymbol{P}, Q$ such that $\boldsymbol{B} \otimes \boldsymbol{A} = \boldsymbol{P}(\boldsymbol{A} \otimes \boldsymbol{B})Q$, see [10].

Exercise 4.5 Prove (4.10).

The following *mixed product rule* is an essential tool for dealing with Kronecker products and sums.

Lemma 4.6 *Suppose $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are rectangular matrices with dimensions so that the products $\boldsymbol{AC}$ and $\boldsymbol{BD}$ are defined. Then the product $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D})$ is defined and*

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD}). \tag{4.11}$$

Proof. If $\boldsymbol{B} \in \mathbb{R}^{r,t}$ and $\boldsymbol{D} \in \mathbb{R}^{t,s}$ for some integers r, s, t then

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \begin{bmatrix} \boldsymbol{A}b_{1,1} & \cdots & \boldsymbol{A}b_{1,t} \\ \vdots & & \vdots \\ \boldsymbol{A}b_{r,1} & \cdots & \boldsymbol{A}b_{r,t} \end{bmatrix} \begin{bmatrix} \boldsymbol{C}d_{1,1} & \cdots & \boldsymbol{C}d_{1,s} \\ \vdots & & \vdots \\ \boldsymbol{C}d_{t,1} & \cdots & \boldsymbol{C}d_{t,s} \end{bmatrix}.$$

Thus for all i, j

$$((\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}))_{i,j} = \boldsymbol{AC} \sum_{k=1}^{t} b_{i,k} d_{k,j} = (\boldsymbol{AC})(\boldsymbol{BD})_{i,j} = ((\boldsymbol{AC}) \otimes (\boldsymbol{BD}))_{i,j}.$$

□

The eigenvalues and eigenvectors of a Kronecker product can easily be determined if one knows the corresponding quantities for each of the factors in the product.

Lemma 4.7 *Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are square matrices. Then the eigenvalues of $\boldsymbol{A} \otimes \boldsymbol{B}$ are products of eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$, and the eigenvectors of $\boldsymbol{A} \otimes \boldsymbol{B}$ are Kronecker products of eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$. More precisely, if $\boldsymbol{A} \in \mathbb{R}^{r,r}$ and $\boldsymbol{B} \in \mathbb{R}^{s,s}$ and*

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \quad i = 1, \ldots, r, \quad \boldsymbol{B}\boldsymbol{v}_j = \mu_j \boldsymbol{v}_j, \quad j = 1, \ldots, s,$$

then

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = \lambda_i \mu_j (\boldsymbol{u}_i \otimes \boldsymbol{v}_j), \quad i = 1, \ldots, r, \quad j = 1, \ldots, s. \tag{4.12}$$

Proof. Using (4.10) and (4.11) the proof is a one liner. For all i, j

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = (\boldsymbol{A}\boldsymbol{u}_i) \otimes (\boldsymbol{B}\boldsymbol{v}_j) = (\lambda_i \boldsymbol{u}_i) \otimes (\mu_j \boldsymbol{v}_j) = (\lambda_i \mu_j)(\boldsymbol{u}_i \otimes \boldsymbol{v}_j).$$

□

Consider next a Kronecker sum.

Lemma 4.8 *For positive integers r, s let $\boldsymbol{A} \in \mathbb{R}^{r,r}$ and $\boldsymbol{B} \in \mathbb{R}^{s,s}$. Then the eigenvalues of the Kronecker sum $\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B}$ are all sums of eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$, and the eigenvectors of $\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B}$ are all Kronecker products of eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$. More precisely, if*

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \quad i = 1, \ldots, r, \quad \boldsymbol{B}\boldsymbol{v}_j = \mu_j \boldsymbol{v}_j, \quad j = 1, \ldots, s,$$

then

$$(\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B})(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = (\lambda_i + \mu_j)(\boldsymbol{u}_i \otimes \boldsymbol{v}_j), \quad i = 1, \ldots, r, \quad j = 1, \ldots, s. \tag{4.13}$$

Proof. Since $\boldsymbol{I}_s \boldsymbol{v}_j = \boldsymbol{v}_j$ for $j = 1, \ldots, s$ and $\boldsymbol{I}_r \boldsymbol{u}_i = \boldsymbol{u}_i$ for $i = 1, \ldots, r$ we obtain by Lemma 4.7 for all i, j

$$(\boldsymbol{A} \otimes \boldsymbol{I}_s)(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = \lambda_i (\boldsymbol{u}_i \otimes \boldsymbol{v}_j), \quad \text{and} \quad (\boldsymbol{I}_r \otimes \boldsymbol{B})(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = \mu_j (\boldsymbol{u}_i \otimes \boldsymbol{v}_j).$$

The result now follows by summing these relations. □

In many cases the Kronecker product and sum inherit properties of their factors.

Lemma 4.9

1. *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular then $\boldsymbol{A} \otimes \boldsymbol{B}$ is nonsingular. Moreover $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$.*
2. *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric then $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{A} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{B}$ are symmetric.*
3. *If one of $\boldsymbol{A}$, $\boldsymbol{B}$ is symmetric positive definite and the other is symmetric positive semidefinite then $\boldsymbol{A} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{B}$ is symmetric positive definite.*

Proof. Suppose that $\boldsymbol{A} \in \mathbb{R}^{r,r}$ and $\boldsymbol{B} \in \mathbb{R}^{s,s}$. 1. follows from the mixed product rule giving

$$\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)\big(\boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}\big) = \big(\boldsymbol{A}\boldsymbol{A}^{-1}\big) \otimes \big(\boldsymbol{B}\boldsymbol{B}^{-1}\big) = \boldsymbol{I}_r \otimes \boldsymbol{I}_s = \boldsymbol{I}_{rs}.$$

Thus $\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)$ is nonsingular with the indicated inverse. 2. and the symmetry part of 3. follow immediately from (4.10). Suppose $\boldsymbol{A}$ is positive definite and $\boldsymbol{B}$

is positive semidefinite. Then $\boldsymbol{A}$ has positive eigenvalues and $\boldsymbol{B}$ has nonnegative eigenvalues. By Lemma 4.8 the eigenvalues of $\boldsymbol{A} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{B}$ are all positive and 3. follows. □

In (4.5) we derived the matrix equation $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$ for the unknowns $\boldsymbol{V}$ in the discrete Poisson problem. With some effort we converted this matrix equation to a linear system in standard form $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} = \boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}$, $\boldsymbol{x} = \text{vec}(\boldsymbol{V})$, and $\boldsymbol{b} = \text{vec}(\boldsymbol{F})$. This conversion could have been carried out with less effort using the following result.

Lemma 4.10 *Suppose $\boldsymbol{A} \in \mathbb{R}^{r,r}$, $\boldsymbol{B} \in \mathbb{R}^{s,s}$, and $\boldsymbol{F}, \boldsymbol{V} \in \mathbb{R}^{r,s}$. Then we have*

$$(\boldsymbol{A} \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{V}) = \text{vec}(\boldsymbol{F}) \quad \Leftrightarrow \quad \boldsymbol{AVB}^T = \boldsymbol{F}, \tag{4.14}$$

$$(\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{V}) = \text{vec}(\boldsymbol{F}) \quad \Leftrightarrow \quad \boldsymbol{AV} + \boldsymbol{VB}^T = \boldsymbol{F}. \tag{4.15}$$

Proof. We partition $\boldsymbol{V}$, $\boldsymbol{F}$, and $\boldsymbol{B}^T$ by columns as $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s]$, $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_s]$ and $\boldsymbol{B}^T = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_s]$. Then we have

$$\begin{aligned}
&(\boldsymbol{A} \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{V}) = \text{vec}(\boldsymbol{F}) \\
&\qquad \Leftrightarrow \quad \begin{bmatrix} \boldsymbol{A}b_{11} & \cdots & \boldsymbol{A}b_{1s} \\ \vdots & & \vdots \\ \boldsymbol{A}b_{s1} & \cdots & \boldsymbol{A}b_{ss} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_s \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_1 \\ \vdots \\ \boldsymbol{f}_s \end{bmatrix} \\
&\qquad \Leftrightarrow \quad \boldsymbol{A}[\sum_j b_{1j}\boldsymbol{v}_j, \ldots, \sum_j b_{sj}\boldsymbol{v}_j] = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_s] \\
&\qquad \Leftrightarrow \quad \boldsymbol{A}[\boldsymbol{Vb}_1, \ldots, \boldsymbol{Vb}_s] = \boldsymbol{F} \quad \Leftrightarrow \quad \boldsymbol{AVB}^T = \boldsymbol{F}.
\end{aligned}$$

This proves (4.14). (4.15) follows immediately from (4.14) as follows

$$\begin{aligned}
&(\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{V}) = \text{vec}(\boldsymbol{F}) \\
&\quad \Leftrightarrow \quad (\boldsymbol{AVI}_s^T + \boldsymbol{I}_r\boldsymbol{VB}^T) = \boldsymbol{F} \quad \Leftrightarrow \quad \boldsymbol{AV} + \boldsymbol{VB}^T = \boldsymbol{F}.
\end{aligned}$$

□

For more on Kronecker products see [10].

4.3 Properties of the 1D and 2D Test Matrices

We can apply these results to the 2D test matrix $\boldsymbol{T}_2$. We first consider the 1D test matrix. The eigenvectors of $\boldsymbol{T}_1$ are the columns of the sine matrix defined by

$$\boldsymbol{S} = \big[\sin\frac{jk\pi}{m+1}\big]_{j,k=1}^{m} \in \mathbb{R}^{m,m}. \tag{4.16}$$

For $m = 3$

$$\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3] = \begin{bmatrix} \sin\frac{\pi}{4} & \sin\frac{2\pi}{4} & \sin\frac{3\pi}{4} \\ \sin\frac{2\pi}{4} & \sin\frac{4\pi}{4} & \sin\frac{6\pi}{4} \\ \sin\frac{3\pi}{4} & \sin\frac{6\pi}{4} & \sin\frac{9\pi}{4} \end{bmatrix} = \begin{bmatrix} t & 1 & t \\ 1 & 0 & -1 \\ t & -1 & t \end{bmatrix}, \; t := \frac{1}{\sqrt{2}}.$$

Lemma 4.11 *Suppose $\boldsymbol{T}_1 = (t_{kj})_{k,j} = \text{tridiag}(a, d, a) \in \mathbb{R}^{m,m}$ with $m \geq 2$, $a, d \in \mathbb{R}$, and let $h = 1/(m+1)$.*

1. *We have $\boldsymbol{T}_1 \boldsymbol{s}_j = \lambda_j \boldsymbol{s}_j$ for $j = 1, \ldots, m$, where*

$$\boldsymbol{s}_j = [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^T, \tag{4.17}$$
$$\lambda_j = d + 2a\cos(j\pi h). \tag{4.18}$$

2. *The eigenvalues are distinct and the eigenvectors are orthogonal*

$$\boldsymbol{s}_j^T \boldsymbol{s}_k = \frac{1}{2h}\delta_{j,k}, \quad j, k = 1, \ldots, m. \tag{4.19}$$

Proof. We find

$$\begin{aligned}(\boldsymbol{T}_1 \boldsymbol{s}_j)_k &= \sum_{l=1}^{m} t_{k,l} \sin\big(lj\pi h\big) = a\big[\sin\big((k-1)j\pi h\big) + \sin\big((k+1)j\pi h\big)\big] + d\sin\big(kj\pi h\big) \\ &= \big(d + 2a\cos(j\pi h)\big)\sin\big(kj\pi h\big) = \lambda_j s_{k,j},\end{aligned}$$

and 1. follows. Since $j\pi h = j\pi/(m+1) \in (0, \pi)$ for $j = 1, \ldots, m$ and the cosine function is strictly monotone decreasing on $(0, \pi)$ the eigenvalues are distinct, and since $\boldsymbol{T}_1$ is symmetric it follows from Lemma 4.12 below that the eigenvectors $\boldsymbol{s}_j$ are orthogonal. To finish the proof of (4.19) we compute the square of the Euclidian norm of each $\boldsymbol{s}_j$ as follows:

$$\begin{aligned}\boldsymbol{s}_j^T \boldsymbol{s}_j &= \sum_{k=1}^{m} \sin^2(kj\pi h) = \sum_{k=0}^{m} \sin^2(kj\pi h) = \frac{1}{2}\sum_{k=0}^{m}(1 - \cos(2kj\pi h)) \\ &= \frac{m+1}{2} - \frac{1}{2}\sum_{k=0}^{m}\cos(2kj\pi h)) = \frac{m+1}{2},\end{aligned}$$

since the last cosine sum is zero. We show this by summing a geometric series of complex exponentials. With $i = \sqrt{-1}$ we find

$$\sum_{k=0}^{m}\cos(2kj\pi h) + i\sum_{k=0}^{m}\sin(2kj\pi h) = \sum_{k=0}^{m} e^{2ikj\pi h} = \frac{e^{2i(m+1)j\pi h} - 1}{e^{2ij\pi h} - 1} = \frac{e^{2ij\pi} - 1}{e^{2ij\pi h} - 1} = 0,$$

and (4.19) follows. ∎

Lemma 4.12 *The eigenvalues of a Hermitian matrix are real. Moreover, eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Proof. Suppose $\boldsymbol{A}^* = \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\boldsymbol{x} \neq 0$. We multiply both sides of $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ by $\boldsymbol{x}^*$ and divide by $\boldsymbol{x}^*\boldsymbol{x}$ to obtain $\lambda = \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}$. Taking complex conjugates we find $\overline{\lambda} = \lambda^* = \frac{(\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x})^*}{(\boldsymbol{x}^*\boldsymbol{x})^*} = \frac{\boldsymbol{x}^*\boldsymbol{A}^*\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \lambda$, and λ is real.

Suppose in addition that $(\mu, \boldsymbol{y})$ is another eigenpair for $\boldsymbol{A}$ with $\mu \neq \lambda$. Multiplying $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ by $\boldsymbol{y}^*$ gives

$$\lambda\boldsymbol{y}^*\boldsymbol{x} = \boldsymbol{y}^*\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{x}^*\boldsymbol{A}^*\boldsymbol{y})^* = (\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{y})^* = (\mu\boldsymbol{x}^*\boldsymbol{y})^* = \mu\boldsymbol{y}^*\boldsymbol{x},$$

using that μ is real. Since $\lambda \neq \mu$ it follows that $\boldsymbol{y}^*\boldsymbol{x} = 0$ which means that $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal. ∎

It is now easy to find the eigenpairs of the 2D test matrix and determine when it is positive definite.

Theorem 4.13 *For fixed $m \geq 2$ let $\boldsymbol{T}_2$ be the matrix given by* (4.9) *and let $h = 1/(m+1)$.*

1. *We have $\boldsymbol{T}_2\boldsymbol{x}_{j,k} = \lambda_{j,k}\boldsymbol{x}_{j,k}$ for $j,k = 1,\ldots,m$, where*

$$\boldsymbol{x}_{j,k} = \boldsymbol{s}_j \otimes \boldsymbol{s}_k, \tag{4.20}$$
$$\boldsymbol{s}_j = [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^T, \tag{4.21}$$
$$\lambda_{j,k} = 2d + 2a\cos(j\pi h) + 2a\cos(k\pi h). \tag{4.22}$$

2. *The eigenvectors are orthogonal*

$$\boldsymbol{x}_{j,k}^T\boldsymbol{x}_{p,q} = \frac{1}{4h^2}\delta_{j,p}\delta_{k,q}, \quad j,k,p,q = 1,\ldots,m. \tag{4.23}$$

3. *$\boldsymbol{T}_2$ is symmetric positive definite if $d > 0$ and $d \geq 2|a|$.*

4. *The Poisson and averaging matrix are symmetric positive definite.*

Proof.

1. follows from Lemma 4.11 and Lemma 4.8 since $\boldsymbol{T}_2 = \boldsymbol{T}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}_1$. Using the transpose rule, the mixed product rule and (4.19) we find for $j,k,p,q = 1,\ldots,m$

$$\big(\boldsymbol{s}_j \otimes \boldsymbol{s}_k\big)^T\big(\boldsymbol{s}_p \otimes \boldsymbol{s}_q\big) = \big(\boldsymbol{s}_j^T \otimes \boldsymbol{s}_k^T\big)\big(\boldsymbol{s}_p \otimes \boldsymbol{s}_q\big) = \big(\boldsymbol{s}_j^T\boldsymbol{s}_p\big) \otimes \big(\boldsymbol{s}_k^T\boldsymbol{s}_q\big) = \frac{1}{4h^2}\delta_{j,p}\delta_{k,q}$$

and 2. follows. Since $\boldsymbol{T}_2$ is symmetric 3. will follow if the eigenvalues are positive. But this is true if $d > 0$ and $d \geq 2|a|$ and this holds both for both choices $a = -1$, $d = 2$ and $a = 1/5$, $d = 5/18$. Thus the matrices in 4. are positive definite. ∎

Exercise 4.14 *Write down the eigenvalues of $\boldsymbol{T} = \text{tridiag}(-1, 2, -1)$ using Lemma 4.11 and conclude that $\boldsymbol{T}$ is symmetric positive definite.*

Exercise 4.15 *Use Lemma 4.11 to show that the matrix* $\boldsymbol{T}_1 := tridiag(a, d, a) \in \mathbb{R}^{n,n}$ *is symmetric positive definite if* $d > 0$ *and* $d \geq 2|a|$.

Exercise 4.16 *For* $m = 2$ *the matrix* (4.9) *is given by*

$$\boldsymbol{A} = \begin{bmatrix} 2d & a & a & 0 \\ a & 2d & 0 & a \\ a & 0 & 2d & a \\ 0 & a & a & 2d \end{bmatrix}.$$

Show that $\lambda = 2a + 2d$ *is an eigenvalue corresponding to the eigenvector* $x = [1, 1, 1, 1]^T$. *Verify that apart from a scaling of the eigenvector this agrees with* (4.22) *and* (4.21) *for* $j = k = 1$ *and* $m = 2$.

Exercise 4.17 Consider the following 9 point difference approximation to the Poisson problem $-\nabla^2 u = f$, $u = 0$ on the boundary of the unit square (cf. (4.1))

$$\begin{array}{lrcl} \text{(a)} & -(\square_h v)_{j,k} & = & (\mu f)_{j,k} \quad j,k = 1,\ldots,m \\ \text{(b)} & v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1} & = & 0, \qquad j,k = 0,1,\ldots,m+1, \\ \text{(c)} & -(\square_h v)_{j,k} & = & [20v_{j,k} - 4v_{j-1,k} - 4v_{j,k-1} - 4v_{j+1,k} - 4v_{j,k+1} \\ & & & \quad - v_{j-1,k-1} - v_{j+1,k-1} - v_{j-1,k+1} - v_{j+1,k+1}]/(6h^2), \\ \text{(d)} & (\mu f)_{j,k} & = & [8f_{j,k} + f_{j-1,k} + f_{j,k-1} + f_{j+1,k} + f_{j,k+1}]/12. \end{array} \tag{4.24}$$

a) Write down the 4-by-4 system we obtain for $m = 2$.

b) Find $v_{j,k}$ for $j, k = 1, 2$, if $f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y)$ and $m = 2$. Answer: $v_{j,k} = 5\pi^2/66$.

It can be shown that (4.24) defines an $O(h^4)$ approximation to (4.1).

Exercise 4.18 Consider the nine point difference approximation to (4.1) given by (4.24) in Problem 4.17.

a) Show that (4.24) is equivalent to the matrix equation

$$\boldsymbol{T}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T} - \frac{1}{6}\boldsymbol{T}\boldsymbol{V}\boldsymbol{T} = h^2\mu\boldsymbol{F}. \tag{4.25}$$

Here $\mu\boldsymbol{F}$ has elements $(\mu f)_{j,k}$ given by (4.24d).

b) Show that the standard form of the matrix equation (4.25) is $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} = \boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T} - \frac{1}{6}\boldsymbol{T} \otimes \boldsymbol{T}$, $\boldsymbol{x} = \text{vec}(\boldsymbol{V})$, and $\boldsymbol{b} = h^2\text{vec}(\mu\boldsymbol{F})$.

Exercise 4.19 Consider the biharmonic equation

$$\begin{array}{ll} \nabla^4 u(s,t) = \nabla^2(\nabla^2 u(s,t)) = f(s,t) & (s,t) \in \Omega, \\ u(s,t) = 0, \quad \nabla^2 u(s,t) = 0 & (s,t) \in \partial\Omega. \end{array} \tag{4.26}$$

Here Ω is the open unit square. The condition $\nabla^2 u = 0$ is called the *Navier boundary condition*. Moreover, $\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy}$.

a) Let $v = -\nabla^2 u$. Show that (4.26) can be written as a system

$$\begin{array}{rcll} -\nabla^2 v(s,t) & = & f(s,t) & (s,t) \in \Omega \\ -\nabla^2 u(s,t) & = & v(s,t) & (s,t) \in \Omega \\ u(s,t) & = & v(s,t) = 0 & (s,t) \in \partial\Omega. \end{array} \tag{4.27}$$

b) Discretizing, using (4.2), with $\boldsymbol{T} = \mathrm{diag}(-1, 2, -1) \in \mathbb{R}^{\mathrm{m},\mathrm{m}}$, $h = 1/(m+1)$, and $\boldsymbol{F} = (f(jh, kh))_{j,k=1}^{m}$ we get two matrix equations

$$\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}, \qquad \boldsymbol{TU} + \boldsymbol{UT} = h^2\boldsymbol{V}.$$

Show that

$$(\boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T})\mathrm{vec}(\boldsymbol{V}) = h^2\mathrm{vec}(\boldsymbol{F}), \quad (\boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T})\mathrm{vec}(\boldsymbol{U}) = h^2\mathrm{vec}(\boldsymbol{V}).$$

and hence $\boldsymbol{A} = (\boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T})^2$ is the matrix for the standard form of the discrete biharmonic equation.

c) Show that with $n = m^2$ the vector form and standard form of the systems in b) can be written

$$\boldsymbol{T}^2\boldsymbol{U} + 2\boldsymbol{TUT} + \boldsymbol{UT}^2 = h^4\boldsymbol{F} \quad \text{and} \quad \boldsymbol{Ax} = \boldsymbol{b}, \tag{4.28}$$

where $\boldsymbol{A} = \boldsymbol{T}^2 \otimes \boldsymbol{I} + 2\boldsymbol{T} \otimes \boldsymbol{T} + \boldsymbol{I} \otimes \boldsymbol{T}^2 \in \mathbb{R}^{n,n}$, $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{U})$, and $\boldsymbol{b} = h^4\,\mathrm{vec}(\boldsymbol{F})$.

d) Determine the eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ in c) and show that it is symmetric positive definite. Also determine the bandwidth of $\boldsymbol{A}$.

e) Suppose we want to solve the standard form equation $\boldsymbol{Ax} = \boldsymbol{b}$. We have two representations for the matrix $\boldsymbol{A}$, the product one in b) and the one in c). Which one would you prefer for a basis of an algorithm? Why?

Chapter 5

Fast Direct Solution of a Large Linear System

5.1 Algorithms for a Banded Positive Definite System

In this chapter we present a fast method for solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is the Poisson matrix (4.7). Thus for $n = 3$

$$\boldsymbol{A} = \left[\begin{array}{rrr|rrr|rrr} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ \hline -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ \hline 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{array}\right] = \left[\begin{array}{ccc} \boldsymbol{T}+2\boldsymbol{I} & -\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{T}+2\boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{T}+2\boldsymbol{I} \end{array}\right],$$

where $\boldsymbol{T} = \text{tridiag}(-1, 2, -1)$. For this matrix we know by now that

1. It is symmetric positive definite.
2. It is banded.
3. It is block-tridiagonal.
4. We know the eigenvalues and eigenvectors of $\boldsymbol{A}$.
5. The eigenvectors are orthogonal.

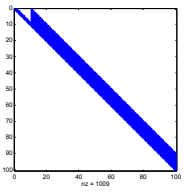


Figure 5.1. *Fill-inn in the Cholesky factor of the Poisson matrix ($n = 100$).*

5.1.1 Cholesky Factorization

Since $\boldsymbol{A}$ is symmetric positive definite we can use the Cholesky factorization Algorithm 3.37. Since $\boldsymbol{A}$ is banded with bandwidth $d = \sqrt{n}$ the complexity of this factorization is $O(nd^2) = O(n^2)$. We need to store $\boldsymbol{A}$ possibly in sparse form.

The nonzero elements in $\boldsymbol{R}$ are shown in Figure 5.1. Note that the zeros between the diagonals in $\boldsymbol{A}$ have become nonzero in $\boldsymbol{R}$. This is known as **fill-inn**.

5.1.2 Block LU Factorization of a Block Tridiagonal Matrix

The Poisson matrix has a block tridiagonal structure. Consider finding the block LU factorization of a block tridiagonal matrix. We are looking for a factorization of the form

$$\begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{C}_1 & & & \\ \boldsymbol{A}_2 & \boldsymbol{D}_2 & \boldsymbol{C}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{A}_{m-1} & \boldsymbol{D}_{m-1} & \boldsymbol{C}_{m-1} \\ & & & \boldsymbol{A}_m & \boldsymbol{D}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & & & \\ \boldsymbol{L}_2 & \boldsymbol{I} & & \\ & \ddots & \ddots & \\ & & \boldsymbol{L}_m & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_1 & \boldsymbol{C}_1 & & \\ & \ddots & \ddots & \\ & & \boldsymbol{R}_{m-1} & \boldsymbol{C}_{m-1} \\ & & & \boldsymbol{R}_m \end{bmatrix}. \tag{5.1}$$

Here $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_m$ and $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_m$ are square matrices while $\boldsymbol{A}_2, \ldots, \boldsymbol{A}_m$ and $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_{m-1}$ can be rectangular.

Using block multiplication the formulas (2.5) generalize to

$$\boldsymbol{R}_1 = \boldsymbol{D}_1, \quad \boldsymbol{L}_k = \boldsymbol{A}_k \boldsymbol{R}_{k-1}^{-1}, \quad \boldsymbol{R}_k = \boldsymbol{D}_k - \boldsymbol{L}_k \boldsymbol{C}_{k-1}, \quad k = 2, 3, \ldots, m. \tag{5.2}$$

To solve the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ we partition $\boldsymbol{b}$ conformaly with $\boldsymbol{A}$ in the form $\boldsymbol{b}^T = [\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_m^T]$. The formulas for solving $\boldsymbol{L}\boldsymbol{y} = \boldsymbol{b}$ and $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$ are as follows:

$$\begin{aligned} &\boldsymbol{y}_1 = \boldsymbol{b}_1, && \boldsymbol{y}_k = \boldsymbol{b}_k - \boldsymbol{L}_k \boldsymbol{y}_{k-1}, && k = 2, 3, \ldots, m, \\ &\boldsymbol{x}_m = \boldsymbol{R}_m^{-1} \boldsymbol{y}_m, && \boldsymbol{x}_k = \boldsymbol{R}_k^{-1}(\boldsymbol{y}_k - \boldsymbol{C}_k \boldsymbol{x}_{k+1}), && k = m-1, \ldots, 2, 1. \end{aligned} \tag{5.3}$$

The solution is then $\boldsymbol{x}^T = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T]$. To find $\boldsymbol{L}_k$ in (5.2) we solve the linear systems $\boldsymbol{L}_k \boldsymbol{R}_{k-1} = \boldsymbol{A}_k$. Similarly we need to solve a linear system to find $\boldsymbol{x}_k$ in (5.3).

The number of arithmetic operations using block factorizations is $O(n^2)$, asymptotically the same as for Cholesky factorization. However we only need to store the $m \times m$ blocks and using matrix operations can be an advantage.

5.1.3 Other Methods

Other methods include

- Iterative methods. We study this in Chapters 9, 10, 11.
- Multigrid. See [5].
- Fast solvers based on diagonalization and the Fast Fourier Transform. See Sections 5.2, 5.3.

5.2 A Fast Poisson Solver based on Diagonalization

The algorithm we now derive will only require $O(n^{3/2})$ flops and we only need to work with matrices of order m. Using the Fast Fourier Transform the number of flops can be reduced further to $O(n \log n)$.

To start we recall that $\boldsymbol{Ax} = \boldsymbol{b}$ can be written as a matrix equation in the form (cf. (4.5))

$$\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F} \quad \text{with} \quad h = 1/(m+1),$$

where $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m,m}$ is the second derivative matrix, $\boldsymbol{V} = (v_{jk}) \in \mathbb{R}^{m,m}$ are the unknowns, and $\boldsymbol{F} = (f_{jk}) = (f(jh, kh)) \in \mathbb{R}^{m,m}$ contains function values.

Recall that the eigenpairs of $\boldsymbol{T}$ are given by

$$\begin{aligned}
\boldsymbol{T}\boldsymbol{s}_j &= \lambda_j \boldsymbol{s}_j, \quad j = 1, \ldots, m, \\
\boldsymbol{s}_j &= [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^T, \\
\lambda_j &= 2 - 2\cos(j\pi h) = 4\sin^2(j\pi h/2), \quad h = 1/(m+1), \\
\boldsymbol{s}_j^T \boldsymbol{s}_k &= \delta_{jk}/(2h) \text{ for all } j, k.
\end{aligned}$$

Let

$$\boldsymbol{S} := [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m] = \left[\sin(jk\pi h)\right]_{j,k=1}^{m} \in \mathbb{R}^{m,m}, \quad \boldsymbol{D} = \text{diag}(\lambda_1, \ldots, \lambda_m). \tag{5.4}$$

Then $\boldsymbol{TS} = \boldsymbol{SD}$ and $\boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{S}^2 = \boldsymbol{I}/(2h)$. Define $\boldsymbol{X} \in \mathbb{R}^{m,m}$ by $\boldsymbol{V} = \boldsymbol{SXS}$, where $\boldsymbol{V}$ is the solution of $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$. Then

$$\begin{aligned}
&\boldsymbol{TV} + \boldsymbol{VT} = h^2 F \\
\overset{\boldsymbol{V}=\boldsymbol{SXS}}{\Longleftrightarrow}\ &\boldsymbol{TSXS} + \boldsymbol{SXST} = h^2\boldsymbol{F} \\
\overset{\boldsymbol{S}(\ )\boldsymbol{S}}{\Longleftrightarrow}\ &\boldsymbol{STSXS}^2 + \boldsymbol{S}^2\boldsymbol{XSTS} = h^2\boldsymbol{SFS} \\
\overset{\boldsymbol{TS}=\boldsymbol{SD}}{\Longleftrightarrow}\ &\boldsymbol{S}^2\boldsymbol{DXS}^2 + \boldsymbol{S}^2\boldsymbol{XS}^2\boldsymbol{D} = h^2\boldsymbol{SFS} \\
\overset{\boldsymbol{S}^2=\boldsymbol{I}/(2h)}{\Longleftrightarrow}\ &\boldsymbol{DX} + \boldsymbol{XD} = 4h^4\boldsymbol{SFS}.
\end{aligned}$$

An equation of the form $\boldsymbol{DX} + \boldsymbol{XD} = \boldsymbol{B}$, where $\boldsymbol{D}$ is diagonal is easy to solve. If $\boldsymbol{D} = \text{diag}(\lambda_j)$ we obtain for each element the equation $\lambda_j x_{jk} + x_{jk}\lambda_k = b_{jk}$ so $x_{jk} = b_{jk}/(\lambda_j + \lambda_k)$ for all j, k.

We now get the following algorithm to find the exact solution of $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$.

Algorithm 5.1 (Fast Poisson Solver) We solve the Poisson problem $-\nabla^2 u = f$ on $\Omega = (0,1)^2$ and $u = 0$ on $\partial\Omega$ using the 5-point scheme, i. e., let $m \in \mathbb{N}$, $h = 1/(m+1)$, and $\boldsymbol{F} = (f(jh, kh)) \in \mathbb{R}^{m,m}$. We compute $\boldsymbol{V} \in \mathbb{R}^{m,m}$, where $v_{jk} \approx u(jh, kh)$ by solving the equation $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$ using diagonalization of $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m,m}$.

```
function V=fastpoisson(F)
m=length(F); h=1/(m+1); hv=pi*h*(1:m)';
sigma=sin(hv/2).^2;
S=sin(hv*(1:m));
G=S*F*S;
X=h^4*G./(sigma*ones(1,m)+ ones(m,1)*sigma');
V=zeros(m+2,m+2);
V(2:m+1,2:m+1)=S*X*S;
```

The formulas are fully vectorized and for convenience we have used $\sigma_j := \lambda_j/4$ instead of λ_j. Since the statement "X=h^4*G./(sigma*ones(1,m)+ ones(m,1)*sigma')" only requires $O(m^2)$ flops the complexity of this algorithm is for large m determined by the 4 m-by-m matrix multiplications and is given by $O(4 \times 2m^3) = O(8n^{3/2})$. [2]

5.3 A Fast Poisson Solver based on the Discrete Sine and Fourier Transforms

In Algorithm 5.1 we need to compute the product of the sine matrix $\boldsymbol{S} \in \mathbb{R}^{m,m}$ given by (5.4) and a matrix $\boldsymbol{A} \in \mathbb{R}^{m,m}$. Since the matrices are m-by-m this will normally require $O(m^3)$ operations. In this section we show that it is possible to calculate the products $\boldsymbol{SA}$ and $\boldsymbol{AS}$ in $O(m^2 \log_2 m)$ operations.

We need to discuss certain transforms known as the Discrete Sine Transform, the Discrete Fourier Transform and the Fast Fourier Transform. These transforms are of independent interest. They have applications to signal processing and image analysis, and are often used when one is dealing with discrete samples of data on a computer.

5.3.1 The Discrete Sine Transform (DST)

Given $\boldsymbol{v} = [v_1, \ldots, v_m]^T \in \mathbb{R}^m$ we say that the vector $\boldsymbol{w} = [w_1, \ldots, w_m]^T$ given by

$$w_j = \sum_{k=1}^{m} \sin\left(\frac{jk\pi}{m+1}\right) v_k, \quad j = 1, \ldots, m$$

is the **Discrete Sine Transform** (DST) of $\boldsymbol{v}$. In matrix form we can write the DST as the matrix times vector $\boldsymbol{w} = \boldsymbol{Sv}$, where $\boldsymbol{S}$ is the sine matrix given by (5.4). We

[2] It is possible to compute $\boldsymbol{V}$ using only two matrix multiplications and hence reduce the complexity to $O(4n^{3/2})$. This is detailed in Problem 5.4.

can then identify the matrix $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{A}$ as the DST of $\boldsymbol{A} \in \mathbb{R}^{m,n}$, i.e. as the DST of the columns of $\boldsymbol{A}$. The product $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{S}$ can also be interpreted as a DST. Indeed, since $\boldsymbol{S}$ is symmetric we have $\boldsymbol{B} = (\boldsymbol{S}\boldsymbol{A}^T)^T$ which means that $\boldsymbol{B}$ is the transpose of the DST of the rows of $\boldsymbol{A}$. It follows that we can compute the unknowns $\boldsymbol{V}$ in Algorithm 5.1 by carrying out Discrete Sine Transforms on 4 m-by-m matrices in addition to the computation of $\boldsymbol{X}$.

5.3.2 The Discrete Fourier Transform (DFT)

The fast computation of the DST is based on its relation to the Discrete Fourier Transform (DFT) and the fact that the DFT can be computed by a technique known as the Fast Fourier Transform (FFT). To define the DFT let for $N \in \mathbb{N}$

$$\omega_N = \exp^{-2\pi i/N} = \cos(2\pi/N) - i\sin(2\pi/N), \tag{5.5}$$

where $i = \sqrt{-1}$ is the imaginary unit. Given $\boldsymbol{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$ we say that $\boldsymbol{z} = [z_1, \ldots, z_N]^T$ given by

$$z_j = \sum_{k=1}^{N} \omega_N^{(j-1)(k-1)} y_k, \quad j = 1, \ldots, N$$

is the **Discrete Fourier Transform** (DFT) of $\boldsymbol{y}$. We can write this as a matrix times vector product $\boldsymbol{z} = \boldsymbol{F}_N \boldsymbol{y}$, where the matrix $\boldsymbol{F}_N$ is given by

$$\boldsymbol{F}_N = \left(\omega_N^{(j-1)(k-1)}\right)_{j,k=1}^{N} \in \mathbb{C}^{N,N}. \tag{5.6}$$

This matrix is known as the **Fourier matrix**. If $\boldsymbol{A} \in \mathbb{R}^{N,m}$ we say that $\boldsymbol{B} = \boldsymbol{F}_N \boldsymbol{A}$ is the DFT of $\boldsymbol{A}$.

As an example, since

$$\omega_4 = \exp^{-2\pi i/4} = \cos(\pi/2) - i\sin(\pi/2) = -i$$

we find

$$\boldsymbol{F}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega_4 & \omega_4^2 & \omega_4^3 \\ 1 & \omega_4^2 & \omega_4^4 & \omega_4^6 \\ 1 & \omega_4^3 & \omega_4^6 & \omega_4^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}. \tag{5.7}$$

The following lemma shows how the Discrete Sine Transform of order m can be computed from the Discrete Fourier Transform of order $2m + 2$.

Lemma 5.2 *Given a positive integer m and a vector $\boldsymbol{x} \in \mathbb{R}^m$. Component k of $\boldsymbol{S}\boldsymbol{x}$ is equal to $i/2$ times component $k + 1$ of $\boldsymbol{F}_{2m+2}\boldsymbol{z}$ where*

$$\boldsymbol{z} = [0, x_1, \ldots, x_m, 0, -x_m, -x_{m-1}, \ldots, -x_1]^T \in \mathbb{R}^{2m+2}.$$

In symbols

$$(\boldsymbol{S}\boldsymbol{x})_k = \frac{i}{2} (\boldsymbol{F}_{2m+2}\boldsymbol{z})_{k+1}, \quad k = 1, \ldots, m.$$

Proof. Let $\omega = \omega_{2m+2} = e^{-2\pi i/(2m+2)} = e^{-\pi i/(m+1)}$. Component $k+1$ of $\boldsymbol{F}_{2m+2}\boldsymbol{z}$ is given by

$$\begin{aligned}(\boldsymbol{F}_{2m+2}\boldsymbol{z})_{k+1} &= \sum_{j=1}^{m} x_j\omega^{jk} - \sum_{j=1}^{m} x_j\omega^{(2m+2-j)k} \\ &= \sum_{j=1}^{m} x_j(\omega^{jk} - \omega^{-jk}) \\ &= -2i\sum_{j=1}^{m} x_j \sin\left(\frac{jk\pi}{m+1}\right) = -2i(\boldsymbol{S}_m\boldsymbol{x})_k.\end{aligned}$$

Dividing both sides by $-2i$ proves the lemma. □

It follows that we can compute the DST of length m by extracting m components from the DFT of length $N = 2m + 2$.

5.3.3 The Fast Fourier Transform (FFT)

From a linear algebra viewpoint the Fast Fourier Transform is a quick way to compute the matrix- vector product $\boldsymbol{F}_N\boldsymbol{y}$. Suppose N is even. The key to the FFT is a connection between $\boldsymbol{F}_N$ and $\boldsymbol{F}_{N/2}$ which makes it possible to compute the FFT of order N as two FFT's of order $N/2$. By repeating this process we can reduce the number of flops to compute a DFT from $O(N^2)$ to $O(N\log_2 N)$.

Suppose N is even. The connection between $\boldsymbol{F}_N$ and $\boldsymbol{F}_{N/2}$ involves a permutation matrix $\boldsymbol{P}_N \in \mathbb{R}^{N,N}$ given by

$$\boldsymbol{P}_N = [\boldsymbol{e}_1, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_{N-1}, \boldsymbol{e}_2, \boldsymbol{e}_4, \ldots, \boldsymbol{e}_N],$$

where the $\boldsymbol{e}_k = (\delta_{j,k})$ are unit vectors. If $\boldsymbol{A}$ is a matrix with N columns $[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ then

$$\boldsymbol{A}\boldsymbol{P}_N = [\boldsymbol{a}_1, \boldsymbol{a}_3, \ldots, \boldsymbol{a}_{N-1}, \boldsymbol{a}_2, \boldsymbol{a}_4, \ldots, \boldsymbol{a}_N],$$

i.e. post multiplying $\boldsymbol{A}$ by $\boldsymbol{P}_N$ permutes the columns of $\boldsymbol{A}$ so that all the odd-indexed columns are followed by all the even-indexed columns. For example we have from (5.7)

$$\boldsymbol{P}_4 = [\boldsymbol{e}_1\ \boldsymbol{e}_3\ \boldsymbol{e}_2\ \boldsymbol{e}_4] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{F}_4\boldsymbol{P}_4 = \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1 \\ 1 & -1 & -i & i \\ \hline 1 & 1 & -1 & -1 \\ 1 & -1 & i & -i \end{array}\right],$$

where we have indicated a certain block structure of $\boldsymbol{F}_4\boldsymbol{P}_4$. These blocks can be related to the 2-by-2 matrix $\boldsymbol{F}_2$. We define the diagonal scaling matrix $\boldsymbol{D}_2$ by

$$\boldsymbol{D}_2 = \mathrm{diag}(1, \omega_4) = \begin{bmatrix} 1 & 0 \\ 1 & -i \end{bmatrix}.$$

Since $\omega_2 = \exp^{-2\pi i/2} = -1$ we find

$$\boldsymbol{F}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{D}_2\boldsymbol{F}_2 = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix},$$

and we see that

$$\boldsymbol{F}_4\boldsymbol{P}_4 = \left[\begin{array}{c|c} \boldsymbol{F}_2 & \boldsymbol{D}_2\boldsymbol{F}_2 \\ \hline \boldsymbol{F}_2 & -\boldsymbol{D}_2\boldsymbol{F}_2 \end{array}\right].$$

This result holds in general.

Theorem 5.3 *If $N = 2m$ is even then*

$$\boldsymbol{F}_{2m}\boldsymbol{P}_{2m} = \left[\begin{array}{c|c} \boldsymbol{F}_m & \boldsymbol{D}_m\boldsymbol{F}_m \\ \hline \boldsymbol{F}_m & -\boldsymbol{D}_m\boldsymbol{F}_m \end{array}\right], \tag{5.8}$$

where

$$\boldsymbol{D}_m = \mathrm{diag}(1, \omega_N, \omega_N^2, \ldots, \omega_N^{m-1}). \tag{5.9}$$

Proof. Fix integers j, k with $0 \le j, k \le m-1$ and set $p = j+1$ and $q = k+1$. Since $\omega_m^m = 1$, $\omega_N^2 = \omega_m$, and $\omega_N^m = -1$ we find by considering elements in the four sub-blocks in turn

$$\begin{array}{llllll} (\boldsymbol{F}_{2m}\boldsymbol{P}_{2m})_{p,q} & = & \omega_N^{j(2k)} & = & \omega_m^{jk} & = & (\boldsymbol{F}_m)_{p,q}, \\ (\boldsymbol{F}_{2m}\boldsymbol{P}_{2m})_{p+m,q} & = & \omega_N^{(j+m)(2k)} & = & \omega_m^{(j+m)k} & = & (\boldsymbol{F}_m)_{p,q}, \\ (\boldsymbol{F}_{2m}\boldsymbol{P}_{2m})_{p,q+m} & = & \omega_N^{j(2k+1)} & = & \omega_N^j\omega_m^{jk} & = & (\boldsymbol{D}_m\boldsymbol{F}_m)_{p,q}, \\ (\boldsymbol{F}_{2m}\boldsymbol{P}_{2m})_{p+m,q+m} & = & \omega_N^{(j+m)(2k+1)} & = & -\omega_N^{j(2k+1)} & = & (-\boldsymbol{D}_m\boldsymbol{F}_m)_{p,q}. \end{array}$$

It follows that the four m-by-m blocks of $\boldsymbol{F}_{2m}\boldsymbol{P}_{2m}$ have the required structure. ∎

Using Theorem 5.3 we can carry out the DFT as a block multiplication. Let $\boldsymbol{y} \in \mathbb{R}^{2m}$ and set $\boldsymbol{w} = \boldsymbol{P}_{2m}^T\boldsymbol{y} = [\boldsymbol{w}_1, \boldsymbol{w}_2]^T$, where $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^m$. Then

$$\begin{aligned} \boldsymbol{F}_{2m}\boldsymbol{y} &= \boldsymbol{F}_{2m}\boldsymbol{P}_{2m}\boldsymbol{P}_{2m}^T\boldsymbol{y} = \boldsymbol{F}_{2m}\boldsymbol{P}_{2m}\boldsymbol{w} \\ &= \left[\begin{array}{c|c} \boldsymbol{F}_m & \boldsymbol{D}_m\boldsymbol{F}_m \\ \hline \boldsymbol{F}_m & -\boldsymbol{D}_m\boldsymbol{F}_m \end{array}\right] \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}_1 + \boldsymbol{q}_2 \\ \boldsymbol{q}_1 - \boldsymbol{q}_2 \end{bmatrix}, \end{aligned}$$

where

$$\boldsymbol{q}_1 = \boldsymbol{F}_m\boldsymbol{w}_1, \quad \text{and} \quad \boldsymbol{q}_2 = \boldsymbol{D}_m(\boldsymbol{F}_m\boldsymbol{w}_2).$$

In order to compute $\boldsymbol{F}_{2m}\boldsymbol{y}$ we need to compute $\boldsymbol{F}_m\boldsymbol{w}_1$ and $\boldsymbol{F}_m\boldsymbol{w}_2$. Note that $\boldsymbol{w}_1^T = [y_1, y_3, \ldots, y_{N-1}]$, while $\boldsymbol{w}_2^T = [y_2, y_4, \ldots, y_N]$. This follows since $\boldsymbol{w}^T = [\boldsymbol{w}_1^T, \boldsymbol{w}_2^T] = \boldsymbol{y}^T\boldsymbol{P}_{2m}$ and post multiplying a vector by $\boldsymbol{P}_{2m}$ moves odd indexed components to the left of all the even indexed components.

We have seen that by combining two FFT's of order m we obtain an FFT of order $2m$. If $N = 2^k$ then this process can be applied recursively as in the following Matlab function:

Algorithm 5.4 (Recursive FFT) For $\boldsymbol{y} \in \mathbb{C}^n$ we compute the Fourier transform $\boldsymbol{z} = \boldsymbol{F}_n \boldsymbol{y}$.

```
function z=fftrec(y)
n=length(y);
if n==1
  z=y;
else
  q1=fftrec(y(1:2:n-1));
  q2=exp(-2*pi*i/n).^(0:n/2-1).*fftrec(y(2:2:n));
  z=[q1+q2 q1-q2];
end
```

Such a recursive version of FFT is useful for testing purposes, but is much too slow for large problems. A challenge for FFT code writers is to develop nonrecursive versions and also to handle efficiently the case where N is not a power of two. We refer to [23] for further details.

The complexity of the FFT is given by $\gamma N \log_2 N$ for some constant γ independent of N. To show this for the special case when N is a power of two let x_k be the complexity (the number of flops) when $N = 2^k$. Since we need two FFT's of order $N/2 = 2^{k-1}$ and a multiplication with the diagonal matrix $\boldsymbol{D}_{N/2}$, it is reasonable to assume that $x_k = 2x_{k-1} + \gamma 2^k$ for some constant γ independent of k. Since $x_0 = 0$ we obtain by induction on k that $x_k = \gamma k 2^k$. Indeed, this holds for $k = 0$ and if $x_{k-1} = \gamma(k-1)2^{k-1}$ then $x_k = 2x_{k-1} + \gamma 2^k = 2\gamma(k-1)2^{k-1} + \gamma 2^k = \gamma k 2^k$. Reasonable implementations of FFT typically have $\gamma \approx 5$, see [23].

The efficiency improvement using the FFT to compute the DFT is spectacular for large N. The direct multiplication $\boldsymbol{F}_N \boldsymbol{y}$ requires $O(8n^2)$ flops since complex arithmetic is involved. Assuming that the FFT uses $5N \log_2 N$ flops we find for $N = 2^{20} \approx 10^6$ the ratio

$$\frac{8N^2}{5N \log_2 N} \approx 84000.$$

Thus if the FFT takes one second of computing time and the computing time is proportional to the number of flops then the direct multiplication would take something like 84000 seconds or 23 hours.

5.3.4 A Poisson Solver based on the FFT

We now have all the ingredients to compute the matrix products $\boldsymbol{SA}$ and $\boldsymbol{AS}$ using FFT's of order $2m + 2$ where m is the order of $\boldsymbol{S}$ and $\boldsymbol{A}$. This can then be used for quick computation of the exact solution $\boldsymbol{V}$ of the discrete Poisson problem in Algorithm 5.1. We first compute $\boldsymbol{H} = \boldsymbol{SF}$ using Lemma 5.2 and m FFT's, one for each of the m columns of $\boldsymbol{F}$. We then compute $\boldsymbol{G} = \boldsymbol{HS}$ by m FFT's, one for each of the rows of $\boldsymbol{H}$. After $\boldsymbol{X}$ is determined we compute $\boldsymbol{Z} = \boldsymbol{SX}$ and $\boldsymbol{V} = \boldsymbol{ZS}$ by another $2m$ FFT's. In total the work amounts to $4m$ FFT's of order $2m+2$. Since one FFT requires $O(\gamma(2m+2)\log_2(2m+2)$ flops the $4m$ FFT's amount to

$$8\gamma m(m+1)\log_2(2m+2) \approx 8\gamma m^2 \log_2 m = 4\gamma n \log_2 n,$$

where $n = m^2$ is the size of the linear system $\boldsymbol{Ax} = \boldsymbol{b}$ we would be solving if Cholesky factorization was used. This should be compared to the $O(8n^{3/2})$ flops used in Algorithm 5.1 requiring 4 straightforward matrix multiplications with $\boldsymbol{S}$. What is faster will depend heavily on the programming of the FFT and the size of the problem. We refer to [23] for other efficient ways to implement the DST.

5.4 Problems

Exercise 5.1 Show that the Fourier matrix $\boldsymbol{F}_4$ is symmetric, but not Hermitian.

Exercise 5.2 Verify Lemma 5.2 directly when $m = 1$.

Exercise 5.3 Show that the exact solution of the discrete Poisson equation (4.3) and (4.4) can be written $\boldsymbol{V} = (v_{i,j})_{i,j=1}^{m}$, where

$$v_{ij} = \frac{1}{(m+1)^4} \sum_{p=1}^{m} \sum_{r=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} \frac{\sin\left(\frac{ip\pi}{m+1}\right) \sin\left(\frac{jr\pi}{m+1}\right) \sin\left(\frac{kp\pi}{m+1}\right) \sin\left(\frac{lr\pi}{m+1}\right)}{\left[\sin\left(\frac{p\pi}{2(m+1)}\right)\right]^2 + \left[\sin\left(\frac{r\pi}{2(m+1)}\right)\right]^2} \, f_{p,r}.$$

Exercise 5.4 Algorithm 5.1 involves multiplying a matrix by $\boldsymbol{S}$ four times. In this problem we show that it is enough to multiply by $\boldsymbol{S}$ two times. We achieve this by diagonalizing only the second $\boldsymbol{T}$ in $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$.

(a) Show that

$$\boldsymbol{TX} + \boldsymbol{XD} = \boldsymbol{C}, \text{ where } \boldsymbol{X} = \boldsymbol{VS}, \text{ and } \boldsymbol{C} = h^2\boldsymbol{FS}.$$

(b) Show that

$$(\boldsymbol{T} + \lambda_j \boldsymbol{I})\boldsymbol{x}_j = \boldsymbol{c}_j \quad j = 1, \ldots, m, \tag{5.10}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$ and $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m]$ and $\lambda_j = 4\sin^2(j\pi h/2)$. Thus we can find $\boldsymbol{X}$ by solving m linear systems, one for each of the columns of $\boldsymbol{X}$. Recall that a tridiagonal $m \times m$ system can be solved by (2.5) and (2.6) in $8m - 7$ flops. Give an algorithm to find $\boldsymbol{X}$ which only requires $O(\delta m^2)$ flops for some constant δ independent of m.

(c) Describe a method to compute $\boldsymbol{V}$ which only requires $O(4m^3) = O(4n^{3/2})$ flops.

(d) Describe a method based on the Fast Fourier Transform which requires $O(\gamma n \log_2 n)$ where γ is the same constant as mentioned at the end of the last section.

Exercise 5.5 Consider the equation

$$\boldsymbol{TV} + \boldsymbol{VT} - \frac{1}{6}\boldsymbol{TVT} = h^2\mu\boldsymbol{F},$$

that was derived in Exercise 4.18 for the 9-point scheme. Define the matrix $\boldsymbol{X}$ by $\boldsymbol{V} = \boldsymbol{SXS} = (x_{j,k})$ where $\boldsymbol{V}$ is the solution of (4.25). Show that

$$\boldsymbol{DX} + \boldsymbol{XD} - \frac{1}{6}\boldsymbol{DXD} = 4h^4\boldsymbol{G}, \text{ where } \boldsymbol{G} = \boldsymbol{S}\mu\boldsymbol{FS},$$

and that

$$x_{j,k} = \frac{h^4 g_{j,k}}{\sigma_j + \sigma_k - \frac{2}{3}\sigma_j\sigma_k}, \text{ where } \sigma_j = \sin^2\left((j\pi h)/2\right) \text{ for } j,k = 1,2,\ldots,m.$$

Show that $\sigma_j + \sigma_k - \frac{2}{3}\sigma_j\sigma_k > 0$ for $j,k = 1,2,\ldots,m$. Conclude that the matrix $\boldsymbol{A}$ in Exercise 4.18 b) is symmetric positive definite and that (4.24) always has a solution $\boldsymbol{V}$.

Exercise 5.6 Derive an algorithm for solving (4.24) which for large m requires essentially the same number of operations as in Algorithm 5.1. (We assume that $\mu\boldsymbol{F}$ already has been formed).

Exercise 5.7 For the biharmonic problem we derived in Exercise 4.19 the equation

$$\boldsymbol{T}^2\boldsymbol{U} + 2\boldsymbol{TUT} + \boldsymbol{UT}^2 = h^4\boldsymbol{F}.$$

Define the matrix $\boldsymbol{X} = (x_{j,k})$ by $\boldsymbol{U} = \boldsymbol{SXS}$ where $\boldsymbol{U}$ is the solution of (4.28). Show that

$$\boldsymbol{D}^2\boldsymbol{X} + 2\boldsymbol{DXD} + \boldsymbol{XD}^2 = 4h^6\boldsymbol{G}, \text{ where } \boldsymbol{G} = \boldsymbol{SFS},$$

and that

$$x_{j,k} = \frac{h^6 g_{j,k}}{4(\sigma_j + \sigma_k)^2}, \text{ where } \sigma_j = \sin^2\left((j\pi h)/2\right) \text{ for } j,k = 1,2,\ldots,m.$$

Exercise 5.8 Use Exercise 5.7 to derive an algorithm

```
function U=simplefastbiharmonic(F)
```

which requires only $O(\delta n^{3/2})$ operations to find $\boldsymbol{U}$ in Problem 4.19. Here δ is some constant independent of n.

Exercise 5.9 In Exercise 5.8 compute the solution $\boldsymbol{U}$ corresponding to $\boldsymbol{F} =$ `ones(m,m)`. For some small m's check that you get the same solution obtained by solving the standard form $\boldsymbol{Ax} = \boldsymbol{b}$ in (4.28). You can use $\boldsymbol{x} = \boldsymbol{A}\backslash\boldsymbol{b}$ for solving $\boldsymbol{Ax} = \boldsymbol{b}$. Use `F(:)` to vectorize a matrix and `reshape(x,m,m)` to turn a vector $\boldsymbol{x} \in \mathbb{R}^{m^2}$ into an $m \times m$ matrix. Make a plot of $\boldsymbol{U}$ for say $m = 50$.

Exercise 5.10 Repeat Exercises 4.19, 5.8 and 5.9 using the nine point rule (4.24) to solve the system (4.27).

Part II

Some Matrix Theory

Chapter 6

Orthonormal Eigenpairs and the Schur Form

A matrix is said to have **orthonormal (orthogonal) eigenpairs** if the eigenvectors are orthonormal (orthogonal). Two examples are the 2. derivative matrix $\boldsymbol{T}$ in (2.3) and the discrete Poisson matrix, cf. Lemma 4.13. In this chapter we characterize the family of matrices that have orthonormal eigenpairs. These matrices are called **normal matrices** and they contain the symmetric, Hermitian, and unitary matrices among their members.

If $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ and $\boldsymbol{S} = \boldsymbol{U} \in \mathbb{C}^{n,n}$ is unitary, then $\boldsymbol{S}^{-1} = \boldsymbol{U}^*$ and $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$. In this case we say that $\boldsymbol{B}$ is **unitary similar** to $\boldsymbol{A}$. In the real case where $\boldsymbol{A}$ and $\boldsymbol{U}$ are real matrices and $\boldsymbol{U}$ is orthogonal, we have $\boldsymbol{S}^{-1} = \boldsymbol{U}^T$ and $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$. Unitary and orthogonal transformations are important in numerical algorithms since they are insensitive to noise in the elements of the matrix.

If $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ then $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{B}$. If $\boldsymbol{B} = \operatorname{diag}(\lambda_j)$ is diagonal and $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$, then $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_j$ for $j = 1, \ldots, n$, and it follows that the columns of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{A}$. Conversely, if $\boldsymbol{A}$ has orthonormal eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, then $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{B}$ or $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$, where the columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$ is diagonal. Thus $\boldsymbol{A}$ is unitary similar to a diagonal matrix if and only if $\boldsymbol{A}$ has a set of orthonormal eigenvectors.

6.1 The Schur Form

Not every matrix can be diagonalized by a similarity transformation, see Theorems D.19, D.20, and D.27. But it can be triangularized, even by a unitary similarity transformation.

Theorem 6.1 (Schur Triangularization) *For each $\boldsymbol{A} \in \mathbb{C}^{n,n}$ there exists a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{n,n}$ such that $\boldsymbol{R} := \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ is upper triangular.*

Proof. We use induction on n. For $n = 1$ the matrix $\boldsymbol{U}$ is the 1×1 identity matrix. Assume that the theorem is true for matrices of order k and suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$, where $n := k+1$. Let $(\lambda_1, \boldsymbol{v}_1)$ be an eigenpair for $\boldsymbol{A}$ with $\|\boldsymbol{v}_1\|_2 = 1$. By

Theorem A.55 we can extend $\boldsymbol{v}_1$ to an orthonormal basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ for $\mathbb{C}^n$. The matrix $\boldsymbol{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{C}^{n,n}$ is unitary, and the first column of the product $\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V}$ is

$$\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V}\boldsymbol{e}_1 = \boldsymbol{V}^*\boldsymbol{A}\boldsymbol{v}_1 = \lambda_1\boldsymbol{V}^*\boldsymbol{v}_1 = \lambda_1\boldsymbol{e}_1.$$

It follows that

$$\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V} = \left[\begin{array}{c|c} \lambda_1 & \boldsymbol{x}^* \\ \hline \boldsymbol{0} & \boldsymbol{M} \end{array}\right], \text{ for some } \boldsymbol{M} \in \mathbb{C}^{k,k} \text{ and } \boldsymbol{x} \in \mathbb{C}^k. \tag{6.1}$$

By the induction hypothesis there is a unitary matrix $\boldsymbol{W}_1 \in \mathbb{C}^{k,k}$ such that $\boldsymbol{W}_1^*\boldsymbol{M}\boldsymbol{W}_1$ is upper triangular. Define

$$\boldsymbol{W} = \left[\begin{array}{c|c} 1 & \boldsymbol{0}^* \\ \hline \boldsymbol{0} & \boldsymbol{W}_1 \end{array}\right] \text{ and } \boldsymbol{U} = \boldsymbol{V}\boldsymbol{W}.$$

Then $\boldsymbol{W}$ and $\boldsymbol{U}$ (cf. Theorem B.26) are unitary and

$$\begin{aligned} \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{W}^*(\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V})\boldsymbol{W} &= \left[\begin{array}{c|c} 1 & \boldsymbol{0}^* \\ \hline \boldsymbol{0} & \boldsymbol{W}_1^* \end{array}\right] \left[\begin{array}{c|c} \lambda_1 & \boldsymbol{x}^* \\ \hline \boldsymbol{0} & \boldsymbol{M} \end{array}\right] \left[\begin{array}{c|c} 1 & \boldsymbol{0}^* \\ \hline \boldsymbol{0} & \boldsymbol{W}_1 \end{array}\right] \\ &= \left[\begin{array}{c|c} \lambda_1 & \boldsymbol{x}^*\boldsymbol{W}_1 \\ \hline \boldsymbol{0} & \boldsymbol{W}_1^*\boldsymbol{M}\boldsymbol{W}_1 \end{array}\right], \end{aligned}$$

is upper triangular. □

By using the unitary transformation $\boldsymbol{V}$ on the $n \times n$ matrix $\boldsymbol{A}$, we obtain a matrix $\boldsymbol{M}$ of order $n-1$. $\boldsymbol{M}$ has the same eigenvalues as $\boldsymbol{A}$ except λ. Thus we can find another eigenvalue of $\boldsymbol{A}$ by working with a smaller matrix $\boldsymbol{M}$ and where one occurrence of λ has been eliminated. This is an example of a **deflation** technique which is very useful in numerical work.

If $\boldsymbol{A}$ has complex eigenvalues then $\boldsymbol{U}$ will be complex even if $\boldsymbol{A}$ is real. The following is a real version of Theorem 6.1.

Theorem 6.2 *For each $\boldsymbol{A} \in \mathbb{R}^{n,n}$ with real eigenvalues there exists an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n,n}$ such that $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$ is upper triangular.*

Proof. Consider the proof of Theorem 6.1. Since $\boldsymbol{A}$ and λ_1 are real the eigenvector $\boldsymbol{v}_1$ is real and the matrix $\boldsymbol{W}$ is real and orthogonal. By the induction hypothesis $\boldsymbol{V}$ is real and orthogonal. But then also $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{W}$ is real and orthogonal. □

Exercise 6.3 *Show that the Schur triangulation of $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 2 \end{smallmatrix}\right]$ is $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \left[\begin{smallmatrix} -1 & -2 \\ 0 & 4 \end{smallmatrix}\right]$, where $\boldsymbol{U} = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right]$.*

From the Schur triangulation $\boldsymbol{R} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ we obtain the **Schur factorization** $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{U}^*$. The matrices $\boldsymbol{U}$ and $\boldsymbol{R}$ are called the **Schur factors**.

A real matrix with complex eigenvalues cannot be reduced to triangular form by an orthogonal similarity transformation. Indeed, if $\boldsymbol{R} = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$ is triangular,

one of the diagonal elements of $\boldsymbol{R}$ (one of the eigenvalues of $\boldsymbol{A}$) will be complex. But then $\boldsymbol{U}$ cannot be real. How far can we reduce a real matrix $\boldsymbol{A}$ by an orthogonal similarity transformation? To study this we note that the complex eigenvalues of $\boldsymbol{A}$ occur in conjugate pairs, $\lambda = \mu + i\nu$, $\overline{\lambda} = \mu - i\nu$, where μ, ν are real. The real 2×2 matrix

$$\boldsymbol{M} = \begin{bmatrix} \mu & \nu \\ -\nu & \mu \end{bmatrix} \tag{6.2}$$

has eigenvalues λ and $\overline{\lambda}$. We say that a matrix is quasi-triangular if it is block triangular with only 1×1 and 2×2 blocks on the diagonal. Moreover, no 2×2 block should have real eigenvalues. As an example consider

$$\boldsymbol{R} = \left[\begin{array}{cc|c|cc} 2 & 1 & 3 & 4 & 5 \\ -1 & 2 & 4 & 3 & 2 \\ \hline 0 & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & -1 & 1 \end{array}\right].$$

$\boldsymbol{R}$ has three diagonal blocks:

$$\boldsymbol{D}_1 = \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{D}_3 = \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}.$$

By Theorem D.3 the eigenvalues of $\boldsymbol{R}$ are the union of the eigenvalues of $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$. The eigenvalues of $\boldsymbol{D}_1$ are $2+i$ and $2-i$, while $\boldsymbol{D}_2$ has eigenvalue 1, and $\boldsymbol{D}_3$ has the same eigenvalues as $\boldsymbol{D}_1$. Thus $\boldsymbol{R}$ has one real eigenvalue 1 corresponding to the 1×1 block and complex eigenvalues $2+i$, $2-i$ with multiplicity 2 corresponding to the two 2×2 blocks.

For a proof that any $\boldsymbol{A} \in \mathbb{R}^{n,n}$ can be brought to quasi-triangular form by a real orthogonal similarity transformation see Section 6.4.

6.1.1 The Spectral Theorem

The special cases where $\boldsymbol{A}$ is Hermitian or real and symmetric deserve special attention.

Theorem 6.4 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is Hermitian. Then $\boldsymbol{A}$ has real eigenvalues $\lambda_1, \ldots, \lambda_n$. Moreover, there is a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{n,n}$ such that $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. For the columns $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ of $\boldsymbol{U}$ we have $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for $j = 1, \ldots, n$. Thus $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ are orthonormal eigenvectors of $\boldsymbol{A}$.*

Proof. That the eigenvalues are real was shown in Lemma 4.12. By Theorem 6.1 there is a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{n,n}$ so that $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ is upper triangular. Since $\boldsymbol{A}^* = \boldsymbol{A}$, we have $\boldsymbol{B}^* = \boldsymbol{B}$. But then $\boldsymbol{B}$ must be diagonal. The columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ of $\boldsymbol{U}$ satisfies $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for all j and are orthonormal eigenvectors of $\boldsymbol{A}$. ∎

The following real version is known as the Spectral Theorem.

Theorem 6.5 *Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ and $\boldsymbol{A}^T = \boldsymbol{A}$. Then $\boldsymbol{A}$ has real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Moreover, there is an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n,n}$ such that*

$$\boldsymbol{U}^T \boldsymbol{A} \boldsymbol{U} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

For the columns $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ of $\boldsymbol{U}$ we have $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for $j = 1, \ldots, n$. Thus $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ are orthonormal eigenvectors of $\boldsymbol{A}$.

Proof. The proof is similar to Theorem 6.4. Since $\boldsymbol{A}^* = \boldsymbol{A}$ it follows from Theorem 6.4 that the eigenvalues are real. By Theorem 6.2 there is a matrix $\boldsymbol{U} \in \mathbb{R}^{n,n}$ with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ so that $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$ is upper triangular. Since $\boldsymbol{A}^T = \boldsymbol{A}$, we have $\boldsymbol{B}^T = \boldsymbol{B}$. But then $\boldsymbol{B}$ must be diagonal. The columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ of $\boldsymbol{U}$ satisfies $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for all j and are orthonormal eigenvectors of $\boldsymbol{A}$. □

Example 6.6 *The orthogonal diagonalization of $\boldsymbol{A} = \left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right]$ is $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \operatorname{diag}(1, 3)$, where $\boldsymbol{U} = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right]$.*

Exercise 6.7 *Suppose $\boldsymbol{C} = \boldsymbol{A} + i\boldsymbol{B}$, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n,n}$. Show that $\boldsymbol{C}$ is skew-Hermitian if and only if $\boldsymbol{A}^T = -\boldsymbol{A}$ and $\boldsymbol{B}^T = \boldsymbol{B}$.*

Exercise 6.8 *Show that any eigenvalue of a skew-Hermitian matrix is purely imaginary.*

6.2 Normal Matrices

It is possible to characterize matrices that have a diagonal Schur form.

Definition 6.9 (Normal Matrix) *A matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is said to be* **normal** *if $\boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{A}^*\boldsymbol{A}$.*

Examples of normal matrices are

1. $\boldsymbol{A}^* = \boldsymbol{A}$, (Hermitian)
2. $\boldsymbol{A}^* = -\boldsymbol{A}$, (Skew-Hermitian)
3. $\boldsymbol{A}^* = \boldsymbol{A}^{-1}$, (Unitary)
4. $\boldsymbol{A} = \boldsymbol{D}$. (Diagonal)

For real matrices "Hermitian" and "symmetric" are synonyms.

The following theorem says that a matrix has orthonormal eigenpairs if and only if it is normal.

Theorem 6.10 *A matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is unitary similar with a diagonal matrix if and only if it is normal.*

Proof. If $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$, with $\boldsymbol{B}$ diagonal, and $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$, then

$$\boldsymbol{A}\boldsymbol{A}^* = (\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^*)(\boldsymbol{U}\boldsymbol{B}^*\boldsymbol{U}^*) = \boldsymbol{U}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{U}^* \text{ and}$$
$$\boldsymbol{A}^*\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{B}^*\boldsymbol{U}^*)(\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^*) = \boldsymbol{U}\boldsymbol{B}^*\boldsymbol{B}\boldsymbol{U}^*.$$

Now $\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{B}^*\boldsymbol{B}$ since $\boldsymbol{B}$ is diagonal, and $\boldsymbol{A}$ is normal.

Suppose $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^*$. By Theorem 6.1 we can find $\boldsymbol{U}$ with $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ such that $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ is upper triangular. Since $\boldsymbol{A}$ is normal $\boldsymbol{B}$ is normal. Indeed,

$$\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}^*\boldsymbol{U} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{U} = \boldsymbol{U}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{B}^*\boldsymbol{B}.$$

The proof is complete if we can show that an upper triangular normal matrix $\boldsymbol{B}$ must be diagonal. The diagonal elements in $\boldsymbol{E} := \boldsymbol{B}^*\boldsymbol{B}$ and $\boldsymbol{F} := \boldsymbol{B}\boldsymbol{B}^*$ are given by

$$e_{ii} = \sum_{k=1}^{n} \bar{b}_{ki}b_{ki} = \sum_{k=1}^{i} |b_{ki}|^2 \text{ and } f_{ii} = \sum_{k=1}^{n} b_{ik}\bar{b}_{ik} = \sum_{k=i}^{n} |b_{ik}|^2.$$

The result now follows by equating e_{ii} and f_{ii} for $i = 1, 2, \ldots, n$. In particular for $i = 1$ we have $|b_{11}|^2 = |b_{11}|^2 + |b_{12}|^2 + \cdots + |b_{1n}|^2$, so $b_{1k} = 0$ for $k = 2, 3, \ldots, n$. Suppose $b_{jk} = 0$ for $j = 1, \ldots, i-1$, $k = j+1, \ldots, n$. Then

$$e_{ii} = \sum_{k=1}^{i} |b_{ki}|^2 = |b_{ii}|^2 = \sum_{k=i}^{n} |b_{ik}|^2 = f_{ii}$$

so $b_{ik} = 0$, $k = i+1, \ldots, n$. By induction on the rows we see that $\boldsymbol{B}$ is diagonal. □

6.3 The Rayleigh Quotient and Minmax Theorems

6.3.1 The Rayleigh Quotient

The Rayleigh quotient is an important tool when studying eigenvalues.

Definition 6.11 *For $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and any $\boldsymbol{x} \in \mathbb{C}^n$ the quantity $R(\boldsymbol{x}) = R_{\boldsymbol{A}}(\boldsymbol{x}) := \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}$ is called a* **Rayleigh quotient** *for $\boldsymbol{A}$.*

If $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}$ then $R(\boldsymbol{x}) = \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \lambda$.

Exercise 6.12 *More generally for $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and any $\boldsymbol{y}, \boldsymbol{x} \in \mathbb{C}^n$ with $\boldsymbol{y}^*\boldsymbol{x} \neq 0$ the quantity $R(\boldsymbol{y}, \boldsymbol{x}) = R_{\boldsymbol{A}}(\boldsymbol{y}, \boldsymbol{x}) := \frac{\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{y}^*\boldsymbol{x}}$ is also called a* **Rayleigh quotient** *for $\boldsymbol{A}$. Show that if $(\lambda, \boldsymbol{x})$ is a (right) eigenpair for $\boldsymbol{A}$ then $R(\boldsymbol{y}, \boldsymbol{x}) = \lambda$ for any $\boldsymbol{y}$ with $\boldsymbol{y}^*\boldsymbol{x} \neq 0$. Also show that if $(\lambda, \boldsymbol{y})$ is a left eigenpair for $\boldsymbol{A}$ then $R(\boldsymbol{y}, \boldsymbol{x}) = \lambda$ for any $\boldsymbol{x}$ with $\boldsymbol{y}^*\boldsymbol{x} \neq 0$.*

The following lemma gives some useful formulas.

Lemma 6.13 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ be an orthonormal basis for a subspace $\mathcal{S} \subset \mathbb{C}^n$. If $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ with $\boldsymbol{x} = \sum_{j=1}^k c_j \boldsymbol{u}_j$ and $\boldsymbol{y} = \sum_{j=1}^k d_j \boldsymbol{u}_j$, then*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^* \boldsymbol{y} = \sum_{j=1}^k \overline{c}_j d_j, \tag{6.3}$$

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^* \boldsymbol{x}} = \frac{\sum_{j=1}^k \lambda_j |c_j|^2}{\sum_{j=1}^k |c_j|^2} = \sum_{j=1}^k \lambda_j |c_j|^2, \text{ if } \|\boldsymbol{x}\|_2 = 1. \tag{6.4}$$

Proof. We have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \sum_{i=1}^k c_i \boldsymbol{u}_i, \sum_{j=1}^k d_j \boldsymbol{u}_j \rangle = \sum_{i=1}^k \sum_{j=1}^k \overline{c}_i d_j \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \sum_{j=1}^k \overline{c}_j d_j$$

and (6.3) follows. Since $\boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^k c_j \boldsymbol{A}\boldsymbol{u}_j = \sum_{j=1}^k c_j \lambda_j \boldsymbol{u}_j$ we obtain from (6.3) that $\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle = \sum_{j=1}^k \lambda_j |c_j|^2$ and also $\|\boldsymbol{x}\|_2^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \sum_{j=1}^k |c_j|^2$. This shows both equalities in (6.4). □

The Rayleigh quotient is especially useful when the matrix $\boldsymbol{A}$ is Hermitian. Since $\boldsymbol{A}$ is normal it has orthonormal eigenpairs $\{(\lambda_1, \boldsymbol{u}_1), \ldots, (\lambda_n, \boldsymbol{u}_n)\}$ and the eigenvalues are real and can be ordered, say $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. In this case we have for any $i \leq k$ and $c_i, \ldots, c_k$ not all zero

$$\lambda_k \leq \frac{\sum_{j=i}^k \lambda_j |c_j|^2}{\sum_{j=i}^k |c_j|^2} \leq \lambda_i, \text{ if } \lambda_i \geq \lambda_{i+1} \geq \cdots \geq \lambda_k. \tag{6.5}$$

Indeed, to show the lower (upper) bound we replace all λ's in the numerator $\sum_{j=i}^k \lambda_j |c_j|^2$ by λ_k (λ_i). From (6.5) it follows that the value of the Rayleigh quotient for a Hermitian matrix must lie between the smallest and largest eigenvalue. Since $R(\boldsymbol{u}_1) = \lambda_1$ and $R(\boldsymbol{u}_n) = \lambda_n$ we can express the smallest and largest eigenvalue in terms of extrema of the Rayleigh quotient.

$$\lambda_n = \min_{\substack{\boldsymbol{x} \in \mathbb{C}^n \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) \leq \max_{\substack{\boldsymbol{x} \in \mathbb{C}^n \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) = \lambda_1 \tag{6.6}$$

6.3.2 Minmax and Maxmin Theorems

More generally we have a minmax and maxmin characterization of the eigenvalues of a Hermitian matrix. In the following theorem $\mathcal{S}$ is a subspace of $\mathbb{C}^n$ of the indicated dimension.

Theorem 6.14 (The Courant-Fischer Theorem) *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is Hermitian with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ ordered so that $\lambda_1 \geq \cdots \geq \lambda_n$. Then for $k = 1, \ldots, n$*

$$\lambda_k = \min_{\dim(\mathcal{S}) = n-k+1} \max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) = \max_{\dim(\mathcal{S}) = k} \min_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}). \tag{6.7}$$

Proof. We prove the maxmin version and leave the minmax version as an exercise. Let $\{(\lambda_1, \boldsymbol{u}_1), \ldots, (\lambda_n, \boldsymbol{u}_n)\}$ be orthonormal eigenpairs for $\boldsymbol{A}$. Fix k. We will show that $\max\min R \leq \lambda_k$ and $\max\min R \geq \lambda_k$, where $\max\min R$ is shorthand for the expression after the second equality in (6.7). Let $\mathcal{S}$ be any subspace of $\mathbb{C}^n$ of dimension k and define $\mathcal{S}' = \operatorname{span}\{\boldsymbol{u}_k, \ldots, \boldsymbol{u}_n\}$. Since $\mathcal{S} + \mathcal{S}' \subset \mathbb{C}^n$ we have $\dim(\mathcal{S} + \mathcal{S}') \leq n$ and we can use (A.6) to find

$$\dim(\mathcal{S} \cap \mathcal{S}') = \dim(\mathcal{S}) + \dim(\mathcal{S}') - \dim(\mathcal{S} + \mathcal{S}') \geq k + (n - k + 1) - n = 1,$$

and it follows that $\mathcal{S} \cap \mathcal{S}'$ is nonempty. Let $\boldsymbol{y} \in \mathcal{S} \cap \mathcal{S}' = \sum_{j=k}^{n} d_j \boldsymbol{u}_j$. By (6.4) applied to $\mathcal{S}'$ we find

$$R(\boldsymbol{y}) = \frac{\sum_{j=k}^{n} \lambda_j |d_j|^2}{\sum_{j=k}^{n} |d_j|^2} \leq \lambda_k.$$

This implies that $\min_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) \leq \lambda_k$ and therefore, since $\mathcal{S}$ is arbitrary, $\max\min R \leq \lambda_k$. To show the inequality in the opposite direction we use the subspace $\mathcal{S}_k := \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$. Suppose $\boldsymbol{x} = \sum_{j=1}^{k} c_j \boldsymbol{u}_j$ is any nonzero element in $\mathcal{S}_k$. Then

$$R(\boldsymbol{x}) = \frac{\sum_{j=1}^{k} \lambda_j |c_j|^2}{\sum_{j=1}^{k} |c_j|^2} \geq \lambda_k.$$

Since $\boldsymbol{x}$ is arbitrary we obtain $\min_{\substack{\boldsymbol{x} \in \mathcal{S}_k \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) \geq \lambda_k$ and therefore $\max\min R \geq \lambda_k$. □

Exercise 6.15 *Modify the proof of the maxmin verson of the Courant-Fischer theorem to prove the minmax version.*

Using Theorem 6.14 we can prove inequalities of eigenvalues without knowing the eigenvectors and we can get both upper and lower bounds.

Corollary 6.16 *Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^{n,n}$ be Hermitian with eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$, and $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n$, respectively. If $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$ then*

$$\alpha_i + \beta_n \leq \gamma_i \leq \alpha_i + \beta_1, \text{ for } i = 1, 2, \ldots, n. \tag{6.8}$$

Proof. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be orthonormal eigenvectors for $\boldsymbol{A}$ and let for fixed i, $\mathcal{S} := \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n-i+1}\}$. By Theorem 6.14 and (6.5) we obtain

$$\gamma_i \leq \max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R_{\boldsymbol{C}}(\boldsymbol{x}) \leq \max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R_{\boldsymbol{A}}(\boldsymbol{x}) + \max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R_{\boldsymbol{B}}(\boldsymbol{x}) = \alpha_i + \max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R_{\boldsymbol{B}}(\boldsymbol{x}) \leq \alpha_i + \beta_1,$$

and this proves the upper inequality. For the lower one we define $\boldsymbol{D} := -\boldsymbol{B}$ and observe that $-\beta_n$ is the largest eigenvalue of $\boldsymbol{D}$. Since $\boldsymbol{A} = \boldsymbol{C} + \boldsymbol{D}$ it follows from the result just proved that $\alpha_i \leq \gamma_i - \beta_n$, which is the same as the lower inequality. □

In many applications of this result the matrix $\boldsymbol{B}$ will be small and then the theorem states that the eigenvalues of $\boldsymbol{C}$ are close to those of $\boldsymbol{A}$. Moreover, it associates a unique eigenvalue of $\boldsymbol{A}$ with each eigenvalue of $\boldsymbol{C}$.

Exercise 6.17 *Show that in Corollary 6.16, if $\boldsymbol{B}$ is symmetric positive semidefinite then $\gamma_i \geq \alpha_i$.*

6.3.3 The Hoffman-Wielandt Theorem

We can also give a bound involving all eigenvalues.

Theorem 6.18 (Hoffman-Wielandt Theorem) *Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$ are both normal matrices with eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_n$, respectively. Then there is a permutation $i_1, \ldots, i_n$ of $1, 2, \ldots, n$ such that*

$$\sum_{j=1}^{n} |\mu_{i_j} - \lambda_j|^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij} - b_{ij}|^2. \tag{6.9}$$

Taking $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{E}$ this shows, in terms of absolute error, that as long as $\boldsymbol{A} + \boldsymbol{E}$ is normal, i. e., we perturb in a "normal way", then the eigenvalue problem for a normal matrix is well conditioned. Small perturbation in the elements of $\boldsymbol{A}$ lead to small changes in the eigenvalues.

For a proof of this theorem see [[19], p. 190]. For a Hermitian matrix we can use the identity permutation if we order both set of eigenvalues in nonincreasing or nondecreasing order.

Exercise 6.19 *Show that* (6.9) *does not hold for the matrices $\boldsymbol{A} := \left[\begin{smallmatrix} 0 & 0 \\ 0 & 4 \end{smallmatrix}\right]$ and $\boldsymbol{B} := \left[\begin{smallmatrix} -1 & -1 \\ 1 & 1 \end{smallmatrix}\right]$. Why does this not contradict the Hoffman-Wielandt theorem?*

6.4 Proof of the Real Schur Form

In this section we prove the following theorem.

Theorem 6.20 *Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$. Then we can find $\boldsymbol{U} \in \mathbb{R}^{n,n}$ with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ such that $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$ is quasi-triangular.*

Proof. If $\boldsymbol{A}$ has only real eigenvalues, Theorem 6.2 gives the result. Suppose $\lambda = \mu + i\nu$, $\mu, \nu \in \mathbb{R}$, is an eigenvalue of $\boldsymbol{A}$ with $\nu \neq 0$. Let $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, be an eigenvector of $\boldsymbol{A}$ corresponding to λ. Since

$$\boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}(\boldsymbol{x} + i\boldsymbol{y}) = (\mu + i\nu)(\boldsymbol{x} + i\boldsymbol{y}) = \mu\boldsymbol{x} - \nu\boldsymbol{y} + i(\nu\boldsymbol{x} + \mu\boldsymbol{y}),$$

we find by comparing real and imaginary parts

$$\boldsymbol{A}\boldsymbol{x} = \mu\boldsymbol{x} - \nu\boldsymbol{y}, \quad \boldsymbol{A}\boldsymbol{y} = \nu\boldsymbol{x} + \mu\boldsymbol{y}. \tag{6.10}$$

We claim that $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly independent. First we note that $\nu \neq 0$ implies $\boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{y} \neq \boldsymbol{0}$. For if $\boldsymbol{x} = \boldsymbol{0}$ then (6.10) implies that $\boldsymbol{0} = -\nu\boldsymbol{y}$, and hence $\boldsymbol{y} = \boldsymbol{0}$ as

well, contradicting the nonzeroness of the eigenvector. Similarly, if $\boldsymbol{y} = \boldsymbol{0}$ then $\boldsymbol{0} = \nu\boldsymbol{x}$, again resulting in a zero eigenvector. Suppose $\boldsymbol{y} = \alpha\boldsymbol{x}$ for some α. Replacing $\boldsymbol{y}$ by $\alpha\boldsymbol{x}$ in (6.10), we find $\boldsymbol{A}\boldsymbol{x} = (\mu - \alpha\nu)\boldsymbol{x}$ and $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y}/\alpha = (\mu + \nu/\alpha)\boldsymbol{x}$. But then $\mu - \alpha\nu = \mu + \nu/\alpha$ or $\alpha^2 = -1$. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are real, we cannot have both $\boldsymbol{y} = \alpha\boldsymbol{x}$ and $\alpha^2 = -1$. We conclude that $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly independent.

(6.10) can be written in matrix form as

$$\boldsymbol{A}\boldsymbol{X}_1 = \boldsymbol{X}_1\boldsymbol{M}, \quad \boldsymbol{X}_1 = (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{n,2}, \tag{6.11}$$

where $\boldsymbol{M}$ is given by (6.2). By Theorem 12.3 we can find an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{n,n}$ such that

$$\boldsymbol{Q}\boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}$$

where $\boldsymbol{R} \in \mathbb{R}^{2,2}$ is upper triangular. Since $\boldsymbol{X}_1$ has linearly independent columns, $\boldsymbol{R}$ is nonsingular. Let $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n]$ and define

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2] = [\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{q}_3, \ldots, \boldsymbol{q}_n].$$

We find

$$\boldsymbol{Q}\boldsymbol{X} = [\boldsymbol{Q}\boldsymbol{X}_1, \boldsymbol{Q}\boldsymbol{q}_3, \ldots, \boldsymbol{Q}\boldsymbol{q}_n] = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-2} \end{bmatrix}.$$

Since $\boldsymbol{R}$ is nonsingular, $\boldsymbol{Q}\boldsymbol{X}$ and $\boldsymbol{X}$ are nonsingular. Moreover, using (6.11)

$$\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = [\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}_1, \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}_2] = [\boldsymbol{X}^{-1}\boldsymbol{X}_1\boldsymbol{M}, \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}_2] = \begin{bmatrix} \boldsymbol{M} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}$$

for some matrices $\boldsymbol{B} \in \mathbb{R}^{2,n-2}$, $\boldsymbol{C} \in \mathbb{R}^{n-2,n-2}$. Now

$$\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^T = (\boldsymbol{Q}\boldsymbol{X})\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}(\boldsymbol{Q}\boldsymbol{X})^{-1} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-2} \end{bmatrix} \begin{bmatrix} \boldsymbol{M} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-2} \end{bmatrix},$$

or

$$\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^T = \begin{bmatrix} \boldsymbol{R}\boldsymbol{M}\boldsymbol{R}^{-1} & \boldsymbol{R}\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}. \tag{6.12}$$

By Theorem D.15 the 2×2 matrix $\boldsymbol{R}\boldsymbol{M}\boldsymbol{R}^{-1}$ has the same eigenvalues λ and $\overline{\lambda}$ as $\boldsymbol{M}$. The remaining $n-2$ eigenvalues of $\boldsymbol{A}$ are the eigenvalues of $\boldsymbol{C}$.

To complete the proof we use induction on n. The theorem is trivially true for $n = 1$ and $n = 2$. Suppose $n \geq 3$ and it holds for matrices of order $\leq n-1$. Let

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{V}} \end{bmatrix}$$

where $\hat{\boldsymbol{V}} \in \mathbb{R}^{n-2,n-2}$, $\hat{\boldsymbol{V}}^T\hat{\boldsymbol{V}} = \boldsymbol{I}_{n-2}$ and $\hat{\boldsymbol{V}}^T\boldsymbol{C}\hat{\boldsymbol{V}}$ is quasi-triangular. Let $\boldsymbol{U} = \boldsymbol{Q}\boldsymbol{V}$. Then $\boldsymbol{U} \in \mathbb{R}^{n,n}$, $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$ is quasi-triangular. ∎

Chapter 7

The Singular Value Decomposition

The singular value decomposition is useful both for theory and practice. Some of its applications include solving over-determined equations, principal component analysis in statistics, numerical determination of the rank of a matrix, algorithms used in search engines, and the theory of matrices.

7.1 Singular Values and Singular Vectors

We know from Theorem 6.10 that a square matrix $\boldsymbol{A}$ can be diagonalized by a unitary similarity transformation if and only if it is normal. In particular, if $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is normal with eigenvalues $\lambda_1, \ldots, \lambda_n$ then

$$\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \text{ or } \boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*, \text{ where } \boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}. \tag{7.1}$$

In this section we show that any matrix, even a rectangular one, can be diagonalized provided we allow two different unitary matrices. Thus

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \text{ where } \boldsymbol{\Sigma} \text{ is a diagonal matrix, } \boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}, \text{ and } \boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}. \tag{7.2}$$

The diagonal elements of $\boldsymbol{\Sigma}$, are called singular values and the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are singular vectors. The formula $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ is known as the singular value decomposition of $\boldsymbol{A}$.

7.1.1 SVD and SVF

Every matrix has a singular value decomposition (SVD) and a reduced form called the singular value factorization (SVF). To derive these we start with a lemma and a theorem.

Lemma 7.1 *Suppose $m, n \in \mathbb{N}$ and $\boldsymbol{A} \in \mathbb{C}^{m,n}$. The matrix $\boldsymbol{A}^*\boldsymbol{A}$ has eigenpairs $(\lambda_j, \boldsymbol{v}_j)$ for $j = 1, \ldots, n$, where $\boldsymbol{v}_j^*\boldsymbol{v}_k = \delta_{jk}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Moreover,*

$$\sigma_j := \sqrt{\lambda_j} = \|\boldsymbol{A}\boldsymbol{v}_j\|_2, \text{ for } j = 1, \ldots, n. \tag{7.3}$$

Proof. The matrix $\boldsymbol{A}^*\boldsymbol{A} \in \mathbb{C}^{n,n}$ is Hermitian, and by Theorem 6.4 it has real eigenvalues λ_j and orthonormal eigenvectors $\boldsymbol{v}_j$ for $j = 1, \ldots, n$. For each j $\|\boldsymbol{A}\boldsymbol{v}_j\|_2^2 = (\boldsymbol{A}\boldsymbol{v}_j)^*\boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{v}_j^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{v}_j^*\lambda_j\boldsymbol{v}_j = \lambda_j$, since $\boldsymbol{v}_j^*\boldsymbol{v}_j = 1$, and (7.3) follows. □

The nonnegative square roots of the n eigenvalues of $\boldsymbol{A}^*\boldsymbol{A}$ are called the **singular values** of $\boldsymbol{A} \in \mathbb{C}^{m,n}$. They are usually ordered so that

$$\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_n. \tag{7.4}$$

We will show that the number r of positive singular values equals the rank of $\boldsymbol{A}$. Moreover, the eigenvectors of $\boldsymbol{A}^*\boldsymbol{A}$ determine orthonormal bases for the column space $\operatorname{span}(\boldsymbol{A})$ and null space $\ker(\boldsymbol{A})$ of $\boldsymbol{A}$.

Theorem 7.2 *Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and let $(\sigma_j^2, \boldsymbol{v}_j)$ for $j = 1, \ldots, n$ be orthonormal eigenpairs for $\boldsymbol{A}^*\boldsymbol{A}$ with $\sigma_1, \ldots, \sigma_n$ ordered as in* (7.4)*. Then $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ is an orthogonal basis for the column space* $\operatorname{span}(\boldsymbol{A})$ *of $\boldsymbol{A}$ and $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$ is an orthonormal basis for the nullspace* $\ker(\boldsymbol{A})$ *of $\boldsymbol{A}$.*

Proof. The proof will be complete if we can show

1. $\boldsymbol{A}\boldsymbol{v}_j \neq \boldsymbol{0}$ if and only if $1 \leq j \leq r$.
2. $\boldsymbol{x} = \sum_{j=1}^n c_j\boldsymbol{v}_j \Rightarrow \boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^r c_j\boldsymbol{A}\boldsymbol{v}_j$.
3. $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ is orthogonal and nonzero.
4. $\operatorname{span}(\boldsymbol{A}) = \operatorname{span}\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$.
5. $\ker(\boldsymbol{A}) = \operatorname{span}\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$.

1 follows by combining (7.3) and (7.4). 2 is a consequence of 1. 1 and the calculation

$$(\boldsymbol{A}\boldsymbol{v}_j)^*\boldsymbol{A}\boldsymbol{v}_k = \boldsymbol{v}_j^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}_k = \boldsymbol{v}_j^*\sigma_k^2\boldsymbol{v}_k = 0, j \neq k$$

implies 3. Clearly $\boldsymbol{A}\boldsymbol{v}_j \in \operatorname{span}(\boldsymbol{A})$ for $j = 1, \ldots, r$ and $\boldsymbol{v}_j \in \ker(\boldsymbol{A})$ by 1. 3 implies 4. Finally, by 3 $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ are linearly independent. So if $\boldsymbol{x} \in \ker(\boldsymbol{A})$ then by 2 $c_1 = \cdots = c_r = 0$ and 5 follows. □

Every matrix has a singular value decomposition.

Theorem 7.3 (SVD) *Let $m, n \in \mathbb{N}$ and suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ has rank r. Then $\boldsymbol{A}$ has exactly r positive singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Moreover, $\boldsymbol{A}$ has the singular value decomposition*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \quad \boldsymbol{U} \in \mathbb{C}^{m,m},\ \boldsymbol{\Sigma} \in \mathbb{R}^{m,n},\ \boldsymbol{V} \in \mathbb{C}^{n,n},$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary and

$$\boldsymbol{\Sigma} := \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0}_{r,n-r} \\ \boldsymbol{0}_{m-r,r} & \boldsymbol{0}_{m-r,n-r} \end{bmatrix} \in \mathbb{R}^{m,n},\ \textit{where } \boldsymbol{\Sigma}_1 := \operatorname{diag}(\sigma_1, \ldots, \sigma_r). \tag{7.5}$$

Here $\mathbf{0}_{k,l} \in \mathbb{R}^{k,l}$ *is the zero matrix and* $\mathbf{0}_{k,l} = [\,]$ *is the empty matrix, if* $k = 0$ *or* $l = 0$.

If $\boldsymbol{A}$ *is real then* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, *where* $\boldsymbol{U} \in \mathbb{R}^{m,m}$ *and* $\boldsymbol{V} \in \mathbb{R}^{n,n}$ *are orthonormal, and* $\boldsymbol{\Sigma}$ *is again given by* (7.5).

Proof. Suppose $(\lambda_j, \boldsymbol{v}_j)$ for $j = 1, \ldots, n$ are orthonormal eigenpairs for $\boldsymbol{A}^*\boldsymbol{A}$ and define $\boldsymbol{\Sigma}$ by (7.5), where $\sigma_j = \sqrt{\lambda_j}$ for all j. By Theorem 7.2 the set $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ is an orthogonal basis for the column space of $\boldsymbol{A}$ and it follows that r is the number of positive singular values. We turn $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ into an orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ for span($\boldsymbol{A}$) by setting (cf. (7.3))

$$\boldsymbol{u}_j := \frac{\boldsymbol{A}\boldsymbol{v}_j}{\|\boldsymbol{A}\boldsymbol{v}_j\|_2} = \frac{1}{\sigma_j}\boldsymbol{A}\boldsymbol{v}_j, \text{ for } j = 1, \ldots, r.$$

By Theorem 7.2

$$\boldsymbol{A}\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j, \ j = 1, \ldots, r \text{ and } \boldsymbol{A}\boldsymbol{v}_j = \mathbf{0}, \ j = r+1, \ldots, n. \tag{7.6}$$

We extend $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ to an orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ for $\mathbb{C}^m$ and define

$$\boldsymbol{U} := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m] \in \mathbb{C}^{m,m} \text{ and } \boldsymbol{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{C}^{n,n}.$$

Since $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns they are unitary matrices, and from (7.5) and (7.6)

$$\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{U}[\sigma_1\boldsymbol{e}_1, \ldots, \sigma_r\boldsymbol{e}_r, \mathbf{0}, \ldots, \mathbf{0}] = [\sigma_1\boldsymbol{u}_1, \ldots, \sigma_r\boldsymbol{u}_r, \mathbf{0}, \ldots, \mathbf{0}] = [\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_n].$$

Thus $\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{V}$ and since $\boldsymbol{V}$ is unitary we find $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^* = \boldsymbol{A}$.

For a matrix with real elements the eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$ are real and the singular value decomposition takes the stated form. ∎

From the singular value decomposition we obtain a reduced factorization called the singular value factorization and an outer product form of this factorization.

Corollary 7.4 (SVF) *Suppose* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ *is a singular value decomposition of a rank* r *matrix* $\boldsymbol{A} \in \mathbb{C}^{m,n}$. *Then* $\boldsymbol{A}$ *has the singular value factorization*

$$\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^* = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*, \quad \boldsymbol{U}_1 \in \mathbb{C}^{m,r}, \ \boldsymbol{\Sigma}_1 \in \mathbb{R}^{r,r}, \ \boldsymbol{V}_1 \in \mathbb{C}^{n,r},$$

where

$$\begin{aligned} &\boldsymbol{\Sigma}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r), \\ &\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m] = [\boldsymbol{U}_1, \boldsymbol{U}_2], \ \boldsymbol{U}_1 \in \mathbb{C}^{m,r}, \ \boldsymbol{U}_2 \in \mathbb{C}^{m,m-r}, \\ &\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] = [\boldsymbol{V}_1, \boldsymbol{V}_2], \ \boldsymbol{V}_1 \in \mathbb{C}^{n,r}, \ \boldsymbol{V}_2 \in \mathbb{C}^{n,n-r}, \end{aligned} \tag{7.7}$$

Proof. We find

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \begin{bmatrix}\boldsymbol{U}_1, \boldsymbol{U}_2\end{bmatrix} \begin{bmatrix}\boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} \begin{bmatrix}\boldsymbol{V}_1^* \\ \boldsymbol{V}_2^*\end{bmatrix} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^* = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*.$$

□

SVF and SVD are not unique. The singular values are unique since they are the nonnegative square roots of the eigenvalues of $\boldsymbol{A}^*\boldsymbol{A}$. However the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are in general not uniquely given.

7.1.2 Examples

Example 7.5 (Nonsingular matrix) *Derive the following SVD.*

$$\boldsymbol{A} := \frac{1}{25}\begin{bmatrix}11 & 48\\ 48 & 39\end{bmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \frac{1}{5}\begin{bmatrix}3 & -4\\ 4 & 3\end{bmatrix}\begin{bmatrix}3 & 0\\ 0 & 1\end{bmatrix}\frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}. \tag{7.8}$$

Discussion: The eigenpairs of $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \left[\begin{smallmatrix}97 & 96\\ 96 & 153\end{smallmatrix}\right]/25$ *are*

$$\boldsymbol{B}\begin{bmatrix}3\\4\end{bmatrix} = 9\begin{bmatrix}3\\4\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}4\\-3\end{bmatrix} = \begin{bmatrix}4\\-3\end{bmatrix}.$$

Taking square roots and normalizing we find $\sigma_1 = 3$, $\sigma_2 = 1$, $\boldsymbol{v}_1 = \left[\begin{smallmatrix}3\\4\end{smallmatrix}\right]/5$, $\boldsymbol{v}_2 = \left[\begin{smallmatrix}4\\-3\end{smallmatrix}\right]/5$. *Thus* $\boldsymbol{u}_1 := \boldsymbol{A}\boldsymbol{v}_1/\sigma_1 = \left[\begin{smallmatrix}3\\4\end{smallmatrix}\right]/5$ *and* $\boldsymbol{u}_2 := \boldsymbol{A}\boldsymbol{v}_2/\sigma_2 = \left[\begin{smallmatrix}-4\\3\end{smallmatrix}\right]/5$ *and this shows* (7.8). *Since* $m = n = r$ *we have* $\boldsymbol{U}_1 = \boldsymbol{U}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$ *and* $\boldsymbol{V}_1 = \boldsymbol{V}$. *In general the SVD and SVF are the same for a nonsingular matrix. See also Example 7.11 for some further discussion.*

Example 7.6 (Full row rank) *Find the singular value decomposition of*

$$\boldsymbol{A} := \frac{1}{15}\begin{bmatrix}14 & 4 & 16\\ 2 & 22 & 13\end{bmatrix} \in \mathbb{R}^{2,3}.$$

Discussion: The eigenpairs of $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \left[\begin{smallmatrix}8 & 4 & 10\\ 4 & 20 & 14\\ 10 & 14 & 17\end{smallmatrix}\right]/9$ *are*

$$\boldsymbol{B}\begin{bmatrix}1\\2\\2\end{bmatrix} = 4\begin{bmatrix}1\\2\\2\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}2\\-2\\1\end{bmatrix} = 1\begin{bmatrix}2\\-2\\1\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}2\\1\\-2\end{bmatrix} = 0\begin{bmatrix}2\\1\\-2\end{bmatrix}.$$

Thus $r = 2$ *and*

$$\boldsymbol{\Sigma} := \begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}, \quad \boldsymbol{V} := \frac{1}{3}\begin{bmatrix}1 & 2 & 2\\ 2 & -2 & 1\\ 2 & 1 & -2\end{bmatrix}.$$

From (7.6) *we find* $\boldsymbol{u}_1 = \boldsymbol{A}\boldsymbol{v}_1/\sigma_1 = [3,4]^T/5$, *and* $\boldsymbol{u}_2 = \boldsymbol{A}\boldsymbol{v}_2/\sigma_2 = [4,-3]^T/5$ *and*

$$\boldsymbol{U} = \frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}.$$

Since $r = 2$ *it follows that* $\operatorname{rank}(\boldsymbol{A}) = 2$, $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ *is an orthonormal basis for* $\operatorname{span}(\boldsymbol{A})$ *and* $\{\boldsymbol{v}_3\}$ *is an orthonormal basis for* $\ker(\boldsymbol{A})$. *The SVF and outer product form of* $\boldsymbol{A}$ *are*

$$\boldsymbol{A} = \frac{1}{5}\left[\begin{smallmatrix}3 & 4\\ 4 & -3\end{smallmatrix}\right]\left[\begin{smallmatrix}2 & 0\\ 0 & 1\end{smallmatrix}\right]\frac{1}{3}\left[\begin{smallmatrix}1 & 2 & 2\\ 2 & -2 & 1\end{smallmatrix}\right] = 2\frac{1}{15}\left[\begin{smallmatrix}3\\4\end{smallmatrix}\right]\left[\begin{smallmatrix}1 & 2 & 2\end{smallmatrix}\right] + 1\frac{1}{15}\left[\begin{smallmatrix}4\\-3\end{smallmatrix}\right]\left[\begin{smallmatrix}2 & -2 & 1\end{smallmatrix}\right].$$

Example 7.7 (Full column rank) *Find the SVD of*

$$\boldsymbol{A}_1 = \frac{1}{15}\begin{bmatrix}14 & 2\\ 4 & 22\\ 16 & 13\end{bmatrix} \in \mathbb{R}^{3,2}.$$

Since $\boldsymbol{A}_1 = \boldsymbol{A}^T$, where $\boldsymbol{A}$ is the matrix in Example 7.6 we can simply transpose the SVD of $\boldsymbol{A}$ in that example. Thus

$$\boldsymbol{A}_1 = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T = \frac{1}{3}\begin{bmatrix}1 & 2 & 2\\ 2 & -2 & 1\\ 2 & 1 & -2\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}\frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}. \tag{7.9}$$

Alternatively we can follow the recipe from the previous examples. The eigenpairs of

$$\boldsymbol{B}_1 = \boldsymbol{A}_1^T\boldsymbol{A}_1 = \frac{1}{25}\begin{bmatrix}52 & 36\\ 36 & 73\end{bmatrix}$$

are

$$\boldsymbol{B}\begin{bmatrix}3\\4\end{bmatrix} = 4\begin{bmatrix}3\\4\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}4\\-3\end{bmatrix} = 1\begin{bmatrix}4\\-3\end{bmatrix}.$$

Thus $\sigma_1 = 2$, $\sigma_2 = 1$. Now

$$\boldsymbol{u}_1 = \boldsymbol{A}_1\boldsymbol{v}_1/\sigma_1 = [1,2,2]^T/3, \quad \boldsymbol{u}_2 = \boldsymbol{A}_1\boldsymbol{v}_2/\sigma_1 = [2,-2,1]^T/3.$$

Since $m = 3$ we also need $\boldsymbol{u}_3$ which should be orthogonal to $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. $\boldsymbol{u}_3 = [2,1,-2]^T$ is such a vector and we obtain (7.9).

Example 7.8 ($r < n < m$) *Consider*

$$\boldsymbol{A} = \begin{bmatrix}1 & 1\\ 1 & 1\\ 0 & 0\end{bmatrix}.$$

For this matrix all the zero matrices in (7.5) *are nonempty. The eigenpairs of*

$$\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}$$

are

$$\boldsymbol{B}\begin{bmatrix}1\\1\end{bmatrix} = 4\begin{bmatrix}1\\1\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}1\\-1\end{bmatrix} = 0\begin{bmatrix}1\\-1\end{bmatrix}$$

and we find $\sigma_1 = 2$, $\sigma_2 = 0$, Thus $r = 1$, $m = 3$, $n = 2$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_1 & 0\\ 0 & 0\\ 0 & 0\end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = [2], \quad \boldsymbol{V} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}.$$

Now (7.6) *implies* $\boldsymbol{u}_1 = \boldsymbol{A}\boldsymbol{v}_1/\sigma_1 = \boldsymbol{s}_1/\sqrt{2}$, *where* $\boldsymbol{s}_1 = [1,1,0]^T$. *To find the other columns of* $\boldsymbol{U}$ *we extend* $\boldsymbol{s}_1$ *to a basis* $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$ *for* $\mathbb{R}^3$, *apply the Gram-Schmidt orthogonalization process to* $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$, *and then normalize. Choosing the basis*

$$\boldsymbol{s}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{s}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{s}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

we find from (A.16)

$$\boldsymbol{w}_1 = \boldsymbol{s}_1,\ \boldsymbol{w}_2 = \boldsymbol{s}_2 - \frac{\boldsymbol{s}_2^T\boldsymbol{w}_1}{\boldsymbol{w}_1^T\boldsymbol{w}_1}\boldsymbol{w}_1 = \begin{bmatrix} -1/2 \\ 1/2 \\ 0 \end{bmatrix},\ \boldsymbol{w}_3 = \boldsymbol{s}_3 - \frac{\boldsymbol{s}_3^T\boldsymbol{w}_1}{\boldsymbol{w}_1^T\boldsymbol{w}_1}\boldsymbol{w}_1 - \frac{\boldsymbol{s}_3^T\boldsymbol{w}_2}{\boldsymbol{w}_2^T\boldsymbol{w}_2}\boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Normalizing the $\boldsymbol{w}_i$*'s we obtain* $\boldsymbol{u}_1 = \boldsymbol{s}_1/\|\boldsymbol{s}_1\|_2 = [1/\sqrt{2}, 1/\sqrt{2}, 0]^T$, $\boldsymbol{u}_2 = \boldsymbol{s}_2/\|\boldsymbol{s}_2\|_2 = [-1/\sqrt{2}, 1/\sqrt{2}, 0]^T$, *and* $\boldsymbol{u}_3 = \boldsymbol{s}_3/\|\boldsymbol{s}_3\|_2 = [0,0,1]^T$. *Therefore,* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, *where*

$$\boldsymbol{U} := \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3,3}, \quad \boldsymbol{\Sigma} := \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{3,2}, \quad \boldsymbol{V} := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \in \mathbb{R}^{2,2}.$$

Exercise 7.9 *Find the singular value decomposition of the following matrices*

(a) $A = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$.

(b) $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}$.

Exercise 7.10 *Find the singular value decomposition of the following matrices*

(a) $A = \boldsymbol{e}_1$ *the first unit vector in* $\mathbb{R}^m$.

(b) $A = \boldsymbol{e}_n^T$ *the last unit vector in* $\mathbb{R}^n$.

(c) $\boldsymbol{A} = \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}$.

The method we used to find the singular value decomposition in the previous examples and exercises can be suitable for hand calculation with small matrices, but it is not appropriate as a basis for a general purpose numerical method. In particular, the Gram-Schmidt orthogonalization process is not numerically stable, and forming $\boldsymbol{A}^*\boldsymbol{A}$ can lead to extra errors in the computation. State of the art computer implementations of the singular value decomposition use an adapted version of the QR algorithm where the matrix $\boldsymbol{A}^*\boldsymbol{A}$ is not formed. The QR algorithm is discussed in Chapter 15.

7.1.3 Singular Values of Normal and Positive Semidefinite Matrices

The singular values of a normal matrix are the absolute values of its eigenvalues. For if $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is normal with eigenvalues $\lambda_1, \ldots, \lambda_n$ ordered so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$, then it follows from Theorem 6.10 that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$, where $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$, and $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix. We find $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}^*\boldsymbol{D}\boldsymbol{U}^*$, where $\boldsymbol{D}^*\boldsymbol{D} = \operatorname{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$. It follows that $\sigma_i^2 = |\lambda_i|^2$ or $\sigma_i = |\lambda_i|$ for $i = 1, \ldots, n$.

For a symmetric positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ the singular values are identical to the eigenvalues. The factorization $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ above is both a SVD decomposition and factorization provided we have sorted the nonnegative eigenvalues in nondecreasing order.

Example 7.11 *The matrix $\boldsymbol{A}$ in Example 7.5 is normal so that the singular values of $\boldsymbol{A}$ are equal to the absolute value of the eigenvalues of $\boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}$ are $\lambda_1 = 3$ and $\lambda_2 = -1$. Thus $\lambda_2 \neq \sigma_2$.*

7.1.4 A Geometric Interpretation

The singular value decomposition gives insight into the geometry of a linear transformation. Consider the linear transformation $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$ given by $\boldsymbol{z} \to \boldsymbol{A}\boldsymbol{z}$. The function $\boldsymbol{T}$ maps the unit sphere $\mathcal{S} := \{\boldsymbol{z} \in \mathbb{R}^n : \|\boldsymbol{z}\|_2 = 1\}$ onto an ellipsoid in $\mathbb{R}^m$. The singular values are the length of the semiaxes. We describe this in the square nonsingular case. Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the singular value decomposition of $\boldsymbol{A}$. Since $\boldsymbol{A}$ has rank n we have $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ and $\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$. Let $\mathcal{E} := \boldsymbol{A}\mathcal{S} \subset \mathbb{C}^n$ be the image of $\mathcal{S}$ under the transformation $\boldsymbol{T}$. If $\boldsymbol{x} \in \mathcal{E}$ then $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$ for some $\boldsymbol{z} \in \mathcal{S}$ and we find

$$\begin{aligned} 1 = \|\boldsymbol{z}\|_2^2 = \|\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{z}\|_2^2 = \|\boldsymbol{A}^{-1}\boldsymbol{x}\|_2^2 = \|\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{x}\|_2^2 \\ = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{x}\|_2^2 = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\|_2^2 = \frac{y_1^2}{\sigma_1^2} + \cdots + \frac{y_n^2}{\sigma_n^2}, \end{aligned}$$

where $\boldsymbol{y} := \boldsymbol{U}^T\boldsymbol{x}$ and we used $\|\boldsymbol{V}\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$ for a vector $\boldsymbol{v}$. The equation $1 = \frac{y_1^2}{\sigma_1^2} + \cdots + \frac{y_n^2}{\sigma_n^2}$ describes an ellipsoid in $\mathbb{R}^n$ with semiaxes of length σ_j along the unit vectors $\boldsymbol{e}_j$ for $j = 1, \ldots, n$. Since the orthogonal transformation $\boldsymbol{U}\boldsymbol{y} \to \boldsymbol{x}$ preserves length, the image $\mathcal{E} = \boldsymbol{A}\mathcal{S}$ is an ellipsoid with semiaxes along the left singular vectors $\boldsymbol{u}_j = \boldsymbol{U}\boldsymbol{e}_j$ of length σ_j. Since $\boldsymbol{A}\boldsymbol{v}_j = \sigma_j\boldsymbol{u}_j$, the right singular vectors are orthogonal points in $\mathcal{S}$ that are mapped onto the semiaxes of $\mathcal{E}$.

Example 7.12 *Consider the transformation $\boldsymbol{A} : \mathbb{R}^2 \to \mathbb{R}^2$ given by the matrix*

$$\boldsymbol{A} := \frac{1}{25}\begin{bmatrix} 11 & 48 \\ 48 & 39 \end{bmatrix}$$

in Example 7.5. Recall that $\sigma_1 = 3$, $\sigma_2 = 1$, $\boldsymbol{u}_1 = [3, 4]^T/5$ and $\boldsymbol{u}_2 = [-4, 3]^T/5$. The ellipsoids $y_1^2/\sigma_1^2 + y_2^2/\sigma_2^2 = 1$ and $\mathcal{E} = \boldsymbol{A}\mathcal{S}$ are shown in Figure 7.1. Since

$\boldsymbol{y} = \boldsymbol{U}^T\boldsymbol{x} = [3/5x_1 + 4/5x_2, -4/5x_1 + 3/5x_2]^T$, *the equation for the ellipsoid on the right is*

$$\frac{(\frac{3}{5}x_1 + \frac{4}{5}x_2)^2}{9} + \frac{(-\frac{4}{5}x_1 + \frac{3}{5}x_2)^2}{1} = 1,$$

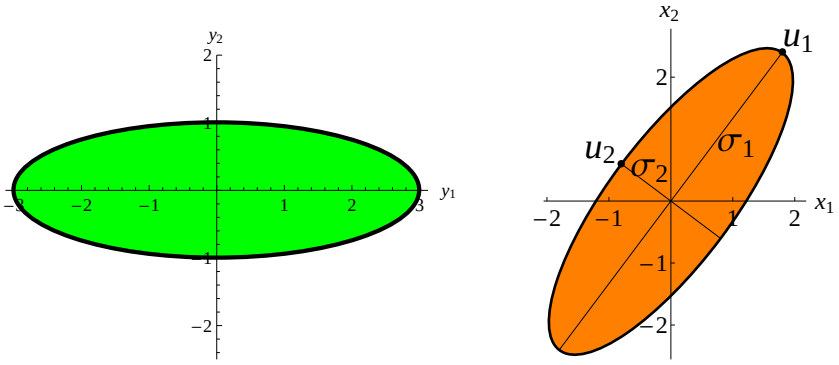


Figure 7.1. *The ellipse $y_1^2/9 + y_2^2 = 1$ (left) and the rotated ellipse $\boldsymbol{A}\mathcal{S}$ (right).*

7.2 Singular Vectors

The columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\boldsymbol{U}$ are called **left singular vectors**. and the columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of $\boldsymbol{V}$ are called **right singular vectors**. These vectors satisfy the following relations.

Theorem 7.13 *If $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the singular value decomposition of $\boldsymbol{A}$ then*

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \text{ and } \boldsymbol{A}^*\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^*. \tag{7.10}$$

If $\boldsymbol{U}$ and $\boldsymbol{V}$ are partitioned as in (7.7) *then*

$$\begin{array}{lll}
1. \ \boldsymbol{A}\boldsymbol{V}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1, & \text{or } \boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i \text{ for } i = 1, \ldots, r, \\
2. \ \boldsymbol{A}\boldsymbol{V}_2 = \boldsymbol{0}, & \text{or } \boldsymbol{A}\boldsymbol{v}_i = \boldsymbol{0} \text{ for } i = r+1, \ldots, n, \\
3. \ \boldsymbol{A}^*\boldsymbol{U}_1 = \boldsymbol{V}_1\boldsymbol{\Sigma}_1, & \text{or } \boldsymbol{A}^*\boldsymbol{u}_i = \sigma_i\boldsymbol{v}_i \text{ for } i = 1, \ldots, r, \\
4. \ \boldsymbol{A}^*\boldsymbol{U}_2 = \boldsymbol{0}, & \text{or } \boldsymbol{A}^*\boldsymbol{u}_i = \boldsymbol{0} \text{ for } i = r+1, \ldots, m.
\end{array} \tag{7.11}$$

Proof. Since $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}$ the first equation in (7.10) follows. Taking conjugate transposes and multiplying by $\boldsymbol{U}$ we have $\boldsymbol{A}^*\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^*$ and the second relation follows. In terms of partitioned matrices, (7.10) gives

$$\boldsymbol{A}\begin{bmatrix}\boldsymbol{V}_1, \boldsymbol{V}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{U}_1, \boldsymbol{U}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}, \quad \boldsymbol{A}^*\begin{bmatrix}\boldsymbol{U}_1, \boldsymbol{U}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{V}_1, \boldsymbol{V}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix},$$

and this leads to the equations in (7.11). ∎

Theorem 7.14 *The singular vectors of $\boldsymbol{A} \in \mathbb{C}^{m,n}$ are orthonormal bases for the four fundamental subspaces of $\boldsymbol{A}$. In particular*

$$\begin{aligned} &1.\ \boldsymbol{U}_1 \text{ is an orthonormal basis for } \operatorname{span}(\boldsymbol{A}),\\ &2.\ \boldsymbol{V}_2 \text{ is an orthonormal basis for } \ker(\boldsymbol{A}),\\ &3.\ \boldsymbol{V}_1 \text{ is an orthonormal basis for } \operatorname{span}(\boldsymbol{A}^*),\\ &4.\ \boldsymbol{U}_2 \text{ is an orthonormal basis for } \ker(\boldsymbol{A}^*). \end{aligned} \tag{7.12}$$

Proof. Since $\boldsymbol{u}_j = \boldsymbol{A}\boldsymbol{v}_j/\sigma_j$ for $j = 1, \ldots, r$ it follows from Theorem 7.2 that the first r left singular vectors of $\boldsymbol{A}$ form an orthonormal basis for $\operatorname{span}(\boldsymbol{A})$ and the last $n - r$ right singular vectors of $\boldsymbol{A}$ form an orthonormal basis for $\ker(\boldsymbol{A})$. The same holds for $\boldsymbol{A}^*$ and we have seen that the left and right singular vectors for $\boldsymbol{A}^*$ are the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively. □

By counting the number of columns in the four submatrices $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1, \boldsymbol{V}_2$, we obtain from Theorem 7.14 a new proof of the following fundamental result (Cf. Theorem B.16).

Corollary 7.15 *Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$. Then*

1. $\operatorname{rank}(\boldsymbol{A}) + \operatorname{null}(\boldsymbol{A}) = n$,
2. $\operatorname{rank}(\boldsymbol{A}) + \operatorname{null}(\boldsymbol{A}^*) = m$,
3. $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}^*)$.

Exercise 7.16 *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be as in Example 7.7. Give orthonormal bases for* $\operatorname{span}(\boldsymbol{B})$ *and* $\ker(\boldsymbol{B})$*and explain why* $\operatorname{span}(\boldsymbol{B}) \oplus \ker(mA)$ *is an orthogonal decomposition of* $\mathbb{R}^3$.

7.2.1 The SVD of $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$

The singular value decomposition of $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$ is related to the spectral decomposition of these matrices.

Theorem 7.17 *Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$ is the singular value decomposition and factorization of $\boldsymbol{A}$. Then a singular value decomposition and factorization of the matrices $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$ are given by*

$$\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^2\boldsymbol{V}_1^* \text{ and } \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\boldsymbol{U}^* = \boldsymbol{U}_1\boldsymbol{\Sigma}_1^2\boldsymbol{U}_1^*. \tag{7.13}$$

Moreover,

$$\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{V}_1 = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^2, \quad \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{V}_2 = \boldsymbol{0}, \tag{7.14}$$

and

$$\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{U}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1^2, \quad \boldsymbol{A}\boldsymbol{A}^*\boldsymbol{U}_2 = \boldsymbol{0}, \tag{7.15}$$

Proof. We compute $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\boldsymbol{V}^*$ and

$$\boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\boldsymbol{V}^* = [\boldsymbol{V}_1, \boldsymbol{V}_2] \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^* \\ \boldsymbol{V}_2^* \end{bmatrix} = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^2\boldsymbol{V}_1^*,$$

with an analogous computation for $\boldsymbol{A}\boldsymbol{A}^*$. The equation (7.15) follows from the computation $\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{U}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1^2\boldsymbol{U}_1^*\boldsymbol{U}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1^2$, $\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{U}_2 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1^2\boldsymbol{U}_1^*\boldsymbol{U}_2 = \boldsymbol{0}$. The proof of (7.14) is analogous. □

Theorem 7.17 leads to

Theorem 7.18 *For any $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have*

1. $\operatorname{rank}\boldsymbol{A} = \operatorname{rank}(\boldsymbol{A}^*\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}\boldsymbol{A}^*)$,
2. $\operatorname{null}(\boldsymbol{A}^*\boldsymbol{A}) = \operatorname{null}\boldsymbol{A}$, *and* $\operatorname{null}(\boldsymbol{A}\boldsymbol{A}^*) = \operatorname{null}(\boldsymbol{A}^*)$,
3. $\operatorname{span}(\boldsymbol{A}^*\boldsymbol{A}) = \operatorname{span}(\boldsymbol{A}^*)$ *and* $\ker(\boldsymbol{A}^*\boldsymbol{A}) = \ker(\boldsymbol{A})$.

Proof. The three matrices $\boldsymbol{A}$, $\boldsymbol{A}^*\boldsymbol{A}$, and $\boldsymbol{A}\boldsymbol{A}^*$ have the same number of nonzero singular values and we obtain 1. Moreover, 2. and 3. follow from Corollary 7.15 and (7.12), respectively, applied to $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$. □

Exercise 7.19 *Let $\boldsymbol{A} \in \mathbb{C}^{m,n}$ with $m \geq n$ have singular values $\sigma_1, \ldots, \sigma_n$, left singular vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m \in \mathbb{C}^m$, and right singular vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{C}^n$. Show that the matrix*

$$\boldsymbol{C} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix}$$

has the $n+m$ eigenpairs

$$\{(\sigma_1, \boldsymbol{p}_1), \ldots, (\sigma_n, \boldsymbol{p}_n)\}, \{(-\sigma_1, \boldsymbol{q}_1), \ldots, (-\sigma_n, \boldsymbol{q}_n)\}, \{(0, \boldsymbol{r}_{n+1}), \ldots, (0, \boldsymbol{r}_m)\},$$

where

$$\boldsymbol{p}_i = \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix}, \quad \boldsymbol{q}_i = \begin{bmatrix} \boldsymbol{u}_i \\ -\boldsymbol{v}_i \end{bmatrix}, \quad \boldsymbol{r}_j = \begin{bmatrix} \boldsymbol{u}_j \\ \boldsymbol{0} \end{bmatrix}, \text{ for } i = 1, \ldots, n \text{ and } j = n+1, \ldots, m.$$

7.3 Determining the Rank of a Matrix

In many elementary linear algebra courses a version of Gaussian elimination, called Gauss-Jordan elimination, is used to determine the rank of a matrix. To carry this out by hand for a large matrix can be a Herculean task and using a computer and floating point arithmetic the result will not be reliable. Entries, which in the final result should have been zero, will have nonzero values because of round-off errors. As an alternative we can use the singular value decomposition to determine rank. Although success is not at all guaranteed, the result will be more reliable than if Gauss-Jordan elimination is used.

By Theorem 7.3 the rank of a matrix is equal to the number of nonzero singular values and if we have computed the singular values, then all we have to do is to count

the nonzero ones. The problem however is the same as for Gaussian elimination. Due to round-off errors none of the computed singular values are likely to be zero.

The following discussion can be used to decide how many of the computed singular values one can set equal to zero. Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ with $m \geq n$ has singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\left[\begin{smallmatrix}\boldsymbol{\Sigma}\\ \boldsymbol{0}\end{smallmatrix}\right]\boldsymbol{V}^*$, where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. We choose $\epsilon > 0$ and let $1 \leq r \leq n$ be the smallest integer such that $\sigma_{r+1}^2 + \cdots + \sigma_n^2 < \epsilon^2$. Define $\boldsymbol{A}' := \boldsymbol{U}\left[\begin{smallmatrix}\boldsymbol{\Sigma}'\\ \boldsymbol{0}\end{smallmatrix}\right]\boldsymbol{V}^*$, where $\boldsymbol{\Sigma}' := \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \in \mathbb{R}^{n,n}$. By Theorem 8.11

$$\|\boldsymbol{A} - \boldsymbol{A}'\|_F = \left\| \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Sigma}' \\ \boldsymbol{0} \end{bmatrix} \right\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2} < \epsilon.$$

Thus $\boldsymbol{A}$ is near a matrix $\boldsymbol{A}'$ of rank r. This can be used to determine rank numerically. We choose an r such that $\sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2}$ is "small". Then we postulate that $\operatorname{rank}(\boldsymbol{A}) = r$ since $\boldsymbol{A}$ is close to a matrix of rank r.

The following theorem shows that of all $m \times n$ matrices of rank r, $\boldsymbol{A}'$ is closest to $\boldsymbol{A}$ measured in the Frobenius norm.

Theorem 7.20 (Best low rank approximation) *Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. For any $r \leq \operatorname{rank}(\boldsymbol{A})$ we have*

$$\|\boldsymbol{A} - \boldsymbol{A}'\|_F = \min_{\substack{\boldsymbol{B} \in \mathbb{R}^{m,n} \\ rank(\boldsymbol{B}) = r}} \|\boldsymbol{A} - \boldsymbol{B}\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2}.$$

For the proof of this theorem we refer to p. 322 of [20].

Exercise 7.21 *Consider the singular value decomposition*

$$A := \begin{bmatrix} 0 & 3 & 3 \\ 4 & 1 & -1 \\ 4 & 1 & -1 \\ 0 & 3 & 3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \end{bmatrix}$$

(a) *Give orthonormal bases for* $\operatorname{span}(A), \operatorname{span}(A^T), \ker(A), \ker(A^T)$ *and* $\operatorname{span}(A)^\perp$.

(b) *Explain why for all matrices $B \in \mathbb{R}^{4,3}$ of rank one we have $\|A - B\|_F \geq 6$.*

(c) *Give a matrix A_1 of rank one such that $\|A - A_1\|_F = 6$.*

Exercise 7.22 *Let $\boldsymbol{A}$ be the $n \times n$ matrix that for $n = 4$ takes the form*

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Thus $\boldsymbol{A}$ is upper triangular with diagonal elements one and all elements above the diagonal equal to -1. Let $\boldsymbol{B}$ be the matrix obtained from $\boldsymbol{A}$ by changing the $(n, 1)$ element from zero to -2^{2-n}.

(a) *Show that* $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$*, where* $\boldsymbol{x} := [2^{n-2}, 2^{n-3}, \ldots, 2^0, 1]^T$*. Conclude that* $\boldsymbol{B}$ *is singular,* $\det(\boldsymbol{A}) = 1$*, and* $\|\boldsymbol{A} - \boldsymbol{B}\|_F = 2^{2-n}$*. Thus even if* $\det(\boldsymbol{A})$ *is not small the matrix* $\boldsymbol{A}$ *is very close to being singular for large* n*.*

(b) *Use Theorem 7.20 to show that the smallest singular vale* σ_n *of* $\boldsymbol{A}$ *is bounded above by* 2^{2-n}*.*

7.4 The Minmax Theorem for Singular Values and the Hoffman-Wielandt Theorem

We have a minmax and maxmin characterization for singular values.

Theorem 7.23 (The Courant-Fischer Theorem for Singular Values) *Suppose* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *has singular values* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *ordered so that* $\sigma_1 \geq \cdots \geq \sigma_n$*. Then for* $k = 1, \ldots, n$

$$\sigma_k = \min_{\dim(\mathcal{S})=n-k+1} \max_{\substack{\boldsymbol{x}\in\mathcal{S} \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\dim(\mathcal{S})=k} \min_{\substack{\boldsymbol{x}\in\mathcal{S} \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}. \tag{7.16}$$

Proof. Since

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = \frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} = \frac{\langle \boldsymbol{x}, \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}$$

denotes the Rayleigh quotient $R_{\boldsymbol{A}^*\boldsymbol{A}}(\boldsymbol{x})$ of $\boldsymbol{A}^*\boldsymbol{A}$, and since the singular values of $\boldsymbol{A}$ are the nonnegative square roots of the eigenvalues of $\boldsymbol{A}^*\boldsymbol{A}$, the results follow from the Courant-Fischer Theorem for eigenvalues, see Theorem 6.14. □

By taking $k = 1$ and $k = n$ in (7.16) we obtain for any $\boldsymbol{A} \in \mathbb{C}^{m,n}$

$$\sigma_1 = \max_{\substack{\boldsymbol{x}\in\mathbb{C}^n \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}, \quad \sigma_n = \min_{\substack{\boldsymbol{x}\in\mathbb{C}^n \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}. \tag{7.17}$$

This follows since the only subspace of $\mathbb{C}^n$ of dimension n is $\mathbb{C}^n$ itself.

The Hoffman-Wielandt Theorem for eigenvalues of Hermitian matrices, Theorem 6.18 can be written

$$\sum_{j=1}^{n} |\mu_j - \lambda_j|^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2 := \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij} - b_{ij}|^2, \tag{7.18}$$

where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$ are both Hermitian matrices with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\mu_1 \geq \cdots \geq \mu_n$, respectively.

For singular values we have a similar result.

Theorem 7.24 (Hoffman-Wielandt Theorem for singular values) *For any* $m, n \in \mathbb{N}$ *and* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m,n}$ *we have*

$$\sum_{j=1}^{n} |\beta_j - \alpha_j|^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2. \tag{7.19}$$

where $\alpha_1 \geq \cdots \geq \alpha_n$ *and* $\beta_1 \geq \cdots \geq \beta_n$ *are the singular values of* $\boldsymbol{A}$ *and* $\boldsymbol{B}$*, respectively.*

Proof. We apply the Hoffman-Wielandt Theorem for eigenvalues to the Hermitian matrices

$$\boldsymbol{C} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix} \text{ and } \boldsymbol{D} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{0} \end{bmatrix} \in \mathbb{C}^{m+n,m+n}.$$

If $\boldsymbol{C}$ and $\boldsymbol{D}$ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{m+n}$ and $\mu_1 \geq \cdots \geq \mu_{m+n}$, respectively then

$$\sum_{j=1}^{m+n} |\lambda_j - \mu_j|^2 \leq \|\boldsymbol{C} - \boldsymbol{D}\|_F^2. \tag{7.20}$$

Suppose $\boldsymbol{A}$ has rank r and SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. We use (7.11) and determine the eigenpairs of $\boldsymbol{C}$ as follows.

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{v}_i \\ \boldsymbol{A}^*\boldsymbol{u}_i \end{bmatrix} = \begin{bmatrix} \alpha_i \boldsymbol{u}_i \\ \alpha_i \boldsymbol{v}_i \end{bmatrix} = \alpha_i \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix}, \quad i = 1, \ldots, r,$$

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i \\ -\boldsymbol{v}_i \end{bmatrix} = \begin{bmatrix} -\boldsymbol{A}\boldsymbol{v}_i \\ \boldsymbol{A}^*\boldsymbol{u}_i \end{bmatrix} = \begin{bmatrix} -\alpha_i \boldsymbol{u}_i \\ \alpha_i \boldsymbol{v}_i \end{bmatrix} = -\alpha_i \begin{bmatrix} \boldsymbol{u}_i \\ -\boldsymbol{v}_i \end{bmatrix}, \quad i = 1, \ldots, r,$$

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{A}^*\boldsymbol{u}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} = 0 \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{0} \end{bmatrix}, \quad i = r+1, \ldots, m,$$

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{v}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{v}_i \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} = 0 \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{v}_i \end{bmatrix}, \quad i = r+1, \ldots, n.$$

Thus $\boldsymbol{C}$ has the $2r$ eigenvalues $\alpha_1, -\alpha_1, \ldots, \alpha_r, -\alpha_r$ and $m+n-2r$ additional zero eigenvalues. Similarly, if $\boldsymbol{B}$ has rank s then $\boldsymbol{D}$ has the $2s$ eigenvalues $\beta_1, -\beta_1, \ldots, \beta_s, -\beta_s$ and $m+n-2s$ additional zero eigenvalues. Let

$$t := \max(r, s).$$

Then

$$\lambda_1 \geq \cdots \geq \lambda_{m+n} = \alpha_1 \geq \cdots \geq \alpha_t \geq 0 = \cdots = 0 \geq -\alpha_t \geq \cdots \geq -\alpha_1,$$
$$\mu_1 \geq \cdots \geq \mu_{m+n} = \beta_1 \geq \cdots \geq \beta_t \geq 0 = \cdots = 0 \geq -\beta_t \geq \cdots \geq -\beta_1.$$

We find

$$\sum_{j=1}^{m+n} |\lambda_j - \mu_j|^2 = \sum_{i=1}^{t} |\alpha_i - \beta_i|^2 + \sum_{i=1}^{t} |-\alpha_i + \beta_i|^2 = 2 \sum_{i=1}^{t} |\alpha_i - \beta_i|^2$$

and

$$\|\boldsymbol{C} - \boldsymbol{D}\|_F^2 = \left\| \begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} - \boldsymbol{B} \\ \boldsymbol{A}^* - \boldsymbol{B}^* & \boldsymbol{0} \end{bmatrix} \right\|_F^2 = \|\boldsymbol{B} - \boldsymbol{A}\|_F^2 + \|(\boldsymbol{B} - \boldsymbol{A})^*\|_F^2 = 2\|\boldsymbol{B} - \boldsymbol{A}\|_F^2.$$

But then (7.20) implies $\sum_{i=1}^{t} |\alpha_i - \beta_i|^2 \leq \|\boldsymbol{B} - \boldsymbol{A}\|_F^2$. Since $t \leq n$ and $\alpha_i = \beta_i = 0$ for $i = t+1, \ldots, n$ we obtain (7.19). □

Chapter 8

Matrix Norms

To measure the size of a matrix we can use a matrix norm. In this chapter we give a systematic study of matrix norms. We start with vector norms. For simplicity we consider only norms on $\mathbb{R}^n$ and $\mathbb{C}^n$.

8.1 Vector Norms

To measure the size of a vector we use norms.

Definition 8.1 (Norm) *A norm in $\mathbb{R}^n(\mathbb{C}^n)$ is a function $\|\cdot\| : \mathbb{R}^n(\mathbb{C}^n) \to \mathbb{R}$ that satisfies for all $\boldsymbol{x}, \boldsymbol{y}$ in $\mathbb{R}^n(\mathbb{C}^n)$ and all a in $\mathbb{R}(\mathbb{C})$*

1. $\|\boldsymbol{x}\| \geq 0$ *with equality if and only if* $\boldsymbol{x} = \boldsymbol{0}$. *(positivity)*
2. $\|a\boldsymbol{x}\| = |a| \, \|\boldsymbol{x}\|$. *(homogeneity)*
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$. *(subadditivity)*

The triples $(\mathbb{R}^n, \mathbb{R}, \|\cdot\|)$ and $(\mathbb{C}^n, \mathbb{C}, \|\cdot\|)$ are examples of **normed vector spaces** *and the inequality 3. is called the* **triangle inequality**.

Since $\|\boldsymbol{x}\| = \|\boldsymbol{x} - \boldsymbol{y} + \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{y}\|$ we obtain $\|\boldsymbol{x} - \boldsymbol{y}\| \geq \|\boldsymbol{x}\| - \|\boldsymbol{y}\|$. By symmetry $\|\boldsymbol{x} - \boldsymbol{y}\| = \|\boldsymbol{y} - \boldsymbol{x}\| \geq \|\boldsymbol{y}\| - \|\boldsymbol{x}\|$ and we obtain the inverse triangle inequality

$$\|\boldsymbol{x} - \boldsymbol{y}\| \geq | \, \|\boldsymbol{x}\| - \|\boldsymbol{y}\| \, |, \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n. \tag{8.1}$$

Consider now some specific vector norms. We define for $p \geq 1$ the p-norms by

$$\|\boldsymbol{x}\|_p := \big(\sum_{j=1}^{n} |x_j|^p \big)^{1/p}, \tag{8.2}$$

$$\|\boldsymbol{x}\|_\infty := \max_{1 \leq j \leq n} |x_j|. \tag{8.3}$$

The most important cases are:

1. $\|\boldsymbol{x}\|_1 = \sum_{j=1}^n |x_j|$, (**the one-norm** or l_1**-norm**)
2. $\|\boldsymbol{x}\|_2 = \left(\sum_{j=1}^n |x_j|^2\right)^{1/2}$, **the two-norm**, l_2**-norm**, or **Euclidian norm**)
3. $\|\boldsymbol{x}\|_\infty = \max_{1\le j\le n}|x_j|$, (**the infinity-norm**, l_∞**-norm**, or **max norm.**)

The infinity norm is related to the other p-norms by

$$\lim_{p\to\infty}\|\boldsymbol{x}\|_p = \|\boldsymbol{x}\|_\infty \text{ for all } \boldsymbol{x}\in\mathbb{C}^n. \tag{8.4}$$

This clearly holds for $\boldsymbol{x} = \boldsymbol{0}$. For $\boldsymbol{x} \neq \boldsymbol{0}$ we write

$$\|\boldsymbol{x}\|_p := \|\boldsymbol{x}\|_\infty \Big(\sum_{j=1}^n \Big(\frac{|x_j|}{\|\boldsymbol{x}\|_\infty}\Big)^p\Big)^{1/p}.$$

Now each term in the sum is not greater than one and at least one term is equal to one and we obtain

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_p \le n^{1/p}\|\boldsymbol{x}\|_\infty, \quad p\ge 1. \tag{8.5}$$

Since $\lim_{p\to\infty} n^{1/p} = 1$ for any $n\in\mathbb{N}$ we see that (8.4) follows.

It can be shown (cf. Appendix H) that the p-norm are norms in $\mathbb{R}^n$ and in $\mathbb{C}^n$ for any p with $1\le p\le\infty$. The triangle inequality $\|\boldsymbol{x}+\boldsymbol{y}\|_p \le \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$ is called **Minkowski's inequality** . To prove it one first establishes **Hölder's inequality**

$$\sum_{j=1}^n |x_j y_j| \le \|\boldsymbol{x}\|_p\|\boldsymbol{y}\|_q, \quad \frac{1}{p}+\frac{1}{q} = 1, \quad \boldsymbol{x},\boldsymbol{y}\in\mathbb{C}^n. \tag{8.6}$$

The relation $\frac{1}{p}+\frac{1}{q}=1$ means that if $p=1$ then $q=\infty$ and if $p=2$ then $q=2$.

(8.5) shows that the infinity norm and any other p-norm can be bounded in terms of each other. We define

Definition 8.2 *Two norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{C}^n$ are equivalent if there are positive constants m and M (depending only on n such that for all vectors $\boldsymbol{x}\in\mathbb{C}^n$ we have*

$$m\|\boldsymbol{x}\| \le \|\boldsymbol{x}\|' \le M\|\boldsymbol{x}\|. \tag{8.7}$$

The following result is proved in Appendix H.

Theorem 8.3 *All vector norms in $\mathbb{C}^n$ are equivalent.*

The inverse triangle inequality (8.1) shows that a norm is a continuous function $\mathbb{C}^n\to\mathbb{R}$.

Exercise 8.4 *Show that $\|\cdot\|p$ is a vector norm in $\mathbb{R}^n$ for $p=1$, $p=\infty$.*

Exercise 8.5 *The set*

$$S_p = \{\boldsymbol{x}\in\mathbb{R}^n : \|\boldsymbol{x}\|_p = 1\}$$

is called the unit sphere in $\mathbb{R}^n$ with respect to p. Draw S_p for $p=1,2,\infty$ for $n=2$.

Exercise 8.6 *Let $1 \leq p$. Produce a vector $\boldsymbol{x}_l$ such that $\|\boldsymbol{x}_l\|_\infty = \|\boldsymbol{x}_l\|_p$ and another vector $\boldsymbol{x}_u$ such that $\|\boldsymbol{x}_u\|_p = n^{1/p}\|\boldsymbol{x}_u\|_p\infty$. Thus the inequalities in (8.4) are sharp.*

Exercise 8.7 *If $1 \leq q \leq p \leq \infty$ then*

$$\|\boldsymbol{x}\|_p \leq \|\boldsymbol{x}\|_q \leq n^{1/q-1/p}\|\boldsymbol{x}\|_p, \quad \boldsymbol{x} \in \mathbb{C}^n.$$

Hint: For the rightmost inequality use Jensen's inequality Cf. Theorem H.2 with $f(z) = z^{p/q}$ and $z_i = |x_i|^q$. For the left inequality consider first $y_i = x_i/\|\boldsymbol{x}\|_\infty$, $i = 1, 2, \ldots, n$.

8.2 Matrix Norms

For simplicity we consider only matrix norms on the vector space $(\mathbb{C}^{m,n}, \mathbb{C})$. All results also holds for $(\mathbb{R}^{m,n}, \mathbb{R})$.

Definition 8.8 (Matrix Norms) *Suppose m, n are positive integers. A function $\|\cdot\| \colon \mathbb{C}^{m,n} \to \mathbb{R}$ is called a* **matrix norm** *on $\mathbb{C}^{m,n}$ if for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m,n}$ and all $c \in \mathbb{C}$*

1. $\|\boldsymbol{A}\| \geq 0$ *with equality if and only if $\boldsymbol{A} = 0$.* *(positivity)*
2. $\|c\boldsymbol{A}\| = |c|\,\|\boldsymbol{A}\|$. *(homogeneity)*
3. $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$. *(subadditivity)*

A matrix norm is simply a vector norm on the finite dimensional vector space $(\mathbb{C}^{m,n}, \mathbb{C})$ of $m \times n$ matrices. Adapting Theorem 8.3 to this special situation gives

Theorem 8.9 *All matrix norms are equivalent. Thus, if $\|\cdot\|$ and $\|\cdot\|'$ are two matrix norms on $\mathbb{C}^{m,n}$ then there are positive constants μ and M such that*

$$\mu\|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|' \leq M\|\boldsymbol{A}\|$$

holds for all $\boldsymbol{A} \in \mathbb{C}^{m,n}$. Moreover, a matrix norm is a continuous function.

8.2.1 The Frobenius Norm

From any vector norm $\|\ \|_V$ on $\mathbb{C}^{mn}$ we can define a matrix norm on $\mathbb{C}^{m,n}$ by $\|\boldsymbol{A}\| := \|\mathrm{vec}(\boldsymbol{A})\|_V$, where $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{C}^{mn}$ is the vector obtained by stacking the columns of $\boldsymbol{A}$ on top of each other. In particular, to the p vector norms for $p = 1, 2, \infty$, we have the corresponding **sum norm**, **Frobenius norm**, and **max norm** defined by

$$\|\boldsymbol{A}\|_S := \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|, \quad \|\boldsymbol{A}\|_F := \Big(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\Big)^{1/2}, \quad \|\boldsymbol{A}\|_M := \max_{i,j}|a_{ij}|. \tag{8.8}$$

Of these norms the Frobenius norm is the most useful. It satisfies the following properties.

Lemma 8.10 *For any matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have*

1. $\|\boldsymbol{A}^*\|_F = \|\boldsymbol{A}\|_F$,
2. $\|\boldsymbol{A}\|_F^2 = \sum_{j=1}^n \|\boldsymbol{a}_{:j}\|_2^2 = \sum_{i=1}^m \|\boldsymbol{a}_{i:}\|_2^2$,
3. $\|\boldsymbol{U}\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{V}\|_F = \|\boldsymbol{A}\|_F$ *for any unitary matrices* $\boldsymbol{U} \in \mathbb{C}^{m,m}$ *and* $\boldsymbol{V} \in \mathbb{C}^{n,n}$,
4. $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F$ *for any* $\boldsymbol{B} \in \mathbb{C}^{n,k}$,
5. $\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A}\|_F \|\boldsymbol{x}\|_2$, *for all* $\boldsymbol{x} \in \mathbb{C}^n$.

Proof.

1. $\|\boldsymbol{A}^*\|_F^2 = \sum_{j=1}^n \sum_{i=1}^m |\overline{a}_{ij}|^2 = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 = \|\boldsymbol{A}\|_F^2$.
2. Obvious.
3. Recall that $\|\boldsymbol{U}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \mathbb{C}^n$ if $\boldsymbol{U}^*\boldsymbol{U} = I$. Applying this to each column $\boldsymbol{a}_{:j}$ of $\boldsymbol{A}$ we find $\|\boldsymbol{U}\boldsymbol{A}\|_F^2 \overset{2.}{=} \sum_{j=1}^n \|\boldsymbol{U}\boldsymbol{a}_{:j}\|_2^2 = \sum_{j=1}^n \|\boldsymbol{a}_{:j}\|_2^2 \overset{2.}{=} \|\boldsymbol{A}\|_F^2$. Similarly, since $\boldsymbol{V}\boldsymbol{V}^* = I$ we find $\|\boldsymbol{A}\boldsymbol{V}\|_F \overset{1.}{=} \|\boldsymbol{V}^*\boldsymbol{A}^*\|_F = \|\boldsymbol{A}^*\|_F \overset{1.}{=} \|\boldsymbol{A}\|_F$.
4. Using Cauchy-Schwarz' inequality and 2. we obtain

$$\|\boldsymbol{A}\boldsymbol{B}\|_F^2 = \sum_{i=1}^n \sum_{j=1}^k \left(\boldsymbol{a}_{i:}^T \boldsymbol{b}_{:j}\right)^2 \leq \sum_{i=1}^n \sum_{j=1}^k \|\boldsymbol{a}_{i:}\|_2^2 \|\boldsymbol{b}_{:j}\|_2^2 = \|\boldsymbol{A}\|_F^2 \|\boldsymbol{B}\|_F^2.$$

5. Since $\|\boldsymbol{v}\|_F = \|\boldsymbol{v}\|_2$ for a vector this follows by taking $k = 1$ and $\boldsymbol{B} = \boldsymbol{x}$ in 4.

□

There is a relation between the Frobenius norm and the singular values.

Theorem 8.11 *We have $\|\boldsymbol{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$, where $\sigma_1, \ldots, \sigma_n$ are the singular values of $\boldsymbol{A}$.*

Proof. Using Lemma 8.10 we find $\|\boldsymbol{A}\|_F \overset{3.}{=} \|\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{V}\|_F = \|\boldsymbol{\Sigma}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$.
□

8.2.2 Consistent and Subordinate Matrix Norms

Since matrices can be multiplied it is useful to have an analogue of subadditivity for matrix multiplication. For square matrices the product $\boldsymbol{A}\boldsymbol{B}$ is defined in a fixed space $\mathbb{C}^{n,n}$, while in the rectangular case matrix multiplication combines matrices in different spaces. The following definition captures this distinction.

Definition 8.12 (Consistent Matrix Norms) *A matrix norm is called* **consistent on** $\mathbb{C}^{n,n}$ *if*

4. $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$ *(submultiplicativity)*

holds for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$*. A matrix norm is* **consistent** *if it is defined on* $\mathbb{C}^{m,n}$ *for all* $m, n \in \mathbb{N}$*, and 4. holds for all matrices* $\boldsymbol{A}, \boldsymbol{B}$ *for which the product* $\boldsymbol{AB}$ *is defined.*

Clearly the three norms in (8.8) are defined for all $m, n \in \mathbb{N}$. From Lemma 8.10 it follows that the Frobenius norm is consistent.

Exercise 8.13 *Show that the sum norm is consistent.*

Exercise 8.14 *Show that the max norm is not consistent by considering* $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$.

Exercise 8.15

(a) *Show that the norm*

$$\|\boldsymbol{A}\| := \sqrt{mn}\|\boldsymbol{A}\|_M, \quad \boldsymbol{A} \in \mathbb{C}^{m,n}$$

is a consistent matrix norm.

(b) *Show that the constant* $\sqrt{mn}$ *can be replaced by* m *and by* n*.*

For a consistent matrix norm on $\mathbb{C}^{n,n}$ we have the inequality

$$\|\boldsymbol{A}^k\| \leq \|\boldsymbol{A}\|^k \text{ for } k \in \mathbb{N}. \tag{8.9}$$

When working with norms we often have to bound the vector norm of a matrix times a vector by the norm of the matrix times the norm of the vector. We have the following definition.

Definition 8.16 (Subordinate Matrix Norms) *Suppose* $m, n \in \mathbb{N}$ *are given, let* $\|\ \|_\alpha$ *on* $\mathbb{C}^m$ *and* $\|\ \|_\beta$ *on* $\mathbb{C}^n$ *be vector norms, and let* $\|\ \|$ *be a matrix norm on* $\mathbb{C}^{m,n}$*. We say that the matrix norm* $\|\ \|$ *is* **subordinate** *to the vector norms* $\|\ \|_\alpha$ *and* $\|\ \|_\beta$ *if* $\|\boldsymbol{Ax}\|_\alpha \leq \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|_\beta$ *for all* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *and all* $\boldsymbol{x} \in \mathbb{C}^n$*. If* $\|\ \|_\alpha = \|\ \|_\beta$ *then we say that* $\|\ \|$ *is subordinate to* $\|\ \|_\alpha$*.*

By Lemma 8.10 we have $\|\boldsymbol{Ax}\|_2 \leq \|\boldsymbol{A}\|_F\|\boldsymbol{x}\|_2$, for all $\boldsymbol{x} \in \mathbb{C}^n$. Thus the Frobenius norm is subordinate to the Euclidian vector norm.

Exercise 8.17 *Show that the sum norm is subordinate to the* l_1*-norm.*

Exercise 8.18 **(a)** *Show that the max norm is subordinate to the* ∞ *and* 1 *norm, i. e.,* $\|\boldsymbol{Ax}\|_\infty \leq \|\boldsymbol{A}\|_M\|\boldsymbol{x}\|_1$ *holds for all* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *and all* $\boldsymbol{x} \in \mathbb{C}^n$*.*

(b) *Show that* $\|\boldsymbol{Ae}_l\|_\infty = \|\boldsymbol{A}\|_M\|\boldsymbol{e}_l\|_1$*, where* $\|\boldsymbol{A}\|_M = |a_{kl}|$*.*

(c) *Show that* $\|\boldsymbol{A}\|_M = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Ax}\|_\infty}{\|\boldsymbol{x}\|_1}$*.*

8.2.3 Operator Norms

Corresponding to vector norms on $\mathbb{C}^n$ and $\mathbb{C}^m$ there is an induced matrix norm on $\mathbb{C}^{m,n}$ which we call the **operator norm**.

Definition 8.19 (Operator Norm) *Suppose $m, n \in \mathbb{N}$ are given and let $\| \ \|_\alpha$ be a vector norm on $\mathbb{C}^m$ and $\| \ \|_\beta$ a vector norm on $\mathbb{C}^n$. For $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we define*

$$\|\boldsymbol{A}\| := \|\boldsymbol{A}\|_{\alpha,\beta} := \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\beta}. \tag{8.10}$$

We call this the (α, β) **operator norm***, the (α, β)-norm, or simply the α-norm if $\alpha = \beta$.*

Before we show that the (α, β)-norm is a matrix norm we make some observations.

1. It is enough to take the max over subsets of $\mathbb{C}^n$. For example

$$\|\boldsymbol{A}\|_{\alpha,\beta} = \max_{\boldsymbol{x}\notin\ker(\boldsymbol{A})} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\beta} = \max_{\|\boldsymbol{x}\|_\beta=1} \|\boldsymbol{A}\boldsymbol{x}\|_\alpha. \tag{8.11}$$

   That we only need to consider $\boldsymbol{x}$'s outside the null space $\ker(\boldsymbol{A})$ of $\boldsymbol{A}$ is obvious. We can take the max over the β-norm unit sphere in $\mathbb{C}^n$ since

$$\max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\beta} = \max_{\boldsymbol{x}\neq 0} \Big\|\boldsymbol{A}\big(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_\beta}\big)\Big\|_\alpha = \max_{\|\boldsymbol{x}\|_\beta=1} \|\boldsymbol{A}\boldsymbol{x}\|_\alpha.$$

2. The operator norm $\|\boldsymbol{A}\|$ is subordinate to the vector norms $\| \ \|_\alpha$ and $\| \ \|_\beta$. Thus

$$\|\boldsymbol{A}\boldsymbol{x}\|_\alpha \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|_\beta \text{ for all } \boldsymbol{A} \in \mathbb{C}^{m,n} \text{ and } \boldsymbol{x} \in \mathbb{C}^n. \tag{8.12}$$

3. We can use max instead of sup in (8.10). This follows by the following compactness argument. Since all vector norms on $\mathbb{C}^n$ are equivalent the unit sphere $\mathcal{S}_\beta := \{\boldsymbol{x} \in \mathbb{C}^n : \|\boldsymbol{x}\|_\beta = 1\}$ is bounded. It is also finite dimensional and closed, and hence compact. Moreover, since the vector norm $\| \ \|_\alpha$ is a continuous function, it follows that the function $f : \mathcal{S}_\beta \to \mathbb{R}$ given by $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|_\alpha$ is continuous. But then f attains its max and min and we have

$$\|\boldsymbol{A}\|_{\alpha,\beta} = \|\boldsymbol{A}\boldsymbol{x}^*\|_\alpha \text{ for some } \boldsymbol{x}^* \in \mathbb{C}^n \text{ with } \|\boldsymbol{x}^*\|_\beta = 1. \tag{8.13}$$

Lemma 8.20 *The operator norm given by* (8.10) *is a matrix norm on $\mathbb{C}^{m,n}$. The operator norm is consistent if the vector norm $\| \ \|_\alpha$ is defined for all $m \in \mathbb{N}$ and $\| \ \|_\beta = \| \ \|_\alpha$.*

Proof. We use (8.11). In 2. and 3. below we take the max over the unit sphere $\mathcal{S}_\beta$.

1. Nonnegativity is obvious. If $\|\boldsymbol{A}\| = 0$ then $\|\boldsymbol{A}\boldsymbol{y}\|_\beta = 0$ for each $\boldsymbol{y} \in \mathbb{C}^n$. In particular, each column $\boldsymbol{A}\boldsymbol{e}_j$ in $\boldsymbol{A}$ is zero. Hence $\boldsymbol{A} = 0$.

2. $\|c\boldsymbol{A}\| = \max_{\boldsymbol{x}} \|c\boldsymbol{A}\boldsymbol{x}\|_\alpha = \max_{\boldsymbol{x}} |c| \, \|\boldsymbol{A}\boldsymbol{x}\|_\alpha = |c| \, \|\boldsymbol{A}\|$.
3. $\|\boldsymbol{A}+\boldsymbol{B}\| = \max_{\boldsymbol{x}} \|(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}\|_\alpha \leq \max_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x}\|_\alpha + \max_{\boldsymbol{x}} \|\boldsymbol{B}\boldsymbol{x}\|_\alpha = \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$.
4. $\|\boldsymbol{A}\boldsymbol{B}\| = \max_{\boldsymbol{B}\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\alpha} = \max_{\boldsymbol{B}\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{B}\boldsymbol{x}\|_\alpha} \frac{\|\boldsymbol{B}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\alpha}$
$\leq \max_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{y}\|_\alpha}{\|\boldsymbol{y}\|_\alpha} \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{B}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\alpha} = \|\boldsymbol{A}\| \, \|\boldsymbol{B}\|$.

□

For any α-norm of the $n \times n$ identity matrix we find

$$\|\boldsymbol{I}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{I}\boldsymbol{x}\|_\alpha}{\|\boldsymbol{x}\|_\alpha} = \max_{\boldsymbol{x}\neq\boldsymbol{0}} 1 = 1.$$

For the Frobenius norm we find $\|\boldsymbol{I}\|_F = \sqrt{n}$, and this shows that the Frobenius norm is not an operator norm for $n > 1$.

8.2.4 The p-Norms

Recall that the p or ℓ_p vector norms (8.2) are given by

$$\|\boldsymbol{x}\|_p := \big(\sum_{j=1}^{n} |x_j|^p\big)^{1/p}, \; p \geq 1, \quad \|\boldsymbol{x}\|_\infty := \max_{1\leq j\leq n} |x_j|.$$

The operator norms $\| \, \|_p$ defined from these p-vector norms are used quite frequently for $p = 1, 2, \infty$. We define for any $1 \leq p \leq \infty$

$$\|\boldsymbol{A}\|_p := \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} = \max_{\|\boldsymbol{y}\|_p=1} \|\boldsymbol{A}\boldsymbol{y}\|_p. \tag{8.14}$$

In the most important cases we have explicit expressions for these norms.

Theorem 8.21 *For $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have*

$$\begin{aligned} \|\boldsymbol{A}\|_1 &:= \max_{1\leq j\leq n} \sum_{k=1}^{m} |a_{k,j}|, && \textit{(max column sum)} \\ \|\boldsymbol{A}\|_2 &:= \sigma_1, && \textit{(largest singular value of } \boldsymbol{A}\textit{)} \\ \|\boldsymbol{A}\|_\infty &:= \max_{1\leq k\leq m} \sum_{j=1}^{n} |a_{k,j}|, && \textit{(max row sum).} \end{aligned} \tag{8.15}$$

The expression $\|\boldsymbol{A}\|_2$ is called the **two-norm** or the **spectral norm** of $\boldsymbol{A}$.

Proof. The result for $p = 2$ follows from the minmax theorem for singular values. Indeed, by (7.17) we have $\sigma_1 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$. For $p = 1, \infty$ we do the following:

(a) We derive a constant K_p such that $\|\boldsymbol{A}\boldsymbol{x}\|_p \leq K_p$ for any $\boldsymbol{x} \in \mathbb{C}^n$ with $\|\boldsymbol{x}\|_p = 1$.

(b) We give an extremal vector $\boldsymbol{y}^* \in \mathbb{C}^n$ with $\|\boldsymbol{y}^*\|_p = 1$ so that $\|\boldsymbol{A}\boldsymbol{y}^*\|_p = K_p$.

It then follows from (8.14) that $\|\boldsymbol{A}\|_p = \|\boldsymbol{A}\boldsymbol{y}^*\|_p = K_p$.

1-norm: Define K_1, c and $\boldsymbol{y}^*$ by $K_1 := \max_{1\le j\le n}\sum_{k=1}^m |a_{kj}| =: \sum_{k=1}^m |a_{kc}|$ and $\boldsymbol{y}^* := \boldsymbol{e}_c$, a unit vector. Then $\|\boldsymbol{y}^*\|_1 = 1$ and we obtain

(a)

$$\|\boldsymbol{A}\boldsymbol{x}\|_1 = \sum_{k=1}^m \Big|\sum_{j=1}^n a_{kj}x_j\Big| \le \sum_{k=1}^m\sum_{j=1}^n |a_{kj}||x_j| = \sum_{j=1}^n \Big(\sum_{k=1}^m |a_{kj}|\Big)|x_j| \le K_1.$$

(b) $\|\boldsymbol{A}\boldsymbol{y}^*\|_1 = K_1$.

∞-norm: Define K_∞, r and $\boldsymbol{y}^*$ by $K_\infty := \max_{1\le k\le m}\sum_{j=1}^m |a_{kj}| =: \sum_{j=1}^n |a_{rj}|$ and $\boldsymbol{y}^* := [e^{-i\theta_1},\dots,e^{-i\theta_n}]^T$, where $a_{rj} = |a_{rj}|e^{i\theta_j}$ for $j = 1,\dots,n$.

(a) $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{1\le k\le m}\big|\sum_{j=1}^n a_{kj}x_j\big| \le \max_{1\le k\le m}\sum_{j=1}^n |a_{kj}||x_j| \le K_\infty$.

(b) $\|\boldsymbol{A}\boldsymbol{y}^*\|_\infty = \max_{1\le k\le m}\big|\sum_{j=1}^n a_{kj}e^{-i\theta_j}\big| = K_\infty$.

The last equality is correct because $\big|\sum_{j=1}^n a_{kj}e^{-i\theta_j}\big| \le \sum_{j=1}^n |a_{kj}| \le K_\infty$ with equality for $k = r$.

□

Example 8.22 *In Example 7.6 we found that the largest singular value of the matrix* $\boldsymbol{A} := \frac{1}{15}\left[\begin{smallmatrix} 14 & 4 & 16 \\ 2 & 22 & 13 \end{smallmatrix}\right]$*, is* $\sigma_1 = 2$*. We find*

$$\|\boldsymbol{A}\|_1 = \frac{29}{15}, \quad \|\boldsymbol{A}\|_2 = 2, \quad \|\boldsymbol{A}\|_\infty = \frac{37}{15}, \quad \|\boldsymbol{A}\|_F = \sqrt{5}.$$

We observe that the values of these norms do not differ by much.

In some cases the spectral norm is equal to an eigenvalue of the matrix.

Theorem 8.23 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *has singular values* $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$ *and eigenvalues* $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$*. Then*

$$\|\boldsymbol{A}\|_2 = \sigma_1 \text{ and } \|\boldsymbol{A}^{-1}\|_2 = \frac{1}{\sigma_n}, \tag{8.16}$$

$$\|\boldsymbol{A}\|_2 = \lambda_1 \text{ and } \|\boldsymbol{A}^{-1}\|_2 = \frac{1}{\lambda_n}, \quad \text{if } \boldsymbol{A} \text{ is symmetric positive definite}, \tag{8.17}$$

$$\|\boldsymbol{A}\|_2 = |\lambda_1| \text{ and } \|\boldsymbol{A}^{-1}\|_2 = \frac{1}{|\lambda_n|}, \text{ if } \boldsymbol{A} \text{ is normal}. \tag{8.18}$$

For the norms of $\boldsymbol{A}^{-1}$ *we assume of course that* $\boldsymbol{A}$ *is nonsingular.*

Proof. Since $1/\sigma_n$ is the largest singular value of $\boldsymbol{A}^{-1}$, (8.16) follows. As shown in Section 7.1.3 the singular values of a symmetric positive definite matrix (normal matrix) are equal to the eigenvalues (absolute value of the eigenvalues). This implies (8.17) and (8.18). □

Exercise 8.24 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular. Use* (8.16) *and* (7.17) *to show that*

$$\|\boldsymbol{A}^{-1}\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{A}\boldsymbol{x}\|_2}.$$

Exercise 8.25 *Let*

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Compute $\|\boldsymbol{A}\|_p$ and $\|\boldsymbol{A}^{-1}\|_p$ for $p = 1, 2, \infty$.

The following result is sometimes useful.

Theorem 8.26 *For any $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have $\|\boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty$.*

Proof. Let $(\sigma_1^2, \boldsymbol{v}_1)$ be an eigenpair for $\boldsymbol{A}^*\boldsymbol{A}$ corresponding to the largest singular value of $\boldsymbol{A}$. Then

$$\|\boldsymbol{A}\|_2^2 \|\boldsymbol{v}_1\|_1 = \sigma_1^2 \|\boldsymbol{v}_1\|_1 = \|\sigma_1^2 \boldsymbol{v}_1\|_1 = \|\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}_1\|_1 \leq \|\boldsymbol{A}^*\|_1 \|\boldsymbol{A}\|_1 \|\boldsymbol{v}_1\|_1.$$

Observing that $\|\boldsymbol{A}^*\|_1 = \|\boldsymbol{A}\|_\infty$ by Theorem 8.21 and canceling $\|\boldsymbol{v}_1\|_1$ proves the result. □

8.2.5 Unitary Invariant Matrix Norms

Definition 8.27 *A matrix norm $\|\ \|$ on $\mathbb{C}^{m,n}$ is called* **unitary invariant** *if $\|\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}\| = \|\boldsymbol{A}\|$ for any $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and any unitary matrices $\boldsymbol{U} \in \mathbb{C}^{m,m}$ and $\boldsymbol{V} \in \mathbb{C}^{n,n}$.*

When an unitary invariant matrix norm is used, the size of a perturbation is not increased by a unitary transformation. Thus if $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary then $\boldsymbol{U}(\boldsymbol{A} + \boldsymbol{E})\boldsymbol{V} = \boldsymbol{U}\boldsymbol{A}\boldsymbol{V} + \boldsymbol{F}$, where $\|\boldsymbol{F}\| = \|\boldsymbol{E}\|$.

It follows from Lemma 8.10 that the Frobenius norm is unitary invariant. We show here that this also holds for the spectral norm. It can be shown that the spectral norm is the only unitary invariant operator norm, see [9] p. 308.

Theorem 8.28 *The Frobenius norm and the spectral norm are unitary invariant. Moreover $\|\boldsymbol{A}^*\|_F = \|\boldsymbol{A}\|_F$ and $\|\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}\|_2$.*

Proof. The results for the Frobenius norm follow from Lemma 8.10. Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and let $\boldsymbol{U} \in \mathbb{C}^{m,m}$ and $\boldsymbol{V} \in \mathbb{C}^{n,n}$ be unitary. Since the 2-vector norm is unitary invariant we obtain

$$\|\boldsymbol{U}\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{U}\boldsymbol{A}\boldsymbol{x}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{A}\|_2.$$

Now $\boldsymbol{A}$ and $\boldsymbol{A}^*$ have the same nonzero singular values, and it follows from Theorem 8.21 that $\|\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}\|_2$. Moreover $\boldsymbol{V}^*$ is unitary. Using these facts we find

$$\|\boldsymbol{A}\boldsymbol{V}\|_2 = \|(\boldsymbol{A}\boldsymbol{V})^*\|_2 = \|\boldsymbol{V}^*\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}\|_2.$$

□

Exercise 8.29 *Show that* $\|\boldsymbol{V}\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_2$ *holds even for a rectangular* $\boldsymbol{V}$ *as long as* $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$.

Exercise 8.30 *Find* $\boldsymbol{A} \in \mathbb{R}^{2,2}$ *and* $\boldsymbol{U} \in \mathbb{R}^{2,1}$ *with* $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ *such that* $\|\boldsymbol{A}\boldsymbol{U}\|_2 < \|\boldsymbol{A}\|_2$. *Thus, in general,* $\|\boldsymbol{A}\boldsymbol{U}\|_2 = \|\boldsymbol{A}\|_2$ *does not hold for a rectangular* $\boldsymbol{U}$ *even if* $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$.

Exercise 8.31 *Show that* $\|\boldsymbol{A}\|_p = \rho(\boldsymbol{A}) := \max|\lambda_i|$ *(the largest eigenvalue of* $\boldsymbol{A}$*),* $1 \le p \le \infty$, *when* $\boldsymbol{A}$ *is a diagonal matrix.*

Exercise 8.32 *A vector* $\boldsymbol{a} \in \mathbb{C}^m$ *can also be considered as a column vector* $\boldsymbol{A} \in \mathbb{C}^{m,1}$.

(a) *Show that the spectral matrix norm (2-norm) of* $\boldsymbol{A}$ *equals the Euclidean vector norm of* $\boldsymbol{a}$.

(b) *Show that* $\|\boldsymbol{A}\|_p = \|\boldsymbol{a}\|_p$ *for* $1 \le p \le \infty$.

Exercise 8.33 *If* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *has elements* a_{ij}, *let* $|\boldsymbol{A}| \in \mathbb{C}^{m,n}$ *be the matrix with elements* $|a_{ij}|$.

(a) *Compute* $|\boldsymbol{A}|$ *if* $\boldsymbol{A} = \begin{bmatrix} 1+i & -2 \\ 1 & 1-i \end{bmatrix}$, $i = \sqrt{-1}$.

(b) *Show that for any* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ $\|\boldsymbol{A}\|_F = \| \, |\boldsymbol{A}| \, \|_F$, $\|\boldsymbol{A}\|_p = \| \, |\boldsymbol{A}| \, \|_p$ *for* $p = 1, \infty$.

(c) *Show that for any* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ $\|\boldsymbol{A}\|_2 \le \| \, |\boldsymbol{A}| \, \|_2$.

(d) *Find a real symmetric* 2×2 *matrix* $\boldsymbol{A}$ *such that* $\|\boldsymbol{A}\|_2 < \| \, |\boldsymbol{A}| \, \|_2$.

Exercise 8.34 *Let* $m, n \in \mathbb{N}$ *and* $\boldsymbol{A} \in \mathbb{C}^{m,n}$. *Show that*

$$\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1} |\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{x}|.$$

8.2.6 Absolute and Monotone Norms

A vector norm on $\mathbb{C}^n$ is called an **absolute norm** if $\|\boldsymbol{x}\| = \| \, |\boldsymbol{x}| \, \|$ for all $\boldsymbol{x} \in \mathbb{C}^n$. Here $|\boldsymbol{x}| := [|x_1|, \ldots, |x_n|]^T$, the absolute values of the components of $\boldsymbol{x}$. Clearly the vector p norms are absolute norms. We state without proof (see Theorem 5.5.10 of [9]) that a vector norm on $\mathbb{C}^n$ is an absolute norm if and only if it is a **monotone norm**, i. e.,

$$|x_i| \le |y_i|, \ i = 1, \ldots, n \Longrightarrow \|\boldsymbol{x}\| \le \|\boldsymbol{y}\|, \text{ for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n.$$

Absolute and monotone matrix norms are defined as for vector norms.

Exercise 8.35 *Show that the Frobenius norm and the* $1, \infty$ *operator norms are absolute norms.*

Exercise 8.36 *Show that the spectral norm is not an absolute norm.*

The study of matrix norms will be continued in Chapter 9.

8.3 The Condition Number with Respect to Inversion

Consider the system of two linear equations

$$\begin{array}{rcrcl} x_1 & + & x_2 & = & 20 \\ x_1 & + & (1-10^{-16})x_2 & = & 20-10^{-15} \end{array}$$

whose exact solution is $x_1 = x_2 = 10$. If we replace the second equation by

$$x_1 + (1+10^{-16})x_2 = 20 - 10^{-15},$$

the exact solution changes to $x_1 = 30$, $x_2 = -10$. Here a small change in one of the coefficients, from $1-10^{-16}$ to $1+10^{-16}$, changed the exact solution by a large amount.

A mathematical problem in which the solution is very sensitive to changes in the data is called **ill-conditioned.** Such problems are difficult to solve on a computer.

In this section we consider what effect a small change (perturbation) in the data $\boldsymbol{A}$,$\boldsymbol{b}$ has on the solution $\boldsymbol{x}$ of a linear system $\boldsymbol{Ax} = \boldsymbol{b}$. Suppose $\boldsymbol{y}$ solves $(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{y} = \boldsymbol{b}+\boldsymbol{e}$ where $\boldsymbol{E}$ is a (small) $n \times n$ matrix and $\boldsymbol{e}$ a (small) vector. How large can $\boldsymbol{y}-\boldsymbol{x}$ be? To measure this we use vector and matrix norms. In this section $\|\ \|$ will denote a vector norm on $\mathbb{C}^n$ and also a submultiplicative matrix norm on $\mathbb{C}^{n,n}$ which in addition is subordinate to the vector norm. Thus for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$ and any $\boldsymbol{x} \in \mathbb{C}^n$ we have

$$\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\| \text{ and } \|\boldsymbol{Ax}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|.$$

This is satisfied if the matrix norm is the operator norm corresponding to the given vector norm, but is also satisfied for the Frobenius matrix norm and the Euclidian vector norm. This follows from Lemma 8.10.

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors in $\mathbb{C}^n$ that we want to compare. The difference $\|\boldsymbol{y}-\boldsymbol{x}\|$ measures the **absolute error** in $\boldsymbol{y}$ as an approximation to $\boldsymbol{x}$, while $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{y}\|$ are measures for the **relative error**.

We consider first a perturbation in the right-hand side $\boldsymbol{b}$.

Theorem 8.37 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *is nonsingular,* $\boldsymbol{b}, \boldsymbol{e} \in \mathbb{C}^n$, $\boldsymbol{b} \neq \boldsymbol{0}$ *and* $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{Ay} = \boldsymbol{b}+\boldsymbol{e}$. *Then*

$$\frac{1}{K(\boldsymbol{A})}\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{b}\|} \le \frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le K(\boldsymbol{A})\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{b}\|}, \quad K(\boldsymbol{A}) = \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|. \tag{8.19}$$

Proof. Subtracting $\boldsymbol{Ax} = \boldsymbol{b}$ from $\boldsymbol{Ay} = \boldsymbol{b}+\boldsymbol{e}$ we have $\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{x}) = \boldsymbol{e}$ or $\boldsymbol{y}-\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{e}$. Combining $\|\boldsymbol{y}-\boldsymbol{x}\| = \|\boldsymbol{A}^{-1}\boldsymbol{e}\| \le \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{e}\|$ and $\|\boldsymbol{b}\| = \|\boldsymbol{Ax}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|$ we obtain the upper bound in (8.19). Combining $\|\boldsymbol{e}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{y}-\boldsymbol{x}\|$ and $\|\boldsymbol{x}\| \le \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b}\|$ we obtain the lower bound. ∎

Consider (8.19). $\|\boldsymbol{e}\|/\|\boldsymbol{b}\|$ is a measure of the size of the perturbation $\boldsymbol{e}$ relative to the size of $\boldsymbol{b}$. The upper bound says that $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{x}\|$ in the worst case can be

$$K(\boldsymbol{A}) = \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|$$

times as large as $\|\boldsymbol{e}\|/\|\boldsymbol{b}\|$. $K(\boldsymbol{A})$ is called the **condition number with respect to inversion of a matrix**, or just the condition number, if it is clear from the context that we are talking about solving linear systems or inverting a matrix. The condition number depends on the matrix $\boldsymbol{A}$ and on the norm used. If $K(\boldsymbol{A})$ is large, $\boldsymbol{A}$ is called **ill-conditioned** (with respect to inversion). If $K(\boldsymbol{A})$ is small, $\boldsymbol{A}$ is called **well-conditioned** (with respect to inversion). We always have $K(\boldsymbol{A}) \ge 1$. For since $\|\boldsymbol{x}\| = \|\boldsymbol{Ix}\| \le \|\boldsymbol{I}\|\|\boldsymbol{x}\|$ for any $\boldsymbol{x}$, by subordinance we have $\|\boldsymbol{I}\| \ge 1$ and therefore by submultiplicativity $\|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\| \ge \|\boldsymbol{AA}^{-1}\| = \|\boldsymbol{I}\| \ge 1$.

Since all matrix norms are equivalent, the dependence of $K(\boldsymbol{A})$ on the norm chosen is less important than the dependence on $\boldsymbol{A}$. Sometimes one chooses the spectral norm when discussing properties of the condition number, and the ℓ_1, ℓ_∞, or Frobenius norm when one wishes to compute it or estimate it.

Explicit expressions for the 2-norm condition number follow from Theorem 8.23.

Theorem 8.38 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular with singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0$ and eigenvalues $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n| > 0$. Then $K_2(\boldsymbol{A}) := \|\boldsymbol{A}\|_2\|\boldsymbol{A}^{-1}\|_2 = \sigma_1/\sigma_n$. Moreover,*

$$K_2(\boldsymbol{A}) = \begin{cases} \lambda_1/\lambda_n, & \text{if } \boldsymbol{A} \text{ is symmetric positive definite,} \\ |\lambda_1|/|\lambda_n|, & \text{if } \boldsymbol{A} \text{ is normal.} \end{cases} \tag{8.20}$$

It follows that $\boldsymbol{A}$ is ill-conditioned with respect to inversion if and only if σ_1/σ_n is large, or λ_1/λ_n is large when $\boldsymbol{A}$ is symmetric positive definite.

Exercise 8.39 *The upper and lower bounds for $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{x}\|$ given by* (8.19) *can be attained for any matrix $\boldsymbol{A}$, but only for special choices of $\boldsymbol{b}$. Suppose $\boldsymbol{y}_{\boldsymbol{A}}$ and $\boldsymbol{y}_{\boldsymbol{A}^{-1}}$ are vectors with $\|\boldsymbol{y}_{\boldsymbol{A}}\| = \|\boldsymbol{y}_{\boldsymbol{A}^{-1}}\| = 1$ and $\|\boldsymbol{A}\| = \|\boldsymbol{Ay}_{\boldsymbol{A}}\|$ and $\|\boldsymbol{A}^{-1}\| = \|\boldsymbol{A}^{-1}\boldsymbol{y}_{\boldsymbol{A}^{-1}}\|$.*

(a) *Show that the upper bound in* (8.19) *is attained if $\boldsymbol{b} = \boldsymbol{Ay}_{\boldsymbol{A}}$ and $\boldsymbol{e} = \boldsymbol{y}_{\boldsymbol{A}^{-1}}$.*

(b) *Show that the lower bound is attained if $\boldsymbol{b} = \boldsymbol{y}_{\boldsymbol{A}^{-1}}$ and $\boldsymbol{e} = \boldsymbol{Ay}_{\boldsymbol{A}}$.*

We consider next a perturbation $\boldsymbol{E}$ in a nonsingular matrix $\boldsymbol{A}$. The following result shows that $\boldsymbol{A}+\boldsymbol{E}$ is nonsingular if $\boldsymbol{E}$ is sufficiently small and that small changes in $\boldsymbol{A}$ give small changes in the inverse if $\boldsymbol{A}$ is well conditioned.

Theorem 8.40 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular and let $\|\cdot\|$ be a consistent matrix norm on $\mathbb{C}^{n,n}$. If $\boldsymbol{E} \in \mathbb{C}^{n,n}$ is so small that $r := \|\boldsymbol{A}^{-1}\boldsymbol{E}\| < 1$ then $\boldsymbol{A}+\boldsymbol{E}$ is nonsingular and*

$$\|(\boldsymbol{A}+\boldsymbol{E})^{-1}\| \leq \frac{\|\boldsymbol{A}^{-1}\|}{1-r}. \tag{8.21}$$

If $r < 1/2$ then

$$\frac{\|(\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1}\|}{\|\boldsymbol{A}^{-1}\|} \leq 2K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}. \tag{8.22}$$

Proof. We show in (8.31) in Section 8.4 that if $\boldsymbol{B} \in \mathbb{C}^{n,n}$ and $\|\boldsymbol{B}\| < 1$ then $\boldsymbol{I}-\boldsymbol{B}$ is nonsingular and

$$\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| \leq \frac{1}{1-\|\boldsymbol{B}\|}. \tag{8.23}$$

Since $r < 1$ the matrix $\boldsymbol{I}-\boldsymbol{B} := \boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{E}$ is nonsingular. Since $(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{A}^{-1}(\boldsymbol{A}+\boldsymbol{E}) = \boldsymbol{I}$ we see that $\boldsymbol{A}+\boldsymbol{E}$ is nonsingular with inverse $(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{A}^{-1}$. Hence, $\|(\boldsymbol{A}+\boldsymbol{E})^{-1}\| \leq \|(\boldsymbol{I}-\boldsymbol{B})^{-1}\|\|\boldsymbol{A}^{-1}\|$ and (8.21) follows from (8.23). From the identity

$$(\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1}\boldsymbol{E}(\boldsymbol{A}+\boldsymbol{E})^{-1}$$

we obtain by (8.21)

$$\|(\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1}\| \leq \|\boldsymbol{A}^{-1}\|\|\boldsymbol{E}\|\|(\boldsymbol{A}+\boldsymbol{E})^{-1}\| \leq K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}\frac{\|\boldsymbol{A}^{-1}\|}{1-r}.$$

Dividing by $\|\boldsymbol{A}^{-1}\|$ and setting $r = 1/2$ proves (8.22). □

We can now show the following upper bounds.

Theorem 8.41 *Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n,n}$, $\boldsymbol{b} \in \mathbb{C}^n$ with $\boldsymbol{A}$ invertible and $\boldsymbol{b} \neq \boldsymbol{0}$. If $r := \|\boldsymbol{A}^{-1}\boldsymbol{E}\| < 1/2$ for some operator norm then $\boldsymbol{A}+\boldsymbol{E}$ is invertible. If $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{y} = \boldsymbol{b}$ then*

$$\frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \leq \|\boldsymbol{A}^{-1}\boldsymbol{E}\| \leq K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}, \tag{8.24}$$

$$\frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq 2K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}.. \tag{8.25}$$

Proof. That the matrix $\boldsymbol{A}+\boldsymbol{E}$ is invertible follows from Theorem 8.40. (8.24) follows easily by taking norms in the equation $\boldsymbol{x}-\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{y}$ and dividing by $\|\boldsymbol{y}\|$. From the identity $\boldsymbol{y}-\boldsymbol{x} = \left((\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1}\right)\boldsymbol{A}\boldsymbol{x}$ we obtain $\|\boldsymbol{y}-\boldsymbol{x}\| \leq \|(\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1}\|\|\boldsymbol{A}\|\|\boldsymbol{x}\|$ and (8.25) follows from (8.21). □

In Theorem 8.41 we gave a bound for the relative error in $\boldsymbol{x}$ as an approximation to $\boldsymbol{y}$, (8.24), and the relative error in $\boldsymbol{y}$ as an approximation to $\boldsymbol{x}$, (8.25). $\|\boldsymbol{E}\|/\|\boldsymbol{A}\|$ is a measure for the size of the perturbation $\boldsymbol{E}$ in $\boldsymbol{A}$ relative to the size of $\boldsymbol{A}$. The condition number again plays a crucial role. $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{y}\|$ can be as

large as $K(\boldsymbol{A})$ times $\|\boldsymbol{E}\|/\|\boldsymbol{A}\|$. It can be shown that the upper bound can be attained for any $\boldsymbol{A}$ and any $\boldsymbol{b}$. In deriving the upper bound we used the inequality $\|\boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{y}\| \leq \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{E}\|\,\|\boldsymbol{y}\|$. For a more or less random perturbation $\boldsymbol{E}$ this is not a severe overestimate for $\|\boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{y}\|$. In the situation where $\boldsymbol{E}$ is due to round-off errors (8.24) can give a fairly realistic estimate for $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{y}\|$.

Suppose we have computed an approximate solution $\boldsymbol{y}$ to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. The vector $\boldsymbol{r}(\boldsymbol{y}) := \boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}$ is called the **residual vector**, or just the residual. We can bound $\boldsymbol{x}-\boldsymbol{y}$ in term of $\boldsymbol{r}$.

Theorem 8.42 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$, $\boldsymbol{b} \in \mathbb{C}^n$, $\boldsymbol{A}$ is nonsingular and $\boldsymbol{b} \neq \boldsymbol{0}$. Let $\boldsymbol{r}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}$ for each $\boldsymbol{y} \in \mathbb{C}^n$. If $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ then*

$$\frac{1}{K(\boldsymbol{A})}\frac{\|\boldsymbol{r}(\boldsymbol{y})\|}{\|\boldsymbol{b}\|} \leq \frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq K(\boldsymbol{A})\frac{\|\boldsymbol{r}(\boldsymbol{y})\|}{\|\boldsymbol{b}\|}. \tag{8.26}$$

Proof. We simply take $\boldsymbol{e} = \boldsymbol{r}(\boldsymbol{y})$ in Theorem 8.37. □

If $\boldsymbol{A}$ is well-conditioned, (8.26) says that $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{x}\| \approx \|\boldsymbol{r}(\boldsymbol{y})\|/\|\boldsymbol{b}\|$. In other words, the accuracy in $\boldsymbol{y}$ is about the same order of magnitude as the residual as long as $\|\boldsymbol{b}\| \approx 1$. If $\boldsymbol{A}$ is ill-conditioned, anything can happen. We can for example have an accurate solution even if the residual is large.

Exercise 8.43 *Let $\|\ \|_p$ be the l_p vector norm and let $cond_p(\boldsymbol{T}) = \|\boldsymbol{T}\|_p\|\boldsymbol{T}^{-1}\|_p$, where $\|\boldsymbol{T}\|_p = \max_{x\neq 0}\|\boldsymbol{T}\boldsymbol{x}\|_p/\|\boldsymbol{x}\|_p$ be the p-condition number of $\boldsymbol{T} \in \mathbb{R}^{m,m}$. In this exercise we find the p-condition numbers for the matrix $\boldsymbol{T} := tridiag(-1,2,-1)$ in terms of $h := 1/(m+1)$. You will need the explicit inverse of $\boldsymbol{T}$ given by* (2.10) *and the eigenvalues given in Lemma 4.11.*

a) *Show that*

$$cond_1(\boldsymbol{T}) = cond_\infty(\boldsymbol{T}) = \frac{1}{2}\begin{cases} h^{-2}, & m \text{ odd},\ m > 1, \\ h^{-2}-1, & m \text{ even}. \end{cases} \tag{8.27}$$

b) *Show that for $p = 2$ we have*

$$cond_2(\boldsymbol{T}) = \cot^2\left(\frac{\pi h}{2}\right) = 1/\tan^2\left(\frac{\pi h}{2}\right).$$

c) *Show the bounds*

$$\frac{4}{\pi^2}h^{-2} - \frac{2}{3} < cond_2(\boldsymbol{T}) < \frac{4}{\pi^2}h^{-2}. \tag{8.28}$$

Hint: For the upper bound use the inequality $\tan x > x$ valid for $0 < x < \pi/2$. For the lower bound we use the inequality $\cot^2 x > \frac{1}{x^2} - \frac{2}{3}$ for $x > 0$. This can be derived for $0 < x < \pi$ by first showing that the second derivative of $\cot^2 x$ is positive and then use Taylor's theorem.

8.4 Convergence and Spectral Radius

We start with some basic notions that we need.

8.4.1 Convergence in $\mathbb{R}^{m,n}$ and $\mathbb{C}^{m,n}$

Definition 8.44 *Consider an infinite sequence of matrices $\{\boldsymbol{A}_k\} = \boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2, \ldots$ in $\mathbb{C}^{m,n}$.*

1. *$\{\boldsymbol{A}_k\}$ is said to converge to the limit $\boldsymbol{A}$ in $\mathbb{C}^{m,n}$ if each element sequence $\{\boldsymbol{A}_k(ij)\}_k$ converges to the corresponding element $\boldsymbol{A}(ij)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.*
2. *$\{\boldsymbol{A}_k\}$ is a* **Cauchy sequence** *if for all $\epsilon > 0$ there is an integer $N \in \mathbb{N}$ such that for each $k, l \geq N$ and all i, j we have $|\boldsymbol{A}_k(ij) - \boldsymbol{A}_l(ij)| \leq \epsilon$.*
3. *$\{\boldsymbol{A}_k\}$ is bounded if there is a constant M such that $|\boldsymbol{A}_k(ij)| \leq M$ for all i, j, k.*

By stacking the columns of $\boldsymbol{A}$ into a vector in $\mathbb{C}^{mn}$ we can use the results in Section A.4 and obtain

Theorem 8.45 1. *A sequence $\{\boldsymbol{A}_k\}$ in $\mathbb{C}^{m,n}$ converges to a matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ if and only if $\lim_{k\to\infty} \|\boldsymbol{A}_k - \boldsymbol{A}\| = 0$ for any matrix norm $\|\cdot\|$.*
2. *A sequence $\{\boldsymbol{A}_k\}$ in $\mathbb{C}^{m,n}$ is convergent if and only if it is a Cauchy sequence.*
3. *Every bounded sequence $\{\boldsymbol{A}_k\}$ in $\mathbb{C}^{m,n}$ has a convergent subsequence.*

8.4.2 The Spectral Radius

We define the **spectral radius** of a matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ as the maximum absolute value of its eigenvalues.

$$\rho(\boldsymbol{A}) := \max_{\lambda \in \sigma(\boldsymbol{A})} |\lambda|. \tag{8.29}$$

Theorem 8.46 *For any matrix norm $\|\cdot\|$ which is consistent on $\mathbb{C}^{n,n}$ and any $\boldsymbol{A} \in \mathbb{C}^{n,n}$ we have $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$.*

Proof. Let $(\lambda, \boldsymbol{x})$ be an eigenpair for $\boldsymbol{A}$ and define $\boldsymbol{X} := [\boldsymbol{x}, \ldots, \boldsymbol{x}] \in \mathbb{C}^{n,n}$. Then $\lambda \boldsymbol{X} = \boldsymbol{A}\boldsymbol{X}$, which implies $|\lambda| \, \|\boldsymbol{X}\| = \|\lambda \boldsymbol{X}\| = \|\boldsymbol{A}\boldsymbol{X}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{X}\|$. Since $\|\boldsymbol{X}\| \neq 0$ we obtain $|\lambda| \leq \|\boldsymbol{A}\|$. □

The inequality $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$ can almost be made into an equality by choosing the norm carefully.

Theorem 8.47 *Let $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and $\epsilon > 0$ be given. There is a consistent matrix norm $\|\cdot\|'$ on $\mathbb{C}^{n,n}$ such that $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|' \leq \rho(\boldsymbol{A}) + \epsilon$.*

Proof. Let $\boldsymbol{A}$ have eigenvalues $\lambda_1, \ldots, \lambda_n$. By the Schur Triangulation Theorem 6.1 there is a unitary matrix $\boldsymbol{U}$ and an upper triangular matrix $\boldsymbol{R} = [r_{ij}]$ such that $\boldsymbol{U}^* \boldsymbol{A} \boldsymbol{U} = \boldsymbol{R}$. For $t > 0$ we define $\boldsymbol{D}_t := \operatorname{diag}(t, t^2, \ldots, t^n) \in \mathbb{R}^{n,n}$, and note that the (i, j) element in $\boldsymbol{D}_t \boldsymbol{R} \boldsymbol{D}_t^{-1}$ is given by $t^{i-j} r_{ij}$ for all i, j. For $n = 3$

$$\boldsymbol{D}_t \boldsymbol{R} \boldsymbol{D}_t^{-1} = \begin{bmatrix} \lambda_1 & t^{-1} r_{12} & t^{-2} r_{13} \\ 0 & \lambda_2 & t^{-1} r_{23} \\ 0 & 0 & \lambda_3 \end{bmatrix}.$$

For each $\boldsymbol{B} \in \mathbb{C}^{n,n}$ and $t > 0$ we define $\|\boldsymbol{B}\|_t := \|\boldsymbol{D}_t\boldsymbol{U}^*\boldsymbol{B}\boldsymbol{U}\boldsymbol{D}_t^{-1}\|_1$. We leave it as an exercise to show that this is a consistent matrix norm on $\mathbb{C}^{n,n}$. We define $\|\boldsymbol{B}\|' := \|\boldsymbol{B}\|_t$, where t is chosen so large that the sum of the absolute values of all off-diagonal elements in $\boldsymbol{D}_t\boldsymbol{R}\boldsymbol{D}_t^{-1}$ is less than ϵ. Then

$$\begin{aligned}\|\boldsymbol{A}\|' &= \|\boldsymbol{D}_t\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{D}_t^{-1}\|_1 = \|\boldsymbol{D}_t\boldsymbol{R}\boldsymbol{D}_t^{-1}\|_1 = \max_{1\le j\le n}\sum_{i=1}^{n}|\left(\boldsymbol{D}_t\boldsymbol{R}\boldsymbol{D}_t^{-1}\right)_{ij}| \\ &\le \max_{1\le j\le n}\left(|\lambda_j| + \epsilon\right) = \rho(\boldsymbol{A}) + \epsilon.\end{aligned}$$

□

Theorem 8.48 *For any $\boldsymbol{A} \in \mathbb{C}^{n,n}$ we have*

$$\lim_{k\to\infty}\boldsymbol{A}^k = 0 \iff \rho(\boldsymbol{A}) < 1.$$

Proof. Suppose $\rho(\boldsymbol{A}) < 1$. By Theorem 8.47 there is a consistent matrix norm $\|\cdot\|$ on $\mathbb{C}^{n,n}$ such that $\|\boldsymbol{A}\| < 1$. But then $\|\boldsymbol{A}^k\| \le \|\boldsymbol{A}\|^k \to 0$ as $k \to \infty$. Hence $\boldsymbol{A}^k \to 0$. Conversely, suppose $(\lambda, \boldsymbol{x})$ is an eigenpair of $\boldsymbol{A}$ with $|\lambda| \ge 1$. Since $\boldsymbol{A}^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$, by Theorem D.3 it follows that $\boldsymbol{A}^k\boldsymbol{x}$ does not tend to zero. But then we cannot have $\boldsymbol{A}^k \to 0$. □

Theorem 8.49 *For any consistent matrix norm $\|\cdot\|$ on $\mathbb{C}^{n,n}$ and any $\boldsymbol{A} \in \mathbb{C}^{n,n}$ we have*

$$\lim_{k\to\infty}\|\boldsymbol{A}^k\|^{1/k} = \rho(\boldsymbol{A}). \tag{8.30}$$

Proof. By Theorems D.3 and 8.46 we obtain $\rho(\boldsymbol{A})^k = \rho(\boldsymbol{A}^k) \le \|\boldsymbol{A}^k\|$ for any $k \in \mathbb{N}$ so that $\rho(\boldsymbol{A}) \le \|\boldsymbol{A}^k\|^{1/k}$. Let $\epsilon > 0$ and consider the matrix $\boldsymbol{B} := (\rho(\boldsymbol{A}) + \epsilon)^{-1}\boldsymbol{A}$. Then $\rho(\boldsymbol{B}) = \rho(\boldsymbol{A})/(\rho(\boldsymbol{A}) + \epsilon) < 1$ and $\|\boldsymbol{B}^k\| \to 0$ by Theorem 8.48 as $k \to \infty$. Choose $N \in \mathbb{N}$ such that $\|\boldsymbol{B}^k\| < 1$ for all $k \ge N$. Then for $k \ge N$

$$\|\boldsymbol{A}^k\| = \|\left(\rho(\boldsymbol{A}) + \epsilon\right)\boldsymbol{B}\big)^k\| = \left(\rho(\boldsymbol{A}) + \epsilon\right)^k\|\boldsymbol{B}^k\| < \left(\rho(\boldsymbol{A}) + \epsilon\right)^k.$$

We have shown that $\rho(\boldsymbol{A}) \le \|\boldsymbol{A}^k\|^{1/k} \le \rho(\boldsymbol{A}) + \epsilon$ for $k \ge N$. Since ϵ is arbitrary the result follows. □

Exercise 8.50 *The convergence $\lim_{k\to\infty}\|\boldsymbol{A}^k\|^{1/k} = \rho(\boldsymbol{A})$ can be quite slow. Consider*

$$\boldsymbol{A} := \begin{bmatrix} \lambda & a & 0 & \cdots & 0 & 0 \\ 0 & \lambda & a & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & a \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \in \mathbb{R}^{n,n}.$$

If $|\lambda| = \rho(\boldsymbol{A}) < 1$ then $\lim_{k\to\infty}\boldsymbol{A}^k = \boldsymbol{0}$ for any $a \in \mathbb{R}$. We show below that the $(1, n)$ element of $\boldsymbol{A}^k$ is given by $f(k) := \binom{k}{n-1}a^{n-1}\lambda^{k-n+1}$ for $k \ge n-1$.

(a) *Make a plot of* $f(k)$ *for* $\lambda = 0.9$, $a = 10$, *and* $k \leq 200$. *Your program should also compute* $\max_k f(k)$. *Use your program to determine how large* k *must be before* $f(k) < 10^{-8}$.

(b) *We can determine the elements of* $\boldsymbol{A}^k$ *explicitly for any* k. *Let* $\boldsymbol{E} := (\boldsymbol{A} - \lambda \boldsymbol{I})/a$. *Show by induction that* $\boldsymbol{E}^k = \left[\begin{smallmatrix} \boldsymbol{0} & \boldsymbol{I}_{n-k} \\ \boldsymbol{0} & \boldsymbol{0} \end{smallmatrix}\right]$ *for* $1 \leq k \leq n-1$ *and that* $\boldsymbol{E}^n = \boldsymbol{0}$.

(c) *We have* $\boldsymbol{A}^k = (a\boldsymbol{E} + \lambda \boldsymbol{I})^k = \sum_{j=0}^{\min\{k,n-1\}} \binom{k}{j} \lambda^{k-j} a^j \boldsymbol{E}_n^j$ *and conclude that the* $(1, n)$ *element is given by* $f(k)$ *for* $k \geq n-1$.

8.4.3 Neumann Series

A geometric series of matrices is known as a Neumann Series.

Theorem 8.51 (Neumann Series) *Suppose* $\boldsymbol{B} \in \mathbb{C}^{n,n}$. *Then*

1. *The series* $\sum_{k=0}^{\infty} \boldsymbol{B}^k$ *converges if and only if* $\rho(\boldsymbol{B}) < 1$.
2. *If* $\rho(\boldsymbol{B}) < 1$ *then* $(\boldsymbol{I} - \boldsymbol{B})$ *is nonsingular and* $(\boldsymbol{I} - \boldsymbol{B})^{-1} = \sum_{k=0}^{\infty} \boldsymbol{B}^k$.
3. *If* $\|\boldsymbol{B}\| < 1$ *for some consistent matrix norm* $\|\cdot\|$ *on* $\mathbb{C}^{n,n}$ *then*

$$\|(\boldsymbol{I} - \boldsymbol{B})^{-1}\| \leq \frac{1}{1 - \|\boldsymbol{B}\|}. \tag{8.31}$$

Proof.

1. Suppose $\rho(\boldsymbol{B}) < 1$. We use Theorem 8.45 and show that the sequence $\{\boldsymbol{A}_m\}$ of partial sums $\boldsymbol{A}_m := \sum_{k=0}^{m} \boldsymbol{B}^k$ is a Cauchy sequence. Let $\epsilon > 0$. By Theorem 8.47 there is a consistent matrix norm $\|\cdot\|$ on $\mathbb{C}^{n,n}$ such that $\|\boldsymbol{B}\| < 1$. Then for $l > m$

$$\|\boldsymbol{A}_l - \boldsymbol{A}_m\| = \|\sum_{k=m+1}^{l} \boldsymbol{B}^k\| \leq \sum_{k=m+1}^{l} \|\boldsymbol{B}\|^k \leq \frac{\|\boldsymbol{B}\|^{m+1}}{1 - \|\boldsymbol{B}\|} \leq \epsilon$$

   provided $m \geq N$ and N is such that $\frac{\|\boldsymbol{B}\|^{N+1}}{1-\|\boldsymbol{B}\|} \leq \epsilon$. Thus $\{\boldsymbol{A}_m\}$ is a Cauchy sequence and hence convergent.

   Conversely, suppose $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{B}$ with $\lambda \geq 1$. Now for $l > m$

$$\|(\boldsymbol{A}_l - \boldsymbol{A}_m)\boldsymbol{x}\| = \|\sum_{k=m+1}^{l} \boldsymbol{B}^k \boldsymbol{x}\| = \|\sum_{k=m+1}^{l} \lambda^k \boldsymbol{x}\| = \|\boldsymbol{x}\| \sum_{k=m+1}^{l} |\lambda|^k \geq |\lambda|^{m+1} \|\boldsymbol{x}\|.$$

   But then $\{\boldsymbol{A}_m\}$ cannot be a Cauchy sequence and hence not convergent.

2. By induction on m it follows that

$$\Big(\sum_{k=0}^{m} \boldsymbol{B}^k\Big)(\boldsymbol{I} - \boldsymbol{B}) = \boldsymbol{I} - \boldsymbol{B}^{m+1}. \tag{8.32}$$

For if $\big(\sum_{k=0}^{m-1} \boldsymbol{B}^k\big)(\boldsymbol{I}-\boldsymbol{B}) = \boldsymbol{I}-\boldsymbol{B}^m$ then

$$\big(\sum_{k=0}^{m} \boldsymbol{B}^k\big)(\boldsymbol{I}-\boldsymbol{B}) = \big(\sum_{k=0}^{m-1} \boldsymbol{B}^k+\boldsymbol{B}^m\big)(\boldsymbol{I}-\boldsymbol{B}) = \boldsymbol{I}-\boldsymbol{B}^m+\boldsymbol{B}^m-\boldsymbol{B}^{m+1} = \boldsymbol{I}-\boldsymbol{B}^{m+1}.$$

Since $\rho(\boldsymbol{B}) < 1$ we conclude that $\boldsymbol{B}^{m+1} \to 0$ and hence taking limits in (8.32) we obtain $\big(\sum_{k=0}^{\infty} \boldsymbol{B}^k\big)(\boldsymbol{I}-\boldsymbol{B}) = \boldsymbol{I}$ which completes the proof of 2.

3. By 1: $\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| = \|\sum_{k=0}^{\infty} \boldsymbol{B}^k\| \le \sum_{k=0}^{\infty} \|\boldsymbol{B}\|^k = \frac{1}{1-\|\boldsymbol{B}\|}$.

□

Exercise 8.52 *Show that* $\|\boldsymbol{B}\|_t := \|\boldsymbol{D}_t\boldsymbol{U}^*\boldsymbol{B}\boldsymbol{U}\boldsymbol{D}_t^{-1}\|_1$ *defined in the proof of Theorem 8.47 is a consistent matrix norm on* $\mathbb{C}^{n,n}$.

Exercise 8.53 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *is nonsingular and* $\boldsymbol{E} \in \mathbb{C}^{n,n}$. *Show that* $\boldsymbol{A}+\boldsymbol{E}$ *is nonsingular if and only if* $\rho(\boldsymbol{A}^{-1}\boldsymbol{E}) < 1$.

Part III

Iterative Methods for Large Linear Systems

Chapter 9

The Classical Iterative Methods

Gaussian elimination and Cholesky factorization are **direct methods**. In absence of rounding errors they find the exact solution using a finite number of arithmetic operations. In an **iterative method** we start with an approximation $\boldsymbol{x}^{(0)}$ to the exact solution $\boldsymbol{x}$ and then compute a sequence $\{\boldsymbol{x}^{(k)}\}$ such that hopefully $\boldsymbol{x}^{(k)} \to \boldsymbol{x}$. Iterative methods are mainly used for large sparse systems, i. e., where many of the elements in the coefficient matrix are zero. The main advantages of iterative methods are reduced storage requirements and ease of implementation. In an iterative method the main work in each iteration is a matrix times vector multiplication, an operation which often does not need storing the matrix, not even in sparse form.

We consider the classical iterative methods of Jacobi, Gauss-Seidel, and an accelerated version of Gauss-Seidel's method called Successive OverRelaxation (SOR). David Young developed in his thesis a beautiful theory describing the convergence rate of SOR, see [26]. We give the main points of this theory specialized to the average- and discrete Poisson matrix. With a careful choice of an acceleration parameter the amount of work using SOR on the discrete Poisson problem is the same as for the fast Poisson solver without FFT. Moreover, SOR is not restricted to constant coefficient methods on a rectangle. However, to obtain fast convergence using SOR it is necessary to have a good estimate for the acceleration parameter.

9.1 Classical Iterative Methods; Component Form

Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular with nonzero diagonal elements and let $\boldsymbol{b} \in \mathbb{C}^n$. Solving the ith equation of $\boldsymbol{Ax} = \boldsymbol{b}$ for x_i, we obtain a fixed-point form of $\boldsymbol{Ax} = \boldsymbol{b}$

$$x_i = \big(-\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j + b_i \big)/a_{ii}, \quad i = 1, 2, \ldots, n. \tag{9.1}$$

Suppose we know an approximation $\boldsymbol{x}^{(k)} = [x_1^{(k)}, \ldots, x_n^{(k)}]^T$ to the exact solution $\boldsymbol{x}$ of $\boldsymbol{Ax} = \boldsymbol{b}$.

1. In **Jacobi's method (J method)** we substitute $\boldsymbol{x}^{(k)}$ into the right hand side of (9.1) and compute a new approximation by

$$x_i^{(k+1)} = \big(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\big)/a_{ii}, \text{ for } i = 1, 2, \ldots, n. \quad (9.2)$$

2. **Gauss-Seidel's method (GS method)** is a modification of Jacobi's method, where we use the new $x_i^{(k+1)}$ immediately after it has been computed.

$$x_i^{(k+1)} = \big(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\big)/a_{ii}, \text{ for } i = 1, 2, \ldots, n. \quad (9.3)$$

3. The **Successive Over Relaxation method (SOR method)** is obtained by introducing an acceleration parameter $0 < \omega < 2$ in the GS method. We write $x_i = \omega x_i + (1-\omega)x_i$ and this leads to the method

$$x_i^{(k+1)} = \omega\big(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i\big)/a_{ii} + (1-\omega)x_i^{(k)}. \quad (9.4)$$

   The SOR method reduces to the Gauss-Seidel method for $\omega = 1$. Denoting the right hand side of (9.3) by $\boldsymbol{x}_{gs}^{(k+1)}$ we can write (9.4) as $\boldsymbol{x}^{(k+1)} = \omega\boldsymbol{x}_{gs}^{(k+1)} + (1-\omega)\boldsymbol{x}^{(k)}$, and we see that $\boldsymbol{x}^{(k+1)}$ is located on the straight line passing through the two points $\boldsymbol{x}_{gs}^{(k+1)}$ and $\boldsymbol{x}^{(k)}$. The restriction $0 < \omega < 2$ is necessary for convergence (cf. Theorem 9.22). Normally we choose the relaxation parameter ω in the range $1 \le \omega < 2$ and then $\boldsymbol{x}^{(k+1)}$ is computed by linear extrapolation, i. e., it is not located between $\boldsymbol{x}_{gs}^{(k+1)}$ and $\boldsymbol{x}^{(k)}$.
4. We mention also briefly the Symmetric Successive Over Relaxation method **SSOR**. One iteration in SSOR consists of two SOR sweeps. A forward SOR sweep (9.4), computing an approximation denoted $\boldsymbol{x}^{(k+1/2)}$ instead of $\boldsymbol{x}^{(k+1)}$, is followed by a back SOR sweep computing

$$x_i^{(k+1)} = \omega\big(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1/2)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k+1)} + b_i\big)/a_{ii} + (1-\omega)x_i^{(k+1/2)} \quad (9.5)$$

   in the order $i = n, n-1, \ldots 1$. The method is slower and more complicated than the SOR method. Its main use is as a symmetric preconditioner. For if $\boldsymbol{A}$ is symmetric then SSOR combines the two SOR steps in such a way that the resulting iteration matrix is similar to a symmetric matrix. We will not discuss this method any further here and refer to Section 11.2 for an alternative example of a preconditioner.

We will refer to the J,GS, and SOR methods as the **classical (iteration) methods**.

9.2 The Discrete Poisson System

Consider the classical methods applied to the discrete Poisson matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ given by (4.7). Let $n = m^2$ and set $h = 1/(m+1)$. In component form the linear system $\boldsymbol{Ax} = \boldsymbol{b}$ can be written (cf. (4.3))

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = h^2 f_{i,j}, \quad i,j = 1,\ldots,m,$$

with homogenous boundary conditions (4.4). Solving for $u_{i,j}$ we obtain

$$u_{i,j} = \big(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} + h^2 f_{i,j}\big)/4, \quad i,j = 1,\ldots,m. \tag{9.6}$$

The J, GS , and SOR methods can now be written

$$\begin{aligned} J:\ v_{ij}^{(k+1)} &= \Big(v_{i-1,j}^{(k)} + v_{i,j-1}^{(k)} + v_{i+1,j}^{(k)} + v_{i,j+1}^{(k)} + h^2 f_{ij}\Big)/4 \\ GS:\ v_{ij}^{(k+1)} &= \Big(v_{i-1,j}^{(k+1)} + v_{i,j-1}^{(k+1)} + v_{i+1,j}^{(k)} + v_{i,j+1}^{(k)} + h^2 f_{ij}\Big)/4 \\ SOR:\ v_{ij}^{(k+1)} &= \omega\Big(v_{i-1,j}^{(k+1)} + v_{i,j-1}^{(k+1)} + v_{i+1,j}^{(k)} + v_{i,j+1}^{(k)} + h^2 f_{ij}\Big)/4 + (1-\omega)v_{i,j}^{(k)}. \end{aligned} \tag{9.7}$$

For GS and SOR we use the **natural ordering** i. e., with i,j in increasing order $i,j = 1,\ldots,m$, while for J any ordering can be used.

Here is a Matlab program to test the convergence of Jacobi's method on the discrete Poisson problem.

Algorithm 9.1 (Jacobi) We carry out Jacobi iterations on the linear system (9.6) with $\boldsymbol{F} = (f_{ij}) \in \mathbb{R}^{m,m}$, starting with $\boldsymbol{V}^{(0)} = \boldsymbol{0} \in \mathbb{R}^{m+2,m+2}$. The output is the number of iterations k, to obtain $\|\boldsymbol{V}^{(k)} - \boldsymbol{U}\|_M := \max_{i,j}|v_{ij}^{(k)} - u_{ij}| < tol$. Here $(u_{ij}) \in \mathbb{R}^{m+2,m+2}$ is the "exact" solution of (9.6) computed using the fast Poisson solver in Algorithm 5.1. We set $k = K+1$ if convergence is not obtained in K iterations.

```
function k=jdp(F,K,tol)
m=length(F);
U=fastpoisson(F);
V=zeros(m+2,m+2); W=V;
E=F/(m+1)^2;
for k=1:K
  for i=2:m+1
    for j=2:m+1
       W(i,j)=(V(i-1,j)+V(i+1,j)+V(i,j-1)...
                  +V(i,j+1)+E(i-1,j-1))/4;
    end
  end
  if max(max(abs(W-U)))<tol,  return
  end
  V=W;
end
k=K+1;
```

	k_{100}	k_{2500}	$k_{10\ 000}$	$k_{40\ 000}$	$k_{160\ 000}$
J	385	8386			
GS	194	4194			
SOR	35	164	324	645	1286

Table 9.1. *The number of iterations k_n to solve the $n \times n$ discrete Poisson problem using the methods of Jacobi, Gauss-Seidel, and SOR (see text) with a tolerance 10^{-8}.*

In Table 9.1 we show the output $k = k_n$ from this algorithm using $\boldsymbol{F} = \texttt{ones}(m, m)$ for $m = 10, 50$, $K = 10^4$, and $tol = 10^{-8}$. We also show the number of iterations for Gauss-Seidel and SOR with a value of ω known as the optimal acceleration parameter $\omega = 2/(1 + \sin(\pi/(m + 1)))$. We will derive this value later. For the GS and SOR methods we have used Algorithm 9.2.

Algorithm 9.2 (SOR) This is the analog of Algorithm 9.1 using GS and SOR instead of J to solve the discrete Poisson problem. w is an acceleration parameter with $0 < w < 2$. For $w = 1$ we obtain Gauss-Seidel's method. The optimal value for the discrete Poisson problem is $w = 2/(1 + \sin(\pi/(m + 1)))$.

```
function k=sordp(F,K,w,tol)
m=length(F);
U=fastpoisson(F);
V=zeros(m+2,m+2);
E=F/(m+1)^2;
for k=1:K
  for i=2:m+1
    for j=2:m+1
       V(i,j)=w*(V(i-1,j)+V(i+1,j)+V(i,j-1)...
              +V(i,j+1)+E(i-1,j-1))/4+(1-w)*V(i,j);
    end
  end
  if max(max(abs(V-U)))<tol,  return
  end
end
k=K+1;
```

We make several remarks about these programs and the results in Table 9.1.

1. The rate (speed) of convergence is quite different for the three methods. The J and GS method converge, but rather slowly. The J method needs about twice as many iterations as the GS method. The improvement using the SOR method with optimal ω is rather spectacular.

2. We show in Section 9.5.1 that the number of iterations k_n for a size n problem is $k_n = O(n)$ for the J and GS method and $k_n = O(\sqrt{n})$ for SOR with optimal

ω. The choice of *tol* will only influence the constants multiplying n or $\sqrt{n}$.

3. From (9.7) it follows that each iteration requires $O(n)$ flops. Thus the number of flops to achieve a given tolerance is $O(k_n \times n)$. Therefore the number of flops for the J and GS method is $O(n^2)$, while it is only $O(n^{3/2})$ for the SOR method with optimal ω. Asymptotically, for J and GS this is the same as using banded Cholesky, while SOR competes with the fast method (without FFT).

4. We do not need to store the coefficient matrix so the storage requirements for these methods on the discrete Poisson problem is $O(n)$, asymptotically the same as for the fast methods. For the GS and SOR method we can store the new $v_{ij}^{(k+1)}$ in the same location as $v_{ij}^{(k)}$. For Jacobi's method we need an extra array. (`W` in Algorithm 9.1).

5. Jacobi's method has the advantage that it can be easily parallelized.

9.3 Matrix Formulations of the Classical Methods

To study convergence it is convenient to use matrix formulations of the classical methods. In general we can construct an iterative method by choosing a nonsingular matrix $\boldsymbol{M}$ and write $\boldsymbol{Ax} = \boldsymbol{b}$ in the equivalent form $\boldsymbol{Bx} = \boldsymbol{c}$, where $\boldsymbol{B} = \boldsymbol{M}^{-1}\boldsymbol{A}$ and $\boldsymbol{c} = \boldsymbol{M}^{-1}\boldsymbol{b}$. The system $\boldsymbol{Bx} = \boldsymbol{c}$ can be written $\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{Bx} + \boldsymbol{c} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{x} + \boldsymbol{c}$, and this defines the iterative method

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}, \quad \boldsymbol{G} = \boldsymbol{I} - \boldsymbol{B} = \boldsymbol{I} - \boldsymbol{M}^{-1}\boldsymbol{A}, \quad \boldsymbol{c} = \boldsymbol{M}^{-1}\boldsymbol{b}. \tag{9.8}$$

Different choices of $\boldsymbol{M}$ leads to different iterative methods. The matrix $\boldsymbol{M}$ can be interpreted in two ways. It is a **preconditioning matrix** since a good choice of $\boldsymbol{M}$ will lead to a system $\boldsymbol{Bx} = \boldsymbol{c}$ with smaller condition number. It is also known as a **splitting matrix**, since if we split $\boldsymbol{A}$ in the form $\boldsymbol{A} = \boldsymbol{M} + (\boldsymbol{A} - \boldsymbol{M})$ then $\boldsymbol{Ax} = \boldsymbol{b}$ can be written $\boldsymbol{Mx} = (\boldsymbol{M} - \boldsymbol{A})\boldsymbol{x} + \boldsymbol{b}$ and this leads to the iterative method

$$\boldsymbol{M}\boldsymbol{x}^{(k+1)} = (\boldsymbol{M} - \boldsymbol{A})\boldsymbol{x}^{(k)} + \boldsymbol{b} \tag{9.9}$$

which is equivalent to (9.8).

The matrix $\boldsymbol{M}$ should be chosen so that $\boldsymbol{G}$ has small spectral radius and such that the linear system (9.9) is easy to solve for $\boldsymbol{x}^{(k+1)}$. These are conflicting demands. $\boldsymbol{M}$ should be an approximation to $\boldsymbol{A}$ to obtain a $\boldsymbol{B}$ with small condition number, but then (9.9) might not be easy to solve for $\boldsymbol{x}^{(k+1)}$.

9.3.1 The Splitting Matrices for the Classical Methods

To describe $\boldsymbol{M}$ for the classical methods we write $\boldsymbol{A}$ as a sum of three matrices, $\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{A_L} - \boldsymbol{A_R}$, where $-\boldsymbol{A_L}$, $\boldsymbol{D}$, and $-\boldsymbol{A_R}$ are the lower, diagonal, and upper

part of $\boldsymbol{A}$, respectively. Thus $\boldsymbol{D} := \operatorname{diag}(a_{11}, \ldots, a_{nn})$,

$$\boldsymbol{A_L} := \begin{bmatrix} 0 & & & \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n,1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix}, \quad \boldsymbol{A_R} := \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & \ddots & \ddots & \vdots \\ & & 0 & -a_{n-1,n} \\ & & & 0 \end{bmatrix}. \tag{9.10}$$

Proposition 9.3 *The splitting matrices $\boldsymbol{M}_J, \boldsymbol{M}_1, \boldsymbol{M}_\omega$ for the J, GS, and SOR method are given by*

$$\boldsymbol{M}_J = \boldsymbol{D}, \quad \boldsymbol{M}_1 = \boldsymbol{D} - \boldsymbol{A}_L, \quad \boldsymbol{M}_\omega = \omega^{-1}\boldsymbol{D} - \boldsymbol{A}_L. \tag{9.11}$$

Proof. The equation $\boldsymbol{Ax} = \boldsymbol{b}$ can be written $\boldsymbol{Dx} - \boldsymbol{A}_L\boldsymbol{x} - \boldsymbol{A}_R\boldsymbol{x} = \boldsymbol{b}$ or $\boldsymbol{Dx} = \boldsymbol{A}_L\boldsymbol{x} + \boldsymbol{A}_R\boldsymbol{x} + \boldsymbol{b}$. This leads to

$$\begin{aligned} J:&\ \boldsymbol{Dx}^{(k+1)} = \boldsymbol{A}_L\boldsymbol{x}^{(k)} + \boldsymbol{A}_R\boldsymbol{x}^{(k)} + \boldsymbol{b}, \\ GS:&\ \boldsymbol{Dx}^{(k+1)} = \boldsymbol{A}_L\boldsymbol{x}^{(k+1)} + \boldsymbol{A}_R\boldsymbol{x}^{(k)} + \boldsymbol{b}, \\ SOR:&\ \boldsymbol{Dx}^{(k+1)} = \omega\big(\boldsymbol{A}_L\boldsymbol{x}^{(k+1)} + \boldsymbol{A}_R\boldsymbol{x}^{(k)} + \boldsymbol{b}\big) + (1-\omega)\boldsymbol{Dx}^{(k)}. \end{aligned} \tag{9.12}$$

Writing these equations in the form (9.9) we obtain (9.11). ∎

Example 9.4 *For the system*

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

we find

$$\boldsymbol{A}_L = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{A}_R = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

and

$$\boldsymbol{M}_J = \boldsymbol{D} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{M}_\omega = \omega^{-1}\boldsymbol{D} - \boldsymbol{A}_L = \begin{bmatrix} 2\omega^{-1} & 0 \\ -1 & 2\omega^{-1} \end{bmatrix}.$$

The iteration matrix $\boldsymbol{G}_\omega = \boldsymbol{I} - \boldsymbol{M}_\omega^{-1}\boldsymbol{A}$ is given by

$$\boldsymbol{G}_\omega = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \omega/2 & 0 \\ \omega^2/4 & \omega/2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1-\omega & \omega/2 \\ \omega(1-\omega)/2 & 1-\omega+\omega^2/4 \end{bmatrix}. \tag{9.13}$$

For the J and GS method we have

$$\boldsymbol{G}_J = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}, \quad \boldsymbol{G}_1 = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix}. \tag{9.14}$$

We could have derived these matrices directly from the component form of the iteration. For example, for the GS method we have the component form

$$x_1^{(k+1)} = \frac{1}{2}x_2^{(k)} + \frac{1}{2}, \quad x_2^{(k+1)} = \frac{1}{2}x_1^{(k+1)} + \frac{1}{2}.$$

Substituting the value of $x_1^{(k+1)}$ *from the first equation into the second equation we find*

$$x_2^{(k+1)} = \frac{1}{2}(\frac{1}{2}x_2^{(k)} + \frac{1}{2}) + \frac{1}{2} = \frac{1}{4}x_2^{(k)} + \frac{3}{4}.$$

Thus

$$\boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} + \begin{bmatrix} 1/2 \\ 3/4 \end{bmatrix} = \boldsymbol{G}_1\boldsymbol{x}^{(k)} + \boldsymbol{c}.$$

9.4 Convergence of Fixed-point Iteration

We have seen that the classical methods can be written in the form (9.8) for a suitable $\boldsymbol{M}$. Starting with $\boldsymbol{x}^{(0)}$ this defines a sequence $\{\boldsymbol{x}^{(k)}\}$ of vectors in $\mathbb{C}^n$. If $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{C}^n$ then $\boldsymbol{x}$ is a solution of $\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{c}$ since

$$\boldsymbol{x} = \lim_{k\to\infty} \boldsymbol{x}^{(k+1)} = \lim_{k\to\infty} (\boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}) = \boldsymbol{G} \lim_{k\to\infty} \boldsymbol{x}^{(k)} + \boldsymbol{c} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{c}.$$

For a general $\boldsymbol{G} \in \mathbb{C}^{n,n}$ and $\boldsymbol{c} \in \mathbb{C}^n$ a solution of $\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x}+\boldsymbol{c}$ is called a **fixed-point** and the iteration $\boldsymbol{x}^{(k+1)} = \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}$ a **fixed-point iteration**. The fixed-point is unique if $\boldsymbol{I} - \boldsymbol{G}$ is nonsingular.

Consider next convergence of fixed-point iteration.

Definition 9.5 *We say that the iterative method* $\boldsymbol{x}^{(k+1)} := \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}$ **converges** *if the sequence* $\{\boldsymbol{x}^{(k)}\}$ *converges for any starting vector* $\boldsymbol{x}^{(0)}$.

To study convergence we consider for $k \geq 0$ the error

$$\boldsymbol{\epsilon}^{(k)} := \boldsymbol{x}^{(k)} - \boldsymbol{x}.$$

Lemma 9.6 *The iterative method* $\boldsymbol{x}^{(k+1)} := \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}$ *converges if and only if* $\lim_{k\to\infty} \boldsymbol{G}^k = \boldsymbol{0}$.

Proof. Subtraction of $\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{c}$ from $\boldsymbol{x}^{(k+1)} = \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}$ leads to cancellation of $\boldsymbol{c}$ and $\boldsymbol{\epsilon}^{(k+1)} = \boldsymbol{G}\boldsymbol{\epsilon}^{(k)}$. By induction $\boldsymbol{\epsilon}^{(k)} = \boldsymbol{G}^k\boldsymbol{\epsilon}^{(0)}$ for $k = 0, 1, 2, \ldots$. It follows that $\boldsymbol{\epsilon}^{(k)} \to 0$ for all $\boldsymbol{\epsilon}^{(0)}$ if and only if $\boldsymbol{G}^k \to \boldsymbol{0}$. ∎

Recall that the spectral radius of a matrix $\boldsymbol{G} \in \mathbb{C}^{n,n}$ with eigenvalues $\lambda_1 \ldots, \lambda_n$ is defined as $\rho(\boldsymbol{G}) = \max_j |\lambda_j|$. Using Theorem 8.46 we obtain the following theorem:

Theorem 9.7 *Suppose* $\boldsymbol{G} \in \mathbb{C}^{n,n}$ *and* $\boldsymbol{c} \in \mathbb{C}^n$. *The iteration* $\boldsymbol{x}^{(k+1)} = \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}$ *converges if and only if* $\rho(\boldsymbol{G}) < 1$.

Since $\rho(\boldsymbol{G}) < \|\boldsymbol{G}\|$ for any consistent matrix norm on $\mathbb{C}^{n,n}$ (cf. Theorem 8.46) we obtain

Corollary 9.8 *If* $\|\boldsymbol{G}\| < 1$ *for some consistent matrix norm, then the iteration* $\boldsymbol{x}^{(k+1)} = \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{c}$ *converges.*

Exercise 9.9 *Show that both Jacobi's method and Gauss-Seidel's method diverge for* $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$.

Exercise 9.10 *Explain why J and GS converge for the cubic spline matrix* $\boldsymbol{N}$ *in Chapter 2. (This is mainly of academic interest since for tridiagonal strictly diagonally dominant matrices Gaussian elimination has complexity* $O(n)$ *and is preferable for such systems.)*

Exercise 9.11 *Show that the J method converges if* $\boldsymbol{A}$ *is strictly diagonally dominant, i. e.,* $|a_{ii}| > \sum_{j\neq i}|a_{ij}|$ *for* $i = 1, \ldots, n$.

Exercise 9.12 *Consider the GS method. Suppose* $r := \max_i r_i < 1$, *where* $r_i = \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|}$. *Show using induction on* i *that* $|\boldsymbol{\epsilon}_j^{(k+1)}| \leq r\|\boldsymbol{\epsilon}^{(k)}\|_\infty$ *for* $j = 1, \ldots, i$. *Conclude that Gauss-Seidel's method is convergent when* $\boldsymbol{A}$ *is strictly diagonally dominant.*

Consider next the **rate of convergence**. Suppose $\|\ \|$ is a matrix norm that is subordinate to a vector norm also denoted by $\|\ \|$. Taking norms in $\boldsymbol{\epsilon}^{(k)} = \boldsymbol{G}^k\boldsymbol{\epsilon}^{(0)}$ we obtain

$$\|\boldsymbol{\epsilon}^{(k)}\| = \|\boldsymbol{G}^k\boldsymbol{\epsilon}^{(0)}\| \leq \|\boldsymbol{G}^k\|\|\boldsymbol{\epsilon}^{(0)}\| \approx \rho(\boldsymbol{G})^k\|\boldsymbol{\epsilon}^{(0)}\|.$$

For the last formula we apply Theorem 8.49 which says that $\lim_{k\to\infty}\|\boldsymbol{G}^k\|^{1/k} = \rho(\boldsymbol{G})$. Thus for fast convergence we should use a $\boldsymbol{G}$ with small spectral radius.

Lemma 9.13 *Suppose* $\rho(\boldsymbol{G}) = 1 - \eta$ *for some* $0 < \eta < 1$, $\|\ \|$ *a consistent matrix norm, and let* $s \in \mathbb{N}$. *Then*

$$\tilde{k} := \frac{\log(10)s}{\eta} \tag{9.15}$$

is an estimate for the smallest number of iterations k *so that* $\rho(\boldsymbol{G})^k \leq 10^{-s}$.

Proof. $\tilde{k}$ is an approximate solution of the equation $\rho(\boldsymbol{G})^k = 10^{-s}$. Indeed, taking logarithms we find $k\log\rho(\boldsymbol{G}) = -s\log 10$. Thus

$$k = -\frac{s\log(10)}{\log(1-\eta)} = \frac{s\log(10)}{\eta + O(\eta^2)} \approx \frac{\log(10)s}{\eta} = \tilde{k}.$$

□

Exercise 9.14 *Consider the iteration in Example 9.4. Show that* $\rho(\boldsymbol{G}_J) = 1/2$. *Then show that* $x_1^{(k)} = x_2^{(k)} = 1 - 2^{-k}$ *for* $k \geq 0$. *Thus the estimate in Lemma 9.13 is exact in this case.*

The convergence $\lim_{k\to\infty}\|\boldsymbol{G}^k\|^{1/k} = \rho(\boldsymbol{G})$ can be quite slow, (cf. Exercise 8.50).

9.4.1 Stopping the Iteration

In Algorithms 9.1 and 9.2 we had access to the exact solution and could stop the iteration when the error was sufficiently small in the infinity norm. The decision when to stop is obviously more complicated when the exact solution is not known. One possibility is to choose a vector norm, keep track of $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|$, and stop when this number is sufficiently small. This must be applied with some care if $\|\boldsymbol{G}\|$ is close to one, as the following result indicates.

Lemma 9.15 *Suppose $\|\boldsymbol{G}\| < 1$ for some consistent matrix norm which is subordinate to a vector norm also denoted by $\|\ \|$. If $\boldsymbol{x}^{(k)} = \boldsymbol{G}\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$ and $\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{c}$. Then*

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \frac{\|\boldsymbol{G}\|}{1 - \|\boldsymbol{G}\|}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|, \quad k \geq 1. \tag{9.16}$$

Proof. We find

$$\begin{aligned}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| &= \|\boldsymbol{G}(\boldsymbol{x}^{(k-1)} - \boldsymbol{x})\| \leq \|\boldsymbol{G}\|\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\| \\ &= \|\boldsymbol{G}\|\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)} + \boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|\boldsymbol{G}\|\big(\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}\| + \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|\big).\end{aligned}$$

Thus $(1 - \|\boldsymbol{G}\|)\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|\boldsymbol{G}\|\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}\|$ which implies (9.16). ∎

Another possibility is to stop when the residual vector $\boldsymbol{r}^{(k)} := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}$ is sufficiently small in some norm. To use the residual vector for stopping it is convenient to write the iterative method (9.8) in an alternative form. If $\boldsymbol{M}$ is the splitting matrix of the method then by (9.9) we have $\boldsymbol{M}\boldsymbol{x}^{(k+1)} = \boldsymbol{M}\boldsymbol{x}^{(k)} - \boldsymbol{A}\boldsymbol{x}^{(k)} + \boldsymbol{b}$. This leads to

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{M}^{-1}\boldsymbol{r}^{(k)}, \quad \boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}. \tag{9.17}$$

Testing on $\boldsymbol{r}^{(k)}$ works fine if $\boldsymbol{A}$ is well conditioned, but Theorem 8.42 shows that the relative error in the solution can be much larger than the relative error in $\boldsymbol{r}^{(k)}$ if $\boldsymbol{A}$ is ill-conditioned.

9.4.2 Richardson's Method (R method)

This method is based on the simple splitting $\boldsymbol{M}_R := \alpha\boldsymbol{I}$, where α is a nonzero scalar. By (9.17) we obtain Richardson's method in the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{-1}\boldsymbol{r}^{(k)}, \quad \boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}. \tag{9.18}$$

If all eigenvalues of $\boldsymbol{A}$ have positive real parts then the R method converges provided α is sufficiently large.

Proposition 9.16 *Suppose all eigenvalues of $\boldsymbol{A}$ have positive real parts and that α is real. Then there is an α_0 such that the R method converges for $\alpha > \alpha_0$. If $\boldsymbol{A}$ has positive eigenvalues $0 < \lambda_n \leq \cdots \leq \lambda_1$ then the spectral radius of*

$$\boldsymbol{G}(\alpha) := \boldsymbol{I} - \alpha^{-1}\boldsymbol{A}$$

is uniquely minimized if $\alpha = \alpha^*$, *where*

$$\alpha^* := \frac{\lambda_1 + \lambda_n}{2}, \text{ and } \rho\big(\boldsymbol{G}(\alpha^*)\big) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}. \tag{9.19}$$

Proof. The eigenvalues of $\boldsymbol{G}(\alpha)$ are

$$\mu_j(\alpha) = 1 - \lambda_j/\alpha, \quad j = 1, \ldots, n,$$

and if $u_j := Re\lambda_j > 0$ then

$$|\mu_j(\alpha)|^2 = (1 - \frac{\lambda_j}{\alpha})(1 - \frac{\overline{\lambda}_j}{\alpha}) = 1 - 2\frac{u_j}{\alpha} + \frac{|\lambda_j|^2}{\alpha^2} = 1 - \frac{|\lambda_j|^2}{\alpha^2}\left(\frac{2\alpha u_j}{|\lambda_j|^2} - 1\right) < 1$$

if $2\alpha > \max_j(|\lambda_j|^2/u_j)$ and the R method converges. We next show that $\rho\big(\boldsymbol{G}(\alpha)\big) > \rho\big(\boldsymbol{G}(\alpha^*)\big)$ if $\alpha \neq \alpha^*$. Indeed, if $\alpha > \alpha^*$ then

$$\rho\big(\boldsymbol{G}(\alpha)\big) \geq \mu_n(\alpha) = 1 - \lambda_n/\alpha > 1 - \lambda_n/\alpha^* = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \rho\big(\boldsymbol{G}(\alpha^*)\big).$$

Next, if $\alpha < \alpha^*$ then

$$-\rho\big(\boldsymbol{G}(\alpha)\big) \leq \mu_1(\alpha) = 1 - \lambda_1/\alpha < 1 - \lambda_1/\alpha^* = -\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = -\rho\big(\boldsymbol{G}(\alpha^*)\big),$$

and again $\rho\big(\boldsymbol{G}(\alpha)\big) > \rho\big(\boldsymbol{G}(\alpha^*)\big)$. ∎

9.5 Convergence of the Classical Methods for the Discrete Poisson Matrix

The matrix $\boldsymbol{A}$ in (4.7) is symmetric positive definite (cf. Theorem 4.13). We show in Theorem 9.23 that the SOR method converges for all $0 < \omega < 2$ if $\boldsymbol{A}$ is symmetric positive definite. So the GS method converges, but the J method does not converge for all symmetric positive definite matrices.

Exercise 9.17 *Show (by finding its eigenvalues) that the matrix*

$$\begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

is symmetric positive definite for $-1/2 < a < 1$, *but that the J method does not converge for* $1/2 < a < 1$.

For the discrete Poisson problem we can determine explicitly the eigenvalues of the iteration matrices and thus not only show convergence, but also estimate the number of iterations necessary to achieve a given accuracy.

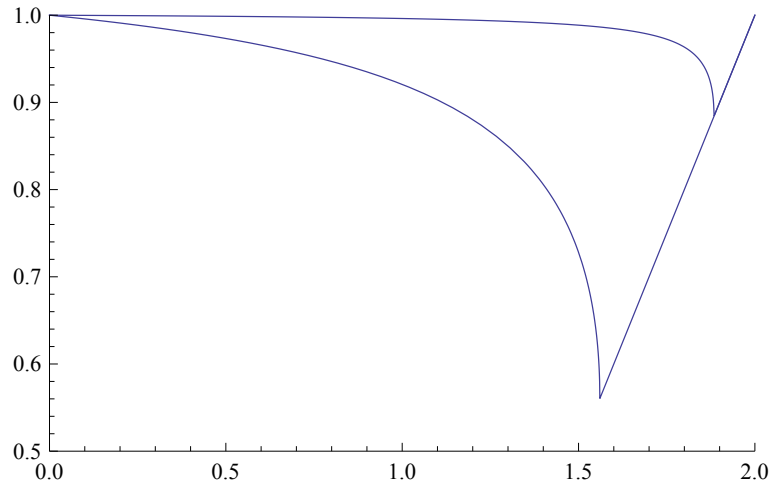


Figure 9.1. $\rho(\boldsymbol{G}_\omega)$ *with* $\omega \in [0,2]$ *for* $n = 100$*, (lower curve) and* $n = 2500$ *(upper curve).*

Recall that by (4.22) the eigenvalues $\lambda_{j,k}$ of $\boldsymbol{A}$ given by (4.7) are

$$\lambda_{j,k} = 4 - 2\cos(j\pi h) - 2\cos(k\pi h), \quad j,k = 1,\ldots,m, \quad h = 1/(m+1).$$

Consider first Jacobi's method. The matrix $\boldsymbol{G}_J = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{A}/4$ has eigenvalues

$$\mu_{j,k} = 1 - \frac{1}{4}\lambda_{j,k} = \frac{1}{2}\cos(j\pi h) + \frac{1}{2}\cos(k\pi h), \quad j,k = 1,\ldots,m. \tag{9.20}$$

It follows that $\rho(\boldsymbol{G}_J) = \cos(\pi h) < 1$ and the J method converges for all starting values and all right hand sides.

For the SOR method it is possible to explicitly determine $\rho(\boldsymbol{G}_\omega)$ for any $\omega \in (0,2)$. The following result will be shown in Section 9.6.

Theorem 9.18 *Consider the SOR iteration* (9.7)*, whith the natural ordering. The spectral radius of* $\boldsymbol{G}_\omega$ *is*

$$\rho(\boldsymbol{G}_\omega) = \begin{cases} \frac{1}{4}\left(\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega-1)}\right)^2, & \text{for } 0 < \omega \le \omega^*, \\ \omega - 1, & \text{for } \omega^* < \omega < 2, \end{cases} \tag{9.21}$$

where $\beta := \rho(\boldsymbol{G}_J)$ *and*

$$\omega^* := \frac{2}{1 + \sqrt{1-\beta^2}} > 1. \tag{9.22}$$

Moreover,

$$\rho(\boldsymbol{G}_\omega) > \rho(\boldsymbol{G}_{\omega^*}) \text{ for } \omega \in (0,2) \setminus \{\omega^*\}. \tag{9.23}$$

	n=100	n=2500	k_{100}	k_{2500}
J	0.959493	0.998103	446	9703
GS	0.920627	0.99621	223	4852
SOR	0.56039	0.88402	32	150

Table 9.2. *Spectral radia for $\boldsymbol{G}_J$, $\boldsymbol{G}_1$, $\boldsymbol{G}_{\omega^*}$ and the smallest integer k_n such that $\rho(\boldsymbol{G})^{k_n} \leq 10^{-8}$.*

A plot of $\rho(\boldsymbol{G}_\omega)$ as a function of $\omega \in (0, 2)$ is shown in Figure 9.1 for $n = 100$ (lower curve) and $n = 2500$ (upper curve). As ω increases the spectral radius of $\boldsymbol{G}_\omega$ decreases monotonically to the minimum ω^*. Then it increases linearly to the value one for $\omega = 2$. We call ω^* the **optimal relaxation parameter**.

For the discrete Poisson problem we have $\beta = \cos(\pi h)$ and it follows from (9.21),(9.22) that

$$\omega^* = \frac{2}{1+\sin(\pi h)}, \quad \rho(\boldsymbol{G}_{\omega^*}) = \omega^* - 1 = \frac{1-\sin(\pi h)}{1+\sin(\pi h)}, \quad h = \frac{1}{m+1}. \tag{9.24}$$

Letting $\omega = 1$ in (9.21) we find $\rho(\boldsymbol{G}_1) = \beta^2 = \rho(\boldsymbol{G}_J)^2 = \cos^2(\pi h)$. Thus, for the discrete Poisson problem the J method needs twice as many iterations as the GS method for a given accuracy.

The values of $\rho(\boldsymbol{G}_J)$, $\rho(\boldsymbol{G}_1)$, and $\rho(\boldsymbol{G}_{\omega^*}) = \omega^* - 1$ are shown in Table 9.2 for $n = 100$ and $n = 2500$. We also show the smallest integer k_n such that $\rho(\boldsymbol{G})^{k_n} \leq 10^{-8}$. This is an estimate for the number of iteration needed to obtain an accuracy of 10^{-8}. These values are comparable to the exact values given in Table 9.1.

9.5.1 Number of Iterations

Let s be a positive integer. We can now estimate the number of iterations k_n to obtain $\rho(\boldsymbol{G})^{k_n} < 10^{-s}$ for the J, GS and SOR method with optimal ω. We use Lemma 9.13 that provided the estimate

$$\tilde{k}_n = \frac{\log(10)s}{\eta}, \quad \rho(\boldsymbol{G}) = 1 - \eta.$$

Note that $h = 1/(m+1) \approx n^{-1/2}$. The estimates we derive agree with those we found numerically in Section 9.2.

- J: $\rho(\boldsymbol{G}_J) = \cos(\pi h) = 1 - \eta$, $\eta = 1 - \cos(\pi h) = \frac{1}{2}\pi^2 h^2 + O(h^4) = \frac{\pi^2}{2}/n + O(n^{-2})$. Thus
$$\tilde{k}_n = \frac{2\log(10)s}{\pi^2} n + O(n^{-1}) = O(n).$$
- GS: $\rho(\boldsymbol{G}_1) = \cos^2(\pi h) = 1 - \eta$, $\eta = 1 - \cos^2(\pi h) = \sin^2 \pi h = \pi^2 h^2 + O(h^4) = \pi^2/n + O(n^{-2})$. Thus
$$\tilde{k}_n = \frac{\log(10)s}{\pi^2} n + O(n^{-1}) = O(n).$$

- SOR: $\rho(\boldsymbol{G}_{\omega^*}) = \frac{1-\sin(\pi h)}{1+\sin(\pi h)} = 1 - 2\pi h + O(h^2)$. Thus

$$\tilde{k}_n = \frac{\log(10)s}{\pi^2}\sqrt{n} + O(n^{-1/2}) = O(\sqrt{n}).$$

Exercise 9.20 *Consider for $a \in \mathbb{C}$*

$$\boldsymbol{x} := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1-a \\ 1-a \end{bmatrix} =: \boldsymbol{G}\boldsymbol{x} + \boldsymbol{c}.$$

Starting with $\boldsymbol{x}^{(0)} = \boldsymbol{0}$ show by induction

$$x_1^{(k)} = x_2^{(k)} = 1 - a^k, \quad k \geq 0,$$

and conclude that the iteration converges to the fixed-point $\boldsymbol{x} = [1,1]^T$ for $|a| < 1$ and diverges for $|a| > 1$. Show that $\rho(\boldsymbol{G}) = 1 - \eta$ with $\eta = 1 - |a|$. Compute the estimate (9.15) *for the rate of convergence for $a = 0.9$ and $s = 16$ and compare with the true number of iterations determined from $|a|^k \leq 10^{-16}$.*

9.6 Convergence Analysis for SOR

The iteration matrix $\boldsymbol{G}_\omega$ for the SOR method can be written in two alternative forms that are both useful for the analysis.

Lemma 9.21 *Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ and $\boldsymbol{D} = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ are both nonsingular. Then*

$$\boldsymbol{G}_\omega = \boldsymbol{I} - (\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{L}})^{-1}\boldsymbol{A} = (\boldsymbol{I} - \omega\boldsymbol{L})^{-1}\big(\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}\big), \tag{9.25}$$

where $\boldsymbol{A}_{\boldsymbol{L}}$ and $\boldsymbol{A}_{\boldsymbol{R}}$ are given by (9.10) *and*

$$\boldsymbol{L} := \boldsymbol{D}^{-1}\boldsymbol{A}_{\boldsymbol{L}}, \quad \boldsymbol{R} := \boldsymbol{D}^{-1}\boldsymbol{A}_{\boldsymbol{R}}, \text{ so that } \boldsymbol{D}^{-1}\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{L} - \boldsymbol{R}. \tag{9.26}$$

Proof. For the first form see (9.8) and Proposition 9.3. Solving the SOR part of (9.12) for $\boldsymbol{x}^{(k+1)}$ gives

$$\boldsymbol{x}^{(k+1)} = \omega\big(\boldsymbol{L}\boldsymbol{x}^{(k+1)} + \boldsymbol{R}\boldsymbol{x}^{(k)} + \boldsymbol{D}^{-1}\boldsymbol{b}\big) + (1-\omega)\boldsymbol{x}^{(k)},$$

or

$$(\boldsymbol{I} - \omega\boldsymbol{L})\boldsymbol{x}^{(k+1)} = \big(\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}\big)\boldsymbol{x}^{(k)} + \omega\boldsymbol{D}^{-1}\boldsymbol{b}.$$

Solving for $\boldsymbol{x}^{(k+1)}$ we obtain $\boldsymbol{x}^{(k+1)} = \boldsymbol{G}_\omega\boldsymbol{x}^{(k)} + \boldsymbol{c}$, where $\boldsymbol{G}_\omega$ is given by the second form in (9.25). ∎

We start with the following convergence result.

Theorem 9.22 *The SOR method diverges if ω is not in the interval $(0, 2)$.*

Proof. Recall that the determinant of a product equals the product of determinants and that the determinant of a triangular matrix equals the product of the diagonal elements. From (9.25) we obtain

$$\det(\boldsymbol{G}_\omega) = \det\left((\boldsymbol{I} - \omega\boldsymbol{L})^{-1}\right)\det\left(\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}\right).$$

Since $\boldsymbol{I} - \omega\boldsymbol{L}$ is lower triangular with ones on the diagonal it follows from Lemma 2.8 that the first determinant equals one. The matrix $\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}$ is upper triangular with $1-\omega$ on the diagonal and therefore its determinant equals $(1-\omega)^n$. It follows that $\det(\boldsymbol{G}_\omega) = (1-\omega)^n$.

Since the determinant of a matrix equals the product of its eigenvalues we must have $|\lambda| \geq |1-\omega|$ for at least one eigenvalue λ of $\boldsymbol{G}_\omega$. We conclude that $\rho(\boldsymbol{G}_\omega) \geq |\omega - 1|$. But then $\rho(\boldsymbol{G}_\omega) \geq 1$ if ω is not in the interval $(0,2)$ and by Theorem 9.7 SOR diverges. ∎

We next show that SOR converges for all $\omega \in (0,2)$ if $\boldsymbol{A}$ is symmetric positive definite.

Theorem 9.23 *SOR converges for a symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ if and only if $0 < \omega < 2$. In particular, Gauss-Seidel's method converges for a symmetric positive definite matrix.*

Proof. By Theorem 9.22 convergence implies $0 < \omega < 2$. Suppose $0 < \omega < 2$. The eigenpair equation $\boldsymbol{G}_\omega\boldsymbol{x} = \lambda\boldsymbol{x}$ can be written $\boldsymbol{x} - (\omega^{-1}\boldsymbol{D} - \boldsymbol{A_L})^{-1}\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ or

$$\boldsymbol{A}\boldsymbol{x} = (\omega^{-1}\boldsymbol{D} - \boldsymbol{A_L})\boldsymbol{y}, \quad \boldsymbol{y} := (1-\lambda)\boldsymbol{x}. \tag{9.27}$$

Since $\boldsymbol{A} = -\boldsymbol{A_L} + \boldsymbol{D} - \boldsymbol{A_R}$ we find

$$(\omega^{-1}\boldsymbol{D} - \boldsymbol{D} + \boldsymbol{A_R})\boldsymbol{y} = (\omega^{-1}\boldsymbol{D} - \boldsymbol{A_L} - \boldsymbol{A})\boldsymbol{y} \overset{(9.27)}{=} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y} = \lambda\boldsymbol{A}\boldsymbol{x},$$

so that by taking inner products and replacing $\boldsymbol{A}_{\boldsymbol{R}}^*$ by $\boldsymbol{A_L}$

$$\begin{aligned}\langle \boldsymbol{y}, \lambda\boldsymbol{A}\boldsymbol{x}\rangle &= \langle \boldsymbol{y}, (\omega^{-1}\boldsymbol{D} - \boldsymbol{D} + \boldsymbol{A_R})\boldsymbol{y}\rangle = \langle (\omega^{-1}\boldsymbol{D} - \boldsymbol{D} + \boldsymbol{A}_{\boldsymbol{R}}^*)\boldsymbol{y}, \boldsymbol{y}\rangle \\ &= \langle (\omega^{-1}\boldsymbol{D} - \boldsymbol{D} + \boldsymbol{A_L})\boldsymbol{y}, \boldsymbol{y}\rangle.\end{aligned} \tag{9.28}$$

Taking inner product with $\boldsymbol{y}$ in (9.27) and adding to (9.28) we obtain

$$\begin{aligned}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \lambda\boldsymbol{A}\boldsymbol{x}\rangle &= \langle (\omega^{-1}\boldsymbol{D} - \boldsymbol{A_L})\boldsymbol{y}, \boldsymbol{y}\rangle + \langle (\omega^{-1}\boldsymbol{D} - \boldsymbol{D} + \boldsymbol{A_L})\boldsymbol{y}, \boldsymbol{y}\rangle \\ &= (2\omega^{-1} - 1)\langle \boldsymbol{D}\boldsymbol{y}, \boldsymbol{y}\rangle = (2\omega^{-1} - 1)(1-\lambda)(1-\overline{\lambda})\langle \boldsymbol{D}\boldsymbol{x}, \boldsymbol{x}\rangle \\ &= (2\omega^{-1} - 1)|1-\lambda|^2\langle \boldsymbol{D}\boldsymbol{x}, \boldsymbol{x}\rangle.\end{aligned}$$

On the other hand, since $\boldsymbol{A}$ is symmetric

$$\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \lambda\boldsymbol{A}\boldsymbol{x}\rangle = (1-\overline{\lambda})\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle + (1-\lambda)\overline{\lambda}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle = (1-|\lambda|^2)\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle.$$

Thus,

$$(2\omega^{-1} - 1)|1-\lambda|^2\langle \boldsymbol{D}\boldsymbol{x}, \boldsymbol{x}\rangle = (1-|\lambda|^2)\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle. \tag{9.29}$$

Since $\boldsymbol{A}$ is symmetric positive definite we observe that also $\boldsymbol{D}$ is symmetric positive definite. Furthermore we cannot have $\lambda = 1$ for then $\boldsymbol{y} = \boldsymbol{0}$ which by (9.27) implies that $\boldsymbol{A}$ is singular. Since $0 < \omega < 2$ implies $\omega^{-1} > 1/2$ the left side of (9.29) is positive and hence the right hand side is positive as well. We conclude that $|\lambda| < 1$. But then $\rho(\boldsymbol{G}_\omega) < 1$ and SOR converges. □

9.7 The Optimal SOR Parameter ω

The following analysis holds both for the discrete Poisson matrix and the averaging matrix given by (4.9). A more general theory is presented in [26]. Consider first how the eigenvalues of $\boldsymbol{G}_J$ and $\boldsymbol{G}_\omega$ are related.

Theorem 9.24 *Consider for $a, d \in \mathbb{R}$ the SOR method applied to the matrix (4.9), where we use the natural ordering. Moreover, assume $\omega \in (0, 2)$.*

1. *If $\lambda \neq 0$ is an eigenvalue of $\boldsymbol{G}_\omega$ then*

$$\mu := \frac{\lambda + \omega - 1}{\omega \lambda^{1/2}} \tag{9.30}$$

   is an eigenvalue of $\boldsymbol{G}_J$.

2. *If μ is an eigenvalue of $\boldsymbol{G}_J$ and λ satisfies the equation*

$$\mu\omega\lambda^{1/2} = \lambda + \omega - 1 \tag{9.31}$$

   then λ is an eigenvalue of $\boldsymbol{G}_\omega$.

Proof. For simplicity of notation we assume that $a = -1$ and $d = 2$. The component equations in this proof hold for $i, j = 1, \ldots, m$. Suppose $(\lambda, \boldsymbol{w})$ is an eigenpair for $\boldsymbol{G}_\omega$. By (9.25) $(\boldsymbol{I} - \omega\boldsymbol{L})^{-1}\big(\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}\big)\boldsymbol{w} = \lambda\boldsymbol{w}$ or

$$(\omega\boldsymbol{R} + \lambda\omega\boldsymbol{L})\boldsymbol{w} = (\lambda + \omega - 1)\boldsymbol{w}. \tag{9.32}$$

Let $\boldsymbol{w} = \mathrm{vec}(\boldsymbol{W})$, where $\boldsymbol{W} \in \mathbb{C}^{m,m}$. Then (9.32) can be written

$$\frac{\omega}{4}(\lambda w_{i-1,j} + \lambda w_{i,j-1} + w_{i+1,j} + w_{i,j+1}) = (\lambda + \omega - 1)w_{i,j}, \tag{9.33}$$

where $w_{i,j} = 0$ if $i \in \{0, m+1\}$ or $j \in \{0, m+1\}$. We claim that $(\mu, \boldsymbol{v})$ is an eigenpair for $\boldsymbol{G}_J$, where μ is given by (9.30) and $\boldsymbol{v} = \mathrm{vec}(\boldsymbol{V})$ with

$$v_{i,j} := \lambda^{-(i+j)/2} w_{i,j}. \tag{9.34}$$

Indeed, replacing $w_{i,j}$ by $\lambda^{(i+j)/2} v_{i,j}$ in (9.33) and cancelling the common factor $\lambda^{(i+j)/2}$ we obtain

$$\frac{\omega}{4}(v_{i-1,j} + v_{i,j-1} + v_{i+1,j} + v_{i,j+1}) = \lambda^{-1/2}(\lambda + \omega - 1)v_{i,j}.$$

But then

$$\boldsymbol{G}_J\boldsymbol{v} = (\boldsymbol{L}+\boldsymbol{R})\boldsymbol{v} = \frac{\lambda+\omega-1}{\omega\lambda^{1/2}} = \mu\boldsymbol{v}.$$

For the converse let $(\mu, \boldsymbol{v})$ be an eigenpair for $\boldsymbol{G}_J$ and let as before $\boldsymbol{v} = \mathrm{vec}(\boldsymbol{V})$, $\boldsymbol{W} = \mathrm{vec}(\boldsymbol{W})$ with $v_{i,j} = \lambda^{-(i+j)/2} w_{i,j}$. The equation $\boldsymbol{G}_J\boldsymbol{v} = \mu\boldsymbol{v}$ can be written

$$\frac{1}{4}(v_{i-1,j} + v_{i,j-1} + v_{i+1,j} + v_{i,j+1}) = \mu v_{i,j}.$$

Let λ be a solution of (9.31). Replacing $v_{i,j}$ by $\lambda^{-(i+j)/2} w_{i,j}$ and canceling $\lambda^{-(i+j)/2}$ we obtain

$$\frac{1}{4}(\lambda^{1/2} w_{i-1,j} + \lambda^{1/2} w_{i,j-1} + \lambda^{-1/2} w_{i+1,j} + \lambda^{-1/2} w_{i,j+1}) = \mu w_{i,j},$$

or, multiplying by $\omega\lambda^{1/2}$

$$\frac{\omega}{4}(\lambda w_{i-1,j} + \lambda w_{i,j-1} + w_{i+1,j} + w_{i,j+1}) = \omega\mu\lambda^{1/2} w_{i,j},$$

Thus, if $\omega\mu^{1/2} = \lambda + \omega - 1$ then by (9.33) $(\lambda, \boldsymbol{w})$ is an eigenpair for $\boldsymbol{G}_\omega$. □

Proof of Theorem 9.18 By (4.22) the eigenvalues of $\boldsymbol{G}_J = \boldsymbol{I} - \boldsymbol{A}/(2d)$ are given by

$$\mu_{j,k} = -a\big(\cos(j\pi h) + \cos(k\pi h)\big)/(2d), \quad j,k = 1,\ldots,m.$$

Thus the eigenvalues are real and if μ is an eigenvalue then $-\mu$ is also an eigenvalue. Thus it is enough to consider positive eigenvalues μ. For simplicity of notation let again $a = -1$ and $d = 2$. Solving (9.31) for λ gives

$$\lambda(\mu) := \frac{1}{4}\Big(\omega\mu \pm \sqrt{(\omega\mu)^2 - 4(\omega-1)}\Big)^2. \tag{9.35}$$

Both roots $\lambda(\mu)$ are eigenvalues of $\boldsymbol{G}_\omega$. The discriminant

$$d(\omega) := (\omega\mu)^2 - 4(\omega-1).$$

is strictly decreasing on $(0,2)$ since

$$d'(\omega) = 2(\omega\mu^2 - 2) < 2(\omega - 2) < 0.$$

Moreover $d(0) = 4 > 0$ and $d(2) = 4\mu^2 - 4 < 0$. As a function of ω, $\lambda(\mu)$ changes from real to complex at

$$\omega = \tilde{\omega}(\mu) := \frac{2}{1+\sqrt{1-\mu^2}}. \tag{9.36}$$

In the complex case we find

$$|\lambda(\mu)| = \frac{1}{4}\Big((\omega\mu)^2 + 4(\omega-1) - (\omega\mu)^2\Big) = \omega - 1, \quad \tilde{\omega}(\mu) < \omega < 2.$$

In the real case both roots of (9.35) are positive and the larger one is

$$\lambda(\mu) = \frac{1}{4}\Big(\omega\mu + \sqrt{(\omega\mu)^2 - 4(\omega-1)}\Big)^2, \quad 0 < \omega \le \tilde{\omega}(\mu). \tag{9.37}$$

Both $\lambda(\mu)$ and $\tilde{\omega}(\mu)$ are strictly increasing as functions of μ. It follows that $|\lambda(\mu)|$ is maximized for $\mu = \rho(\boldsymbol{G}_J) =: \beta$ and for this value of μ we obtain (9.21) for $0 < \omega \le \tilde{\omega}(\beta) = \omega^*$.

Evidently $\rho(\boldsymbol{G}_\omega) = \omega - 1$ is strictly increasing in $\omega^* < \omega < 2$. Equation (9.23) will follow if we can show that $\rho(\boldsymbol{G}_\omega)$ is strictly decreasing in $0 < \omega < \omega^*$. By differentiation

$$\frac{d}{d\omega}\Big(\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega-1)}\Big) = \frac{\beta\sqrt{(\omega\beta)^2 - 4(\omega-1)} + \omega\beta^2 - 2}{\sqrt{(\omega\beta)^2 - 4(\omega-1)}}.$$

Since $\beta^2(\omega^2\beta^2 - 4\omega + 4) < (2 - \omega\beta^2)^2$ the numerator is negative and the strict decrease of $\rho(\boldsymbol{G}_\omega)$ in $0 < \omega < \omega^*$ follows.

Chapter 10

The Conjugate Gradient Method

The conjugate gradient method is an iterative method for solving large sparse linear systems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with a symmetric positive definite coefficient matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$. It can also be used to minimize a quadratic function $Q : \mathbb{R}^n \to \mathbb{R}$ given by $Q(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}$, see the following chapter. We compute a sequence of approximations to the exact solution. Each new approximation $\boldsymbol{x}^{(k+1)}$ is computed from the previous $\boldsymbol{x}^{(k)}$ by a formula of the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}, \tag{10.1}$$

where $\boldsymbol{p}^{(k)}$ is a vector, the **search direction**, and α_k is a scalar determining the **step length**. A characteristic of the method is that the residuals $\boldsymbol{r}^{(k)} := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}$ (the negative gradients of $Q(\boldsymbol{x}^{(k)})$) are orthogonal (or conjugate), i. e., $(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)}) = 0$ for $i \neq j$, where $(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^T\boldsymbol{y}$ is the usual inner product in $\mathbb{R}^n$. This orthogonality property has given the method its name. If $\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{r}^{(n-1)}$ are nonzero then $\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{r}^{(n)}$ are $n+1$ orthogonal vectors in $\mathbb{R}^n$ and $\boldsymbol{r}^{(n)}$ must be zero. It follows that the conjugate gradient method is a direct method. The exact solution is found in a finite number of operations. In practice, however the method is used as an iterative method for large linear systems since the residuals become small quite rapidly. For the Poisson problem the method converges as fast as the SOR-method with optimal acceleration parameter and we do not have to estimate the parameter. The conjugate gradient method was first published as a direct method in [7]. It was only some 20-30 years later that the iterative nature was seriously appreciated.

The number of iterations to achieve a desired accuracy is essentially proportional to the square root of the 2-norm condition number of the coefficient matrix of the linear system. Thus the smaller the condition number the faster the method converges.

Before deriving the method we give the algorithm, discuss implementation and give numerical examples.

10.1 The Conjugate Gradient Algorithm

Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is symmetric positive definite and let $\boldsymbol{b} \in \mathbb{R}^n$. To solve the linear system $\boldsymbol{Ax} = \boldsymbol{b}$ we choose an initial guess $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, set $\boldsymbol{p}^{(0)} := \boldsymbol{r}^{(0)} := \boldsymbol{b} - \boldsymbol{Ax}^{(0)}$ and generate a sequence of vectors $\{\boldsymbol{x}^{(k)}\}$ as follows:

For $k = 0,1,2,\ldots$

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}, \quad \alpha_k := \frac{(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)})}{(\boldsymbol{p}^{(k)}, \boldsymbol{Ap}^{(k)})}, \tag{10.2}$$

$$\boldsymbol{r}^{(k+1)} := \boldsymbol{r}^{(k)} - \alpha_k \boldsymbol{Ap}^{(k)}, \tag{10.3}$$

$$\boldsymbol{p}^{(k+1)} := \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{p}^{(k)}, \quad \beta_k := \frac{(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(k+1)})}{(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)})}. \tag{10.4}$$

Here $(\boldsymbol{u}, \boldsymbol{v}) := \boldsymbol{u}^T \boldsymbol{v}$ is the usual inner product of two vectors.

By induction on k we have $\boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{Ax}^{(k)}$. This follows by definition for $k = 0$, and if it holds for k then by (10.3) and (10.2) $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k \boldsymbol{Ap}^{(k)} = \boldsymbol{b} - \boldsymbol{Ax}^{(k)} - \alpha_k \boldsymbol{Ap}^{(k)} = \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}) = \boldsymbol{b} - \boldsymbol{Ax}^{(k+1)}$.

Example 10.1 *Consider the linear system*

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Starting with $\boldsymbol{x}^{(0)} = \boldsymbol{0}$ *we set* $\boldsymbol{p}^{(0)} = \boldsymbol{r}^{(0)} = \boldsymbol{b} = [1,0]^T$. *Using* (10.2) *we find* $\alpha_0 = \frac{(\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(0)})}{(\boldsymbol{p}^{(0)}, \boldsymbol{Ap}^{(0)})} = \frac{1}{2}$. *Then* $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0 \boldsymbol{p}^{(0)} = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] + \frac{1}{2}\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1/2 \\ 0 \end{smallmatrix}\right]$ *and from* (10.3) *we find* $\boldsymbol{r}^{(1)} = \boldsymbol{r}^{(0)} - \alpha_0 \boldsymbol{Ap}^{(0)} = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right] - \frac{1}{2}\left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 \\ 1/2 \end{smallmatrix}\right]$. *By* (10.4) *we find* $\beta_0 = \frac{(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(1)})}{(\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(0)})} = \frac{1}{4}$ *so that*

$$\boldsymbol{p}^{(1)} = \boldsymbol{r}^{(1)} + \beta_0 \boldsymbol{p}^{(0)} = \left[\begin{smallmatrix} 0 \\ 1/2 \end{smallmatrix}\right] + \frac{1}{4}\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right] = \begin{bmatrix} 1/4 \\ 1/2 \end{bmatrix}.$$

Continuing with the next iteration we obtain $\alpha_1 = \frac{(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(1)})}{(\boldsymbol{p}^{(1)}, \boldsymbol{Ap}^{(1)})} = \frac{2}{3}$ *and* $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \alpha_1 \boldsymbol{p}^{(1)} = \left[\begin{smallmatrix} 1/2 \\ 0 \end{smallmatrix}\right] + \frac{2}{3}\begin{bmatrix} 1/4 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$. *Since* $\boldsymbol{r}^{(2)} = \boldsymbol{r}^{(1)} - \alpha_1 \boldsymbol{Ap}^{(1)} = \boldsymbol{0}$ *this is the exact solution found in* $n = 2$ *iterations.*

Exercise 10.2 *Do one iteration with the conjugate gradient method when* $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. *(Answer:* $\boldsymbol{x}^{(1)} = \frac{(\boldsymbol{b},\boldsymbol{b})}{(\boldsymbol{b},\boldsymbol{Ab})}\boldsymbol{b}$.*)*

Exercise 10.3 *Do two conjugate gradient iterations for the system*

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$$

starting with $\boldsymbol{x}^{(0)} = \boldsymbol{0}$.

The formulas in (10.2)-(10.4) and the previous discussion form a basis for an algorithm.

Algorithm 10.4 (Conjugate Gradient Iteration) The symmetric positive definite linear system $\boldsymbol{Ax} = \boldsymbol{b}$ is solved by the conjugate gradient method. $\boldsymbol{x}$ is a starting vector for the iteration. The iteration is stopped when $||\boldsymbol{r}^{(k)}||_2/||\boldsymbol{r}^{(0)}||_2 \leq \text{tol}$ or $k > \text{itmax}$. K is the number of iterations used.

```
function [x,K]=cg(A,b,x,tol,itmax)
r=b-A*x; p=r; rho=r'*r;
rho0=rho; for k=0:itmax
    if sqrt(rho/rho0)<= tol
        K=k; return
    end
    t=A*p; a=rho/(p'*t);
    x=x+a*p; r=r-a*t;
    rhos=rho; rho=r'*r;
    p=r+(rho/rhos)*p;
end
K=itmax+1;
```

The work involved in each iteration is

1. one matrix times vector ($\boldsymbol{t} = \boldsymbol{Ap}$),
2. two inner products (($\boldsymbol{p}, \boldsymbol{t}$) and ($\boldsymbol{r}, \boldsymbol{r}$)),
3. three vector-plus-scalar-times-vector ($\boldsymbol{x} = \boldsymbol{x} + a\boldsymbol{p}$, $\boldsymbol{r} = \boldsymbol{r} - a\boldsymbol{t}$ and $\boldsymbol{p} = \boldsymbol{r} + (rho/rhos)\boldsymbol{p}$),

The dominating part is the computation of $\boldsymbol{t} = \boldsymbol{Ap}$.

10.2 Numerical Example

We test the method on the example used in Chapter 9. The matrix is given by the Kronecker sum $\boldsymbol{T}_2 := \boldsymbol{T}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}_1$ where $\boldsymbol{T}_1 = \text{tridiag}_m(a, d, a)$. We recall that this matrix is symmetric positive definite if $d > 0$ and $d \geq 2|a|$. We set $h = 1/(m+1)$ and $\boldsymbol{f} = [1, \ldots, 1]^T \in \mathbb{R}^n$.

Note that for our test problems $\boldsymbol{T}_2$ only has $O(5n)$ nonzero elements. Therefore, taking advantage of the sparseness of $\boldsymbol{T}_2$ we can compute $\boldsymbol{t}$ in Algorithm 10.4 in $O(n)$ flops. With such an implementation the total number of flops in one iteration is $O(n)$. We also note that it is not necessary to store the matrix $\boldsymbol{T}_2$.

To use the Conjugate Gradient Algorithm on the test matrix for large n it is advantageous to use a matrix equation formulation. We define matrices $\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{P}, \boldsymbol{B}, \boldsymbol{T} \in \mathbb{R}^{m,m}$ by $\boldsymbol{x} = \text{vec}(\boldsymbol{V})$, $\boldsymbol{r} = \text{vec}(\boldsymbol{R})$, $\boldsymbol{p} = \text{vec}(\boldsymbol{P})$, $\boldsymbol{t} = \text{vec}(\boldsymbol{T})$, and $h^2\boldsymbol{f} = \text{vec}(\boldsymbol{B})$. Then $\boldsymbol{T}_2\boldsymbol{x} = h^2\boldsymbol{f} \Longleftrightarrow \boldsymbol{T}_1\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T}_1 = \boldsymbol{B}$, and $\boldsymbol{t} = \boldsymbol{T}_2\boldsymbol{p} \Longleftrightarrow \boldsymbol{T} = \boldsymbol{T}_1\boldsymbol{P} + \boldsymbol{P}\boldsymbol{T}_1$.

This leads to the following algorithm for testing the conjugate gradient algorithm.

n	2 500	10 000	40 000	1 000 000	4 000 000
K	13	12	11	9	8

Table 10.6. *The number of iterations K for the averaging problem on a $\sqrt{n} \times \sqrt{n}$ grid for various n*

n	2 500	10 000	40 000	160 000
K	80	160	321	647
$K/\sqrt{n}$	1.6	1.6	1.61	1.62

Table 10.7. *The number of iterations K for the Poisson problem on a $\sqrt{n} \times \sqrt{n}$ grid for various n*

Algorithm 10.5 (Testing Conjugate Gradient)
$\boldsymbol{A} = \text{tridiag}_m(a, d, a) \otimes \boldsymbol{I}_m + \boldsymbol{I}_m \otimes \text{tridiag}_m(a, d, a) \in \mathbb{R}^{m^2, m^2}$

```
function [V,K]=cgtest(m,a,d,tol,itmax)
R=ones(m)/(m+1)^2; rho=sum(sum(R.*R)); rho0=rho; P=R;
V=zeros(m,m); T1=sparse(tridiagonal(a,d,a,m));
for k=1:itmax
    if sqrt(rho/rho0)<= tol
        K=k; return
    end
    T=T1*P+P*T1; a=rho/sum(sum(P.*T)); V=V+a*P; R=R-a*T;
    rhos=rho; rho=sum(sum(R.*R)); P=R+(rho/rhos)*P;
end
K=itmax+1;
```

Consider first the averaging matrix given by $a = 1/9$ and $d = 5/18$. Starting with $\boldsymbol{x}^{(0)} = \boldsymbol{0}$ and $tol = 10^{-8}$ we obtain the values in Table 10.6.

The convergence is quite rapid. Note that each iteration only requires $O(n)$ flops and since it appears that the number of iterations can be bounded independently of n, we solve the problem in $O(n)$ operations. This is the best we can do for a problem with n unknowns.

Consider next the Poisson problem corresponding to $a = -1$ and $d = 2$. Again starting with $\boldsymbol{x}^{(0)} = \boldsymbol{0}$ and $tol = 10^{-8}$ and using CG in the form of Algorithm 10.5 we list K, the required number of iterations, and $K/\sqrt{n}$. We obtain the values in Table 10.7.

The results show that K is much smaller than n and appears to be proportional to $\sqrt{n}$. This is the same speed as for SOR and we don't have to estimate any acceleration parameter.

We will show in Section 10.4 that the number of iterations to achieve $\|\boldsymbol{r}\|_2/\|\boldsymbol{r}\|_0 \leq tol$ is bounded by the square root of the 2-norm condition number of $\boldsymbol{T}_2$.

For the averaging problem it follows from (4.22) that the largest and smallest eigenvalue of $\boldsymbol{T}_2$ are $\lambda_{max} = \frac{5}{9} + \frac{4}{9}\cos(\pi h)$ and $\lambda_{min} = \frac{5}{9} - \frac{4}{9}\cos(\pi h)$. Thus

$$\text{cond}_2(\boldsymbol{T}_2) = \frac{\lambda_{max}}{\lambda_{min}} = \frac{5 + 4\cos(\pi h)}{5 - 4\cos(\pi h)} \leq 9.$$

Thus the condition number is independent of n and the number of iterations can be bounded independently of n.

For the Poisson problem we find

$$\text{cond}_2(\boldsymbol{T}_2) = \frac{\lambda_{max}}{\lambda_{min}} = \frac{1 + \cos(\pi h)}{1 - \cos(\pi h)} = \text{cond}_2(\boldsymbol{T}) = O(n)$$

and we solve the discrete Poisson problem in $O(n^{3/2})$ flops. Again this is the same as for the SOR method and for the fast method without the FFT. In comparison the Cholesky Algorithm requires $O(n^2)$ flops both for the averaging and the Poisson problem.

10.3 Derivation and Basic Properties

Let $\boldsymbol{A} \in \mathbb{R}^{n,n}$ be symmetric positive definite. We will use two inner products on $\mathbb{R}^n$

1. $(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^T\boldsymbol{y}$
2. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}$.

The first product is the usual inner product corresponding to the Euclidian norm, while the second product, called the $\boldsymbol{A}$-product or the energy product, is an inner product since $\boldsymbol{A}$ is symmetric positive definite.

Exercise 10.8 *Show that the $\boldsymbol{A}$-inner product is an inner product.*

We note that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = (\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}).$$

The associated norm

$$\|x\|_{\boldsymbol{A}} := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

is called the **$\boldsymbol{A}$-norm** or **energy norm** of $\boldsymbol{x}$. Two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are **orthogonal** if $(\boldsymbol{x}, \boldsymbol{y}) = 0$ and **$\boldsymbol{A}$-orthogonal** if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Suppose $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ is an initial approximation to the solution of the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and let $\boldsymbol{r}^{(0)} := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$ be the corresponding residual. We consider the **Krylov subspaces** $\mathbb{W}_k$ of $\mathbb{R}^n$ defined by $\mathbb{W}_0 = \{\boldsymbol{0}\}$ and

$$\mathbb{W}_k = \text{span}(\boldsymbol{r}^{(0)}, \boldsymbol{A}\boldsymbol{r}^{(0)}, \boldsymbol{A}^2\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{r}^{(0)}), \quad k = 1, 2, 3, \cdots.$$

The Krylov spaces are nested subspaces

$$\mathbb{W}_0 \subset \mathbb{W}_1 \subset \mathbb{W}_2 \subset \cdots \subset \mathbb{W}_n \subset \mathbb{R}^n$$

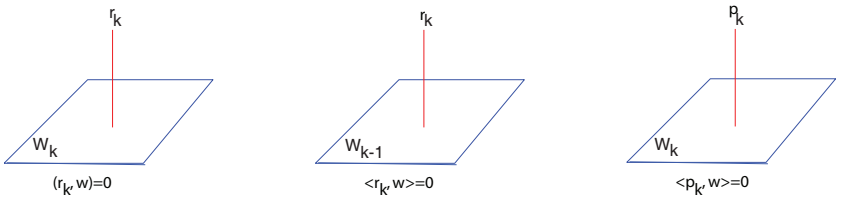


Figure 10.10. *Orthogonality in the conjugate gradient algorithm.*

with $\dim(\mathbb{W}_k) \leq k$ for all $k \geq 0$. We also note that if $\boldsymbol{w} \in \mathbb{W}_k$ then $\boldsymbol{A}\boldsymbol{w} \in \mathbb{W}_{k+1}$. This implies

$$\boldsymbol{r}^{(k-1)}, \boldsymbol{p}^{(k-1)}, \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(0)} \in \mathbb{W}_k, \quad k = 1, 2, \ldots. \tag{10.5}$$

For since $\boldsymbol{p}^{(0)} = \boldsymbol{r}^{(0)}$ and $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} = \alpha_0 \boldsymbol{p}^{(0)}$ this holds for $k = 1$ and it holds for any $k \geq 1$ by induction.

Theorem 10.9 *Suppose $\boldsymbol{r}^{(j)} \neq 0$ for $j = 0, 1, \ldots, k$. Then*

1. *$\{\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k)}\}$ is an orthogonal basis for $\mathbb{W}_{k+1}$.*
2. *$\{\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(k)}\}$ is an $\boldsymbol{A}$-orthogonal basis for $\mathbb{W}_{k+1}$.*

Proof. We show using induction on k that

1. $\{\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k)}\}$ is an orthogonal basis for $\mathbb{W}_{k+1}$.
2. $$\boldsymbol{p}^{(j)} = \boldsymbol{r}^{(j)} - \sum_{i=0}^{j-1} \frac{\langle \boldsymbol{r}^{(j)}, \boldsymbol{p}^{(i)} \rangle}{\langle \boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i)} \rangle} \boldsymbol{p}^{(i)}, \quad j \leq k. \tag{10.6}$$

Thus, $\{\boldsymbol{p}^{(0)}, \ldots, \boldsymbol{p}^{(k)}\}$ is the result of applying the Gram-Schmidt orthogonalization process to the linearly independent residuals $\{\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{r}^{(k)}\}$ using the inner product $\langle \cdot, \cdot \rangle$ (cf. Theorem A.50).

The claims hold for $k = 0$. If they hold for k then $(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(i)}) = \langle \boldsymbol{p}^{(k)}, \boldsymbol{p}^{(i)} \rangle = 0$ for $i < k$. For $j = 0, 1, \ldots, k$

$$\begin{aligned}
(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(j)}) &= (\boldsymbol{r}^{(k)} - \alpha_k \boldsymbol{A}\boldsymbol{p}^{(k)}, \boldsymbol{r}^{(j)}) \\
&= (\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(j)}) - \alpha_k \langle \boldsymbol{p}^{(k)}, \boldsymbol{p}^{(j)} - \beta_{j-1} \boldsymbol{p}^{(j-1)} \rangle \\
&= (\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(j)}) - \alpha_k \langle \boldsymbol{p}^{(k)}, \boldsymbol{p}^{(j)} \rangle.
\end{aligned}$$

This is zero for $j < k$ and vanishes for $j = k$ by the formula for α_k. Since $\boldsymbol{r}^{(j)}$ is nonzero and $\boldsymbol{r}^{(j)} \in \mathbb{W}_{j+1} \subset \mathbb{W}_{k+1}$ for $j \leq k$ Claim 1. follows.

By what we have shown $\langle \boldsymbol{r}^{(k+1)}, \boldsymbol{w}\rangle = 0$ for $\boldsymbol{w} \in \mathbb{W}_k$. But then $\langle \boldsymbol{r}^{(k+1)}, \boldsymbol{p}^{(i)}\rangle = 0$ for $i \le k-1$ since $\boldsymbol{p}^{(i)} \in \mathbb{W}_{i+1} \subset \mathbb{W}_k$ for $i \le k-1$. Now (10.6) follows for $j = k+1$ since

$$\begin{aligned}
\boldsymbol{r}^{(k+1)} - \sum_{i=0}^{k} \frac{\langle \boldsymbol{r}^{(k+1)}, \boldsymbol{p}^{(i)}\rangle}{\langle \boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i)}\rangle}\boldsymbol{p}^{(i)} &= \boldsymbol{r}^{(k+1)} - \frac{\langle \boldsymbol{r}^{(k+1)}, \boldsymbol{p}^{(k)}\rangle}{\langle \boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)}\rangle}\boldsymbol{p}^{(k)} \\
&= \boldsymbol{r}^{(k+1)} - \frac{(\boldsymbol{r}^{(k+1)}, \boldsymbol{A}\boldsymbol{p}^{(k)})}{\langle \boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)}\rangle}\boldsymbol{p}^{(k)} \\
&= \boldsymbol{r}^{(k+1)} - \frac{(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(k)} - \boldsymbol{r}^{(k+1)})}{\alpha_k \langle \boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)}\rangle}\boldsymbol{p}^{(k)} \\
&= \boldsymbol{r}^{(k+1)} + \frac{(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(k+1)})}{(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)})}\boldsymbol{p}^{(k)} \\
&= \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{p}^{(k)} = \boldsymbol{p}^{(k+1)}.
\end{aligned}$$

□

Since any $\boldsymbol{w} \in \mathbb{W}_k$ is a linear combination of $\{\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k-1)}\}$ and also $\{\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(k-1)}\}$ Theorem 10.9 implies

$$(\boldsymbol{r}^{(k)}, \boldsymbol{w}) = \langle \boldsymbol{p}^{(k)}, \boldsymbol{w}\rangle = 0, \quad \boldsymbol{w} \in \mathbb{W}_k. \tag{10.7}$$

These orthogonal properties are illustrated in Figure 10.10.

The orthogonality of the residuals implies the following best approximation property of the sequence of iterates$\{\boldsymbol{x}^{(k)}\}$ in the conjugate gradient algorithm.

Corollary 10.11 *Suppose $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is symmetric positive definite and $\{\boldsymbol{x}^{(k)}\}$ is generate by the conjugate gradient algorithm. Then $\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(0)}$ is the best approximation to $\boldsymbol{x} - \boldsymbol{x}^{(0)}$ in the $\boldsymbol{A}$-norm*

$$\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\|_{\boldsymbol{A}} = \min_{\boldsymbol{w} \in \mathbb{W}_k} \|\boldsymbol{x} - \boldsymbol{x}^{(0)} - \boldsymbol{w}\|_{\boldsymbol{A}}. \tag{10.8}$$

Proof. By (10.7)

$$0 = (\boldsymbol{r}^{(k)}, \boldsymbol{w}) = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^{(k)}, \boldsymbol{w}) = \langle \boldsymbol{x} - \boldsymbol{x}^{(k)}, \boldsymbol{w}\rangle = \langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{w}\rangle, \quad \boldsymbol{w} \in \mathbb{W}_k,$$

where $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{x}^{(0)}$ and $\boldsymbol{p} := \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(0)}$. By (10.5) it follows that $\boldsymbol{p} \in \mathbb{W}_k$. Thus, $\boldsymbol{p}$ is the $\boldsymbol{A}$-orthogonal projection of $\boldsymbol{v}$ into $\mathbb{W}_k$, and since $\boldsymbol{v} - \boldsymbol{p} = \boldsymbol{x} - \boldsymbol{x}^{(k)}$ the result follows from Theorem A.52. □

Exercise 10.12 *Consider the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ where*

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \quad \textit{and} \quad \boldsymbol{b} = \begin{bmatrix} 4 \\ 0 \\ 0 \end{bmatrix}.$$

a) *Determine the vectors defining the Krylov spaces for* $k \leq 3$ *taking as initial approximation* $\boldsymbol{x} = \boldsymbol{0}$. *Answer:* $[\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \boldsymbol{A}^2\boldsymbol{b}] = \begin{bmatrix} 4 & 8 & 20 \\ 0 & -4 & -16 \\ 0 & 0 & 4 \end{bmatrix}$.

b) *Carry out three CG-iterations on* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. *Answer:*

$$[\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}] = \begin{bmatrix} 0 & 2 & 8/3 & 3 \\ 0 & 0 & 4/3 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$[\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \boldsymbol{r}^{(3)}] = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4/3 & 0 \end{bmatrix},$$

$$[\boldsymbol{A}\boldsymbol{p}^{(0)}, \boldsymbol{A}\boldsymbol{p}^{(1)}, \boldsymbol{A}\boldsymbol{p}^{(2)}] = \begin{bmatrix} 8 & 0 & 0 \\ -4 & 3 & 0 \\ 0 & -2 & 16/9 \end{bmatrix},$$

$$[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}, \boldsymbol{p}^{(3)}] = \begin{bmatrix} 4 & 1 & 4/9 & 0 \\ 0 & 2 & 8/9 & 0 \\ 0 & 0 & 12/9 & 0 \end{bmatrix},$$

c) *Verify that*

- $dim(\mathbb{W}_k) = k$ *for* $k = 0, 1, 2, 3$.
- $\boldsymbol{x}^{(3)}$ *is the exact solution of* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.
- $\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{r}^{(k-1)}$ *is an orthogonal basis for* $\mathbb{W}_k$ *for* $k = 1, 2, 3$.
- $\boldsymbol{p}^{(0)}, \ldots, \boldsymbol{p}^{(k-1)}$ *is an* $\boldsymbol{A}$*-orthogonal basis for* $\mathbb{W}_k$ *for* $k = 1, 2, 3$.
- $\{\|\boldsymbol{r}^{(k)}\|\}$ *is monotonically decreasing.*
- $\{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|\}$ *is monotonically decreasing.*

Exercise 10.13 *Study the proof of Lemma 10.21 which shows that for the Euclidean norm*

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\|_2 \leq \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|_2, \quad k \geq 1.$$

Exercise 10.14 *Consider solving the least squares problem by using the conjugate gradient method on the normal equations* $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$. *Explain why only the following modifications in Algorithm 10.4 are necessary*

1. *r=A'(b-A*x); p=r;*
2. *a=rho/(t'*t);*
3. *r=r-a*A'*t;*

Note that the condition number of the normal equations is $cond_2(\boldsymbol{A})^2$, *the square of the condition number of* $\boldsymbol{A}$.

10.4 Convergence

The main result in this section is the following theorem.

Theorem 10.15 *Suppose we apply the conjugate gradient method to a symmetric positive definite system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Then the $\boldsymbol{A}$-norms of the errors satisfy*

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(k)}||_{\boldsymbol{A}}}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_{\boldsymbol{A}}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \quad \textit{for} \quad k \geq 0,$$

where $\kappa = cond_2(\boldsymbol{A}) = \lambda_{max}/\lambda_{min}$ is the 2-norm condition number of $\boldsymbol{A}$.

This theorem explains what we observed in the previous section. Namely that the number of iterations is linked to $\sqrt{\kappa}$, the square root of the condition number of $\boldsymbol{A}$. Indeed, the following corollary gives an upper bound for the number of iterations in terms of $\sqrt{\kappa}$.

Corollary 10.16 *If for some $\epsilon > 0$ we have $k \geq \frac{1}{2}\log(\frac{2}{\epsilon})\sqrt{\kappa}$ then $\frac{\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m A}{\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_{\boldsymbol{A}}} \leq \epsilon$.*

We prove Theorem 10.15 for $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. By Corollary 10.11 $\boldsymbol{x}^{(k)}$ is the best approximation to the solution $\boldsymbol{x}$ in the $\boldsymbol{A}$-norm. We convert this best approximation property into a best approximation problem involving polynomials. In the following we let Π_k denote the class of univariate polynomials of degree $\leq k$ with real coefficients.

Theorem 10.17 *Suppose $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is symmetric positive definite with eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding orthonormal eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$. Then*

$$||\boldsymbol{x} - \boldsymbol{x}^{(k)}||_{\boldsymbol{A}}^2 = \min_{\substack{Q \in \Pi_k \\ Q(0)=1}} \sum_{j=1}^{n} \frac{\sigma_j^2}{\lambda_j} Q(\lambda_j)^2, \tag{10.9}$$

where the σ_j's are the coefficients when $\boldsymbol{b}$ is expanded in terms of the basis of eigenvectors of $\boldsymbol{A}$, $\boldsymbol{b} = \sum_{j=1}^{n} \sigma_j \boldsymbol{u}_j$.

Proof. If $\boldsymbol{w} \in \mathbb{W}_k = \text{span}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{b})$ then for some $a_0, \ldots, a_{k-1}$

$$\boldsymbol{w} = \sum_{j=0}^{k-1} a_j \boldsymbol{A}^j \boldsymbol{b} = P(\boldsymbol{A})\boldsymbol{b},$$

where

$$P(\boldsymbol{A}) = a_0 I + a_1 \boldsymbol{A} + a_2 \boldsymbol{A}^2 + \cdots + a_{k-1}\boldsymbol{A}^{k-1}$$

is a matrix polynomial corresponding to the ordinary polynomial $P(t) = a_0 + a_1 t + \cdots + a_{k-1}t^{k-1}$ of degree $\leq k-1$. Then

$$\begin{aligned} ||\boldsymbol{x} - \boldsymbol{w}||_{\boldsymbol{A}}^2 &= \big(\boldsymbol{x} - \boldsymbol{w}, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{w})\big) \\ &= \big(\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{w}), \boldsymbol{b} - \boldsymbol{A}\boldsymbol{w}\big) \\ &= \big(\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}P(\boldsymbol{A})\boldsymbol{b}), \boldsymbol{b} - \boldsymbol{A}P(\boldsymbol{A})\boldsymbol{b}\big) \\ &= \big(\boldsymbol{A}^{-1}Q(\boldsymbol{A})\boldsymbol{b}, Q(\boldsymbol{A})\boldsymbol{b}\big), \end{aligned} \tag{10.10}$$

where $Q(\boldsymbol{A}) = I - \boldsymbol{A}P(\boldsymbol{A})$ is another matrix polynomial corresponding to the polynomial $Q(t) = 1 - tP(t)$. Observe that $Q \in \Pi_k$ and $Q(0) = 1$. Using the eigenvector expansion for $\boldsymbol{b}$ we obtain

$$Q(\boldsymbol{A})\boldsymbol{b} = \sum_{j=1}^{n} \sigma_j Q(\boldsymbol{A})\boldsymbol{u}_j = \sum_{j=1}^{n} \sigma_j Q(\lambda_j)\boldsymbol{u}_j. \tag{10.11}$$

The last equality follows from (D.3). We also have

$$\boldsymbol{A}^{-1}Q(\boldsymbol{A})\boldsymbol{u}_j = Q(\lambda_j)\boldsymbol{A}^{-1}\boldsymbol{u}_j = \frac{Q(\lambda_j)}{\lambda_j}\boldsymbol{u}_j. \tag{10.12}$$

Combining (10.10),(10.11), and (10.12) we find

$$\begin{aligned} ||\boldsymbol{x} - \boldsymbol{w}||_{\boldsymbol{A}}^2 &= \big(\boldsymbol{A}^{-1}Q(\boldsymbol{A})\boldsymbol{b}, Q(\boldsymbol{A})\boldsymbol{b}\big) \\ &= \Big(\sum_{i=1}^{n} \sigma_i \frac{Q(\lambda_i)}{\lambda_i}\boldsymbol{u}_i, \sum_{j=1}^{n} \sigma_j Q(\lambda_j)\boldsymbol{u}_j\Big) \\ &= \sum_{i,j} \sigma_i\sigma_j \frac{Q(\lambda_i)Q(\lambda_j)}{\lambda_i}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \sum_{j=1}^{n} \sigma_j^2 \frac{Q(\lambda_j)^2}{\lambda_j}. \end{aligned}$$

Minimizing over $\boldsymbol{w}$ is the same as minimizing over all $Q \in \Pi_k$ with $Q(0) = 1$ and the proof is complete. □

We will use the following weaker form of Theorem 10.17 to estimate the rate of convergence.

Corollary 10.18 *Suppose $[a, b]$ with $0 < a < b$ is an interval containing all the eigenvalues of $\boldsymbol{A}$. Then for all $Q \in \Pi_k$ with $Q(0) = 1$ we have*

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(k)}||_{\boldsymbol{A}}}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_{\boldsymbol{A}}} \leq \max_{a \leq x \leq b} |Q(x)|.$$

Proof. In the proof of Theorem 10.17 we showed that to each $\boldsymbol{w} \in \mathbb{W}_k$ there corresponds a polynomial $Q \in \Pi_k$ with $Q(0) = 1$ such that

$$||\boldsymbol{x} - \boldsymbol{w}||_{\boldsymbol{A}}^2 = \sum_{j=1}^{n} \sigma_j^2 \frac{Q(\lambda_j)^2}{\lambda_j}.$$

Taking $\boldsymbol{w} = \boldsymbol{x}^{(0)}$ we find $||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_{\boldsymbol{A}}^2 = \sum_{j=1}^{n} \frac{\sigma_j^2}{\lambda_j}$. Therefore, by Theorem 10.17 for any $\boldsymbol{w} \in \mathbb{W}_k$

$$||\boldsymbol{x} - \boldsymbol{x}^{(k)}||_{\boldsymbol{A}}^2 \leq ||\boldsymbol{x} - \boldsymbol{w}||_{\boldsymbol{A}}^2 \leq \max_{a \leq x \leq b} |Q(x)|^2 \sum_{j=1}^{n} \frac{\sigma_j^2}{\lambda_j} = \max_{a \leq x \leq b} |Q(x)|^2 ||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_{\boldsymbol{A}}^2$$

and the result follows by taking square roots. □

We will apply Corollary 10.18 with $Q(x)$ a suitably shifted and normalized version of the Chebyshev poynomial. Recall that the Chebyshev polynomial of degree n is defined recursively by

$$T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t), \quad n \geq 1$$

starting with $T_0(t) = 1$ and $T_1(t) = t$. Thus $T_2(t) = 2t^2 - 1$, $T_3(t) = 4t^3 - 3t$ etc. In general T_n is a polynomial of degree n. There are some convenient closed form expressions for T_n.

Lemma 10.19 *For $n \geq 0$*

1. $T_n(t) = \cos(n\arccos t)$ *for* $t \in [-1, 1]$,
2. $T_n(t) = \frac{1}{2}\left[\left(t + \sqrt{t^2-1}\right)^n + \left(t + \sqrt{t^2-1}\right)^{-n}\right]$ *for* $|t| \geq 1$.

Proof. 1. With $P_n(t) = \cos(n \arccos t)$ we have $P_n(t) = \cos n\phi$, where $t = \cos\phi$. Therefore

$$P_{n+1}(t) + P_{n-1}(t) = \cos(n+1)\phi + \cos(n-1)\phi = 2\cos\phi\cos n\phi = 2tP_n(t)$$

and it follows that P_n satisfies the same recurrence relation as T_n. Since $P_0 = T_0$ and $P_1 = T_1$ we have $P_n = T_n$ for all $n \geq 0$.

2. Fix t with $|t| \geq 1$ and let $x_n := T_n(t)$ for $n \geq 0$. The recurrence relation for the Chebyshev polynomials can then be written

$$x_{n+1} - 2tx_n + x_{n-1} = 0 \text{ for } n \geq 1, \text{ with } x_0 = 1, x_1 = t. \tag{10.13}$$

To find x_n we insert $x_n = z^n$ into (10.13) and obtain $z^{n+1} - 2tz^n + z^{n-1} = 0$ or $z^2 - 2tz + 1 = 0$. Let z_1 and z_2 be the roots of this quadratic equation. Then z_1^n, z_2^n and more generally $c_1 z_1^n + c_2 z_2^n$ are solutions of (10.13) for any constants c_1 and c_2. We find these constants from the initial conditions $x_0 = c_1 + c_2 = 1$ and $x_1 = c_1 z_1 + c_2 z_2 = t$. Since $z_1 + z_2 = 2t$ the solution is $c_1 = c_2 = \frac{1}{2}$. Solving the quadratic equation we find $z_1 = \alpha := t + \sqrt{t^2 - 1}$ and $z_2 = \alpha^{-1}$. It follows that $x_n = T_n(t) = \frac{1}{2}(\alpha^n + \alpha^{-n})$ which is the same as 2. □

Exercise 10.20 *Show that*

$$T_n(t) = \cosh(n\,\mathit{arccosh}\, t) \text{ for } |t| \geq 1,$$

where arccosh is the inverse function of $\cosh x := (e^x + e^{-x})/2$.

Proof of Theorem 10.15.

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\boldsymbol{A}$ and let $k \geq 0$. We apply Corollary 10.18 with $a = \min \lambda_j$, $b = \max \lambda_j$, and

$$Q(x) = T_k\left(\frac{b+a-2x}{b-a}\right) / T_k\left(\frac{b+a}{b-a}\right). \tag{10.14}$$

Note that Q is admissible since $Q \in \Pi_k$ with $Q(0) = 1$. By Lemma 10.19

$$\max_{a \le x \le b} \left| T_k\left(\frac{b+a-2x}{b-a}\right)\right| = \max_{-1 \le t \le 1} \left|T_k(t)\right| = 1. \tag{10.15}$$

Moreover with $t = (b+a)/(b-a)$ we have

$$t + \sqrt{t^2-1} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}, \quad \kappa = b/a.$$

Thus again by Lemma 10.19 we find

$$T_k\left(\frac{b+a}{b-a}\right) = \frac{1}{2}\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right] \ge \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k. \tag{10.16}$$

Using (10.15) and (10.16) in (10.14) completes the proof. □

Proof of Corollary 10.16.

Proof. The inequality

$$\frac{x-1}{x+1} < e^{-2/x} \quad \text{for} \quad x > 1 \tag{10.17}$$

follows from the familiar series expansion of the exponential function. Indeed, with $y = 1/x$ we find

$$e^{2/x} = e^{2y} = \sum_{k=0}^{\infty} \frac{(2y)^k}{k!} < 1 + 2\sum_{k=1}^{\infty} y^k = \frac{1+y}{1-y} = \frac{x+1}{x-1}$$

and (10.17) follows. By Theorem 10.15 we then find

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(k)}||_{\boldsymbol{A}}}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_{\boldsymbol{A}}} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k < 2e^{-2k/\sqrt{\kappa}}.$$

Solving the inequality $2e^{-2k/\sqrt{\kappa}} < \epsilon$ leads immediately to the result. □

The Euclidian norm of the residuals $\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}$ in the conjugate gradient iteration decreases monotonically (cf. Exercise 10.13). The following lemma shows that the Euclidian norm of the errors $\boldsymbol{x} - \boldsymbol{x}^{(k)}$ are also monotonically decreasing.

Lemma 10.21 *Let $\boldsymbol{x}$ be the exact solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, define $\boldsymbol{\epsilon}_k = \boldsymbol{x} - \boldsymbol{x}^{(k)}$ for $k \ge 0$ and let $\|\ \|$ denote the Euclidian vector norm. If $\boldsymbol{p}^{(j)} \ne 0$ for $j \le k$ then $\|\boldsymbol{\epsilon}_{k+1}\|_2 < \|\boldsymbol{\epsilon}_k\|_2$. More precisely,*

$$\|\boldsymbol{\epsilon}_{k+1}\|_2^2 = \|\boldsymbol{\epsilon}_k\|_2^2 - \frac{\|\boldsymbol{p}^{(k)}\|_2^2}{\|\boldsymbol{p}^{(k)}\|_{\boldsymbol{A}}^2}\left(\|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2 + \|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2\right). \tag{10.18}$$

Proof. Set

$$\rho_j := \|\boldsymbol{r}^{(j)}\|_2^2 \quad \text{and} \quad \pi_j := \|\boldsymbol{p}^{(j)}\|_{\boldsymbol{A}}^2,\ j \geq 0$$

and let m be the smallest integer such that $\|\boldsymbol{\epsilon}_m\|_2 = 0$. Since $\boldsymbol{p}^{(j)} \neq 0$ for $j \leq k$ we have $\dim\mathbb{W}_{k+1} = k+1$ which implies that $\boldsymbol{r}^{(k)} \neq 0$ and hence $m > k$. For $j < m$

$$\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} + \alpha_j\boldsymbol{p}^{(j)} = \boldsymbol{x}^{(j-1)} + \alpha_{j-1}\boldsymbol{p}^{(j-1)} + \alpha_j\boldsymbol{p}^{(j)} = \cdots = \boldsymbol{x}^{(0)} + \sum_{i=0}^{j}\alpha_i\boldsymbol{p}^{(i)}$$

so that

$$\boldsymbol{\epsilon}_j = \boldsymbol{x}^{(m)} - \boldsymbol{x}^{(j)} = \sum_{i=j}^{m-1}\alpha_i\boldsymbol{p}^{(i)}, \quad \alpha_i = \frac{\rho_i}{\pi_i}. \tag{10.19}$$

For $j > k$

$$(\boldsymbol{p}^{(j)}, \boldsymbol{p}^{(k)}) = (\boldsymbol{r}^{(j)} + \beta_{j-1}\boldsymbol{p}^{(j-1)}, \boldsymbol{p}^{(k)}) = \beta_{j-1}(\boldsymbol{p}^{(j-1)}, \boldsymbol{p}^{(k)}) = \cdots = \beta_{j-1}\cdots\beta_k(\boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)})$$

and since $\beta_{j-1}\cdots\beta_k = \rho_j/\rho_k$ we obtain

$$(\boldsymbol{p}^{(j)}, \boldsymbol{p}^{(k)}) = \frac{\rho_j}{\rho_k}(\boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)}), \quad j \geq k. \tag{10.20}$$

By $\boldsymbol{A}$-orthogonality and (10.19)

$$\|\boldsymbol{\epsilon}_j\|_{\boldsymbol{A}}^2 = \langle \sum_{i=j}^{m-1}\alpha_i\boldsymbol{p}^{(i)}, \sum_{i=j}^{m-1}\alpha_i\boldsymbol{p}^{(i)}\rangle = \sum_{i=j}^{m-1}\alpha_i^2\pi_i = \sum_{i=j}^{m-1}\frac{\rho_i^2}{\pi_i}. \tag{10.21}$$

Now

$$\begin{aligned}\|\boldsymbol{\epsilon}_k\|_2^2 &= \|\boldsymbol{\epsilon}_{k+1} + \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_2^2 = \|\boldsymbol{\epsilon}_{k+1} + \alpha_k\boldsymbol{p}^{(k)}\|_2^2 \\ &= \|\boldsymbol{\epsilon}_{k+1}\|_2^2 + \alpha_k\big(2\big(\boldsymbol{p}^{(k)}, \boldsymbol{\epsilon}_{k+1}\big) + \alpha_k\|\boldsymbol{p}^{(k)}\|_2^2\big).\end{aligned} \tag{10.22}$$

and moreover

$$\begin{aligned}&\alpha_k\big(2\big(\boldsymbol{p}^{(k)}, \boldsymbol{\epsilon}_{k+1}\big) + \alpha_k\|\boldsymbol{p}^{(k)}\|_2^2\big) \overset{(10.19)}{=} \alpha_k\big(2\sum_{j=k+1}^{m-1}\alpha_j\big(\boldsymbol{p}^{(j)}, \boldsymbol{p}^{(k)}\big) + \alpha_k\|\boldsymbol{p}^{(k)}\|_2^2\big) \\ &\overset{(10.20)}{=} \alpha_k\big(2\sum_{j=k+1}^{m-1}\alpha_j\frac{\rho_j}{\rho_k}\|\boldsymbol{p}^{(k)}\|_2^2 + \alpha_k\|\boldsymbol{p}^{(k)}\|_2^2\big) = \frac{\|\boldsymbol{p}^{(k)}\|_2^2}{\pi_k}\Big(\sum_{j=k}^{m-1}\frac{\rho_j^2}{\pi_j} + \sum_{j=k+1}^{m-1}\frac{\rho_j^2}{\pi_j}\Big) \\ &\overset{(10.21)}{=} \frac{\|\boldsymbol{p}^{(k)}\|_2^2}{\pi_k}\big(\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2 + \|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2\big).\end{aligned}$$

Inserting this in (10.22) proves the lemma. ∎

Chapter 11

Minimization and Preconditioning

We continue the study of the conjugate gradient method. Recall that the rate of convergence depends on the square root of the condition number of the coefficient matrix. For problems with a large condition number the convergence can be slow. For such problems a *preconditioned conjugate gradient method* is often used, and we consider this method here.

The conjugate gradient method can also be used as a minimization algorithm and we start discussing some aspects of minimization of quadratic functions.

11.1 Minimization

If $\boldsymbol{A}$ is symmetric positive definite then the quadratic function

$$Q(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}, \quad \boldsymbol{x} \in \mathbb{R}^n$$

has a unique global minimum $\boldsymbol{x}^* \in \mathbb{R}^n$ which is found by setting the gradient $\boldsymbol{g}(\boldsymbol{x}) := \nabla Q(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$ equal to zero (cf. Appendix G). So we find the minimum as a solution of the linear system $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{b}$. We see also that the gradient of $Q(\boldsymbol{x})$ is equal to the residual of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, i.e. $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{r}(\boldsymbol{x})$.

A general class of minimization algorithms for Q is given as follows:

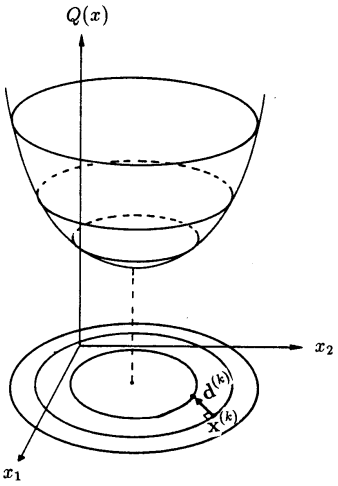
1. Choose $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.
2. For $k = 0, 1, 2, \ldots$
   (a) Choose a "search direction" $\boldsymbol{d}^{(k)}$.
   (b) Choose a "step length" σ_k.
   (c) $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \sigma_k \boldsymbol{d}^{(k)}$.

We would like to generate a sequence $\{\boldsymbol{x}^{(k)}\}$ of points such that $\{\boldsymbol{x}^{(k)}\}$ converges quickly to the minimum $\boldsymbol{x}$ of Q.

We can think of $Q(\boldsymbol{x})$ as a paraboloid. To see this, let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal, be the spectral decomposition of $\boldsymbol{A}$ and change variables to $\boldsymbol{v} = [v_1, \ldots, v_n] := \boldsymbol{U}^T\boldsymbol{x}$ and $\boldsymbol{c} := \boldsymbol{U}^T\boldsymbol{b} = [c_1, \ldots, c_n]$. Then

$$Q(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x} = \frac{1}{2}\boldsymbol{v}^T\boldsymbol{D}\boldsymbol{v} - \boldsymbol{c}^T\boldsymbol{v} = \frac{1}{2}\sum_{j=1}^{n}\lambda_j v_j^2 - \sum_{j=1}^{n} c_j v_j.$$

In particular for $n = 2$ we have $z := \frac{1}{2}\lambda_1 v_1^2 + \frac{1}{2}\lambda_2 v_2^2 - c_1 v_1 - c_2 v_2$ and since λ_1 and λ_2 are positive this is the equation for a paraboloid in (v_1, v_2, z) space as shown in the following figure.



Suppose $\boldsymbol{x}^{(k)} \approx \boldsymbol{x}^*$. To find a better approximation to the minimum we choose a search direction $\boldsymbol{d}^{(k)}$ and go from $\boldsymbol{x}^{(k)}$ along $\boldsymbol{d}^{(k)}$ a certain distance determined by σ_k. To see how σ_k and $\boldsymbol{d}^{(k)}$ should be chosen, we note that

$$Q(\boldsymbol{x}^{(k+1)}) = Q(\boldsymbol{x}^{(k)}) + \sigma_k(\boldsymbol{d}^{(k)}, \boldsymbol{r}^{(k)}) + \frac{1}{2}\sigma_k^2\langle\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)}\rangle, \tag{11.1}$$

where $\boldsymbol{r}^{(k)} = \boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}$. Since $\boldsymbol{A}$ is symmetric positive definite, we have $\sigma_k^2\langle\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)}\rangle > 0$ for all nonzero σ_k and $\boldsymbol{d}^{(k)}$. In order to make $Q(\boldsymbol{x}^{(k+1)})$ smaller than $Q(\boldsymbol{x}^{(k)})$, we must at least pick σ_k and $\boldsymbol{d}^{(k)}$ such that $\sigma_k(\boldsymbol{d}^{(k)}, \boldsymbol{r}^{(k)}) < 0$. For such a direction we can determine the step length $\sigma_k = \sigma_k^*$ such that $Q(\boldsymbol{x}^{(k+1)})$ is as small as possible, i.e.

$$Q(\boldsymbol{x}^{(k+1)}) = \min_{\sigma\in\mathbb{R}} Q(\boldsymbol{x}^{(k)} + \sigma\boldsymbol{d}^{(k)}).$$

Differentiating with respect to σ_k in (11.1) and setting the right-hand side equal to zero, we find

$$\sigma_k^* := -\frac{(\boldsymbol{d}^{(k)}, \boldsymbol{r}^{(k)})}{\langle\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)}\rangle}. \tag{11.2}$$

We find $\frac{\partial^2 Q}{\partial \sigma_k^2} = \langle \boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)} \rangle > 0 \Longrightarrow Q(\boldsymbol{x}^{(k)} + \sigma_k^* \boldsymbol{d}^{(k)}) = \min_{\sigma \in \mathbb{R}} Q(\boldsymbol{x}^{(k)} + \sigma \boldsymbol{d}^{(k)})$ and σ_k^* is called **optimal** with respect to $\boldsymbol{d}^{(k)}$.

In the method of **Steepest Descent** we choose $\boldsymbol{d}^{(k)} = -\boldsymbol{r}^{(k)}$ and $\sigma_k = \sigma_k^*$ so that

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)})}{\langle \boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k)} \rangle} \boldsymbol{r}^{(k)}, \quad k = 0, 1, 2, \ldots. \tag{11.3}$$

The method of steepest descent will converge very slowly if $\boldsymbol{A}$ is ill-conditioned. For then the ratio of the smallest and biggest eigenvalue becomes large and the paraboloid becomes very distorted. In this case the residuals need not point in the direction of the minimum. It can be shown that the number of iterations is proportional to the two-norm condition number $\lambda_{max}/\lambda_{min}$ of $\boldsymbol{A}$.

Consider now the conjugate gradient method. Here we choose $\boldsymbol{A}$-orthogonal search directions $\boldsymbol{d}^{(k)} = -\boldsymbol{p}^{(k)}$. Since by (10.1) $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}$ where $\alpha_k = (\boldsymbol{p}^{(k)}, \boldsymbol{r}^{(k)})/(\boldsymbol{p}^{(k)}, \boldsymbol{A}\boldsymbol{p}^{(k)})$, we see that the step length $-\alpha_k$ is optimal with respect to $-\boldsymbol{p}^{(k)}$. Moreover the gradients $\{\boldsymbol{r}^{(k)}\}$ are orthogonal. It can also be shown that

$$Q(\boldsymbol{x}^{(k+1)}) = \min_{w \in \mathcal{W}_{k+1}} Q(\boldsymbol{x}^{(0)} + w) \tag{11.4}$$

and in the next section we show that the number of iterations is proportional to the square root of the two-norm condition number of $\boldsymbol{A}$. So the conjugate gradient minimization algorithm converges much faster that the method of steepest descent for problems where the ratio $\lambda_{max}/\lambda_{min}$ is large.

Conjugate gradient like algorithms can be used to minimize more general functions than Q, see [15].

Exercise 11.1 *Show that* $(\boldsymbol{r}^{(k)}, \boldsymbol{r}^{(k+1)} = 0$ *in the method of steepest descent. Does this mean that all the residuals are orthogonal?*

Exercise 11.2 *Let* $Q(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}^T \boldsymbol{x}$ *have a minimum at* $\boldsymbol{x}^* \in \mathbb{R}^n$.

a) *Show that* $Q(\boldsymbol{x}) = \|\boldsymbol{x}^* - \boldsymbol{x}\|_{\boldsymbol{A}}^2 - \|\boldsymbol{x}^*\|_{\boldsymbol{A}}^2$ *for any* $\boldsymbol{x} \in \mathbb{R}^n$.

b) *Show* (11.4).

11.2 Preconditioning

For problems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ of size n where both n and $\text{cond}_2(\boldsymbol{A})$ are large it is often possible to improve the performance of the conjugate gradient method by using a technique known as **pre-conditioning**. Instead of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ we consider an equivalent system $\boldsymbol{B}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{b}$, where $\boldsymbol{B}$ is nonsingular and $\text{cond}_2(\boldsymbol{B}\boldsymbol{A})$ is smaller than $\text{cond}_2(\boldsymbol{A})$. We cannot use CG on $\boldsymbol{B}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{b}$ directly since $\boldsymbol{B}\boldsymbol{A}$ in general is not symmetric even if both $\boldsymbol{A}$ and $\boldsymbol{B}$ are. But if $\boldsymbol{B}$ is symmetric positive definite then we can apply CG to a symmetrized system and then transform the recurrence formulae to an iterative method for the original system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. This iterative method is

known as the **pre-conditioned conjugate gradient method**. We shall see that the convergence properties of this method is determined by the eigenvalues of $\boldsymbol{BA}$.

Suppose $\boldsymbol{B}$ is symmetric positive definite. By Theorem 3.33 there is a non-singular matrix $\boldsymbol{C}$ such that $\boldsymbol{B} = \boldsymbol{C}^T\boldsymbol{C}$. ($\boldsymbol{C}$ is only needed for the derivation and will never be computed). Now

$$\boldsymbol{BAx} = \boldsymbol{Bb} \Leftrightarrow \boldsymbol{C}^T(\boldsymbol{CAC}^T)\boldsymbol{C}^{-T}\boldsymbol{x} = \boldsymbol{C}^T\boldsymbol{Cb} \Leftrightarrow (\boldsymbol{CAC}^T)\boldsymbol{y} = \boldsymbol{Cb},\ \&\ \boldsymbol{x} = \boldsymbol{C}^T\boldsymbol{y}.$$

We have 3 linear systems

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{11.5}$$
$$\boldsymbol{BAx} = \boldsymbol{Bb} \tag{11.6}$$
$$(\boldsymbol{CAC}^T)\boldsymbol{y} = \boldsymbol{Cb},\ \&\ \boldsymbol{x} = \boldsymbol{C}^T\boldsymbol{y}. \tag{11.7}$$

Note that (11.5) and (11.7) are symmetric positive definite linear systems. In addition to being symmetric positive definite the matrix $\boldsymbol{CAC}^T$ is similar to $\boldsymbol{BA}$. Indeed,

$$\boldsymbol{C}^T(\boldsymbol{CAC}^T)\boldsymbol{C}^{-T} = \boldsymbol{BA}.$$

Thus $\boldsymbol{CAC}^T$ and $\boldsymbol{BA}$ have the same eigenvalues. Therefore if we apply the conjugate gradient method to (11.7) then the rate of convergence will be determined by the eigenvalues of $\boldsymbol{BA}$.

We apply the conjugate gradient method to $(\boldsymbol{CAC}^T)\boldsymbol{y} = \boldsymbol{Cb}$. Denoting the search direction by $\boldsymbol{q}^{(k)}$ and the residual by $\boldsymbol{z}^{(k)} = \boldsymbol{CAC}^T\boldsymbol{y}^{(k)} - \boldsymbol{Cb}$ we obtain the following from (10.2), (10.3), and (10.4).

$$\begin{aligned}
&\boldsymbol{y}^{(k+1)} = \boldsymbol{y}^{(k)} + \alpha_k\boldsymbol{q}^{(k)}, \quad \alpha_k = (\boldsymbol{z}^{(k)}, \boldsymbol{z}^{(k)})/(\boldsymbol{q}^{(k)}, (\boldsymbol{CAC}^T)\boldsymbol{q}^{(k)}),\\
&\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} + \alpha_k(\boldsymbol{CAC}^T)\boldsymbol{q}^{(k)},\\
&\boldsymbol{q}^{(k+1)} = \boldsymbol{z}^{(k+1)} + \beta_k\boldsymbol{q}^{(k)}, \quad \beta_k = (\boldsymbol{z}^{(k+1)}, \boldsymbol{z}^{(k+1)})/(\boldsymbol{z}^{(k)}, \boldsymbol{z}^{(k)}).
\end{aligned}$$

With

$$\boldsymbol{x}^{(k)} := \boldsymbol{C}^T\boldsymbol{y}^{(k)}, \quad \boldsymbol{p}^{(k)} := \boldsymbol{C}^T\boldsymbol{q}^{(k)}, \quad \boldsymbol{s}^{(k)} := \boldsymbol{C}^T\boldsymbol{z}^{(k)}, \quad \boldsymbol{r}^{(k)} := \boldsymbol{C}^{-1}\boldsymbol{z}^{(k)} \tag{11.8}$$

this can be transformed into

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{p}^{(k)}, \quad \alpha_k = (\boldsymbol{s}^{(k)}, \boldsymbol{r}^{(k)})/\langle\boldsymbol{p}^{(k)}, \boldsymbol{p}^{(k)}\rangle, \tag{11.9}$$
$$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} + \alpha_k\boldsymbol{Ap}^{(k)}, \tag{11.10}$$
$$\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} + \alpha_k\boldsymbol{BAp}^{(k)}, \tag{11.11}$$
$$\boldsymbol{p}^{(k+1)} = \boldsymbol{s}^{(k+1)} + \beta_k\boldsymbol{p}^{(k)}, \quad \beta_k = (\boldsymbol{s}^{(k+1)}, \boldsymbol{r}^{(k+1)})/(\boldsymbol{s}^{(k)}, \boldsymbol{r}^{(k)}). \tag{11.12}$$

Here $\boldsymbol{x}^{(k)}$ will be an approximation to the solution $\boldsymbol{x}$ of $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{r}^{(k)} = \boldsymbol{Ax}^{(k)} - \boldsymbol{b}$ is the residual in the original system and $\boldsymbol{s}^{(k)} = \boldsymbol{BAx}^{(k)} - \boldsymbol{Bb}$ is the residual in the preconditioned system. This follows since by (11.8)

$$\boldsymbol{r}^{(k)} = \boldsymbol{C}^{-1}\boldsymbol{z}^{(k)} = \boldsymbol{C}^{-1}\boldsymbol{CAC}^T\boldsymbol{y}^{(k)} - \boldsymbol{b} = \boldsymbol{Ax}^{(k)} - \boldsymbol{b}$$

and $\boldsymbol{s}^{(k)} = \boldsymbol{C}^T\boldsymbol{z}^{(k)} = \boldsymbol{C}^T\boldsymbol{C}\boldsymbol{r}^{(k)} = \boldsymbol{B}\boldsymbol{r}^{(k)}$. We now have the following preconditioned conjugate gradient algorithm for obtaining an approximation $\boldsymbol{x}^{(k)}$ to the solution of a symmetric positive definite system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

Algorithm 11.3 (Preconditioned Conjugate Gradient Algorithm)

1. *Choose a starting vector* $\boldsymbol{x}^{(0)}$ *(for example* $\boldsymbol{x}^{(0)} = 0$*)*
2. $\boldsymbol{r}_0 = \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{b},\ \boldsymbol{p}_0 = \boldsymbol{s}_0 = \boldsymbol{B}\boldsymbol{r}_0$
3. $\rho_0 = (\boldsymbol{s}_0, \boldsymbol{r}_0);\ k = 0$
4. *while* $\sqrt{\rho_k/\rho_0} > \epsilon \quad \& \quad k < kmax$
   - 4.1*a* $\boldsymbol{t}_k = \boldsymbol{A}\boldsymbol{p}^{(k)}$
   - 4.1*b* $\boldsymbol{w}_k = \boldsymbol{B}\boldsymbol{t}_k$
   - 4.2 $\alpha_k = \rho_k/(\boldsymbol{p}^{(k)}, \boldsymbol{t}_k)$
   - 4.3 $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{p}^{(k)}$
   - 4.4*a* $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} + \alpha_k\boldsymbol{t}_k \quad (\boldsymbol{r}^{(k)} = \boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})$
   - 4.4*b* $\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} + \alpha_k\boldsymbol{w}_k \quad (\boldsymbol{s}^{(k)} = \boldsymbol{B}\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{B}\boldsymbol{b})$
   - 4.5 $\rho_{k+1} = (\boldsymbol{s}^{(k+1)}, \boldsymbol{r}^{(k+1)})$
   - 4.6 $\boldsymbol{p}^{(k+1)} = \boldsymbol{s}^{(k+1)} + \dfrac{\rho_{k+1}}{\rho_k}\boldsymbol{p}^{(k)}$
   - 4.7 $k = k + 1$

This algorithm is quite similar to Algorithm 10.4. The main additional work is contained in statement 4.1b. We'll discuss this further in connection with an example.

We have the following convergence result for this algorithm.

Theorem 11.4 *Suppose we apply a symmetric positive definite preconditioner* $\boldsymbol{B}$ *to the symmetric positive definite system* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$*. Then the quantities* $\boldsymbol{x}^{(k)}$ *computed in Algorithm 11.3 satisfy the following bound:*

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(k)}||_{\boldsymbol{A}}}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_{\boldsymbol{A}}} \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k, \quad \textit{for} \quad k \geq 0,$$

where $\kappa = \lambda_{max}/\lambda_{min}$ *is the ratio of the largest and smallest eigenvalue of* $\boldsymbol{B}\boldsymbol{A}$*.*

Proof. Since Algorithm 11.3 is equivalent to solving (11.7) by the conjugate gradient method Theorem 10.15 implies that

$$\frac{||\boldsymbol{y} - \boldsymbol{y}^{(k)}||_{\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T}}{||\boldsymbol{y} - \boldsymbol{y}_0||_{\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T}} \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k, \quad \text{for} \quad k \geq 0,$$

where $\boldsymbol{y}^{(k)}$ is the conjugate gradient approximation to the solution $\boldsymbol{y}$ of (11.7) and κ is the ratio of the largest and smallest eigenvalue of $\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T$. Since $\boldsymbol{B}\boldsymbol{A}$ and $\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T$

are similar this is the same as the κ in the theorem. By (11.8) we have

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{y}^{(k)}\|^2_{\boldsymbol{CAC}^T} &= (\boldsymbol{y}-\boldsymbol{y}^{(k)}, \boldsymbol{CAC}^T(\boldsymbol{y}-\boldsymbol{y}^{(k)})) \\ &= (\boldsymbol{C}^T(\boldsymbol{y}-\boldsymbol{y}^{(k)}), \boldsymbol{AC}^T(\boldsymbol{y}-\boldsymbol{y}^{(k)})) = \|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|^2_{\boldsymbol{A}}\end{aligned}$$

and the proof is complete. □

We conclude that $\boldsymbol{B}$ should satisfy the following requirements for a problem of size n:

1. The eigenvalues of $\boldsymbol{BA}$ should be located in a narrow interval. Preferably one should be able to bound the length of the interval independently of n.
2. The evaluation of $\boldsymbol{Bx}$ for a given vector $\boldsymbol{x}$ should not be expensive in storage and flops, ideally $O(n)$ for both.

11.3 Preconditioning Example

Throughout this section we use the same grid and notation as in Section 2.4. Let $h = 1/(m+1)$.

We recall the Poisson problem

$$-\nabla^2 u = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x,y) \quad \text{for} \quad (x,y) \in \Omega = (0,1)^2 \tag{11.13}$$

$$u = 0 \text{ on } \partial\Omega,$$

where f is a given function, Ω is the unit square in the plane, and $\partial\Omega$ is the boundary of Ω. For numerical solution we have the **discrete Poisson problem** which can either be written as a matrix equation

$$\begin{aligned} h^2 f_{j,k} &= 4v_{j,k} - v_{j-1,k} - v_{j+1,k} - v_{j,k-1} - v_{j,k+1}, \quad j,k = 1,\ldots,m \\ v_{0,k} &= v_{m+1,k} = v_{j,0} = v_{j,m+1} = 0, \quad j,k = 0,1,\ldots,m+1, \end{aligned}$$

or as a system $\boldsymbol{A}_p\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{x} = \text{vec}(v_{i,j})$, $\boldsymbol{b} = h^2\text{vec}(f_{i,j})$ and the elements $a_{i,j}$ of $\boldsymbol{A}_p$ are given by

$$\begin{aligned} a_{ii} &= 4, && i = 1,\ldots,n \\ a_{i+1,i} = a_{i,i+1} &= -1, && i = 1,\ldots,n-1, \quad i \neq m, 2m, \ldots, (m-1)m \\ a_{i+m,i} = a_{i,i+m} &= -1, && i = 1,\ldots,n-m \\ a_{ij} &= 0, && \text{otherwise.} \end{aligned}$$

11.3.1 A Banded Matrix

Consider the problem

$$\begin{aligned} -\tfrac{\partial}{\partial x}\big(c(x,y)\tfrac{\partial u}{\partial x}\big) - \tfrac{\partial}{\partial y}\big(c(x,y)\tfrac{\partial u}{\partial y}\big) &= f(x,y) && (x,y) \in \Omega = (0,1)^2 \\ u(x,y) &= 0 && (x,y) \in \partial\Omega. \end{aligned} \tag{11.14}$$

Here Ω is the open unit square while $\partial\Omega$ is the boundary of Ω. The functions f and c are given and we seek a function $u = u(x,y)$ such that (11.14) holds. We assume that c and f are defined and continuous on Ω and that $c(x,y) > 0$ for all $(x,y) \in \Omega$. The problem (11.14) reduces to the Poisson problem in the special case where $c(x,y) = 1$ for $(x,y) \in \Omega$.

As for the Poisson problem we solve (11.14) numerically on a grid of points

$$\{(jh, kh) : \ j,k = 0, 1, \ldots, m+1\}, \quad \text{where} \quad h = 1/(m+1),$$

and where m is a positive integer. Let (x,y) be one of the interior grid points. For univariate functions f, g we use the central difference approximations

$$\begin{aligned}\frac{\partial}{\partial t}\Big(f(t)\frac{\partial}{\partial t}g(t)\Big) &\approx \Big(f(t+\frac{h}{2})\frac{\partial}{\partial t}g(t+h/2) - f(t-\frac{h}{2})\frac{\partial}{\partial t}g(t-\frac{h}{2})\Big)/h \\ &\approx \Big(f(t+\frac{h}{2})\big(g(t+h)-g(t)\big) - f(t-\frac{h}{2})\big(g(t)-g(t-h)\big)\Big)/h^2\end{aligned}$$

to obtain

$$\frac{\partial}{\partial x}\big(c\frac{\partial u}{\partial x}\big)_{j,k} \approx \frac{c_{j+\frac{1}{2},k}(v_{j+1,k}-v_{j,k}) - c_{j-\frac{1}{2},k}(v_{j,k}-v_{j-1,k})}{h^2}$$

and

$$\frac{\partial}{\partial y}\big(c\frac{\partial u}{\partial y}\big)_{j,k} \approx \frac{c_{j,k+\frac{1}{2}}(v_{j,k+1}-v_{j,k}) - c_{j,k-\frac{1}{2}}(v_{j,k}-v_{j,k-1})}{h^2},$$

where $c_{p,q} = c(ph, qh)$ and $v_{j,k} \approx u(jh, kh)$. With these approximations the discrete analog of (11.14) turns out to be

$$\begin{array}{rcll} -(\boldsymbol{P}_h v)_{j,k} &=& h^2 f_{j,k} & j,k = 1,\ldots,m \\ v_{j,k} &=& 0 & j = 0, m+1 \text{ all } k \text{ or } k = 0, m+1 \text{ all j,} \end{array} \tag{11.15}$$

where

$$\begin{aligned} -(\boldsymbol{P}_h v)_{j,k} &= (c_{j,k-\frac{1}{2}} + c_{j-\frac{1}{2},k} + c_{j+\frac{1}{2},k} + c_{j,k+\frac{1}{2}})v_{j,k} \\ &\quad - c_{j,k-\frac{1}{2}}v_{j,k-1} - c_{j-\frac{1}{2},k}v_{j-1,k} - c_{j+\frac{1}{2},k}v_{j+1,k} - c_{j,k+\frac{1}{2}}v_{j,k+1} \end{aligned} \tag{11.16}$$

and $f_{j,k} = f(jh, kh)$.

As before we let $\boldsymbol{V} = (v_{j,k}) \in \mathbb{R}^{m,m}$ and $\boldsymbol{F} = (f_{j,k}) \in \mathbb{R}^{m,m}$. The corresponding linear system can be written $\boldsymbol{Ax} = \boldsymbol{b}$ where $\boldsymbol{x} = \text{vec}(\boldsymbol{V})$, $\boldsymbol{b} = h^2\text{vec}(\boldsymbol{F})$, and the n-by-n coefficient matrix $\boldsymbol{A}$ is given by

$$\begin{array}{rcll} a_{i,i} &=& c_{j_i,k_i-\frac{1}{2}} + c_{j_i-\frac{1}{2},k_i} + c_{j_i+\frac{1}{2},k_i} + c_{j_i,k_i+\frac{1}{2}}, & i = 1,2,\ldots,n \\ a_{i+1,i} &=& a_{i,i+1} = -c_{j_i+\frac{1}{2},k_i}, & i \bmod m \neq 0 \\ a_{i+m,i} &=& a_{i,i+m} = -c_{j_i,k_i+\frac{1}{2}}, & i = 1,2,\ldots,n-m \\ a_{i,j} &=& 0 & \text{otherwise,} \end{array} \tag{11.17}$$

where (j_i, k_i) with $1 \le j_i, k_i \le m$ is determined uniquely from the equation $i = j_i + (k_i - 1)m$ for $i = 1,\ldots,n$. When $c(x,y) = 1$ for all $(x,y) \in \Omega$ then we recover the Poisson matrix.

In general we cannot write $\boldsymbol{A}$ as a matrix equation of the form (4.15). But we can show that $\boldsymbol{A}$ is symmetric and it is positive definite as long as the function c is positive on Ω. Recall that a matrix $\boldsymbol{A}$ is positive definite if $x^T\boldsymbol{A}\boldsymbol{x} > 0$ for all $x \neq 0$.

Theorem 11.1 *If $c(x,y) > 0$ for $(x,y) \in \Omega$ then the matrix $\boldsymbol{A}$ given by (11.17) is symmetric positive definite.*

Proof.

To each $x \in \mathbb{R}^n$ there corresponds a matrix $\boldsymbol{V} \in \mathbb{R}^{m,m}$ such that $x = \text{vec}(\boldsymbol{V})$. We claim that

$$x^T\boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^{m}\sum_{k=0}^{m} c_{j,k+\frac{1}{2}}\big(v_{j,k+1} - v_{j,k}\big)^2 + \sum_{k=1}^{m}\sum_{j=0}^{m} c_{j+\frac{1}{2},k}\big(v_{j+1,k} - v_{j,k}\big)^2, \quad (11.18)$$

where $v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1} = 0$ for $j,k = 0,1,\ldots,m+1$. Since $c_{j+\frac{1}{2},k}$ and $c_{j,k+\frac{1}{2}}$ correspond to values of c in Ω for the values of j,k in the sums it follows that they are positive and from (11.18) we see that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $x \in \mathbb{R}^n$. Moreover if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0$ then all quadratic factors are zero and $v_{j,k+1} = v_{j,k}$ for $k = 0,1,\ldots,m$ and $j = 1,\ldots,m$. Now $v_{j,0} = v_{j,m+1} = 0$ implies that $\boldsymbol{V} = \boldsymbol{0}$ and hence $x = 0$. Thus $\boldsymbol{A}$ is symmetric positive definite.

It remains to prove (11.18). From the connection between (11.16) and (11.17) we have

$$\begin{aligned}
\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= \sum_{j=1}^{m}\sum_{k=1}^{m} -(\boldsymbol{P}_h v)_{j,k} v_{j,k} \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m} \Big(c_{j,k-\frac{1}{2}} v_{j,k}^2 + c_{j-\frac{1}{2},k} v_{j,k}^2 + c_{j+\frac{1}{2},k} v_{j,k}^2 + c_{j,k+\frac{1}{2}} v_{j,k}^2 \\
&\qquad - c_{j,k-\frac{1}{2}} v_{j,k-1} v_{j,k} - c_{j,k+\frac{1}{2}} v_{j,k} v_{j,k+1} \\
&\qquad - c_{j-\frac{1}{2},k} v_{j-1,k} v_{j,k} - c_{j+\frac{1}{2},k} v_{j,k} v_{j+1,k} \Big).
\end{aligned}$$

Using the homogenous boundary conditions we have

$$\begin{aligned}
\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j,k-\frac{1}{2}} v_{j,k}^2 &= \sum_{j=1}^{m}\sum_{k=0}^{m} c_{j,k+\frac{1}{2}} v_{j,k+1}^2, \\
\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j,k-\frac{1}{2}} v_{j,k-1} v_{j,k} &= \sum_{j=1}^{m}\sum_{k=0}^{m} c_{j,k+\frac{1}{2}} v_{j,k+1} v_{j,k}, \\
\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j-\frac{1}{2},k} v_{j,k}^2 &= \sum_{k=1}^{m}\sum_{j=0}^{m} c_{j+\frac{1}{2},k} v_{j+1,k}^2, \\
\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j-\frac{1}{2},k} v_{j-,k} v_{j,k} &= \sum_{k=1}^{m}\sum_{j=0}^{m} c_{j+\frac{1}{2},k} v_{j+1,k} v_{j,k}.
\end{aligned}$$

n	2500	10000	22500	40000	62500
K	222	472	728	986	1246
$K/\sqrt{n}$	4.44	4.72	4.85	4.93	4.98
K_{pre}	22	23	23	23	23

Table 11.2. *The number of iterations K (no preconditioning) and K_{pre} (with preconditioning) for the problem (11.14) using the discrete Poisson problem as a preconditioner.*

It follows that

$$\begin{aligned}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = &\sum_{j=1}^{m}\sum_{k=0}^{m} c_{j,k+\frac{1}{2}}\left(v_{j,k}^2 + v_{j,k+1}^2 - 2v_{j,k}v_{j,k+1}\right)\\ &+ \sum_{k=1}^{m}\sum_{j=0}^{m} c_{j+\frac{1}{2},k}\left(v_{j,k}^2 + v_{j+1,k}^2 - 2v_{j,k}v_{j+1,k}\right)\end{aligned}$$

and (11.18) follows. □ □

11.3.2 Preconditioning

Consider solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is given by (11.17) and $b \in \mathbb{R}^n$. Since $\boldsymbol{A}$ is positive definite it is nonsingular and the system has a unique solution $x \in \mathbb{R}^n$. Moreover we can use either Cholesky factorization or the block tridiagonal solver to find $\boldsymbol{x}$. Since the bandwidth of $\boldsymbol{A}$ is $m = \sqrt{n}$ both of these methods require $O(n^2)$ flops for large n.

If we choose $c(x, y) \equiv 1$ in (11.14), we get the Poisson problem (11.13). With this in mind, we may think of the coefficient matrix $\boldsymbol{A}_p$ arising from the discretization of the Poisson problem as an approximation to the matrix (11.17). This suggests using $\boldsymbol{B} = \boldsymbol{A}_p^{-1}$, the inverse of the discrete Poisson matrix as a preconditioner for the system (11.15).

Consider Algorithm 11.3. With this preconditioner Statement 4.1b can be written $\boldsymbol{A}_p\boldsymbol{w}_k = \boldsymbol{t}_k$.

In Section 5.2 we developed a Simple fast Poisson Solver, Cf. Algorithm 5.1. This method can be utilized to solve $\boldsymbol{A}_p\boldsymbol{w}_k = \boldsymbol{t}_k$.

Consider the specific problem where

$$c(x, y) = e^{-x+y} \text{ and } f(x, y) = 1.$$

We have used Algorithm 10.4 (conjugate gradient without preconditioning), and Algorithm 11.3 (conjugate gradient with preconditioning) to solve the problem (11.14). We used $\boldsymbol{x}^{(0)} = 0$ and $\epsilon = 10^{-8}$. The results are shown in Table 11.2.

Without preconditioning the number of iterations still seems to be more or less proportional to $\sqrt{n}$ although the convergence is slower than for the constant

coefficient problem. Using preconditioning speeds up the convergence considerably. The number of iterations appears to be bounded independently of n. This illustrates that preconditioning is needed when solving nontrivial problems.

Using a preconditioner increases the work in each iteration. For the present example the number of flops in each iteration changes from $O(n)$ without preconditioning to $O(n^{3/2})$ or $O(n \log_2 n)$ with preconditioning. This is not a large increase and both the number of iterations and the computing time is reduced drastically.

Let us finally show that the number $\kappa = \lambda_{max}/\lambda_{min}$ which determines the rate of convergence for the preconditioned conjugate gradient method applied to (11.14) can be bounded independently of n.

Theorem 11.3 *Suppose $0 < c_0 \le c(x,y) \le c_1$ for all $(x,y) \in [0,1]^2$. For the eigenvalues of the matrix $\boldsymbol{BA} = \boldsymbol{A}_p^{-1}\boldsymbol{A}$ just described we have*

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \le \frac{c_1}{c_0}.$$

Proof.

Suppose $\boldsymbol{A}_p^{-1}\boldsymbol{A}\boldsymbol{x} = \lambda x$ for some $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$. Then $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{A}_p x$. Multiplying this by $\boldsymbol{x}^T$ and solving for λ we find

$$\lambda = \frac{x^T \boldsymbol{A}\boldsymbol{x}}{x^T \boldsymbol{A}_p x}.$$

We computed $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ in (11.18) and we obtain $\boldsymbol{x}^T\boldsymbol{A}_p\boldsymbol{x}$ by setting all the c's there equal to one

$$\boldsymbol{x}^T\boldsymbol{A}_p x = \sum_{i=1}^{m}\sum_{j=0}^{m}\left(v_{i,j+1} - v_{i,j}\right)^2 + \sum_{j=1}^{m}\sum_{i=0}^{m}\left(v_{i+1,j} - v_{i,j}\right)^2.$$

Thus $\boldsymbol{x}^T\boldsymbol{A}_p x > 0$ and bounding all the c's in (11.18) from below by c_0 and above by c_1 we find

$$c_0(x^T\boldsymbol{A}_p x) \le x^T\boldsymbol{A}\boldsymbol{x} \le c_1(x^T\boldsymbol{A}_p x)$$

which implies that $c_0 \le \lambda \le c_1$ for all eigenvalues λ of $\boldsymbol{BA} = \boldsymbol{A}_p^{-1}\boldsymbol{A}$. □

Using $c(x,y) = e^{-x+y}$ as above, we find $c_0 = e^{-2}$ and $c_1 = 1$. Thus $\kappa \le e^2 \approx 7.4$, a quite acceptable matrix condition which explains the convergence results from our numerical experiment.

Part IV

Orthonormal Transformations and Least Squares

Chapter 12

Orthonormal Transformations

Transformations by elementary lower triangular matrices is used in Gaussian elimination to reduce a matrix to triangular form (cf. Appendix E). Elementary lower triangular matrices are not the only kind of transformations which can be used for such a task. In this chapter we study how transformations by orthogonal matrices can be used to reduce a rectangular matrix to upper triangular (also called upper trapezoidal) form. This lead to a decomposition of the matrix known as a QR decomposition and a compact form which we refer to as a QR factorization. Orthonormal transformations have the advantage that they preserve the Euclidian norm of a vector, and the spectral norm and Frobenius norm of a matrix. Indeed, if $\boldsymbol{Q} \in \mathbb{R}^{m,m}$ is an orthogonal matrix then $\|\boldsymbol{Q}\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$, $\|\boldsymbol{Q}\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_2$, and $\|\boldsymbol{Q}\boldsymbol{A}\|_F = \|\boldsymbol{A}\|_F$ for any vector $\boldsymbol{v} \in \mathbb{R}^m$ and any matrix $\boldsymbol{A} \in \mathbb{R}^{m,n}$, (cf. Lemma 8.10 and Theorem 8.28). This means that when an orthogonal transformation is applied to an inaccurate vector or matrix then the error will not grow. Thus in general an orthogonal transformation is numerically stable. The QR factorization can be used to solve least squares problems and linear equations. We consider linear equations in this chapter and least squares problems in Chapter 13.

12.1 The QR Decomposition and QR Factorization.

Definition 12.1 *Let $\boldsymbol{A} \in \mathbb{C}^{m,n}$ with $m \geq n \geq 1$. We say that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is a* **QR decomposition** *of $\boldsymbol{A}$ if $\boldsymbol{Q} \in \mathbb{C}^{m,m}$ is square and unitary and*

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0}_{m-n,n} \end{bmatrix}$$

where $\boldsymbol{R}_1 \in \mathbb{C}^{n,n}$ is upper triangular and $\boldsymbol{0}_{m-n,n} \in \mathbb{C}^{m-n,n}$ is the zero matrix. We call $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ a **QR factorization** *of $\boldsymbol{A}$ if $\boldsymbol{Q} \in \mathbb{C}^{m,n}$ has orthonormal columns and $\boldsymbol{R} \in \mathbb{C}^{n,n}$ is upper triangular.*

A QR factorization is obtained from a QR decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ by simply using the first n columns of $\boldsymbol{Q}$ and the first n rows of $\boldsymbol{R}$. Indeed, if we partition $\boldsymbol{Q}$

as $[\boldsymbol{Q}_1, \boldsymbol{Q}_2]$ and $\boldsymbol{R} = \left[\begin{smallmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{smallmatrix}\right]$, where $\boldsymbol{Q}_1 \in \mathbb{R}^{m,n}$ and $\boldsymbol{R}_1 \in \mathbb{R}^{n,n}$ then $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ is a QR factorization of $\boldsymbol{A}$. On the other hand a QR factorization $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ of $\boldsymbol{A}$ can be turned into a QR decomposition by extending the set of columns $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}$ of $\boldsymbol{Q}_1$ into an orthonormal basis $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n, \boldsymbol{q}_{n+1}, \ldots, \boldsymbol{q}_m\}$ for $\mathbb{R}^m$ and adding $m-n$ rows of zeros to $\boldsymbol{R}_1$. We then obtain the QR decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m]$ and $\boldsymbol{R} = \left[\begin{smallmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{smallmatrix}\right]$.

Example 12.2 *An example of a QR decomposition is*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 7 \\ 1 & -1 & -4 \\ 1 & -1 & 2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \times \begin{bmatrix} 2 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix} = QR,$$

while a QR factorization $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ is obtained by dropping the last column of $\boldsymbol{Q}$ and the last row of $\boldsymbol{R}$ so that

$$\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} 2 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix} = \boldsymbol{Q}_1\boldsymbol{R}_1$$

Consider existence and uniqueness.

Theorem 12.3 *Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ with $m \geq n \geq 1$. Then $\boldsymbol{A}$ has a QR decomposition and a QR factorization. The QR factorization is unique if $\boldsymbol{A}$ has linearly independent columns and $\boldsymbol{R}$ has positive diagonal elements.*

Proof. The general case can be proved using Householder transformations, see Theorem 12.19. Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has linearly independent columns. By Corollary 3.24 the matrix $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric positive definite, and by Theorem 3.32 it has a Cholesky factorization $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{R}$, where $\boldsymbol{R} \in \mathbb{R}^{n,n}$ is upper triangular and nonsingular. The matrix $\boldsymbol{Q} := \boldsymbol{A}\boldsymbol{R}^{-1}$ has orthonormal columns since $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{R}^{-T}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{R}^{-1} = \boldsymbol{R}^{-T}\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{R}^{-1} = \boldsymbol{I}$. But then $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is a QR factorization of $\boldsymbol{A}$. This shows existence. For uniqueness, if $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is a QR factorization of $\boldsymbol{A}$ and $\boldsymbol{R}$ has positive diagonal elements then $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^T\boldsymbol{R}$ is the Cholesky factorization of $\boldsymbol{A}^T\boldsymbol{A}$. Since the Cholesky factorization is unique it follows that $\boldsymbol{R}$ is unique and hence $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{R}^{-1}$ is unique. □

The QR factorization can be used to prove a classical determinant inequality.

Theorem 12.4 (Hadamard's Inequality) *For any $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{C}^{n,n}$ we have*

$$|\det(\boldsymbol{A})| \leq \prod_{j=1}^{n} \|\boldsymbol{a}_j\|_2. \tag{12.1}$$

Equality holds if and only if $\boldsymbol{A}$ has a zero column or the columns of $\boldsymbol{A}$ are orthogonal.

Proof. Let $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ be a QR factorization of $\boldsymbol{A}$. Since

$$1 = \det(\boldsymbol{I}) = \det(\boldsymbol{Q}^*\boldsymbol{Q}) = \det(\boldsymbol{Q}^*)\det(\boldsymbol{Q}) = |\det(\boldsymbol{Q})|^2$$

we have $|\det(\boldsymbol{Q})| = 1$. Let $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n]$. Then $(\boldsymbol{A}^*\boldsymbol{A})_{jj} = \|\boldsymbol{a}_j\|_2^2 = (\boldsymbol{R}^*\boldsymbol{R})_{jj} = \|\boldsymbol{r}_j\|_2^2$, and

$$|\det(\boldsymbol{A})| = |\det(\boldsymbol{Q}\boldsymbol{R})| = |\det(\boldsymbol{R})| = \prod_{j=1}^{n} |r_{jj}| \leq \prod_{j=1}^{n} \|\boldsymbol{r}_j\|_2 = \prod_{j=1}^{n} \|\boldsymbol{a}_j\|_2.$$

The inequality is proved. If equality holds then either $\det(\boldsymbol{A}) = 0$ and $\boldsymbol{A}$ has a zero column, or $\det(\boldsymbol{A}) \neq 0$ and $r_{jj} = \|\boldsymbol{r}_j\|_2$ for $j = 1, \ldots, n$. This happens if and only if $\boldsymbol{R}$ is diagonal. But then $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{R}^*\boldsymbol{R}$ is diagonal, which means that the columns of $\boldsymbol{A}$ are orthogonal. □

Exercise 12.5

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{Q} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 2 & 2 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Show that $\boldsymbol{Q}$ is orthogonal and that $\boldsymbol{Q}\boldsymbol{R}$ is a QR decomposition of $\boldsymbol{A}$. Find a QR factorization of $\boldsymbol{A}$.

12.1.1 QR and Gram-Schmidt

The Gram-Schmidt orthogonalization of the columns of $\boldsymbol{A}$ can be used to find the QR factorization of $\boldsymbol{A}$.

Theorem 12.6 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has rank n and define*

$$\boldsymbol{v}_1 = \boldsymbol{a}_1, \quad \boldsymbol{v}_j = \boldsymbol{a}_j - \sum_{i=1}^{j-1} \frac{\boldsymbol{a}_j^T\boldsymbol{v}_i}{\boldsymbol{v}_i^T\boldsymbol{v}_i}\boldsymbol{v}_i, \text{ for } j = 2, \ldots, n. \tag{12.2}$$

Let

$$\boldsymbol{Q}_1 := [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n], \quad \boldsymbol{q}_j = \frac{\boldsymbol{v}_j}{\|\boldsymbol{v}_j\|_2}, \quad j = 1, \ldots, n,$$

$$\boldsymbol{R}_1 := \begin{bmatrix} \|\boldsymbol{v}_1\|_2 & \boldsymbol{a}_2^T\boldsymbol{v}_1 & \boldsymbol{a}_3^T\boldsymbol{v}_1 & \cdots & \boldsymbol{a}_{n-1}^T\boldsymbol{v}_1 & \boldsymbol{a}_n^T\boldsymbol{v}_1 \\ 0 & \|\boldsymbol{v}_1\|_2 & \boldsymbol{a}_3^T\boldsymbol{v}_2 & \cdots & \boldsymbol{a}_{n-1}^T\boldsymbol{v}_2 & \boldsymbol{a}_n^T\boldsymbol{v}_2 \\ & & \ddots & & & \vdots \\ & & & & \|\boldsymbol{v}_{n-1}\|_2 & \boldsymbol{a}_n^T\boldsymbol{v}_{n-1} \\ & & & & & \|\boldsymbol{v}_n\|_2 \end{bmatrix}. \tag{12.3}$$

Then $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ is the unique QR factorization of $\boldsymbol{A}$.

Proof. $\boldsymbol{Q}_1$ is well defined and with orthonormal columns since by Theorem A.50 $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}$ is an orthonormal basis for span($\boldsymbol{A}$). That $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ follows by writing (12.2) in the form

$$\boldsymbol{a}_1 = \|\boldsymbol{v}_1\|_2\boldsymbol{q}_1, \quad \boldsymbol{a}_j = \sum_{i=1}^{j-1}(\boldsymbol{a}_j^T\boldsymbol{v}_i)\boldsymbol{q}_i + \|\boldsymbol{v}_j\|_2\boldsymbol{q}_j, \quad j = 2, \ldots, n.$$

Clearly $\boldsymbol{R}_1$ has positive diagonal elements and the factorization is unique. □

Exercise 12.7 *Construct $\boldsymbol{Q}_1$ and $\boldsymbol{R}_1$ in Example 12.2 using Gram-Schmidt orthogonalization.*

12.2 The Householder Transformation

The Gram-Schmidt orthogonalization process should not be used to compute the QR factorization numerically. The columns of $\boldsymbol{Q}_1$ computed in floating point arithmetic using Gram-Schmidt orthogonalization will often be far from orthogonal. There is a modified version of Gram-Schmidt which behaves better numerically, but this will not be considered here, see [2]. Instead we consider Householder transformations.

Definition 12.8 *A matrix $\boldsymbol{H} \in \mathbb{R}^{n,n}$ of the form*

$$\boldsymbol{H} := \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T, \text{ where } \boldsymbol{u} \in \mathbb{R}^n \text{ and } \boldsymbol{u}^T\boldsymbol{u} = 2$$

is called a **Householder transformation**. *The name* **elementary reflector** *is also used.*

For $n = 2$ we find $\boldsymbol{H} = \begin{bmatrix} 1-u_1^2 & -u_1u_2 \\ -u_2u_1 & 1-u_2^2 \end{bmatrix}$. A Householder transformation is symmetric and orthogonal. Indeed, $\boldsymbol{H}^T = (\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)^T = \boldsymbol{H}$ and

$$\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{H}^2 = (\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T) = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T + \boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{u})\boldsymbol{u}^T = \boldsymbol{I}.$$

There are several ways to represent a Householder transformation. Householder used $\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T$, where $\boldsymbol{u}^T\boldsymbol{u} = 1$. For any nonzero $\boldsymbol{v} \in \mathbb{R}^n$ the matrix

$$\boldsymbol{H} := \boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}} \tag{12.4}$$

is a Householder transformation. In fact $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T$, where $\boldsymbol{u} := \sqrt{2}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}$.

The main use of Householder transformations is to produce zeros in vectors. We start with

Lemma 12.9 *Suppose $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$ and $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{y} \neq \boldsymbol{0}$. Then $\left(\boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}\right)\boldsymbol{x} = \boldsymbol{y}$.*

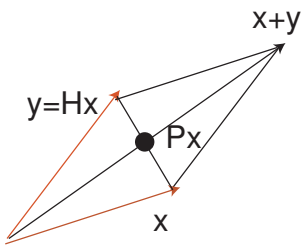


Figure 12.1. *The Householder transformation*

Proof. Since $\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{y}$ we have

$$\boldsymbol{v}^T\boldsymbol{v} = (\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}) = 2\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{y}^T\boldsymbol{x} = 2\boldsymbol{v}^T\boldsymbol{x}. \tag{12.5}$$

But then $\left(\boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}\right)\boldsymbol{x} = \boldsymbol{x} - \frac{2\boldsymbol{v}^T\boldsymbol{x}}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{v} = \boldsymbol{y}$. □

A geometric interpretation of this lemma is shown in Figure 12.1. We have

$$\boldsymbol{H} = \boldsymbol{I} - \frac{2\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}} = \boldsymbol{P} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}, \text{ where } \boldsymbol{P} := \boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}},$$

and

$$\boldsymbol{P}\boldsymbol{x} = \boldsymbol{x} - \frac{\boldsymbol{v}^T\boldsymbol{x}}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v} \overset{(12.5)}{=} \boldsymbol{x} - \frac{1}{2}\boldsymbol{v} = \frac{1}{2}(\boldsymbol{x}+\boldsymbol{y}).$$

It follows that $\boldsymbol{H}\boldsymbol{x}$ is the reflected image of $\boldsymbol{x}$. The mirror contains the vector $\boldsymbol{x}+\boldsymbol{y}$ and has normal $\boldsymbol{x}-\boldsymbol{y}$.

Exercise 12.10 *Show that* $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$ *implies that* $\boldsymbol{x}-\boldsymbol{y}$ *is orthogonal to* $\boldsymbol{x}+\boldsymbol{y}$ *and conclude that* $\boldsymbol{P}\boldsymbol{x}$ *is the orthogonal projection of* $\boldsymbol{x}$ *into the subspace* $\operatorname{span}(\boldsymbol{x}+\boldsymbol{y})$.

We can introduce zeros in components $2, 3, \ldots$ in a vector $\boldsymbol{x}$ by picking $\alpha^2 = \boldsymbol{x}^T\boldsymbol{x}$ and $\boldsymbol{y} := \alpha\boldsymbol{e}_1$ in Lemma 12.9. The equation $\alpha^2 = \boldsymbol{x}^T\boldsymbol{x}$ has two solutions $\alpha = +\|\boldsymbol{x}\|_2$ and $\alpha = -\|\boldsymbol{x}\|_2$. We want to develop an algorithm which defines a Householder transformation for any nonzero $\boldsymbol{x}$. We achieve this by choosing α to have opposite sign of x_1. Then $v_1 = x_1 - \alpha \neq 0$ so $\boldsymbol{v} \neq \boldsymbol{0}$. Another advantage of this choice is that we avoid cancelation in the subtraction in the first component of $\boldsymbol{v} = \boldsymbol{x} - \alpha\boldsymbol{e}_1$. This leads to a numerically stable algorithm.

Lemma 12.11 *For a nonzero vector* $\boldsymbol{x} \in \mathbb{R}^n$ *we define*

$$\alpha := \begin{cases} -\|\boldsymbol{x}\|_2 & \text{if } x_1 > 0 \\ +\|\boldsymbol{x}\|_2 & \text{otherwise,} \end{cases} \tag{12.6}$$

and

$$\boldsymbol{H} := \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T \text{ with } \boldsymbol{u} = \frac{\boldsymbol{x}/\alpha - \boldsymbol{e}_1}{\sqrt{1 - x_1/\alpha}}. \tag{12.7}$$

Then $\boldsymbol{H}$ *is a Householder transformation and* $\boldsymbol{H}\boldsymbol{x} = \alpha\boldsymbol{e}_1$.

Proof. Let $\boldsymbol{y} := \alpha\boldsymbol{e}_1$ and $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{y}$. If $x_1 > 0$ then $y_1 = \alpha < 0$, while if $x_1 \leq 0$ then $y_1 = \alpha > 0$. It follows that $\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{y}$ and $\boldsymbol{v} \neq \boldsymbol{0}$. By Lemma 12.9 we have $\boldsymbol{H}\boldsymbol{x} = \alpha\boldsymbol{e}_1$, where $\boldsymbol{H} = \boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}$ is a Householder transformation. Since

$$0 < \boldsymbol{v}^T\boldsymbol{v} = (\boldsymbol{x} - \alpha\boldsymbol{e}_1)^T(\boldsymbol{x} - \alpha\boldsymbol{e}_1) = \boldsymbol{x}^T\boldsymbol{x} - 2\alpha x_1 + \alpha^2 = 2\alpha(\alpha - x_1),$$

we find

$$\boldsymbol{H} = \boldsymbol{I} - \frac{2(\boldsymbol{x} - \alpha\boldsymbol{e}_1)(\boldsymbol{x} - \alpha\boldsymbol{e}_1)^T}{2\alpha(\alpha - x_1)} = \boldsymbol{I} - \frac{(\boldsymbol{x}/\alpha - \boldsymbol{e}_1)(\boldsymbol{x}/\alpha - \boldsymbol{e}_1)^T}{1 - x_1/\alpha} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T.$$

□

Example 12.12 *For* $\boldsymbol{x} := [1, 2, 2]^T$ *we have* $\|\boldsymbol{x}\|_2 = 3$ *and since* $x_1 > 0$ *we choose* $\alpha = -3$. *We find* $\boldsymbol{u} = -[2, 1, 1]^T/\sqrt{3}$ *and*

$$\boldsymbol{H} = \boldsymbol{I} - \frac{1}{3}\begin{bmatrix}2\\1\\1\end{bmatrix}\begin{bmatrix}2 & 1 & 1\end{bmatrix} = \frac{1}{3}\begin{bmatrix}-1 & -2 & -2\\-2 & 2 & -1\\-2 & -1 & 2\end{bmatrix}.$$

The formulas in Lemma 12.11 are implemented in the following algorithm from [17].

Algorithm 12.13 (Generate a Householder transformation) To given $\boldsymbol{x} \in \mathbb{R}^n$ the following algorithm computes $a = \alpha$ and the vector $\boldsymbol{u}$ so that $(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{x} = \alpha\boldsymbol{e}_1$.

```
function [u,a]=housegen(x)
a=norm(x); u=x;
if  a==0
    u(1)=sqrt(2); return;
end
if u(1)>0
    a=-a;
end
u=u/a; u(1)=u(1)-1;
u=u/sqrt(-u(1));
```

If $\boldsymbol{x} = \boldsymbol{0}$ then any $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_2 = \sqrt{2}$ can be used in the Householder transformation. In the algorithm we use $\boldsymbol{u} = \sqrt{2}\boldsymbol{e}_1$ in this case.

Exercise 12.14 *Determine* $\boldsymbol{H}$ *in Algorithm 12.13 when* $\boldsymbol{x} = \boldsymbol{e}_1$.

Householder transformations can also be used to zero out only the lower part of a vector. Suppose $\boldsymbol{y} \in \mathbb{R}^k$, $\boldsymbol{z} \in \mathbb{R}^{n-k}$ and $\alpha^2 = \boldsymbol{z}^T\boldsymbol{z}$. Consider finding a Householder transformation $\boldsymbol{H}$ such that $\boldsymbol{H}\left[\begin{smallmatrix}\boldsymbol{y}\\\boldsymbol{z}\end{smallmatrix}\right] = \left[\begin{smallmatrix}\boldsymbol{y}\\\alpha\boldsymbol{e}_1\end{smallmatrix}\right]$. Let $\hat{\boldsymbol{u}}$ and α be the

output of Algorithm 12.13 called with $\boldsymbol{x} = \boldsymbol{z}$, i. e., $[\hat{\boldsymbol{u}}, \alpha] = \text{housegen}(\boldsymbol{z})$ and set $\boldsymbol{u}^T = [\boldsymbol{0}^T, \hat{\boldsymbol{u}}^T]$. Then

$$\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} - \begin{bmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{u}} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} & \hat{\boldsymbol{u}}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{H}} \end{bmatrix},$$

where $\hat{\boldsymbol{H}} = \boldsymbol{I} - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T$. Since $\boldsymbol{u}^T\boldsymbol{u} = \hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} = 2$ we see that $\boldsymbol{H}$ and $\hat{\boldsymbol{H}}$ are Householder transformations.

Exercise 12.15 *Construct a Householder transformation $\boldsymbol{Q}$ such that $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{y}$ in the following cases.*

a) $\boldsymbol{x} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}.$

b) $\boldsymbol{x} = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix}.$

Exercise 12.16 *Show that a 2×2 Householder transformation can be written in the form*

$$\boldsymbol{Q} = \begin{bmatrix} -\cos\phi & \sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

Find $\boldsymbol{Q}\boldsymbol{x}$ if $\boldsymbol{x} = [\cos\phi, \sin\phi]^T$.

Exercise 12.17 a) *Find Householder transformations $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{3,3}$ such that*

$$\boldsymbol{Q}_2\boldsymbol{Q}_1\boldsymbol{A} = \boldsymbol{Q}_2\boldsymbol{Q}_1 \begin{bmatrix} 1 & 0 & 1 \\ -2 & -1 & 0 \\ 2 & 2 & 1 \end{bmatrix}$$

is upper triangular.

b) *Find the QR factorization of $\boldsymbol{A}$ where $\boldsymbol{R}$ has positive diagonal elements.*

12.3 Householder Triangulation

Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$. We treat the cases $m > n$ and $m \leq n$ separately and consider first $m > n$. We describe how to find a sequence $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_n$ of orthogonal matrices such that

$$\boldsymbol{H}_n\boldsymbol{H}_{n-1}\cdots\boldsymbol{H}_1\boldsymbol{A} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix},$$

and where $\boldsymbol{R}_1$ is upper triangular. Here each $\boldsymbol{H}_k$ is a Householder transformation. Since the product of orthogonal matrices is orthogonal and each $\boldsymbol{H}_k$ is symmetric we obtain the QR decomposition of $\boldsymbol{A}$ in the form

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}, \text{ where } \boldsymbol{Q} := \boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_n \text{ and } \boldsymbol{R} := \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}. \tag{12.8}$$

Define $\boldsymbol{A}_1 = \boldsymbol{A}$ and suppose for $k \geq 1$ that $\boldsymbol{A}_k$ is upper triangular in its first $k-1$ columns so that $\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{0} & \boldsymbol{D}_k \end{bmatrix}$, where $\boldsymbol{B}_k \in \mathbb{R}^{k-1,k-1}$ is upper triangular and $\boldsymbol{D}_k \in \mathbb{R}^{n-k+1,n-k+1}$. Let $\hat{\boldsymbol{H}}_k = \boldsymbol{I} - \hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^T$ be a Householder transformation which zero out the first column in $\boldsymbol{D}_k$ under the diagonal, so that $\hat{\boldsymbol{H}}_k(\boldsymbol{D}_k \boldsymbol{e}_1) = \alpha_k \boldsymbol{e}_1$. Set $\boldsymbol{H}_k := \begin{bmatrix} \boldsymbol{I}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{H}}_k \end{bmatrix}$. Then $\boldsymbol{A}_{k+1} := \boldsymbol{H}_k \boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{0} & \hat{\boldsymbol{H}}_k \boldsymbol{D}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{k+1} & \boldsymbol{C}_{k+1} \\ \boldsymbol{0} & \boldsymbol{D}_{k+1} \end{bmatrix}$, where $\boldsymbol{B}_{k+1} \in \mathbb{R}^{k,k}$ is upper triangular and $\boldsymbol{D}_{k+1} \in \mathbb{R}^{n-k,n-k}$. Thus $\boldsymbol{A}_{k+1}$ is upper triangular in its first k columns and the reduction has been carried one step further. At the end $\boldsymbol{R} := \boldsymbol{A}_{n+1} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}$, where $\boldsymbol{R}_1$ is upper triangular and $\boldsymbol{R} = \boldsymbol{H}_n \cdots \boldsymbol{H}_2 \boldsymbol{H}_1 \boldsymbol{A}$. Thus $\boldsymbol{A} = \boldsymbol{H}_1 \cdots \boldsymbol{H}_n \boldsymbol{R}$ and we obtain (12.8).

The process just described can be illustrated as follows when $m = 4$ and $n = 3$ using so called **Wilkinson diagrams**.

$$\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix} \xrightarrow{\boldsymbol{H}_1} \left[\begin{array}{c|cc} r_{11} & r_{12} & r_{13} \\ \hline 0 & x & x \\ 0 & x & x \\ 0 & x & x \end{array}\right] \xrightarrow{\boldsymbol{H}_2} \left[\begin{array}{cc|c} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ \hline 0 & 0 & x \\ 0 & 0 & x \end{array}\right] \xrightarrow{\boldsymbol{H}_3} \left[\begin{array}{ccc} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \\ \hline 0 & 0 & 0 \end{array}\right].$$

$$\boldsymbol{A}_1 = \boldsymbol{D}_1 \qquad \boldsymbol{A}_2 = \begin{bmatrix} \boldsymbol{B}_2 & \boldsymbol{C}_2 \\ \boldsymbol{0} & \boldsymbol{D}_2 \end{bmatrix} \qquad \boldsymbol{A}_3 = \begin{bmatrix} \boldsymbol{B}_3 & \boldsymbol{C}_3 \\ \boldsymbol{0} & \boldsymbol{D}_3 \end{bmatrix} \qquad \boldsymbol{A}_4 = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}$$

The transformation is applied to the lower right block.

The process can also be applied to $\boldsymbol{A} \in \mathbb{R}^{m,n}$ if $m \leq n$. In this case $m-1$ Householder transformations will suffice and we obtain

$$\boldsymbol{H}_{m-1} \cdots \boldsymbol{H}_1 \boldsymbol{A} = [\boldsymbol{R}_1, \boldsymbol{S}_1] = \boldsymbol{R}, \tag{12.9}$$

where $\boldsymbol{R}_1$ is upper triangular and $\boldsymbol{S}_1 \in \mathbb{R}^{m,n-m}$.

In an algorithm we can store most of the vectors $\hat{\boldsymbol{u}}_k = [u_{kk}, \ldots, u_{mk}]^T$ and $\boldsymbol{R}_1$ in $\boldsymbol{A}$. However, the elements u_{kk} in $\hat{\boldsymbol{u}}_k$ and r_{kk} in $\boldsymbol{R}_1$ have to compete for the diagonal in $\boldsymbol{A}$. For $m = 4$ and $n = 3$ the two possibilities look as follows:

$$\boldsymbol{A} = \begin{bmatrix} u_{11} & r_{12} & r_{13} \\ u_{21} & u_{22} & r_{23} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \end{bmatrix} \text{ or } \boldsymbol{A} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ u_{21} & r_{22} & r_{23} \\ u_{31} & u_{32} & r_{33} \\ u_{41} & u_{42} & u_{43} \end{bmatrix}.$$

Whatever alternative is chosen the loser has to be stored in a separate vector. In the following algorithm we store $r : kk$ in the diagonal of $\boldsymbol{A}$, while the vectors $\hat{\boldsymbol{u}}$ are not stored.

Algorithm 12.18 (Householder Triangulation of a matrix) Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ with $m \geq n$ and $\boldsymbol{B} \in \mathbb{R}^{m,r}$. The algorithm uses `housegen` to compute Householder transformations $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_s$, where $s = min(n, m-1)$ such that $\boldsymbol{R} = \boldsymbol{H}_s ... \boldsymbol{H}_1 \boldsymbol{A}$ is upper trapezoidal and $\boldsymbol{C} = \boldsymbol{H}_s ... \boldsymbol{H}_1 \boldsymbol{B}$. If $\boldsymbol{B} = \boldsymbol{I} \in \mathbb{R}^{m,m}$ then $\boldsymbol{C}^T \boldsymbol{R}$ is the QR decomposition of $\boldsymbol{A}$. If $\boldsymbol{B}$ is the empty matrix then $\boldsymbol{C}$ is the empty matrix with m rows and 0 columns.

```
 function [R,C] = housetriang(A,B)
[m,n]=size(A); r=size(B,2);    A=[A,B];
for k=1:min(n,m-1)
    [v,A(k,k)]=housegen(A(k:m,k));
    C=A(k:m,k+1:n+r); A(k:m,k+1:n+r)=C-v*(v'*C);
end
R=triu(A(:,1:n)); C=A(:,n+1:n+r);
```

The function `housegen(x)` returns a Householder transformation for any $\boldsymbol{x} \in R^n$. Thus in Algorithm 12.18 we obtain a QR decomposition $\boldsymbol{A} = \boldsymbol{QR}$, where $\boldsymbol{Q} = \boldsymbol{H}_1 \ldots \boldsymbol{H}_r$, is orthogonal and $r = \min\{n, m-1\}$. Thus a QR factorization always exists and we have proved

Theorem 12.19 *Any $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has a QR decomposition and a QR factorization.*

The bulk of the work in Algorithm 12.18 is the computation of $\boldsymbol{C} - \boldsymbol{v} * (\boldsymbol{v}^T * \boldsymbol{C})$ for each k. It can be determined from the following lemma.

Lemma 12.20 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^n$. The computation of $\boldsymbol{A} - \boldsymbol{u}(\boldsymbol{u}^T \boldsymbol{A})$ and $\boldsymbol{A} - (\boldsymbol{Av})\boldsymbol{v}^T$ both cost $O(4mn)$ flops.*

Proof. It costs $O(2mn)$ flops to compute $\boldsymbol{w}^T := \boldsymbol{u}^T \boldsymbol{A}$, $O(mn)$ flops to compute $\boldsymbol{W} = \boldsymbol{uw}^T$ and $O(mn)$ flops for the final subtraction $\boldsymbol{A} - \boldsymbol{W}$, a total of $O(4mn)$ flops. Taking transpose we obtain the same count for $\boldsymbol{A} - (\boldsymbol{Av})\boldsymbol{v}^T$. □

Since $\boldsymbol{C} \in \mathbb{R}^{m-k+1, n+r-k}$ and $m \geq n$ the cost of computing the update $\boldsymbol{C} - \boldsymbol{v} * (\boldsymbol{v}^T * \boldsymbol{C})$ is $4(m-k)(n+r-k)$ flops. This implies that the work in Algorithm 12.18 can be estimated as

$$\int_0^n 4(m-k)(n+r-k)dk = 2m(n+r)^2 - \frac{2}{3}(n+r)^3. \tag{12.10}$$

12.3.1 QR and Linear Systems

Algorithm 12.18 can be used to solve linear systems . If $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is nonsingular and $\boldsymbol{b} \in \mathbb{R}^n$ then the output will be the upper triangular matrix $\boldsymbol{R} = \boldsymbol{H}_{n-1} ... \boldsymbol{H}_1 \boldsymbol{A} \in \mathbb{R}^{n,n}$ and $\boldsymbol{C} = \boldsymbol{c} = \boldsymbol{H}_{n-1} ... \boldsymbol{H}_1 \boldsymbol{b} \in \mathbb{R}^n$. So

$$\boldsymbol{Ax} = \boldsymbol{b} \Rightarrow \boldsymbol{H}_{n-1} ... \boldsymbol{H}_1 \boldsymbol{A} = \boldsymbol{H}_{n-1} ... \boldsymbol{H}_1 \boldsymbol{b},$$

and the solution of $\boldsymbol{Ax} = \boldsymbol{b}$ is found by solving the upper triangular linear system $\boldsymbol{Rx} = \boldsymbol{c}$.

Algorithm 12.18 can be used as an alternative to Gaussian elimination. Recall that in Gaussian elimination we compute an LU factorization of $\boldsymbol{A}$ and then find the solution by solving two triangular systems. The two methods are similar since they both reduce $\boldsymbol{A}$ to upper triangular form using certain transformations.

Which method is better? Here is a short discussion

- Advantages with Householder:
  - Always works for nonsingular systems.
  - Row interchanges are not necessary.
  - Numerically stable.
- Advantages with Gauss
  - Half the number of flops compared to Householder.
  - Row interchanges are often not necessary.
  - Usually stable, (but no guarantee).
  - In general better than Householder for banded and sparse matrices.

Linear systems can be constructed where Gaussian elimination will fail numerically even if row interchanges are used. On the other hand the transformations used in Householder triangulation are orthogonal so the method is quite stable. So why is Gaussian elimination more popular than Householder triangulation? One reason is that the number of flops in (12.10) when $m = n$ is given by $2n(n+1)^2 - \frac{2}{3}(n+1)^3 \approx 2n^3 - \frac{2}{3}n^3 = 4n^3/3$, while the count for Gaussian elimination is half of that. Numerical stability can be a problem with Gaussian elimination, but years and years of experience shows that it works well for most practical problems and pivoting is often not necessary. Tradition might also play a role.

12.4 Givens Rotations

In some applications, the matrix we want to triangulate has a special structure. Suppose for example that $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is square and upper Hessenberg as illustrated by a Wilkinson diagram for $n = 4$

$$\boldsymbol{A} = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix}.$$

Only one element in each column needs to be annihilated and a full Householder transformation will be inefficient. In this case we can use a simpler transformation.

Definition 12.21 *A* **plane rotation** *(also called a* **Given's rotation***) is a matrix of the form*

$$\boldsymbol{P} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \text{ where } c^2 + s^2 = 1.$$

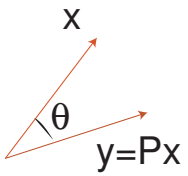


Figure 12.2. *A plane rotation.*

A plane rotation is orthogonal and there is a unique angle $\theta \in [0, 2\pi)$ such that $c = \cos\theta$ and $s = \sin\theta$. Moreover, the identity matrix is a plane rotation corresponding to $\theta = 0$.

Exercise 12.22 *Show that if* $\boldsymbol{x} = \left[\begin{smallmatrix} r\cos\alpha \\ r\sin\alpha \end{smallmatrix}\right]$ *then* $\boldsymbol{P}\boldsymbol{x} = \left[\begin{smallmatrix} r\cos(\alpha-\theta) \\ r\sin(\alpha-\theta) \end{smallmatrix}\right]$. *Thus* $\boldsymbol{P}$ *rotates a vector* $\boldsymbol{x}$ *in the plane an angle* θ *clockwise. See Figure 12.2.*

Suppose

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \neq \boldsymbol{0}, \quad c := \frac{x_1}{r}, \quad s := \frac{x_2}{r}, \quad r := \|\boldsymbol{x}\|_2.$$

Then

$$\boldsymbol{P}\boldsymbol{x} = \frac{1}{r}\begin{bmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{r}\begin{bmatrix} x_1^2 + x_2^2 \\ 0 \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix},$$

and we have introduced a zero in $\boldsymbol{x}$. We can take $\boldsymbol{P} = \boldsymbol{I}$ when $\boldsymbol{x} = \boldsymbol{0}$.

For an n-vector $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \leq i < j \leq n$ we define a **rotation in the** i,j**-plane** as a matrix $\boldsymbol{P}_{ij} = (p_{kl}) \in \mathbb{R}^{n,n}$ by $p_{kl} = \delta_{kl}$ except for positions ii, jj, ij, ji, which are given by

$$\begin{bmatrix} p_{ii} & p_{ij} \\ p_{ji} & p_{jj} \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \text{ where } c^2 + s^2 = 1.$$

Premultiplying a matrix by a rotation in the i,j plane changes only rows i and j of the matrix, while postmultiplying the matrix by such a rotation only changes column i and j. In particular, if $\boldsymbol{B} = \boldsymbol{P}_{ij}\boldsymbol{A}$ and $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{P}_{ij}$ then $\boldsymbol{B}(k,:) = \boldsymbol{A}(k,:)$, $\boldsymbol{C}(:,k) = \boldsymbol{A}(:,k)$ for all $k \neq i, j$ and

$$\begin{bmatrix} \boldsymbol{B}(i,:) \\ \boldsymbol{B}(j,:) \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}\begin{bmatrix} \boldsymbol{A}(i,:) \\ \boldsymbol{A}(j,:) \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{C}(:,i) & \boldsymbol{C}(:,j) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}(:,i) & \boldsymbol{A}(:,j) \end{bmatrix}\begin{bmatrix} c & s \\ -s & c \end{bmatrix}. \tag{12.11}$$

An upper Hessenberg matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ can be transformed to upper triangular form using rotations $\boldsymbol{P}_{i,i+1}$ for $i = 1, \ldots, n-1$. For $n = 4$ the process can be illustrated as follows.

$$\boldsymbol{A} = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} \overset{\boldsymbol{P}_{12}}{\rightarrow} \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ \boldsymbol{0} & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} \overset{\boldsymbol{P}_{23}}{\rightarrow} \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ 0 & r_{22} & r_{23} & r_{24} \\ 0 & \boldsymbol{0} & x & x \\ 0 & 0 & x & x \end{bmatrix} \overset{\boldsymbol{P}_{34}}{\rightarrow} \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ 0 & r_{22} & r_{23} & r_{24} \\ 0 & 0 & r_{33} & r_{34} \\ 0 & 0 & \boldsymbol{0} & r_{44} \end{bmatrix}.$$

For an algorithm see Exercise 12.23.

Exercise 12.23 *Let $\boldsymbol{A} \in \mathbb{R}^{n,n}$ be upper Hessenberg and nonsingular, and let $\boldsymbol{b} \in \mathbb{R}^n$. The following algorithm solves the linear system $\boldsymbol{Ax} = \boldsymbol{b}$ using rotations $\boldsymbol{P}_{k,k+1}$ for $k = 1, \ldots, n-1$. Determine the number of flops of this algorithm.*

Algorithm 12.24 (Upper Hessenberg linear system) Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is nonsingular and upper Hessenberg and that $\boldsymbol{b} \in \mathbb{R}^n$. This algorithm uses Given's rotations to solve the linear system $\boldsymbol{Ax} = \boldsymbol{b}$. It uses Algorithm E.7.

```
function x=rothesstri(A,b)
n=length(A); A=[A b];
for k=1:n-1
    r=norm([A(k,k),A(k+1,k)]);
    if r>0
        c=A(k,k)/r; s=A(k+1,k)/r;
        A([k k+1],k+1:n+1)=[c s;-s c]*A([k k+1],k+1:n+1);
    end
    A(k,k)=r; A(k+1,k)=0;
end
x=backsolve(A(:,1:n),A(:,n+1));
```

Chapter 13

Least Squares

13.1 The Pseudo-Inverse and Orthogonal Projections

13.1.1 The Pseudo-Inverse

Suppose $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$ is a singular value factorization of $\boldsymbol{A} \in \mathbb{C}^{m,n}$. The matrix $\boldsymbol{A}^\dagger \in \mathbb{C}^{n,m}$ given by

$$\boldsymbol{A}^\dagger := \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^* \tag{13.1}$$

is called the **pseudo-inverse** of $\boldsymbol{A}$. It is independent of the particular factorization used to define it. We show this in Exercises 13.1, 13.2. In terms of the singular value decomposition we have

$$\boldsymbol{A}^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{U}^*, \text{ where } \boldsymbol{\Sigma}^\dagger := \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1} & \mathbf{0}_{r,m-r} \\ \mathbf{0}_{n-r,r} & \mathbf{0}_{n-r,m-r} \end{bmatrix}.$$

If $\boldsymbol{A}$ is square and nonsingular then $\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{I}$ and $\boldsymbol{A}^\dagger$ is the usual inverse of $\boldsymbol{A}$. Any matrix has a pseudoinverse, and so $\boldsymbol{A}^\dagger$ is a generalization of the usual inverse.

Exercise 13.1 *Show that $\boldsymbol{B} := \boldsymbol{A}^\dagger$ satisfies (1) $\boldsymbol{A}\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A}$, (2) $\boldsymbol{B}\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}$, (3) $(\boldsymbol{B}\boldsymbol{A})^* = \boldsymbol{B}\boldsymbol{A}$, and (4) $(\boldsymbol{A}\boldsymbol{B})^* = \boldsymbol{A}\boldsymbol{B}$.*

Conversely, Exercise 13.2 shows that if $\boldsymbol{B} \in \mathbb{C}^{n,m}$ satisfies the four equations in Exercise 13.1 then $\boldsymbol{B} = \boldsymbol{A}^\dagger$. Thus $\boldsymbol{A}^\dagger$ is uniquely defined by these axioms and is independent of the particular singular value factorization used to define it.

Exercise 13.2 *Given $\boldsymbol{A} \in \mathbb{C}^{m,n}$, and suppose $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^{n,m}$ satisfy*

$$\begin{array}{rclcrcl} \boldsymbol{A}\boldsymbol{B}\boldsymbol{A} & = & \boldsymbol{A} & (1) & \boldsymbol{A}\boldsymbol{C}\boldsymbol{A} & = & \boldsymbol{A}, \\ \boldsymbol{B}\boldsymbol{A}\boldsymbol{B} & = & \boldsymbol{B} & (2) & \boldsymbol{C}\boldsymbol{A}\boldsymbol{C} & = & \boldsymbol{C}, \\ (\boldsymbol{A}\boldsymbol{B})^H & = & \boldsymbol{A}\boldsymbol{B} & (3) & (\boldsymbol{A}\boldsymbol{C})^H & = & \boldsymbol{A}\boldsymbol{C}, \\ (\boldsymbol{B}\boldsymbol{A})^H & = & \boldsymbol{B}\boldsymbol{A} & (4) & (\boldsymbol{C}\boldsymbol{A})^H & = & \boldsymbol{C}\boldsymbol{A}. \end{array}$$

Verify the following proof that $\boldsymbol{B} = \boldsymbol{C}$.

$$\begin{aligned}\boldsymbol{B} &= (\boldsymbol{B}\boldsymbol{A})\boldsymbol{B} = (\boldsymbol{A}^H)\boldsymbol{B}^H\boldsymbol{B} = (\boldsymbol{A}^H\boldsymbol{C}^H)\boldsymbol{A}^H\boldsymbol{B}^H\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{B}^H)\boldsymbol{B} \\ &= \boldsymbol{C}\boldsymbol{A}(\boldsymbol{B}\boldsymbol{A}\boldsymbol{B}) = (\boldsymbol{C})\boldsymbol{A}\boldsymbol{B} = \boldsymbol{C}(\boldsymbol{A}\boldsymbol{C})\boldsymbol{A}\boldsymbol{B} = \boldsymbol{C}\boldsymbol{C}^H\boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{B}) \\ &= \boldsymbol{C}\boldsymbol{C}^H(\boldsymbol{A}^H\boldsymbol{B}^H\boldsymbol{A}^H) = \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{A}^H) = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C} = \boldsymbol{C}.\end{aligned}$$

Exercise 13.3 *Show that the matrices* $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ *and* $\boldsymbol{B} = \frac{1}{4}\left[\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{smallmatrix}\right]$ *satisfy the axioms in Exercise 13.1. Thus we can conclude that* $\boldsymbol{B} = \boldsymbol{A}^\dagger$ *without computing the singular value decomposition of* $\boldsymbol{A}$.

Exercise 13.4 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *has linearly independent columns. Show that* $\boldsymbol{A}^H\boldsymbol{A}$ *is nonsingular and* $\boldsymbol{A}^\dagger = (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H$. *If* $\boldsymbol{A}$ *has linearly independent rows, then show that* $\boldsymbol{A}\boldsymbol{A}^H$ *is nonsingular and* $\boldsymbol{A}^\dagger = \boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{A}^H)^{-1}$.

Exercise 13.5 *Show that* $\boldsymbol{u}^\dagger = (\boldsymbol{u}^H\boldsymbol{u})^{-1}\boldsymbol{u}^H$ *if* $\boldsymbol{u} \in \mathbb{C}^{n,1}$ *is nonzero.*

Exercise 13.6 *If* $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^H$ *where* $\boldsymbol{u} \in \mathbb{C}^m$, $\boldsymbol{v} \in \mathbb{C}^n$ *are nonzero, show that*

$$\boldsymbol{A}^\dagger = \frac{1}{\alpha}\boldsymbol{A}^H, \quad \alpha = \|\boldsymbol{u}\|_2^2\|\boldsymbol{v}\|_2^2.$$

Exercise 13.7 *Show that* $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)^\dagger = \operatorname{diag}(\lambda_1^\dagger, \ldots, \lambda_n^\dagger)$ *where*

$$\lambda_i^\dagger = \begin{cases} 1/\lambda_i, & \lambda_i \neq 0 \\ 0 & \lambda_i = 0. \end{cases}$$

Exercise 13.8 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{m,n}$. *Show that*

a) $(\boldsymbol{A}^H)^\dagger = (\boldsymbol{A}^\dagger)^H$.

b) $(\boldsymbol{A}^\dagger)^\dagger = \boldsymbol{A}$.

c) $(\alpha\boldsymbol{A})^\dagger = \frac{1}{\alpha}\boldsymbol{A}^\dagger, \ \alpha \neq 0$.

Exercise 13.9 *Suppose* $k, m, n \in \mathbb{N}$, $\boldsymbol{A} \in \mathbb{C}^{m,n}$, $\boldsymbol{B} \in \mathbb{C}^{n,k}$. *Suppose* $\boldsymbol{A}$ *has linearly independent columns and* $\boldsymbol{B}$ *has linearly independent rows.*

a) *Show that* $(\boldsymbol{A}\boldsymbol{B})^\dagger = \boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$. *Hint: Let* $\boldsymbol{E} = \boldsymbol{A}\boldsymbol{F}$, $\boldsymbol{F} = \boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$. *Show by using* $\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^\dagger = \boldsymbol{I}$ *that* $\boldsymbol{F}$ *is the pseudo-inverse of* $\boldsymbol{E}$.

b) *Find* $\boldsymbol{A} \in \mathbb{R}^{1,2}$, $\boldsymbol{B} \in \mathbb{R}^{2,1}$ *such that* $(\boldsymbol{A}\boldsymbol{B})^\dagger \neq \boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$.

Exercise 13.10 *Show that* $\boldsymbol{A}^H = \boldsymbol{A}^\dagger$ *if and only if all singular values of* $\boldsymbol{A}$ *are either zero or one.*

13.1.2 Orthogonal Projections

The singular value decomposition and the pseudo-inverse can be used to compute orthogonal projections into the subspaces $\text{span}(\boldsymbol{A})$ and $\ker(\boldsymbol{A}^*)$.

We start by recalling some facts about sums, direct sums and orthogonal sums of subspaces (Cf. Chapter A). Suppose $\mathcal{S}$ and $\mathcal{T}$ are subspaces, and $\langle\cdot,\cdot\rangle$ an inner product on $\mathbb{R}^n$ or $\mathbb{C}^n$. We define

- **Sum:** $\mathcal{X} := \mathcal{S} + \mathcal{T} := \{\boldsymbol{s} + \boldsymbol{t} : \boldsymbol{s} \in \mathcal{S} \text{ and } \boldsymbol{t} \in \mathcal{T}\}$.
- **Direct Sum:** $\mathcal{S} + \mathcal{T}$ is called a direct sum if $\mathcal{S} \cap \mathcal{T} = \{\boldsymbol{0}\}$. We write $\mathcal{S} \oplus \mathcal{T}$ instead of $\mathcal{S} + \mathcal{T}$ in this case.
- **Orthogonal Sum:** $\mathcal{S} \oplus \mathcal{T}$ is an orthogonal sum if $\langle \boldsymbol{s}, \boldsymbol{t}\rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$ and all $\boldsymbol{t} \in \mathcal{T}$.
- **Orthogonal Complement:** $\mathcal{T} = \mathcal{S}^\perp := \{\boldsymbol{x} \in \mathcal{X} : \langle \boldsymbol{s}, \boldsymbol{x}\rangle = 0 \text{ for all } \boldsymbol{s} \in \mathcal{S}\}$.
- If $\mathcal{S} \oplus \mathcal{T}$ is an orthogonal sum and $\boldsymbol{v} = \boldsymbol{s} + \boldsymbol{t} \in \mathcal{S} \oplus \mathcal{T}$ with $\boldsymbol{s} \in \mathcal{S}$ and $\boldsymbol{t} \in \mathcal{T}$ then $\boldsymbol{s}$ and $\boldsymbol{t}$ are called the **orthogonal projections** of $\boldsymbol{v}$ into $\mathcal{S}$ and $\mathcal{T}$. The orthogonal projection is unique.

We recall that

- $\mathcal{S} + \mathcal{T} = \mathcal{T} + \mathcal{S}$ and $\mathcal{S} + \mathcal{T}$ is a subspace of $\mathbb{R}^n$ or $\mathbb{C}^n$.
- $\dim(\mathcal{S} + \mathcal{T}) = \dim \mathcal{S} + \dim \mathcal{T} - \dim(\mathcal{S} \cap \mathcal{T})$.
- $\dim(\mathcal{S} \oplus \mathcal{T}) = \dim \mathcal{S} + \dim \mathcal{T}$.
- Every $\boldsymbol{v} \in \mathcal{S} \oplus \mathcal{T}$ can be decomposed uniquely as $\boldsymbol{v} = \boldsymbol{s} + \boldsymbol{t}$, where $\boldsymbol{s} \in \mathcal{S}$ and $\boldsymbol{t} \in \mathcal{T}$.
- $\mathbb{C}^m = \text{span}(\boldsymbol{A}) \oplus \ker(\boldsymbol{A}^*)$ is an orthogonal sum for any $\boldsymbol{A} \in \mathbb{C}^{m,n}$.

Theorem 13.11 *Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $\boldsymbol{b} \in \mathbb{C}^m$. Then*

$$\boldsymbol{b}_1 := \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{b} \tag{13.2}$$

is the orthogonal projection of $\boldsymbol{b}$ into $\text{span}(\boldsymbol{A})$*, and*

$$\boldsymbol{b}_2 := (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\dagger)\boldsymbol{b} \tag{13.3}$$

is the orthogonal projection of $\boldsymbol{b}$ into the orthogonal complement $\ker(\boldsymbol{A}^*)$ *of* $\text{span}(\boldsymbol{A})$*.*

Proof. Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is a singular value decomposition of $\boldsymbol{A} \in \mathbb{C}^{m,n}$. Then

$$\boldsymbol{b} = \boldsymbol{U}\boldsymbol{U}^*\boldsymbol{b} = [\boldsymbol{U}_1\ \boldsymbol{U}_2]\begin{bmatrix}\boldsymbol{U}_1^* \\ \boldsymbol{U}_2^*\end{bmatrix}\boldsymbol{b} = \boldsymbol{U}_1\boldsymbol{U}_1^*\boldsymbol{b} + \boldsymbol{U}_2\boldsymbol{U}_2^*\boldsymbol{b} =: \boldsymbol{b}_1 + \boldsymbol{b}_2,$$

where $\boldsymbol{b}_j = \boldsymbol{U}_j\boldsymbol{c}_j$ and $\boldsymbol{c}_j := \boldsymbol{U}_j^*\boldsymbol{b}$ for $j = 1, 2$. Since $\boldsymbol{U}_1(\boldsymbol{U}_2)$ is an orthonormal basis for $\text{span}(\boldsymbol{A})$ ($\ker(\boldsymbol{A}^*)$), we have $\boldsymbol{b}_1(\boldsymbol{b}_2) \in \text{span}(\boldsymbol{A})$ ($\ker(\boldsymbol{A}^*)$). Now $\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{U}_1\boldsymbol{U}_1^*$, and then $\boldsymbol{b}_2 = \boldsymbol{b} - \boldsymbol{b}_1 = (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\dagger)\boldsymbol{b}$. ∎

Example 13.12 *The singular value decomposition of* $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ *is* $\boldsymbol{A} = \boldsymbol{I}_3 \boldsymbol{A} \boldsymbol{I}_2$. *Thus* $\boldsymbol{U}_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ *and* $\boldsymbol{U}_2 = \left[\begin{smallmatrix} 0 \\ 0 \\ 1 \end{smallmatrix}\right]$. *Moreover* $\boldsymbol{A}^\dagger = \boldsymbol{I}_2 \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right] \boldsymbol{I}_3 = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right]$. *If* $\boldsymbol{b} = \left[\begin{smallmatrix} b_1 \\ b_2 \\ b_3 \end{smallmatrix}\right]$, *then* $\boldsymbol{b}_1 = \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{b} = \boldsymbol{U}_1 \boldsymbol{U}_1^T \boldsymbol{b} = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right] \boldsymbol{b} = \left[\begin{smallmatrix} b_1 \\ b_2 \\ 0 \end{smallmatrix}\right]$ *and* $\boldsymbol{b}_2 = (\boldsymbol{I}_3 - \boldsymbol{A}\boldsymbol{A}^\dagger)\boldsymbol{b} = \boldsymbol{U}_2 \boldsymbol{U}_2^T \boldsymbol{b} = \left[\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right] \boldsymbol{b} = \left[\begin{smallmatrix} 0 \\ 0 \\ b_3 \end{smallmatrix}\right]$.

Exercise 13.13 *Show that if* $\boldsymbol{A}$ *has rank* n *then* $\boldsymbol{A}(\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{b}$ *is the projection of* $\boldsymbol{b}$ *into* $\operatorname{span}(\boldsymbol{A})$. *(Cf. Exercise 13.4.)*

Exercise 13.14 *Consider the linear system* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *where* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *has rank* $r > 0$ *and* $\boldsymbol{b} \in \mathbb{C}^n$. *Let*

$$\boldsymbol{U}^H \boldsymbol{A} \boldsymbol{V} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

represent the singular value decomposition of $\boldsymbol{A}$.

a) *Let* $\boldsymbol{c} = [c_1, \ldots, c_n]^T = \boldsymbol{U}^H \boldsymbol{b}$ *and* $\boldsymbol{y} = [y_1, \ldots, y_n]^T = \boldsymbol{V}^H \boldsymbol{x}$. *Show that* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *if and only if*

$$\begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{y} = \boldsymbol{c}.$$

b) *Show that* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *has a solution* $\boldsymbol{x}$ *if and only if* $c_{r+1} = \cdots = c_n = 0$.

c) *Deduce that a linear system* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *has either no solution, one solution or infinitely many solutions.*

Exercise 13.15 *For any* $\boldsymbol{A} \in \mathbb{C}^{m,n}$, $\boldsymbol{b} \in \mathbb{C}^n$ *show that one and only one of the following systems has a solution*

$$(1) \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \qquad (2) \quad \boldsymbol{A}^H \boldsymbol{y} = \boldsymbol{0},\ \boldsymbol{y}^H \boldsymbol{b} \neq 0.$$

In other words either $\boldsymbol{b} \in \operatorname{span}(\boldsymbol{A})$, *or we can find* $\boldsymbol{y} \in \ker(\boldsymbol{A}^H)$ *such that* $\boldsymbol{y}^H \boldsymbol{b} \neq 0$. *This is called* **Fredholms alternative**.

13.2 The Least Squares Problem

Let $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $\boldsymbol{b} \in \mathbb{C}^m$ be given. Consider the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ of m equations in n unknowns. If $m > n$, we have more equations than unknowns and there might be no vector $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Let $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \in \mathbb{C}^m$. We can then pick a vector norm $\|\cdot\|$ and look for $\boldsymbol{x} \in \mathbb{C}^n$ which minimizes $\|\boldsymbol{r}(\boldsymbol{x})\|$. The choice $\|\cdot\| = \|\cdot\|_2$, the Euclidean norm, is particularly convenient and will be studied here.

Definition 13.16 *Given $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $\boldsymbol{b} \in \mathbb{C}^m$. We call an $\boldsymbol{x} \in \mathbb{C}^n$ which minimizes $\|\boldsymbol{r}(\boldsymbol{x})\|_2^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ a least squares solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. We set*

$$E(\boldsymbol{x}) := \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{r}(\boldsymbol{x})\|_2^2.$$

To find $\boldsymbol{x}$ which minimizes $E(\boldsymbol{x})$ is called the least squares problem.

Since the square root function is monotone minimizing $E(\boldsymbol{x})$ or $\sqrt{E(\boldsymbol{x})}$ is equivalent.

Theorem 13.17 *The least squares problem always has a solution. The solution is unique if and only if $\boldsymbol{A}$ has linearly independent columns. Moreover, the following are equivalent.*

1. *$\boldsymbol{x}$ is a solution of the least squares problem.*
2. *$\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^*\boldsymbol{b}$*
3. *$\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}$, for some $\boldsymbol{z} \in \ker(\boldsymbol{A})$, and where $\boldsymbol{A}^\dagger$ is the pseudo-inverse of $\boldsymbol{A}$.*

We have $\|\boldsymbol{x}\|_2 \geq \|\boldsymbol{A}^\dagger\boldsymbol{b}\|_2$ for all solutions $\boldsymbol{x}$ of the least squares problem.

Proof. Let $\boldsymbol{b} = \boldsymbol{b}_1 + \boldsymbol{b}_2$, where $\boldsymbol{b}_1 \in \operatorname{span}(\boldsymbol{A})$ and $\boldsymbol{b}_2 \in \ker(\boldsymbol{A}^*)$ are the orthogonal projections into $\operatorname{span}(\boldsymbol{A})$ and $\ker(\boldsymbol{A}^*)$, respectively (see Theorem 13.11). Since $\boldsymbol{b}_2^*\boldsymbol{v} = 0$ for any $\boldsymbol{v} \in \operatorname{span}(\boldsymbol{A})$ we have $\boldsymbol{b}_2^*(\boldsymbol{b}_1 - \boldsymbol{A}\boldsymbol{x}) = 0$ for any $\boldsymbol{x} \in \mathbb{C}^m$. Therefore, for $\boldsymbol{x} \in \mathbb{C}^n$,

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = \|(\boldsymbol{b}_1 - \boldsymbol{A}\boldsymbol{x}) + \boldsymbol{b}_2\|_2^2 = \|\boldsymbol{b}_1 - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\boldsymbol{b}_2\|_2^2 \geq \|\boldsymbol{b}_2\|_2^2,$$

with equality if and only if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$. Since $\boldsymbol{b}_1 \in \operatorname{span}(\boldsymbol{A})$ we can always find such an $\boldsymbol{x}$ and existence follows.
$1 \iff 2$: By what we have shown $\boldsymbol{x}$ solves the least squares problem if and only if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$ so that $\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1 + \boldsymbol{b}_2 - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_2 \in \ker(\boldsymbol{A}^*)$, or $\boldsymbol{A}^*(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{0}$.
$1 \implies 3$: Suppose $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$ and define$\boldsymbol{z} := \boldsymbol{x} - \boldsymbol{A}^\dagger\boldsymbol{b}$. Then $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b} = \boldsymbol{b}_1 - \boldsymbol{b}_1 = \boldsymbol{0}$ and $\boldsymbol{z} \in \ker(\boldsymbol{A})$.
$3 \implies 1$: If $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}$ with $\boldsymbol{z} \in \ker(\boldsymbol{A})$ then $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{A}\boldsymbol{z} = \boldsymbol{b}_1$.

If $\boldsymbol{A}$ has linearly independent columns then $\ker(\boldsymbol{A}) = \{\boldsymbol{0}\}$ and $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b}$ is the unique solution.

Suppose $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}$, with $\boldsymbol{z} \in \ker(\boldsymbol{A})$ is a solution. To show the minimum norm property $\|\boldsymbol{x}\|_2 \geq \|\boldsymbol{A}^\dagger\boldsymbol{b}\|_2$ we recall that if the right singular vectors of $\boldsymbol{A}$ are partitioned as $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n] = [\boldsymbol{V}_1, \boldsymbol{V}_2]$, then $\boldsymbol{V}_2$ is a basis for $\ker(\boldsymbol{A})$. Moreover, $\boldsymbol{V}_2^*\boldsymbol{V}_1 = \boldsymbol{0}$ since $\boldsymbol{V}$ has orthonormal columns. If $\boldsymbol{A}^\dagger = \boldsymbol{V}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_1^*$ and $\boldsymbol{z} \in \ker(\boldsymbol{A})$ then $\boldsymbol{z} = \boldsymbol{V}_2\boldsymbol{y}$ for some $\boldsymbol{y} \in \mathbb{C}^{n-r}$ and we obtain

$$\boldsymbol{z}^*\boldsymbol{A}^\dagger\boldsymbol{b} = \boldsymbol{y}^*\boldsymbol{V}_2^*\boldsymbol{V}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_1^*\boldsymbol{b} = \boldsymbol{0}.$$

Thus $\boldsymbol{z}$ and $\boldsymbol{A}^\dagger\boldsymbol{b}$ are orthogonal so that $\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}\|_2^2 = \|\boldsymbol{A}^\dagger\boldsymbol{b}\|_2^2 + \|\boldsymbol{z}\|_2^2 \geq \|\boldsymbol{A}^\dagger\boldsymbol{b}\|_2^2$. ∎

The linear system

$$\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^*\boldsymbol{b}$$

is called the **normal equations**. It is a linear system of n equations in n unknowns. If $\boldsymbol{A}$ is real then the coefficient matrix $\boldsymbol{A}^T\boldsymbol{A}$ is nonsingular and hence symmetric positive definite if and only if $\boldsymbol{A}$ has linearly independent columns.

Before discussing numerical methods for solving the least squares problem we consider some examples.

13.3 Examples

Example 13.18 *We choose n functions $\phi_1, \phi_2, \ldots, \phi_n$ defined for $t \in \{t_1, t_2, \ldots, t_m\}$ and positive numbers $w_1, \ldots, w_m$. Typical examples of functions might be polynomials, trigonometric functions, exponential functions, or splines. We want to find $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ such that*

$$E(\boldsymbol{x}) := \sum_{i=1}^{m} w_i \left[\sum_{j=1}^{n} x_j \phi_j(t_i) - y_i \right]^2$$

is as small as possible. Let $p(t) := \sum_{j=1}^{n} x_j \phi_j(t)$. The numbers w_i are called weights. If y_i is an accurate observation, we can choose a large weight w_i. This will force $p(t_i) - y_i$ to be small. Similarly, a small w_i will allow $p(t_i) - y_i$ to be large. If an estimate for the standard deviation δy_i in y_i is known for each i, we can choose $w_i = 1/(\delta y_i)^2$, $i = 1, 2, \ldots, m$. Let $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $\boldsymbol{b} \in \mathbb{R}^m$ have elements $a_{i,j} = \sqrt{w_i}\phi_j(t_i)$ and $b_i = \sqrt{w_i} y_i$. Then

$$(\boldsymbol{A}\boldsymbol{x})_i = \sqrt{w_i} \sum_{j=1}^{n} x_j \phi_j(t_i),$$

$$E(\boldsymbol{x}) = \sum_{i=1}^{m} [(\boldsymbol{A}\boldsymbol{x})_i - b_i]^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2.$$

and we have a least squares problem.

The i, j element $b_{i,j}$ in $\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{A}$ and the ith component c_i in $\boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{b}$ take the form

$$\begin{array}{ll} b_{i,j} & = \sum_{k=1}^{m} a_{k,i} a_{k,j} = \sum_{k=1}^{m} w_k \phi_i(t_k) \phi_j(t_k), \\ c_i & = \sum_{k=1}^{m} w_k^{1/2} y_k \phi_i(t_k). \end{array} \tag{13.4}$$

In particular, if $n = 2, w_i = 1, i = 1, \ldots, m$, $\phi_1(t) = 1$, and $\phi_2(t) = t$, the normal equations can be written

$$\begin{bmatrix} m & \sum t_k \\ \sum t_k & \sum t_k^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sum y_k \\ \sum t_k y_k \end{bmatrix}. \tag{13.5}$$

Here k ranges from 1 to m in the sums. This 2×2 system is symmetric positive definite and is easily solved for x_1 and x_2.

Example 13.19 *With the data*

x	1.0	2.0	3.0	4.0
y	3.1	1.8	1.0	0.1

we try a least squares fit of the form

$$p(t) = x_1 + x_2 t.$$

We can find x_1 and x_2 by solving the linear system (13.5)*. In this case we obtain*

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 10.1 \end{bmatrix}. \tag{13.6}$$

The solution is $x_1 = 3.95$ and $x_2 = -0.98$. The data and the polynomial $p(t)$ are shown in Figure 13.1.

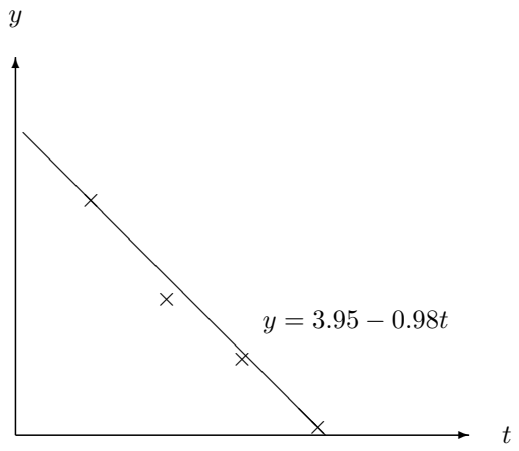


Figure 13.1. *A least squares fit to data.*

Example 13.20 *Suppose we have a simple input/output model. To every input $\boldsymbol{u} \in \mathbb{R}^n$ we obtain an output $y \in \mathbb{R}$. Assuming we have a linear relation*

$$y = \boldsymbol{u}^T \boldsymbol{x} = \sum_{i=1}^{n} u_i x_i,$$

between $\boldsymbol{u}$ and y, how can we determine $\boldsymbol{x}$?

Performing $m \geq n$ experiments we obtain a table of values

$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	$\cdots$	$\boldsymbol{u}_m$
y	y_1	y_2	$\cdots$	y_m

.

We would like to find $\boldsymbol{x}$ *such that*

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \vdots \\ \boldsymbol{u}_m^T \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \boldsymbol{b}.$$

We can estimate $\boldsymbol{x}$ *by solving the least squares problem* $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.

Exercise 13.21 *Suppose* $(t_i, y_i)_{i=1}^m$ *are* m *points in the plane. We consider the over-determined systems*

$$\begin{array}{llcl} \textbf{(i)} & x_1 & = & y_1 \\ & x_1 & = & y_2 \\ & & \vdots & \\ & x_1 & = & y_m \end{array} \qquad \begin{array}{llcl} \textbf{(ii)} & x_1 + t_1 x_2 & = & y_1 \\ & x_1 + t_2 x_2 & = & y_2 \\ & & \vdots & \\ & x_1 + t_m x_2 & = & y_m \end{array}$$

a) *Find the normal equations for* (**i**) *and the least squares solution.*

b) *Find the normal equations for* (**ii**) *and give a geometric interpretation of the least squares solution.*

Exercise 13.22 *Related to* (**ii**) *in Exercise 13.21 we have the overdetermined system*

$$\textbf{(iii)} \quad x_1 + (t_i - \hat{t})x_2 = y_i, \quad i = 1, 2, \ldots, m,$$

where $\hat{t} = (t_1 + \cdots + t_m)/m$.

a) *Find the normal equations for* (**iii**) *and give a geometric interpretation of the least squares solution.*

b) *Fit a straight line to the points* (t_i, y_i)*:* $(998.5, 1)$*,* $(999.5, 1.9)$*,* $(1000.5, 3.1)$ *and* $(1001.5, 3.5)$ *using* **a)**. *Draw a sketch of the solution.*

Exercise 13.23 *In this problem we derive an algorithm to fit a circle* $(t - c_1)^2 + (y - c_2)^2 = r^2$ *to* $m \geq 3$ *given points* $(t_i, y_i)_{i=1}^m$ *in the* (t, y)*-plane. We obtain the overdetermined system*

$$(t_i - c_1)^2 + (y_i - c_2)^2 = r^2, \ i = 1, \ldots, m, \tag{13.7}$$

of m *equations in the three unknowns* c_1, c_2 *and* r*. This system is nonlinear, but it can be solved from the linear system*

$$t_i x_1 + y_i x_2 + x_3 = t_i^2 + y_i^2, \ i = 1, \ldots, m, \tag{13.8}$$

and then setting $c_1 = x_1/2$*,* $c_2 = x_2/2$ *and* $r^2 = c_1^2 + c_2^2 + x_3$.

a) *Derive* (13.8) *from* (13.7). *Explain how we can find* c_1, c_2, r *once* $[x_1, x_2, x_3]$ *is determined.*

b) *Formulate* (13.8) *as a linear least squares problem for suitable* $\boldsymbol{A}$ *and* $\boldsymbol{b}$.

c) *Does the matrix* $\boldsymbol{A}$ *in* **b)** *have linearly independent columns?*

d) *Use* (13.8) *to find the circle passing through the three points* $(1,4),(3,2),(1,0)$.

13.4 Numerical Solution using the Normal Equations

We assume that $\boldsymbol{A}$ and $\boldsymbol{b}$ are real and that $\boldsymbol{A}$ has linearly independent columns. The coefficient matrix $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A}$ in the normal equations is symmetric positive definite, and we can solve these equations using the $\boldsymbol{R}^T\boldsymbol{R}$ factorization of $\boldsymbol{B}$.

Consider forming the normal equations. We can use either a column oriented- or a row oriented approach. To derive these we partition $\boldsymbol{A}$ in terms of columns or rows as

$$\boldsymbol{A} = [\boldsymbol{a}_{:1}, \ldots, \boldsymbol{a}_{:n}] = \begin{bmatrix} \boldsymbol{a}_{1:}^T \\ \vdots \\ \boldsymbol{a}_{m:}^T \end{bmatrix}.$$

We then find

1. $(\boldsymbol{A}^T\boldsymbol{A})_{ij} = \boldsymbol{a}_{:i}^T\boldsymbol{a}_{:j}, \quad (\boldsymbol{A}^T\boldsymbol{b})_i = \boldsymbol{a}_{:i}^T\boldsymbol{b},$ (inner product form),
2. $\boldsymbol{A}^T\boldsymbol{A} = \sum_{i=1}^m \boldsymbol{a}_{i:}\boldsymbol{a}_{i:}^T, \quad \boldsymbol{A}^T\boldsymbol{b} = \sum_{i=1}^m b_i\boldsymbol{a}_{i:},$ (outer product form).

The outer product form is suitable for large problems since it uses only one pass through the data importing one row of mA at a time from some separate storage.

Consider the number of operations to compute $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A}$. We need $2m$ flops to find each $\boldsymbol{a}_{:i}^T\boldsymbol{a}_{:j}$. Since $\boldsymbol{B}$ is symmetric we only need to compute $n(n+1)/2$ such inner products. It follows that $\boldsymbol{B}$ can be computed in $O(mn^2)$ flops. The computation of $\boldsymbol{B}$ using outer products can also be done in $O(mn^2)$ flops by computing only one half of $\boldsymbol{A}$. In conclusion the number of operations are $O(mn^2)$ to find $\boldsymbol{B}$, $2mn$ to find $\boldsymbol{A}^T\boldsymbol{b}$, $O(n^3/3)$ to find $\boldsymbol{R}$, $O(n^2)$ to solve $\boldsymbol{R}^T\boldsymbol{y} = \boldsymbol{c}$ and $O(n^2)$ to solve $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$. Since $m \geq n$, the bulk of the work is to find $\boldsymbol{B}$.

A problem with the normal equation approach is that the linear system can be poorly conditioned. In fact the 2-norm condition number of $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A}$ is the square of the condition number of $\boldsymbol{A}$. This follows, since the singular values of $\boldsymbol{B}$ are the square of the singular values of $\boldsymbol{A}$. If $\boldsymbol{A}$ is ill-conditioned, this could make the normal equation approach problematic. One difficulty which can be encountered is that the computed $\boldsymbol{A}^T\boldsymbol{A}$ might not be positive definite. See Problem 13.31 for an example.

13.5 Numerical Solution using the QR Factorization

Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has rank n and let $\boldsymbol{b} \in \mathbb{R}^m$. The QR factorization can be used to solve the least squares problem $\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$. Suppose $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ is a

QR factorization of $\boldsymbol{A}$. Since $\boldsymbol{Q}_1$ has orthonormal columns we find

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 = \boldsymbol{R}_1^T\boldsymbol{R}_1, \quad \boldsymbol{A}^T\boldsymbol{b} = \boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{b}.$$

Since $\boldsymbol{A}$ has rank n the matrix $\boldsymbol{R}_1^T$ is nonsingular and can be canceled. Thus

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b} \Longrightarrow \boldsymbol{R}_1\boldsymbol{x} = \boldsymbol{c}_1, \quad \boldsymbol{c}_1 := \boldsymbol{Q}_1^T\boldsymbol{b}.$$

We can use Householder transformations or Givens rotations to find $\boldsymbol{R}_1$ and $\boldsymbol{c}_1$. Consider using the Householder triangulation algorithm Algorithm 12.18. We find $\boldsymbol{R} = \boldsymbol{Q}^T\boldsymbol{A}$ and $\boldsymbol{c} = \boldsymbol{Q}^T\boldsymbol{b}$, where $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is the QR decomposition of $\boldsymbol{A}$. The matrices $\boldsymbol{R}_1$ and $\boldsymbol{c}_1$ are located in the first n rowws of $\boldsymbol{R}$ and $\boldsymbol{c}$.

Thus have the following method to solve the full rank least squares problem.

1. `[R,c]=housetriang(A,b).`
2. Solve $\boldsymbol{R}(1:n, 1:n)\boldsymbol{x} = \boldsymbol{c}(1:n)$ for the least squares solution $\boldsymbol{x}$.

Example 13.24 *Consider the least squares problem with*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 7 \\ 1 & -1 & -4 \\ 1 & -1 & 2 \end{bmatrix} \text{ and } \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

This is the matrix in Example 12.2. The least squares solution $\boldsymbol{x}$ is found by solving the system

$$\begin{bmatrix} 2 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

and we find $\boldsymbol{x} = [1, 0, 0]^T$.

Using Householder triangulation is a useful alternative to normal equations for solving full rank least squares problems. The 2 norm condition number for the system $\boldsymbol{R}_1\boldsymbol{x} = \boldsymbol{c}_1$ is $K_2(\boldsymbol{R}_1) = K_2(\boldsymbol{Q}_1\boldsymbol{R}_1) = K_2(\boldsymbol{A})$, and as discussed in the previous section this is the square root of $K_2(\boldsymbol{A}^T\boldsymbol{A})$, the condition number for the normal equations. Thus if $\boldsymbol{A}$ is mildly ill-conditioned the normal equations can be quite ill-conditioned and solving the normal equations can give inaccurate results. On the other hand Algorithm 12.18 is quite stable.

But using Householder transformations requires more work. The leading term in the number of flops in Algorithm 12.18 is approximately $2mn^2 - 2n^3/3$, (cf. (12.10)), while the number of flops needed to form the normal equations, taking advantage of symmetry is $O(mn^2)$. Thus for m much larger than n using Householder triangulation requires twice as many flops as the an approach based on the normal equations. Also, Householder triangulation have problems taking advantage of the structure in sparse problems.

13.6 Numerical Solution using the Singular Value Factorization

This method can be used even if $\boldsymbol{A}$ does not have full rank. It requires knowledge of the pseudo-inverse of $\boldsymbol{A}$. By Theorem 13.17

$$\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{z}$$

is a least squares solution for any $\boldsymbol{z} \in \ker(\boldsymbol{A})$.

Example 13.25 *The pseudo-inverse of* $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$ *is* $\boldsymbol{A}^\dagger = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$. *(cf. Example 13.3. Moreover,* $[-1, 1]^T$ *is a basis for* $\ker(\boldsymbol{A})$. *If* $\boldsymbol{b} = [b_1, b_2, b_3]^T$, *then for any* $z \in \mathbb{R}$ *the vector*

$$\boldsymbol{x} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + z\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

is a solution of $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ *and this gives all solutions.*

When $\operatorname{rank}(\boldsymbol{A})$ is less than the number of columns of $\boldsymbol{A}$ then $\ker(\boldsymbol{A}) \neq \{\boldsymbol{0}\}$, and we have a choice of $\boldsymbol{z}$. One possible choice is $\boldsymbol{z} = \boldsymbol{0}$ giving the minimal norm solution $\boldsymbol{A}^\dagger \boldsymbol{b}$. (Cf. Theorem 13.17.)

13.7 Perturbation Theory for Least Squares

In this section we consider what effect small changes in the data $\boldsymbol{A}, \boldsymbol{b}$ have on the solution $\boldsymbol{x}$ of the least squares problem $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$.

If $\boldsymbol{A}$ has linearly independent columns then we can write the least squares solution $\boldsymbol{x}$ (the solution of $\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^H\boldsymbol{b}$) as

$$\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b}, \quad \boldsymbol{A}^\dagger := (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H.$$

13.7.1 Perturbing the right hand side

Let us now consider the effect of a perturbation in $\boldsymbol{b}$ on $\boldsymbol{x}$.

Theorem 13.26 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *has linearly independent columns, and let* $\boldsymbol{b}, \boldsymbol{e} \in \mathbb{C}^m$. *Let* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ *be the solutions of* $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ *and* $\min\|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b} - \boldsymbol{e}\|_2$. *Finally, let* $\boldsymbol{b}_1$, $\boldsymbol{e}_1$ *be the projections of* $\boldsymbol{b}$ *and* $\boldsymbol{e}$ *on* $\operatorname{span}(\boldsymbol{A})$. *If* $\boldsymbol{b}_1 \neq \boldsymbol{0}$, *we have for any operator norm*

$$\frac{1}{K(\boldsymbol{A})}\frac{\|\boldsymbol{e}_1\|}{\|\boldsymbol{b}_1\|} \leq \frac{\|\boldsymbol{y} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq K(\boldsymbol{A})\frac{\|\boldsymbol{e}_1\|}{\|\boldsymbol{b}_1\|}, \quad K(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^\dagger\|. \tag{13.9}$$

Proof. Subtracting $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b}$ from $\boldsymbol{y} = \boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{A}^\dagger \boldsymbol{e}$ we have $\boldsymbol{y} - \boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{e}$. Since $\boldsymbol{A}^H \boldsymbol{e} = \boldsymbol{A}^H \boldsymbol{e}_1$, we have $\boldsymbol{A}^\dagger \boldsymbol{e} = \boldsymbol{A}^\dagger \boldsymbol{e}_1$. Thus $\|\boldsymbol{y} - \boldsymbol{x}\| = \|\boldsymbol{A}^\dagger \boldsymbol{e}_1\| \leq \|\boldsymbol{A}^\dagger\| \|\boldsymbol{e}_1\|$. Moreover, $\|\boldsymbol{b}_1\| = \|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \|\boldsymbol{x}\|$. Therefore $\|\boldsymbol{y} - \boldsymbol{x}\| / \|\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \|\boldsymbol{A}^\dagger\| \|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$ proving the rightmost inequality. From $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{e}_1$ and $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b}_1$ we obtain the leftmost inequality. □

(13.9) is analogous to the bound (8.19) for linear systems. We see that the number $K(\boldsymbol{A}) = \|\boldsymbol{A}\| \|\boldsymbol{A}^\dagger\|$ generalizes the condition number $\|\boldsymbol{A}\| \|\boldsymbol{A}^{-1}\|$ for a square matrix. The main difference between (13.9) and (8.19) is however that $\|\boldsymbol{e}\| / \|\boldsymbol{b}\|$ in (8.19) has been replaced by $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$, the projections of $\boldsymbol{e}$ and $\boldsymbol{b}$ on span$(\boldsymbol{A})$. If $\boldsymbol{b}$ lies almost entirely in $N(\boldsymbol{A}^H)$, i.e. $\|\boldsymbol{b}\| / \|\boldsymbol{b}_1\|$ is large, $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$ can be much larger than $\|\boldsymbol{e}\| / \|\boldsymbol{b}\|$. This is illustrated in Figure 13.2. If $\boldsymbol{b}$ is almost orthogonal to span$(\boldsymbol{A})$, $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$ will normally be much larger than $\|\boldsymbol{e}\| / \|\boldsymbol{b}\|$. Note that $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$ is also present in the lower bound.

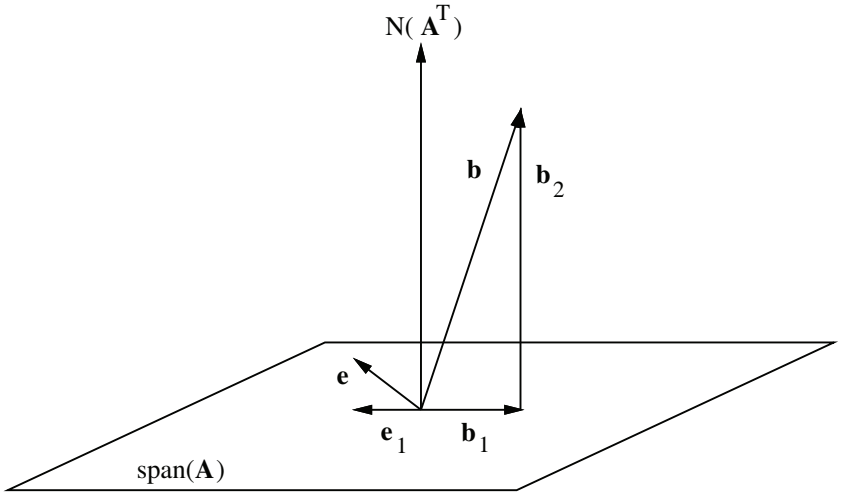


Figure 13.2. *Graphical interpretation of the bounds in Theorem 13.26.*

Example 13.27 *Suppose*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 10^{-4} \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} 10^{-6} \\ 0 \\ 0 \end{bmatrix}.$$

For this example we can compute $K(\boldsymbol{A})$ by finding $\boldsymbol{A}^\dagger$ explicitly. Indeed,

$$\boldsymbol{A}^T \boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad (\boldsymbol{A}^T \boldsymbol{A})^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$
$$\boldsymbol{A}^\dagger = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Thus $K_\infty(\boldsymbol{A}) = \|\boldsymbol{A}\|\infty\|\boldsymbol{A}^\dagger\|\infty = 2 \cdot 2 = 4$ *is quite small.*

Consider now the projections $\boldsymbol{b}_1$ *and* $\boldsymbol{b}_2$*. We have* $\text{span}(\boldsymbol{A}) = \text{span}(\boldsymbol{e}_1, \boldsymbol{e}_2) \subset \mathbb{R}^2$*, and* $N(\boldsymbol{A}^T) = \text{span}(\boldsymbol{e}_3) \subset \mathbb{R}^3$*. The projection* $\boldsymbol{b}_1$ *on* $\text{span}(\boldsymbol{A})$ *is* $\boldsymbol{b}_1 = (10^{-4}, 0)^T$*. Since* $\|\boldsymbol{b}\|\infty/\|\boldsymbol{b}_1\|\infty = 10^4$ *is large, we expect that the solutions* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *of* $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ *and* $\min\|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b} - \boldsymbol{e}\|_2$ *will differ by much more than* $\|\boldsymbol{e}\|\infty/\|\boldsymbol{b}\|\infty = 10^{-6}$*. To check this we compute* $\boldsymbol{x}$ *and* $\boldsymbol{y}$*. These can be found by either solving the normal equations or by solving* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$*,* $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}_1 + \boldsymbol{e}_1$*. This gives* $\boldsymbol{x} = \boldsymbol{b}_1 = (10^{-4}, 0)^T$ *and* $\boldsymbol{y} = \boldsymbol{b}_1 + \boldsymbol{e}_1 = (10^{-4} + 10^{-6}, 0)^T$*. We find*

$$\frac{\|\boldsymbol{x} - \boldsymbol{y}\|\infty}{\|\boldsymbol{x}\|\infty} = \frac{10^{-6}}{10^{-4}} = 10^{-2},$$

and this is indeed much larger than $\|\boldsymbol{e}\|\infty/\|\boldsymbol{b}\|\infty = 10^{-6}$*, but equals* $\|\boldsymbol{e}_1\|\infty/\|\boldsymbol{b}_1\|\infty$*.*

(13.9) *takes the form*

$$\frac{1}{4}10^{-2} \leq \frac{\|\boldsymbol{y} - \boldsymbol{x}\|\infty}{\|\boldsymbol{x}\|\infty} \leq 4 \cdot 10^{-2}.$$

Exercise 13.28 *Let*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

a) *Determine the projections* $\boldsymbol{b}_1$ *and* $\boldsymbol{b}_2$ *of* $\boldsymbol{b}$ *on* $\text{span}(\boldsymbol{A})$ *and* $N(\boldsymbol{A}^T)$*.*

b) *Compute* $K(\boldsymbol{A}) = \|\boldsymbol{A}\|_2\|\boldsymbol{A}^\dagger\|_2$*.*

For each $\boldsymbol{A}$ we can find $\boldsymbol{b}$ and $\boldsymbol{e}$ so that we have equality in the upper bound in (13.9). The lower bound is best possible in a similar way.

Exercise 13.29 a) *Let* $\boldsymbol{A} \in \mathbb{C}^{m,n}$*. Show that we have equality to the right in* (13.9) *if* $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{y}_A$*,* $\boldsymbol{e}_1 = \boldsymbol{y}_{A^\dagger}$ *where* $\|\boldsymbol{A}\boldsymbol{y}_A\| = \|\boldsymbol{A}\|$*,* $\|\boldsymbol{A}^\dagger\boldsymbol{y}_{A^\dagger}\| = \|\boldsymbol{A}^\dagger\|$*.*

b) *Show that we have equality to the left if we switch* $\boldsymbol{b}$ *and* $\boldsymbol{e}$ *in a).*

c) *Let* $\boldsymbol{A}$ *be as in Example 13.27. Find extremal* $\boldsymbol{b}$ *and* $\boldsymbol{e}$ *when the* l_∞ *norm is used.*

13.7.2 Perturbing the matrix

The analysis of the effects of a perturbation $\boldsymbol{E}$ in $\boldsymbol{A}$ is quite difficult. The following result is stated without proof, see [12, p. 51]. For other estimates see [2] and [19].

Theorem 13.30 *Suppose* $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{m,n}$*,* $m > n$*, where* $\boldsymbol{A}$ *has linearly independent columns and* $\alpha := 1 - \|\boldsymbol{E}\|_2\|\boldsymbol{A}^\dagger\|_2 > 0$*. Then* $\boldsymbol{A} + \boldsymbol{E}$ *has linearly independent columns. Let* $\boldsymbol{b} = \boldsymbol{b}_1 + \boldsymbol{b}_2 \in \mathbb{C}^m$ *where* $\boldsymbol{b}_1$ *and* $\boldsymbol{b}_2$ *are the projections on* $\text{span}(\boldsymbol{A})$ *and* $N(\boldsymbol{A}^H)$ *respectively. Suppose* $\boldsymbol{b}_1 \neq \boldsymbol{0}$*. Let* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *be the solutions of* $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ *and* $\min\|(\boldsymbol{A} + \boldsymbol{E})\boldsymbol{y} - \boldsymbol{b}\|_2$*. Then*

$$\rho = \frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2}{\|\boldsymbol{x}\|_2} \leq \frac{1}{\alpha}K(1 + \beta K)\frac{\|\boldsymbol{E}\|_2}{\|\boldsymbol{A}\|_2}, \quad \beta = \frac{\|\boldsymbol{b}_2\|_2}{\|\boldsymbol{b}_1\|_2}, \quad K = \|\boldsymbol{A}\|_2\|\boldsymbol{A}^\dagger\|_2. \tag{13.10}$$

(13.10) says that the relative error in $\boldsymbol{y}$ as an approximation to $\boldsymbol{x}$ can be at most $K(1+\beta K)/\alpha$ times as large as the size $\|\boldsymbol{E}\|_2/\|\boldsymbol{A}\|_2$ of the relative perturbation in $\boldsymbol{A}$. If $\boldsymbol{b}$ lies almost entirely in span($\boldsymbol{A}$), β will be small, and we have $\rho \leq \frac{1}{\alpha}K\|\boldsymbol{E}\|_2/\|\boldsymbol{A}\|_2$. This corresponds to the estimate (8.25) for linear systems. If β is not small, the term $\frac{1}{\alpha}K^2\beta\|\boldsymbol{E}\|_2/\|\boldsymbol{A}\|_2$ will dominate. In other words, the condition number is roughly $K(\boldsymbol{A})$ if β is small and $K(\boldsymbol{A})^2\beta$ if β is not small. Note that β is large if $\boldsymbol{b}$ is almost orthogonal to span($\boldsymbol{A}$) and that $\boldsymbol{b}_2 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ is the residual of $\boldsymbol{x}$.

Exercise 13.31 *Consider the least squares problems where*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1+\epsilon \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}, \quad \epsilon \in \mathbb{R}.$$

a) *Find the normal equations and the exact least squares solution.*

b) *Suppose ϵ is small and we replace the $(2,2)$ entry $3+2\epsilon+\epsilon^2$ in $\boldsymbol{A}^T\boldsymbol{A}$ by $3+2\epsilon$. (This will be done in a computer if $\epsilon < \sqrt{u}$, u being the round-off unit). Solve $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$ for $\boldsymbol{x}$ and compare with the $\boldsymbol{x}$ found in a). (We will get a much more accurate result using the QR factorization or the singular value decomposition on this problem).*

13.8 Perturbation Theory for Singular Values

In this section we consider what effect a small change in the matrix $\boldsymbol{A}$ has on the singular values.

We recall the Hoffman-Wielandt Theorem for singular values, Theorem 7.24. If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m,n}$ are rectangular matrices with singular values $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$, then

$$\sum_{j=1}^{n} |\alpha_j - \beta_j|^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2.$$

This shows that the singular values of a matrix are well conditioned. Changing the Frobenius norm of a matrix by small amount only changes the singular values by a small amount.

Using the 2-norm we have a similar result involving only one singular value.

Theorem 13.32 *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m,n}$ be rectangular matrices with singular values $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$. Then*

$$|\alpha_j - \beta_j| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2, \text{ for } j = 1, 2, \ldots, n. \tag{13.11}$$

Proof. Fix j and let $\mathcal{S}$ be the $n-j+1$ dimensional subspace for which the minimum in Theorem 7.23 is obtained for $\boldsymbol{A}$. Then

$$\alpha_j = \max_{\substack{\boldsymbol{x}\in\mathcal{S} \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|(\boldsymbol{B} + (\boldsymbol{A} - \boldsymbol{B}))\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leq \max_{\substack{\boldsymbol{x}\in\mathcal{S} \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{B}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} + \max_{\substack{\boldsymbol{x}\in\mathcal{S} \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leq \beta_j + \|\boldsymbol{A} - \boldsymbol{B}\|_2.$$

By symmetry we obtain $\beta_j \leq \alpha_j + \|\boldsymbol{A} - \boldsymbol{B}\|_2$ and the proof is complete. ∎

The following result is an analogue of Theorem 8.40.

Theorem 13.33 *Let $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{m,n}$ have singular values $\alpha_1 \geq \cdots \geq \alpha_n$ and $\epsilon_1 \geq \cdots \geq \epsilon_n$. If $\|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{E}\|_2 < 1$ then*

1. $\operatorname{rank}(\boldsymbol{A} + \boldsymbol{E}) \geq \operatorname{rank}(\boldsymbol{A})$,
2. $\|(\boldsymbol{A} + \boldsymbol{E})^\dagger\|_2 \leq \frac{\|\boldsymbol{A}^\dagger\|_2}{1 - \|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{E}\|_2} = \frac{1}{\alpha_r - \epsilon_1}$,

where r is the rank of $\boldsymbol{A}$.

Proof. Suppose $\boldsymbol{A}$ has rank r and let $\boldsymbol{B} := \boldsymbol{A} + \boldsymbol{E}$ have singular values $\beta_1 \geq \cdots \geq \beta_n$. In terms of singular values the inequality $\|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{E}\|_2 < 1$ can be written $\epsilon_1/\alpha_r < 1$ or $\alpha_r > \epsilon_1$. By Theorem 13.32 we have $\alpha_r - \beta_r \leq \epsilon_1$, which implies $\beta_r \geq \alpha_r - \epsilon_1 > 0$, and this shows that $\operatorname{rank}(\boldsymbol{A} + \boldsymbol{E} > r$. To prove 2., the inequality $\beta_r \geq \alpha_r - \epsilon_1$ implies that

$$\|(\boldsymbol{A} + \boldsymbol{E})^\dagger\|_2 \leq \frac{1}{\beta_r} \leq \frac{1}{\alpha_r - \epsilon_1} = \frac{1/\alpha_r}{1 - \epsilon_1/\alpha_r} = \frac{\|\boldsymbol{A}^\dagger\|_2}{1 - \|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{E}\|_2}$$

∎

Part V

Eigenvalues and Eigenvectors

Chapter 14

Numerical Eigenvalue Problems

In this and the next chapter we consider numerical methods for finding one or more of the eigenvalues and eigenvectors of a matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$. Maybe the first method which comes to mind is to form the characteristic polynomial $\pi_{\boldsymbol{A}}$ of $\boldsymbol{A}$, and then use a polynomial root finder, like Newton's method to determine one or several of the eigenvalues.

It turns out that this is not suitable as an all purpose method. One reason is that a small change in one of the coefficients of $\pi_{\boldsymbol{A}}(\lambda)$ can lead to a large change in the roots of the polynomial. For example, if $\pi_{\boldsymbol{A}}(\lambda :) = \lambda^{16}$ and $q(\lambda) = \lambda^{16} - 10^{-16}$ then the roots of $\pi_{\boldsymbol{A}}$ are all equal to zero, while the roots of q are $\lambda_j = 10^{-1}e^{2\pi ij/16}$, $j = 1, \ldots, 16$. The roots of q have absolute value 0.1 and a perturbation in one of the polynomial coefficients of magnitude 10^{-16} has led to an error in the roots of approximately 0.1. The situation can be somewhat remedied by representing the polynomials using a different basis.

We will see that for many matrices the eigenvalues are less sensitive to perturbations in the elements of the matrix. In this text we will only consider methods which work directly with the matrix.

14.1 Perturbation of Eigenvalues

In this section we study the following problem. Given matrices $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n,n}$, where we think of $\boldsymbol{E}$ as a pertubation of $\boldsymbol{A}$. By how much do the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{A} + \boldsymbol{E}$ differ? Not surprisingly this problem is more complicated than the corresponding problem for linear systems.

We illustrate this by considering two examples. Suppose $\boldsymbol{A}_0 := \boldsymbol{0}$ is the zero matrix. If $\lambda \in \sigma(\boldsymbol{A}_0 + \boldsymbol{E}) = \sigma(\boldsymbol{E})$, then $|\lambda| \leq \|\boldsymbol{E}\|_\infty$ by Theorem 8.46, and any zero eigenvalue of $\boldsymbol{A}_0$ is perturbed by at most $\|\boldsymbol{E}\|_\infty$. On the other hand consider

for $\epsilon > 0$ the matrices

$$\boldsymbol{A}_1 := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \boldsymbol{E} := \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \epsilon & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} = \epsilon \boldsymbol{e}_n \boldsymbol{e}_1^T.$$

The characteristic polynomial of $\boldsymbol{A}_1 + \boldsymbol{E}$ is $\pi(\lambda) := (-1)^n(\lambda^n - \epsilon)$, and the zero eigenvalues of $\boldsymbol{A}_1$ are perturbed by the amount $|\lambda| = \|\boldsymbol{E}\|_\infty^{1/n}$. Thus, for $n = 16$, a perturbation of say $\epsilon = 10^{-16}$ gives a change in eigenvalue of 0.1.

The following theorem shows that a dependence $\|\boldsymbol{E}\|_\infty^{1/n}$ is the worst that can happen.

Theorem 14.1 (Elsner's Theorem) *Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n,n}$. To every $\mu \in \sigma(\boldsymbol{A} + \boldsymbol{E})$ there is a $\lambda \in \sigma(\boldsymbol{A})$ such that*

$$|\mu - \lambda| \le \big(\|\boldsymbol{A}\|_2 + \|\boldsymbol{A} + \boldsymbol{E}\|_2\big)^{1-1/n} \|\boldsymbol{E}\|_2^{1/n}. \tag{14.1}$$

Proof. Suppose $\boldsymbol{A}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ and let λ_1 be one which is closest to μ. Let $\boldsymbol{u}_1$ with $\|\boldsymbol{u}_1\|_2 = 1$ be an eigenvector corresponding to μ, and extend $\boldsymbol{u}_1$ to an orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ of $\mathbb{C}^n$. Note that

$$\|(\mu \boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_1\|_2 = \|(\boldsymbol{A} + \boldsymbol{E})\boldsymbol{u}_1 - \boldsymbol{A}\boldsymbol{u}_1\|_2 = \|\boldsymbol{E}\boldsymbol{u}_1\|_2 \le \|\boldsymbol{E}\|_2,$$
$$\prod_{j=2}^{n} \|(\mu \boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_j\|_2 \le \prod_{j=2}^{n} (|\mu| + \|\boldsymbol{A}\boldsymbol{u}_j\|_2) \le \big(\|(\boldsymbol{A} + \boldsymbol{E})\|_2 + \|\boldsymbol{A}\|_2\big)^{n-1}.$$

Using this and Hadamard's inequality (12.1) we find

$$\begin{aligned} |\mu - \lambda_1|^n &\le \prod_{j=1}^{n} |\mu - \lambda_j| = |\det(\mu \boldsymbol{I} - \boldsymbol{A})| = |\det\big((\mu \boldsymbol{I} - \boldsymbol{A})[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]\big)| \\ &\le \|(\mu \boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_1\|_2 \prod_{j=2}^{n} \|(\mu \boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_j\|_2 \le \|\boldsymbol{E}\|_2 \big(\|(\boldsymbol{A} + \boldsymbol{E})\|_2 + \|\boldsymbol{A}\|_2\big)^{n-1}. \end{aligned}$$

The result follows by taking nth roots in this inequality. ∎

It follows from this theorem that the eigenvalues depend continuously on the elements of the matrix. The factor $\|\boldsymbol{E}\|_2^{1/n}$ shows that this dependence is almost, but not quite, differentiable. As an example, the eigenvalues of the matrix $\left[\begin{smallmatrix} 1 & 1 \\ \epsilon & 1 \end{smallmatrix}\right]$ are $1 \pm \sqrt{\epsilon}$ and this expression is not differentiable at $\epsilon = 0$.

Recall that a matrix is nondefective if the eigenvectors form a basis for $\mathbb{C}^n$. For nondefective matrices we can get rid of the annoying exponent $1/n$ in $\|\boldsymbol{E}\|_2$. The following theorem is proved in Section 14.4. For a more general discussion see [19].

Theorem 14.2 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ has linearly independent eigenvectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ be the eigenvector matrix. Suppose $\boldsymbol{E} \in \mathbb{C}^{n,n}$ and let μ be an eigenvalue of $\boldsymbol{A} + \boldsymbol{E}$. Then we can find an eigenvalue λ of $\boldsymbol{A}$ such that*

$$|\lambda - \mu| \leq K_p(\boldsymbol{X}) \|\boldsymbol{E}\|_p, \quad 1 \leq p \leq \infty, \text{ where } K_p(\boldsymbol{X}) := \|\boldsymbol{X}\|_p \, \|\boldsymbol{X}^{-1}\|_p. \tag{14.2}$$

The equation (14.2) shows that for a nondefective matrix the absolute error can be magnified by at most $K_p(\boldsymbol{X})$, the condition number of the eigenvector matrix with respect to inversion. If $K_p(\boldsymbol{X})$ is small then a small perturbation changes the eigenvalues by small amounts.

Even if we get rid of the factor $1/n$, the equation (14.2) illustrates that it can be difficult or sometimes impossible to compute accurate eigenvalues and eigenvectors of matrices with almost linearly dependent eigenvectors. On the other hand the eigenvalue problem for normal matrices is better conditioned. Indeed, if $\boldsymbol{A}$ is normal then it has a set of orthonormal eigenvectors and the eigenvector matrix is unitary. If we restrict attention to the 2-norm then $K_2(\boldsymbol{X}) = 1$ and (14.2) implies the following result.

Theorem 14.3 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is normal and let μ be an eigenvalue of $\boldsymbol{A} + \boldsymbol{E}$ for some $\boldsymbol{E} \in \mathbb{C}^{n,n}$. Then we can find an eigenvalue λ of $\boldsymbol{A}$ such that $|\lambda - \mu| \leq \|\boldsymbol{E}\|_2$.*

For an even stronger result for Hermitian matrices see Corollary 6.16. We conclude that the situation for the absolute error in an eigenvalue of a Hermitian matrix is quite satisfactory. Small perturbations in the elements are not magnified in the eigenvalues.

14.1.1 Gerschgorin's Theorem

The following theorem is useful for locating eigenvalues of an arbitrary square matrix.

Theorem 14.4 (Gerschgorin's Circle Theorem) *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$. Define for $i = 1, 2, \ldots, n$*

$$R_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\}, \quad r_i := \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,$$

$$C_j = \{z \in \mathbb{C} : |z - a_{jj}| \leq c_j\}, \quad c_j := \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|.$$

Then any eigenvalue of $\boldsymbol{A}$ lies in $R \cap C$ where $R = R_1 \cup R_2 \cup \cdots \cup R_n$ and $C = C_1 \cup C_2 \cup \cdots \cup C_n$.

Proof. Suppose $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}$. We claim that $\lambda \in R_i$, where i is such that $|x_i| = \|\boldsymbol{x}\|_\infty$. Indeed, $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$ implies that $\sum_j a_{ij} x_j = \lambda x_i$ or

$(\lambda - a_{ii})x_i = \sum_{j\neq i} a_{ij}x_j$. Dividing by x_i and taking absolute values we find

$$|\lambda - a_{ii}| = |\sum_{j\neq i} a_{ij}x_j/x_i| \leq \sum_{j\neq i} |a_{ij}||x_j/x_i| \leq r_i$$

since $|x_j/x_i| \leq 1$ for all j. Thus $\lambda \in R_i$.

Since λ is also an eigenvalue of $\boldsymbol{A}^T$, it must be in one of the row disks of $\boldsymbol{A}^T$. But these are the column disks C_j of $\boldsymbol{A}$. Hence $\lambda \in C_j$ for some j. □

The set R_i is a subset of the complex plane consisting of all points inside a circle with center at a_{ii} and radius r_i, c.f. Figure 14.1. R_i is called a (Gerschgorin) row disk.

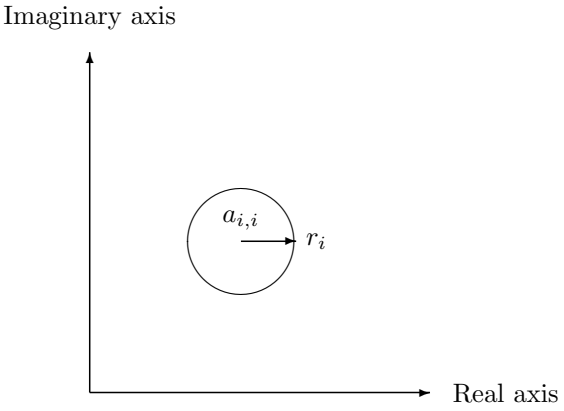


Figure 14.1. *The Gerschgorin disk R_i.*

An eigenvalue λ lies in the union of the row disks $R_1, \ldots, R_n$ and also in the union of the column disks $C_1, \ldots, C_n$. If $\boldsymbol{A}$ is Hermitian then $R_i = C_i$ for $i = 1, 2, \ldots, n$. Moreover, in this case the eigenvalues of $\boldsymbol{A}$ are real, and the Gerschgorin disks can be taken to be intervals on the real line.

Example 14.5 *Let $\boldsymbol{T} = tridiag(-1, 2, -1) \in \mathbb{R}^{m,m}$ be the second derivative matrix. Since $\boldsymbol{A}$ is Hermitian we have $R_i = C_i$ for all i and the eigenvalues are real. We find*

$$R_1 = R_m = \{z \in \mathbb{R} : |z-2| \leq 1\}, \text{ and } R_i = \{z \in \mathbb{R} : |z-2| \leq 2\}, \quad i = 2, 3, \ldots, m-1.$$

We conclude that $\lambda \in [0, 4]$ for any eigenvalue λ of $\boldsymbol{T}$. To check this, we recall that by Lemma 4.11 the eigenvalues of $\boldsymbol{T}$ are given by

$$\lambda_j = 4\left[\sin\frac{j\pi}{2(m+1)}\right]^2, \quad j = 1, 2, \ldots, m.$$

When m is large the smallest eigenvalue $4\left[\sin\frac{\pi}{2(m+1)}\right]^2$ is very close to zero and the largest eigenvalue $4\left[\sin\frac{m\pi}{2(m+1)}\right]^2$ is very close to 4. Thus Gerschgorin,s theorem gives a remarkably good estimate for large m.

Sometimes some of the Gerschgorin disks are distinct and we have

Corollary 14.6 *If p of the Gerschgorin row disks are disjoint from the others, the union of these disks contains precisely p eigenvalues. The same result holds for the column disks.*

Proof. Consider a family of matrices

$$\boldsymbol{A}(t) := \boldsymbol{D} + t(\boldsymbol{A} - \boldsymbol{D}), \quad \boldsymbol{D} := \operatorname{diag}(a_{11}, \ldots, a_{nn}), \quad t \in [0,1].$$

We have $\boldsymbol{A}(0) = \boldsymbol{D}$ and $\boldsymbol{A}(1) = \boldsymbol{A}$. As a function of t, every eigenvalue of $\boldsymbol{A}(t)$ is a continuous function of t. This follows from Theorem 14.1, see Exercise 14.7. The row disks $R_i(t)$ of $\boldsymbol{A}(t)$ have radius proportional to t, indeed

$$R_i(t) = \{z \in \mathbb{C} : |z - a_{ii}| \le t r_i\}, \quad r_i := \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

Clearly $0 \le t_1 < t_2 \le 1$ implies $R_i(t_1) \subset R_i(t_2)$ and $R_i(1)$ is a row disk of $\boldsymbol{A}$ for all i. Suppose $\bigcup_{k=1}^{p} R_{i_k}(1)$ are disjoint from the other disks of $\boldsymbol{A}$ and set $R^p(t) := \bigcup_{k=1}^{p} R_{i_k}(t)$ for $t \in [0,1]$. Now $R^p(0)$ contains only the p eigenvalues $a_{i_1,i_1}, \ldots, a_{i_p,i_p}$ of $\boldsymbol{A}(0) = \boldsymbol{D}$. As t increases from zero to one the set $R^p(t)$ is disjoint from the other row disks of $\boldsymbol{A}$ and by the continuity of the eigenvalues cannot loose or gain eigenvalues. It follows that $R^p(1)$ must contain p eigenvalues of $\boldsymbol{A}$. ∎

Exercise 14.7 *Suppose $t_1, t_2 \in [0,1]$ and that μ is an eigenvalue of $\boldsymbol{A}(t_2)$. Show, using Theorem 14.1 with $\boldsymbol{A} = \boldsymbol{A}(t_1)$ and $\boldsymbol{E} = \boldsymbol{A}(t_2) - \boldsymbol{A}(t_1)$, that $\boldsymbol{A}(t_1)$ has an eigenvalue λ such that*

$$|\lambda - \mu| \le C(t_2 - t_1)^{1/n}, \text{ where } C \le 2\big(\|\boldsymbol{D}\|_2 + \|\boldsymbol{A} - \boldsymbol{D}\|_2\big).$$

Thus, as a function of t, every eigenvalue of $\boldsymbol{A}(t)$ is a continuous function of t.

Example 14.8 *Consider the matrix $\boldsymbol{A} = \begin{bmatrix} 1 & \epsilon_1 & \epsilon_2 \\ \epsilon_3 & 2 & \epsilon_4 \\ \epsilon_5 & \epsilon_6 & 3 \end{bmatrix}$, where $|\epsilon_i| \le 10^{-15}$ all i. By Corollary 14.6 the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $\boldsymbol{A}$ are distinct and satisfy $|\lambda_j - j| \le 2 \times 10^{-15}$ for $j = 1, 2, 3$.*

Exercise 14.9 *Consider the matrix*

$$\boldsymbol{A} = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}.$$

Show using Gerschgorin's theorem that $\boldsymbol{A}$ *is nonsingular.*

Exercise 14.10 *Show using Gerschgorin,s theorem that a strictly diagonally dominant matrix* $\boldsymbol{A}$ *(*$|a_{i,i}| > \sum_{j \neq i} |a_{i,j}|$ *for all* i*) is nonsingular.*

14.2 Unitary Similarity Transformation of a Matrix into Upper Hessenberg Form

Before attempting to find eigenvalues and eigenvectors of a matrix (exceptions are made for certain sparse matrices), it is often advantageous to reduce it by similarity transformations to a simpler form. Orthogonal similarity transformations are particularly important since they are insensitive to noise in the elements of the matrix. In this section we show how this reduction can be done.

Recall that a matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is upper Hessenberg if $a_{i,j} = 0$ for $j = 1, 2, \ldots, i-2$, $i = 3, 4, \ldots, n$. We will reduce $\boldsymbol{A} \in \mathbb{R}^{n,n}$ to upper Hessenberg form by unitary similarity transformations. Let $\boldsymbol{A}_1 = \boldsymbol{A}$ and define $\boldsymbol{A}_{k+1} = \boldsymbol{H}_k \boldsymbol{A}_k \boldsymbol{H}_k$ for $k = 1, 2, \ldots, n-2$. Here $\boldsymbol{H}_k$ is a Householder transformation chosen to introduce zeros in the elements of column k of $\boldsymbol{A}_k$ under the subdiagonal. The final matrix $\boldsymbol{A}_{n-1}$ will be upper Hessenberg.

If $\boldsymbol{A}_1 = \boldsymbol{A}$ is symmetric, the matrix $\boldsymbol{A}_{n-1}$ will be symmetric and tridiagonal. For if $\boldsymbol{A}_k^T = \boldsymbol{A}_k$ then

$$\boldsymbol{A}_{k+1}^T = (\boldsymbol{H}_k \boldsymbol{A}_k \boldsymbol{H}_k)^T = \boldsymbol{H}_k \boldsymbol{A}_k^T \boldsymbol{H}_k = \boldsymbol{A}_{k+1}.$$

Since $\boldsymbol{A}_{n-1}$ is upper Hessenberg and symmetric, it must be tridiagonal.

To describe the reduction to upper Hessenberg or tridiagonal form in more detail we partition $\boldsymbol{A}_k$ as follows

$$\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{D}_k & \boldsymbol{E}_k \end{bmatrix}.$$

Suppose $\boldsymbol{B}_k \in \mathbb{R}^{k,k}$ is upper Hessenberg, and the first $k-1$ columns of $\boldsymbol{D}_k \in \mathbb{R}^{n-k,k}$ are zero, i.e. $\boldsymbol{D}_k = [\boldsymbol{0}, \boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{d}_k]$. Let $\boldsymbol{V}_k = \boldsymbol{I} - \boldsymbol{v}_k \boldsymbol{v}_k^T \in \mathbb{R}^{n-k,n-k}$ be a Householder transformation such that $\boldsymbol{V}_k \boldsymbol{d}_k = \alpha_k \boldsymbol{e}_1$, where $\alpha_k^2 = \boldsymbol{d}_k^T \boldsymbol{d}_k$. Define

$$\boldsymbol{H}_k = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_k \end{bmatrix} \in \mathbb{R}^{n,n}.$$

The matrix $\boldsymbol{H}_k$ is a Householder transformation, and we find

$$\boldsymbol{A}_{k+1} = \boldsymbol{H}_k \boldsymbol{A}_k \boldsymbol{H}_k = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{D}_k & \boldsymbol{E}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_k \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \boldsymbol{V}_k \\ \boldsymbol{V}_k \boldsymbol{D}_k & \boldsymbol{V}_k \boldsymbol{E}_k \boldsymbol{V}_k \end{bmatrix}.$$

Now $\boldsymbol{V}_k \boldsymbol{D}_k = [\boldsymbol{V}_k \boldsymbol{0}, \ldots, \boldsymbol{V}_k \boldsymbol{0}, \boldsymbol{V}_k \boldsymbol{d}_k] = (\boldsymbol{0}, \ldots, \boldsymbol{0}, \alpha_k \boldsymbol{e}_1)$. Moreover, the matrix $\boldsymbol{B}_k$ is not affected by the $\boldsymbol{H}_k$ transformation. Therefore the upper left $(k+1) \times (k+1)$ corner of $\boldsymbol{A}_{k+1}$ is upper Hessenberg and the reduction is carried one step further. The reduction stops with $\boldsymbol{A}_{n-1}$ which is upper Hessenberg.

To find $\boldsymbol{A}_{k+1}$ we use Algorithm 12.13 to find $\boldsymbol{v}_k$ and α_k. We store $\boldsymbol{v}_k$ in the kth column of a matrix $\boldsymbol{L}$ as $\boldsymbol{L}(k+1:n, k) = \boldsymbol{v}_k$. This leads to the following algorithm.

Algorithm 14.11 (Householder reduction to Hessenberg form) This algorithm uses Householder similarity transformations to reduce a matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ to upper Hessenberg form. The reduced matrix $\boldsymbol{B}$ is tridiagonal if $\boldsymbol{A}$ is symmetric. Details of the transformations are stored in a lower triangular matrix $\boldsymbol{L}$. The elements of $\boldsymbol{L}$ can be used to assemble an orthonormal matrix $\boldsymbol{Q}$ such that $\boldsymbol{B} = \boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q}$. Algorithm 12.13 is used in each step of the reduction.

```
function [L,B] = hesshousegen(A)
n=length(A); L=zeros(n,n); B=A;
for k=1:n-2
    [v,B(k+1,k)]=housegen(B(k+1:n,k));
    L(k+1:n,k)=v; B(k+2:n,k)=zeros(n-k-1,1);
    C=B(k+1:n,k+1:n); B(k+1:n,k+1:n)=C-v*(v'*C);
    C=B(1:n,k+1:n); B(1:n,k+1:n)=C-(C*v)*v';
end
```

Exercise 14.12 *Show that the number of flops for Algorithm 14.11 is* $O(\frac{10}{3}n^3)$.

We can use the output of Algorithm 14.11 to assemble the matrix $\boldsymbol{Q} \in \mathbb{R}^{n,n}$ such that $\boldsymbol{Q}$ is orthonormal and $\boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q}$ is upper Hessenberg. We need to compute the product $\boldsymbol{Q} = \boldsymbol{H}_1 \boldsymbol{H}_2 \cdots \boldsymbol{H}_{n-2}$, where $\boldsymbol{H}_k = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} - \boldsymbol{v}_k \boldsymbol{v}_k^T \end{bmatrix}$ and $\boldsymbol{v}_k \in \mathbb{R}^{n-k}$. Since $\boldsymbol{v}_1 \in \mathbb{R}^{n-1}$ and $\boldsymbol{v}_{n-2} \in \mathbb{R}^2$ it is most economical to assemble the product from right to left. We compute

$$\boldsymbol{Q}_{n-1} = \boldsymbol{I} \text{ and } \boldsymbol{Q}_k = \boldsymbol{H}_k \boldsymbol{Q}_{k+1} \text{ for } k = n-2, n-3, \ldots, 1.$$

Suppose $\boldsymbol{Q}_{k+1}$ has the form $\left[\begin{smallmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_k \end{smallmatrix}\right]$, where $\boldsymbol{U}_k \in \mathbb{R}^{n-k,n-k}$. Then

$$\boldsymbol{Q}_k = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} - \boldsymbol{v}_k \boldsymbol{v}_k^T \end{bmatrix} * \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_k - \boldsymbol{v}_k (\boldsymbol{v}_k^T \boldsymbol{U}_k) \end{bmatrix}.$$

This leads to the following algorithm.

Algorithm 14.13 (Assemble Householder transformations) Suppose `[L,B] = hesshousegen(A)` is the output of Algorithm 14.11. This algorithm assembles an orthonormal matrix $\boldsymbol{Q}$ from the columns of $\boldsymbol{L}$ such that $\boldsymbol{B} = \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ is upper Hessenberg.

```
function Q = accumulateQ(L)
n=length(L); Q=eye(n);
for k=n-2:-1:1
    v=L(k+1:n,k); C=Q(k+1:n,k+1:n);
    Q(k+1:n,k+1:n)=C-v*(v'*C);
end
```

Exercise 14.14 *Show that the number of flops required by Algorithm 14.13 is* $O(\frac{4}{3}n^3)$.

Exercise 14.15 *If* $\boldsymbol{A}$ *is symmetric we can modify Algorithm 14.11 as follows. To find* $\boldsymbol{A}_{k+1}$ *from* $\boldsymbol{A}_k$ *we have to compute* $\boldsymbol{V}_k\boldsymbol{E}_k\boldsymbol{V}_k$ *where* $\boldsymbol{E}_k$ *is symmetric. Dropping subscripts we have to compute a product of the form* $\boldsymbol{G} = (\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{E}(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T)$. *Let* $\boldsymbol{w} := \boldsymbol{E}\boldsymbol{v}$, $\beta := \frac{1}{2}\boldsymbol{v}^T\boldsymbol{w}$ *and* $\boldsymbol{z} := \boldsymbol{w} - \beta\boldsymbol{v}$. *Show that* $\boldsymbol{G} = \boldsymbol{E} - \boldsymbol{v}\boldsymbol{z}^T - \boldsymbol{z}\boldsymbol{v}^T$. *Since* $\boldsymbol{G}$ *is symmetric, only the sub- or superdiagonal elements of* $\boldsymbol{G}$ *need to be computed. Computing* $\boldsymbol{G}$ *in this way, it can be shown that we need* $O(4n^3/3)$ *operations to tridiagonalize a symmetric matrix by orthonormal similarity transformations. This is less than half the work to reduce a nonsymmetric matrix to upper Hessenberg form. We refer to [18] for a detailed algorithm.*

14.3 Computing a Selected Eigenvalue of a Symmetric Matrix

Let $\boldsymbol{A} \in \mathbb{R}^{n,n}$ be symmetric with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. In this section we consider a method to compute an approximation to the mth eigenvalue λ_m for some $1 \leq m \leq n$. Using Householder similarity transformations as outlined in the previous section we can assume that $\boldsymbol{A}$ is symmetric and tridiagonal.

$$\boldsymbol{A} = \begin{bmatrix} d_1 & c_1 & & & \\ c_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-2} & d_{n-1} & c_{n-1} \\ & & & c_{n-1} & d_n \end{bmatrix}. \tag{14.3}$$

Suppose one of the off-diagonal elements is equal to zero, say $c_i = 0$. We then have $\boldsymbol{A} = \left[\begin{smallmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{smallmatrix}\right]$, where

$$\boldsymbol{A}_1 = \begin{bmatrix} d_1 & c_1 & & & \\ c_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{i-2} & d_{i-1} & c_{i-1} \\ & & & c_{i-1} & d_i \end{bmatrix} \text{ and } \boldsymbol{A}_2 = \begin{bmatrix} d_{i+1} & c_{i+1} & & & \\ c_{i+1} & d_{i+2} & c_{i+2} & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-2} & d_{n-1} & c_{n-1} \\ & & & c_{n-1} & d_n \end{bmatrix}.$$

Thus $\boldsymbol{A}$ is block diagonal and each diagonal block is tridiagonal. By 6. of Theorem D.3 we can split the eigenvalue problem into two smaller problems involving $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. We assume that this reduction has been carried out so that $\boldsymbol{A}$ is irreducible, i. e., $c_i \neq 0$ for $i = 1, \ldots, n-1$.

We first show that irreducibility implies that the eigenvalues are distinct.

Lemma 14.16 *An irreducible, tridiagonal and symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ has n real and distinct eigenvalues.*

Proof. Let $\boldsymbol{A}$ be given by (14.3). By Theorem 6.5 the eigenvalues are real. Define for $x \in \mathbb{R}$ the polynomial $p_k(x) := \det(x\boldsymbol{I}_k - \boldsymbol{A}_k)$ for $k = 1, \ldots, n$, where $\boldsymbol{A}_k$ is the upper left $k \times k$ corner of $\boldsymbol{A}$ (the leading principal submatrix of order k). The eigenvalues of $\boldsymbol{A}$ are the roots of the polynomial p_n. Using the last column to expand for $k \geq 2$ the determinant $p_{k+1}(x)$ we find

$$p_{k+1}(x) = (x - d_{k+1})p_k(x) - c_k^2 p_{k-1}(x). \tag{14.4}$$

Since $p_1(x) = x - d_1$ and $p_2(x) = (x - d_2)(x - d_1) - c_1^2$ this also holds for $k = 0, 1$ if we define $p_{-1}(x) = 0$ and $p_0(x) = 1$. For M sufficiently large we have

$$p_2(-M) > 0, \quad p_2(d_1) < 0, \quad p_2(+M) > 0.$$

Since p_2 is continuous there are $y_1 \in (-M, d_1)$ and $y_2 \in (d_1, M)$ such that $p_2(y_1) = p_2(y_2) = 0$. It follows that the root d_1 of p_1 separates the roots of p_2, so y_1 and y_2 must be distinct. Consider next

$$p_3(x) = (x - d_3)p_2(x) - c_2^2 p_1(x) = (x - d_3)(x - y_1)(x - y_2) - c_2^2(x - d_1).$$

Since $y_1 < d_1 < y_2$ we have for M sufficiently large

$$p_3(-M) < 0, \quad p_3(y_1) > 0, \quad p_3(y_2) < 0, \quad p_3(+M) > 0.$$

Thus the roots x_1, x_2, x_3 of p_3 are separated by the roots y_1, y_2 of p_2. In the general case suppose for $k \geq 2$ that the roots $z_1, \ldots, z_{k-1}$ of p_{k-1} separate the roots $y_1, \ldots, y_k$ of p_k. Choose M so that $y_0 := -M < y_1$, $y_{k+1} := M > y_k$. Then

$$y_0 < y_1 < z_1 < y_2 < z_2 \cdots < z_{k-1} < y_k < y_{k+1}.$$

We claim that for M sufficiently large

$$p_{k+1}(y_j) = (-1)^{k+1-j}|p_{k+1}(y_j)| \neq 0, \text{ for } j = 0, 1, \ldots, k+1.$$

This holds for $j = 0, k+1$, and for $j = 1, \ldots, k$ since

$$p_{k+1}(y_j) = -c_k^2 p_{k-1}(y_j) = -c_k^2 (y_j - z_1) \cdots (y_j - z_{k-1}).$$

It follows that the roots $x_1, \ldots, x_{k+1}$ are separated by the roots $y_1, \ldots, y_k$ of p_k and by induction the roots of p_n (the eigenvalues of $\boldsymbol{A}$) are distinct. □

14.3.1 The Inertia Theorem

We say that two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$ are **congruent** if $\boldsymbol{A} = \boldsymbol{E}^H \boldsymbol{B} \boldsymbol{E}$ for some nonsingular matrix $\boldsymbol{E} \in \mathbb{C}^{n,n}$. By Theorem 6.10 a Hermitian matrix $\boldsymbol{A}$ is both congruent and similar to a diagonal matrix $\boldsymbol{D}$, $\boldsymbol{U}^H \boldsymbol{A} \boldsymbol{U} = \boldsymbol{D}$ where $\boldsymbol{U}$ is unitary. The eigenvalues of $\boldsymbol{A}$ are the diagonal elements of $\boldsymbol{D}$. Let $\pi(\boldsymbol{A})$, $\zeta(\boldsymbol{A})$ and $\upsilon(\boldsymbol{A})$ denote the number of positive, zero and negative eigenvalues of $\boldsymbol{A}$. If $\boldsymbol{A}$ is Hermitian then all eigenvalues are real and $\pi(\boldsymbol{A}) + \zeta(\boldsymbol{A}) + \upsilon(\boldsymbol{A}) = n$.

Theorem 14.17 (Sylvester's Inertia Theorem) *If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$ are Hermitian and congruent then $\pi(\boldsymbol{A}) = \pi(\boldsymbol{B})$, $\zeta(\boldsymbol{A}) = \zeta(\boldsymbol{B})$ and $\upsilon(\boldsymbol{A}) = \upsilon(\boldsymbol{B})$.*

Proof. Suppose $\boldsymbol{A} = \boldsymbol{E}^H \boldsymbol{B} \boldsymbol{E}$, where $\boldsymbol{E}$ is nonsingular. Assume first that $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal matrices. Suppose $\pi(\boldsymbol{A}) = k$ and $\pi(\boldsymbol{B}) = m < k$. We shall show that this leads to a contradiction. Let $\boldsymbol{E}_1$ be the upper left $m \times k$ corner of $\boldsymbol{E}$. Since $m < k$, we can find a nonzero $\boldsymbol{x}$ such that $\boldsymbol{E}_1 \boldsymbol{x} = \boldsymbol{0}$ (cf. Lemma B.5). Let $\boldsymbol{y}^T = [\boldsymbol{x}^T, \boldsymbol{0}^T] \in \mathbb{C}^n$, and $\boldsymbol{z} = [z_1, \ldots, z_n]^T = \boldsymbol{E}\boldsymbol{y}$. Then $z_i = 0$ for $i = 1, 2, \ldots, m$. If $\boldsymbol{A}$ has positive eigenvalues $\lambda_1, \ldots, \lambda_k$ and $\boldsymbol{B}$ has eigenvalues $\mu_1, \ldots, \mu_n$, where $\mu_i \leq 0$ for $i \geq m+1$ then

$$\boldsymbol{y}^H \boldsymbol{A} \boldsymbol{y} = \sum_{i=1}^{n} \lambda_i |y_i|^2 = \sum_{i=1}^{k} \lambda_i |x_i|^2 > 0.$$

But

$$\boldsymbol{y}^H \boldsymbol{A} \boldsymbol{y} = \boldsymbol{y}^H \boldsymbol{E}^H \boldsymbol{B} \boldsymbol{E} \boldsymbol{y} = \boldsymbol{z}^H \boldsymbol{B} \boldsymbol{z} = \sum_{i=m+1}^{n} \mu_i |z_i|^2 \leq 0,$$

a contradiction.

We conclude that $\pi(\boldsymbol{A}) = \pi(\boldsymbol{B})$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal. Moreover, $\upsilon(\boldsymbol{A}) = \pi(-\boldsymbol{A}) = \pi(-\boldsymbol{B}) = \upsilon(\boldsymbol{B})$ and $\zeta(\boldsymbol{A}) = n - \pi(\boldsymbol{A}) - \upsilon(\boldsymbol{A}) = n - \pi(\boldsymbol{B}) - \upsilon(\boldsymbol{B}) = \zeta(\boldsymbol{B})$. This completes the proof for diagonal matrices.

Let in the general case $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ be unitary matrices such that $\boldsymbol{U}_1^H \boldsymbol{A} \boldsymbol{U}_1 = \boldsymbol{D}_1$ and $\boldsymbol{U}_2^H \boldsymbol{B} \boldsymbol{U}_2 = \boldsymbol{D}_2$ where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices. Since $\boldsymbol{A} = \boldsymbol{E}^H \boldsymbol{B} \boldsymbol{E}$, we find $\boldsymbol{D}_1 = \boldsymbol{F}^H \boldsymbol{D}_2 \boldsymbol{F}$ where $\boldsymbol{F} = \boldsymbol{U}_2^H \boldsymbol{E} \boldsymbol{U}_1$ is nonsingular. Thus $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are congruent diagonal matrices. But since $\boldsymbol{A}$ and $\boldsymbol{D}_1$, $\boldsymbol{B}$ and $\boldsymbol{D}_2$ have the same eigenvalues, we find $\pi(\boldsymbol{A}) = \pi(\boldsymbol{D}_1) = \pi(\boldsymbol{D}_2) = \pi(\boldsymbol{B})$. Similar results hold for ζ and υ. □

Corollary 14.18 *Suppose* $\boldsymbol{A} = \text{tridiag}(c_i, d_i, c_i) \in \mathbb{R}^{n,n}$ *is symmetric and that* $\alpha \in \mathbb{R}$ *is such that* $\boldsymbol{A} - \alpha\boldsymbol{I}$ *has an* LDL^T *factorization, i.e.* $\boldsymbol{A} - \alpha\boldsymbol{I} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ *where* $\boldsymbol{L}$ *is unit lower triangular and* $\boldsymbol{D}$ *is diagonal. Then the number of eigenvalues of* $\boldsymbol{A}$ *strictly less than* α *equals the number of negative diagonal elements in* $\boldsymbol{D}$*. The diagonal elements* $d_1(\alpha), \ldots, d_n(\alpha)$ *in* $\boldsymbol{D}$ *can be computed recursively as follows*

$$d_1(\alpha) = d_1 - \alpha, \;\; d_k(\alpha) = d_k - \alpha - c_{k-1}^2/d_{k-1}(\alpha), \;\; k = 2, 3, \ldots, n. \tag{14.5}$$

Proof. Since the diagonal elements in $\boldsymbol{R}$ in an LU factorization equal the diagonal elements in $\boldsymbol{D}$ in an $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ factorization we see that the formulas in (14.5) follows immediately from (2.5). Since $\boldsymbol{L}$ is nonsingular, $\boldsymbol{A} - \alpha\boldsymbol{I}$ and $\boldsymbol{D}$ are congruent. By the previous theorem $v(\boldsymbol{A} - \alpha\boldsymbol{I}) = v(\boldsymbol{D})$, the number of negative diagonal elements in $\boldsymbol{D}$. If $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ then $(\boldsymbol{A} - \alpha\boldsymbol{I})\boldsymbol{x} = (\lambda - \alpha)\boldsymbol{x}$, and $\lambda - \alpha$ is an eigenvalue of $\boldsymbol{A} - \alpha\boldsymbol{I}$. But then $v(\boldsymbol{A} - \alpha\boldsymbol{I})$ equals the number of eigenvalues of $\boldsymbol{A}$ which are less than α. □

Exercise 14.19 *Consider the matrix in Exercise 14.9. Determine the number of eigenvalues greater than 4.5.*

Exercise 14.20 *Let for* $n \in \mathbb{N}$

$$\boldsymbol{A}_n = \begin{bmatrix} 10 & 1 & 0 & \cdots & 0 \\ 1 & 10 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 10 & 1 \\ 0 & \cdots & 0 & 1 & 10 \end{bmatrix}.$$

a) *Let* d_k *be the diagonal elements of* $\boldsymbol{D}$ *in an* LDL^T *factorization of* $\boldsymbol{A}_n$*. Show that* $5 + \sqrt{24} < d_k \leq 10$, $k = 1, 2, \ldots, n$.

b) *Show that* $D_n = \det(\boldsymbol{A}_n) > (5 + \sqrt{24})^n$*. Give* $n_0 \in \mathbb{N}$ *such that your computer gives an overflow when* D_{n_0} *is computed in floating point arithmetic.*

Exercise 14.21 *(Simultaneous diagonalization of two symmetric matrices by a congruence transformation). Let* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n,n}$ *where* $\boldsymbol{A}^T = \boldsymbol{A}$ *and* $\boldsymbol{B}$ *is symmetric positive definite. Let* $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{D}\boldsymbol{U}$ *where* $\boldsymbol{U}$ *is orthonormal and* $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_n)$*. Let* $\hat{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^T\boldsymbol{D}^{-1/2}$ *where* $\boldsymbol{D}^{-1/2} = \text{diag}(d_1^{-1/2}, \ldots, d_n^{-1/2})$.

a) *Show that* $\hat{\boldsymbol{A}}$ *is symmetric. Let* $\hat{\boldsymbol{A}} = \hat{\boldsymbol{U}}^T\hat{\boldsymbol{D}}\hat{\boldsymbol{U}}$ *where* $\hat{\boldsymbol{U}}$ *is orthonormal and* $\hat{\boldsymbol{D}}$ *is diagonal. Set* $\boldsymbol{E} = \boldsymbol{U}^T\boldsymbol{D}^{-1/2}\hat{\boldsymbol{U}}^T$.

b) *Show that* $\boldsymbol{E}$ *is nonsingular and that* $\boldsymbol{E}^T\boldsymbol{A}\boldsymbol{E} = \hat{\boldsymbol{D}}$, $\boldsymbol{E}^T\boldsymbol{B}\boldsymbol{E} = \boldsymbol{I}$.

14.3.2 Approximating λ_m

Corollary 14.18 can be used to determine the mth eigenvalue of $\boldsymbol{A}$, where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. For this we use interval bisection. Using Gerschgorin's theorem we first find an interval $[a, b]$, such that $[a, b)$ contains the eigenvalues of $\boldsymbol{A}$. Let for $x \in [a, b]$

$$\rho(x) := \#\{k : d_k(x) < 0 \text{ for } k = 1, \ldots, n\}$$

be the number of eigenvalues of $\boldsymbol{A}$ which are strictly less than x. Clearly $\rho(a) = 0$, $\rho(b) = n$ and $\rho(e) - \rho(d)$ is the number of eigenvalues in $[d, e)$. Let $c = (a + b)/2$ and $k := \rho(c)$. If $k \geq m$ then $\lambda_m \leq c$ and $\lambda_m \in [a, c]$, while if $k < m$ then $\lambda_m \geq c$ and $\lambda_m \in [c, b]$. Continuing with the interval containing λ_m we generate a sequence $\{[a_j, b_j]\}$ of intervals, each containing λ_m and $b_j - a_j = 2^{-j}(b - a)$.

As it stands this method will fail if in (14.5) one of the $d_k(\alpha)$ is zero. One possibility is to replace such a $d_k(\alpha)$ by a suitable small number, say $\delta_k = \pm|c_k|\epsilon_M$, where the negative sign is used if $c_k < 0$, and ϵ_M is the Machine epsilon, typically 2×10^{-16} for Matlab. This replacement is done if $|d_k(\alpha)| < |\delta_k|$.

Exercise 14.22 *Suppose $\boldsymbol{A} = tridiag(\boldsymbol{c}, \boldsymbol{d}, \boldsymbol{c})$ is symmetric and tridiagonal with elements $d_1, \ldots, d_n$ on the diagonal and $c_1, \ldots, c_{n-1}$ on the neighboring subdiagonals. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of $\boldsymbol{A}$. We shall write a program to compute one eigenvalue λ_m for a given m using bisection and the method outlined in Section 14.3.2.*

a) *Write a* `function k=count(c,d,x)` *which for given x counts the number of eigenvalues of $\boldsymbol{A}$ strictly less than x. Use the replacement described above if one of the $d_j(x)$ is close to zero.*

b) *Write a* `function lambda=findeigv(c,d,m)` *which first estimates an interval $[a, b]$ containing all eigenvalues of $\boldsymbol{A}$ and then generates a sequence $\{[a_k, b_k]\}$ of intervals each containing λ_m. Iterate until $b_k - a_k \leq (b - a)\epsilon_M$, where ϵ_M is Matlab's machine epsilon* `eps`*. Typically $\epsilon_M \approx 2.22 \times 10^{-16}$.*

c) *Test the program on $\boldsymbol{T} := tridiag(-1, 2, -1)$ of size 100. Compare the exact value of λ_5 with your result and the result obtained by using Matlab's built-in function* `eig`*.*

Exercise 14.23 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is upper Hessenberg and $x \in \mathbb{C}$. We will study two algorithms to compute $f(x) = \det(\boldsymbol{A} - x\boldsymbol{I})$.*

a) *Show that Gaussian elimination without pivoting requires $O(1 * n^2)$ flops.*

b) *Show that the number of flops is the same if partial pivoting is used.*

c) *Estimate the number of flops if Given's rotations are used.*

d) *Compare the two methods discussing advantages and disadvantages.*

14.4 Perturbation Proofs

We first show that the p-norm of a diagonal matrix is equal to its spectral radius.

Lemma 14.24 *If $\boldsymbol{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix then $\|\boldsymbol{A}\|_p = \rho(\boldsymbol{A})$ for $1 \le p \le \infty$.*

Proof. For any $\boldsymbol{x} \in \mathbb{C}^n$ and $p < \infty$ we have

$$\|\boldsymbol{A}\boldsymbol{x}\|_p = \|[\lambda_1 x_1, \ldots, \lambda_n x_n]^T\|_p = \big(\sum_{j=1}^{n} |\lambda_j|^p |x_j|^p\big)^{1/p} \le \rho(\boldsymbol{A})\|\boldsymbol{x}\|_p.$$

Thus $\|\boldsymbol{A}\|_p = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} \le \rho(\boldsymbol{A})$. But from Theorem 8.46 we have $\rho(\boldsymbol{A}) \le \|\boldsymbol{A}\|_p$ and the proof is complete for $p < \infty$. ∎

Exercise 14.25 *Give a direct proof that $\|\boldsymbol{A}\|_\infty = \rho(\boldsymbol{A})$ if $\boldsymbol{A}$ is diagonal.*

Suppose now $(\mu, \boldsymbol{x})$ is an approximation to an eigenpair of a matrix $\boldsymbol{A}$. One way to check the accuracy is to compute the residual $\boldsymbol{r} := \boldsymbol{A}\boldsymbol{x} - \mu\boldsymbol{x}$. For an exact eigenpair the residual is zero and we could hope that a small residual implies an accurate eigenpair.

Theorem 14.26 (Absolute errors) *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ has linearly independent eigenvectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ be the eigenvector matrix. To any $\mu \in \mathbb{C}$ and $\boldsymbol{x} \in \mathbb{C}^n$ with $\|\boldsymbol{x}\|_p = 1$ we can find an eigenvalue λ of $\boldsymbol{A}$ such that*

$$|\lambda - \mu| \le K_p(\boldsymbol{X})\|\boldsymbol{r}\|_p, \quad 1 \le p \le \infty, \tag{14.6}$$

where $\boldsymbol{r} := \boldsymbol{A}\boldsymbol{x} - \mu\boldsymbol{x}$ and $K_p(\boldsymbol{X}) := \|\boldsymbol{X}\|_p \, \|\boldsymbol{X}^{-1}\|_p$. If for some $\boldsymbol{E} \in \mathbb{C}^{n,n}$ it holds that $(\mu, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A} + \boldsymbol{E}$, then we can find an eigenvalue λ of $\boldsymbol{A}$ such that

$$|\lambda - \mu| \le K_p(\boldsymbol{X})\|\boldsymbol{E}\|_p, \quad 1 \le p \le \infty, \tag{14.7}$$

Proof. If $\mu \in \sigma(\boldsymbol{A})$ then we can take $\lambda = \mu$ and (14.6), (14.7) hold trivially. So assume $\mu \notin \sigma(\boldsymbol{A})$. Since $\boldsymbol{A}$ is nondefective it can be diagonalized, we have $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{-1}$, where $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $(\lambda_j, \boldsymbol{x}_j)$ are the eigenpairs of $\boldsymbol{A}$ for $j = 1, \ldots, n$. Define $\boldsymbol{D}_1 := \boldsymbol{D} - \mu\boldsymbol{I}$. Then $\boldsymbol{D}_1^{-1} = \operatorname{diag}\big((\lambda_1 - \mu)^{-1}, \ldots, (\lambda_n - \mu)^{-1}\big)$ exists and

$$\boldsymbol{X}\boldsymbol{D}_1^{-1}\boldsymbol{X}^{-1}\boldsymbol{r} = \big(\boldsymbol{X}(\boldsymbol{D} - \mu\boldsymbol{I})\boldsymbol{X}^{-1}\big)^{-1}\boldsymbol{r} = (\boldsymbol{A} - \mu\boldsymbol{I})^{-1}(\boldsymbol{A} - \mu\boldsymbol{I})\boldsymbol{x} = \boldsymbol{x}.$$

Using this and Lemma 14.24 we obtain

$$1 = \|\boldsymbol{x}\|_p = \|\boldsymbol{X}\boldsymbol{D}_1^{-1}\boldsymbol{X}^{-1}\boldsymbol{r}\|_p \le \|\boldsymbol{D}_1^{-1}\|_p K_p(\boldsymbol{X})\|\boldsymbol{r}\|_p = \frac{K_p(\boldsymbol{X})\|\boldsymbol{r}\|_p}{\min_j |\lambda_j - \mu|}.$$

But then (14.6) follows. If $(\boldsymbol{A} + \boldsymbol{E})\boldsymbol{x} = \mu\boldsymbol{x}$ then $\boldsymbol{0} = \boldsymbol{A}\boldsymbol{x} - \mu\boldsymbol{x} + \boldsymbol{E}\boldsymbol{x} = \boldsymbol{r} + \boldsymbol{E}\boldsymbol{x}$. But then $\|\boldsymbol{r}\|_p = \|-\boldsymbol{E}\boldsymbol{x}\|_p \le \|\boldsymbol{E}\|_p$. Inserting this in (14.6) proves (14.7). ∎

For the accuracy of an eigenvalue of small magnitude we are interested in the size of the relative error.

Theorem 14.27 (Relative errors) *Suppose in Theorem 14.26 that $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular. To any $\mu \in \mathbb{C}$ and $\boldsymbol{x} \in \mathbb{C}^n$ with $\|\boldsymbol{x}\|_p = 1$, we can find an eigenvalue λ of $\boldsymbol{A}$ such that*

$$\frac{|\lambda - \mu|}{|\lambda|} \leq K_p(\boldsymbol{X}) K_p(\boldsymbol{A}) \frac{\|\boldsymbol{r}\|_p}{\|\boldsymbol{A}\|_p}, \quad 1 \leq p \leq \infty, \tag{14.8}$$

where $\boldsymbol{r} := \boldsymbol{A}\boldsymbol{x} - \mu\boldsymbol{x}$. If for some $\boldsymbol{E} \in \mathbb{C}^{n,n}$ it holds that $(\mu, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A} + \boldsymbol{E}$, then we can find an eigenvalue λ of $\boldsymbol{A}$ such that

$$\frac{|\lambda - \mu|}{|\lambda|} \leq K_p(\boldsymbol{X}) \|\boldsymbol{A}^{-1}\boldsymbol{E}\|_p \leq K_p(\boldsymbol{X}) K_p(\boldsymbol{A}) \frac{\|\boldsymbol{E}\|_p}{\|\boldsymbol{A}\|_p}, \quad 1 \leq p \leq \infty, \tag{14.9}$$

Proof. Applying Theorem 8.46 to $\boldsymbol{A}^{-1}$ we have for any $\lambda \in \sigma(\boldsymbol{A})$

$$\frac{1}{\lambda} \leq \|\boldsymbol{A}^{-1}\|_p = \frac{K_p(\boldsymbol{A})}{\|\boldsymbol{A}\|_p}$$

and (14.8) follows from (14.6). To prove (14.9) we define the matrices $\boldsymbol{B} := \mu\boldsymbol{A}^{-1}$ and $\boldsymbol{F} := -\boldsymbol{A}^{-1}\boldsymbol{E}$. If $(\lambda_j, \boldsymbol{x})$ are the eigenpairs for $\boldsymbol{A}$ then $\left(\frac{\mu}{\lambda_j}, \boldsymbol{x}\right)$ are the eigenpairs for $\boldsymbol{B}$ for $j = 1, \ldots, n$. Since $(\mu, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A} + \boldsymbol{E}$ we find

$$(\boldsymbol{B} + \boldsymbol{F} - \boldsymbol{I})\boldsymbol{x} = (\mu\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{E} - \boldsymbol{I})\boldsymbol{x} = \boldsymbol{A}^{-1}\big(\mu\boldsymbol{I} - (\boldsymbol{E} + \boldsymbol{A})\big)\boldsymbol{x} = \boldsymbol{0}.$$

Thus $(1, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{B} + \boldsymbol{F}$. Applying Theorem 14.26 to this eigenvalue we can find $\lambda \in \sigma(\boldsymbol{A})$ such that $|\frac{\mu}{\lambda} - 1| \leq K_p(\boldsymbol{X})\|\boldsymbol{F}\|_p = K_p(\boldsymbol{X})\|\boldsymbol{A}^{-1}\boldsymbol{E}\|_p$ which proves the first estimate in (14.9). The second inequality in (14.9) follows from the submultiplicativity of the p-norm. ∎

Chapter 15

Some Methods for Computing Eigenvalues

15.1 The Power Method

Let $\boldsymbol{A} \in \mathbb{C}^{n,n}$ have eigenpairs $(\lambda_j, \boldsymbol{v}_j)$, $j = 1, \ldots, n$. Given $\boldsymbol{z}_0 \in \mathbb{C}^n$ we assume that

$$\begin{aligned} &(i) \quad |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|, \\ &(ii) \quad \boldsymbol{z}_0^T \boldsymbol{v}_1 \neq 0 \\ &(iii) \quad \boldsymbol{A} \text{ has linearly independent eigenvectors.} \end{aligned} \tag{15.1}$$

The first assumption means that $\boldsymbol{A}$ has a dominant eigenvalue λ_1 of algebraic multiplicity one. The second assumption says that $\boldsymbol{z}_0$ has a component in the direction $\boldsymbol{v}_1$. The third assumption is not necessary, but is included in order to simply the analysis.

The **power method** is a technique to compute the dominant eigenvector $\boldsymbol{v}_1$ of $\boldsymbol{A}$. As a by product we can also find the corresponding eigenvalue. We define a sequence $\{\boldsymbol{z}_k\}$ of vectors in $\mathbb{C}^n$ by

$$\boldsymbol{z}_k := \boldsymbol{A}^k \boldsymbol{z}_0 = \boldsymbol{A}\boldsymbol{z}_{k-1}, \quad k = 1, 2, \ldots. \tag{15.2}$$

To see what happens let $\boldsymbol{z}_0 = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n$, where by assumption (ii) of (15.1) we have $c_1 \neq 0$. Since $\boldsymbol{A}^k\boldsymbol{v}_j = \lambda_j^k\boldsymbol{v}_j$ for all j we see that

$$\boldsymbol{z}_k = c_1\lambda_1^k\boldsymbol{v}_1 + c_2\lambda_2^k\boldsymbol{v}_2 + \cdots + c_n\lambda_n^k\boldsymbol{v}_n, \quad k = 0, 1, 2, \ldots. \tag{15.3}$$

Dividing by λ_1^k we find

$$\frac{\boldsymbol{z}_k}{\lambda_1^k} = c_1\boldsymbol{v}_1 + c_2\Big(\frac{\lambda_2}{\lambda_1}\Big)^k\boldsymbol{v}_2 + \cdots + c_n\Big(\frac{\lambda_n}{\lambda_1}\Big)^k\boldsymbol{v}_n, \quad k = 0, 1, 2, \ldots. \tag{15.4}$$

Assumption (i) of (15.1) implies that $(\lambda_j/\lambda_1)^k \to 0$ as $k \to \infty$ for all $j \geq 2$ and we obtain

$$\lim_{k\to\infty} \frac{\boldsymbol{z}_k}{\lambda_1^k} = c_1\boldsymbol{v}_1, \tag{15.5}$$

the dominant eigenvector of $\boldsymbol{A}$. It can be shown that this also holds for defective matrices as long as (i) and (ii) of (15.1) hold, see for example page 58 of [18].

In practice we need to scale the iterates $\boldsymbol{z}_k$ somehow and we normally do not know λ_1. Instead we choose a norm on $\mathbb{C}^n$, set $\boldsymbol{x}_0 = \boldsymbol{z}_0/\|\boldsymbol{z}_0\|$ and generate for $k = 1, 2, \ldots$ unit vectors as follows:

$$\begin{array}{ll} (i) & \boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} \\ (ii) & \boldsymbol{x}_k = \boldsymbol{y}_k/\|\boldsymbol{y}_k\|. \end{array} \tag{15.6}$$

Lemma 15.1 *Suppose* (15.1) *holds. Then*

$$\lim_{k\to\infty} \left(\frac{|\lambda_1|}{\lambda_1}\right)^k \boldsymbol{x}_k = \frac{c_1}{|c_1|}\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}.$$

In particular, if $\lambda_1 > 0$ and $c_1 > 0$ then the sequence $\{\boldsymbol{x}_k\}$ will converge to the eigenvector $\boldsymbol{u}_1 := \boldsymbol{v}_1/\|\boldsymbol{v}_1\|$ of unit length.

Proof. By induction on k it follows that $\boldsymbol{x}_k = \boldsymbol{z}_k/\|\boldsymbol{z}_k\|$ for all $k \geq 0$, where $\boldsymbol{z}_k = \boldsymbol{A}^k\boldsymbol{z}_0$. Indeed, this holds for $k = 1$, and if it holds for $k-1$ then $\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} = \boldsymbol{A}\boldsymbol{z}_{k-1}/\|\boldsymbol{z}_{k-1}\| = \boldsymbol{z}_k/\|\boldsymbol{z}_{k-1}\|$ and $\boldsymbol{x}_k = (\boldsymbol{z}_k/\|\boldsymbol{z}_{k-1}\|)(\|\boldsymbol{z}_{k-1}\|/\|\boldsymbol{z}_k\|) = \boldsymbol{z}_k/\|\boldsymbol{z}_k\|$. But then

$$\boldsymbol{x}_k = \frac{\boldsymbol{z}_k}{\|\boldsymbol{z}_k\|} = \frac{c_1\lambda_1^k}{|c_1\lambda_1^k|}\frac{\boldsymbol{v}_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{v}_2 + \cdots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{v}_n}{\|\boldsymbol{v}_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{v}_2 + \cdots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{v}_n\|}, \quad k = 0, 1, 2, \ldots,$$

and this implies the lemma. ∎

Suppose we know an approximate eigenvector $\boldsymbol{u}$ of $\boldsymbol{A}$, but not the corresponding eigenvalue μ. One way of estimating μ is to minimize the Euclidian norm of the residual $r(\lambda) := \boldsymbol{A}\boldsymbol{u} - \lambda\boldsymbol{u}$.

Theorem 15.2 *Let $\boldsymbol{A} \in \mathbb{C}^{n,n}$, $\boldsymbol{u} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$, and let $\rho : \mathbb{C} \to \mathbb{R}$ be given by $\rho(\lambda) = \|\boldsymbol{A}\boldsymbol{u} - \lambda\boldsymbol{u}\|_2$. Then ρ is minimized when $\lambda := \frac{\boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}^*\boldsymbol{u}}$, the Rayleigh quotient for $\boldsymbol{A}$.*

Proof. It is equivalent to minimize $E(\lambda) := \rho^2(\lambda)$. Now

$$E(\lambda) = \boldsymbol{u}^T\boldsymbol{u}\lambda^2 - 2\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}\lambda + \boldsymbol{u}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{u}.$$

We see that E is a quadratic polynomial and since $\boldsymbol{u}^T\boldsymbol{u} > 0$, E has a unique minimum λ, where $E'(\lambda) = 0$. The solution of $E'(\lambda) = 0$ is given by $\lambda = \frac{\boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}^*\boldsymbol{u}}$. ∎

Using Rayleigh quotients we can incorporate the calculation of the eigenvalue into the power iteration. We can then compute the residual and stop the iteration

when the residual is sufficiently small. The estimate (14.8) can give us some insight. Recall that if $\boldsymbol{A}$ is nonsingular and nondefective with eigenvector matrix $\boldsymbol{X}$ and $(\mu, \boldsymbol{u})$ is an approximate eigenpair with $\|\boldsymbol{u}\|_2 = 1$, then we can find an eigenvalue λ of $\boldsymbol{A}$ such that

$$\frac{|\lambda - \mu|}{|\lambda|} \leq K_2(\boldsymbol{X}) K_2(\boldsymbol{A}) \frac{\|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\|_2}{\|\boldsymbol{A}\|_2}.$$

Thus if the relative residual is small and both $\boldsymbol{A}$ and $\boldsymbol{X}$ are well conditioned then the relative error in the eigenvalue will be small.

This discussion leads to the power method with Rayleigh quotient computation.

Algorithm 15.3 (The Power Method) Given $\boldsymbol{A} \in \mathbb{C}^{n,n}$, a starting vector $\boldsymbol{z} \in \mathbb{C}^n$, a maximum number K of iterations, and a convergence tolerance *tol*. The power method combined with a Rayleigh quotient estimate for the eigenvalue is used to compute a dominant eigenpair $(l, \boldsymbol{x})$ of $\boldsymbol{A}$ with $\|\boldsymbol{x}\|_2 = 1$. The integer *it* returns the number of iterations needed in order for $\|\boldsymbol{A}\boldsymbol{x} - l\boldsymbol{x}\|_2 / \|\boldsymbol{A}\|_F < tol$. If no such eigenpair is found in K iterations the value $it = K + 1$ is returned.

```
function [l,x,it]=powerit(A,z,K,tol)
af=norm(A,'fro'); x=z/norm(z);
for k=1:K
    y=A*x; l=x'*y;
    if norm(y-l*x)/af<tol
        it=k; x=y/norm(y); return
    end
    x=y/norm(y);
end
it=K+1;
```

Example 15.4 *We try* `powerit` *on the three matrices*

$$\boldsymbol{A}_1 := \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \boldsymbol{A}_2 := \begin{bmatrix} 1.7 & -0.4 \\ 0.15 & 2.2 \end{bmatrix}, \text{ and } \boldsymbol{A}_3 = \begin{bmatrix} 1 & 2 \\ -3 & 4 \end{bmatrix}.$$

In each case we start with the random vector $\boldsymbol{z} = [0.6602, 0.3420]$ *and* $tol = 10^{-6}$. *For* $\boldsymbol{A}_1$ *we get convergence in* 7 *iterations, for* $\boldsymbol{A}_2$ *it takes* 174 *iterations, and for* $\boldsymbol{A}_3$ *we do not get convergence.*

The matrix $\boldsymbol{A}_3$ *does not have a dominant eigenvalue since the two eigenvalues are complex conjugate of each other. Thus the basic condition (i) of* (15.1) *is not satisfied and the power method diverges. The enormous difference in the rate of convergence for* $\boldsymbol{A}_1$ *and* $\boldsymbol{A}_2$ *can be explained by looking at* (15.4). *The rate of convergence depends on the ratio* $\frac{|\lambda_2|}{|\lambda_1|}$. *If this ratio is small then the convergence is fast, while it can be quite slow if the ratio is close to one. The eigenvalues of* $\boldsymbol{A}_1$ *are* $\lambda_1 = 5.3723$ *and* $\lambda_2 = -0.3723$ *giving a quite small ratio of* 0.07 *and the convergence is fast. On the other hand the eigenvalues of* $\boldsymbol{A}_2$ *are* $\lambda_1 = 2$ *and* $\lambda_2 = 1.9$ *and the corresponding ratio is* 0.95 *resulting in slow convergence.*

A variant of the power method is the **shifted power method** In this method we choose a number s and apply the power method to the matrix $\boldsymbol{A} - s\boldsymbol{I}$. The number s is called a shift since it shifts an eigenvalue λ of $\boldsymbol{A}$ to $\lambda - s$ of $\boldsymbol{A} - s\boldsymbol{I}$. Sometimes the convergence can be faster if the shift is chosen intelligently. For example, if we apply the shifted power method to $\boldsymbol{A}_2$ in Example 15.4 with shift 1.8 then with the same starting vector and *tol* as above we get convergence in 17 iterations instead of 174 for the unshifted algorithm.

15.1.1 The Inverse Power Method

Another variant of the power method with Rayleigh quotient is the **inverse power method**. This method can be used to determine any eigenpair $(\lambda, \boldsymbol{x})$ of $\boldsymbol{A}$ as long as λ has algebraic multiplicity one. In the inverse power method we apply the power method to the inverse matrix $(\boldsymbol{A} - s\boldsymbol{I})^{-1}$, where s is a shift. If $\boldsymbol{A}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ in no particular order then $(\boldsymbol{A} - s\boldsymbol{I})^{-1}$ has eigenvalues

$$\mu_1(s) = (\lambda_1 - s)^{-1}, \mu_2(s) = (\lambda_2 - s)^{-1}, \ldots, \mu_n(s) = (\lambda_n - s)^{-1}.$$

Suppose λ_1 is a simple eigenvalue of $\boldsymbol{A}$. Then $\lim_{s\to\lambda_1} |\mu_1(s)| = \infty$, while $\lim_{s\to\lambda_1} \mu_j(s) = (\lambda_j - \lambda_1)^{-1} < \infty$ for $j = 2, \ldots, n$. Hence, by choosing s sufficiently close to λ_1 the inverse power method will converge to that eigenvalue.

For the inverse power method (15.6) is replaced by

$$\begin{aligned} &(i) \quad (\boldsymbol{A} - s\boldsymbol{I})\boldsymbol{y}_k = \boldsymbol{x}_{k-1} \\ &(ii) \quad \boldsymbol{x}_k = \boldsymbol{y}_k / \|\boldsymbol{y}_k\|. \end{aligned} \tag{15.7}$$

Note that we solve the linear system rather than computing the inverse matrix. Normally the PLU factorization of $\boldsymbol{A} - s\boldsymbol{I}$ is precomputed in order to speed up the iteration.

A variant of the inverse power method is known simply as **Rayleigh quotient iteration**. In this method we change the shift from iteration to iteration, using the previous Rayleigh quotient $\boldsymbol{s}_{k-1}$ as the current shift. In each iteration we need to compute the following quantities

$$\begin{aligned} &(i) \quad (\boldsymbol{A} - s_{k-1}\boldsymbol{I})\boldsymbol{y}_k = \boldsymbol{x}_{k-1}, \\ &(ii) \quad \boldsymbol{x}_k = \boldsymbol{y}_k / \|\boldsymbol{y}_k\|, \\ &(iii) \quad \boldsymbol{s}_k = \boldsymbol{x}_k^* \boldsymbol{A}\boldsymbol{x}_k, \\ &(iv) \quad \boldsymbol{r}_k = \boldsymbol{A}\boldsymbol{x}_k - s_k\boldsymbol{x}_k. \end{aligned}$$

We can avoid the calculation of $\boldsymbol{A}\boldsymbol{x}_k$ in (iii) and (iv). Let

$$\rho_k := \frac{\boldsymbol{y}_k^* \boldsymbol{x}_{k-1}}{\boldsymbol{y}_k^* \boldsymbol{y}_k}, \quad \boldsymbol{w}_k := \frac{\boldsymbol{x}_{k-1}}{\|\boldsymbol{y}_k\|_2}.$$

Then

$$s_k = \frac{\boldsymbol{y}_k^* \boldsymbol{A}\boldsymbol{y}_k}{\boldsymbol{y}_k^* \boldsymbol{y}_k} = s_{k-1} + \frac{\boldsymbol{y}_k^* (\boldsymbol{A} - s_{k-1}\boldsymbol{I})\boldsymbol{y}_k}{\boldsymbol{y}_k^* \boldsymbol{y}_k} = s_{k-1} + \frac{\boldsymbol{y}_k^* \boldsymbol{x}_{k-1}}{\boldsymbol{y}_k^* \boldsymbol{y}_k} = s_{k-1} + \rho_k,$$

$$\boldsymbol{r}_k = \boldsymbol{A}\boldsymbol{x}_k - s_k\boldsymbol{x}_k = \frac{\boldsymbol{A}\boldsymbol{y}_k - (s_{k-1} + \rho_k)\boldsymbol{y}_k}{\|\boldsymbol{y}_k\|_2} = \frac{\boldsymbol{x}_{k-1} - \rho_k \boldsymbol{y}_k}{\|\boldsymbol{y}_k\|_2} = \boldsymbol{w}_k - \rho_k \boldsymbol{x}_k.$$

k	1	2	3	4	5
$\|\boldsymbol{r}\|_2$	1.0e+000	7.7e-002	1.6e-004	8.2e-010	2.0e-020
$\|s-\lambda_1\|$	3.7e-001	-1.2e-002	-2.9e-005	-1.4e-010	-2.2e-016

Table 15.7. *Quadratic convergence of Rayleigh quotient iteration.*

Another problem is that the linear system in i) becomes closer and closer to singular as s_k converges to the eigenvalue. Thus the system becomes more and more ill-conditioned and we can expect large errors in the computed $\boldsymbol{y}_k$. This is indeed true, but we are lucky. Most of the error occurs in the direction of the eigenvector and this error disappears when we normalize $\boldsymbol{y}_k$ in ii). Miraculously, the normalized eigenvector will be quite accurate.

We obtain the following algorithm.

Algorithm 15.5 (Rayleigh quotient iteration) Given an approximation $(s, \boldsymbol{x})$ to an eigenpair $(\lambda, \boldsymbol{v})$ of a matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$. This algorithm computes a hopefully better approximation to $(\lambda, \boldsymbol{v})$ by doing one Rayleigh quotient iteration. The length nr of the new residual is also returned

```
function [x,s,nr]=rayleighit(A,x,s)
n=length(x);
y=(A-s*eye(n,n))\x;
yn=norm(y);
w=x/yn;
x=y/yn;
rho=x'*w;
s=s+rho;
nr=norm(w-rho*x);
```

Since the shift changes from iteration to iteration the computation of $\boldsymbol{y}$ in `rayleighit` will require $O(n^3)$ flops for a full matrix. For such a matrix it might pay to reduce it to a upper Hessenberg form or tridiagonal form before starting the iteration. However, if we have a good approximation to an eigenpair then only a few iterations are necessary to obtain close to machine accuracy.

If Rayleigh quotient iteration converges the convergence will be quadratic and sometimes even cubic. We illustrate this with an example.

Example 15.6 *The matrix* $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$ *has an eigenvalue* $\lambda_1 = (5-\sqrt{33})/2 \approx -0.37$. *We test the rate of convergence by calling* `rayleighit` *5 times starting with* $\boldsymbol{x} = [1,1]^T$ *and* $s = 0$. *The results are shown in Table 15.7. The errors are approximately squared in each iteration indicating quadratic convergence.*

15.2 The QR Algorithm

The QR algorithm is an iterative method to compute all eigenvalues and eigenvectors of a matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$. The matrix is reduced to triangular form by a sequence of unitary similarity transformations computed from the QR factorization of $\boldsymbol{A}$. Recall that for a square matrix the QR factorization and the QR decomposition are the same. If $\boldsymbol{A} = \boldsymbol{QR}$ is a QR factorization then $\boldsymbol{Q} \in \mathbb{C}^{n,n}$ is unitary, $\boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{R} \in \mathbb{C}^{n,n}$ is upper triangular.

The basic QR algorithm takes the following form:

$$\begin{array}{|l|}\hline \boldsymbol{A}_1 = \boldsymbol{A} \\ \text{for } k = 1, 2, \ldots \\ \quad \boldsymbol{Q}_k\boldsymbol{R}_k = \boldsymbol{A}_k \qquad (\text{QR factorization of } \boldsymbol{A}_k) \\ \quad \boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k. \\ \text{end} \\ \hline \end{array} \tag{15.8}$$

Here are two examples to illustrate the convergence.

Example 15.8 *We start with*

$$\boldsymbol{A}_1 = \boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \left(\frac{1}{\sqrt{5}}\begin{bmatrix} -2 & -1 \\ -1 & 2 \end{bmatrix}\right) * \left(\frac{1}{\sqrt{5}}\begin{bmatrix} -5 & -4 \\ 0 & 3 \end{bmatrix}\right) = \boldsymbol{Q}_1\boldsymbol{R}_1$$

and obtain

$$\boldsymbol{A}_2 = \boldsymbol{R}_1\boldsymbol{Q}_1 = \frac{1}{5}\begin{bmatrix} -5 & -4 \\ 0 & 3 \end{bmatrix} * \begin{bmatrix} -2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 2.8 & -0.6 \\ -0.6 & 1.2 \end{bmatrix}.$$

Continuing we find

$$\boldsymbol{A}_3 = \begin{bmatrix} 2.997 & -0.074 \\ -0.074 & 1.0027 \end{bmatrix}, \quad \boldsymbol{A}_9 = \begin{bmatrix} 3.0000 & -0.0001 \\ -0.0001 & 1.0000 \end{bmatrix}$$

$\boldsymbol{A}_9$ *is almost diagonal and contains the eigenvalues* $\lambda_1 = 3$ *and* $\lambda_2 = 1$ *on the diagonal.*

Example 15.9 *Applying the QR iteration* (15.8) *to the matrix*

$$A_1 = A = \begin{bmatrix} 0.9501 & 0.8913 & 0.8214 & 0.9218 \\ 0.2311 & 0.7621 & 0.4447 & 0.7382 \\ 0.6068 & 0.4565 & 0.6154 & 0.1763 \\ 0.4860 & 0.0185 & 0.7919 & 0.4057 \end{bmatrix}$$

we obtain

$$A_{14} = \left[\begin{array}{r|rr|r} 2.323 & 0.047223 & -0.39232 & -0.65056 \\ \hline -2.1e-10 & 0.13029 & 0.36125 & 0.15946 \\ -4.1e-10 & -0.58622 & 0.052576 & -0.25774 \\ \hline 1.2e-14 & 3.3e-05 & -1.1e-05 & 0.22746 \end{array}\right].$$

This matrix is almost quasi-triangular and estimates for the eigenvalues $\lambda_1, \ldots, \lambda_4$ of $\boldsymbol{A}$ can now easily be determined from the diagonal blocks of $\boldsymbol{A}_{14}$. The 1×1 blocks give us two real eigenvalues $\lambda_1 \approx 2.323$ and $\lambda_4 \approx 0.2275$. The middle 2×2 block has complex eigenvalues resulting in $\lambda_2 \approx 0.0914 + 0.4586i$ and $\lambda_3 \approx 0.0914 - 0.4586i$. From Gerschgorin's circle theorem 14.4 and Corollary 14.6 it follows that the approximations to the real eigenvalues are quite accurate. We would also expect the complex eigenvalues to have small absolute errors.

15.2.1 The Relation to the Power Method

In the basic QR algorithm we obtain the QR factorization of the powers $\boldsymbol{A}^k$ as follows:

Theorem 15.10 *Let $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_k$ and $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_k$ be the matrices generated by the basic QR algorithm* (15.8)*. Then the products*

$$\tilde{\boldsymbol{Q}}_k := \boldsymbol{Q}_1 \cdots \boldsymbol{Q}_k \text{ and } \tilde{\boldsymbol{R}}_k := \boldsymbol{R}_k \cdots \boldsymbol{R}_1 \text{ for } k \geq 1 \tag{15.9}$$

are the matrices in a QR factorization $\boldsymbol{A}^k = \tilde{\boldsymbol{Q}}_k \tilde{\boldsymbol{R}}_k$ of $\boldsymbol{A}^k$.

Proof. The proof is by induction on k. Clearly $\tilde{\boldsymbol{Q}}_1 \tilde{\boldsymbol{R}}_1 = \boldsymbol{Q}_1 \boldsymbol{R}_1 = \boldsymbol{A}_1$. Suppose $\tilde{\boldsymbol{Q}}_{k-1} \tilde{\boldsymbol{R}}_{k-1} = \boldsymbol{A}^{k-1}$ for some $k \geq 2$. Since $\boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{A}_k = \tilde{\boldsymbol{Q}}_{k-1}^* \boldsymbol{A} \tilde{\boldsymbol{Q}}_{k-1}$ we find

$$\tilde{\boldsymbol{Q}}_k \tilde{\boldsymbol{R}}_k = \tilde{\boldsymbol{Q}}_{k-1}(\boldsymbol{Q}_k \boldsymbol{R}_k)\tilde{\boldsymbol{R}}_{k-1} = \tilde{\boldsymbol{Q}}_{k-1} \boldsymbol{A}_k \tilde{\boldsymbol{R}}_{k-1} = (\tilde{\boldsymbol{Q}}_{k-1} \tilde{\boldsymbol{Q}}_{k-1}^*) \boldsymbol{A} \tilde{\boldsymbol{Q}}_{k-1} \tilde{\boldsymbol{R}}_{k-1} = \boldsymbol{A}^k.$$

□

Since $\tilde{\boldsymbol{R}}_k$ is upper triangular, its first column is a multiple of $\boldsymbol{e}_1$ so that

$$\boldsymbol{A}^k \boldsymbol{e}_1 = \tilde{\boldsymbol{Q}}_k \tilde{\boldsymbol{R}}_k \boldsymbol{e}_1 = \tilde{r}_{11}^{(k)} \tilde{\boldsymbol{Q}}_k \boldsymbol{e}_1 \text{ or } \tilde{\boldsymbol{q}}_1^{(k)} := \tilde{\boldsymbol{Q}}_k \boldsymbol{e}_1 = \frac{1}{\tilde{r}_{11}^{(k)}} \boldsymbol{A}^k \boldsymbol{e}_1.$$

Since $\|\tilde{\boldsymbol{q}}_1^{(k)}\|_2 = 1$ the first column of $\tilde{\boldsymbol{Q}}_k$ is the result of applying the normalized power iteration (15.6) to the starting vector $\boldsymbol{x}_0 = \boldsymbol{e}_1$. If this iteration converges we conclude that the first column of $\tilde{\boldsymbol{Q}}_k$ must converge to a dominant eigenvector of $\boldsymbol{A}$. It can be shown that the first column of $\boldsymbol{A}_k$ must then converge to $\lambda_1 \boldsymbol{e}_1$, where λ_1 is a dominant eigenvalue of $\boldsymbol{A}$. This is clearly what happens in Examples 15.8 and 15.9.

15.2.2 A convergence theorem

There is no theorem which proves convergence of the QR algorithm in general. The following theorem shows convergence under somewhat restrictive assumptions.

Theorem 15.11 *Suppose in the basic QR algorithm* (15.8) *that*

1. *$\boldsymbol{A} \in \mathbb{R}^{n,n}$ can be diagonalized, $\boldsymbol{X}^{-1} \boldsymbol{A} \boldsymbol{X} = \boldsymbol{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.*

2. *The eigenvalues* $\lambda_1, \ldots, \lambda_n$ *are real with* $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$.

3. *The inverse of the eigenvector matrix has an LU factorization* $\boldsymbol{X}^{-1} = \boldsymbol{LR}$.

Let $\tilde{\boldsymbol{Q}}_k = \boldsymbol{Q}_1 \ldots \boldsymbol{Q}_k$ *for* $k \geq 1$. *Then there is a diagonal matrix* $\boldsymbol{D}_k$ *with diagonal elements* ± 1 *such that* $\tilde{\boldsymbol{Q}}_k \boldsymbol{D}_k \to \boldsymbol{Q}$, *where* $\boldsymbol{Q}^T \boldsymbol{AQ}$ *is triangular and* $\boldsymbol{Q}$ *is the Q-factor in the QR factorization of the eigenvector matrix* $\boldsymbol{X}$.

Proof. In this proof we assume that every QR factorization has an $\boldsymbol{R}$ with positive diagonal elements so that the factorization is unique. Let $\boldsymbol{X} = \boldsymbol{QR}$ be the QR factorization of $\boldsymbol{X}$. We observe that $\boldsymbol{Q}^T \boldsymbol{AQ}$ is upper triangular. For since $\boldsymbol{X}^{-1}\boldsymbol{AX} = \boldsymbol{\Lambda}$ we have $\boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{AQR} = \boldsymbol{\Lambda}$ so that $\boldsymbol{Q}^T\boldsymbol{AQ} = \boldsymbol{R\Lambda R}^{-1}$ is upper triangular. Since $\boldsymbol{A}_{k+1} = \tilde{\boldsymbol{Q}}_k^T \boldsymbol{A} \tilde{\boldsymbol{Q}}_k$, it is enough to show that $\tilde{\boldsymbol{Q}}_k \boldsymbol{D}_k \to \boldsymbol{Q}$ for some diagonal matrix $\boldsymbol{D}_k$ with diagonal elements ± 1.

We define the nonsingular matrices

$$\boldsymbol{F}_k := \boldsymbol{R\Lambda}^k \boldsymbol{L\Lambda}^{-k} \boldsymbol{R}^{-1} = \hat{\boldsymbol{Q}}_k \hat{\boldsymbol{R}}_k, \quad \boldsymbol{G}_k := \hat{\boldsymbol{R}}_k \boldsymbol{R\Lambda}^k \boldsymbol{R}, \quad \boldsymbol{D}_k := \operatorname{diag}\big(\frac{\delta_1}{|\delta_1|}, \ldots, \frac{\delta_n}{|\delta_n|}\big),$$

where $\delta_1, \ldots, \delta_n$ are the diagonal elements in the upper triangular matrix $\boldsymbol{G}_k$ and $\boldsymbol{F}_k = \hat{\boldsymbol{Q}}_k \hat{\boldsymbol{R}}_k$ is the QR factorization of $\boldsymbol{F}_k$. Then

$$\begin{aligned} \boldsymbol{A}^k &= \boldsymbol{X\Lambda}^k \boldsymbol{X}^{-1} = \boldsymbol{QR\Lambda}^k \boldsymbol{LR} = \boldsymbol{Q}(\boldsymbol{R\Lambda}^k \boldsymbol{L\Lambda}^{-k} \boldsymbol{R}^{-1})(\boldsymbol{R\Lambda}^k \boldsymbol{R}) \\ &= \boldsymbol{QF}_k(\boldsymbol{R\Lambda}^k \boldsymbol{R}) = \boldsymbol{Q}\hat{\boldsymbol{Q}}_k \hat{\boldsymbol{R}}_k(\boldsymbol{R\Lambda}^k \boldsymbol{R}) = (\boldsymbol{Q}\hat{\boldsymbol{Q}}_k \boldsymbol{D}_k^{-1})(\boldsymbol{D}_k \boldsymbol{G}_k), \end{aligned}$$

and this is the QR factorization of $\boldsymbol{A}^k$. Indeed, $\boldsymbol{Q}\hat{\boldsymbol{Q}}_k \boldsymbol{D}_k^{-1}$ is a product of orthonormal matrices and therefore orthonormal. Moreover $\boldsymbol{D}_k \boldsymbol{G}_k$ is a product of upper triangular matrices and therefore upper triangular. Note that $\boldsymbol{D}_k$ is chosen so that this matrix has positive diagonal elements. By Theorem 15.10 $\boldsymbol{A}^k = \tilde{\boldsymbol{Q}}_k \tilde{\boldsymbol{R}}_k$ is also the QR factorization of $\boldsymbol{A}^k$, and we must have $\tilde{\boldsymbol{Q}}_k = \boldsymbol{Q}\hat{\boldsymbol{Q}}_k \boldsymbol{D}_k^{-1}$ or $\tilde{\boldsymbol{Q}}_k \boldsymbol{D}_k = \boldsymbol{Q}\hat{\boldsymbol{Q}}_k$. The theorem will follow if we can show that $\hat{\boldsymbol{Q}}_k \to \boldsymbol{I}$.

The matrix $\boldsymbol{\Lambda}^k \boldsymbol{L\Lambda}^{-k}$ is lower triangular with elements $(\frac{\lambda_i}{\lambda_j})^k l_{ij}$ on and under the diagonal. Thus for $n = 3$

$$\boldsymbol{\Lambda}^k \boldsymbol{L\Lambda}^{-k} = \begin{bmatrix} 1 & 0 & 0 \\ (\frac{\lambda_2}{\lambda_1})^k l_{21} & 1 & 0 \\ (\frac{\lambda_3}{\lambda_1})^k l_{31} & (\frac{\lambda_3}{\lambda_2})^k l_{32} & 1 \end{bmatrix}.$$

By Assumption 2. it follows that $\boldsymbol{\Lambda}^k \boldsymbol{L\Lambda}^{-k} \to \boldsymbol{I}$, and hence $\boldsymbol{F}_k \to \boldsymbol{I}$. Since $\hat{\boldsymbol{R}}_k^T \hat{\boldsymbol{R}}_k$ is the Cholesky factorization of $\boldsymbol{F}_k^T \boldsymbol{F}_k$ it follows that $\hat{\boldsymbol{R}}_k^T \hat{\boldsymbol{R}}_k \to \boldsymbol{I}$. By the continuity of the Cholesky factorization it holds $\hat{\boldsymbol{R}}_k \to \boldsymbol{I}$ and hence $\hat{\boldsymbol{R}}_k^{-1} \to \boldsymbol{I}$. But then $\hat{\boldsymbol{Q}}_k = \boldsymbol{F}_k \hat{\boldsymbol{R}}_k^{-1} \to \boldsymbol{I}$. ∎

Exercise 15.12 *Use Theorem 8.40 to show that* $\hat{\boldsymbol{R}}_k \to \boldsymbol{I}$ *implies* $\hat{\boldsymbol{R}}_k^{-1} \to \boldsymbol{I}$.

15.2.3 The Shifted QR Algorithms

Like in the inverse power method it is possible to speed up the convergence by introducing shifts. The **explicitly shifted QR algorithm** works as follows:

$$
\begin{aligned}
&\boldsymbol{A}_1 = \boldsymbol{A} \\
&\text{for } k = 1, 2, \ldots \\
&\quad \text{Choose a shift } s_k \\
&\quad \boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{A}_k - s_k \boldsymbol{I} \qquad (\text{QR factorization of } \boldsymbol{A}_k - s\boldsymbol{I}) \\
&\quad \boldsymbol{A}_{k+1} = \boldsymbol{R}_k \boldsymbol{Q}_k + s_k \boldsymbol{I}. \\
&\text{end}
\end{aligned}
\tag{15.10}
$$

We will not develop the practical details of an implementation of this algorithm. We contend ourselves with the following remarks. See [18] for a detailed discussion and algorithms.

1. $\boldsymbol{A}_{k+1}$ is unitary similar to $\boldsymbol{A}_k$. For since $\boldsymbol{R}_k = \boldsymbol{Q}_k^*(\boldsymbol{A}_k - s_k\boldsymbol{I})$ we find $\boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k + s_k\boldsymbol{I} = \boldsymbol{Q}_k^*(\boldsymbol{A}_k - s_k\boldsymbol{I})\boldsymbol{Q}_k + s_k\boldsymbol{I} = \boldsymbol{Q}_k^*\boldsymbol{A}_k\boldsymbol{Q}_k$.
2. Before applying this algorithm we reduce $\boldsymbol{A}$ to upper Hessenberg form using Algorithm 14.11.
3. If $\boldsymbol{A}$ is upper Hessenberg then all matrices $\{\boldsymbol{A}_k\}_{k\geq 1}$ will be upper Hessenberg. This follows since $\boldsymbol{Q}_k = (\boldsymbol{A}_k - s_k\boldsymbol{I})\boldsymbol{R}^{-1}$ implies $\boldsymbol{A}_{k+1} = \boldsymbol{R}_k(\boldsymbol{A}_k - s_k\boldsymbol{I})\boldsymbol{R}_k^{-1} + s_k\boldsymbol{I} = \boldsymbol{R}_k\boldsymbol{A}_k\boldsymbol{R}_k^{-1}$. This product of two upper triangular matrices and an upper Hessenberg matrix is upper Hessenberg.
4. Givens rotations is used to compute the QR factorization of $\boldsymbol{A}_k - s_k\boldsymbol{I}$.
5. To compute $\boldsymbol{A}_{k+1}$ from $\boldsymbol{A}_k$ requires $O(n^2)$ flops if $\boldsymbol{A}_k$ is upper Hessenberg and $O(n)$ flops if $\boldsymbol{A}_k$ is tridiagonal.
6. The shifted QR algorithm is related to the power method, cf. Theorem 15.10.
7. The equation $\boldsymbol{A} - s_k\boldsymbol{I} = \boldsymbol{Q}_k\boldsymbol{R}_k$ implies that $(\boldsymbol{A} - s_k\boldsymbol{I})^T\boldsymbol{q}_k = r_{nn}^k\boldsymbol{e}_n$, where $\boldsymbol{q}_k$ is the last column of $\boldsymbol{Q}_k$ and r_{nn}^k is the (n,n) element in $\boldsymbol{R}_k$. Thus $\boldsymbol{q}_k$ is the result of one iteration of the inverse power method to $\boldsymbol{A}^T$ with shift s_k.
8. If a subdiagonal element $a_{i+1,i}$ of an upper Hessenberg matrix $\boldsymbol{A}$ is equal to zero, then the eigenvalues of $\boldsymbol{A}$ are the union of the eigenvalues of the two smaller matrices $A(1:i, 1:i)$ and $A(i+1:n, i+1:n)$. Thus if during the iteration the $(i+1,i)$ element of $\boldsymbol{A}_k$ is sufficiently small then we can continue the iteration on the two smaller submatrices separately. This splitting occurs often in practice and can with a proper implementation reduce the computation time considerably.
9. The shift $s_k := \boldsymbol{e}_n^T\boldsymbol{A}_k\boldsymbol{e}_n$ is called the **Rayleigh quotient shift**.
10. The eigenvalue of the lower right 2×2 corner of $\boldsymbol{A}_k$ closest to the n,n element of $\boldsymbol{A}_k$ is called the**Wilkinson shift**. This shift can be used to find complex eigenvalues of a real matrix.
11. The convergence is very fast and at least quadratic both for the Rayleigh quotient shift and the Wilkinson shift.

12. By doing two QR iterations at a time it is possible to find both real and complex eigenvalues without using complex arithmetic. The corresponding algorithm is called the **implicitly shifted QR algorithm**
13. After having computed the eigenvalues we can compute the eigenvectors in steps. First we find the eigenvectors of the triangular or quasi-triangular matrix. We then compute the eigenvectors of the upper Hessenberg matrix and finally we get the eigenvectors of $\boldsymbol{A}$.
14. Practical experience indicates that only $O(n)$ iterations are needed to find all eigenvalues of $\boldsymbol{A}$. Thus both the explicit- and implicit shift QR algorithms are normally $O(n^3)$ algorithms.

Part VI

Appendix

Appendix A

Vectors

This chapter contains a review of vector space concepts that will be useful in this text. we start by introducing a vector space. To define a vector space we need a field $\mathbb{F}$, a set of vectors $\mathcal{V}$, a way to combine vectors called **vector addition**, and a way to combine elements of $\mathbb{F}$ and $\mathcal{V}$ called **scalar multiplication**. In the first part of this section $\mathbb{F}$ will be an arbitrary field, but later the field will be the set of real or complex numbers with the usual arithmetic operations.

A.1 Vector Spaces

Definition A.1 *A* **field** *is a set $\mathbb{F}$ together with two operations $+, \cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ such that for all $a, b, c \in \mathbb{F}$ the following arithmetic rules hold*

(A0) *there exists an element $0 \in \mathbb{F}$ such that $a + 0 = a$.*

(Am) *there exists an element $(-a) \in \mathbb{F}$ such that $a + (-a) = 0$. We define subtraction as $a - b := a + (-b)$.*

(Aa) $a + (b + c) = (a + b) + c$.

(Ac) $a + b = b + a$.

(M1) *there exists an element $1 \in \mathbb{F}$ such that $a \cdot 1 = a$.*

(Mi) *if $a \neq 0$ then there exists an element $a^{-1} \in \mathbb{F}$ such that $a \cdot a^{-1} = 1$.*

(Ma) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.

(Mc) $a \cdot b = b \cdot a$.

(D) $a \cdot (b + c) = a \cdot b + a \cdot c$.

The requirements **(A0), (Am), (Aa)** are the axioms for a group. They state that $(\mathbb{F}, +)$ is a **group**, and since in addition **(Ac)** holds then $(\mathbb{F}, +)$ is by defintion an

abelian group. The axioms **(M1), (Mi), (Ma), (Mc)** state that $(\mathbb{F}\setminus\{0\},\cdot)$ is an abelian group. Often we drop the dot and write ab for the product $a\cdot b$. Examples of fields are $\mathbb{R}$ or $\mathbb{C}$ with ordinary addition and multiplication.

Definition A.2 *A* **vector space** *over a field $\mathbb{F}$ is a set $\mathcal{V}$ together with two operations vector addition, $+:\mathcal{V}\times\mathcal{V}\to\mathcal{V}$ and scalar multiplication, $\cdot:\mathbb{F}\times\mathcal{V}\to\mathcal{V}$ such that for all $a,b\in\mathbb{F}$ and $\boldsymbol{v},\boldsymbol{w}\in\mathcal{V}$ the following hold*

(V) *$(\mathcal{V},+)$ is an abelian group.*

(Va) $(a\cdot b)\cdot\boldsymbol{v}=a\cdot(b\cdot\boldsymbol{v})$.

(Vd1) $(a+b)\cdot\boldsymbol{v}=a\cdot\boldsymbol{v}+b\cdot\boldsymbol{v}$.

(Vd2) $a\cdot(\boldsymbol{v}+\boldsymbol{w})=a\cdot\boldsymbol{v}+a\cdot\boldsymbol{w}$.

(M1) $1\cdot\boldsymbol{v}=\boldsymbol{v}$.

We denote a vector space by $(\mathcal{V},\mathbb{F})$ or by $\mathcal{V}$ if the underlying field is clear from the context.

Definition A.3 *Let $(\mathcal{V},\mathbb{F})$ be a vector space and $\mathcal{S}$ a nonempty subset of $\mathcal{V}$. Then $(\mathcal{S},\mathbb{F})$ is a* **subspace** *of $(\mathcal{V},\mathbb{F})$ if $(\mathcal{S},\mathbb{F})$ is itself a vector space.*

It follows that $(\mathcal{S},\mathbb{F})$ is a subspace of $(\mathcal{V},\mathbb{F})$ if $\mathcal{S}$ is closed under vector addition and scalar multiplication, i.e. $a\boldsymbol{s}_1+b\boldsymbol{s}_2\in\mathcal{S}$ for all $a,b\in\mathbb{F}$ and all $\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{S}$. For any vector space $(\mathcal{V},\mathbb{F})$ the two sets $\{\mathbf{0}\}$, consisting only of the zero element in $\mathcal{V}$, and $\mathcal{V}$ itself are subspaces. They are called the **trivial subspaces**.

Here are some examples of vector spaces.

Example A.4 (The Vector Spaces $\mathbb{R}^n$ and $\mathbb{C}^n$) *In the following chapters we will deal almost exclusively with the vector spaces $\mathbb{R}^n=(\mathbb{R}^n,\mathbb{R})$, $\mathbb{C}^n=(\mathbb{C}^n,\mathbb{C})$ and their subspaces. Addition and scalar multiplication are defined by*

$$\boldsymbol{v}+\boldsymbol{w}=\begin{bmatrix}v_1+w_1\\ \vdots\\ v_n+w_n\end{bmatrix},\quad a\boldsymbol{v}=\begin{bmatrix}av_1\\ \vdots\\ av_n\end{bmatrix}.$$

Example A.5 (Subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$) *For a given vector $\boldsymbol{x}\in\mathbb{R}^n$ let $\mathcal{S}=\{t\boldsymbol{x}:t\in\mathbb{R}\}$. Then $\mathcal{S}$ is a subspace of $\mathbb{R}^n$, in fact it represents a straight line passing through the origin. For $n=2$ it can be shown that all nontrivial subspaces of $\mathbb{R}^2$ are of this form. For $n=3$ the nontrivial subspaces are all lines and all planes containing $\{\mathbf{0}\}$.*

Example A.6 (The Vector Space $C(I)$) *Let $\mathbb{F} = \mathbb{R}$ and let $C(I)$ be the set of all real valued functions $f : I \to \mathbb{R}$ which are defined and continuous on an interval $I \subset \mathbb{R}$. Here the vectors are functions in $C(I)$. Vector addition and scalar multiplication are defined for all $f, g \in C(I)$ and all $a \in \mathbb{R}$ by*

$$(f+g)(x) := f(x) + g(x), \quad (af)(x) := af(x), \text{ for all } x \in I.$$

$C(I) = \big(C(I), \mathbb{R}\big)$ is a vector space since

- *the sum of two continuous functions is continuous,*
- *a constant times a continuous function is continuous*
- *vector addition and scalar multiplication are defined point-wise, so the axioms for a vector space follows from properties of real numbers.*

Example A.7 (The Vector Space Π_n) *Let $\Pi_n(I)$ be the set of all polynomials of degree at most n defined on a subset $I \subset \mathbb{R}$ or $I \subset \mathbb{C}$. We write simply Π_n if $I = \mathbb{R}$ or $I = \mathbb{C}$. With pointwise addition and scalar multiplication defined as in Example A.6 the set $(\Pi_n(I), \mathbb{R})$ is a subspace of $(C(I), \mathbb{R})$.*

Definition A.8 (Linear Combinations) *The sum $c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdot + c_n\boldsymbol{v}_n$ with $c_i \in \mathbb{F}$ and $\boldsymbol{v}_i \in \mathcal{V}$ for $i = 1, \ldots, n$ is called a* **linear combination** *of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. We say that the linear combination is nontrivial if at least one of the c_i's is nonzero. The set*

$$\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} := \{c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n : c_i \in \mathbb{F},\ i = 1, \ldots, n\}$$

spanned by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{V}$ is a subspace of $(\mathcal{V}, \mathbb{F})$. A vector space $\mathcal{V}$ is called **finite dimensional** *if it has a finite spanning set; i.e. there exist $n \in \mathbb{N}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ in $\mathcal{V}$ such that $\mathcal{V} = \operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$.*

Exercise A.9 *Show that the $\mathbf{0}$ of vector addition is unique and that $\{\mathbf{0}\}$ is a subspace.*

Exercise A.10 *Show that $0 \cdot \boldsymbol{x} = \mathbf{0}$ for any $\boldsymbol{x} \in \mathcal{V}$.*

Exercise A.11 *Show that $\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a subspace.*

Exercise A.12 *Show that $\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is the smallest subspace containing the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.*

A.2 Linear Independence and Bases

Definition A.13 *Let $\mathcal{X} := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a set of vectors in a vector space $(\mathcal{V}, \mathbb{F})$. We say that $\mathcal{X}$ is* **linearly dependent** *if we can find a nontrivial linear combination which is equal to zero. We say that $\mathcal{X}$ is* **linearly independent** *if it is not linearly dependent. In other words*

$$c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n = \mathbf{0} \text{ for some } c_1, \ldots, c_n \in \mathbb{F} \implies c_1 = \cdots = c_n = 0.$$

The elements in a set of linearly independent vectors must all be nonzero and we have

Lemma A.14 *Suppose $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ span a vector space V and that $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are linearly independent vectors in $\mathcal{V}$. Then $k \leq n$.*

Proof. Suppose $k > n$. Write $\boldsymbol{w}_1$ as a linear combination of elements from the set $\mathcal{X}_0 := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$, say $\boldsymbol{w}_1 = c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n$. Since $\boldsymbol{w}_1 \neq \mathbf{0}$ not all the c's are equal to zero. Pick a nonzero c, say c_{i_1}. Then $\boldsymbol{v}_{i_1}$ can be expressed as a linear combination of $\boldsymbol{w}_1$ and the remaining $\boldsymbol{v}$'s. So the set $\mathcal{X}_1 := \{\boldsymbol{w}_1, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i_1-1}, \boldsymbol{v}_{i_1+1}, \ldots, \boldsymbol{v}_n\}$ must also be a spanning set for $\mathcal{V}$. We repeat this for $\boldsymbol{w}_2$ and $\mathcal{X}_1$. In the linear combination $\boldsymbol{w}_2 = d_{i_1}\boldsymbol{w}_1 + \sum_{j \neq i_1} d_j\boldsymbol{v}_j$, we must have $d_{i_2} \neq 0$ for some i_2. Moreover $i_2 \neq i_1$ for otherwise $\boldsymbol{w}_2 = d_1\boldsymbol{w}_1$ contradicting the linear independence of the $\boldsymbol{w}$'s. So the set $\mathcal{X}_2$ consisting of the $\boldsymbol{v}$'s with $\boldsymbol{v}_{i_1}$ replaced by $\boldsymbol{w}_1$ and $\boldsymbol{v}_{i_2}$ replaced by $\boldsymbol{w}_2$ is again a spanning set for $\mathcal{V}$. Repeating this process $n-2$ more times we obtain a spanning set $\mathcal{X}_n$ where all the v's have been replaced by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$. Since $k > n$ we can then write $\boldsymbol{w}_k$ as a linear combination of $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ contradicting the linear independence of the $\boldsymbol{w}$'s. We conclude that $k \leq n$. ∎

Definition A.15 *A finite set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ in a vector space $(\mathcal{V}, \mathbb{F})$ is a* **basis** *for $(\mathcal{V}, \mathbb{F})$ if*

1. $\text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} = \mathcal{V}$.
2. *$\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is linearly independent.*

Theorem A.16 *Suppose $(\mathcal{V}, \mathbb{F})$ is a vector space and that $S := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a spanning set for $\mathcal{V}$. Then we can find a subset $\{\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}\}$ of S that forms a basis for $\mathcal{V}$.*

Proof. If $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is linearly dependent we can express one of the $\boldsymbol{v}$'s as a nontrivial linear combination of the remaining $\boldsymbol{v}$'s and drop that $\boldsymbol{v}$ from the spanning set. Continue this process until the remaining $\boldsymbol{v}$'s are linearly independent. They still span the vector space and therefore form a basis. ∎

Corollary A.17 *A vector space is finite dimensional if and only if it has a basis.*

Proof. Let $\mathcal{V} = \text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a finite dimensional vector space. By Theorem A.16 $\mathcal{V}$ has a basis. Conversely, if $\mathcal{V} = \text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis then it is by defintion a finite spanning set. □

Theorem A.18 *Every basis for a vector space $\mathcal{V}$ has the same number of elements. This number is called the* **dimension** *of the vector space and denoted $\dim \mathcal{V}$.*

Proof. Suppose $\mathcal{X} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ and $\mathcal{Y} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$ are two bases for $\mathcal{V}$. By Lemma A.14 we have $k \leq n$. Using the same Lemma with $\mathcal{X}$ and $\mathcal{Y}$ switched we obtain $n \leq k$. We conclude that $n = k$. □

The set of unit vectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ form a basis for both $\mathbb{R}^n$ and $\mathbb{C}^n$. The dimension of the trivial subspace $\{\boldsymbol{0}\}$ is defined to be zero.

Theorem A.19 *Every linearly independent set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ in a finite dimensional vector space $\mathcal{V}$ can be enlarged to a basis for $\mathcal{V}$.*

Proof. If $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ does not span $\mathcal{V}$ we can enlarge the set by one vector $\boldsymbol{v}_{k+1}$ which cannot be expressed as a linear combination of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$. The enlarged set is also linearly independent. Continue this process. Since the space is finite dimensional it must stop after a finite number of steps. □

It is convenient to introduce a matrix transforming a basis in a subspace into a basis for the space itself.

Lemma A.20 *Suppose $\mathcal{S}$ is a subspace of a finite dimensional vector space $(\mathcal{V}, \mathbb{F})$ and let $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be a basis for $\mathcal{S}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ a basis for $\mathcal{V}$. Then each $\boldsymbol{s}_j$ can be expressed as a linear combination of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, say*

$$\boldsymbol{s}_j = \sum_{i=1}^{m} a_{ij} \boldsymbol{v}_i \text{ for } j = 1, \ldots, n. \tag{A.1}$$

If $\boldsymbol{x} \in \mathcal{S}$ then $\boldsymbol{x} = \sum_{j=1}^{n} c_j \boldsymbol{s}_j = \sum_{i=1}^{m} b_i \boldsymbol{v}_i$ for some coefficients $\boldsymbol{b} := [b_1, \ldots, \boldsymbol{b}_m]^T$, $\boldsymbol{c} := [c_1, \ldots, c_n]^T$. Moreover $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{c}$, where $\boldsymbol{A} = [a_{ij}] \in \mathbb{C}^{m,n}$. The matrix $\boldsymbol{A}$ has linearly independent columns.

Proof. (A.1) holds since $\boldsymbol{s}_j \in \mathcal{V}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ spans $\mathcal{V}$. Since $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ is a basis for $\mathcal{S}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ a basis for $\mathcal{V}$ every $\boldsymbol{x} \in \mathcal{S}$ can be written $\boldsymbol{x} = \sum_{j=1}^{n} c_j \boldsymbol{s}_j = \sum_{i=1}^{m} b_i \boldsymbol{v}_i$ for some scalars (c_j) and (b_i). But then

$$\boldsymbol{x} = \sum_{j=1}^{n} c_j \boldsymbol{s}_j \overset{(A.1)}{=} \sum_{j=1}^{n} c_j \big(\sum_{i=1}^{m} a_{ij} \boldsymbol{v}_i \big) = \sum_{i=1}^{m} \big(\sum_{j=1}^{n} a_{ij} c_j \big) \boldsymbol{v}_i = \sum_{i=1}^{m} b_i \boldsymbol{v}_i.$$

Since $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ is linearly independent it follows that $b_i = \sum_{j=1}^{n} a_{ij} c_j$ for $i = 1, \ldots, m$ or $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{c}$. Finally, to show that $\boldsymbol{A}$ has linearly independent columns

suppose $\boldsymbol{b} := \boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$ for some $\boldsymbol{c} = [c_1, \ldots, c_n]^T$. Define $\boldsymbol{x} := \sum_{j=1}^{n} c_j \boldsymbol{s}_j$. Then $\boldsymbol{x} = \sum_{i=1}^{m} b_i \boldsymbol{v}_i$ and since $\boldsymbol{b} = \boldsymbol{0}$ we have $\boldsymbol{x} = \boldsymbol{0}$. But since $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ is linearly independent we have $\boldsymbol{c} = \boldsymbol{0}$. □

The matrix $\boldsymbol{A}$ in Lemma A.20 is called a **change of basis matrix**.

Exercise A.21 *Show that the elements in a linearly independent set must be nonzero.*

Exercise A.22 *Show that the set of unit vectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ form a basis both for $\mathbb{R}^n$ and for $\mathbb{C}^n$. Why does this show that the dimension of $\mathbb{R}^n$ and $\mathbb{C}^n$ is n?*

A.3 Operations on Subspaces

Let $\mathcal{R}$ and $\mathcal{S}$ be two subsets of a vector space $(\mathcal{V}, \mathbb{F})$ and let a be a scalar. The **sum**, **multiplication by scalar**, **union**, and **intersection** of $\mathcal{R}$ and $\mathcal{S}$ are defined by

$$\mathcal{R} + \mathcal{S} := \{\boldsymbol{r} + \boldsymbol{s} : \boldsymbol{r} \in \mathcal{R} \text{ and } \boldsymbol{s} \in \mathcal{S}\}, \tag{A.2}$$

$$a\mathcal{S} := \{a\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}\}, \tag{A.3}$$

$$\mathcal{R} \cup \mathcal{S} := \{\boldsymbol{x} : \boldsymbol{x} \in \mathcal{R} \text{ or } \boldsymbol{x} \in \mathcal{S}\}. \tag{A.4}$$

$$\mathcal{R} \cap \mathcal{S} := \{\boldsymbol{x} : \boldsymbol{x} \in \mathcal{R} \text{ and } \boldsymbol{x} \in \mathcal{S}\}. \tag{A.5}$$

Exercise A.23 *Let $\mathcal{R} = \{(x, y) : x^2 + y^2 \leq 1\}$ be the unit disc in $\mathbb{R}^2$ and set $\mathcal{S} = \{(x, y) : (x - \frac{1}{2})^2 + y^2 \leq 1\}$. Find $\mathcal{R} + \mathcal{S}, 2\mathcal{S}, \mathcal{R} \cup \mathcal{S},$, and $\mathcal{R} \cap \mathcal{S}$.*

A.3.1 Sums and intersections of subspaces

In many cases $\mathcal{R}$ and $\mathcal{S}$ will be subspaces. Then $a\mathcal{S} = \mathcal{S}$ and both the sum and intersection of two subspaces is a subspace of $(\mathcal{V}, \mathbb{F})$. Note however that the union $\mathcal{R} \cup \mathcal{S}$ of two subspaces is not necessarily a subspace.

Exercise A.24 *Let $\mathcal{R}$ and $\mathcal{S}$ be two subspaces of a vector space $(\mathcal{V}, \mathbb{F})$. Show that $a\mathcal{S} = \mathcal{S}$ and that both $\mathcal{R} + \mathcal{S}$ and $\mathcal{R} \cap \mathcal{S}$ are subspaces of $(\mathcal{V}, \mathbb{F})$.*

Example A.25 *For given vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ with $\boldsymbol{x}$ and $\boldsymbol{y}$ linearly independent let $\mathcal{R} = \text{span}\{\boldsymbol{x}\}$ and $\mathcal{S} = \text{span}\{\boldsymbol{y}\}$. Then $\mathcal{R}$ and $\mathcal{S}$ are subspaces of $\mathbb{R}^n$. For $n = 2$ we have $\mathcal{R} + \mathcal{S} = \mathbb{R}^2$, while for $n = 3$ the sum represents a plane passing through the origin. We also see that $\mathcal{R} \cap \mathcal{S} = \{\boldsymbol{0}\}$ and that $\mathcal{R} \cup \mathcal{S}$ is not a subspace.*

Exercise A.26 *Show the statements made in Example A.25.*

Theorem A.27 *Let $\mathcal{R}$ and $\mathcal{S}$ be two subspaces of a vector space $(\mathcal{V}, \mathbb{F})$. Then*

$$\dim(\mathcal{R} + \mathcal{S}) = \dim(\mathcal{R}) + \dim(\mathcal{S}) - \dim(\mathcal{R} \cap \mathcal{S}). \tag{A.6}$$

Proof. Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$ be a basis for $\mathcal{R} \cap \mathcal{S}$, where $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\} = \emptyset$, the empty set, in the case $\mathcal{R} \cap \mathcal{S} = \{\boldsymbol{0}\}$. We use Theorem A.19 to extend $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$ to a basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_q\}$ for $\mathcal{R}$ and a basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_t\}$ for $\mathcal{S}$. Every $\boldsymbol{x} \in \mathcal{R} + \mathcal{S}$ can be written as a linear combination of $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_q, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_t\}$ so these vectors span $\mathcal{R} + \mathcal{S}$. We show that they are linearly independent and hence a basis. Suppose $\boldsymbol{u} + \boldsymbol{r} + \boldsymbol{s} = \boldsymbol{0}$, where $\boldsymbol{u} := \sum_{j=1}^{p} \alpha_j \boldsymbol{u}_j$, $\boldsymbol{r} := \sum_{j=1}^{q} \rho_j \boldsymbol{r}_j$, and $\boldsymbol{s} := \sum_{j=1}^{t} \sigma_j \boldsymbol{s}_j$. Now $\boldsymbol{r} = -(\boldsymbol{u} + \boldsymbol{s})$ belongs to both $\mathcal{R}$ and to $\mathcal{S}$ and hence $\boldsymbol{r} \in \mathcal{R} \cap \mathcal{S}$. Therefore $\boldsymbol{r}$ can be written as a linear combination of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ say $\boldsymbol{r} := \sum_{j=1}^{p} \beta_j \boldsymbol{u}_j$ and at the same time as a linear combination of $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_q$. But then $\boldsymbol{0} = \sum_{j=1}^{p} \beta_j \boldsymbol{u}_j - \sum_{j=1}^{q} \rho_j \boldsymbol{r}_j$ and since $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_q\}$ is linearly independent we must have $\beta_1 = \cdots = \beta_p = \rho_1 = \cdots = \rho_q = 0$ and hence $\boldsymbol{r} = \boldsymbol{0}$. We now have $\boldsymbol{u} + \boldsymbol{s} = \boldsymbol{0}$ and by linear independence of $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_t\}$ we obtain $\alpha_1 = \cdots = \alpha_p = \sigma_1 = \cdots = \sigma_t = 0$. We have shown that the vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_q, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_t\}$ constitute a basis for $\mathcal{R} + \mathcal{S}$. The result now follows from a simple calculation

$$\dim(\mathcal{R} + \mathcal{S}) = p + q + t = (p + q) + (p + t) - p = \dim(\mathcal{R}) + \dim(\mathcal{S}) - \dim(\mathcal{R} \cap \mathcal{S}).$$

□

From this theorem it follows that $\dim(\mathcal{R} + \mathcal{S}) = \dim(\mathcal{R}) + \dim(\mathcal{S})$ provided $\mathcal{R} \cap \mathcal{S} = \{\boldsymbol{0}\}$.

Definition A.28 (Direct Sum) *Let $\mathcal{R}$ and $\mathcal{S}$ be two subspaces of a vector space $(\mathcal{V}, \mathbb{F})$. If $\mathcal{R} \cap \mathcal{S} = \{\boldsymbol{0}\}$ then the subspace $\mathcal{R} + \mathcal{S}$ is called a* **direct sum** *and denoted $\mathcal{R} \oplus \mathcal{S}$. The subspaces $\mathcal{R}$ and $\mathcal{S}$ are called* **complementary** *in the subspace $\mathcal{R} \oplus \mathcal{S}$.*

Theorem A.29 *Let $\mathcal{R}$ and $\mathcal{S}$ be two subspaces of a vector space $(\mathcal{V}, \mathbb{F})$ and assume $\mathcal{R} \cap \mathcal{S} = \{\boldsymbol{0}\}$. Every $\boldsymbol{x} \in \mathcal{R} \oplus \mathcal{S}$ can be decomposed uniquely in the form $\boldsymbol{x} = \boldsymbol{r} + \boldsymbol{s}$, where $\boldsymbol{r} \in \mathcal{R}$ and $\boldsymbol{s} \in \mathcal{S}$. If $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k\}$ is a basis for $\mathcal{R}$ and $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ is a basis for $\boldsymbol{S}$ then $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ is a basis for $\mathcal{R} \oplus \mathcal{S}$.*

Proof. To show uniqueness, suppose we could write $\boldsymbol{x} = \boldsymbol{r}_1 + \boldsymbol{s}_1 = \boldsymbol{r}_2 + \boldsymbol{s}_2$ for $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \mathcal{R}$ and $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{S}$. Then $\boldsymbol{r}_1 - \boldsymbol{r}_2 = \boldsymbol{s}_2 - \boldsymbol{s}_1$ and it follows that $\boldsymbol{r}_1 - \boldsymbol{r}_2$ and $\boldsymbol{s}_2 - \boldsymbol{s}_1$ belong to both $\mathcal{R}$ and $\mathcal{S}$ and hence to $\mathcal{R} \cap \mathcal{S}$. But then $\boldsymbol{r}_1 - \boldsymbol{r}_2 = \boldsymbol{s}_2 - \boldsymbol{s}_1 = \boldsymbol{0}$ so $\boldsymbol{r}_1 = \boldsymbol{r}_2$ and $\boldsymbol{s}_2 = \boldsymbol{s}_1$. Thus uniqueness follows. Suppose $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k\}$ is a basis for $\mathcal{R}$ and $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ is a basis for $\boldsymbol{S}$. Since $\dim(\mathcal{R} + \mathcal{S}) = \dim(\mathcal{R}) + \dim(\mathcal{S})$ the vectors $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ span $\mathcal{R} + \mathcal{S}$. To show linear independence suppose $\sum_{j=1}^{k} \rho_j \boldsymbol{r}_j + \sum_{j=1}^{n} \sigma_j \boldsymbol{s}_j = \boldsymbol{0}$. The first sum belongs to $\mathcal{R}$ and the second to $\mathcal{S}$ and the sum is a decomposition of $\boldsymbol{0}$. By uniqueness of the decomposition both sums must be zero. But then $\rho_1 = \cdots = \rho_k = \sigma_1 = \cdots = \sigma_n = 0$ and linear independence follows. □

A.3.2 The quotient space

For the sum of two sets we write $\boldsymbol{x}+\mathcal{S} := \{\boldsymbol{x}+\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}\}$ when one of the sets is a singleton set $\{\boldsymbol{x}\}$. Suppose $\mathcal{S}$ is a subspace of a vector space $(\mathcal{X}, \mathbb{F})$. Since $a\mathcal{S} = \mathcal{S}$ we have

$$a(\boldsymbol{x}+\mathcal{S}) + b(\boldsymbol{y}+\mathcal{S}) = (a\boldsymbol{x}+b\boldsymbol{y}) + \mathcal{S}, \text{ for all } a, b \in \mathbb{F} \text{ and all } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}.$$

The set

$$\mathcal{X}/\mathcal{S} := \{\boldsymbol{x}+\mathcal{S} : \boldsymbol{x} \in \mathcal{X}\} \tag{A.7}$$

is a vector space if we define

$$a(\boldsymbol{x}+\mathcal{S}) + b(\boldsymbol{y}+\mathcal{S}) := (a\boldsymbol{x}+b\boldsymbol{y}) + \mathcal{S}, \text{ for all } a, b \in \mathbb{F} \text{ and all } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}.$$

The space $\mathcal{X}/\mathcal{S}$ is called the **quotient space** of $\mathcal{X}$ by $\mathcal{S}$. The zero element in $\mathcal{X}/\mathcal{S}$ is $\mathcal{S}$ itself. Moreover, if $\boldsymbol{x}+\mathcal{S} = \boldsymbol{y}+\mathcal{S}$ then $\boldsymbol{x}-\boldsymbol{y} \in \mathcal{S}$.

Exercise A.30 *Show that $\mathcal{X}/\mathcal{S}$ is a vector space.*

Theorem A.31 *Suppose $\mathcal{S}$ is a subspace of a finite dimensional vector space $(\mathcal{X}, \mathbb{F})$. Then*

$$\dim(\mathcal{S}) + \dim\left(\mathcal{X}/\mathcal{S}\right) = \dim(\mathcal{X}). \tag{A.8}$$

Proof. Let $n := \dim(\mathcal{X})$, $k = \dim(\mathcal{S})$, and let $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ be a basis for $\mathcal{S}$. By Theorem A.19 we can extend it to a basis $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k, \boldsymbol{t}_{k+1}, \ldots, \boldsymbol{t}_n\}$ for $\mathcal{X}$. The result will follow if we can show that $\{\boldsymbol{t}_{k+1}+\mathcal{S}, \ldots, \boldsymbol{t}_n+\mathcal{S}\}$ is a basis for $\mathcal{X}/\mathcal{S}$. Recall that the zero element in $\mathcal{X}/\mathcal{S}$ is $\mathcal{S}$. To show linear independence suppose $\sum_{j=k+1}^{n} a_j(\boldsymbol{t}_j+\mathcal{S}) = \mathcal{S}$ for some $a_{k+1}, \ldots, a_n$ in $\mathbb{F}$. Since $\sum_{j=k+1}^{n} a_j\mathcal{S} = \mathcal{S}$ and the zero element in $\mathcal{X}/\mathcal{S}$ is unique we must have $\sum_{j=k+1}^{n} a_j\boldsymbol{t}_j = \boldsymbol{0}$ which implies that $a_{k+1} = \cdots = a_n = 0$ by linear independence of the $\boldsymbol{t}$'s. It remains to show that $\text{span}\{\boldsymbol{t}_{k+1}+\mathcal{S}, \ldots, \boldsymbol{t}_n+\mathcal{S}\} = \mathcal{X}/\mathcal{S}$. Suppose $\boldsymbol{x}+\mathcal{S} \in \mathcal{X}/\mathcal{S}$. For some $a_1, \ldots, a_n$ we have $\boldsymbol{x} = \boldsymbol{x}_1+\boldsymbol{x}_2$, where $\boldsymbol{x}_1 = \sum_{j=1}^{k} a_j\boldsymbol{s}_j$ and $\boldsymbol{x}_2 = \sum_{j=k+1}^{n} a_j\boldsymbol{t}_j$. Since $\boldsymbol{x}_1+\mathcal{S} = \mathcal{S}$ we have $\boldsymbol{x}+\mathcal{S} = \boldsymbol{x}_2+\mathcal{S} = \sum_{j=k+1}^{n} a_j\boldsymbol{t}_j + \mathcal{S} = \sum_{j=k+1}^{n} a_j(\boldsymbol{t}_j+\mathcal{S}) \in \mathcal{X}/\mathcal{S}$. □

A.4 Convergence of Vectors

Consider an infinite sequence $\{\boldsymbol{x}_k\} = \boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ of vectors in $\mathbb{R}^n$. This sequence converges to zero if and only if each component sequence $\boldsymbol{x}_k(j)$ converges to zero for $j = 1, \ldots, n$. In terms of the natural basis we have $\boldsymbol{x}_k = \sum_{j=1}^{n} \boldsymbol{x}_k(j)\boldsymbol{e}_k$ and another way of stating convergence to zero is that in terms of the basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ for $\mathbb{R}^n$ each coefficient $\boldsymbol{x}_k(j)$ of $\boldsymbol{x}_k$ converges to zero.

Consider now a more general vector space.

Definition A.32 *Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a basis for a finite dimensional vector space $(\mathcal{V}, \mathbb{F})$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, and let $\{\boldsymbol{x}_k\}$ be an infinite sequence of vectors in*

$\mathcal{V}$ with basis coefficients $\{\boldsymbol{c}_k\}$, i.e. $\boldsymbol{x}_k = \sum_{j=1}^n c_{kj}\boldsymbol{v}_j$ for each k. We say that $\{\boldsymbol{x}_k\}$ converges to zero, or have the limit zero, if $\lim_{k\to\infty} c_{kj} = 0$ for $j = 1, \ldots, n$. We say that $\{\boldsymbol{x}_k\}$ converge to the limit $\boldsymbol{x}$ in $\mathcal{V}$ if $\boldsymbol{x}_k - \boldsymbol{x}$ converges to zero. We write this as $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}$ or $\boldsymbol{x}_k \to \boldsymbol{x}$ (as $k \to \infty$).

This definition is actually independent of the basis chosen. If $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ is another basis for $\mathcal{V}$ and $\boldsymbol{x}_k = \sum_{j=1}^n b_{kj}\boldsymbol{w}_j$ for each k then from Lemma A.20 $\boldsymbol{b}_k = \boldsymbol{A}\boldsymbol{c}_k$ for some nonsingular matrix $\boldsymbol{A}$ independent of k. Hence $\boldsymbol{c}_k \to \boldsymbol{0}$ if and only if $\boldsymbol{b}_k \to \boldsymbol{0}$. If $\{a_k\}$ and $\{b_k\}$ are sequences of scalars and $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ are sequences of vectors such that $\{a_k\} \to a$, $\{b_k\} \to b$, $\{\boldsymbol{x}_k\} \to \boldsymbol{x}$, and $\{\boldsymbol{y}_k\} \to \boldsymbol{y}$ then $\{a_k\boldsymbol{x}_k + b_k\boldsymbol{y}_k\} \to a\boldsymbol{x} + b\boldsymbol{y}$. This shows that scalar multiplication and vector addition are continuous functions with respect to this notion of limit.

Corresponding to a basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$, we define

$$\|\boldsymbol{x}\|_c := \max_{1\le j\le n} |c_j| \text{ where } \boldsymbol{x} = \sum_{j=1}^n c_j\boldsymbol{v}_j.$$

We leave as an exercise to show that this is a norm on $\mathcal{V}$. Recall that any two norms on $\mathcal{V}$ are equivalent. This implies that for any other norm $\|\cdot\|$ on $\mathcal{V}$ there are positive constants α, β such that any $\boldsymbol{x} = \sum_{j=1}^n c_j\boldsymbol{v}_j$ satisfy

$$\|\boldsymbol{x}\| \le \alpha \max_{1\le j\le n} |c_j| \text{ and } |c_j| \le \beta\|\boldsymbol{x}\| \text{ for } j = 1, \ldots, n. \tag{A.9}$$

Suppose now $(\mathcal{V}, \mathbb{F}, \|\cdot\|)$ is a normed vector space with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. The notion of limit can then be stated in terms of convergence in norm.

Theorem A.33 *In a normed vector space we have $\boldsymbol{x}_k \to \boldsymbol{x}$ if and only if $\lim_{k\to\infty}\|\boldsymbol{x}_k - \boldsymbol{x}\| = 0$.*

Proof. Suppose $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis for the vector space and assume $\boldsymbol{x}_k, \boldsymbol{x} \in \mathcal{V}$. Then $\boldsymbol{x}_k - \boldsymbol{x} = \sum_{j=1}^n c_{kj}\boldsymbol{v}_j$ for some scalars c_{kj} By (A.9) we see that

$$\frac{1}{\beta}\max_{k,j}|c_{kj}| \le \|\boldsymbol{x}_k - \boldsymbol{x}\| \le \alpha\max_{k,j}|c_{kj}|$$

and hence $\|\boldsymbol{x}_k - \boldsymbol{x}\| \to 0 \Leftrightarrow \lim_k c_{kj} \to 0$ for each $j \Leftrightarrow \boldsymbol{x}_k \to \boldsymbol{x}$. ∎

Since all vector norms are equivalent we have convergence in any norm we can define on a finite dimensional vector space.

Definition A.34 *Let $(\mathcal{V}, \mathbb{F}, \|\cdot\|)$ be a normed vector space and let $\{\boldsymbol{x}_k\}$ in $\mathcal{V}$ be an infinite sequence.*

1. *$\{\boldsymbol{x}_k\}$ is a* **Cauchy sequence** *if $\lim_{k,l\to\infty}(\boldsymbol{x}_k - \boldsymbol{x}_l) = 0$ or equivalently $\lim_{k,l\to\infty}\|\boldsymbol{x}_k - \boldsymbol{x}_l\| = 0$. More precisely, for each $\epsilon > 0$ there is an integer $N \in \mathbb{N}$ such that for each $k, l \ge N$ we have $\|\boldsymbol{x}_k - \boldsymbol{x}_l\| \le \epsilon$.*
2. *The normed vector space is said to be* **complete** *if every Cauchy sequence converges to a point in the space.*

3. $\{\boldsymbol{x}_k\}$ *is called* **bounded** *if there is a positive number* M *such that* $\|\boldsymbol{x}_k\| \leq M$ *for all* k.

4. $\{\boldsymbol{x}_{n_k}\}$ *is said to be a* **subsequence** *of* $\{\boldsymbol{x}_k\}_{k\geq 0}$ *if* $0 \leq n_0 < n_1 < n_2 \cdots$.

Theorem A.35 *In a finite dimensional vector space* $\mathcal{V}$ *the following hold:*

1. *A sequence in* $\mathcal{V}$ *is convergent if and only if it is a Cauchy sequence.*
2. $\mathcal{V}$ *is complete.*
3. *Every bounded sequence in* $\mathcal{V}$ *has a convergent subsequence.*

Proof.

1. Suppose $\boldsymbol{x}_k \to \boldsymbol{x}$. By the triangle inequality $\|\boldsymbol{x}_k - \boldsymbol{x}_l\| \leq \|\boldsymbol{x}_k - \boldsymbol{x}\| + \|\boldsymbol{x}_l - \boldsymbol{x}\|$ and hence $\|\boldsymbol{x}_k - \boldsymbol{x}_l\| \to 0$. Conversely, let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a basis for $\mathcal{V}$ and $\{\boldsymbol{x}_k\}$ a Cauchy sequence with $\boldsymbol{x}_k = \sum_{j=1}^n c_{kj}\boldsymbol{v}_j$ for each k. Then $\boldsymbol{x}_k - \boldsymbol{x}_l = \sum_{j=1}^n (c_{kj} - c_{lj})\boldsymbol{v}_j$ and since $\lim_{k,l\to\infty}(\boldsymbol{x}_k - \boldsymbol{x}_l) = 0$ we have by definition of convergence $\lim_{k,l\to\infty}(c_{kj} - c_{lj}) = 0$ for $j = 1, \ldots, n$. Thus for each j we have a Cauchy-sequence $\{c_{kj}\} \in \mathbb{C}$ and since $\mathbb{C}$ is complete $\{c_{kj}\}$ converges to some $c_j \in \mathbb{C}$. But then $\boldsymbol{x}_k \to \boldsymbol{x} := \sum_{j=1}^n c_j\boldsymbol{v}_j \in \mathcal{V}$.
2. $\mathcal{V}$ is complete since we just showed that every Cauchy sequence converges to a point in the space.
3. Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a basis for $\mathcal{V}$ and $\{\boldsymbol{x}_k\}$ be a bounded sequence with $\boldsymbol{x}_k = \sum_{j=1}^n c_{kj}\boldsymbol{v}_j$ for each k. By (A.9) each coefficient sequence $\{c_{kj}\}_k$ is a bounded sequence of complex numbers and therefore, by a well known property of complex numbers, has a convergent subsequence. In particular the sequence of $\boldsymbol{v}_1$ coefficients $\{c_{k1}\}$ has a convergent subsequence $c_{k_i,1}$. For the second component the sequence $\{c_{k_i,2}\}$ has a convergent subsequence, say $c_{l_i,2}$. Continuing with $j = 3, \ldots, n$ we obtain integers $0 \leq m_0 < m_1 < \cdots$ such that $\{c_{m_i,j}\}$ is a convergent subsequence of c_{kj} for $j = 1, \ldots, n$. But then $\{\boldsymbol{x}_{m_i}\}$ is a convergent subsequence of $\{\boldsymbol{x}_k\}$.

□

A.4.1 Convergence of Series of Vectors

Consider now an infinite series $\sum_{m=0}^\infty \boldsymbol{y}_m$ of vectors in a vector space $(\mathcal{V}, \mathbb{F})$ with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. We say that the series converges if the sequence of partial sums $\{\boldsymbol{x}_k\}$ given by $\boldsymbol{x}_k = \sum_{m=0}^k \boldsymbol{y}_m$ converges. A sufficient condition for convergence is that $\sum_{m=0}^\infty \|\boldsymbol{y}_m\|$ converges for some vector norm. We say that the series converges **absolutely** if this is the case. Note that $\|\sum_{m=0}^\infty \boldsymbol{y}_m\| \leq \sum_{m=0}^\infty \|\boldsymbol{y}_m\|$, and absolute convergence in one norm implies absolute convergence in any norm by Theorem 8.3. In an absolute convergent series we may change the order of the terms without changing the value of the sum.

Exercise A.36 *Show that if* $\{a_k\} \to a$, $\{b_k\} \to b$, $\{\boldsymbol{x}_k\} \to \boldsymbol{x}$, *and* $\{\boldsymbol{y}_k\} \to \boldsymbol{y}$ *then* $\{a_k\boldsymbol{x}_k + b_k\boldsymbol{y}_k\} \to a\boldsymbol{x} + b\boldsymbol{y}$.

Exercise A.37 *Show that* $\|\cdot\|_c$ *is a norm.*

A.5 Inner Products

An **inner product** or **scalar product** in a vector space $(\mathcal{V}, \mathbb{F})$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, is a function $\langle\cdot,\cdot\rangle$ mapping pairs of vectors into a scalar. We consider first the case where $\mathbb{F} = \mathbb{R}$.

Definition A.38 *An inner product in a vector space* $(\mathcal{V}, \mathbb{R})$ *is a function* $\mathcal{V}\times\mathcal{V} \to \mathbb{R}$ *satisfying for all* $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$ *and all* $a, b \in \mathbb{R}$ *the following conditions:*

1. $\langle \boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$ *with equality if and only if* $\boldsymbol{x} = \boldsymbol{0}$. *(positivity)*
2. $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{y}, \boldsymbol{x}\rangle$ *(symmetry)*
3. $\langle a\boldsymbol{x} + b\boldsymbol{y}, \boldsymbol{z}\rangle = a\langle \boldsymbol{x}, \boldsymbol{z}\rangle + b\langle \boldsymbol{y}, \boldsymbol{z}\rangle$. *(linearity)*

The triple $(\mathcal{V}, \mathbb{R}, \langle\cdot,\cdot,\rangle)$ *is called a* **real inner product space**

The **standard inner product in** $\mathcal{V} = \mathbb{R}^n$ is given by $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \boldsymbol{x}^T\boldsymbol{y}$. It is clearly an inner product in $\mathbb{R}^n$.

When the field of scalars is $\mathbb{C}$ the inner product is complex valued and properties 2. and 3. are altered as follows:

Definition A.39 *An inner product in a vector space* $(\mathcal{V}, \mathbb{C})$ *is a function* $\mathcal{V}\times\mathcal{V} \to \mathbb{C}$ *satisfying for all* $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$ *and all* $a, b \in \mathbb{C}$ *the following conditions:*

1. $\langle \boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$ *with equality if and only if* $\boldsymbol{x} = \boldsymbol{0}$. *(positivity)*
2. $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \overline{\langle \boldsymbol{y}, \boldsymbol{x}\rangle}$ *(skew symmetry)*
3. $\langle a\boldsymbol{x} + b\boldsymbol{y}, \boldsymbol{z}\rangle = a\langle \boldsymbol{x}, \boldsymbol{z}\rangle + b\langle \boldsymbol{y}, \boldsymbol{z}\rangle$. *(linearity)*

The triple $(\mathcal{V}, \mathbb{C}, \langle\cdot,\cdot,\rangle)$ *is called a* **complex inner product space**

Note the complex conjugate in 2. and that (Cf. Exercise A.45)

$$\langle \boldsymbol{x}, a\boldsymbol{y} + b\boldsymbol{z}\rangle = \overline{a}\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \overline{b}\langle \boldsymbol{x}, \boldsymbol{z}\rangle. \tag{A.10}$$

The **standard inner product in** $\mathbb{C}^n$ is given by $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \boldsymbol{x}^*\boldsymbol{y} = \sum_{j=1}^n \overline{x_j}y_j$. It is clearly an inner product in $\mathbb{C}^n$.

Suppose now $(\mathcal{V}, \mathbb{F}, \langle\cdot,\cdot,\rangle)$ is an inner product space with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. We define the **inner product norm** by

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}, \quad \boldsymbol{x} \in \mathcal{V}.$$

For any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$ and scalar $a \in \mathbb{F}$ we have (Cf. Exercises A.44 and A.45) by linearity and symmetry the expansion

$$\|\boldsymbol{x} + a\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + 2a\langle \boldsymbol{x}, \boldsymbol{y}\rangle + a^2\|\boldsymbol{y}\|^2 \qquad \text{(real case)}, \tag{A.11}$$

$$= \|\boldsymbol{x}\|^2 + 2\mathrm{Re}\langle \boldsymbol{x}, a\boldsymbol{y}\rangle + |a|^2\|\boldsymbol{y}\|^2 \quad \text{(complex case)}, \tag{A.12}$$

where $\mathrm{Re}\, z$ and $\mathrm{Im}\, z$ denotes the real- and imaginary part of the complex number z.

In the complex case we can write the inner product of two vectors as a sum of inner product norms. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$ it follows from (A.12) that

$$4\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2 + i\|\boldsymbol{x} - i\boldsymbol{y}\|^2 - i\|\boldsymbol{x} + i\boldsymbol{y}\|^2, \tag{A.13}$$

where $i = \sqrt{-1}$ and we used that $\mathrm{Im}(z) = \mathrm{Re}(-iz)$ for any $z \in \mathbb{C}$.

To show that the inner product norm is a norm in $(\mathcal{V}, \mathbb{R})$ we need the triangle inequality. To show it we start with a famous inequality.

Theorem A.40 (Cauchy-Schwarz inequality) *For any $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space*

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$$

with equality if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent.

Proof. The inequality is trivial if $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = 0$ so assume $\langle \boldsymbol{x}, \boldsymbol{y}\rangle \neq 0$. Suppose first $\langle \boldsymbol{x}, \boldsymbol{y}\rangle \in \mathbb{R}$. We define the scalar $a := -\frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|^2}$, and use (A.11) to obtain $0 \leq \|\boldsymbol{x} + a\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 - \big(\langle \boldsymbol{x}, \boldsymbol{y}\rangle\big)^2/\|\boldsymbol{y}\|^2$. Thus the inequality follows in the real case. Suppose next $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ is complex valued, say $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = re^{i\phi}$. We define $b := e^{-i\phi}$ and observe that $b\langle \boldsymbol{x}, \boldsymbol{y}\rangle = r$ is real valued and $|b| = 1$. Using the real case of the Cauchy-Schwarz inequality we find

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = |b\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = |\langle b\boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|b\boldsymbol{x}\|\|\boldsymbol{y}\| = \|\boldsymbol{x}\|\|\boldsymbol{y}\|$$

which proves the inequality also in the complex case. We have equality if and only if $\boldsymbol{x} + a\boldsymbol{y} = \boldsymbol{0}$ which means that $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent. □

Theorem A.41 (Triangle Inequality) *For any $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space*

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|.$$

Proof. From the Cauchy-Schwarz inequality it follows that $\mathrm{Re}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$. Using this on the inner product term in (A.12) with $a = 1$ we get

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq \|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2 = (\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2.$$

Taking square roots completes the proof. □

Theorem A.42 (Parallelogram Identity) *For all $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space*

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 + \|\boldsymbol{x}-\boldsymbol{y}\|^2 = 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2.$$

Proof. We set $a = \pm 1$ in the inner product expansion (A.12) and add the two equations. □

In the real case the Cauchy-Schwarz inequality implies that $-1 \leq \frac{\langle \boldsymbol{x},\boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \leq 1$ for nonzero $\boldsymbol{x}$ and $\boldsymbol{y}$ so there is a unique angle θ in $[0,\pi]$ such that

$$\cos\theta = \frac{\langle \boldsymbol{x},\boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}. \tag{A.14}$$

This defines the **angle** between vectors in a real inner product space.

Exercise A.43 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has linearly independent columns. Show that $\langle \boldsymbol{x},\boldsymbol{y}\rangle := \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{y}$ defines an inner product on $\mathbb{R}^n$.*

Exercise A.44 *Show* (A.11)

Exercise A.45 *Show* (A.10) *and* (A.12).

Exercise A.46 *Show* (A.13)

Exercise A.47 *Show that in the complex case there is a unique angle θ in $[0,\pi/2]$ such that*

$$\cos\theta = \frac{|\langle \boldsymbol{x},\boldsymbol{y}\rangle|}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}. \tag{A.15}$$

A.6 Orthogonality

As in the previous section we assume that $(\mathcal{V}, \mathbb{F}, \langle\cdot,\cdot,\rangle\rangle$ is an inner product space with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. Also $\|\cdot\|$ denotes the inner product norm.

Definition A.48 (Orthogonality) *Two vectors $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space are called* **orthogonal** *or* **perpendicular**, *denoted as $\boldsymbol{x} \perp \boldsymbol{y}$, if $\langle \boldsymbol{x},\boldsymbol{y}\rangle = 0$. The vectors are* **orthonormal** *if in addition $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$.*

For orthogonal vectors it follows from (A.12) that the Pythagorean theorem holds

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2, \quad \text{if} \quad \boldsymbol{x} \perp \boldsymbol{y}.$$

Definition A.49 (Orthogonal- and Orthonormal Bases) *A set of vectors* $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ *in a subspace* $\mathcal{S}$ *of a real or complex inner product space is called an* **orthogonal basis** *for* $\mathcal{S}$ *if it is a basis for* $\mathcal{S}$ *and* $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0$ *for* $i \neq j$*. It is an* **orthonormal basis** *for* $\mathcal{S}$ *if it is a basis for* $\mathcal{S}$ *and* $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij}$ *for all* i, j*.*

A basis for an inner product space can be turned into an orthogonal- or orthonormal basis for the subspace by the following construction.

Theorem A.50 (Gram-Schmidt) *Let* $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ *be a basis for a real or complex inner product space* $(\mathcal{S}, \mathbb{F}, \langle \cdot, \cdot \rangle)$*. Define*

$$\boldsymbol{v}_1 := \boldsymbol{s}_1, \quad \boldsymbol{v}_j := \boldsymbol{s}_j - \sum_{i=1}^{j-1} \frac{\langle \boldsymbol{s}_j, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i, \quad j = 2, \ldots, k. \tag{A.16}$$

Then $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ *is an orthogonal basis for* $\mathcal{S}$ *and the normalized vectors*

$$\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\} := \{\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}, \ldots, \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}\}$$

is an orthonormal basis for $\mathcal{S}$*.*

Proof. To show that $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ is an orthogonal basis for $\mathcal{S}$ we use induction on k. Let $S_j := \operatorname{span}\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_j\}$ for $j = 1, \ldots, k$. Clearly $\boldsymbol{v}_1 = \boldsymbol{s}_1$ is an orthogonal basis for S_1. Suppose for some $j \geq 2$ that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{j-1}$ is an orthogonal basis for S_{j-1} and let $\boldsymbol{v}_j$ be given by (A.16) as a linear combination of $\boldsymbol{s}_j$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{j-1}$. Replacing each of these $\boldsymbol{v}_i$ by a linear combination of $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{j-1}$ we obtain $\boldsymbol{v}_j = \sum_{i=1}^{j} a_i \boldsymbol{s}_i$ for some $a_0, \ldots, a_j$ with $a_j = 1$. Since $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_j$ are linearly independent and $a_j \neq 0$ we deduce that $\boldsymbol{v}_j \neq 0$. By the induction hypothesis

$$\langle \boldsymbol{v}_j, \boldsymbol{v}_l \rangle = \langle \boldsymbol{s}_j, \boldsymbol{v}_l \rangle - \sum_{i=1}^{j-1} \frac{\langle \boldsymbol{s}_j, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \langle \boldsymbol{v}_i, \boldsymbol{v}_l \rangle = \langle \boldsymbol{s}_j, \boldsymbol{v}_l \rangle - \frac{\langle \boldsymbol{s}_j, \boldsymbol{v}_l \rangle}{\langle \boldsymbol{v}_l, \boldsymbol{v}_l \rangle} \langle \boldsymbol{v}_l, \boldsymbol{v}_l \rangle = 0$$

for $l = 1, \ldots, j-1$. Thus $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j$ is an orthogonal basis for S_j.

If $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ is an orthogonal basis for $\mathcal{S}$ then clearly $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ is an orthonormal basis for $\mathcal{S}$. □

Theorem A.51 (Orthogonal Projection) *Let* $\mathcal{S}$ *be a subspace of a finite dimensional real or complex inner product space* $(\mathcal{V}, \mathbb{F}, \langle \cdot, \cdot, \rangle)$*. To each* $\boldsymbol{x} \in \mathcal{V}$ *there is a unique vector* $\boldsymbol{p} \in \mathcal{S}$ *such that*

$$\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0, \quad \textit{for all } \boldsymbol{s} \in \mathcal{S}. \tag{A.17}$$

If $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ *is an orthogonal basis for* $\boldsymbol{S}$ *then*

$$\boldsymbol{p} = \sum_{i=1}^{k} \frac{\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i. \tag{A.18}$$

Proof. Define $\boldsymbol{p}$ by (A.18). Then

$$\langle \boldsymbol{p}, \boldsymbol{v}_j \rangle = \sum_{i=1}^{k} \frac{\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \frac{\langle \boldsymbol{x}, \boldsymbol{v}_j \rangle}{\langle \boldsymbol{v}_j, \boldsymbol{v}_j \rangle} \langle \boldsymbol{v}_j, \boldsymbol{v}_j \rangle = \langle \boldsymbol{x}, \boldsymbol{v}_j \rangle$$

so that by linearity $\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{v}_j \rangle = 0$ for $j = 1, \ldots, k$. But then $\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$. This shows existence of a $\boldsymbol{p}$ satisfying (A.17). For uniqueness suppose $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathcal{S}$ and $\langle \boldsymbol{x} - \boldsymbol{p}_1, \boldsymbol{s} \rangle = \langle \boldsymbol{x} - \boldsymbol{p}_2, \boldsymbol{s} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$. Then $\langle \boldsymbol{x} - \boldsymbol{p}_1, \boldsymbol{s} \rangle - \langle \boldsymbol{x} - \boldsymbol{p}_2, \boldsymbol{s} \rangle = \langle \boldsymbol{p}_2 - \boldsymbol{p}_1, \boldsymbol{s} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$ and in particular $\langle \boldsymbol{p}_2 - \boldsymbol{p}_1, \boldsymbol{p}_2 - \boldsymbol{p}_1 \rangle = 0$ which implies that $\boldsymbol{p}_2 - \boldsymbol{p}_1 = \boldsymbol{0}$ or $\boldsymbol{p}_1 = \boldsymbol{p}_2$. □

Theorem A.52 (Best Approximation) *Let $\mathcal{S}$ be a subspace of a finite dimensional real or complex inner product space $(\mathcal{V}, \mathbb{F}, \langle \cdot, \cdot, \rangle)$. Let $x \in \mathcal{V}$, and $\boldsymbol{p} \in \mathcal{S}$. The following statements are equivalent*

1. $\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0, \quad$ *for all* $\boldsymbol{s} \in \boldsymbol{S}$.

2. $\|\boldsymbol{x} - \boldsymbol{s}\| > \|\boldsymbol{x} - \boldsymbol{p}\|$ *for all* $\boldsymbol{s} \in \mathcal{S}$ *with* $\boldsymbol{s} \neq \boldsymbol{p}$.

Proof. Suppose 1. holds and that $\boldsymbol{s} \neq \boldsymbol{p}$. Using Pythagoras for inner products we have

$$\|\boldsymbol{x} - \boldsymbol{s}\|^2 = \|(\boldsymbol{x} - \boldsymbol{p}) + (\boldsymbol{p} - \boldsymbol{s})\|^2 = \|\boldsymbol{x} - \boldsymbol{p}\|^2 + \|\boldsymbol{p} - \boldsymbol{s}\|^2 > \|\boldsymbol{x} - \boldsymbol{p}\|^2.$$

Conversely, suppose 2. holds. Pick any nonzero $\boldsymbol{s} \in \mathcal{S}$ and define the scalar $a := -\mathrm{Re}\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle / \|\boldsymbol{s}\|^2$. Using (A.12) and the minimality of $\boldsymbol{p}$ we obtain

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{p}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{p} + a\boldsymbol{s}\|^2 &= \|\boldsymbol{x} - \boldsymbol{p}\|^2 + 2a\mathrm{Re}\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle + a^2 \|\boldsymbol{s}\|^2 \\ &= \|\boldsymbol{x} - \boldsymbol{p}\|^2 - \big(\mathrm{Re}\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle\big)/\|\boldsymbol{s}\|^2. \end{aligned}$$

This can only be true if $\mathrm{Re}\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$. Since $\boldsymbol{s} \in \mathcal{S}$ implies that $i\boldsymbol{s} \in \mathcal{S}$, where $i = \sqrt{-1}$, we have

$$0 = \mathrm{Re}\langle \boldsymbol{x} - \boldsymbol{p}, i\boldsymbol{s} \rangle = \mathrm{Re}\big(-i\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle \big) = \mathrm{Im}\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle$$

and hence $\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s} \rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$. □

The vector $\boldsymbol{p}$ is called the **orthogonal projection** of $\boldsymbol{x}$ into $\mathcal{S}$ with respect to $\langle \cdot, \cdot \rangle$., and denoted by $\boldsymbol{p} = P_{\mathcal{S}}\boldsymbol{x}$.

In terms of an orthogonal basis $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ for $\mathcal{S}$ we have the representation

$$\boldsymbol{s} = \sum_{i=1}^{k} \frac{\langle \boldsymbol{s}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i, \text{ all } \boldsymbol{s} \in \mathcal{S}. \tag{A.19}$$

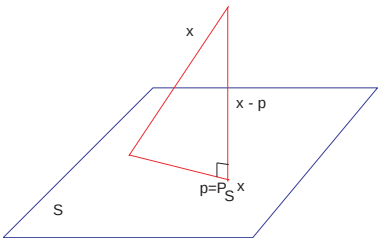


Figure A.1. *The orthogonal projection of $\boldsymbol{x}$ into $\mathcal{S}$.*

A.7 Projections and Orthogonal Complements

Theorem A.53 *Let $\mathcal{S}$ be a subspace in a real or complex inner product space $(\mathcal{V}, \mathbb{F}, \langle\cdot,\cdot,\rangle)$ and let $P_{\mathcal{S}} : \mathbb{R}^n \to \mathcal{S}$ be the operator mapping a vector $\boldsymbol{x} \in \mathcal{V}$ into the orthogonal projection $\boldsymbol{p}$ in $\mathcal{S}$. Then $P_{\mathcal{S}}$ is a* **linear projection operator**, *i.e.*

1. *$P_{\mathcal{S}}(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha P_{\mathcal{S}}\boldsymbol{x} + \beta P_{\mathcal{S}}\boldsymbol{y}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$ and all $\alpha, \beta \in \mathbb{F}$.*
2. *$P_{\mathcal{S}}^2 = P_{\mathcal{S}}$, i.e. $P_{\mathcal{S}}(P_{\mathcal{S}}\boldsymbol{x}) = P_{\mathcal{S}}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{V}$.*

Proof.

1. Let $\boldsymbol{p} := P_{\mathcal{S}}\boldsymbol{x}$ and $\boldsymbol{q} := P_{\mathcal{S}}\boldsymbol{y}$. Then $\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s}\rangle = 0$ and $\langle \boldsymbol{y} - \boldsymbol{q}, \boldsymbol{s}\rangle = 0$ for all $\boldsymbol{s} \in \mathcal{S}$, and by linearity of the inner product

$$\langle \alpha\boldsymbol{x} + \beta\boldsymbol{y} - (\alpha\boldsymbol{p} + \beta\boldsymbol{q}), \boldsymbol{s}\rangle = \alpha\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{s}\rangle + \beta\langle \boldsymbol{y} - \boldsymbol{q}, \boldsymbol{s}\rangle = 0.$$

   But then $\alpha\boldsymbol{p} + \beta\boldsymbol{q} = \alpha P_{\mathcal{S}}\boldsymbol{x} + \beta P_{\mathcal{S}}\boldsymbol{y}$ is the orthogonal projection of $\alpha\boldsymbol{x} + \beta\boldsymbol{y}$ into $\mathcal{S}$ and 1. follows.
2. Since $\boldsymbol{p} = P_{\mathcal{S}}\boldsymbol{x} \in \mathcal{S}$ the uniqueness implies that $P_{\mathcal{S}}\boldsymbol{p} = \boldsymbol{p}$ which gives 2.

□

Definition A.54 (Orthogonal Complement) *Let $\mathcal{S}$ be a subspace in a real or complex inner product space $(\mathcal{V}, \mathbb{F}, \langle\cdot,\cdot,\rangle)$. The Orthogonal Complement of $\mathcal{S}$, which is denoted by $\mathcal{S}^\perp$, consists of all vectors in $\mathcal{V}$ that are orthogonal to every $\boldsymbol{s} \in \mathcal{S}$. In other words*

$$\boldsymbol{x} \in \mathcal{S}^\perp \iff \langle \boldsymbol{x}, \boldsymbol{s}\rangle = 0, \text{ for all } \boldsymbol{s} \in \mathcal{S}.$$

Clearly $\mathcal{S}^\perp$ is a subspace of $\mathcal{V}$.

Theorem A.55 (Orthogonal Decomposition) *For each subspace $\mathcal{S}$ of a real or complex inner product space we have the direct sum decomposition $\mathcal{V} = \mathcal{S} \oplus \mathcal{S}^\perp$. If $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ is a basis for $\mathcal{S}$ and $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$ is a basis for $\boldsymbol{S}^\perp$ then $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$ is a basis for $\mathcal{S} \oplus \mathcal{S}^\perp$. In particular, any orthonormal basis for $\mathcal{S}$ can be extended to an orthonormal basis for $\mathcal{V}$.*

Proof. If $\boldsymbol{x} \in \mathcal{S} \cap \mathcal{S}^\perp$ then $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ so $\boldsymbol{x} = \mathbf{0}$.This means that $\mathcal{S} \cap \mathcal{S}^\perp = \{0\}$ and $\mathcal{S} \oplus \mathcal{S}^\perp$ is a direct sum. Every $\boldsymbol{x} \in \mathbb{R}^n$ can be decomposed as $\boldsymbol{x} = P_{\mathcal{S}}\boldsymbol{x} + (\boldsymbol{x} - P_{\mathcal{S}}\boldsymbol{x})$ where $P_{\mathcal{S}}\boldsymbol{x} \in \mathcal{S}$ and $\boldsymbol{x} - P_{\mathcal{S}}\boldsymbol{x} \in \mathcal{S}^\perp$. Since we are dealing with a direct sum it follows from Theorem A.29 that any basis $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ for $\mathcal{S}$ and any basis $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$ for $\mathcal{S}^\perp$ can be combined into a basis for $\mathcal{V}$. If $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ is an orthonormal basis for $\mathcal{S}$ then we apply the Gram-Schmidt process to $(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$ to obtain a combined orthonormal basis for $\mathcal{V}$. □

Exercise A.56 *Show that* $\left(\mathcal{S}^\perp\right)^\perp = \mathcal{S}$ *for any subspace* $\mathcal{S}$ *of a real or complex inner product space.*

Appendix B

Matrices

In this chapter we review some topics related to matrices. In Section B.1 we study block-multiplication, a basic tool in matrix analysis. We then review the transpose matrix, linear systems, and inverse matrices. We end the chapter with some basic facts about orthonormal-, and unitary matrices.

Some matrices with many zeros have names indicating their "shape". Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ or $\boldsymbol{A} \in \mathbb{C}^{n,n}$. Then $\boldsymbol{A}$ is

- **diagonal** if $a_{ij} = 0$ for $i \neq j$.
- **upper triangular** or **right triangular** if $a_{ij} = 0$ for $i > j$.
- **lower triangular** or **left triangular** if $a_{ij} = 0$ for $i < j$.
- **upper Hessenberg** if $a_{ij} = 0$ for $i > j + 1$.
- **lower Hessenberg** if $a_{ij} = 0$ for $i < j + 1$.
- **tridiagonal** if $a_{ij} = 0$ for $|i - j| > 1$.
- **lower banded** with bandwidth p if $a_{ij} = 0$ for $i > j + p$.
- **upper banded** with bandwidth q if $a_{ij} = 0$ for $i < j + q$.
- **banded** with bandwidth $p + q + 1$ if $\boldsymbol{A}$ is both lower banded with bandwidth p and upper banded with bandwidth q.
- **block upper triangular** if there is an integer k such that $a_{ij} = 0$ for $i = k + 1, \ldots, n$ and $j = 1, \ldots, k$.
- **block lower triangular** if $\boldsymbol{A}^T$ is block upper triangular.

B.1 Arithmetic Operations and Block Multiplication

The arithmetic operations on rectangular matrices are

- **matrix addition** $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$ if $c_{ij} = a_{ij} + b_{ij}$ for all i, j and $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are matrices of the same dimension.
- **multiplication by a scalar** $\boldsymbol{C} = \alpha \boldsymbol{A}$, where $c_{ij} = \alpha a_{ij}$ for all i, j.

- **multiplication by another matrix** $\boldsymbol{C} = \boldsymbol{AB}$, $\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B}$ or $\boldsymbol{C} = \boldsymbol{A} * \boldsymbol{B}$, where $\boldsymbol{A} \in \mathbb{C}^{m,p}$, $\boldsymbol{B} \in \mathbb{C}^{p,n}$, $\boldsymbol{C} \in \mathbb{C}^{m,n}$, and $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$.
- **entry-by-entry matrix operations**(add a dot) $\boldsymbol{C} = \boldsymbol{A}.*\boldsymbol{B}$ and $D = \boldsymbol{A}./\boldsymbol{B}$, and $E = \boldsymbol{A}.\wedge r$ where all matrices are of the same dimension and $c_{ij} = a_{ij} b_{ij}$, $d_{ij} = a_{ij}/b_{ij}$ and $e_{ij} = a_{ij}^{r}$ for all i, j and suitable r. The entry-by-entry product $\boldsymbol{C} = \boldsymbol{A}.*\boldsymbol{B}$ is known as the **Schur product** and also the **Hadamard product**.

Example B.1 (The Vector Space of $m \times n$ matrices) *On the set $\mathbb{C}^{m,n}$ of $m \times n$ matrices we define vector addition as matrix addition and scalar multiplication as a scalar times a matrix. Then $\mathbb{C}^{m,n} = (\mathbb{C}^{m,n}, \mathbb{C})$ is a vector space. Of course $\mathbb{R}^{m,n} = (\mathbb{R}^{m,n}, \mathbb{R})$ is also a vector space.*

A rectangular matrix $\boldsymbol{A}$ can be partitioned into submatrices by drawing horizontal lines between selected rows and vertical lines between selected columns. For example, the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

can be partitioned as

$$(i)\ \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} = \left[\begin{array}{c|cc} 1 & 2 & 3 \\ \hline 4 & 5 & 6 \\ 7 & 8 & 9 \end{array}\right], \quad (ii)\ \left[\boldsymbol{a}_{:1}, \boldsymbol{a}_{:2}, \boldsymbol{a}_{:3}\right] = \left[\begin{array}{c|c|c} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array}\right],$$

$$(iii)\ \begin{bmatrix} \boldsymbol{a}_{1:}^{T} \\ \boldsymbol{a}_{2:}^{T} \\ \boldsymbol{a}_{3:}^{T} \end{bmatrix} = \left[\begin{array}{ccc} 1 & 2 & 3 \\ \hline 4 & 5 & 6 \\ \hline 7 & 8 & 9 \end{array}\right], \quad (iv)\ \left[\boldsymbol{A}_{11}, \boldsymbol{A}_{12}\right] = \left[\begin{array}{c|cc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array}\right].$$

In (i) the matrix $\boldsymbol{A}$ is divided into four submatrices

$$\boldsymbol{A}_{11} = \begin{bmatrix} 1 \end{bmatrix}, \ \boldsymbol{A}_{12} = \begin{bmatrix} 2, 3 \end{bmatrix}, \ \boldsymbol{A}_{21} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \text{ and } \boldsymbol{A}_{22} = \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix},$$

while in (ii) and (iii) $\boldsymbol{A}$ has been partitioned into columns and rows, respectively. The submatrices in a partition is often referred to as **blocks** and a partitioned matrix is sometimes called a **block matrix**.

We will make heavily use of block multiplication of matrices. In the following we assume that $\boldsymbol{A} \in \mathbb{C}^{m,p}$ and $\boldsymbol{B} \in \mathbb{C}^{p,n}$. We have the following rules and observations for block multiplication.

1. If $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{:1}, \ldots, \boldsymbol{b}_{:n} \end{bmatrix}$ is partitioned into columns then the partition of the product $\boldsymbol{AB}$ into columns is

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{Ab}_{:1}, \boldsymbol{Ab}_{:2}, \ldots, \boldsymbol{Ab}_{:n} \end{bmatrix}.$$

In particular, if I is the identity matrix of order p then

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{I} = \boldsymbol{A}\left[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p\right] = \left[\boldsymbol{A}\boldsymbol{e}_1, \boldsymbol{A}\boldsymbol{e}_2, \ldots, \boldsymbol{A}\boldsymbol{e}_p\right]$$

and we see that column j of $\boldsymbol{A}$ can be written $\boldsymbol{A}\boldsymbol{e}_j$ for $j = 1, \ldots, p$.

2. Similarly, if $\boldsymbol{A}$ is partitioned into rows then

$$\boldsymbol{A}\boldsymbol{B} = \begin{bmatrix} a_{1:}^T \\ a_{2:}^T \\ \vdots \\ a_{m:}^T \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} a_{1:}^T\boldsymbol{B} \\ a_{2:}^T\boldsymbol{B} \\ \vdots \\ a_{m:}^T\boldsymbol{B} \end{bmatrix}$$

and taking $\boldsymbol{A} = \boldsymbol{I}_p$ it follows that row i of $\boldsymbol{B}$ can be written $\boldsymbol{e}_i^T\boldsymbol{B}$ for $i = 1, \ldots, p$.

3. It is often useful to write the matrix-vector product $\boldsymbol{A}\boldsymbol{x}$ as a linear combination of the columns of $\boldsymbol{A}$

$$\boldsymbol{A}\boldsymbol{x} = x_1\boldsymbol{a}_{:1} + x_2\boldsymbol{a}_{:2} + \cdots + x_p\boldsymbol{a}_{:p}.$$

One way to see that this is correct is to partition $\boldsymbol{A}$ into columns and $\boldsymbol{x}$ into rows.

4. If $\boldsymbol{B} = \left[\boldsymbol{B}_1, \boldsymbol{B}_2\right]$, where $\boldsymbol{B}_1 \in \mathbb{C}^{p,r}$ and $\boldsymbol{B}_2 \in \mathbb{C}^{p,n-r}$ then

$$\boldsymbol{A}\left[\boldsymbol{B}_1, \boldsymbol{B}_2\right] = \left[\boldsymbol{A}\boldsymbol{B}_1, \boldsymbol{A}\boldsymbol{B}_2\right].$$

This follows from Rule 1. by an appropriate grouping of the columns of $\boldsymbol{B}$.

5. If $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}$, where $\boldsymbol{A}_1 \in \mathbb{C}^{k,p}$ and $\boldsymbol{A}_2 \in \mathbb{C}^{m-k,p}$ then

$$\begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_1\boldsymbol{B} \\ \boldsymbol{A}_2\boldsymbol{B} \end{bmatrix}.$$

This follows from Rule 2. by a grouping of the rows of $\boldsymbol{A}$.

6. If $\boldsymbol{A} = \left[\boldsymbol{A}_1, \boldsymbol{A}_2\right]$ and $\boldsymbol{B} = \left[\boldsymbol{B}_1, \boldsymbol{B}_2\right]$, where $\boldsymbol{A}_1 \in \mathbb{C}^{m,s}$, $\boldsymbol{A}_2 \in \mathbb{C}^{m,p-s}$, $\boldsymbol{B}_1 \in \mathbb{C}^{s,n}$ and $\boldsymbol{B}_2 \in \mathbb{C}^{p-s,n}$ then

$$\left[\boldsymbol{A}_1, \boldsymbol{A}_2\right] \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} = \left[\boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2\right].$$

Indeed, $(\boldsymbol{A}\boldsymbol{B})_{ij} = \sum_{j=1}^{p} a_{ik}b_{kj} = \sum_{j=1}^{s} a_{ik}b_{kj} + \sum_{j=s+1}^{p} a_{ik}b_{kj} = (\boldsymbol{A}_1\boldsymbol{B}_1)_{ij} + (\boldsymbol{A}_2\boldsymbol{B}_2)_{ij} = (\boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2)_{ij}$.

7. If $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$ then

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix},$$

provided the vertical partition in $\boldsymbol{A}$ matches the horizontal one in $\boldsymbol{B}$, i.e. the number of columns in $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{21}$ equals the number of rows in $\boldsymbol{B}_{11}$ and $\boldsymbol{B}_{12}$. To show this we use Rule 4. to obtain

$$\boldsymbol{AB} = \left[\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} \\ \boldsymbol{B}_{21} \end{bmatrix}, \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{12} \\ \boldsymbol{B}_{22} \end{bmatrix} \right].$$

We complete the proof using Rules 5. and 6.

8. Consider finally the general case. If all the matrix products in

$$\boldsymbol{C}_{ij} = \sum_{k=1}^{s} \boldsymbol{A}_{ik}\boldsymbol{B}_{kj}, \quad i = 1, \ldots, p,\ j = 1, \ldots, q$$

are well defined then

$$\begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1s} \\ \vdots & & \vdots \\ \boldsymbol{A}_{p1} & \cdots & \boldsymbol{A}_{ps} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \cdots & \boldsymbol{B}_{1q} \\ \vdots & & \vdots \\ \boldsymbol{B}_{s1} & \cdots & \boldsymbol{B}_{sq} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{11} & \cdots & \boldsymbol{C}_{1q} \\ \vdots & & \vdots \\ \boldsymbol{C}_{p1} & \cdots & \boldsymbol{C}_{pq} \end{bmatrix},$$

where

$$\boldsymbol{C}_{ij} = \sum_{k=1}^{s} \boldsymbol{A}_{ik}\boldsymbol{B}_{kj}, \quad i = 1, \ldots, p,\ j = 1, \ldots, q.$$

The requirements are that

- the number of columns in $\boldsymbol{A}$ is equal to the number of rows in $\boldsymbol{B}$.
- the position of the vertical partition lines in $\boldsymbol{A}$ has to mach the position of the horizontal partition lines in $\boldsymbol{B}$. The horizontal lines in $\boldsymbol{A}$ and the vertical lines in $\boldsymbol{B}$ can be anywhere.

B.2 The Transpose Matrix

The **transpose** of $\boldsymbol{A} \in \mathbb{C}^{m,n}$ is a matrix $\boldsymbol{B} \in \mathbb{C}^{n,m}$, where $b_{ij} = a_{ji}$ for all i, j. Thus the rows of $\boldsymbol{A}$ are the columns of $\boldsymbol{B}$ and vice versa. The transpose of $\boldsymbol{A}$ is denoted $\boldsymbol{A}^T$. Three important properties of the transpose are

1. $(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$.
2. $(\boldsymbol{AC})^T = \boldsymbol{C}^T\boldsymbol{A}^T$.
3. $(\boldsymbol{A}^T)^T = \boldsymbol{A}$.

Here $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m,n}$ and $\boldsymbol{C} \in \mathbb{C}^{n,k}$, where k, m, n are any positive integers.

Consider now the real case $\boldsymbol{A} \in \mathbb{R}^{m,n}$. A useful characterization is the following:

Theorem B.2 *Let $\langle x, y\rangle := \boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{m} x_i y_i$ be the usual inner product on $\mathbb{R}^m$ For any $\boldsymbol{A} \in \mathbb{R}^{m,n}$ we have $\langle \boldsymbol{x}, \boldsymbol{Ay}\rangle = \langle \boldsymbol{A}^T\boldsymbol{x}, \boldsymbol{y}\rangle$, all $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{y} \in \mathbb{R}^n$. If $\langle \boldsymbol{x}, \boldsymbol{Ay}\rangle = \langle \boldsymbol{Bx}, \boldsymbol{y}\rangle$ holds for some $\boldsymbol{B} \in \mathbb{R}^{n,m}$ and all $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{y} \in \mathbb{R}^n$ then $\boldsymbol{B} = \boldsymbol{A}^T$.*

Proof. For any $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$

$$\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle = \sum_{i=1}^{m} x_i \big(\sum_{j=1}^{n} a_{ij} y_j \big) = \sum_{j=1}^{n} \big(\sum_{i=1}^{m} x_i a_{ij} \big) y_j = \sum_{j=1}^{n} (\boldsymbol{A}^T \boldsymbol{x})_j y_j = \langle \boldsymbol{A}^T \boldsymbol{x}, \boldsymbol{y} \rangle.$$

If we choose $\boldsymbol{x} = \boldsymbol{e}_i$ and $\boldsymbol{y} = \boldsymbol{e}_j$ then $a_{ij} = \langle \boldsymbol{e}_i, \boldsymbol{A}\boldsymbol{e}_j \rangle = \langle \boldsymbol{B}\boldsymbol{e}_i, \boldsymbol{e}_j \rangle = b_{ji}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$ so $\boldsymbol{B} = \boldsymbol{A}^T$. □

The **Hermitian transpose** or **conjugate transpose** of $\boldsymbol{A} \in \mathbb{C}^{m,n}$ is the matrix $\boldsymbol{B} \in \mathbb{C}^{n,m}$ given by $\boldsymbol{B} = (\overline{\boldsymbol{A}})^T$ Here $\overline{z} = x - iy$ denotes the complex conjugate of $z = x + iy$, where $i = \sqrt{-1}$ is the imaginary unit and $x, y \in \mathbb{R}$. Moreover $\overline{\boldsymbol{A}}$ is obtained from $\boldsymbol{A}$ by taking the complex conjugate of all its entries. The Hermitian transpose of $\boldsymbol{A}$ is denoted $\boldsymbol{A}^H$. The Hermitian transpose enjoys the same properties as the transpose:

1. $(\boldsymbol{A} + \boldsymbol{B})^H = \boldsymbol{A}^H + \boldsymbol{B}^H$.
2. $(\boldsymbol{A}\boldsymbol{C})^H = \boldsymbol{C}^H \boldsymbol{A}^H$.
3. $(\boldsymbol{A}^H)^H = \boldsymbol{A}$.

Again $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m,n}$ and $\boldsymbol{C} \in \mathbb{C}^{n,k}$, where k, m, n are any positive integers.

We obtain the same characterization in the complex case.

Theorem B.3 *Let $\langle x, y \rangle := \boldsymbol{x}^H \boldsymbol{y} = \sum_{i=1}^{m} \overline{x}_i y_i$ be the usual inner product in $\mathbb{C}^m$ For any $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have $\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle = \langle \boldsymbol{A}^H \boldsymbol{x}, \boldsymbol{y} \rangle$, all $\boldsymbol{x} \in \mathbb{C}^m$, $\boldsymbol{y} \in \mathbb{C}^n$. If $\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle = \langle \boldsymbol{B}\boldsymbol{x}, \boldsymbol{y} \rangle$ holds for some $\boldsymbol{B} \in \mathbb{C}^{n,m}$ and all $\boldsymbol{x} \in \mathbb{C}^m$, $\boldsymbol{y} \in \mathbb{C}^n$ then $\boldsymbol{B} = \boldsymbol{A}^H$.*

Exercise B.4 *Use Theorem B.3 to show that $(\boldsymbol{A}\boldsymbol{C})^H = \boldsymbol{C}^H \boldsymbol{A}^H$ and $(\boldsymbol{A}^H)^H = \boldsymbol{A}$.*

B.3 Linear Systems

Consider a linear system

$$\begin{array}{ccccc} a_{11}x_1 + a_{12}x_2 + & \cdots & + a_{1n}x_n = & b_1 \\ a_{21}x_1 + a_{22}x_2 + & \cdots & + a_{2n}x_n = & b_2 \\ \vdots \quad\quad \vdots & & \vdots \quad\quad & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + & \cdots & + a_{mn}x_n = & b_m \end{array}$$

of m equations in n unknowns. Here for all i, j, the coefficients a_{ij}, the unknowns x_j, and the components of the right hand sides b_i, are real or complex numbers. The system can be written as a vector equation

$$x_1 \boldsymbol{a}_1 + x_2 \boldsymbol{a}_2 + \cdots + x_n \boldsymbol{a}_n = \boldsymbol{b},$$

where $\boldsymbol{a}_j = \big[a_{1j}, \ldots, \boldsymbol{a}_{mj}\big]^T \in \mathbb{C}^m$ for $j = 1, \ldots, n$ and $\boldsymbol{b} = \big[b_1, \ldots, b_m\big]^T$. It can also be written as a matrix equation

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \boldsymbol{b}.$$

The system is **homogenous** if $\boldsymbol{b} = 0$ and it is said to be **underdetermined**, **square**, or **overdetermined** if $m < n$, $m = n$, or $m > n$, respectively.

A linear system may have a unique solution, infinitely many solutions, or no solution. To discuss this we first consider a homogenous underdetermined system.

Lemma B.5 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}(\mathbb{C}^{m,n})$ with $m < n$. Then there is a nonzero $\boldsymbol{x} \in \mathbb{R}^n(\mathbb{C}^n)$ such that $\boldsymbol{Ax} = \boldsymbol{0}$.*

Proof. Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}(\mathbb{C}^{m,n})$ with $m < n$. The n columns of $\boldsymbol{A}$ span a subspace of $\mathbb{R}^m(\mathbb{C}^m)$. Since $\mathbb{R}^m(\mathbb{C}^m)$ has dimension m the dimension of this subspace is at most m. By Lemma A.14 the columns of $\boldsymbol{A}$ must be linearly dependent. It follows that there is a nonzero $\boldsymbol{x} \in \mathbb{R}^n(\mathbb{C}^n)$ such that $\boldsymbol{Ax} = \boldsymbol{0}$. □

Consider now a square linear system. The following definition is essential.

Definition B.6 *A square matrix $\boldsymbol{A}$ is said to be* **non-singular** *if the only solution of the homogenous system $\boldsymbol{Ax} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$. The matrix is* **singular** *if it is not non-singular.*

Theorem B.7 *Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$. The linear system $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} \in \mathbb{R}^n(\mathbb{C}^n)$ for any $\boldsymbol{b} \in \mathbb{R}^n(\mathbb{C}^n)$ if and only if the matrix $\boldsymbol{A}$ is non-singular.*

Proof. Suppose $\boldsymbol{A}$ is non-singular. We define $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{A}\ \boldsymbol{b}\end{bmatrix} \in \mathbb{R}^{n,n+1}(\mathbb{C}^{n,n+1})$ by adding a column to $\boldsymbol{A}$. By Lemma B.5 there is a nonzero $\boldsymbol{z} \in \mathbb{R}^{n+1}(\mathbb{C}^{n+1})$ such that $\boldsymbol{Bz} = \boldsymbol{0}$. If we write $\boldsymbol{z} = \begin{bmatrix}\tilde{\boldsymbol{z}} \\ z_{n+1}\end{bmatrix}$ where $\tilde{\boldsymbol{z}} = \begin{bmatrix}z_1, \ldots, z_n\end{bmatrix}^T \in \mathbb{R}^n(\mathbb{C}^n)$ and $z_{n+1} \in \mathbb{R}(\mathbb{C})$, then

$$\boldsymbol{Bz} = [\boldsymbol{A}\ \boldsymbol{b}]\begin{bmatrix}\tilde{\boldsymbol{z}} \\ z_{n+1}\end{bmatrix} = \boldsymbol{A}\tilde{\boldsymbol{z}} + z_{n+1}\boldsymbol{b} = \boldsymbol{0}.$$

We cannot have $z_{n+1} = 0$ for then $\boldsymbol{A}\tilde{\boldsymbol{z}} = \boldsymbol{0}$ for a nonzero $\tilde{\boldsymbol{z}}$ contradicting the non-singularity of $\boldsymbol{A}$. Define $\boldsymbol{x} := -\tilde{\boldsymbol{z}}/z_{n+1}$. Then

$$\boldsymbol{Ax} = -\boldsymbol{A}\big(\frac{\tilde{\boldsymbol{z}}}{z_{n+1}}\big) = -\frac{1}{z_{n+1}}\boldsymbol{A}\tilde{\boldsymbol{z}} = -\frac{1}{z_{n+1}}\big(-z_{n+1}\boldsymbol{b}\big) = \boldsymbol{b}$$

so $\boldsymbol{x}$ is a solution.

Suppose $\boldsymbol{Ax} = \boldsymbol{b}$ and $\boldsymbol{Ay} = \boldsymbol{b}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n(\mathbb{C}^n)$. Then $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{0}$ and since $\boldsymbol{A}$ is non-singular we conclude that $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0}$ or $\boldsymbol{x} = \boldsymbol{y}$. Thus the solution is unique.

Conversely, if $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution for any $\boldsymbol{b} \in \mathbb{R}^n(\mathbb{C}^n)$ then $\boldsymbol{Ax} = \boldsymbol{0}$ has a unique solution which must be $\boldsymbol{x} = \boldsymbol{0}$. Thus $\boldsymbol{A}$ is non-singular. □

B.4 The Inverse matrix

Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ is a square matrix. A matrix $\boldsymbol{B} \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ is called a **right inverse** of $\boldsymbol{A}$ if $\boldsymbol{AB} = \boldsymbol{I}$. A matrix $\boldsymbol{C} \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ is said to be a **left inverse** of $\boldsymbol{A}$ if $\boldsymbol{CA} = \boldsymbol{I}$. We say that $\boldsymbol{A}$ is **invertible** if it has both a left- and a right inverse. If $\boldsymbol{A}$ has a right inverse $\boldsymbol{B}$ and a left inverse $\boldsymbol{C}$ then

$$\boldsymbol{C} = \boldsymbol{CI} = \boldsymbol{C}(\boldsymbol{AB}) = (\boldsymbol{CA})\boldsymbol{B} = \boldsymbol{IB} = \boldsymbol{B}$$

and this common inverse is called the **inverse** of $\boldsymbol{A}$ and denoted $\boldsymbol{A}^{-1}$. Thus the inverse satisfies $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$.

We want to characterize the class of invertible matrices and start with a lemma.

Lemma B.8 *If $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ with $\boldsymbol{AB} = \boldsymbol{C}$ then $\boldsymbol{C}$ is non-singular if and only if both $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular.*

Proof. Suppose both $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular and let $\boldsymbol{Cx} = \boldsymbol{0}$. Then $\boldsymbol{ABx} = \boldsymbol{0}$ and since $\boldsymbol{A}$ is non-singular we see that $\boldsymbol{Bx} = \boldsymbol{0}$. Since $\boldsymbol{B}$ is non-singular we have $\boldsymbol{x} = \boldsymbol{0}$. We conclude that $\boldsymbol{C}$ is non-singular.

For the converse suppose first that $\boldsymbol{B}$ is singular and let $\boldsymbol{x} \in \mathbb{R}^n(\mathbb{C}^n)$ be a nonzero vector so that $\boldsymbol{Bx} = \boldsymbol{0}$. But then $\boldsymbol{Cx} = (\boldsymbol{AB})\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{Bx}) = \boldsymbol{A0} = \boldsymbol{0}$ so $\boldsymbol{C}$ is singular. Finally suppose $\boldsymbol{B}$ is non-singular, but $\boldsymbol{A}$ is singular. Let $\tilde{\boldsymbol{x}}$ be a nonzero vector such that $\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{0}$. By Theorem B.7 there is a vector $\boldsymbol{x}$ such that $\boldsymbol{Bx} = \tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ is nonzero since $\tilde{\boldsymbol{x}}$ is nonzero. But then $\boldsymbol{Cx} = (\boldsymbol{AB})\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{Bx}) = \boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{0}$ for a nonzero vector $\boldsymbol{x}$ and $\boldsymbol{C}$ is singular. □

Theorem B.9 *A square matrix is invertible if and only if it is non-singular.*

Proof. Suppose first $\boldsymbol{A}$ is a non-singular matrix. By Theorem B.7 each of the linear systems $\boldsymbol{Ab}_i = \boldsymbol{e}_i$ has a unique solution $\boldsymbol{b}_i$ for $i = 1, \ldots, n$. Let $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\end{bmatrix}$. Then $\boldsymbol{AB} = \begin{bmatrix}\boldsymbol{Ab}_1, \ldots, \boldsymbol{Ab}_n\end{bmatrix} = \begin{bmatrix}\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\end{bmatrix} = \boldsymbol{I}$ so that $\boldsymbol{A}$ has a right inverse $\boldsymbol{B}$. By Lemma B.8 $\boldsymbol{B}$ is non-singular since $\boldsymbol{I}$ is non-singular and $\boldsymbol{AB} = \boldsymbol{I}$. Since $\boldsymbol{B}$ is non-singular we can use what we have shown for $\boldsymbol{A}$ to conclude that $\boldsymbol{B}$ has a right inverse $\boldsymbol{C}$, i.e. $\boldsymbol{BC} = \boldsymbol{I}$. But then $\boldsymbol{AB} = \boldsymbol{BC} = \boldsymbol{I}$ so $\boldsymbol{B}$ has both a right inverse and a left inverse which must be equal so $\boldsymbol{A} = \boldsymbol{C}$. Since $\boldsymbol{BC} = \boldsymbol{I}$ we have $\boldsymbol{BA} = \boldsymbol{I}$ so $\boldsymbol{B}$ is also a left inverse of $\boldsymbol{A}$ and $\boldsymbol{A}$ is invertible.

Conversely, if $\boldsymbol{A}$ is invertible then it has a right inverse $\boldsymbol{B}$ and since $\boldsymbol{AB} = \boldsymbol{I}$ and $\boldsymbol{I}$ is non-singular we again use Lemma B.8 to conclude that $\boldsymbol{A}$ is non-singular. □

The theorem shows that we can use the terms "non-singular" and "invertible" interchangeably. If $\boldsymbol{B}$ is a right inverse or a left inverse of $\boldsymbol{A}$ then it follows from Lemma B.8 that $\boldsymbol{A}$ is non-singular. Thus to verify that some matrix $\boldsymbol{B}$ is an inverse of another matrix $\boldsymbol{A}$ it is enough to show that $\boldsymbol{B}$ is either a left inverse of a right inverse of $\boldsymbol{A}$. This calculation also proves that $\boldsymbol{A}$ is non-singular. We use this observation to give simple proofs of the following results.

Corollary B.10 *Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ are non-singular and c is a nonzero constant.*

1. *$\boldsymbol{A}^{-1}$ is non-singular and $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.*
2. *$\boldsymbol{C} = \boldsymbol{AB}$ is non-singular and $\boldsymbol{C}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.*
3. *$\boldsymbol{A}^T$ is non-singular and $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T =: \boldsymbol{A}^{-T}$.*
4. *$c\boldsymbol{A}$ is non-singular and $(c\boldsymbol{A})^{-1} = \frac{1}{c}\boldsymbol{A}^{-1}$.*

Proof.

1. Since $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ the matrix $\boldsymbol{A}$ is a right inverse of $\boldsymbol{A}^{-1}$. Thus $\boldsymbol{A}^{-1}$ is non-singular and $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
2. We note that $(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1})(\boldsymbol{AB}) = \boldsymbol{B}^{-1}(\boldsymbol{A}^{-1}\boldsymbol{A})\boldsymbol{B} = \boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}$. Thus $\boldsymbol{AB}$ is invertible with the indicated inverse since it has a left inverse.
3. Now $\boldsymbol{I} = \boldsymbol{I}^T = (\boldsymbol{A}^{-1}\boldsymbol{A})^T = \boldsymbol{A}^T(\boldsymbol{A}^{-1})^T$ showing that $(\boldsymbol{A}^{-1})^T$ is a right inverse of $\boldsymbol{A}^T$.
4. The matrix $\frac{1}{c}\boldsymbol{A}^{-1}$ is a one sided inverse of $c\boldsymbol{A}$.

□

Exercise B.11 *Show that*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \alpha \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \quad \alpha = \frac{1}{ad - bc},$$

for any a, b, c, d such that $ad - bc \neq 0$.

Exercise B.12 *Find the inverse of*

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Exercise B.13 *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$, and $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n,m}$ for some $n, m \in \mathbb{N}$. If $(\boldsymbol{I} + \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}$ exists then*

$$(\boldsymbol{A} + \boldsymbol{BC}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I} + \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}^T\boldsymbol{A}^{-1}.$$

B.5 Rank, Nullity, and the Fundamental Subspaces

Recall that the column space (or span) and the null space (kernel) of a matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ are defined by

$$\begin{aligned} \operatorname{span} \boldsymbol{A} &:= \{\boldsymbol{y} \in \mathbb{C}^m : \boldsymbol{y} = \boldsymbol{Ax},\ \boldsymbol{x} \in \mathbb{R}^n\} \\ \ker \boldsymbol{A} &:= \{\boldsymbol{x} \in \mathbb{C}^n : \boldsymbol{Ax} = \boldsymbol{0}\}. \end{aligned}$$

These sets are subspaces of $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively. The four subspaces span $\boldsymbol{A}$, $\ker \boldsymbol{A}$, span $\boldsymbol{A}^*$ and $\ker \boldsymbol{A}^*$ are known as the four **fundamental subspaces** of a matrix. The dimension of the column space of $\boldsymbol{A}$ is called the **rank** of $\boldsymbol{A}$ and denoted $\operatorname{rank} \boldsymbol{A}$. The dimension $\dim \ker \boldsymbol{A}$ of the null space is called the **nullity** of $\boldsymbol{A}$ and denoted $\operatorname{null} \boldsymbol{A}$.

Recall that the orthogonal complement $\mathcal{S}^\perp$ of a subspace $\mathcal{S}$ of $\mathbb{C}^n$ is $\{\boldsymbol{t} \in \mathbb{C}^n : \langle \boldsymbol{s}, \boldsymbol{t} \rangle = 0 \text{ for all } \boldsymbol{s} \in \mathcal{S}\}$. For $\mathcal{S} = \operatorname{span} \boldsymbol{A}$ we have

Theorem B.14 *The orthogonal complement of the column space of a matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ is the null space of $\boldsymbol{A}^*$. We have the orthogonal decomposition*

$$\mathbb{C}^m = \operatorname{span} \boldsymbol{A} \oplus \ker \boldsymbol{A}^*. \tag{B.1}$$

Proof. We first show that

$$\operatorname{span}(\boldsymbol{A})^\perp = \ker(\boldsymbol{A}^*) := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{A}^* \boldsymbol{y} = \boldsymbol{0}\}.$$

Suppose $\boldsymbol{c} \in \operatorname{span}(\boldsymbol{A})$. Then $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{R}^n$. If $\boldsymbol{y} \in \ker(\boldsymbol{A}^*)$ then $\langle \boldsymbol{y}, \boldsymbol{c} \rangle = \langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{x} \rangle = \langle \boldsymbol{A}^* \boldsymbol{y}, \boldsymbol{x} \rangle = 0$. Thus $\ker(\boldsymbol{A}^*) \subset \operatorname{span}(\boldsymbol{A})^\perp$. To show that $\operatorname{span}(\boldsymbol{A})^\perp \subset \ker(\boldsymbol{A}^*)$ we pick any $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{A})^\perp$. Then $\langle \boldsymbol{A}^* \boldsymbol{y}, \boldsymbol{x} \rangle = \langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{x} \rangle = 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$ which means that $\boldsymbol{y} \in \ker(\boldsymbol{A}^*)$. The orthogonal decomposition (B.1) now follows from Theorem A.55. □

The following formula for the rank of a product of two matrices will also be useful.

Lemma B.15 *If $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $\boldsymbol{B} \in \mathbb{C}^{n,p}$ for some $m, n, p \in \mathbb{N}$ then*

$$\operatorname{rank}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{rank} \boldsymbol{B} - \dim(\ker \boldsymbol{A} \cap \operatorname{span} \boldsymbol{B}).$$

Proof. Pick a basis $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ for $\ker \boldsymbol{A} \cap \operatorname{span} \boldsymbol{B}$ and extend it to a basis $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k, \boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_l\}$ for $\operatorname{span} \boldsymbol{B}$. The result will follow if we can show that $Y := \{\boldsymbol{A}\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{A}\boldsymbol{x}_l\}$ is a basis for $\operatorname{span}(\boldsymbol{A}\boldsymbol{B})$.

(i) Y is linearly independent. For if $\sum c_j \boldsymbol{A}\boldsymbol{x}_j := \sum_{j=k+1}^{l} c_j \boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{0}$ then $\boldsymbol{A}\big(\sum c_j \boldsymbol{x}_j\big) = \boldsymbol{0}$, and hence $\sum c_j \boldsymbol{x}_j \in \ker \boldsymbol{A} \cap \operatorname{span} \boldsymbol{B}$. But then $\sum_{j=k+1}^{l} c_j \boldsymbol{x}_j = \sum_{j=1}^{k} c_j \boldsymbol{s}_j$ for some $c_1, \ldots, c_k$, and by linear independence we have $c_1 = \cdots = c_l = 0$.

(ii) $\operatorname{span}\ Y \subset \operatorname{span}(\boldsymbol{A}\boldsymbol{B})$. Suppose $\boldsymbol{y} = \sum c_j \boldsymbol{A}\boldsymbol{x}_j \in \operatorname{span}\ Y$. Since $\boldsymbol{x}_j \in \operatorname{span}(\boldsymbol{B})$ we have $\boldsymbol{x}_j = \boldsymbol{B}\boldsymbol{z}_j$, for some $\boldsymbol{z}_j$, $j = k+1, \ldots, l$. But then $\boldsymbol{y} = \sum_j c_j \boldsymbol{A}\boldsymbol{B}\boldsymbol{z}_j \in \operatorname{span}(\boldsymbol{A}\boldsymbol{B})$.

(iii) $\operatorname{span}(\boldsymbol{A}\boldsymbol{B}) \subset \operatorname{span}\ Y$. If $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{A}\boldsymbol{B})$ then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{x} \in \operatorname{span} \boldsymbol{B}$. Since $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k, \boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_l\}$ is a basis for $\operatorname{span} \boldsymbol{B}$ we have $\boldsymbol{x} = \sum_{j=1}^{k} c_j \boldsymbol{s}_j + \sum_{j=k+1}^{l} c_j \boldsymbol{x}_j$ for some $c_1, \ldots, c_l$ and $\boldsymbol{s}_j \in \ker(\boldsymbol{A})$. But then

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^{k} c_j \boldsymbol{A}\boldsymbol{s}_j + \sum_{j=k+1}^{l} c_j \boldsymbol{A}\boldsymbol{x}_j = \sum_{j=k+1}^{l} c_j \boldsymbol{A}\boldsymbol{x}_j \in \operatorname{span}(Y).$$

□

Consider now the four fundamental subspaces.

Theorem B.16 *For any matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have*

1. $\operatorname{rank} \boldsymbol{A} + \operatorname{null} \boldsymbol{A} = n$,
2. $\operatorname{rank} \boldsymbol{A} + \operatorname{null} \boldsymbol{A}^* = m$,
3. $\operatorname{rank} \boldsymbol{A} = \operatorname{rank} \boldsymbol{A}^*$.

Proof.
1. Taking $\boldsymbol{B}$ to be the identity matrix in Lemma B.15 we obtain $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank} \boldsymbol{I} - \dim(\ker \boldsymbol{A} \cap \operatorname{span} \boldsymbol{I}) = n - \dim(\ker \boldsymbol{A} \cap \mathbb{C}^n) = n - \dim \ker \boldsymbol{A}$.
2. This follows from Theorems A.55 and B.14.
3. If we apply **2**. to $\boldsymbol{A}^*$ we obtain $\dim \operatorname{span} \boldsymbol{A}^* + \dim \ker \boldsymbol{A} = n$. But then $\operatorname{rank} \boldsymbol{A} = \dim \operatorname{span} \boldsymbol{A} \overset{1.}{=} n - \dim \ker \boldsymbol{A} = n - (n - \dim \operatorname{span} \boldsymbol{A}^*) = \dim \operatorname{span} \boldsymbol{A}^* = \operatorname{rank}(\boldsymbol{A}^*)$.
□

To derive some further results about rank and nullity we start with a definition:

Definition B.17 (Equivalent matrices) *Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m,n}$. We say that $\boldsymbol{A}$ is equivalent to $\boldsymbol{B}$, denoted $\boldsymbol{A} \sim \boldsymbol{B}$, if $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{Y}$ for some non-singular matrices $\boldsymbol{X} \in \mathbb{C}^{m,m}$ and $\boldsymbol{Y} \in \mathbb{C}^{n,n}$.*

Exercise B.18 *Show that $\sim$ is an equivalence relation, i. e.,*

(i) $\boldsymbol{A} \sim \boldsymbol{A}$,

(ii) *if $\boldsymbol{A} \sim \boldsymbol{B}$ then $\boldsymbol{B} \sim \boldsymbol{A}$,*

(iii) *if $\boldsymbol{A} \sim \boldsymbol{B}$ and $\boldsymbol{B} \sim C$ then $\boldsymbol{A} \sim C$.*

For any subspace $\mathcal{S}$ of $\mathbb{C}^n$ and $\boldsymbol{B} \in \mathbb{C}^{m,n}$ we define $\boldsymbol{B}\mathcal{S} := \{\boldsymbol{B}\boldsymbol{s} \colon \boldsymbol{s} \in \mathcal{S}\}$.

Exercise B.19 *Show that $\boldsymbol{B}\mathcal{S}$ is a subspace of $\mathbb{C}^m$.*

Exercise B.20 *Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and that $X \in \mathbb{C}^{m,m}$ and $\boldsymbol{Y} \in \mathbb{C}^{n,n}$ are non-singular. Show that*

1. $\operatorname{span}(\boldsymbol{A}) = \operatorname{span}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{X}^{-1} \operatorname{span}(X\boldsymbol{A})$,
2. $\ker \boldsymbol{A} = \boldsymbol{Y} \ker(\boldsymbol{A}\boldsymbol{Y}) = \ker(\boldsymbol{X}\boldsymbol{A})$,
3. $\operatorname{rank}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{Y}) = \operatorname{rank}(\boldsymbol{A})$,
4. $\operatorname{null}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{Y}) = \operatorname{null} \boldsymbol{A}$.

For the rank of a general product we have

Theorem B.21 *Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and $\boldsymbol{B} \in \mathbb{C}^{n,p}$ for some $m, n, p \in \mathbb{N}$. Then*

$$\operatorname{rank}(\boldsymbol{AB}) \leq \min\{\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(\boldsymbol{B})\}.$$

Proof. Since $\operatorname{span}(\boldsymbol{AB}) \subset \operatorname{span}(\boldsymbol{A})$ we have $\operatorname{rank}(\boldsymbol{AB}) \leq \operatorname{rank}(\boldsymbol{A})$. Now $\operatorname{span}(\boldsymbol{B}^*\boldsymbol{A}^*) \subset \operatorname{span}(\boldsymbol{B}^*)$. Therefore $\operatorname{rank}(\boldsymbol{AB}) = \operatorname{rank}((\boldsymbol{AB})^*) = \operatorname{rank}(\boldsymbol{B}^*\boldsymbol{A}^*) \leq \operatorname{rank}(\boldsymbol{B}^*) = \operatorname{rank}(\boldsymbol{B})$. ∎

We end this section with the following useful result.

Theorem B.22 *If the matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ has rank r then there is at least one non-singular $r \times r$ submatrix in $\boldsymbol{A}$. Moreover there are no non-singular submatrices of larger order.*

Proof. We use Theorem A.16 twice. There is a subset $\{\boldsymbol{a}_{\cdot j_1}, \ldots, \boldsymbol{a}_{\cdot j_r}\}$ of the columns of $\boldsymbol{A}$ which forms a basis for $\operatorname{span}(\boldsymbol{A})$. Consider the matrix $\boldsymbol{B}^* \in \mathbb{C}^{m,r}$, where $\boldsymbol{B} = [\boldsymbol{a}_{\cdot j_1}, \ldots, \boldsymbol{a}_{\cdot j_r}]$. Since $r = \operatorname{rank}(\boldsymbol{B}) = \operatorname{rank}(\boldsymbol{B}^*)$ there is a subset $\{i_1, \ldots, i_r\}$ of $\{1, \ldots, m\}$ such that columns $i_1, \ldots, i_r$ of $\boldsymbol{B}^*$ form a basis for $\operatorname{span}(\boldsymbol{B}^*)$. But then rows $i_1, \ldots, i_r$ of $\boldsymbol{B}$ are linearly independent, defining a non-singular $r \times r$ submatrix in $\boldsymbol{A}$. Suppose M is a non-singular submatrix in $\boldsymbol{A}$ of order k. The columns in $\boldsymbol{A}$ corresponding to the columns in M are linearly independent and hence $k \leq r$. ∎

B.6 Linear Transformations and Matrices

Let $(\mathcal{X}, \mathbb{F})$ and $(\mathcal{Y}, \mathbb{F})$ be vector spaces over the same field $\mathbb{F}$. A mapping $\boldsymbol{T} : \mathcal{X} \to \mathcal{Y}$ is called **linear** if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and all $a \in \mathbb{F}$ we have

1. $\boldsymbol{T}(\boldsymbol{x} + \boldsymbol{y}) = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{T}\boldsymbol{y}$, (additivity)
2. $\boldsymbol{T}(a\boldsymbol{x}) = a\boldsymbol{T}\boldsymbol{x}$. (homogeneity)

If $\mathcal{Y}$ is the vector space of all functions $f : \mathbb{R} \to \mathbb{R}$ and $\mathcal{X}$ is the space of all differentiable functions $f : \mathbb{R} \to \mathbb{R}$, then the mapping $\boldsymbol{T} : \mathcal{X} \to \mathcal{Y}$ given by $\boldsymbol{T}f := df/dx$ is a linear transformation from $\mathcal{X}$ to $\mathcal{Y}$. The mapping $\boldsymbol{T}$ given by $(\boldsymbol{T}f)(x) := \int_0^x f(t)dt$ is a linear transformation from the space $\mathcal{X}$ of all continuous functions $f : \mathbb{R} \to \mathbb{R}$ into $\mathcal{X}$.

Linear transformations are not the main emphasis of this text and we will only consider briefly the special case where $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$. The same results hold for the complex case $\mathcal{X} = \mathbb{C}^n$ and $\mathcal{Y} = \mathbb{C}^m$. Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$. The mapping $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$ given by $\boldsymbol{T}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}$ is clearly additive and homogenous. Thus it is a linear mapping. It turns out that all linear mappings $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$ are of this form.

Theorem B.23 *Every linear map from $\mathbb{R}^n \to \mathbb{R}^m$ can be written in the form $\boldsymbol{T} = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{A} \in \mathbb{R}^{m,n}$.*

Proof. Suppose $\boldsymbol{x} \in \mathbb{R}^n$. Then $\boldsymbol{x} = \sum_{j=1}^n x_j \boldsymbol{e}_j$ and by linearity

$$\boldsymbol{T}\boldsymbol{x} = \boldsymbol{T}\big(\sum_{j=1}^n x_j \boldsymbol{e}_j\big) = \sum_{j=1}^n x_j \boldsymbol{T}\boldsymbol{e}_j = \sum_{j=1}^n x_j \boldsymbol{a}_j = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] = [\boldsymbol{T}\boldsymbol{e}_1, \ldots, \boldsymbol{T}\boldsymbol{e}_n] \in \mathbb{R}^{m,n}$. □

Let

$$\begin{aligned} \operatorname{span} \boldsymbol{T} &:= \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y} = \boldsymbol{T}\boldsymbol{x} \text{ for some } \boldsymbol{x} \in \mathbb{R}^n\}, \\ \ker \boldsymbol{T} &:= \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{T}\boldsymbol{x} = \boldsymbol{0}\}, \end{aligned} \tag{B.2}$$

be the **span** and **kernel** of the linear transformation $\boldsymbol{T}$. The sets $\operatorname{span} \boldsymbol{T}$ and $\ker \boldsymbol{T}$ are subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

Theorem B.24 *Suppose $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation. Then For any matrix $\boldsymbol{A} \in \mathbb{C}^{m,n}$ we have*

$$\dim \operatorname{span} \boldsymbol{T} + \dim \ker \boldsymbol{T} = n.$$

Proof. This follows from Theorem B.16 since $\boldsymbol{T}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}$ for some matrix $\boldsymbol{A}$. □

Much more can be said about linear transformations and matrices. We refer to any book on linear algebra.

B.7 Orthonormal and Unitary Matrices

Definition B.25 *A matrix $\boldsymbol{Q} \in \mathbb{R}^{n,n}$ is said to be* **orthonormal** *if $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$.*

Since the columns of an orthonormal matrix are orthonormal, we have chosen the term "orthonormal matrix" although "orthogonal matrix" is more common in the classical literature.

Theorem B.26 *Suppose $\boldsymbol{Q} \in \mathbb{R}^{n,n}$. The following is equivalent:*

1. *$\boldsymbol{Q}$ is orthonormal,*
2. *the columns of $\boldsymbol{Q}$ form an orthonormal basis for $\mathbb{R}^n$.*
3. *$\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$*
4. *$\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$*
5. *the columns of $\boldsymbol{Q}^T$ (rows of $\boldsymbol{Q}$) form an orthonormal basis for $\mathbb{R}^n$,*
6. *$\langle \boldsymbol{Q}\boldsymbol{x}, \boldsymbol{Q}\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T\boldsymbol{y}$ is the usual inner product on $\mathbb{R}^n$.*

We also have

(i) *The product $\boldsymbol{Q}_1\boldsymbol{Q}_2$ of two orthonormal matrices is orthonormal.*

(ii) *If $\boldsymbol{Q}$ is orthonormal then $\|\boldsymbol{Q}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Proof. Let $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ be the columns of $\boldsymbol{Q}$.

1$\Leftrightarrow$ 2 This follows since $(\boldsymbol{Q}^T\boldsymbol{Q})_{ij} = \langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle$ for all i, j.

1$\Leftrightarrow$ 3 Since $\boldsymbol{Q}^T$ is a left inverse of $\boldsymbol{Q}$ it follows from the discussion after Theorem B.9 that $\boldsymbol{Q}$ is invertible and $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$.

3$\Leftrightarrow$ 4 Since $\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}$ the definition of the inverse matrix implies that $\boldsymbol{Q}^T$ is a right inverse of $\boldsymbol{Q}$ so that $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$.

4$\Leftrightarrow$ 5 This follows since $\boldsymbol{Q}^T$ is orthonormal and $(\boldsymbol{Q}^T)^T = \boldsymbol{Q}$

1$\Leftrightarrow$ 6 If $\boldsymbol{Q}$ is orthonormal then by Theorem B.2 we have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{Q}\boldsymbol{x}, \boldsymbol{Q}\boldsymbol{y} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Conversely, taking $\boldsymbol{x} = \boldsymbol{e}_i$ and $\boldsymbol{y} = \boldsymbol{e}_j$ we find $(\boldsymbol{Q}^T\boldsymbol{Q})_{ij} = \langle \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{e}_i, \boldsymbol{e}_j \rangle = \langle \boldsymbol{Q}\boldsymbol{e}_i, \boldsymbol{Q}\boldsymbol{e}_j \rangle = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = \delta_{ij}$ for all $i, j = 1, \ldots, n$.

Suppose $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are orthonormal. Then $(\boldsymbol{Q}_1\boldsymbol{Q}_2)^T\boldsymbol{Q}_1\boldsymbol{Q}_2 = \boldsymbol{Q}_2^T\boldsymbol{Q}_1^T\boldsymbol{Q}_1\boldsymbol{Q}_2 = \boldsymbol{I}$ so the product $\boldsymbol{Q}_1\boldsymbol{Q}_2$ is orthonormal. Using 6. with $\boldsymbol{y} = \boldsymbol{x}$ we obtain **(ii)**. □

Consider now the complex case.

Definition B.27 *A matrix $U \in \mathbb{C}^{n,n}$ is said to be* **unitary** *if $U^*U = \boldsymbol{I}$.*

Note that a real unitary matrix is orthonormal.

Theorem B.28 *Suppose $U \in \mathbb{C}^{n,n}$. The following is equivalent:*

1. *U is unitary,*
2. *the columns of U form an orthonormal basis for $\mathbb{C}^n$.*
3. *$U^{-1} = U^*$*
4. *$UU^* = \boldsymbol{I}$*
5. *the columns of U^* (rows of U) form an orthonormal basis for $\mathbb{C}^n$,*
6. *$\langle U\boldsymbol{x}, U\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$, where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^*\boldsymbol{y}$ is the usual inner product on $\mathbb{C}^n$.*
7. *$\|U\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \mathbb{C}^n$.*

The product U_1U_2 of two unitary matrices is unitary.

Proof. That 1-6 are equivalent is similar to the proof of the real case. Clearly 6 implies 7. That 7 implies 6 follows from the fact that we can write a complex inner product as a sum of norms. (Cf. (A.13).) □

Exercise B.29 *Show in Theorem B.28 that 7. implies 6.*

Appendix C

Determinants

The first systematic treatment of determinants was given by Cauchy in 1812. He adopted the word "determinant" which was introduced by Gauss in 1801. The first use of determinants was made by Leibniz in 1693 in a letter to De L'Hôspital. By the beginning of the 20th century the theory of determinants filled four volumes of almost 2000 pages (Muir, 1906–1923. Historic references can be found in this work). The main use of determinants in this text will be to study the characteristic polynomial of a matrix.

In this section we give the elementary properties of determinants that we need.

C.1 Permutations

For $n \in \mathbb{N}$, let $N_n = \{1, 2, \ldots, n\}$. A *permutation* is a function $\sigma : N_n \to N_n$ which is one-to-one and onto. That is, $\{\sigma(1), \sigma(2), \ldots, \sigma(n)\}$ is a rearrangement of $\{1, 2, \ldots, n\}$. If $n = 2$, there are two permutations $\{1, 2\}$ and $\{2, 1\}$, while for $n = 3$ we have six permutations $\{1, 2, 3\}$, $\{1, 3, 2\}$, $\{2, 1, 3\}$, $\{2, 3, 1\}$, $\{3, 1, 2\}$ and $\{3, 2, 1\}$. We denote the set of all permutations on N_n by S_n. There are $n!$ elements in S_n.

If σ, τ are two permutations in S_n, we can define their product $\sigma\tau$ as

$$\sigma\tau = \{\sigma(\tau(1)), \sigma(\tau(2)), \ldots, \sigma(\tau(n))\}.$$

For example if $\sigma = \{1, 3, 2\}$ and $\tau = \{3, 2, 1\}$, then $\sigma\tau = \{\sigma(3), \sigma(2), \sigma(1)\} = \{2, 3, 1\}$, while $\tau\sigma = \{\tau(1), \tau(3), \tau(2)\} = \{3, 1, 2\}$. Thus in general $\sigma\tau \neq \tau\sigma$. It is easily shown that the product of two permutations σ, τ is a permutation, i.e. $\sigma\tau : N_n \to N_n$ is one-to-one and onto.

The permutation $\epsilon = \{1, 2, \ldots, n\}$ is called the *identity permutation* in S_n. We have $\epsilon\sigma = \sigma\epsilon = \sigma$ for all $\sigma \in S_n$.

Since each $\sigma \in S_n$ is one-to-one and onto, it has a unique inverse σ^{-1}. To define $\sigma^{-1}(j)$ for $j \in N_n$, we find the unique i such that $\sigma(i) = j$. Then $\sigma^{-1}(j) = i$. We have $\sigma^{-1}\sigma = \sigma\sigma^{-1} = \epsilon$. As an example, if $\sigma = \{2, 3, 1\}$ then $\sigma^{-1} = \{3, 1, 2\}$, and $\sigma^{-1}\sigma = \sigma\sigma^{-1} = \{1, 2, 3\} = \epsilon$.

With each $\sigma \in S_n$ we can associate a + or − sign. We define

$$\text{sign}(\sigma) = \frac{g(\sigma)}{|g(\sigma)|},$$

where

$$g(\sigma) = \prod_{i=2}^{n} (\sigma(i) - \sigma(1))(\sigma(i) - \sigma(2)) \cdots (\sigma(i) - \sigma(i-1)).$$

For example if $\epsilon = \{1, 2, 3, 4\}$ and $\sigma = \{4, 3, 1, 2\}$, then

$$\begin{aligned} g(\epsilon) &= (2-1)(3-1)(3-2)(4-1)(4-2)(4-3) = 1! \cdot 2! \cdot 3! > 0, \\ g(\sigma) &= (3-4)(1-4)(1-3)(2-4)(2-3)(2-1) \\ &= (-1)(-3)(-2)(-2)(-1) \cdot 1 = -1! \cdot 2! \cdot 3! < 0. \end{aligned}$$

Thus $\text{sign}(\epsilon) = +1$ and $\text{sign}(\sigma) = -1$.

$g(\sigma)$ contains one positive factor $(2-1)$ and five negative ones. The negative factors are called *inversions*. The number of inversions equals the number of times a bigger integer precedes a smaller one in σ. That is, in $\{4, 3, 1, 2\}$ 4 precedes 3, 1 and 2 (three inversions corresponding to the negative factors $(3-4)$, $(1-4)$ and $(2-4)$ in $g(\sigma)$), and 3 precedes 1 and 2 ($(1-3)$ and $(2-3)$ in $g(\sigma)$). This makes it possible to compute $\text{sign}(\sigma)$ without actually writing down $g(\sigma)$.

In general, the sign function has the following properties

1. $\text{sign}(\epsilon) = 1$.
2. $\text{sign}(\sigma\tau) = \text{sign}(\sigma)\text{sign}(\tau)$ for $\sigma, \tau \in S_n$.
3. $\text{sign}(\sigma^{-1}) = \text{sign}(\sigma)$ for $\sigma \in S_n$.

Since all factors in $g(\epsilon)$ are positive, we have $g(\epsilon) = |g(\epsilon)|$ and $\text{sign}(\epsilon) = 1$. This proves 1. To prove 2 we first note that for any S_n

$$\text{sign}(\sigma) = \frac{g(\sigma)}{g(\epsilon)}.$$

Since $g(\sigma)$ and $g(\epsilon)$ contain the same factors apart from signs and $g(\epsilon) > 0$, we have $|g(\sigma)| = g(\epsilon)$. Now

$$\text{sign}(\sigma\tau) = \frac{g(\sigma\tau)}{g(\epsilon)} = \frac{g(\sigma\tau)}{g(\tau)} \frac{g(\tau)}{g(\epsilon)} = \frac{g(\sigma\tau)}{g(\tau)} \text{sign}(\tau).$$

We have to show that $g(\sigma\tau)/g(\tau) = g(\sigma)/g(\epsilon)$. We write $g(\sigma)/g(\epsilon)$ in the form

$$\frac{g(\sigma)}{g(\epsilon)} = \prod_{i=2}^{n} \prod_{j=1}^{i-1} r_\sigma(i, j), \qquad r_\sigma(i, j) = \frac{\sigma(i) - \sigma(j)}{i - j}.$$

Now

$$\frac{g(\sigma\tau)}{g(\tau)} = \frac{\prod_{i=2}^{n} (\sigma(\tau(i)) - \sigma(\tau(1))) \cdots (\sigma(\tau(i)) - \sigma(\tau(i-1)))}{\prod_{i=2}^{n} (\tau(i) - \tau(1)) \cdots (\tau(i) - \tau(i-1))} = \prod_{i=2}^{n} \prod_{j=1}^{i-1} r_\sigma(\tau(i), \tau(j)).$$

τ is a permutation so $g(\sigma)/g(\epsilon)$ and $g(\sigma\tau)/g(\tau)$ contain the same factors. Moreover, the sign of the factors are the same since $r(i,j) = r(j,i)$ for all $i \neq j$. Thus $g(\sigma)/g(\epsilon) = g(\sigma\tau)/g(\tau)$, and 2 is proved. Finally, 3 follows from 1 and 2; $1 = \text{sign}(\epsilon) = \text{sign}(\sigma\sigma^{-1}) = \text{sign}(\sigma)\text{sign}(\sigma^{-1})$ so that σ and σ^{-1} have the same sign.

Exercise C.1 *Show that $\rho(\sigma\tau) = (\rho\sigma)\tau$ for $\rho, \sigma, \tau \in S_n$, i.e. multiplication of permutations is associative. (In fact, we have*

1. *Multiplication is associative.*
2. *There exists an identity permutation ϵ.*
3. *Every permutation has an inverse.*

Thus the set S_n of permutations is a group with respect to multiplication. S_n is called the symmetric group of degree *n).*

C.2 Basic Properties of Determinants

For any $A \in \mathbb{C}^{n,n}$ the determinant of A is defined the number

$$\det(A) = \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{\sigma(1),1} a_{\sigma(2),2} \cdots a_{\sigma(n),n}. \tag{C.1}$$

This sum ranges of all $n!$ permutations of $\{1, 2, \ldots, n\}$. We also denote the determinant by (Cayley, 1841)

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

From the definition we have

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}.$$

The first term on the right corresponds to the identity permutation ϵ given by $\epsilon(i) = i$, $i = 1, 2$. The second term comes from the permutation $\sigma = \{2, 1\}$. For $n = 3$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13}.$$

The following is a list of properties of determinants.

1. **Triangular matrix** The determinant of a triangular matrix is the product of the diagonal elements. $\det(A) = a_{11}a_{22}\cdots a_{nn}$. In particular $\det(I) = 1$.

2. **Transpose** $\det(A^T) = \det(A)$.
3. **Homogeneity** For any $\beta_i \in \mathbb{C}$, $i = 1, 2, \ldots, n$, we have

$$\det\big(\,[\beta_1 \boldsymbol{a}_1, \beta_2 \boldsymbol{a}_2, \ldots, \beta_n \boldsymbol{a}_n)]\,\big) = \beta_1\beta_2\cdots\beta_n \det\big(\,[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)]\,\big).$$

4. **Permutation of columns** If $\tau \in S_n$ then

$$\det(\boldsymbol{B}) := \det[(\boldsymbol{a}_{\tau(1)}, \boldsymbol{a}_{\tau(2)}, \ldots, \boldsymbol{a}_{\tau(n)})] = \operatorname{sign}(\tau)\det[(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)].$$

5. **Additivity**

$$\begin{aligned}\det\big(\,[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k-1}, \boldsymbol{a}_k + \boldsymbol{a}'_k, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_n]\,\big)\\ = \det\big(\,[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]\,\big) + \det\big(\,[\boldsymbol{a}_1, \ldots, \boldsymbol{a}'_k \ldots, \boldsymbol{a}_n]\,\big).\end{aligned}$$

6. **Singular matrix** $\det(A) = 0$ if and only if A is singular.
7. **Product rule** If $A, B \in \mathbb{C}^{n,n}$ then $\det(AB) = \det(A)\det(B)$.
8. **Block triangular** If A is block triangular with diagonal blocks B and C then $\det(A) = \det(B)\det(C)$.

Proof.

1. If $\sigma \neq \epsilon$, we can find distinct integers i and j such that $\sigma(i) > i$ and $\sigma(j) < j$. But then $a_{\sigma(i),i} = 0$ if $\boldsymbol{A}$ is upper triangular and $a_{\sigma(j),j} = 0$ if $\boldsymbol{A}$ is lower triangular. Hence

$$\det(\boldsymbol{A}) = \operatorname{sign}(\epsilon) a_{\epsilon(1),1} a_{\epsilon(2),2} \cdots a_{\epsilon(n),n} = a_{1,1}a_{2,2}\cdots a_{n,n}.$$

   Since the identity matrix is triangular with all diagonal elements equal to one, we have that $\det(\boldsymbol{I}) = 1$.
2. By definition of $\boldsymbol{A}^T$ and the det-function

$$\det(\boldsymbol{A}^T) = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)}.$$

   Consider an element $a_{i,\sigma(i)}$. If $\sigma(i) = j$ then

$$a_{i,\sigma(i)} = a_{\sigma^{-1}(j),j}.$$

   Since $\sigma(1), \sigma(2), \ldots, \sigma(n)$ ranges through $\{1, 2, \ldots, n\}$, we obtain

$$\begin{aligned}\det(\boldsymbol{A}^T) &= \textstyle\sum_{\sigma\in S_n} \operatorname{sign}(\sigma) a_{\sigma^{-1}(1),1} a_{\sigma^{-1}(2),2} \cdots a_{\sigma^{-1}(n),n}\\ &= \textstyle\sum_{\sigma\in S_n} \operatorname{sign}(\sigma^{-1}) a_{\sigma^{-1}(1),1} a_{\sigma^{-1}(2),2} \cdots a_{\sigma^{-1}(n),n}\\ &= \textstyle\sum_{\sigma^{-1}\in S_n} \operatorname{sign}(\sigma^{-1}) a_{\sigma^{-1}(1),1} a_{\sigma^{-1}(2),2} \cdots a_{\sigma^{-1}(n),n}\\ &= \det(\boldsymbol{A}).\end{aligned}$$

3. This follows immediately from the definition of $\det[(\beta_1\boldsymbol{a}_1, \beta_2\boldsymbol{a}_2, \ldots, \beta_n\boldsymbol{a}_n)]$.

4. We have

$$\det(\boldsymbol{B}) = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{\sigma(1),\tau(1)} a_{\sigma(2),\tau(2)} \cdots a_{\sigma(n),\tau(n)}.$$

Fix i in $\{1, 2, \ldots, n\}$. Let $k = \sigma(i)$ and $m = \tau(i)$. Then $\tau^{-1}(m) = i$ and $\sigma(\tau^{-1}(m)) = k$. Hence

$$a_{\sigma(i),\tau(i)} = a_{k,m} = a_{\sigma\tau^{-1}(m),m}.$$

Moreover, $\operatorname{sign}(\sigma) = \operatorname{sign}(\tau)\operatorname{sign}(\sigma\tau^{-1})$. Thus

$$\det(\boldsymbol{B}) = \operatorname{sign}(\tau) \sum_{\sigma \in S_n} \operatorname{sign}(\sigma\tau^{-1}) a_{\sigma\tau^{-1}(1),1} a_{\sigma\tau^{-1}(2),2} \cdots a_{\sigma\tau^{-1}(n),n}.$$

But as σ ranges over S_n, $\sigma\tau^{-1}$ also ranges over S_n. Hence

$$\det(\boldsymbol{B}) = \operatorname{sign}(\tau) \det[(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)].$$

5. This follows at once from the definition.
6. We observe that the determinant of a matrix is equal to the product of the eigenvalues and that a matrix is singular if and only if zero is an eigenvalue (cf. Theorems D.6, D.7). But then the result follows.
7. To better understand the general proof, we do the 2×2 case first. Let $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2)$, $\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2)$. Then

$$\boldsymbol{AB} = (\boldsymbol{A}\boldsymbol{b}_1, \boldsymbol{A}\boldsymbol{b}_2) = (b_{1,1}\boldsymbol{a}_1 + b_{2,1}\boldsymbol{a}_2, b_{1,2}\boldsymbol{a}_1 + b_{2,2}\boldsymbol{a}_2).$$

Using the additivity, we obtain

$$\begin{aligned} \det(\boldsymbol{AB}) \;=\; & \det(b_{1,1}\boldsymbol{a}_1, b_{1,2}\boldsymbol{a}_1) + \det(b_{2,1}\boldsymbol{a}_2, b_{1,2}\boldsymbol{a}_1) \\ & + \det(b_{1,1}\boldsymbol{a}_1, b_{2,2}\boldsymbol{a}_2) + \det(b_{2,1}\boldsymbol{a}_2, b_{2,2}\boldsymbol{a}_2). \end{aligned}$$

Next we have by homogeneity

$$\begin{aligned} \det(\boldsymbol{AB}) \;=\; & b_{1,1}b_{1,2}\det(\boldsymbol{a}_1, \boldsymbol{a}_1) + b_{2,1}b_{1,2}\det(\boldsymbol{a}_2, \boldsymbol{a}_1) \\ & + b_{1,1}b_{2,2}\det(\boldsymbol{a}_1, \boldsymbol{a}_2) + b_{2,1}b_{2,2}\det(\boldsymbol{a}_2, \boldsymbol{a}_2). \end{aligned}$$

Property 6 implies that $\det(\boldsymbol{a}_1, \boldsymbol{a}_1) = \det(\boldsymbol{a}_2, \boldsymbol{a}_2) = 0$. Using Property 4, we obtain $\det(\boldsymbol{a}_2, \boldsymbol{a}_1) = -\det(\boldsymbol{a}_1, \boldsymbol{a}_2)$ and

$$\det(\boldsymbol{AB}) = (b_{1,1}b_{2,2} - b_{2,1}b_{1,2})\det(\boldsymbol{a}_1, \boldsymbol{a}_2) = \det(\boldsymbol{B})\det(\boldsymbol{A}).$$

The proof for $n > 2$ follows the $n = 2$ case step by step. Let $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n) = \boldsymbol{AB}$. Then

$$\boldsymbol{c}_i = \boldsymbol{A}\boldsymbol{b}_i = b_{1,i}\boldsymbol{a}_1 + b_{2,i}\boldsymbol{a}_2 + \cdots + b_{n,i}\boldsymbol{a}_n, \qquad i = 1, 2, \ldots, n.$$

Using the additivity, we obtain

$$\det(\boldsymbol{AB}) = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_n=1}^{n} \det[(b_{i_1,1}\boldsymbol{a}_{i_1}, b_{i_2,2}\boldsymbol{a}_{i_2}, \ldots, b_{i_n,n}\boldsymbol{a}_{i_n})].$$

Next we have by homogeneity

$$\det(\boldsymbol{AB}) = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n} b_{i_1,1}b_{i_2,2}\cdots b_{i_n,n}\det[(\boldsymbol{a}_{i_1},\boldsymbol{a}_{i_2},\ldots,\boldsymbol{a}_{i_n})].$$

Property 6 implies that $\det[(\boldsymbol{a}_{i_1},\ldots,\boldsymbol{a}_{i_n})] = 0$ if any two of the indices $i_1,\ldots,i_n$ are equal. Therefore we only get a contribution to the sum whenever $i_1,\ldots,i_n$ is a permutation of $\{1,2,\ldots,n\}$. Thus

$$\det(\boldsymbol{AB}) = \sum_{\sigma\in S_n} b_{\sigma(1),1}\cdots b_{\sigma(n),n}\det[(\boldsymbol{a}_{\sigma(1)},\ldots,\boldsymbol{a}_{\sigma(n)})].$$

By Property 4 we obtain

$$\det(\boldsymbol{AB}) = \sum_{\sigma\in S_n} \operatorname{sign}(\tau) b_{\sigma(1),1}\cdots b_{\sigma(n),n}\det[(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n)].$$

According to the definition of $\det(\boldsymbol{B})$ this is equal to $\det(\boldsymbol{B})\det(\boldsymbol{A})$.

8. Suppose $\boldsymbol{A}$ is block upper triangular. Let

$$S_{n,k} = \{\sigma\in S_n : \sigma(i)\le k \text{ if } i\le k, \text{ and } \sigma(i)\ge k{+}1 \text{ if } i\ge k{+}1\}.$$

We claim that $a_{\sigma(1),1}\cdots a_{\sigma(n),n} = 0$ if $\sigma\notin S_{n,k}$, because if $\sigma(i) > k$ for some $i\le k$ then $a_{\sigma(i),i} = 0$ since it lies in the zero part of $\boldsymbol{A}$. If $\sigma(i)\le k$ for some $i\ge k{+}1$, we must have $\sigma(j) > k$ for some $j\le k$ to make "room" for $\sigma(i)$, and $a_{\sigma(j),j} = 0$. It follows that

$$\det(\boldsymbol{A}) = \sum_{\sigma\in S_{n,k}} \operatorname{sign}(\sigma) a_{\sigma(1),1}\cdots a_{\sigma(n),n}.$$

Define

$$\rho(i) = \begin{cases} \sigma(i) & i = 1,\ldots,k \\ i & i = k{+}1,\ldots,n, \end{cases} \qquad \tau(i) = \begin{cases} i & i = 1,\ldots,k \\ \sigma(i) & i = k{+}1,\ldots,n. \end{cases}$$

If $\sigma\in S_{n,k}$, ρ and τ will be permutations. Moreover, $\sigma = \rho\tau$. Define $\hat\rho$ and $\hat\tau$ in S_k and S_{n-k} respectively by $\hat\rho(i) = \rho(i)$, $i = 1,\ldots,k$, and $\hat\tau(i) = \tau(i{+}k) - k$ for $i = 1,\ldots,n{-}k$. As σ ranges over $S_{n,k}$, $\hat\rho$ and $\hat\tau$ will take on all values in S_k and S_{n-k} respectively. Since $\operatorname{sign}(\hat\rho) = \operatorname{sign}(\rho)$ and $\operatorname{sign}(\hat\tau) = \operatorname{sign}(\tau)$, we find

$$\operatorname{sign}(\sigma) = \operatorname{sign}(\rho)\operatorname{sign}(\tau) = \operatorname{sign}(\hat\rho)\operatorname{sign}(\hat\tau).$$

Then

$$\begin{aligned} \det(\boldsymbol{A}) &= \textstyle\sum_{\hat\rho\in S_k}\sum_{\hat\tau\in S_{n-k}} \operatorname{sign}(\hat\rho)\operatorname{sign}(\hat\tau) b_{\hat\rho(1),1}\cdots b_{\hat\rho(k),k} d_{\hat\tau(1),1}\cdots d_{\hat\tau(n-k),n-k} \\ &= \det(\boldsymbol{B})\det(\boldsymbol{D}). \end{aligned}$$

□

C.3 The Adjoint Matrix and Cofactor Expansion

We start with a useful formula for the solution of a linear system.

Let $\boldsymbol{A}_j(\boldsymbol{b})$ denote the matrix obtained from $\boldsymbol{A}$ by replacing the jth column of $\boldsymbol{A}$ by $\boldsymbol{b}$. For example,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \quad \boldsymbol{A}_1(\boldsymbol{b}) = \begin{pmatrix} 3 & 2 \\ 6 & 1 \end{pmatrix}, \quad \boldsymbol{A}_2(\boldsymbol{b}) = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix},$$
$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{I}_1(\boldsymbol{x}) = \begin{pmatrix} x_1 & 0 \\ x_2 & 1 \end{pmatrix}, \quad \boldsymbol{I}_2(\boldsymbol{x}) = \begin{pmatrix} 1 & x_1 \\ 0 & x_2 \end{pmatrix}.$$

Theorem C.2 (Cramers rule (1750)) *Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ with $\det(\boldsymbol{A}) \neq 0$ and $\boldsymbol{b} \in \mathbb{C}^n$. Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ be the unique solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Then*

$$x_j = \frac{\det(\boldsymbol{A}_j(\boldsymbol{b}))}{\det(\boldsymbol{A})}, \quad j = 1, 2, \ldots, n.$$

Proof. Since $1 = \det(\boldsymbol{I}) = \det(\boldsymbol{A}\boldsymbol{A}^{-1}) = \det(\boldsymbol{A})\det(\boldsymbol{A}^{-1})$ we have $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$. Then

$$\begin{aligned} \frac{\det(\boldsymbol{A}_j(\boldsymbol{b}))}{\det(\boldsymbol{A})} &= \det(\boldsymbol{A}^{-1}\boldsymbol{A}_j(\boldsymbol{b})) \\ &= \det([\boldsymbol{A}^{-1}\boldsymbol{a}_1, \ldots, \boldsymbol{A}^{-1}\boldsymbol{a}_{j-1}, \boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1}\boldsymbol{a}_{j+1}, \ldots, \boldsymbol{A}^{-1}\boldsymbol{a}_n]) \\ &= \det([\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{j-1}, \boldsymbol{x}, \boldsymbol{e}_{j+1}, \ldots, \boldsymbol{e}_n]) = x_j, \end{aligned}$$

where we used Property 8 for the last equality. ∎

Exercise C.3 *Solve the following system by Cramers rule:*

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

Let $\boldsymbol{A}_{i,j}$ denote the submatrix of $\boldsymbol{A}$ obtained by deleting the ith row and jth column of $\boldsymbol{A}$. For example,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad \boldsymbol{A}_{1,1} = \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix}, \quad \boldsymbol{A}_{1,2} = \begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix},$$
$$\boldsymbol{A}_{2,1} = \begin{bmatrix} 2 & 3 \\ 8 & 9 \end{bmatrix}, \quad \boldsymbol{A}_{2,2} = \begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix}, \quad \text{etc.}$$

Definition C.4 (Cofactor and Adjoint) *For $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and $1 \leq i, j \leq n$ the determinant $\det(\boldsymbol{A}_{ij})$ is called the* **cofactor** *of a_{ij}. The matrix $adj(\boldsymbol{A}) \in \mathbb{C}^{n,n}$ with elements $(-1)^{i+j}\det(\boldsymbol{A}_{j,i})$ is called the* **adjoint** *of $\boldsymbol{A}$.*

Exercise C.5 *Show that if*

$$\boldsymbol{A} = \begin{bmatrix} 2 & -6 & 3 \\ 3 & -2 & -6 \\ 6 & 3 & 2 \end{bmatrix},$$

then

$$adj(\boldsymbol{A}) = \begin{bmatrix} 14 & 21 & 42 \\ -42 & -14 & 21 \\ 21 & -42 & 14 \end{bmatrix}.$$

Moreover,

$$adj(\boldsymbol{A})\boldsymbol{A} = \begin{bmatrix} 343 & 0 & 0 \\ 0 & 343 & 0 \\ 0 & 0 & 343 \end{bmatrix} = \det(\boldsymbol{A})\boldsymbol{I}.$$

Theorem C.6 (The inverse as an adjoint) *If $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular then*

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})}\ adj(\boldsymbol{A}).$$

Proof. Let $\boldsymbol{A}^{-1} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_j = [x_{1j}, \ldots, x_{nj}]^T$. The equation $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ implies that $\boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{e}_j$ for $j = 1, \ldots, n$ and by Cramer's rule

$$x_{ij} = \frac{\det(\boldsymbol{A}_i(\boldsymbol{e}_j))}{\det(\boldsymbol{A})} = (-1)^{i+j}\frac{\det(\boldsymbol{A}_{ji})}{\det(\boldsymbol{A})},\ j = 1, 2, \ldots, n.$$

For the last equality we first interchange the first and ith column of $\boldsymbol{A}_i(\boldsymbol{e}_j)$. By Property 4 it follows that $\det(\boldsymbol{A}_i(\boldsymbol{e}_j)) = (-1)^{i-1}\det\big([\boldsymbol{e}_j, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n]\big)$. We then interchange row j and row 1. Using Property 8 we obtain

$$\det(\boldsymbol{A}_i(\boldsymbol{e}_j)) = (-1)^{i+j-2}\det(\boldsymbol{A}_{ji}) = (-1)^{i+j}\det(\boldsymbol{A}_{ji}).$$

□

Corollary C.7 *For any $\boldsymbol{A} \in \mathbb{C}^{n,n}$ we have*

$$\boldsymbol{A}\ adj(\boldsymbol{A}) = \ adj(\boldsymbol{A})\boldsymbol{A} = \det(\boldsymbol{A})\boldsymbol{I}. \tag{C.2}$$

Proof. If $\boldsymbol{A}$ is nonsingular then (C.2) follows from Theorem C.6. We simply multiply by $\boldsymbol{A}$ from the left and from the right. Suppose next that $\boldsymbol{A}$ is singular with m zero eigenvalues $\lambda_1, \ldots, \lambda_m$ and nonzero eigenvalues $\lambda_{m+1}, \ldots, \lambda_n$. We define $\epsilon_0 := \min_{m+1 \le j \le n}|\lambda_j|$. For any $\epsilon \in (0, \epsilon_0)$ the matrix $\boldsymbol{A} + \epsilon\boldsymbol{I}$ has nonzero eigenvalues $\epsilon, \ldots, \epsilon, \lambda_{m+1}+\epsilon, \ldots, \lambda_n+\epsilon$ and hence is nonsingular. By what we have proved

$$(\boldsymbol{A} + \epsilon\boldsymbol{I})\ \mathrm{adj}(\boldsymbol{A} + \epsilon\boldsymbol{I}) = \ \mathrm{adj}(\boldsymbol{A} + \epsilon\boldsymbol{I})(\boldsymbol{A} + \epsilon\boldsymbol{I}) = \det(\boldsymbol{A} + \epsilon\boldsymbol{I})\boldsymbol{I}. \tag{C.3}$$

Since the elements in $\boldsymbol{A} + \epsilon\boldsymbol{I}$ and $\mathrm{adj}(\boldsymbol{A} + \epsilon\boldsymbol{I})$ depend continuously on ϵ we can take limits in (C.3) to obtain (C.2). □

Corollary C.8 (Cofactor expansion) *For any $\boldsymbol{A} \in \mathbb{C}^{n,n}$ we have*

$$\det(\boldsymbol{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij}) \text{ for } i = 1, \ldots, n, \tag{C.4}$$

$$\det(\boldsymbol{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij}) \text{ for } j = 1, \ldots, n. \tag{C.5}$$

Proof. By (C.2) we have $\boldsymbol{A}\,\mathrm{adj}(\boldsymbol{A}) = \det(\boldsymbol{A})\boldsymbol{I}$. But then $\det(\boldsymbol{A}) = \boldsymbol{e}_i^T \boldsymbol{A}\mathrm{adj}(\boldsymbol{A})\boldsymbol{e}_i = \sum_{j=1}^{n}(-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij})$ which is (C.4). Applying this row expansion to $\boldsymbol{A}^T$ we find $\det(\boldsymbol{A}^T) = \sum_{j=1}^{n}(-1)^{i+j} a_{ji} \det(\boldsymbol{A}_{ji})$. Switching the roles of i and j proves (C.5). □

C.4 Computing Determinants

A determinant of an n-by-n matrix computed from the definition can contain up to $n!$ terms and we need other methods to compute determinants.

A matrix can be reduced to upper triangular form using elementary row operations. We can then use Property 1. to compute the determinant. The elementary operations using either rows or columns are

1. Interchanging two rows(columns).
2. Multiply a row(column) by a scalar α.
3. Add a constant multiple of one row(column) to another row(column).

Let $\boldsymbol{B}$ be the result of performing an elementary operation on $\boldsymbol{A}$. For the three elementary operations the numbers $\det(\boldsymbol{A})$ and $\det(\boldsymbol{B})$ are related as follows.

1. $\det(\boldsymbol{B}) = -\det(\boldsymbol{A})$ (from Property 4.)
2. $\det(\boldsymbol{B}) = \alpha \det(\boldsymbol{A})$ (from Property 3.)
3. $\det(\boldsymbol{B}) = \det(\boldsymbol{A})$(from Properties 5., 7.)

It follows from Property 2. that it is enough to show this for column operations. The proof of 1. and 2. are immediate. For 3. suppose we add α times column k to column i for some $k \neq i$. Then using Properties 5. and 7. we find

$$\begin{aligned}\det(\boldsymbol{B}) &= \det\left(\left[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{a}_i + \alpha\boldsymbol{a}_k, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n\right]\right) \\ &\overset{5.}{=} \det(\boldsymbol{A}) + \det\left(\left[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \alpha\boldsymbol{a}_k, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n\right]\right) \overset{7.}{=} \det(A)\end{aligned}$$

To compute the value of a determinant it is often convenient to use row- or column operations to introduce zeros in a row or column of $\boldsymbol{A}$ and then use one of the cofactor expansions in Corollary C.8.

Example C.9 The equation for a straight line through two points (x_1, y_1) and

(x_2, y_2) in the plane can be written as the equation

$$\det(\boldsymbol{A}) := \begin{vmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix} = 0$$

involving a determinant of order 3. We can compute this determinant using row operations of type 3. Subtracting row 2 from row 3 and then row 1 from row 2 we obtain

$$\begin{vmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix} = \begin{vmatrix} 1 & x & y \\ 0 & x_1 - x & y_1 - y \\ 0 & x_2 - x_1 & y_2 - y_1 \end{vmatrix} = (x_1 - x)(y_2 - y_1) - (y_1 - y)(x_2 - x_1).$$

Rearranging the equation $\det(\boldsymbol{A}) = 0$ we obtain

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

which is the slope form of the equation of a straight line.

Exercise C.10 *Show that the equation for the plane through the points* (x_1, y_1, z_1), (x_2, y_2, z_2) *and* (x_3, y_3, z_3) *is*

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0.$$

Exercise C.11 *Let* $P_i = (x_i, y_i)$, $i = 1, 2, 3$, *be three points in the plane defining a triangle* T. *Show that the area of* T *is*

$$A(T) = \frac{1}{2} \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix}$$

Hint: $A(T) = A(ABP_3P_1) + A(P_3BCP_2) - A(P_1ACP_2)$, *c.f. Figure C.12.*

Exercise C.13 *Show that*

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{i>j} (x_i - x_j),$$

where $\prod_{i>j}(x_i - x_j) = \prod_{i=2}^{n}(x_i - x_1)(x_i - x_2)\cdots(x_i - x_{i-1})$. *This determinant is called the* Van der Monde determinant. *Hint: Subtract* x_n^k *times column* k *from column* $k+1$ *for* $k = n-1, n-2, \ldots, 1$.

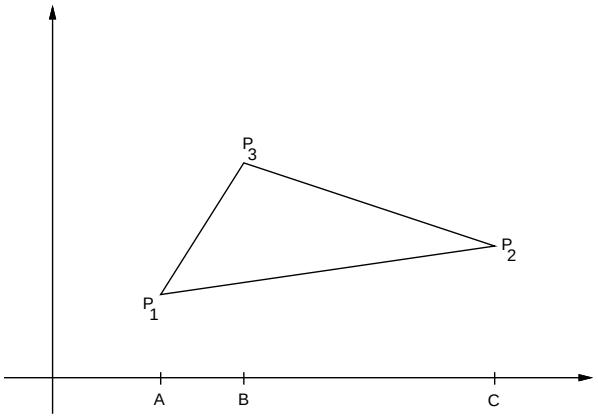


Figure C.12. *The triangle T defined by the three points P_1, P_2 and P_3.*

Exercise C.14 *(Cauchy 1842). Let $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]^T$ be in $\mathbb{R}^n$.*

a) *Consider the matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ with elements $a_{i,j} = 1/(\alpha_i + \beta_j)$, $i, j = 1, 2, \ldots, n$. Show that*

$$\det(\boldsymbol{A}) = Pg(\boldsymbol{\alpha})g(\boldsymbol{\beta})$$

where $P = \prod_{i=1}^n \prod_{j=1}^n a_{ij}$, and for $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_n]^T$

$$g(\boldsymbol{\gamma}) = \prod_{i=2}^{n} (\gamma_i - \gamma_1)(\gamma_i - \gamma_2) \cdots (\gamma_i - \gamma_{i-1})$$

Hint: Multiply the ith row of $\boldsymbol{A}$ by $\prod_{j=1}^n (\alpha_i + \beta_j)$ for $i = 1, 2, \ldots, n$. Call the resulting matrix $\boldsymbol{C}$. Each element of $\boldsymbol{C}$ is a product of $n-1$ factors $\alpha_r + \beta_s$. Hence $\det(\boldsymbol{C})$ is a sum of terms where each term contain precisely $n(n-1)$ factors $\alpha_r + \beta_s$. Thus $\det(\boldsymbol{C}) = q(\alpha, \beta)$ where q is a polynomial of degree at most $n(n-1)$ in α_i and β_j. Since $\det(\boldsymbol{A})$ and therefore $\det(\boldsymbol{C})$ vanishes if $\alpha_i = \alpha_j$ for some $i \neq j$ or $\beta_r = \beta_s$ for some $r \neq s$, we have that $q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ must be divisible by each factor in $g(\boldsymbol{\alpha})$ and $g(\boldsymbol{\beta})$. Since $g(\boldsymbol{\alpha})$ and $g(\boldsymbol{\beta})$ is a polynomial of degree $n(n-1)$, we have

$$q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = kg(\boldsymbol{\alpha})g(\boldsymbol{\beta})$$

for some constant k independent of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Show that $k = 1$ by choosing $\beta_i + \alpha_i = 0$, $i = 1, 2, \ldots, n$.

b) *Notice that the cofactor of any element in the above matrix $\boldsymbol{A}$ is the determinant of a matrix of similar form. Use the cofactor and determinant of $\boldsymbol{A}$ to represent the elements of $\boldsymbol{A}^{-1} = (b_{j,k})$. Answer:*

$$b_{j,k} = (\alpha_k + \beta_j)A_k(-\beta_j)B_j(-\alpha_k),$$

where

$$A_k(x) = \prod_{s \neq k} \left(\frac{\alpha_s - x}{\alpha_s - \alpha_k} \right), \qquad B_k(x) = \prod_{s \neq k} \left(\frac{\beta_s - x}{\beta_s - \beta_k} \right).$$

Exercise C.15 *Let $\boldsymbol{H}_n = (h_{i,j})$ be the $n \times n$ matrix with elements $h_{i,j} = 1/(i+j-1)$. Use Exercise C.14 to show that the elements $t^n_{i,j}$ in $\boldsymbol{T}_n = \boldsymbol{H}_n^{-1}$ are given by*

$$t^n_{i,j} = \frac{f(i)f(j)}{i+j-1},$$

where

$$f(i+1) = \left(\frac{i^2 - n^2}{i^2} \right) f(i), \quad i = 1, 2, \ldots, \quad f(1) = -n.$$

C.5 Some Useful Determinant Formulas

Suppose $\boldsymbol{A} \in \mathbb{C}^{m,n}$ and suppose for an integer $r \leq \min\{m, n\}$ that $\boldsymbol{i} = \{i_1, \ldots, i_r\}$ and $\boldsymbol{j} = \{j_1, \ldots, j_r\}$ are integers with $1 \leq i_1 < i_2 < \cdots < i_r \leq m$ and $1 \leq j_1 < j_2 < \cdots < j_r$. We let

$$\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j}) = \begin{bmatrix} a_{i_1,j_1} & \cdots & a_{i_1,j_r} \\ \vdots & & \vdots \\ a_{i_r,j_1} & \cdots & a_{i_r,j_r} \end{bmatrix}$$

be the submatrix of $\boldsymbol{A}$ consisting of rows $i_1, \ldots, i_r$ and columns $j_1, \ldots, j_r$. The following formula bears a strong resemblance to the formula for matrix multiplication.

Theorem C.16 (Cauchy-Binet formula) *Let $\boldsymbol{A} \in \mathbb{C}^{m,p}$, $\boldsymbol{B} \in \mathbb{C}^{p,n}$ and $\boldsymbol{C} = \boldsymbol{AB}$. Suppose $1 \leq r \leq \min\{m, n, p\}$ and let $\boldsymbol{i} = \{i_1, \ldots, i_r\}$ and $\boldsymbol{j} = \{j_1, \ldots, j_r\}$ be integers with $1 \leq i_1 < i_2 < \cdots < i_r \leq m$ and $1 \leq j_1 < j_2 < \cdots < j_r \leq n$. Then*

$$\det\big(\boldsymbol{C}(\boldsymbol{i}, \boldsymbol{j})\big) = \sum_{\boldsymbol{k}} \det\big(\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{k})\big) \det\big(\boldsymbol{B}(\boldsymbol{k}, \boldsymbol{j})\big), \tag{C.6}$$

where we sum over all $\boldsymbol{k} = \{k_1, \ldots, k_r\}$ with $1 \leq k_1 < k_2 < \cdots < k_r \leq p$.

Appendix D

Eigenvalues and Eigenvectors

Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is a square matrix, $\lambda \in \mathbb{C}$ and $\boldsymbol{x} \in \mathbb{C}^n$. We say that $(\lambda, \boldsymbol{x})$ is an **eigenpair** for $\boldsymbol{A}$ if $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ and $\boldsymbol{x}$ is nonzero. The scalar λ is called an **eigenvalue** and $\boldsymbol{x}$ is said to be an **eigenvector**. If $(\lambda, \boldsymbol{x})$ is an eigenpair then $(\lambda, \alpha\boldsymbol{x})$ is an eigenpair for any $\alpha \in \mathbb{C}$ with $\alpha \neq 0$. An eigenvector is a special vector that is mapped by $\boldsymbol{A}$ into a vector parallel to itself. The length is increased if $|\lambda| > 1$ and decreased if $|\lambda| < 1$. The set of distinct eigenvalues is called the **spectrum** of $\boldsymbol{A}$ and is denoted by $\sigma(\boldsymbol{A})$.

D.1 The Characteristic Polynomial

D.1.1 The characteristic equation

Lemma D.1 *For any $\boldsymbol{A} \in \mathbb{C}^{n,n}$ we have $\lambda \in \sigma(\boldsymbol{A}) \Longleftrightarrow \det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$.*

Proof. Suppose $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}$. The equation $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ can be written $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. Since $\boldsymbol{x}$ is nonzero the matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$ must be singular with a zero determinant. Conversely, if $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ then $\boldsymbol{A} - \lambda\boldsymbol{I}$ is singular and $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$ for some nonzero $\boldsymbol{x} \in \mathbb{C}^n$. Thus $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ and $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}$. □

We observe that $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ if and only if $\det(\lambda\boldsymbol{I} - \boldsymbol{A}) = 0$.The equation $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ or equivalently $\det(\lambda\boldsymbol{I} - \boldsymbol{A}) = 0$ is called the **characteristic equation** of $\boldsymbol{A}$.

Definition D.2 *The function $\pi_{\boldsymbol{A}} \colon \mathbb{C} \to \mathbb{C}$ given by $\pi_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{I})$ is called the* **characteristic polynomial** *of $\boldsymbol{A}$.*

To see that $\pi_{\boldsymbol{A}}$ is in fact a polynomial let us take a closer look at this function. For $n = 3$ we have

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix}.$$

Expanding this determinant by the first column we find

$$\begin{aligned} \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (a_{11} - \lambda) \begin{vmatrix} a_{22} - \lambda & a_{23} \\ a_{32} & a_{33} - \lambda \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} - \lambda \end{vmatrix} \\ + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} - \lambda & a_{23} \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda)(a_{33} - \lambda) + r(\lambda) \end{aligned}$$

for some polynomial r of degree at most one. In general

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda) + r(\lambda), \tag{D.1}$$

where each term in $r(\lambda)$ has at most $n - 2$ factors containing λ. It follows that r is a polynomial of degree at most $n - 2$, $\pi_{\boldsymbol{A}}$ is a polynomial of exact degree n, and the eigenvalues are the roots of this polynomial.

By the fundamental theorem of algebra an $n \times n$ matrix has precisely n eigenvalues $\lambda_1, \ldots, \lambda_n$ some of which might be complex even if $\boldsymbol{A}$ is real. The complex eigenpairs of a real matrix occur in complex conjugate pairs. Indeed, taking the complex conjugate on both sides of the equation $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$ with $\boldsymbol{A}$ real gives $\boldsymbol{A}\overline{\boldsymbol{x}} = \overline{\lambda}\overline{\boldsymbol{x}}$.

The following result will be useful.

Theorem D.3 *Suppose $(\mu, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A} \in \mathbb{C}^{n,n}$. Then*

1. *If $\boldsymbol{A}$ is nonsingular then $(\mu^{-1}, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}^{-1}$.*
2. *$(\mu^k, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}^k$ for $k \in \mathbb{N}$.*
3. *If p given by $p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k$ is a polynomial, then $(p(\mu), \boldsymbol{x})$ is an eigenpair for the matrix $p(\boldsymbol{A}) := a_0 \boldsymbol{I} + a_1 \boldsymbol{A} + a_2 \boldsymbol{A}^2 + \cdots + a_k \boldsymbol{A}^k$.*
4. *μ is an eigenvalue for $\boldsymbol{A}^T$, in fact $\pi_{\boldsymbol{A}^T} = \pi_{\boldsymbol{A}}$.*
5. *$\overline{\mu}$ is an eigenvalue for $\boldsymbol{A}^*$, in fact $\pi_{\boldsymbol{A}^*}(\overline{\lambda}) = \overline{\pi_{\boldsymbol{A}}(\lambda)}$ for all $\lambda \in \mathbb{C}$.*
6. *If $\boldsymbol{A} = \left[\begin{smallmatrix} \boldsymbol{B} & \boldsymbol{C} \\ 0 & \boldsymbol{D} \end{smallmatrix}\right]$ is block triangular then $\pi_{\boldsymbol{A}} = \pi_{\boldsymbol{B}} \cdot \pi_{\boldsymbol{D}}$.*

Proof.

1. $\boldsymbol{A}\boldsymbol{x} = \mu \boldsymbol{x} \implies \boldsymbol{A}^{-1}\boldsymbol{x} = \mu^{-1}\boldsymbol{x}$.

2. We use induction on k. The case $k = 1$ is trivial and if $\boldsymbol{A}^{k-1}\boldsymbol{x} = \mu^{k-1}\boldsymbol{x}$ then $\boldsymbol{A}^k \boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}^{k-1}\boldsymbol{x} = \mu^{k-1}\boldsymbol{A}\boldsymbol{x} = \mu^k \boldsymbol{x}$.

3. $p(\boldsymbol{A})\boldsymbol{x} = \sum_{j=0}^k a_j \boldsymbol{A}^j \boldsymbol{x} \overset{2.}{=} \sum_{j=0}^k a_j \mu^j \boldsymbol{x} = p(\mu)\boldsymbol{x}$.

4. Since $\det(\boldsymbol{B}^T) = \det(\boldsymbol{B})$ for any matrix $\boldsymbol{B}$ we find for any $\lambda \in \mathbb{C}$

$$\pi_{\boldsymbol{A}^T}(\lambda) = \det(\boldsymbol{A}^T - \lambda \boldsymbol{I}) = \det\left((\boldsymbol{A} - \lambda \boldsymbol{I})^T\right) = \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \pi_{\boldsymbol{A}}(\lambda).$$

Thus $\boldsymbol{A}^T$ and $\boldsymbol{A}$ have the same characteristic polynomial and hence the same eigenvalues.

5. We have $\pi_{\boldsymbol{A}^*}(\overline{\lambda}) \stackrel{4}{=} \pi_{\overline{\boldsymbol{A}}}(\overline{\lambda}) = \det(\overline{\boldsymbol{A}} - \overline{\lambda}\boldsymbol{I}) = \overline{\det(\boldsymbol{A} - \lambda \boldsymbol{I})} = \overline{\pi_{\boldsymbol{A}}(\lambda)}$. Thus $\pi_{\boldsymbol{A}}(\lambda) = 0 \Leftrightarrow \pi_{\boldsymbol{A}^*}(\overline{\lambda}) = 0$ and the result follows.

6. By Property 8 of determinants

$$\pi_{\boldsymbol{A}}(\lambda) = \begin{vmatrix} \boldsymbol{B} - \lambda \boldsymbol{I} & \boldsymbol{C} \\ 0 & \boldsymbol{D} - \lambda \boldsymbol{I} \end{vmatrix} = \det(\boldsymbol{B} - \lambda \boldsymbol{I}) \det(\boldsymbol{D} - \lambda \boldsymbol{I}) = \pi_{\boldsymbol{B}}(\lambda) \cdot \pi_{\boldsymbol{D}}(\lambda).$$

□

In general it is not easy to find all eigenvalues of a matrix. One notable exception is a triangular matrix.

Theorem D.4 *The eigenvalues of a triangular matrix are given by its diagonal elements.*

Proof. If $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is triangular then $\boldsymbol{A} - \lambda \boldsymbol{I}$ is also triangular with diagonal elements $a_{ii} - \lambda$ for $i = 1, \ldots, n$. But then the roots of $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \prod_{i=1}^{n}(a_{ii} - \lambda) = 0$ are $\lambda_i = a_{ii}$ for $i = 1, \ldots, n$. □

To find the eigenvectors of a triangular matrix requires more work. Indeed, the eigenvectors are nontrivial solutions of a homogenous triangular linear system with at least one zero on the diagonal.

Example D.5 *The* 3×3 *matrix* $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{smallmatrix}\right]$ *has the eigenvalue* $\lambda = 1$. *The homogenous triangular linear system for an eigenvector* $\boldsymbol{x} = [x_1, x_2, x_3]^T$ *is*

$$(\boldsymbol{A} - \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \text{ or } \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

We find $x_2 = x_3 = 0$ *so any eigenvector must be a multiple of* $\boldsymbol{e}_1$.

There are two useful relations between the elements of a matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and its eigenvalues $\lambda_1, \ldots, \lambda_n$.

Theorem D.6 *For any* $\boldsymbol{A} \in \mathbb{C}^{n,n}$

$$\operatorname{trace}(\boldsymbol{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n, \quad \det(\boldsymbol{A}) = \lambda_1 \lambda_2 \cdots \lambda_n, \tag{D.2}$$

where the **trace** *of* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *is the sum of its diagonal elements*

$$\operatorname{trace}(\boldsymbol{A}) := a_{11} + a_{22} + \cdots + a_{nn} \tag{D.3}$$

and $\det(\boldsymbol{A})$ *is the determinant of* $\boldsymbol{A}$.

Proof. We compare two different expansion of $\pi_{\boldsymbol{A}}$. On the one hand from (D.1) we find

$$\pi_{\boldsymbol{A}}(\lambda) = (-1)^n\lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_0,$$

where $c_{n-1} = (-1)^{n-1}\operatorname{trace}(\boldsymbol{A})$ and $c_0 = \pi_{\boldsymbol{A}}(0) = \det(\boldsymbol{A})$. On the other hand

$$\pi_{\boldsymbol{A}}(\lambda) = (\lambda_1 - \lambda)\cdots(\lambda_n - \lambda) = (-1)^n\lambda^n + d_{n-1}\lambda^{n-1} + \cdots + d_0,$$

where $d_{n-1} = (-1)^{n-1}(\lambda_1 + \cdots + \lambda_n)$ and $d_0 = \lambda_1\cdots\lambda_n$. Since $c_j = d_j$ for all j we obtain (D.2). ∎

For a 2×2 matrix the characteristic equation takes the convenient form

$$\lambda^2 - \operatorname{trace}(\boldsymbol{A})\lambda + \det(\boldsymbol{A}) = 0. \tag{D.4}$$

Thus, if $\boldsymbol{A} = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right]$ then $\operatorname{trace}(\boldsymbol{A}) = 4$, $\det(\boldsymbol{A}) = 3$ so that $\pi_{\boldsymbol{A}}(\lambda) = \lambda^2 - 4\lambda + 3$.

In terms of eigenvalues we have an additional characterization of a singular matrix.

Theorem D.7 *The matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is singular if and only if zero is an eigenvalue.*

Proof. Zero is an eigenvalue if and only if $\pi_{\boldsymbol{A}}(0) = \det(\boldsymbol{A}) = 0$ which happens if and only if $\boldsymbol{A}$ is singular. ∎

Exercise D.8 *Find eigenvalues and eigenvectors of* $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix}$.

Exercise D.9 *Let $\lambda \in \sigma(A)$ where $A^2 = A \in \mathbb{C}^{n,n}$. Show that $\lambda = 0$ or $\lambda = 1$. (A matrix is called* idempotent *if $A^2 = A$).*

Exercise D.10 *Let $\lambda \in \sigma(A)$ where $A^k = 0$ for some $k \in \mathbb{N}$. Show that $\lambda = 0$. (A matrix $A \in \mathbb{C}^{n,n}$ such that $A^k = 0$ for some $k \in \mathbb{N}$ is called* nilpotent*).*

Exercise D.11 *Let $\lambda \in \sigma(A)$ where $A^H A = I$. Show that $|\lambda| = 1$.*

Exercise D.12 *Suppose $A \in \mathbb{C}^{n,n}$ is singular. Then we can find $\epsilon_0 > 0$ such that $A + \epsilon I$ is nonsingular for all $\epsilon \in (0, \epsilon_0)$. Hint:* $\det(A) = \lambda_1\lambda_2\cdots\lambda_n$, *where λ_i are the eigenvalues of A.*

Exercise D.13 *For $q_i \in \mathbb{C}$ let $f(\lambda) = \lambda^n + q_{n-1}\lambda^{n-1} + \cdots + q_0$ be a polynomial of degree n in λ. We derive two matrices which have $(-1)^n f$ as its characteristic polynomial.*

a) *Show that* $f = (-1)^n \pi_A$ *where*

$$A = \begin{bmatrix} -q_{n-1} & -q_{n-2} & \cdots & -q_1 & -q_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

A is called the companion matrix *of f.*

b) *Show that* $f = (-1)^n \pi_{A'}$ *where*

$$A' = \begin{bmatrix} 0 & 0 & \cdots & 0 & -q_0 \\ 1 & 0 & \cdots & 0 & -q_1 \\ 0 & 1 & \cdots & 0 & -q_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -q_{n-1} \end{bmatrix}.$$

Thus A' can also be regarded as a companion matrix for f.

D.2 Similarity Transformations

Row operations can be used to reduce a matrix to triangular form, but row operations change the eigenvalues of a matrix. We need a transformation which can be used to simplify a matrix without changing the eigenvalues.

Definition D.14 *Two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n,n}$ are said to be* **similar** *if there is a nonsingular matrix $\boldsymbol{S} \in \mathbb{C}^{n,n}$ such that $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$. The transformation $\boldsymbol{A} \to \boldsymbol{B}$ is called a* **similarity transformation**.

A similarity transformation does not change the eigenvalues.

Theorem D.15 *Similar matrices have the same characteristic polynomial and therefore the same eigenvalues.*

Proof. Let $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$. By properties of determinants

$$\begin{aligned} \pi_{\boldsymbol{B}}(\lambda) &= \det(\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} - \lambda\boldsymbol{I}) = \det\left(\boldsymbol{S}^{-1}(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{S}\right) \\ &= \det(\boldsymbol{S}^{-1})\det(\boldsymbol{A} - \lambda\boldsymbol{I})\det(\boldsymbol{S}) = \det(\boldsymbol{S}^{-1}\boldsymbol{S})\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \pi_{\boldsymbol{A}}(\lambda). \end{aligned}$$

But then $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same characteristic polynomial and hence the same eigenvalues. ∎

Consider next what a similarity transformation does to the eigenvectors.

Theorem D.16 1. *$(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ if and only if $(\lambda, \boldsymbol{S}\boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}$.*

2. *The columns of* $\boldsymbol{S}$ *are eigenvectors of* $\boldsymbol{A}$ *if and only if* $\boldsymbol{B}$ *is diagonal.*

Proof.

1. $\boldsymbol{Bx} = \lambda\boldsymbol{x} \Leftrightarrow \boldsymbol{S}^{-1}\boldsymbol{ASx} = \lambda\boldsymbol{x} \Leftrightarrow \boldsymbol{A}(\boldsymbol{Sx}) = \lambda(\boldsymbol{Sx})$, and $\boldsymbol{Sx} \neq 0$ since $\boldsymbol{S}$ is nonsingular.
2. Suppose $\boldsymbol{A}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ and let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ be the columns of $\boldsymbol{S}$. If $\boldsymbol{B}$ is diagonal then $(\lambda_i, \boldsymbol{e}_i)$ is an eigenpair for $\boldsymbol{B}$ and $(\lambda_i, \boldsymbol{Se}_i) = (\lambda_i, \boldsymbol{s}_i)$ is an eigenpair for $\boldsymbol{A}$ for $i = 1, \ldots, n$. Conversely, if $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{AS}$ and the columns $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ of $\boldsymbol{S}$ are eigenvectors of $\boldsymbol{A}$ then $\boldsymbol{As}_i = \lambda_i\boldsymbol{s}_i$ for $i = 1, \ldots, n$. But then $\boldsymbol{AS} = \boldsymbol{SC}$, where $\boldsymbol{C} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal. Thus $\boldsymbol{C} = \boldsymbol{S}^{-1}\boldsymbol{AS} = \boldsymbol{B}$ is diagonal.

□

The following result is sometimes useful.

Theorem D.17 *For any* $\boldsymbol{A} \in \mathbb{C}^{m,n}$ *and* $\boldsymbol{B} \in \mathbb{C}^{n,m}$ *the matrices* $\boldsymbol{AB}$ *and* $\boldsymbol{BA}$ *have the same spectrum. More precisely,*

$$\lambda^n \pi_{\boldsymbol{AB}}(\lambda) = \lambda^m \pi_{\boldsymbol{BA}}(\lambda), \quad \lambda \in \mathbb{C}.$$

Proof. Define block matrices of order $n + m$ by

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{AB} & 0 \\ \boldsymbol{B} & 0 \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} 0 & 0 \\ \boldsymbol{B} & \boldsymbol{BA} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{A} \\ 0 & \boldsymbol{I}_n \end{bmatrix}.$$

By Property 6. of Theorem D.3 we have $\pi_{\boldsymbol{E}}(\lambda) = \lambda^n \pi_{\boldsymbol{AB}}(\lambda)$ and $\pi_{\boldsymbol{F}}(\lambda) = \lambda^m \pi_{\boldsymbol{BA}}(\lambda)$. But $\boldsymbol{ES} = \boldsymbol{SF}$ so $\boldsymbol{E}$ and $\boldsymbol{F}$ are similar and have the same characteristic polynomial by the proof of Theorem D.15. □

D.3 Linear Independence of Eigenvectors

Definition D.18 *A square matrix* $\boldsymbol{A}$ *is* **diagonalizable** *if it is similar to a diagonal matrix,* $\boldsymbol{S}^{-1}\boldsymbol{AS} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Since $\boldsymbol{S}$ is nonsingular its columns are eigenvectors of $\boldsymbol{A}$, and Theorem D.16 implies the following result.

Theorem D.19 *A matrix is diagonalizable if and only if its eigenvectors form a basis for* $\mathbb{R}^n$ *or* $\mathbb{C}^n$.

A matrix with distinct eigenvalues can be diagonalized.

Theorem D.20 *Eigenvectors corresponding to distinct eigenvalues are linearly independent.*

Proof. Suppose $(\lambda_1, \boldsymbol{x}_1), \ldots, (\lambda_k, \boldsymbol{x}_k)$ are eigenpairs for $\boldsymbol{A} \in \mathbb{C}^{n,n}$ with $\lambda_i \neq \lambda_j$ for $i \neq j$. Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are linearly dependent. Let $m \leq k$ be the smallest positive integer so that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are linearly dependent. Since $\boldsymbol{x}_1 \neq 0$ we see that $m \geq 2$. For some nonzero $(c_1, \ldots, c_n)$ we have

$$\sum_{j=1}^{m} c_j \boldsymbol{x}_j = \boldsymbol{0}. \tag{D.5}$$

Applying $\boldsymbol{A}$ to this equation we obtain by linearity $\sum_{j=1}^{m} c_j \lambda_j \boldsymbol{x}_j = \boldsymbol{0}$. From this relation we subtract λ_m times (D.5) and find $\sum_{j=1}^{m-1} c_j(\lambda_j - \lambda_m)\boldsymbol{x}_j = \boldsymbol{0}$. But since $\lambda_j - \lambda_m \neq 0$ for $j = 1, \ldots, m-1$ and at least one $c_j \neq 0$ for $j < m$ we see that $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m-1}\}$ is linearly dependent, contradicting the minimality of m. □

Corollary D.21 *If $\boldsymbol{A} \in \mathbb{C}^{n,n}$ has distinct eigenvalues then the corresponding eigenvectors form a basis for $\mathbb{C}^n$.*

Proof. By the previous theorem the n eigenvectors are linearly independent. Since n is the dimension of $\mathbb{C}^n$ the eigenvectors form a basis. □

For a matrix with multiple eigenvalues the situation is more complicated. We have seen that any eigenvector of the 3×3 matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{smallmatrix}\right]$ is a multiple of $\boldsymbol{e}_1$. Thus this matrix does not have a set of linearly independent eigenvectors. On the other hand the unit matrix has a basis of eigenvectors, namely the unit vectors.

In order to characterize the matrices with eigenvectors which form a basis we have to count carefully the multiplicity of the eigenvalues. We consider two kinds of multiplicities called algebraic and geometric multiplicities. The algebraic multiplicity of an eigenvalue λ is simply the multiplicity of λ as a root in the characteristic polynomial. More formally we state:

Definition D.22 *We say that an eigenvalue λ of $\boldsymbol{A}$ has* **algebraic multiplicity** $a = a(\lambda) = a_{\boldsymbol{A}}(\lambda)$ *if $\pi_{\boldsymbol{A}}(z) = (z - \lambda)^a p(z)$, where $p(z) \neq 0$. The eigenvalue λ is simple (double, triple) if a is equal to one (two, three). A complex number z which is not an eigenvalue is defined to have algebraic multiplicity $a_{\boldsymbol{A}}(z) = 0$.*

To define the second kind of multiplicity we consider for each $\lambda \in \sigma(\boldsymbol{A})$ the nullspace

$$\ker(\boldsymbol{A} - \lambda \boldsymbol{I}) := \{\boldsymbol{x} \in \mathbb{R}^n : (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}\} \tag{D.6}$$

of $\boldsymbol{A} - \lambda \boldsymbol{I}$. This set consists of all eigenvectors of $\boldsymbol{A}$ corresponding to the eigenvalue λ. If $\boldsymbol{x}, \boldsymbol{y} \in \ker(\boldsymbol{A} - \lambda \boldsymbol{I})$ and α, β are scalars then $\alpha \boldsymbol{x} + \beta \boldsymbol{y} \in \ker(\boldsymbol{A} - \lambda \boldsymbol{I})$. So this nullspace is a subspace of $\mathbb{C}^n$. The dimension of the subspace must be at least one since $\boldsymbol{A} - \lambda \boldsymbol{I}$ is singular.

Definition D.23 *The* **geometric multiplicity** $g = g(\lambda) = g_{\boldsymbol{A}}(\lambda)$ *of an eigenvalue λ of $\boldsymbol{A}$ is the dimension of the nullspace* $\ker(\boldsymbol{A} - \lambda \boldsymbol{I})$.

Example D.24 *The $n \times n$ identity matrix has the eigenvalue $\lambda = 1$ with $\pi_I(\lambda) = (1-\lambda)^n$. Since $\boldsymbol{I} - \lambda\boldsymbol{I}$ is the zero matrix when $\lambda = 1$, the nullspace of $\boldsymbol{I} - \lambda\boldsymbol{I}$ is all of n-space and it follows that $a = g = n$. On the other hand the 3×3 matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{smallmatrix}\right]$ has the eigenvalue $\lambda = 1$ with $a = 3$ and $g = 1$.*

The geometric multiplicity of an eigenvalue is always bounded above by the algebraic multiplicity of the eigenvalue.

Theorem D.25 *For any square matrix $\boldsymbol{A}$ and any $\lambda \in \sigma(\boldsymbol{A})$ we have $g_{\boldsymbol{A}}(\lambda) \leq a_{\boldsymbol{A}}(\lambda)$.*

Proof. Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_g\}$ with $g := g_\lambda(\boldsymbol{A})$, be an orthonormal basis for $\ker(\boldsymbol{A} - \lambda\boldsymbol{I})$ and extend this set to an orthonormal basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ for $\mathbb{C}^n$. Then the matrix $\boldsymbol{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{C}^{n,n}$ is unitary and $\boldsymbol{V}^{-1} = \boldsymbol{V}^*$. Partition $\boldsymbol{V}$ as $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2]$, where $\boldsymbol{V}_1 := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_g]$ and $\boldsymbol{V}_2 := [\boldsymbol{v}_{g+1}, \ldots, \boldsymbol{v}_n]$. Then $\boldsymbol{A}\boldsymbol{V}_1 = \lambda\boldsymbol{V}_1$, $\boldsymbol{V}_1^*\boldsymbol{V}_1 = \boldsymbol{I}_g$, $\boldsymbol{V}_2^*\boldsymbol{V}_1 = 0$, and

$$\boldsymbol{B} := \boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1^* \\ \boldsymbol{V}_2^* \end{bmatrix} \boldsymbol{A} \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_1^*\boldsymbol{A}\boldsymbol{V}_1 & \boldsymbol{V}_1^*\boldsymbol{A}\boldsymbol{V}_2 \\ \boldsymbol{V}_2^*\boldsymbol{A}\boldsymbol{V}_1 & \boldsymbol{V}_2^*\boldsymbol{A}\boldsymbol{V}_2 \end{bmatrix} = \begin{bmatrix} \lambda\boldsymbol{I}_g & \boldsymbol{V}_1^*\boldsymbol{A}\boldsymbol{V}_2 \\ 0 & \boldsymbol{V}_2^*\boldsymbol{A}\boldsymbol{V}_2 \end{bmatrix}.$$

Since $\boldsymbol{B}$ is block triangular Property 6 of Theorem D.3 implies that $\pi_{\boldsymbol{B}}(z) = (z - \lambda)^g \pi_{\boldsymbol{V}_2^*\boldsymbol{A}\boldsymbol{V}_2}(z)$. But then $a_{\boldsymbol{B}}(\lambda) \geq g$. Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are similar they have the same characteristic polynomial, and it follows that $a_{\boldsymbol{A}}(\lambda) = a_{\boldsymbol{B}}(\lambda) \geq g_{\boldsymbol{A}}(\lambda)$. □

Definition D.26 *An eigenvalue where $g_{\boldsymbol{A}}(\lambda) < a_{\boldsymbol{A}}(\lambda)$ is said to be* **defective** *. A matrix is* **defective** *if at least one of its eigenvalues is defective.*

Theorem D.27 *A matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ has n linearly independent eigenvectors if and only if the algebraic and geometric multiplicity of all eigenvalues are the same.*

Proof. Suppose $\boldsymbol{A}$ has distinct eigenvalues $\mu_1, \ldots, \mu_r$ with algebraic multiplicities $a_1, \ldots, a_r$ and geometric multiplicities $g_1, \ldots, g_r$. Suppose $\{\boldsymbol{v}_{j1}, \ldots, \boldsymbol{v}_{j,g_j}\}$ is a basis for $\ker(\boldsymbol{A} - \mu_j\boldsymbol{I})$ for $j = 1, \ldots, r$. We claim that the combined set $\{\boldsymbol{v}_{jk}\}_{k=1,j=1}^{g_j,r}$ is linearly independent. We show this using induction on r. Suppose $\{\boldsymbol{v}_{jk}\}_{k=1,j=1}^{g_j,r-1}$ is linearly independent and assume

$$\sum_{j=1}^{r}\sum_{k=1}^{g_j} a_{jk}\boldsymbol{v}_{jk} = \boldsymbol{0} \text{ for some scalars } a_{jk}. \tag{D.7}$$

We multiply this equation by $(\boldsymbol{A} - \mu_r\boldsymbol{I})$ and obtain by linearity

$$\boldsymbol{0} = \sum_{j=1}^{r}\sum_{k=1}^{g_j} a_{jk}(\boldsymbol{A} - \mu_r\boldsymbol{I})\boldsymbol{v}_{jk} = \sum_{j=1}^{r}\sum_{k=1}^{g_j} a_{jk}(\mu_j - \mu_r)\boldsymbol{v}_{jk} = \sum_{j=1}^{r-1}\sum_{k=1}^{g_j} a_{jk}(\mu_j - \mu_r)\boldsymbol{v}_{jk}.$$

By the induction hypothesis all these a_{jk} vanish and in (D.7) we are left with $\sum_{k=1}^{g_r} a_{rk}\boldsymbol{v}_{rk} = \boldsymbol{0}$. Since these $\boldsymbol{v}'s$ form a basis for $\ker(\boldsymbol{A} - \mu_r\boldsymbol{I})$ we also have $a_{rk} = 0$ for $k = 1, \ldots, g_r$. (This also proves the induction hypothesis for $r = 1$.) Thus $\{\boldsymbol{v}_{jk}\}_{k=1,j=1}^{g_j,r}$ is linearly independent and it follows that the number of linearly independent eigenvectors is equal to $\sum_j g_j$. Since $g_j \leq a_j$ for all j and $\sum_j a_j = n$ we have $\sum_j g_j = n$ if and only if $a_j = g_j$ for $j = 1, \ldots, r$. □

D.4 Left Eigenvectors

Definition D.28 *A nonzero vector $\boldsymbol{y} \in \mathbb{C}^n$ corresponding to an eigenvalue λ of $\boldsymbol{A}$ is called a* **left eigenvector** *of $\boldsymbol{A}$ if $\boldsymbol{y}^*\boldsymbol{A} = \lambda\boldsymbol{y}^*$. We say that $(\lambda, \boldsymbol{y})$ is a* **left eigenpair** *of $\boldsymbol{A}$.*

Note that $\boldsymbol{y}^*\boldsymbol{A} = \lambda\boldsymbol{y}^*$ if and only if $\boldsymbol{A}^*\boldsymbol{y} = \overline{\lambda}\boldsymbol{y}$. It follows from Theorem D.3 that if $\boldsymbol{y}^*\boldsymbol{A} = \lambda\boldsymbol{y}^*$ then λ must be an eigenvalue for $\boldsymbol{A}$, while a left eigenvector $\boldsymbol{y}$ is an eigenvector for $\boldsymbol{A}^*$. If we need to make a distinction then an ordinary eigenvector, eigenpair is called a **right eigenvector** and **right eigenpair**, respectively.

Left- and right eigenvectors corresponding to distinct eigenvalues are orthogonal.

Theorem D.29 *Suppose $(\mu, \boldsymbol{y})$ and $(\lambda, \boldsymbol{x})$ are left and right eigenpairs of $\boldsymbol{A} \in \mathbb{C}^{n,n}$. If $\lambda \neq \mu$ then $\boldsymbol{y}^*\boldsymbol{x} = 0$.*

Proof. Using the eigenpair relation in two ways we obtain $\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{y}^*\boldsymbol{x} = \mu\boldsymbol{y}^*\boldsymbol{x}$ and we conclude that $\boldsymbol{y}^*\boldsymbol{x} = 0$. □

The case where $\lambda = \mu$ is more complicated. For example, the matrix $\boldsymbol{A} := \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ has one eigenvalue $\lambda = 1$ of algebraic multiplicity two, one right eigenvector $\boldsymbol{x} = \boldsymbol{e}_1$ and one left eigenvector $\boldsymbol{y} = \boldsymbol{e}_2$. Thus $\boldsymbol{y}^*\boldsymbol{x} = 0$. Two sufficient conditions guaranteeing that $\boldsymbol{y}^*\boldsymbol{x} \neq 0$ are given in the following theorem.

Theorem D.30 *Suppose $\boldsymbol{y}$ and $\boldsymbol{x}$ are left- and right eigenvectors corresponding to the same eigenvalue λ of $\boldsymbol{A} \in \mathbb{C}^{n,n}$. Then $\boldsymbol{y}^*\boldsymbol{x} \neq 0$ in the following two cases:*

1. *$\boldsymbol{A}$ can be diagonalized.*
2. *The algebraic multiplicity of λ is equal to one.*

Proof.

1. Suppose $\boldsymbol{Y}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\boldsymbol{Y} = \boldsymbol{X}^{-1}$. Partition $\boldsymbol{Y}$ by rows and $\boldsymbol{X}$ by columns as

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^* \\ \vdots \\ \boldsymbol{y}_n^* \end{bmatrix}, \quad \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n].$$

Since $\boldsymbol{YA} = \boldsymbol{DY}$ and $\boldsymbol{AX} = \boldsymbol{XD}$, we see that $\boldsymbol{y}_i$ is a left eigenvector and $\boldsymbol{x}_i$ is a right eigenvector corresponding to λ_i for $i = 1, \ldots, n$. But since $\boldsymbol{YX} = I$ we have $\boldsymbol{y}_i^* \boldsymbol{x}_i = 1$ for all i.

2. Assume that $\|\boldsymbol{x}\|_2 = 1$. We have (cf. (6.1))

$$\boldsymbol{V}^* \boldsymbol{A} \boldsymbol{V} = \left[\begin{array}{c|c} \lambda & \boldsymbol{z}^* \\ \hline \boldsymbol{0} & \boldsymbol{M} \end{array}\right],$$

where $\boldsymbol{V}$ is unitary and $\boldsymbol{V}\boldsymbol{e}_1 = \boldsymbol{x}$. Let $\boldsymbol{u} := \boldsymbol{V}^* \boldsymbol{y}$. Then

$$(\boldsymbol{V}^* \boldsymbol{A}^* \boldsymbol{V})\boldsymbol{u} = \boldsymbol{V}^* \boldsymbol{A}^* \boldsymbol{y} = \overline{\lambda} \boldsymbol{V}^* \boldsymbol{y} = \overline{\lambda} \boldsymbol{u},$$

so $(\overline{\lambda}, \boldsymbol{u})$ is an eigenpair of $\boldsymbol{V}^* \boldsymbol{A}^* \boldsymbol{V}$. But then $\boldsymbol{y}^* \boldsymbol{x} = \boldsymbol{u}^* \boldsymbol{V}^* \boldsymbol{V} \boldsymbol{e}_1$. Suppose that $\boldsymbol{u}^* \boldsymbol{e}_1 = 0$, i. e., $\boldsymbol{u} = \left[\begin{smallmatrix} 0 \\ \boldsymbol{v} \end{smallmatrix}\right]$ for some nonzero $\boldsymbol{v} \in \mathbb{C}^{n-1}$. Then

$$\boldsymbol{V}^* \boldsymbol{A}^* \boldsymbol{V} \boldsymbol{u} = \left[\begin{array}{c|c} \overline{\lambda} & \boldsymbol{0}^* \\ \hline \boldsymbol{z} & \boldsymbol{M}^* \end{array}\right] \begin{bmatrix} 0 \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{M}^* \boldsymbol{v} \end{bmatrix} = \overline{\lambda} \begin{bmatrix} 0 \\ \boldsymbol{v} \end{bmatrix}$$

and by Theorem D.3 it follows that λ is an eigenvalue of $\boldsymbol{M}$. But this is impossible since λ has algebraic multiplicity one and the eigenvalues of $\boldsymbol{A}$ are the union of λ and the eigenvalues of $\boldsymbol{M}$.

□

Corollary D.31 *If $\boldsymbol{A} \in \mathbb{C}^{n,n}$ has linearly independent right eigenvectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ then $\boldsymbol{A}$, also has linearly independent left eigenvectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. For any $\boldsymbol{v} \in \mathbb{C}^n$ we have*

$$\boldsymbol{v} = \sum_{j=1}^{n} (\boldsymbol{y}_j^* \boldsymbol{v}) \boldsymbol{x}_j = \sum_{k=1}^{n} (\boldsymbol{x}_k^* \boldsymbol{v}) \boldsymbol{y}_k. \tag{D.8}$$

Proof. From the proof of the previous theorem we have $\boldsymbol{y}_k^* \boldsymbol{x}_j = \delta_{kj}$ for all j, k. So if $\boldsymbol{v} = \sum_{j=1}^{n} c_j \boldsymbol{x}_j$, then $\boldsymbol{y}_k^* \boldsymbol{v} = \sum_{j=1}^{n} c_j \boldsymbol{y}_k^* \boldsymbol{x}_j = c_k$ for $k = 1, \ldots, n$. The proof of the second formula is similar. □

Appendix E

Gaussian Elimination

Gaussian elimination is the classical method for solving n linear equations in n unknowns. In component form the system is

$$\begin{array}{cccc} a_{11}x_1 + a_{12}x_2 + & \cdots & + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + & \cdots & + a_{2n}x_n = b_2, \\ \vdots \quad\quad \vdots & & \vdots \quad\quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + & \cdots & + a_{nn}x_n = b_n, \end{array}$$

and in matrix form

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \boldsymbol{b}.$$

We recall (see Definition B.6 and Theorem B.7) that the square system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for all right hand sides $\boldsymbol{b}$ if and only if $\boldsymbol{A}$ is nonsingular, i. e., the homogeneous system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ only has the solution $\boldsymbol{x} = \boldsymbol{0}$. We recall (cf. Theorem B.9) that a square matrix is invertible if and only if $\boldsymbol{A}$ is nonsingular, and the solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ can be written $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$, where $\boldsymbol{A}^{-1}$ is the inverse of $\boldsymbol{A}$. However, for large systems it is inefficient to compute $\boldsymbol{x}$ in this way. For an example see (2.10) and the discussion about the matrix $\boldsymbol{T}$ there. We also note (Lemma B.8) that if $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are square matrices, then $\boldsymbol{A}$ is nonsingular if and only if both $\boldsymbol{B}$ and $\boldsymbol{C}$ are nonsingular and in that case $\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}$.

The entries of $\boldsymbol{A}$ and $\boldsymbol{b}$ can be either real or complex numbers. For simplicity and ease of exposition we assume real entries.

In Gaussian elimination with no row interchanges we compute a triangular factorization of the coefficient matrix $\boldsymbol{A}$. This factorization is known as an LU factorization[3] of $\boldsymbol{A}$. In this chapter we discuss some theoretical and algorithmic

[3]We normally denote an upper triangular matrix by $\boldsymbol{R}$, but we respect common practice and most often we refer to the factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$ as an LU factorization of $\boldsymbol{A}$

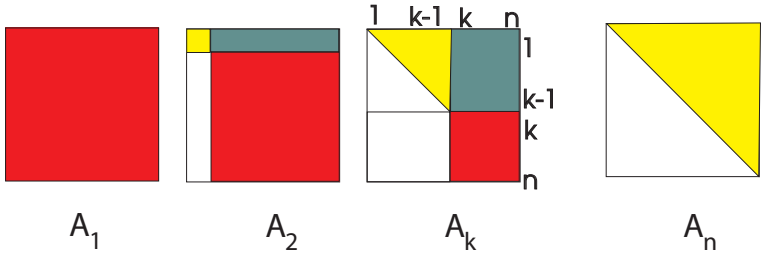


Figure E.1. *Gaussian elimination*

aspects of Gaussian elimination We consider also Gaussian elimination with row interchanges.

E.1 Gaussian Elimination and LU factorization

In **Gaussian elimination** without row interchanges we start with a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and generate a sequence of equivalent systems $\boldsymbol{A}^{(k)}\boldsymbol{x} = \boldsymbol{b}^{(k)}$ for $k = 1, \ldots, n$, where $\boldsymbol{A}^{(1)} = \boldsymbol{A}$, $\boldsymbol{b}^{(1)} = \boldsymbol{b}$, and $\boldsymbol{A}^{(k)}$ has zeros under the diagonal in its first $k-1$ columns. Thus $\boldsymbol{A}^{(n)}$ is upper triangular and the system $\boldsymbol{A}^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$ is easy to solve. The process is illustrated in Figure E.1.

The matrix $\boldsymbol{A}^{(k)}$ takes the form

$$\boldsymbol{A}^{(k)} = \left[\begin{array}{ccc|ccccc} a^1_{1,1} & \cdots & a^1_{1,k-1} & a^1_{1,k} & \cdots & a^1_{1,j} & \cdots & a^1_{1,n} \\ & \ddots & \vdots & \vdots & & \vdots & & \vdots \\ & & a^{k-1}_{k-1,k-1} & a^{k-1}_{k-1,k} & \cdots & a^{k-1}_{k-1,j} & \cdots & a^{k-1}_{k-1,n} \\ \hline & & & a^k_{k,k} & \cdots & a^k_{k,j} & \cdots & a^k_{k,n} \\ & & & \vdots & & \vdots & & \vdots \\ & & & a^k_{i,k} & \cdots & a^k_{i,j} & \cdots & a^k_{i,n} \\ & & & \vdots & & \vdots & & \vdots \\ & & & a^k_{n,k} & \cdots & a^k_{n,j} & \cdots & a^k_{n,n} \end{array}\right]. \tag{E.1}$$

The process transforming $\boldsymbol{A}^{(k)}$ into $\boldsymbol{A}^{(k+1)}$ for $k = 1, \ldots, n-1$ can be described as follows.

$$\boxed{\begin{array}{l} \text{for } i = k+1 : n \\ \quad l^k_{ik} = a^k_{ik}/a^k_{kk} \\ \quad \text{for } j = k : n \\ \qquad a^{k+1}_{ij} = a^k_{ij} - l^k_{ik} a^k_{kj} \end{array}} \tag{E.2}$$

For $j = k$ it follows from (E.2) that $a^{k+1}_{ik} = a^k_{ik} - \frac{a^k_{ik}}{a^k_{kk}} a^k_{kk} = 0$ for $i = k+1, \ldots, n$. Thus $\boldsymbol{A}^{(k+1)}$ will have zeros under the diagonal in its first k columns and

the elimination is carried one step further. The numbers l_{ik}^k in (E.2) are called **multipliers**.

Alternatively, we can describe the transformation $\boldsymbol{A}^{(k)} \rightarrow \boldsymbol{A}^{(k+1)}$ as a multiplication of $\boldsymbol{A}^{(k)}$ by a matrix known as an **elementary lower triangular matrix** .

Definition E.1 *For $1 \leq k \leq n-1$ and $\boldsymbol{l}_k = [l_{k+1,k}, \ldots, l_{n,k}]^T \in \mathbb{R}^{n-k}$ we define the matrix $\boldsymbol{M}_k \in \mathbb{R}^{n,n}$ by*

$$\boldsymbol{M}_k := \boldsymbol{I} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{l}_k \end{bmatrix} \boldsymbol{e}_k^T = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -l_{k+1,k} & 1 & \cdots & 0 \\ \vdots & & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -l_{n,k} & 0 & \cdots & 1 \end{bmatrix}, \tag{E.3}$$

where $\boldsymbol{0}$ is the zero vector in $\mathbb{R}^k$. We call $\boldsymbol{M}_k$ an elementary lower triangular matrix.

We have

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{M}_k \boldsymbol{A}^{(k)}, \text{ for } k = 1, \ldots, n-1, \tag{E.4}$$

where $\boldsymbol{M}_k \in \mathbb{R}^{n,n}$ is an elementary lower triangular matrix of the form (E.3) with $l_{ik} = l_{ik}^k$ given by (E.2) for $i = k+1, \ldots, n$.

Exercise E.2 *Show* (E.4).

Gaussian elimination with no row interchanges is valid if and only if the **pivots** a_{kk}^k are nonzero for $k = 1, \ldots, n-1$.

Theorem E.3 *We have $a_{k,k}^k \neq 0$ for $k = 1, \ldots, n-1$ if and only if the leading principal submatrices*

$$\boldsymbol{A}_k = \begin{bmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{bmatrix}$$

of $\boldsymbol{A}$ are nonsingular for $k = 1, \ldots, n-1$.

Proof. Let $\boldsymbol{B}_k = \boldsymbol{A}_{k-1}^{(k)}$ be the upper left $k-1$ corner of $\boldsymbol{A}^{(k)}$ given by (E.1). Observe that the entries of the matrix $\boldsymbol{B}_k$ is computed from $\boldsymbol{A}$ by using only entries from $\boldsymbol{A}_{k-1}$ and that only row-operations preserving non-singularity are used. It follows that $\boldsymbol{A}_{k-1}$ is nonsingular if and only if $\boldsymbol{B}_k$ is nonsingular. By Lemma 2.8 $\boldsymbol{B}_k$ is nonsingular if and only if $a_{ii}^{(i)} \neq 0$, $i = 1, \ldots, k-1$. We conclude that $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n-1}$ are nonsingular if and only if $\boldsymbol{B}_2, \ldots, \boldsymbol{B}_n$ are nonsingular which is equivalent to $a_{kk}^{(k)} \neq 0$ for $k = 1, \ldots, n-1$. ∎

Gaussian elimination is a way to compute the LU factorization of the coefficient matrix.

Theorem E.4 *Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ and that $\boldsymbol{A}_k$ is nonsingular for $k = 1, \ldots, n-1$. Then Gaussian elimination with no row interchanges results in an LU factorization of $\boldsymbol{A} \in \mathbb{R}^{n,n}$. In particular$\boldsymbol{A} = \boldsymbol{LR}$, where*

$$\boldsymbol{L} = \begin{bmatrix} 1 & & & \\ l_{21}^1 & 1 & & \\ \vdots & & \ddots & \\ l_{n1}^1 & l_{n2}^2 & \cdots & 1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} a_{11}^1 & \cdots & a_{1n}^1 \\ & \ddots & \vdots \\ & & a_{nn}^n \end{bmatrix}, \tag{E.5}$$

where the l_{ij}^j and a_{ij}^i are given by (E.2).

Proof. From (E.2) we have for all i, j

$$l_{ik}a_{kj}^k = a_{ij}^k - a_{ij}^{k+1} \text{ for } k < \min(i,j), \text{ and } l_{ij}a_{jj}^j = a_{ij}^j \text{ for } i > j.$$

Thus for $i \leq j$ we find

$$(\boldsymbol{LR})_{ij} = \sum_{k=1}^{n} l_{ik}u_{kj} = \sum_{k=1}^{i-1} l_{ik}a_{kj}^k + a_{ij}^i = \sum_{k=1}^{i-1} \left(a_{ij}^k - a_{ij}^{k+1}\right) + a_{ij}^i = a_{ij}^1 = a_{ij},$$

while for $i > j$

$$(\boldsymbol{LR})_{ij} = \sum_{k=1}^{n} l_{ik}u_{kj} = \sum_{k=1}^{j-1} l_{ik}a_{kj}^k + l_{ij}a_{jj}^j = \sum_{k=1}^{j-1} \left(a_{ij}^k - a_{ij}^{k+1}\right) + a_{ij}^j = a_{ij}.$$

□

Note that this Theorem holds even if $\boldsymbol{A}$ is singular. Since $\boldsymbol{L}$ is nonsingular the matrix $\boldsymbol{R}$ is singular, and we must have $a_{nn}^n = 0$ when $\boldsymbol{A}$ is singular.

E.1.1 Algoritms

Consider next an algorithm to find the LU factorization of $\boldsymbol{A}$ using Gaussian elimination with no row interchanges. Storing both the entries l_{ij}^j and a_{ij}^i in A we can write (E.2) as follows for $k = 1, \ldots, n-1$.

$$\boxed{\begin{array}{l} \text{for } i = k+1 : n \\ \quad a_{ik} = a_{ik}/a_{kk} \\ \quad \text{for } j = k+1 : n \\ \quad\quad a_{ij} = a_{ij} - a_{ik}a_{kj} \end{array}} \tag{E.6}$$

We can write (E.6) using outer product notation. We have

$$\begin{bmatrix} a_{k+1,k+1} & \cdots & a_{k+1,n} \\ \vdots & & \vdots \\ a_{n,k+1} & \cdots & a_{n,n} \end{bmatrix} = \begin{bmatrix} a_{k+1,k+1} & \cdots & a_{k+1,n} \\ \vdots & & \vdots \\ a_{n,k+1} & \cdots & a_{n,n} \end{bmatrix} - \begin{bmatrix} a_{k+1,k} \\ \vdots \\ a_{n,k} \end{bmatrix} \begin{bmatrix} a_{k,k+1} \cdots a_{k,n} \end{bmatrix}.$$

The result is a matrix of order $n-k$.

This leads to the following algorithm.

Algorithm E.5 (lufactor) Given $\boldsymbol{A} \in \mathbb{R}^{n,n}$ with $\boldsymbol{A}_k \in \mathbb{R}^{k,k}$ nonsingular for $k = 1, \ldots, n-1$. This algorithm computes an LU factorization of $\boldsymbol{A}$ using Gaussian elimination without row interchanges.

```
function [L,R]=lufactor(A)
n=length(A); for k=1:n-1
    kn=k+1:n;
    A(kn,k)=A(kn,k)/A(k,k);
    A(kn,kn)=A(kn,kn)-A(kn,k)*A(k,kn);
end
L=eye(n,n)+tril(A,-1);
R=triu(A);
```

Once we have an LU factorization of $\boldsymbol{A}$ the system $\boldsymbol{Ax} = \boldsymbol{b}$ is solved easily in two steps. Since $\boldsymbol{LRx} = \boldsymbol{b}$ we have $\boldsymbol{Ly} = \boldsymbol{b}$, where $\boldsymbol{y} := \boldsymbol{Rx}$. We first solve $\boldsymbol{Ly} = \boldsymbol{b}$ for $\boldsymbol{y}$ and then $\boldsymbol{Rx} = \boldsymbol{y}$ for $\boldsymbol{x}$. Consider solving a system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is lower triangular with nonzero diagonal entries. For $n = 3$ we have

$$\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

From the first equation we find $x_1 = b_1/a_{11}$. Solving the second equation for x_2 we obtain $x_2 = (b_2 - a_{21}x_1)/a_{11}$. Finally the third equation gives $x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33}$. This process is known as forward substitution and we arrive at the following algorithm.

Algorithm E.6 (forwardsolve) Given a nonsingular lower triangular matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ and $\boldsymbol{b} \in \mathbb{R}^n$. An $\boldsymbol{x} \in \mathbb{R}^n$ is computed so that $\boldsymbol{Ax} = \boldsymbol{b}$.

```
function x=forwardsolve(A,b)
n=length(b); x=b(:);
for k=1:n
    x(k)=(x(k)-A(k,1:k-1)*x(1:k-1))/A(k,k);
end
```

A system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is upper triangular must be solved 'bottom-up'. We first find x_n from the last equation and then move upwards for the remaining unknowns. We have the following algorithm.

Algorithm E.7 (backsolve) Given a nonsingular upper triangular matrix $\boldsymbol{A} \in \mathbb{R}^{n,n}$ and $\boldsymbol{b} \in \mathbb{R}^n$. An $\boldsymbol{x} \in \mathbb{R}^n$ is computed so that $\boldsymbol{Ax} = \boldsymbol{b}$.

```
function x=backsolve(A,b)
n=length(b); x=b(:);
for k=n:-1:1
    x(k)=(x(k)-A(k,k+1:n)*x(k+1:n))/A(k,k);
end
```

E.1.2 Operation count

We define a **flop** (floating point operation) as one of the floating point arithmetic operations, ie. multiplication, division, addition and subtraction. We denote by **nflops** the total number of flops in an algorithm, i.e. the the sum of all multiplications, divisions, additions and subtractions. For a problem of size n the number nflops will often be a polynomial in n. For example, we will show below that an LU factorization requires $\frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{n}{6}$ flops. For large values of n the highest term $\frac{2}{3}n^3$ dominates and we usually say that nflops $= O(\frac{2}{3}n^3)$ ignoring lower order terms. We sometimes say that nflops $= O(n^3)$ if we do not bother with the constant (in this case $2/3$) in front of the n^3 term.

In many implementations the computing time T_A for an algorithm A applied to a large problem is proportional to $N_A :=$ nflops. If this is true then we typically have $T_A = \alpha N_A$, where α is in the range 10^{-12} to 10^{-9} on a modern computer.

Consider now $N_{LU} :=$ nflops for LU factorization. Let M, D, A, S be the number of multiplications, divisions, additions, and subtractions. We first do an exact count. From (E.6) we find

- $M = \sum_{k=1}^{n-1}(n-k)^2 = \sum_{m=1}^{n-1} m^2 = \frac{1}{3}n(n-1)(n-\frac{1}{2})$
- $D = \sum_{m=1}^{n-1} m = \frac{1}{2}n(n-1), \quad S = M, \quad A = 0.$

Thus

$$N_{LU} = M + D + A + S = \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n = O(\frac{2}{3}n^3)$$

There is a quick way to arrive at the leading term $2n^3/3$. We only consider the arithmetic operations contributing to the leading term. Then we replace sums by integrals letting the summation indices be continuous variables and adjust limits of integration in an insightful way to simplify the calculation. In the Gaussian elimination case the contribution to the leading term only comes from M and S and we find

$$M + S = 2\sum_{k=1}^{n-1}(n-k)^2 \approx 2\int_1^{n-1}(n-k)^2 dk \approx 2\int_0^{n}(n-k)^2 dk = \frac{2}{3}n^3.$$

This is the correct leading term and we obtain $N_{LU} = O(2n^3/3)$ which is reasonably correct for large values of n.

Consider next forward and backward substitution. Counting flops and letting $N_S := N_F + N_B$ we find

$$N_S \approx \int_1^n 2(k-1)dk + \int_1^n 2(n-k)dk \approx \int_0^n 2kdk + \int_0^n 2(n-k)dk = 2n^2.$$

Comparing N_{LU} and N_S we see that LU factorization is an $O(n^3)$ process while the solution stage only require $O(n^2)$ flops. This leads to dramatic differences in computing time as illustrated in the following table:

n	T_{LU}	T_S
10^3	1s	0.003s
10^4	17min.	0.3s
10^6	32 years	51min

Here we have assumed that the computing time for the LU factorization is $T_{LU} = 10^{-9}n^3$ and the computing time for the forward and bacwards substittution is $T_S = 3 \times 10^{-9}n^2$ corresponding to $\alpha = 3 \times 10^{-9}/2$.

To further illustrate the difference between n^3 and n^2 for large n suppose we want to solve m systems $\boldsymbol{A}_j\boldsymbol{x}_j = \boldsymbol{b}_j$ for $j = 1, \ldots, m$, where $\boldsymbol{A}_j \in \mathbb{R}^{n,n}$ and $\boldsymbol{b}_j \in \mathbb{R}^n$. We need $m(\frac{2}{3}n^3 + 2n^2)$ flops for this. Thus if $n = 10^4$ and $m = 100$ the table gives a computing time of approximately 1700min. Suppose now $\boldsymbol{A}_j = \boldsymbol{A}$, i.e. we have the same coefficient matrix in all systems. We can then write the m systems more compactly as $\boldsymbol{AX} = \boldsymbol{B}$, where $\boldsymbol{A} \in \mathbb{R}^{n,n}$, $\boldsymbol{B} \in \mathbb{R}^{n,m}$ and the matrix $\boldsymbol{X} \in \mathbb{R}^{n,m}$ is the unknown. To solve $\boldsymbol{AX} = \boldsymbol{B}$ we first compute the LU factorization of $\boldsymbol{A}$ and then apply forward and backward substitution to the columns of $\boldsymbol{B}$. If $n = 10^4$ the computing time for this would be 17min for the LU factorization and 30s for the solution phase.

E.2 Pivoting

We have seen that Gaussian elimination without row interchanges is only well defined if the leading principal minors $\boldsymbol{A}_k \in \mathbb{R}^{k,k}$ are nonsingular for $k = 1, \ldots, n-1$.

Suppose now $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is nonsingular. We can still solve a linear system with $\boldsymbol{A}$ if we incorporate row interchanges. Interchanging two rows (and/or two columns) during Gaussian elimination is known as **pivoting**. The entry which is moved to the diagonal position (k, k) is called the **pivot entry** or **pivot** for short. Gaussian elimination with row interchanges can be described as follows.

1. **Choose** $r_k \geq k$ so that $a^k_{r_k,k} \neq 0$.
2. **Interchange** rows r_k and k of $\boldsymbol{A}^{(k)}$.
3. **Eliminate** by computing l^k_{ik} and a^{k+1}_{ij} using (E.2).

We have seen that the elimination step can be described as multiplying the current matrix by an elementary transformation matrix $\boldsymbol{M}_k$ given by (E.3). But before we can multiply by $\boldsymbol{M}_k$ we have to interchange rows. This can be described in terms of permutation matrices.

E.2.1 Permutation matrices

Definition E.8 *A* **permutation matrix** *is a matrix of the form*

$$P = \boldsymbol{I}(:, \boldsymbol{p}) = [\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_n}] \in \mathbb{R}^{n,n},$$

where $\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_n}$ *is a permutation of the unit vectors* $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n \in \mathbb{R}^n$.

Every permutation $\boldsymbol{p} = [i_1, \ldots, i_n]^T$ of the integers $1, 2, \ldots, n$ gives rise to a permutation matrix and vice versa. Post-multiplying a matrix $\boldsymbol{A}$ by a permutation matrix results in a permutation of the columns, while pre-multiplying by a permutation matrix gives a permutation of the rows. In symbols

$$\boldsymbol{AP} = \boldsymbol{A}(:, \boldsymbol{p}), \quad \boldsymbol{P}^T\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{p}, :). \tag{E.7}$$

Indeed, $\boldsymbol{AP} = (\boldsymbol{Ae}_{i_1}, \ldots, \boldsymbol{Ae}_{i_n}) = \boldsymbol{A}(:, \boldsymbol{p})$ and $\boldsymbol{P}^T\boldsymbol{A} = (\boldsymbol{A}^T\boldsymbol{P})^T = (\boldsymbol{A}^T(:, \boldsymbol{p}))^T = \boldsymbol{A}(\boldsymbol{p}, :)$.

Since $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$ the inverse of $\boldsymbol{P}$ is equal to its transpose, $\boldsymbol{P}^{-1} = \boldsymbol{P}^T$ and $\boldsymbol{PP}^T = \boldsymbol{I}$ as well. Thus a permutation matrix is an orthonormal matrix.

We will use a particularly simple permutation matrix.

Definition E.9 *We define a* **(j,k)-Interchange Matrix** $\boldsymbol{I}_{jk}$ *by interchanging column* j *and* k *of the identity matrix.*

Since $\boldsymbol{I}_{jk} = \boldsymbol{I}_{kj}$, and we obtain the identity by applying $\boldsymbol{I}_{jk}$ twice, we see that $\boldsymbol{I}_{jk}^2 = \boldsymbol{I}$ and an interchange matrix is symmetric and equal to its own inverse. Pre-multiplying a matrix by an interchange matrix interchanges two rows of the matrix, while post-multiplication interchanges two columns.

E.2.2 Gaussian elimination works mathematically

The process going from $\boldsymbol{A}^{(k)}$ to $\boldsymbol{A}^{(k+1)}$ can be written

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{M}_k\boldsymbol{P}_k\boldsymbol{A}^{(k)}, \text{ for } k = 1, \ldots, n-1, \tag{E.8}$$

where $\boldsymbol{P}_k = I_{r_k,k} \in \mathbb{R}^{n,n}$ is a permutation matrix interchanging rows k and r_k of $\boldsymbol{A}^{(k)}$ and $\boldsymbol{M}_k \in \mathbb{R}^{n,n}$ is an elementary lower triangular matrix of the form (E.3) with $l_{ik} = l_{ik}^k$ given by (E.2) for $i = k+1, \ldots, n$.

If $\boldsymbol{A}$ is nonsingular then Gaussian elimination can always be carried to completion by using suitable row interchanges. To show this, suppose by induction on k that $\boldsymbol{A}^{(k)}$ is nonsingular. Since $\boldsymbol{A}^{(1)} = \boldsymbol{A}$ this holds for $k = 1$. By Lemma 2.7 the lower right diagonal block in $\boldsymbol{A}^{(k)}$ is nonsingular. But then at least one entry in the first column of that block must be nonzero and it follows that r_k exists so that $a_{r_k,k}^k \neq 0$. But then the matrices $\boldsymbol{P}_k$ and $\boldsymbol{M}_k$ in (E.8) are well defined. By Lemma 2.8 the matrix $\boldsymbol{M}_k$ is nonsingular and since a permutation matrix is nonsingular it follows from Lemma B.8 that $\boldsymbol{A}^{(k+1)}$ is nonsingular. We conclude that $\boldsymbol{A}^{(k)}$ is nonsingular for $k = 1, \ldots, n$.

E.2.3 Pivot strategies

Rp to now we have said nothing about what rows in $\boldsymbol{A}$ to interchange during the elimination. We start with an example illustrating that small pivots should be avoided.

Example E.10 *Applying Gaussian elimination without row interchanges to the linear system*

$$\begin{aligned} 10^{-4}x_1 + 2x_2 &= 4 \\ x_1 + x_2 &= 3 \end{aligned}$$

we obtain the upper triangular system

$$\begin{aligned} 10^{-4}x_1 + 2x_2 &= 4 \\ (1 - 2 \times 10^4)x_2 &= 3 - 4 \times 10^4 \end{aligned}$$

The exact solution is

$$x_2 = \frac{-39997}{-19999} \approx 2, \quad x_1 = \frac{4 - 2x_2}{10^{-4}} = \frac{20000}{19999} \approx 1.$$

Suppose we round the result of each arithmetic operation to three digits. The solutions $\mathrm{fl}(x_1)$ *and* $\mathrm{fl}(x_2)$ *computed in this way is*

$$\mathrm{fl}(x_2) = 2, \quad \mathrm{fl}(x_1) = 0.$$

The computed value 0 of x_1 *is completely wrong. Suppose instead we apply Gaussian elimination to the same system, but where we have interchanged the equations. The system is*

$$\begin{aligned} x_1 + x_2 &= 3 \\ 10^{-4}x_1 + 2x_2 &= 4 \end{aligned}$$

and we obtain the upper triangular system

$$\begin{aligned} x_1 + x_2 &= 3 \\ (2 - 10^{-4})x_2 &= 4 - 3 \times 10^{-4} \end{aligned}$$

Now the solution is computed as follows

$$x_2 = \frac{3.9997}{1.9999} \approx 2, \quad x_1 = 3 - x_2 \approx 1.$$

In this case rounding each calculation to three digits produces $\mathrm{fl}(x_1) = 1$ *and* $\mathrm{fl}(x_2) = 2$ *which is quite satisfactory since it is the exact solution rounded to three digits.*

We briefly describe the two most common pivoting strategies. The choice

$$a^k_{r_k,k} := \max\{|a^k_{i,k}| : k \leq i \leq n\}$$

with r_k the smallest such index in case of a tie, is known as **partial pivoting**. It is possible to interchange both rows and columns. The choice

$$a^k_{r_k,s_k} := \max\{|a^k_{i,j}| : k \le i,j \le n\}$$

with r_k, s_k the smallest such indices in case of a tie, is known as **complete pivoting**. Complete pivoting is known to be more stable, but requires a lot of search and is seldom used in practice.

E.3 The PLU-Factorization

Consider now Gaussian elimination with row pivoting. We can keep track of the row interchanges using **pivot vectors** $\boldsymbol{p}_k$. We define

$$\boldsymbol{p} := \boldsymbol{p}_n, \text{ where } \boldsymbol{p}_1 := [1,2,\ldots,n]^T, \text{ and } \boldsymbol{p}_{k+1} := \boldsymbol{I}_{r_k,k}\boldsymbol{p}_k \text{ for } k = 1,\ldots,n-1. \tag{E.9}$$

We obtain $\boldsymbol{p}_{k+1}$ from $\boldsymbol{p}_k$ by interchanging the entries r_k and k in $\boldsymbol{p}_k$. In particular the first $k-1$ components in $\boldsymbol{p}_k$ and $\boldsymbol{p}_{k+1}$ are the same.

There is a close relation between the pivot vectors $\boldsymbol{p}_k$ and the corresponding interchange matrices $\boldsymbol{P}_k := \boldsymbol{I}_{r_k,k}$. Since $\boldsymbol{P}_k\boldsymbol{I}(\boldsymbol{p}_k,:) = \boldsymbol{I}(\boldsymbol{P}_k\boldsymbol{p}_k,:) = \boldsymbol{I}(\boldsymbol{p}_{k+1},:)$ we obtain

$$\boldsymbol{P}^T := \boldsymbol{P}_{n-1}\cdots\boldsymbol{P}_1 = \boldsymbol{I}(\boldsymbol{p},:), \quad \boldsymbol{P} := \boldsymbol{P}_1\boldsymbol{P}_2\cdots\boldsymbol{P}_{n-1} = \boldsymbol{I}(:,\boldsymbol{p}). \tag{E.10}$$

Instead of interchanging the rows of $\boldsymbol{A}$ during elimination we can keep track of the ordering of the rows using the pivot vectors $\boldsymbol{p}_k$. The Gaussian elimination in Section E.1 with entries a^1_{ij} can be described as follows:

```
p = [1,...,n]^T;
for k = 1 : n − 1
    choose r_k ≥ k so that a^k_{p_{r_k},k} ≠ 0.
    p = I_{r_k,k} p
    for i = k + 1 : n
        a^k_{p_i,k} = a^k_{p_i,k} / a^k_{p_k,k}
        for j = k : n
            a^{k+1}_{p_i,j} = a^k_{p_i,j} − a^k_{p_i,k} a^k_{p_k,j}
```

(E.11)

This leads to the following factorization:

Theorem E.11 *Gaussian elimination with row pivoting on a nonsingular matrix* $\boldsymbol{A} \in \mathbb{R}^{n,n}$ *leads to a factorization* $\boldsymbol{A} = \boldsymbol{PLR}$*, where* $\boldsymbol{P}$ *is a permutation matrix,* $\boldsymbol{L}$ *is lower triangular with ones on the diagonal, and* $\boldsymbol{R}$ *is upper triangular. More*

explicitly, $\boldsymbol{P} = \boldsymbol{I}(:, \boldsymbol{p})$*, where* $\boldsymbol{p} = \boldsymbol{I}_{r_{n-1},n-1} \cdots \boldsymbol{I}_{r_1,1}[1, \ldots, n]^T$*, and*

$$\boldsymbol{L} = \begin{bmatrix} 1 & & & \\ a^1_{p_2,1} & 1 & & \\ \vdots & & \ddots & \\ a^1_{p_n,1} & a^2_{p_n,2} & \cdots & 1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} a^1_{p_1,1} & \cdots & a^1_{p_1,n} \\ & \ddots & \vdots \\ & & a^n_{p_n,n} \end{bmatrix}, \tag{E.12}$$

Proof. The proof is analogous to the proof for LU factorization without pivoting. From (E.11) we have for all i, j

$$a^k_{p_i,k} a^k_{p_k,j} = a^k_{p_i,j} - a^{k+1}_{p_i,j} \text{ for } k < \min(i,j), \text{ and } a^k_{p_i,j} a^j_{p_j,j} = a^j_{p_i,j} \text{ for } i > j.$$

Thus for $i \leq j$ we find

$$\begin{aligned} (\boldsymbol{LR})_{ij} &= \sum_{k=1}^{n} l_{i,k} u_{kj} = \sum_{k=1}^{i-1} a^k_{p_i,k} a^k_{p_k,j} + a^i_{p_i,j} \\ &= \sum_{k=1}^{i-1} \left(a^k_{p_i,j} - a^{k+1}_{p_i,j} \right) + a^i_{p_i,j} = a^1_{p_i,j} = a_{p_i,j} = \left(\boldsymbol{P}^T \boldsymbol{A} \right)_{ij}, \end{aligned}$$

while for $i > j$

$$\begin{aligned} (\boldsymbol{LR})_{ij} &= \sum_{k=1}^{n} l_{ik} u_{kj} = \sum_{k=1}^{j-1} a^k_{p_i,k} a^k_{p_k,j} + a^k_{p_i,j} a^j_{p_j,j} \\ &= \sum_{k=1}^{j-1} \left(a^k_{p_i,j} - a^{k+1}_{p_i,j} \right) + a^j_{p_i,j} = a^1_{p_i,j} = a_{p_i,j} = \left(\boldsymbol{P}^T \boldsymbol{A} \right)_{ij}. \end{aligned}$$

□

E.4 An Algorithm for Finding the PLU-Factorization

Using pivot vectors we can compute the PLU factorization of $\boldsymbol{A}$ without physically interchanging the entries a^k_{ij}. As is clear from (E.12) we can store the entries of $\boldsymbol{L}$ and $\boldsymbol{R}$ in $\boldsymbol{A}$ and work with $\boldsymbol{A}(\boldsymbol{p}_k, :)$. At the end the entries of $\boldsymbol{L}$ and $\boldsymbol{R}$ will be located under and above the diagonal.

In the following algorithm we use partial pivoting.

Algorithm E.12 (PLU factorization) Given a nonsingular $\boldsymbol{A} \in \mathbb{R}^{n,n}$. This algorithm computes a PLU factorization of $\boldsymbol{A}$ using Gaussian elimination with partial pivoting. The permutation matrix $\boldsymbol{P}$ can be recovered form the pivot vector $\boldsymbol{p}$ as $\boldsymbol{P} = \boldsymbol{I}(:, \boldsymbol{p})$.

```
function [p,L,R] = plufactor(A)
n = length(A);
p = 1:n;
for k=1:n-1
   [maxv,r] = max(abs(A(p(k:n),k)));
   p([k r+k-1]) = p([r+k-1 k]);
   ps=p(k+1:n);
   A(ps,k) = A(ps,k)/A(p(k),k);
      A(ps,k+1:n) = A(ps,k+1:n) - A(ps,k)*A(p(k),k+1:n);
end
L = eye(n,n) + tril(A(p,:),-1);
R = triu(A(p,:));
```

Once we have a PLU factorization of $\boldsymbol{A}$ the system $\boldsymbol{Ax} = \boldsymbol{b}$ is solved easily in three steps. Since $\boldsymbol{PLRx} = \boldsymbol{b}$ we have $\boldsymbol{Pz} = \boldsymbol{b}$, $\boldsymbol{Ly} = \boldsymbol{z}$, and $\boldsymbol{Rx} = \boldsymbol{y}$. Using the output [p,L,R] of Algorithm E.12 the solution can be found from Algorithms E.6 and E.7 in two steps.

1. y=forwardsolve(L,b(p));
2. x=backsolve(R,y);

Exercise E.13 In this exercise we develop column oriented vectorized versions of forward and backward substitution. Suppose $L \in \mathbb{R}^{n,n}$ is lower triangular and $R \in \mathbb{R}^{n,n}$ is upper triangular. Consider the system $L\boldsymbol{x} = \boldsymbol{b}$. Suppose after $k-1$ steps of the algorithm we have a reduced system in the form

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{k+1,k} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ l_{nk} & & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_k \\ b_{k+1} \\ \vdots \\ b_n \end{bmatrix}.$$

This system is of order $n-k+1$. The unknowns are $x_k, \ldots, x_n$.

a) We see that $x_k = b_k$ and eliminating x_k from the remaining equations show that we obtain a system of order $n-k$ with unknowns $x_{k+1}, \ldots, x_n$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{k+2,k+1} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ l_{n,k+1} & & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_{k+1} \\ \vdots \\ b_n \end{bmatrix} - x_k \begin{bmatrix} l_{k+1,k} \\ \vdots \\ l_{n,k} \end{bmatrix}.$$

Thus at the kth step, $k = 1, 2, \ldots n$ we set $x_k = b_k$ and update b as follows:

$$b(k+1:n) = b(k+1:n) - x(k) * L(k+1:n, k).$$

b) Suppose now $L \in \mathbb{R}^{n,n}$ is lower triangular, $R \in \mathbb{R}^{n,n}$ is upper triangular and $\boldsymbol{b} \in \mathbb{R}^n$. Justify the following column oriented vectorized algorithms for solving $Lx = b$ and $Rx = b$.

Algorithm E.14 (Forward Substitution (column oriented))

for $k = 1 : n$
  $x(k) = b(k)/L(k, k)$;
  $b(k{+}1{:}n) = b(k{+}1{:}n) - L(k{+}1{:}n, k) * x(k)$;
end

Algorithm E.15 (Backward Substitution (column oriented))

for $k = n : -1 : 1$
  $x(k) = b(k)/R(k, k)$;
  $b(1{:}k{-}1) = b(1{:}k{-}1) - R(1{:}k{-}1, k) * x(k)$;
end

Each algorithm requires n^2 flops.

Appendix F

Computer Arithmetic

F.1 Absolute and Relative Errors

Suppose a and b are real or complex scalars. If b is an approximation to a then there are different ways of measuring the error in b.

Definition F.1 (Absolute Error) *The absolute error in b as an approximation to a is the number $\epsilon := |a - b|$. The number $e := b - a$ is called the error in b as an approximation to a. This is what we have to add to a to get b.*

Note that the absolute error is symmetric in a and b, so that ϵ is also the absolute error in a as an approximation to b

Definition F.2 (Relative Error) *If $a \neq 0$ then the relative error in b as an approximation to a is defined by*

$$\rho = \rho_b := \frac{|b - a|}{|a|}.$$

We say that a and b agree to approximately $-\log_{10} \rho$ digits.

As an example, if $a := 31415.9265$ and $b := 31415.8951$, then $\rho = 0.999493 * 10^{-6}$ and a and b agree to approximately 6 digits.

We have $b = a(1 + r)$ for some r if and only if $\rho = |r|$.

We can also consider the relative error $\rho_a := |a-b|/|b|$ in a as an approximation to b.

Lemma F.3 *If $a, b \neq 0$ and $\rho_b < 1$ then $\rho_a \leq \rho_b/(1 - \rho_b)$.*

Proof. Since $|a|\rho_b = |b - a| \geq |a| - |b|$ we obtain $|b| \geq |a| - |a - b| = (1 - \rho_b)|a|$. Then

$$\rho_a = \frac{|b - a|}{|b|} \leq \frac{|b - a|}{(1 - \rho_b)|a|} = \frac{\rho_b}{1 - \rho_b}.$$

□

If ρ_b is small then ρ_a is small and it does not matter wether we choose ρ_a or ρ_b to discuss relative error.

Exercise F.4 *Compare ρ_a and ρ_b when $a := 3.1415.9265$ and $b := 31415.8951$.*

F.2 Floating Point Numbers

We shall assume that the reader is familiar with different number systems (binary, octal, decimal, hexadecimal) and how to convert from one number system to another. We use $(x)_\beta$ to indicate a number written to the base β. If no parenthesis and subscript are used, the base 10 is understood. For instance,

$$\begin{aligned}(100)_2 &= 4,\\ (0.1)_2 &= 0.5,\\ 0.1 = (0.1)_{10} &= (0.0001100110011001\ldots)_2.\end{aligned}$$

In general,

$$x = (c_m c_{m-1} \ldots c_0 . d_1 d_2 \ldots d_n)_\beta$$

means

$$x = \sum_{i=0}^{m} c_i \beta^i + \sum_{i=1}^{n} d_i \beta^{-i}, \quad 0 \le c_i, d_i \le \beta - 1.$$

We can move the decimal point by adding an exponent:

$$y = x \cdot \beta^e,$$

for example

$$(0.1)_{10} = (1.100110011001\ldots)_2 \cdot 2^{-4}.$$

We turn now to a description of the floating-point numbers. We will only describe a **standard system**, namely the binary IEEE floating-point standard. Although it is not used by all systems, it has been widely adopted and is used in Matlab. For a more complete introduction to the subject see [8],[17].

We denote the real numbers which are represented in our computer by $\mathcal{F}$. The set $\mathcal{F}$ are characterized by three integers t, and $\underline{e}, \overline{e}$. We define

$$\epsilon_M := 2^{-t}, \qquad \text{machine epsilon,} \tag{F.1}$$

and

$$\begin{aligned}&\mathcal{F} := \{0\} \cup \mathcal{S} \cup \mathcal{N}, \text{where}\\ &\mathcal{N} := \mathcal{N}_+ \cup \mathcal{N}_-, \quad \mathcal{N}_+ := \cup_{e=\underline{e}}^{\overline{e}} \mathcal{N}_e, \quad \mathcal{N}_- := -\mathcal{N}_+,\\ &\mathcal{N}_e := \left\{(1.d_1 d_2 \cdots d_t)_2\right\} * 2^e = \{1, 1+\epsilon_M, 1+2\epsilon_M, \ldots, 2-\epsilon_M\} * 2^e,\\ &\mathcal{S} := \mathcal{S}_+ \cup \mathcal{S}_-, \quad \mathcal{S}_+ := \{\epsilon_M, 2\epsilon_M, 3\epsilon_M, \ldots, 1-\epsilon_M\} * 2^{\underline{e}}, \quad \mathcal{S}_- := -\mathcal{S}_+.\end{aligned} \tag{F.2}$$

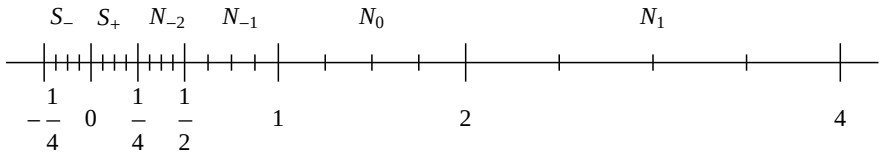


Figure F.1. *Distribution of some positive floating-point numbers*

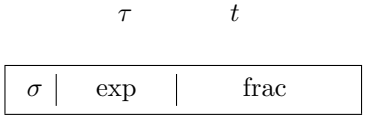
Example F.5 *Suppose* $t := 2$, $\overline{e} = 3$ *and* $\underline{e} := -2$. *Then* $\epsilon_M = 1/4$ *and we find*

$$\mathcal{N}_{-2} = \{\frac{1}{4}, \frac{5}{16}, \frac{3}{8}, \frac{7}{16}\}, \quad \mathcal{N}_{-1} = \{\frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}\}, \quad \mathcal{N}_0 = \{1, \frac{5}{4}, \frac{3}{2}, \frac{3}{4}, \frac{7}{4}\},$$
$$\mathcal{N}_1 = \{2, \frac{5}{2}, 3, \frac{7}{2}\}, \quad \mathcal{N}_2 = \{4, 5, 6, 7\}, \quad \mathcal{N}_3 = \{8, 10, 12, 14\},$$
$$\mathcal{S}_+ = \{\frac{1}{16}, \frac{1}{8}, \frac{3}{16}\}, \quad \mathcal{S}_- = \{-\frac{3}{16}, -\frac{1}{8}, -\frac{1}{16}\}.$$

The position of some of these sets on the real line is shown in Figure F.1

1. The elements of $\mathcal{N}$ are called **normalized (floating-point) numbers**. They consists of three parts, the sign +1 or -1, the **mantissa** $(1.d_1d_2\cdots d_t)_2$, and the **exponent part** 2^e.
2. the elements in $\mathcal{N}_+$ has the sign +1 indicated by the bit $\sigma = 0$ and the elements in $\mathcal{N}_-$ has the sign bit $\sigma = 1$. Thus the sign of a number is $(-1)^\sigma$. The standard system has two zeros $+0$ and -0.
3. The mantissa is a number between 1 and 2. It consists of $t + 1$ binary digits.
4. The number e in the exponent part is restricted to the range $\underline{e} \le e \le \overline{e}$.
5. The positive normalized numbers are located in the interval $[r_m, r_M]$, where
$$r_m := 2^{\underline{e}}, \quad r_M := (2 - \epsilon_M) * 2^{\overline{e}}. \tag{F.3}$$
6. The elements in $\mathcal{S}$ are called **subnormal** or **denormalized**. As for normalized numbers they consists of three parts, but the mantissa is less than one in size. The main use of subnormal numbers is to soften the effect of underflow. If a number is in the range $(0, (1 - \epsilon_M/2) * 2^{\underline{e}})$, then it is rounded to the nearest subnormal number or to zero.
7. Two additional symbols "Inf" and "NaN" are used for special purposes.
8. The symbol **Inf** is used to represent numbers outside the interval $[-r_M, r_M]$ **(overflow)**, and results of arithmetic operations of the form $x/0$, where $x \in \mathcal{N}$. Inf has a sign, +Inf and -Inf.
9. The symbol **NaN** stands for "not a number". a NaN results from illegal operations of the form $0/0$, $0 * \text{Inf}$, Inf/Inf, $\text{Inf} - \text{Inf}$ and so on.
10. The choices of t, $\overline{e}$, and $\underline{e}$ are to some extent determined by the architecture of the computer. A floating-point number, say x, occupies $n := 1 + \tau + t$ bits,

where 1 bit is used for the sign, τ bits for the exponent, and t bits for the fractional part of the mantissa.



Here $\sigma = 0$ if $x > 0$ and $\sigma = 1$ if $x < 0$, and $\text{exp} \in \{0, 1, 2, 3, \ldots, 2^\tau - 1\}$ is an integer. The integer frac is the fractional part $d_1 d_2 \cdots d_t$ of the mantissa. The value of a normalized number in the standard system is

$$x = (-1)^\sigma * (1.\text{frac})_2 * 2^{\text{exp}-b,} \text{ where } b := 2^{\tau-1} - 1. \tag{F.4}$$

The integer b is called the **bias**.

11. To explain the choice of b we note that the extreme values exp = 0 and $\text{exp} = 2^\tau - 1$ are used for special purposes. The value exp = 0 is used for the number zero and the subnormal numbers, while $\text{exp} = 2^\tau - 1$ is used for Inf and NaN. Since $2b = 2^\tau - 2$, the remaining numbers of exp, i. e., $\text{exp} \in \{1, 2, \ldots, 2^\tau - 2\}$ correspond to e in the set $\{1 - b, 2 - b, \ldots, b\}$. Thus in a standard system we have

$$\underline{e} = 1 - b, \quad \overline{e} = b := 2^{\tau-1} - 1. \tag{F.5}$$

12. The most common choices of τ and t are shown in the following table

precision	τ	t	b	$\epsilon_M = 2^{-t}$	$r_m = 2^{1-b}$	r_M
half	5	10	15	9.8×10^{-4}	6.1×10^{-5}	6.6×10^{4}
single	8	23	127	1.2×10^{-7}	1.2×10^{-38}	3.4×10^{38}
double	11	52	1023	2.2×10^{-16}	2.2×10^{-308}	1.8×10^{308}
quad	15	112	16383	1.9×10^{-34}	3.4×10^{-4932}	1.2×10^{4932}

Here b is given by (F.5) and r_M by (F.3) The various lines correspond to a normalized number occupying **half** a word of 32 bits, one word (**single precision**), two words (**double precision**), and 4 words (**quad precision**).

Exercise F.6 *Check the results of the following operations on your computer.* 1^{Inf}, 2^{Inf}, e^{-Inf}, Inf^0, $\log 0$, $\sin(Inf)$, $\arctan(-Inf)$.

F.3 Rounding and Arithmetic Operations

The standard system is a closed system. Every $x \in \mathbb{R}$ has a representation as either a floating-point number, or Inf or NaN, and every arithmetic operation produces a result. We denote the computer representation of a real number x by $\text{fl}(x)$.

F.3.1 Rounding

To represent a real number x there are three cases.

$$\mathrm{fl}(x) = \begin{cases} \text{Inf}, & \text{if } x > r_M, \\ -\text{Inf}, & \text{if } x < -r_M, \\ \text{round to zero}, & \text{otherwise.} \end{cases}$$

To represent a real number with $|x| \leq r_M$ the system chooses a machine number $\mathrm{fl}(x)$ closest to x. This is known as **rounding**. When x is midway between two numbers in $\mathcal{F}$ we can either choose the one of larger magnitude (**round away from zero**), or pick the one with a zero last bit (**round to zero**). The standard system uses round to zero. As an example, if $x = 1 + \epsilon_M/2$, then x is midway between 1 and $1 + \epsilon_M$. Therefore $\mathrm{fl}(x) = 1 + \epsilon_M$ if round away from zero is used, while $\mathrm{fl}(x) = 1$ if x is rounded to zero. This is because the machine representation of 1 has frac $= 0$.

The following lemma gives a bound for the relative error in rounding.

Theorem F.7 *If $r_m \leq |x| \leq r_M$ then*

$$\mathit{fl}(x) = x(1+\delta), \quad |\delta| \leq u_M := \frac{1}{2}\epsilon_M = 2^{-t-1}.$$

Proof. Suppose $2^e < x < 2^{e+1}$. Then $\mathrm{fl}(x) \in \{1, 1+\epsilon_M, 1+2\epsilon_M, \ldots, 2-\epsilon_M\} * 2^e$. These numbers are uniformly spaced with spacing $\epsilon_M * 2^e$ and therefore $|\mathrm{fl}(x) - x| \leq \frac{1}{2}\epsilon_M 2^e \leq \frac{1}{2}\epsilon_M * |x|$. The proof for a negative x is similar. □

The number u_M is called the **rounding unit**.

Exercise F.8 *Show that the upper bound for δ is attained for $x = (1 + \epsilon_M/2) * 2^e$ when round to zero is used. Compute δ when $x = (2 - \epsilon_M/2) * 2^e$.*

F.3.2 Arithmetic Operations

Suppose $x, y \in \mathcal{N}$. In a standard system we have

$$\mathrm{fl}(x \circ y) = (x \circ y)(1+\delta), \quad |\delta| \leq u_M, \quad \circ \in \{+, -, *, /, \sqrt{}\}, \tag{F.6}$$

where u_M is the rounding unit of the system. This means that the computed value is as good as the rounded exact answer. This is usually achieved by using one or several extra digits known as **guard digits** in the calculation.

F.4 Backward Rounding-Error Analysis

The computed sum of two numbers $\alpha_1, \alpha_2 \in \mathcal{N}$ satisfy $\mathrm{fl}(\alpha_1 \circ \alpha_2) = (\alpha_1 + \alpha_2)(1+\delta)$, where $|\delta| \leq u_M$, the rounding unit. If we write this as $\mathrm{fl}(\alpha_1 \circ \alpha_2) = \tilde{\alpha}_1 + \tilde{\alpha}_2$, where $\tilde{\alpha}_i := \alpha_i(1+\delta)$ for $i = 1, 2$, we see that the computed sum is the exact sum of two numbers which approximate the exact summands with small relative error,

$|\delta| \leq u_M$. The error in the addition has been boomeranged back on the data α_1, α_2, and in this context we call δ the **backward error**. A similar interpretation is valid for the other arithmetic operations $-, *, /, \sqrt{}$, and we assume it also holds for the elementary functions $\sin, \cos, \exp, \log$ and so on.

Suppose more generally we want to compute the value of an expression $\phi(\alpha_1, \ldots, \alpha_n)$ Here $\alpha_1, \ldots, \alpha_n \in \mathcal{N}$ are given data, and we are using the arithmetic operations, and implementations of the standard elementary functions, in the computation. A **backward error analysis** consists of showing that the computed result is obtained as the exact result of using data $\boldsymbol{\beta} := [\beta_1, \ldots \beta_n]^T$ instead of $\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_n]$. In symbols

$$\tilde{\phi}(\alpha_1, \ldots, \alpha_n) = \phi(\beta_1, \ldots, \beta_n).$$

If we can show that the relative error in $\boldsymbol{\beta}$ as an approximation to $\boldsymbol{\alpha}$ is $O(u_M)$ either componentwise or norm-wise in some norm, then we say that the algorithm to compute $\phi(\alpha_1, \ldots, \alpha_n)$ is **backward stable**. Normally the constant K in the $O(u_M)$ term will grow with n. Typically $K = p(n)$ for some polynomial p is acceptable, while an exponential growth of K can be problematic.

F.4.1 Computing a Sum

We illustrate this discussion by computing the backward error in the sum of n numbers $s := \alpha_1 + \cdots + \alpha_n$, where $\alpha_i \in \mathcal{N}$ for all i. We have the following algorithm.

```
s_1 := α_1
for k = 2 : n
    s_k := fl(s_{k-1} + α_k)
end
s̃ := s_n
```

Using a standard system we obtain for $n = 3$

$$s_2 = \mathrm{fl}(\alpha_1 + \alpha_2) = \alpha_1(1 + \delta_2) + \alpha_2(1 + \delta_2),$$
$$s_3 = \mathrm{fl}(s_2 + \alpha_3) = s_2(1 + \delta_3) + \alpha_3(1 + \delta_3) = \alpha_1(1 + \eta_1) + \alpha_2(1 + \eta_2) + \alpha_3(1 + \eta_3),$$
$$\eta_1 = \eta_2 = (1 + \delta_2)(1 + \delta_3), \quad \eta_3 = (1 + \delta_3), \quad |\delta_i| \leq u_M.$$

In general, with $\delta_1 := 0$,

$$\tilde{s} = \sum_{i=1}^{n} \alpha_i(1 + \eta_i). \quad \eta_i = (1 + \delta_i) \ldots (1 + \delta_n), \quad |\delta_i| \leq u_M, \quad i = 1, \ldots, n. \tag{F.7}$$

With $\phi(\alpha_1, \ldots, \alpha_n) := \alpha_1 + \cdots + \alpha_n$ this shows that

$$\tilde{s} = \tilde{\phi}(\alpha_1, \ldots, \alpha_n) = \phi(\beta_1, \ldots, \beta_n), \quad \beta_i = \alpha_i(1 + \eta_i). \tag{F.8}$$

The following lemma gives a convenient bound on the η factors.

Lemma F.9 *Suppose for integers k, m with $0 \le m \le k$ and $k \ge 1$ that*

$$1 + \eta_k := \frac{(1+\delta_1)\cdots(1+\delta_m)}{(1+\delta_{m+1})\cdots(1+\delta_k)}, \quad |\delta_j| \le u_M, \quad j = 1, \ldots, k.$$

If $ku_M \le \frac{1}{11}$ then

$$|\eta_k| \le ku'_M, \text{ where } u'_M := 1.1u_M. \tag{F.9}$$

Proof. We first show that

$$ku_M \le \alpha < 1 \Longrightarrow |\eta_k| \le k\frac{u_M}{1-\alpha}. \tag{F.10}$$

For convenience we use $u := u_M$ in the proof. Since $u < 1$ we have $1/(1-u) = 1 + u + u^2/(1-u) > 1 + u$ and we obtain

$$(1-u)^k \le \frac{(1-u)^m}{(1+u)^{k-m}} \le 1 + \eta_k \le \frac{(1+u)^m}{(1-u)^{k-m}} \le (1-u)^{-k}.$$

The proof of (F.10) will be complete if we can show that

$$1 - ku \le (1-u)^k, \quad (1-u)^{-k} \le 1 + ku'.$$

The first inequality is an easy induction on k. If it holds for k, then

$$(1-u)^{k+1} = (1-u)^k(1-u) \ge (1-ku)(1-u) = 1 - (k+1)u + ku^2 \ge 1 - (k+1)u.$$

The second inequality is a consequence of the first,

$$(1-u)^{-k} \le (1-ku)^{-1} = 1 + \frac{ku}{1-ku} \le 1 + \frac{ku}{1-\alpha} = 1 + ku'.$$

Letting $\alpha = \frac{1}{11}$ in (F.10) we obtain (F.9). □

The number $u'_M := 1.1u_M$, corresponding to $\alpha = 1/11$, is called the **adjusted rounding unit** . In the literature many values of α can be found. [17] uses $\alpha = 1/10$ giving $u'_M = 1.12u_M$, while in [8] the value $\alpha = 0.01$ can be found. In the classical work [25] one finds $1/(1-\alpha) = 1.06$.

Let us return to the backward error (F.8) in a sum of n numbers. Since $\delta_1 = 0$ we see that

$$|\eta_1| \le (n-1)u'_M, \quad |\eta_i| \le (n-i+1)u'_M, \text{ for } i = 2, \ldots, n.$$

or more simply

$$|\eta_i| \le (n-1)u'_M, \text{ for } i = 1, \ldots, n. \tag{F.11}$$

This shows that the algorithm for computing a sum is backward stable.

The bounds from a backward rounding-error analysis can be used together with a condition number to bound the actual error in the computed result. To see

this for the sum, we subtract the exact sum $s = \alpha_1 + \cdots + \alpha_n$ from the computed sum $\tilde{s} = \alpha_1(1+\eta_1) + \cdots + \alpha_n(1+\eta_n)$, to get

$$|\tilde{s} - s| = |\alpha_1\eta_1 + \cdots + \alpha_n\eta_n| \leq (|\alpha_1| + \cdots + |\alpha_n|)(n-1)u'_M.$$

Thus the relative error in the computed sum of n numbers is bounded as follows

$$|\frac{\tilde{s}-s}{s}| \leq \kappa(n-1)u'_M, \text{ where } \kappa := \frac{|\alpha_1| + \cdots + |\alpha_n|}{\alpha_1 + \cdots + \alpha_n}. \tag{F.12}$$

This bound shows that the backward error can be magnified by at most κ. The number κ is called the **condition number**. for the sum.

The condition number measures how much a relative error in each of the components in a sum can be magnified in the final sum. The backward error shows how large these relative perturbations can be in the actual algorithm we used to compute the sum. Using backward error analysis and condition number separates the process of estimating the error in the final result into two distinct jobs.

A problem where small relative changes in the data leads to large relative changes in the exact result is called **ill conditioned**. We see that computing a sum can be ill-conditoned if the exact value of the sum is close to zero and some of the individualterms have large absolute values with opposite signs.

F.4.2 Computing an Inner Product

Computing an inner product $p := \alpha_1\gamma_1 + \cdots + \alpha_n\gamma_n$ is also backward stable using the standard algorithm

$p_1 := \text{fl}(\alpha_1\gamma_1)$
for $k = 2 : n$
    $p_k := \text{fl}\big(p_{k-1} + \text{fl}(\alpha_k\gamma_k)\big)$
end
$\tilde{p} := p_n$

For a backward error analysis of this algorithm we only need to modify (F.7) slightly. All we have to do is to add terms $\text{fl}(\alpha_k\gamma_k) = \alpha_k\gamma_k(1+\pi_k)$ to the terms of the sum. The result is

$$\tilde{p} = \sum_{k=1}^{n} \alpha_k\gamma_k(1+\eta_k), \quad \eta_k = (1+\pi_k)(1+\delta_k)\cdots(1+\delta_n), \quad k = 1, \ldots, n,$$

where $\delta_1 = 0$. Thus for the inner product of n terms we obtain

$$|\frac{\tilde{p}-p}{p}| \leq \kappa n u_M, \quad \kappa := \frac{|\alpha_1\gamma_1| + \cdots + |\alpha_n\gamma_n|}{|\alpha_1\gamma_1 + \cdots + \alpha_n\gamma_n|}. \tag{F.13}$$

The computation can be ill conditioned if the exact value is close to zero and some of the components are large in absolute value.

F.4.3 Computing a Matrix Product

Using matrix norms we can bound the backward error in matrix algorithms. Suppose we want to compute the matrix product $\boldsymbol{C} = \boldsymbol{A} * \boldsymbol{B}$. Let n be the number of columns of $\boldsymbol{A}$ and the number of rows of $\boldsymbol{B}$. Each element in $\boldsymbol{C}$ is the inner product of a row of $\boldsymbol{A}$ and a column of $\boldsymbol{B}$. Thus if $\tilde{\boldsymbol{C}}$ is the computed product then from (F.13)

$$|\frac{\tilde{c}_{ij} - c_{ij}}{c_{ij}}| \leq \kappa_{ij} n u'_M, \quad \kappa_{ij} := \frac{|a_1 b_1| + \cdots + |a_n b_n|}{|a_1 b_1 + \cdots + a_n b_n|}, \text{ all } i, j. \tag{F.14}$$

We write this as $|\tilde{c}_{ij} - c_{ij}| \leq \kappa_{ij} |c_{ij}| n u'_M$. Using the infinity matrix norm we find

$$\sum_j |\tilde{c}_{ij} - c_{ij}| \leq n u'_M \sum_j \kappa_{ij} |c_{ij}| \leq \kappa n u'_M \sum_j |c_{ij}| \leq \kappa n u'_M \|\boldsymbol{C}\|_\infty, \text{ all } i,$$

where $\kappa := \max_{ij} \kappa_{ij}$. Maximizing over i we obtain

$$\frac{\|\tilde{\boldsymbol{C}} - \boldsymbol{C}\|_\infty}{\|\boldsymbol{C}\|_\infty} \leq \kappa n u'_M. \tag{F.15}$$

The calculation of a matrix product can be ill conditioned if one or more of the product elements are small and the corresponding inner products have large terms of opposite signs.

Appendix G

Differentiation of Vector Functions

For any sufficiently differentiable $f : \mathbb{R}^n \to \mathbb{R}$ we recall that the partial derivative with respect to the ith variable of f is defined by

$$D_i f(\boldsymbol{x}) := \frac{\partial f(\boldsymbol{x})}{\partial x_i} := \lim_{h \to \mathbf{0}} \frac{f(\boldsymbol{x} + h\boldsymbol{e}_i) - f(\boldsymbol{x})}{h}, \quad \boldsymbol{x} \in \mathbb{R}^n,$$

where $\boldsymbol{e}_i$ is the ith unit vector in $\mathbb{R}^n$. For each $\boldsymbol{x} \in \mathbb{R}^n$ we define the **gradient** $\nabla f(\boldsymbol{x}) \in \mathbb{R}^n$, and the **hessian** $\nabla\nabla^T f(\boldsymbol{x}) \in \mathbb{R}^{n,n}$ of f by

$$\nabla f :== \begin{bmatrix} D_1 f \\ \vdots \\ D_n f \end{bmatrix}, \quad \boldsymbol{H} f := \nabla\nabla^T f := \begin{bmatrix} D_1 D_1 f & \cdots & D_1 D_n f \\ \vdots & & \vdots \\ D_n D_1 & \cdots & D_n D_n f \end{bmatrix}, \tag{G.1}$$

where $\nabla^T f := (\nabla f)^T$ is the row vector gradient. The operators $\nabla\nabla^T$ and $\nabla^T\nabla$ are quite different. Indeed, $\nabla^T\nabla f = D_1^2 f + \cdots + D_n^2 f =: \nabla^2$ the **Laplacian** of f, while $\nabla\nabla^T$ can be thought of as an outer product resulting in a matrix.

Exercise G.1 *For $f, g : \mathbb{R}^n \to \mathbb{R}$ show the product rules*

1. $\nabla(fg) = f\nabla g + g\nabla f, \quad \nabla^T(fg) = f\nabla^T g + g\nabla^T f,$
2. $\nabla\nabla^T(fg) = \nabla f\nabla^T g + \nabla g\nabla^T f + f\nabla\nabla^T g + g\nabla\nabla^T f.$
3. $\nabla^2(fg) = 2\nabla^T f\nabla g + f\nabla^2 g + g\nabla^2 f.$

We define the **Jacobian** of a vector function $\boldsymbol{f} = [f_1, \ldots f_m]^T : \mathbb{R}^n \to \mathbb{R}^m$ as the m, n matrix

$$\nabla^T \boldsymbol{f} := \begin{bmatrix} D_1 f_1 & \cdots & D_n f_1 \\ \vdots & & \vdots \\ D_1 f_m & \cdots & D_n f_m \end{bmatrix}.$$

As an example, if $f(\boldsymbol{x}) = f(x,y) = x^2 - xy + y^2$ and $\boldsymbol{g}(x,y) := [f(x,y),\ x-y]^T$ then

$$\nabla f(x,y) = \begin{bmatrix} 2x-y \\ -x+2y \end{bmatrix}, \quad \nabla^T \boldsymbol{g}(x,y) = \begin{bmatrix} 2x-y & -x+2y \\ 1 & -1 \end{bmatrix},$$
$$\boldsymbol{H}f(x,y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

The second order Taylor expansion in n variables can be expressed in terms of the gradient and the hessian.

Lemma G.2 *Suppose $f \in C^2(\Omega)$, where $\Omega \in \mathbb{R}^n$ contains two points $\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in \Omega$, such that the line segment $L := \{\boldsymbol{x} + t\boldsymbol{h} : t \in (0,1)\} \subset \Omega$. Then*

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \boldsymbol{h}^T \nabla f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{h}^T \nabla \nabla^T f(\boldsymbol{c})\boldsymbol{h}, \text{ for some } \boldsymbol{c} \in L. \tag{G.2}$$

Proof. Let $g : [0,1] \to \mathbb{R}$ be defined by $g(t) := f(\boldsymbol{x} + t\boldsymbol{h})$. Then $g \in C^2[0,1]$ and by the chain rule

$$g(0) = f(\boldsymbol{x}) \quad g(1) = f(\boldsymbol{x}+\boldsymbol{h}),$$
$$g'(t) = \sum_{i=1}^{n} h_i \frac{\partial f(\boldsymbol{x}+t\boldsymbol{h})}{\partial x_i} = \boldsymbol{h}^T \nabla f(\boldsymbol{x}+t\boldsymbol{h}),$$
$$g''(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} h_i h_j \frac{\partial^2 f(\boldsymbol{x}+t\boldsymbol{h})}{\partial x_i \partial x_j} = \boldsymbol{h}^T \nabla \nabla^T f(\boldsymbol{x}+t\boldsymbol{h})\boldsymbol{h}.$$

Inserting these expressions in the second order Taylor expansion

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(u), \text{ for some } u \in (0,1),$$

we obtain (G.2) with $\boldsymbol{c} = \boldsymbol{x} + u\boldsymbol{h}$. □

The gradient and hessian of some functions involving matrices can be found from the following lemma.

Lemma G.3 *For any $m, n \in \mathbb{N}$, $\boldsymbol{B} \in \mathbb{R}^{n,n}$, $\boldsymbol{C} \in \mathbb{R}^{m,n}$, and $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m$ we have*

1. $\nabla(\boldsymbol{y}^T\boldsymbol{C}) = \nabla^T(\boldsymbol{C}\boldsymbol{x}) = \boldsymbol{C}$,
2. $\nabla(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}) = (\boldsymbol{B}+\boldsymbol{B}^T)\boldsymbol{x}, \quad \nabla^T(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{B}+\boldsymbol{B}^T)$,
3. $\nabla\nabla^T(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}) = \boldsymbol{B}+\boldsymbol{B}^T$.

Proof.

1. We find $D_i(\boldsymbol{y}^T\boldsymbol{C}) = \lim_{h\to 0}\frac{1}{h}\big((\boldsymbol{y}+h\boldsymbol{e}_i)^T\boldsymbol{C} - \boldsymbol{y}^T\boldsymbol{C}\big) = \boldsymbol{e}_i^T\boldsymbol{C}$ and $D_i(\boldsymbol{C}\boldsymbol{x}) = \lim_{h\to 0}\frac{1}{h}(\boldsymbol{C}(\boldsymbol{x}+h\boldsymbol{e}_i) - \boldsymbol{C}\boldsymbol{x}) = \boldsymbol{C}\boldsymbol{e}_i$ and 1. follows.

2. Here we find

$$\begin{aligned} D_i(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}) &= \lim_{h\to 0}\frac{1}{h}\big((\boldsymbol{x}+h\boldsymbol{e}_i)^T\boldsymbol{B}(\boldsymbol{x}+h\boldsymbol{e}_i) - \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}\big) \\ &= \lim_{h\to 0}\big(\boldsymbol{e}_i^T\boldsymbol{B}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{e}_i + h\boldsymbol{e}_i^T\boldsymbol{e}_i\big) = \boldsymbol{e}_i^T(\boldsymbol{B}+\boldsymbol{B}^T)\boldsymbol{x}, \end{aligned}$$

   and the first part of 2. follows. Taking transpose we obtain the second part.

3. Combining 1. and 2. we obtain 3.

□

Appendix H

Some Inequalities

In this appendix we derive an inequality for convex functions called Jensen's inequality and use it to show H"older's and Minkowski's inequaltiies.

H.1 Convexity

Definition H.1 (Convex function) *Let $I \subset \mathbb{R}$ be an interval. A function $f : I \to \mathbb{R}$ is called convex if*

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$

for all $x, y \in I$ and all $\lambda \in [0,1]$. The sum $\sum_{j=1}^{n} \lambda_j z_j$ is called a **convex combination** *of $z_1, \ldots, z_n$ if $\lambda_j \geq 0$ for $j = 1, \ldots, n$ and $\sum_{j=1}^{n} \lambda_j = 1$.*

The condition is shown graphically in Figure H.1.

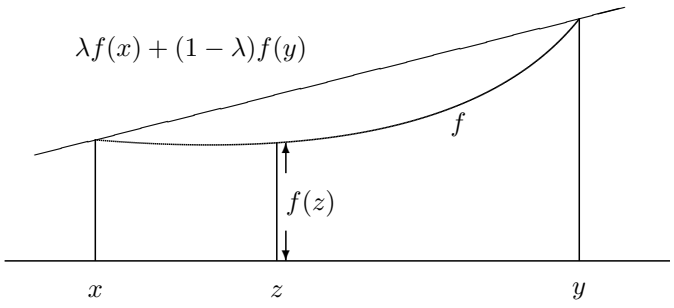


Figure H.1. *A convex function.*

It then follows that the function $-\log x$ is convex on $I = (0, \infty)$.

H.2 Inequalities

Theorem H.2 (Jensen's Inequality) *Suppose $I \in \mathbb{R}$ is an interval and $f : I \to \mathbb{R}$ is convex. Then for all $n \in \mathbb{N}$, all $\lambda_1, \ldots, \lambda_n$ with $\lambda_j \geq 0$ for $j = 1, \ldots, n$ and $\sum_{j=1}^n \lambda_j = 1$, and all $z_1, \ldots, z_n \in I$ we have*

$$f(\sum_{j=1}^n \lambda_j) \leq \sum_{j=1}^n \lambda_j f(z_j).$$

Proof. We use induction on n. The result is trivial for $n = 1$. Let $n \geq 2$, assume the inequality holds for $k = n - 1$, and let λ_j, z_j for $j = 1, \ldots, n$ be given as in the theorem. Since $n \geq 2$ we have $\lambda_i < 1$ for at leas one i so assume without loss of generality that $\lambda_1 < 1$. Define u by $u := \sum_{j=2}^n \frac{\lambda_j}{1-\lambda_1} z_j$. Since $\sum_{j=2}^n \lambda_j = 1 - \lambda_1$ this is a convex combination of k terms and the induction hypothesis implies that $f(u) \leq \sum_{j=2}^n \frac{\lambda_j}{1-\lambda_1} f(z_j)$. But then by the convexity of f

$$f(\sum_{j=1}^n \lambda_j) = f(\lambda_1 z_1 + (1-\lambda_1)u) \leq \lambda_1 f(z_1) + (1-\lambda_1) f(u) \leq \sum_{j=1}^n \lambda_j f(z_j)$$

and the inequality holds for $k + 1 = n$. □

Corollary H.3 (Weighted geometric/arithmetic mean inequality) *Suppose $\sum_{j=1}^n \lambda_j a_j$ is a convex combination of nonnegative numbers $a_1, \ldots, a_n$. Then*

$$a_1^{\lambda_1} a_2^{\lambda_2} \cdots a_n^{\lambda_n} \leq \sum_{j=1}^n \lambda_j a_j, \tag{H.1}$$

where $0^0 := 0$.

Proof. The result is trivial if one or more of the a_j's are zero so assume $a_j > 0$ for all j. We use Jensen's inequality with the convex function $f(x) = -\log x$ on $I = (0, \infty)$. Then

$$-\log\big(\sum_{j=1}^n \lambda_j a_j\big) \leq -\sum_{j=1}^n \lambda_j \log(a_j) = -\log\big(a_1^{\lambda_1} \cdots a_n^{\lambda_n}\big)$$

and since the log function is monotone the inequality follows. □

Taking $\lambda_j = \frac{1}{n}$ for all j in (H.1) we obtain the classical **geometric/arithmetic mean inequality**

$$(a_1 a_2 \cdots a_n)^{\frac{1}{n}} \leq \frac{1}{n} \sum_{j=1}^n a_j. \tag{H.2}$$

Corollary H.4 (Hölder's inequality) *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ and $1 \leq p \leq \infty$*

$$\sum_{j=1}^{n} |x_j y_j| \leq \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q, \text{ where } \frac{1}{p} + \frac{1}{q} = 1.$$

Proof. We leave the proof for $p = 1$ and $p = \infty$ as an exercise so assume $1 < p < \infty$. For any $a, b \geq 0$ the weighted arithmetic/geometric mean inequality implies that

$$a^{\frac{1}{p}} b^{\frac{1}{q}} \leq \frac{1}{p} a + \frac{1}{q} b, \text{ where } \frac{1}{p} + \frac{1}{q} = 1. \tag{H.3}$$

If $\boldsymbol{x} = \boldsymbol{0}$ or $\boldsymbol{y} = \boldsymbol{0}$ there is nothing to prove so assume that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are nonzero. Using H.3 on each term we obtain

$$\frac{1}{\|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q} \sum_{j=1}^{n} |x_j y_j| = \sum_{j=1}^{n} \left(\frac{|x_j|^p}{\|\boldsymbol{x}\|_p^p} \right)^{\frac{1}{p}} \left(\frac{|y_j|^q}{\|\boldsymbol{y}\|_q^q} \right)^{\frac{1}{q}} \leq \sum_{j=1}^{n} \left(\frac{1}{p} \frac{|x_j|^p}{\|\boldsymbol{x}\|_p^p} + \frac{1}{q} \frac{|y_j|^q}{\|\boldsymbol{y}\|_q^q} \right) = 1$$

and the proof of the inequality is complete. ∎

Corollary H.5 (Minkowski's inequality) *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ and $1 \leq p \leq \infty$*

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p \leq \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p.$$

Proof. We leave the proof for $p = 1$ and $p = \infty$ as an exercise so assume $1 < p < \infty$. We write

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p^p = \sum_{j=1}^{n} |x_j + y_j|^p \leq \sum_{j=1}^{n} |x_j| |x_j + y_j|^{p-1} + \sum_{j=1}^{n} |y_j| |x_j + y_j|^{p-1}.$$

We apply Hölder's inequality with exponent p and q to each sum. In view of the relation $(p-1)q = p$ the result is

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p^p \leq \|\boldsymbol{x}\|_p \|\boldsymbol{x} + \boldsymbol{y}\|_p^{p/q} + \|\boldsymbol{x}\|_p \|\boldsymbol{x} + \boldsymbol{y}\|_p^{p/q}.$$

Since $p - \frac{p}{q} = 1$ the inequality follows. ∎

Appendix I

The Jordan Form

I.1 The Jordan Form

We have seen that any square matrix can be triangularized by a unitary similarity transformation. Moreover, any nondefective matrix can be diagonalized. The following question arises. How close to a diagonal matrix can we reduce a defective matrix by a similarity transformation?

Definition I.1 *A* **Jordan block**, *denoted* $\boldsymbol{J}_m(\lambda)$ *is an* $m \times m$ *matrix of the form*

$$\boldsymbol{J}_m(\lambda) := \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

A 3×3 Jordan block has the form $\boldsymbol{J}_3(\lambda) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$. We see that λ is an eigenvalue of $\boldsymbol{J}_m(\lambda)$ and any eigenvector must be a multiple of $\boldsymbol{e}_1$. Thus, the eigenvectors of $\boldsymbol{J}_m(\lambda)$ have algebraic multiplicity m and geometric multiplicity one.

The Jordan canonical form is a decomposition of a matrix into Jordan blocks.

Theorem I.2 *Suppose* $\boldsymbol{A} \in \mathbb{C}^{n,n}$ *has* k *distinct eigenvalues* $\lambda_1, \ldots, \lambda_k$ *of algebraic multiplicities* $a_1, \ldots, a_k$ *and geometric multiplicities* $g_1, \ldots, g_k$. *There is a nonsingular matrix* $\boldsymbol{S} \in \mathbb{C}^{n,n}$ *such that*

$$\boldsymbol{J} := \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} = \operatorname{diag}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k), \text{ with } \boldsymbol{U}_i \in \mathbb{C}^{a_i, a_i}, \tag{I.1}$$

where each $\boldsymbol{U}_i$ *is block diagonal having* g_i *Jordan blocks along the diagonal*

$$\boldsymbol{U}_i = \operatorname{diag}(\boldsymbol{J}_{m_{i,1}}(\lambda_i), \ldots, \boldsymbol{J}_{m_{i,g_i}}(\lambda_i). \tag{I.2}$$

Here $m_{i,1}, \ldots, m_{i,g_i}$ *are unique integers so that* $m_{i,1} \geq m_{i,2} \geq \cdots \geq m_{i,g_i}$ *and* $a_i = \sum_{j=1}^{g_i} m_{i,j}$ *for all* i.

The matrix $\boldsymbol{J}$ in (I.1) is called the **Jordan form** of $\boldsymbol{A}$. As an example consider the Jordan form

$$\boldsymbol{J} := \operatorname{diag}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \begin{bmatrix} 2 & 1 & 0 & & & & & \\ 0 & 2 & 1 & & & & & \\ 0 & 0 & 2 & & & & & \\ & & & 2 & 1 & & & \\ & & & 0 & 2 & & & \\ & & & & & 2 & & \\ & & & & & & 3 & 1 \\ & & & & & & 0 & 3 \end{bmatrix} \in \mathbb{R}^{8,8}. \tag{I.3}$$

The eigenvalues together with their algebraic and geometric multiplicities can be read off directly from the Jordan form.

- $\boldsymbol{U}_1 = \operatorname{diag}(\boldsymbol{J}_3(2), \boldsymbol{J}_2(2), \boldsymbol{J}_1(2))$ and $\boldsymbol{U}_2 = \boldsymbol{J}_2(3)$.
- 2 is an eigenvalue of algebraic multiplicity 6 and geometric multiplicity 3.
- 3 is an eigenvalue of algebraic multiplicity 2 and geometric multiplicity 1.

Each $\boldsymbol{U}_i$ is upper triangular with the eigenvalue λ_i on the diagonal and consists of g_i Jordan blocks. These Jordan blocks can be taken in any order and it is customary to refer to any such block diagonal matrix as the Jordan form of $\boldsymbol{A}$. Thus in the example the matrix

$$\boldsymbol{J} := \begin{bmatrix} 3 & 1 & & & & & & \\ 0 & 3 & & & & & & \\ & & 2 & 1 & & & & \\ & & 0 & 2 & & & & \\ & & & & 2 & & & \\ & & & & & 2 & 1 & 0 \\ & & & & & 0 & 2 & 1 \\ & & & & & 0 & 0 & 2 \end{bmatrix}$$

is also a Jordan form of $\boldsymbol{A}$. In any Jordan form of this $\boldsymbol{A}$ the sizes of the 4 Jordan blocks $\boldsymbol{J}_3(2), \boldsymbol{J}_2(2), \boldsymbol{J}_1(2), \boldsymbol{J}_2(3)$ are uniquely given.

The columns of $\boldsymbol{S}$ are called **principal vectors**. They satisfy the matrix equation $\boldsymbol{AS} = \boldsymbol{SJ}$. As an example, in (I.3) we have $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_8]$ and we find

$$\begin{aligned} &\boldsymbol{A}\boldsymbol{s}_1 = 2\boldsymbol{s}_1, \quad \boldsymbol{A}\boldsymbol{s}_2 = 2\boldsymbol{s}_2 + \boldsymbol{s}_1, \\ &\boldsymbol{A}\boldsymbol{s}_3 = 2\boldsymbol{s}_3, \\ &\boldsymbol{A}\boldsymbol{s}_4 = 2\boldsymbol{s}_4, \quad \boldsymbol{A}\boldsymbol{s}_5 = 2\boldsymbol{s}_5 + \boldsymbol{s}_4, \quad \boldsymbol{A}\boldsymbol{s}_6 = 2\boldsymbol{s}_6 + \boldsymbol{s}_5, \\ &\boldsymbol{A}\boldsymbol{s}_7 = 3\boldsymbol{s}_7, \quad \boldsymbol{A}\boldsymbol{s}_8 = 3\boldsymbol{s}_8 + \boldsymbol{s}_7, \end{aligned}$$

We see that the first principal vector in each Jordan block is an eigenvector of $\boldsymbol{A}$. The remaining principal vectors are not eigenvectors.

Exercise I.3 *For the Jordan form of the matrix* $\boldsymbol{A} = \begin{bmatrix} 3 & 0 & 1 \\ -4 & 1 & -2 \\ -4 & 0 & -1 \end{bmatrix}$ *we have* $\boldsymbol{J} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. *Find* $\boldsymbol{S}$.

Exercise I.4 *Find the Jordan form of the matrix*

$$A = \frac{1}{9} \begin{bmatrix} 10 & 16 & -8 & -5 & 6 & 1 & -3 & 4 \\ -7 & 32 & -7 & -10 & 12 & 2 & -6 & 8 \\ -6 & 12 & 12 & -15 & 18 & 3 & -9 & 12 \\ -5 & 10 & -5 & -2 & 24 & 4 & -12 & 16 \\ -4 & 8 & -4 & -16 & 30 & 14 & -15 & 20 \\ -3 & 6 & -3 & -12 & 9 & 24 & -9 & 24 \\ -2 & 4 & -2 & -8 & 6 & -2 & 15 & 28 \\ -1 & 2 & -1 & -4 & 3 & -1 & -6 & 41 \end{bmatrix}. \tag{I.4}$$

The following lemma is useful when studying powers of matrices.

Lemma I.5 *Let $\boldsymbol{J}$ be the Jordan form of a matrix $\boldsymbol{A} \in \mathbb{C}^{n,n}$ as given in Theorem I.2. Then for $r = 0, 1, 2, \ldots$, $m = 2, 3, \ldots$, and any $\lambda \in \mathbb{C}$*

1. $\boldsymbol{A}^r = \boldsymbol{S}\boldsymbol{J}^r\boldsymbol{S}^{-1}$,
2. $\boldsymbol{J}^r = \operatorname{diag}(\boldsymbol{U}_1^r, \ldots, \boldsymbol{U}_k^r)$,
3. $\boldsymbol{U}_i^r = \operatorname{diag}(\boldsymbol{J}_{m_{i,1}}(\lambda_i)^r, \ldots, \boldsymbol{J}_{m_{i,g_i}}(\lambda_i)^r)$,
4. $\boldsymbol{E}_m^r = \left[\begin{smallmatrix} \boldsymbol{0} & \boldsymbol{I}_{m-r} \\ \boldsymbol{0} & \boldsymbol{0} \end{smallmatrix}\right]$ *for* $1 \le r \le m-1$, *where* $\boldsymbol{E}_m := \boldsymbol{J}_m(\lambda) - \lambda\boldsymbol{I}_m$,
5. $\boldsymbol{E}_m^m = \boldsymbol{0}$.
6. $\boldsymbol{J}_m(\lambda)^r = (\boldsymbol{E}_m + \lambda\boldsymbol{I}_m)^r = \sum_{k=0}^{\min\{r,m-1\}} \binom{r}{k} \lambda^{r-k} \boldsymbol{E}_m^k$

Proof.

1. We have $\boldsymbol{A}^2 = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}\boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1} = \boldsymbol{S}\boldsymbol{J}^2\boldsymbol{S}^{-1}$ and 1. follows by induction on r.
2. This follows since $\boldsymbol{J}$ is block diagonal.
3. Each $\boldsymbol{J}_{m_i,j}$ is block diagonal.
4. We have
$$\boldsymbol{E}_m = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_{m-1} \\ 0 & \boldsymbol{0}^T \end{bmatrix}. \tag{I.5}$$
The result follow for $r = 1$ and for general $r \le m-1$ by induction.
5. $\boldsymbol{E}_m^m = \boldsymbol{E}_m^{m-1}\boldsymbol{E}_m = \boldsymbol{0}$.
6. This follows from the binomial theorem since $\boldsymbol{I}_m$ and $\boldsymbol{E}_m$ commute and $\boldsymbol{E}^m = \boldsymbol{0}$.

□

Exercise I.6 *Determine $\boldsymbol{J}_3^r$ for $r \ge 1$.*

Exercise I.7 *Find $\boldsymbol{J}^{100}$ and $\boldsymbol{A}^{100}$ for the matrix in Exercise I.3.*

I.1.1 The Minimal Polynomial

Let $\boldsymbol{J}$ be the Jordan form of $\boldsymbol{A}$ given in Theorem I.2. Since $\boldsymbol{A}$ and $\boldsymbol{J}$ are similar they have the same characteristic polynomial, and since the Jordan form of $\boldsymbol{A}$ is upper triangular with the eigenvalues of $\boldsymbol{A}$ on the diagonal we have

$$\pi_{\boldsymbol{A}}(\lambda) = \pi_{\boldsymbol{J}}(\lambda) = \prod_{i=1}^{k}\prod_{j=1}^{g_i}(\lambda_i - \lambda)^{m_{ij}}.$$

The polynomials $p_{ij}(\lambda) := (\lambda_i - \lambda)^{m_{ij}}$ are called the **elementary divisors** of $\boldsymbol{A}$. They divide the characteristic polynomial.

Definition I.8 *Suppose $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$ is the Jordan canonical form of $\boldsymbol{A}$. The polynomial*

$$\mu(z) := \prod_{i=1}^{k}(\lambda_i - z)^{m_i} \text{ where } m_i := \max_{1\le j\le g_i} m_{ij},$$

is called the **minimal polynomial** *of $\boldsymbol{A}$.*

Since each factor in $\mu(z)$ is also a factor in $\pi_{\boldsymbol{A}}(z)$, we have the factorization $\pi_{\boldsymbol{A}}(z) = \mu(z)\nu(z)$ for some polynomial $\nu(z)$.

Exercise I.9 *What is the characteristic polynomial and the minimal polynomial of the matrix $\boldsymbol{J}$ in* (I.3)*?*

To see in what way the minimal polynomial is minimal, we consider two matrices defined from the characteristic polynomial π_A and the minimal polynomial. Substituting a matrix for the independent variable in these polynomial we obtain

$$\pi_{\boldsymbol{A}}(\boldsymbol{A}) := \prod_{i=1}^{k}\prod_{j=1}^{g_i}(\lambda_i\boldsymbol{I} - \boldsymbol{A})^{m_{ij}}, \quad \mu(\boldsymbol{A}) := \prod_{i=1}^{k}(\lambda_i\boldsymbol{I} - \boldsymbol{A})^{m_i}. \tag{I.6}$$

By induction it is easy to see that $\mu(\boldsymbol{A})$ and $\pi_{\boldsymbol{A}}(\boldsymbol{A})$ are polynomials in the matrix $\boldsymbol{A}$. Moreover, $\mu(\boldsymbol{A}) = \prod_{i=1}^{k}(\lambda_i\boldsymbol{I} - \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1})^{m_i} = \boldsymbol{S}\mu(\boldsymbol{J})\boldsymbol{S}^{-1}$, so that $\mu(\boldsymbol{A}) = \boldsymbol{0}$ if and only if $\mu(\boldsymbol{J}) = \boldsymbol{0}$. Now,

$$\begin{aligned}\mu(\boldsymbol{J}) &= \prod_{i=1}^{k}(\lambda_i\boldsymbol{I} - \boldsymbol{J})^{m_i} = \prod_{i=1}^{k}\operatorname{diag}\big((\lambda_i\boldsymbol{I} - \boldsymbol{U}_1)^{m_i}, \dots, (\lambda_i\boldsymbol{I} - \boldsymbol{U}_k)^{m_i}\big)\\ &= \operatorname{diag}\big(\prod_{i=1}^{k}(\lambda_i\boldsymbol{I} - \boldsymbol{U}_1)^{m_i}, \dots, \prod_{i=1}^{k}(\lambda_i\boldsymbol{I} - \boldsymbol{U}_k)^{m_i}\big) = \boldsymbol{0},\end{aligned}$$

since $(\lambda_r\boldsymbol{I} - \boldsymbol{U}_r)^{m_r} = \boldsymbol{0}$ for $r = 1, \dots, k$. To show the latter we observe that

$$\begin{aligned}(\lambda_r\boldsymbol{I} - \boldsymbol{U}_r)^{m_r} &= \operatorname{diag}\big((\lambda_r\boldsymbol{I} - \boldsymbol{J}_{m_{r1}})^{m_r}, \dots, (\lambda_r\boldsymbol{I} - \boldsymbol{J}_{m_{r,g_r}})^{m_r}\big)\\ &= \operatorname{diag}(\boldsymbol{E}_{m_{r1}}^{m_r}, \dots, E_{m_{r,g_r}}^{m_r}) = \boldsymbol{0},\end{aligned}$$

by Lemma I.5 and the maximality of m_r.

We have shown that a matrix satisfies its minimal polynomial equation $\mu(\boldsymbol{A}) = \mathbf{0}$. Moreover, the degree of any polynomial p such that $p(\boldsymbol{A}) = \mathbf{0}$ is at least as large as the degree $d = \sum_{i=1}^{k} m_i$ of the minimal polynomial μ. This follows from the proof since any such polynomial must contain the elementary divisors $(\lambda_i - \lambda)^{m_i}$ for $i = 1, \ldots, k$. Since the minimal polynomial divides the characteristic polynomial we obtain as a corollary the **Cayley-Hamilton Theorem** which says that a matrix satisfies its characteristic equation $\pi_{\boldsymbol{A}}(\boldsymbol{A}) = \mathbf{0}$.

Exercise I.10 *Show that* $p(\boldsymbol{B}) = \boldsymbol{S}^{-1} p(\boldsymbol{A}) \boldsymbol{S}$ *for any polynomial* p *and any similar matrices* $\boldsymbol{B} = \boldsymbol{S}^{-1} \boldsymbol{A} \boldsymbol{S}$.

Exercise I.11 *What is the minimal polynomial of the unit matrix and more generally of a diagonalizable matrix?*

Bibliography

[1] Beckenbach, E. F, and R. Bellman, *Inequalities*, Springer Verlag, Berlin, Fourth Printing, 1983.

[2] Björck, Åke, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1995.

[3] Golub, G. H., and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD, third edition, 1996.

[4] Greenbaum, Anne, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[5] Hackbush, Wolfgang, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, Berlin, 1994.

[6] Hestenes, Magnus, *Conjugate Direction Methods in Optimization*, Springer-Verlag, Berlin, 1980.

[7] Hestenes, M. and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research of the National Bureau of Standards **29**(1952), 409–439.

[8] Higham, Nicloas J., *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[9] Horn, Roger A. and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[10] Horn, Roger A. and Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[11] Kato, *Perturbation Theory for Linear Operators*, Pringer.

[12] Lawson, C.L. and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N.J, 1974.

[13] Lax, Peter D., *Linear Algebra*, John Wiley & Sons, New York, 1997.

[14] Leon, Steven J., *Linear Algebra with Applications*, Prentice Hall, NJ, Seventh Edition, 2006.

[15] Meyer, Carl D., *Matrix Analysis and Applied Linear Algebra* , Siam Philadelphia, 2000.

[16] Steel, J. Michael, *The Cauchy-Schwarz Master Class*, Cambridge University Press, Cambridge, UK, 2004.

[17] Stewart, G. G., *Matrix Algorithms Volume I: Basic Decompositions*, Siam Philadelphia, 1998.

[18] Stewart, G. G., *Matrix Algorithms Volume II: Eigensystems*, Siam Philadelphia, 2001.

[19] Stewart, G. G. and Ji-guang Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.

[20] Stewart, G. G., *Introduction to Matrix Computations*, Academic press, New York, 1973.

[21] Trefethen, Lloyd N., and David Bau III, *Numerical Linear Algebra*, Siam Philadelphia, 1997.

[22] Tveito, A., and R. Winther, *Partial Differential Equations*, Springer, Berlin.

[23] Van Loan, Charles, *Computational Frameworks for the Fast Fourier Transform,* Siam Philadelphia, 1992.

[24] Varga, R. S., *Matrix Iterative Analysis/ 2nd Edn.*, Springer Verlag, New York, 2000.

[25] Wilkinson, J. H., *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[26] Young, D. M.,*Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

Index