Lecture Notes for MAT-INF 4130

Tom Lyche

July 2, 2013

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. Most of the chapters correspond approximately to one week of lectures. Earlier versions of this manuscript were converted to LaTeX by Are Magnus Bruaset and Njål Foldnes. A special thanks goes to Christian Schulz and Georg Muntingh who helped me with the exercise sessions and have provided solutions to all problems in this book.

Oslo, 1. July 2013

Tom Lyche

Chapter 0

A Short Review of Linear Algebra

In this introductory chapter we give a compact introduction to linear algebra with emphasis on $\mathbb{R}^n$ and $\mathbb{C}^n$. For a more elementary introduction, see for example the book [22]. We start by introducing the notation used.

0.1 Notation

The following sets and notations will be used in this book.

1. The sets 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. $\mathbb{R}^{m\times n}$ is the set of matrices $\boldsymbol{A}$ with real elements. The integers m and n are the number of rows and columns in the tableau

$$\boldsymbol{A}=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix}.$$

The element in the ith row and jth column of $\boldsymbol{A}$ 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 $\boldsymbol{a}_i^T$ with the risk of some confusion. 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 $z=x+iy=re^{i\phi}=r(\cos\phi+i\sin\phi)$, with $x,y\in\mathbb{R}$, is a complex number then $\overline{z}:=x-iy=re^{-i\phi}=\cos\phi-i\sin\phi$ and $|z|:=\sqrt{\overline{z}z}=\sqrt{x^2+y^2}=r$. $\mathrm{Re}(z):=x$ 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 $\boldsymbol{A}\in\mathbb{C}^{m\times n}$ and $\boldsymbol{x}\in\mathbb{C}^n$. We define complex row vectors using either the transpose $\boldsymbol{x}^T$ 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 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ are matrices of the same size, i. e., with the same number of rows and columns, and $c_{ij}=a_{ij}+b_{ij}$ for all i,j.

- **multiplication by a scalar** $\boldsymbol{C} = \alpha\boldsymbol{A}$, where $c_{ij} = \alpha a_{ij}$ for all i, j.
- **matrix multiplication** $\boldsymbol{C} = \boldsymbol{AB}$, $\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B}$ or $\boldsymbol{C} = \boldsymbol{A} * \boldsymbol{B}$, where $\boldsymbol{A} \in \mathbb{C}^{m\times p}$, $\boldsymbol{B} \in \mathbb{C}^{p\times n}$, $\boldsymbol{C} \in \mathbb{C}^{m\times n}$, 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** $\boldsymbol{C} = \boldsymbol{A} \times \boldsymbol{B}$, $\boldsymbol{D} = \boldsymbol{A}/\boldsymbol{B}$, and $\boldsymbol{E} = \boldsymbol{A} \wedge r$ where all matrices are of the same size 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 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ or $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. The **transpose** $\boldsymbol{A}^T$ and **conjugate transpose** $\boldsymbol{A}^*$ are $n \times m$ matrices with elements $a_{ij}^T = a_{ji}$ and $a_{ij}^* = \overline{a}_{ji}$, respectively. If $\boldsymbol{B}$ is an n, p matrix then $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$ and $(\boldsymbol{AB})^* = \boldsymbol{B}^*\boldsymbol{A}^*$.
11. The **unit vectors** in $\mathbb{R}^n$ and $\mathbb{C}^n$ are denoted by

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

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

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

is the **identity matrix** of order n. Both the collumns and the transpose of the rows of $\boldsymbol{I}$ are the unit vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.
12. Some matrices with many zeros have names indicating their "shape". Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ or $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. 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 $|i - j| > 1$.
- d**-banded** if $a_{ij} = 0$ for $|i - j| > d$.

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

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

$$\boldsymbol{B} = \operatorname{tridiag}(a_i, d_i, c_i) = \operatorname{tridiag}(\boldsymbol{a}, \boldsymbol{d}, \boldsymbol{c}) := \begin{bmatrix} d_1 & c_1 & & & \\ a_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix}.$$

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

14. Suppose $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ 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 $\boldsymbol{A}(\boldsymbol{i}, \boldsymbol{j}) \in \mathbb{C}^{r \times c}$ is the submatrix of $\boldsymbol{A}$ consisting of rows $\boldsymbol{i} := [i_1, \ldots, i_r]$ and columns $\boldsymbol{j} := [j_1, \ldots, j_c]$

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

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

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

0.2 Vector Spaces and Subspaces

Many mathematical systems have analogous properties to vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$.

Definition 0.1 (Real vector space)
A **real vector space** *is a nonempty set $\mathcal{V}$, whose objects are called* **vectors**, *together with two operations $+ : \mathcal{V} \times \mathcal{V} \longrightarrow \mathcal{V}$ and $\cdot : \mathbb{R} \times \mathcal{V} \longrightarrow \mathcal{V}$, called* **addition** *and* **scalar multiplication**, *satisfying the following axioms for all vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ in $\mathcal{V}$ and scalars c, d in $\mathbb{R}$.*

(V1) *The sum $\boldsymbol{u} + \boldsymbol{v}$ is in $\mathcal{V}$,*

(V2) $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$,

(V3) $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$,

(V4) *There is a* **zero vector** $\mathbf{0}$ *such that* $\boldsymbol{u} + \mathbf{0} = \boldsymbol{u}$,

(V5) *For each* $\boldsymbol{u}$ *in* $\mathcal{V}$ *there is a vector* $-\boldsymbol{u}$ *in* $\mathcal{V}$ *such that* $\boldsymbol{u} + (-\boldsymbol{u}) = \mathbf{0}$,

(S1) *The scalar multiple* $c \cdot \boldsymbol{u}$ *is in* $\mathcal{V}$,

(S2) $c \cdot (\boldsymbol{u} + \boldsymbol{v}) = c \cdot \boldsymbol{u} + c \cdot \boldsymbol{v}$,

(S3) $(c + d) \cdot \boldsymbol{u} = c \cdot \boldsymbol{u} + d \cdot \boldsymbol{u}$,

(S4) $c \cdot (d \cdot \boldsymbol{u}) = (cd) \cdot \boldsymbol{u}$,

(S5) $1 \cdot \boldsymbol{u} = \boldsymbol{u}$.

The scalar multiplication symbol $\cdot$ *is often omitted, writing* $c\boldsymbol{v}$ *instead of* $c \cdot \boldsymbol{v}$. *We define* $\boldsymbol{u} - \boldsymbol{v} := \boldsymbol{u} + (-\boldsymbol{v})$. *We call* $\mathcal{V}$ *a* **complex vector space** *if the scalars consist of all complex numbers* $\mathbb{C}$. *In this book a vector space is either real or complex.*

From the axioms it follows that

1. The zero vector is unique.
2. For each $\boldsymbol{u} \in \mathcal{V}$ the **negative** $-\boldsymbol{u}$ of $\boldsymbol{u}$ is unique.
3. $0\boldsymbol{u} = \mathbf{0}$, $c\mathbf{0} = \mathbf{0}$, and $-\boldsymbol{u} = (-1)\boldsymbol{u}$.

Here are some examples

1. The space $\mathbb{R}^n$, where $n \in \mathbb{N}$, is a real vector space.
2. Similarly, $\mathbb{C}^n$ is a complex vector space.
3. Let $\mathcal{D}$ be a subset of $\mathbb{R}$ and $d \in \mathbb{N}$. The set $\mathcal{V}$ of all functions $\boldsymbol{f}, \boldsymbol{g} : \mathcal{D} \to \mathbb{R}^d$ is a real vector space with

$$(\boldsymbol{f} + \boldsymbol{g})(t) := \boldsymbol{f}(t) + \boldsymbol{g}(t), \quad (c\boldsymbol{f})(t) := c\boldsymbol{f}(t), \quad t \in \mathcal{D}, \quad c \in \mathbb{R}.$$

   Two functions $\boldsymbol{f}, \boldsymbol{g}$ in $\mathcal{V}$ are equal if $\boldsymbol{f}(t) = \boldsymbol{g}(t)$ for all $t \in \mathcal{D}$. The zero element is the **zero function** given by $\boldsymbol{f}(t) = \mathbf{0}$ for all $t \in \mathcal{D}$ and the negative of $\boldsymbol{f}$ is given by $-\boldsymbol{f} = (-1)\boldsymbol{f}$. In the following we will use boldface letters for functions only if $d > 1$.

4. For $n \geq 0$ the space Π_n of polynomials of degree at most n consists of all polynomials $p : \mathbb{R} \to \mathbb{R}$, $p : \mathbb{R} \to \mathbb{C}$, or $p : \mathbb{C} \to \mathbb{C}$ of the form

$$p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n, \tag{2}$$

where the coefficients $a_0, \ldots, a_n$ are real or complex numbers. p is called the **zero polynomial** if all coefficients are zero. All other polynomials are said to be **nontrivial**. The **degree** of a nontrivial polynomial p given by (2) is the smallest integer $0 \leq k \leq n$ such that $p(t) = a_0 + \cdots + a_k t^k$ with $a_k \neq 0$. The degree of the zero polynomial is not defined. Π_n is a vector space if we define addition and scalar multiplication as for functions.

Definition 0.2 (Linear combination)
For $n \geq 1$ let $\mathcal{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of vectors in a vector space $\mathcal{V}$ and let $c_1, \ldots, c_n$ be scalars.

1. *The sum $c_1\boldsymbol{x}_1 + \cdots + c_n\boldsymbol{x}_n$ is called a* **linear combination** *of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.*
2. *The linear combination is* **nontrivial** *if $c_j\boldsymbol{x}_j \neq \boldsymbol{0}$ for at least one j.*
3. *The set of all linear combinations of elements in $\mathcal{X}$ is denoted* $\operatorname{span}(\mathcal{X})$.
4. *A vector space is* **finite dimensional** *if it has a finite spanning set; i. e., there exists $n \in \mathbb{N}$ and $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\mathcal{V}$ such that $\mathcal{V} = \operatorname{span}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})$.*

Example 0.3 (Linear combinations)

1. *Any $\boldsymbol{x} = [x_1, \ldots, x_m]^T$ in $\mathbb{C}^m$ can be written as a linear combination of the unit vectors as $\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \cdots + x_m\boldsymbol{e}_m$. Thus, $\mathbb{C}^m = \operatorname{span}(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\})$ and $\mathbb{C}^m$ is finite dimensional. Similarly $\mathbb{R}^m$ is finite dimensional.*
2. *Let $\Pi = \cup_n \Pi_n$ be the space of all polynomials. Π is a vector space that is not finite dimensional. For suppose Π is finite dimensional. Then $\Pi = \operatorname{span}(\{p_1, \ldots, p_m\})$ for some polynomials $p_1, \ldots, p_m$. Let d be an integer such that the degree of p_j is less than d for $j = 1, \ldots, m$. A polynomial of degree d cannot be written as a linear combination of $p_1, \ldots, p_m$, a contradiction.*

0.2.1 Linear independence and bases

Definition 0.4 (Linear independence)
A set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of nonzero vectors in a vector space is **linearly dependent** *if $\boldsymbol{0}$ can be written as a nontrivial linear combination of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Otherwise $\mathcal{X}$ is* **linearly independent**.

A set of vectors $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is linearly independent if and only if

$$c_1\boldsymbol{x}_1 + \cdots + c_n\boldsymbol{x}_n = \mathbf{0} \qquad \Longrightarrow \qquad c_1 = \cdots = c_n = 0. \tag{3}$$

Suppose $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is linearly independent. Then

1. If $\boldsymbol{x} \in \text{span}(\mathcal{X})$ then the scalars $c_1, \ldots, c_n$ in the representation $\boldsymbol{x} = c_1\boldsymbol{x}_1 + \cdots + c_n\boldsymbol{x}_n$ are unique.
2. Any nontrivial linear combination of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is nonzero,

Lemma 0.5 (Linear independence and span)
Suppose $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ span a vector space $\mathcal{V}$ 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 with $i_2 \neq i_1$. 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 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ 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 0.6 (basis)
A finite set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ in a vector space $\mathcal{V}$ is a **basis** *for $\mathcal{V}$ 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 0.7 (Basis subset of a spanning set)
Suppose $\mathcal{V}$ is a vector space and that $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a spanning set for $\mathcal{V}$. Then we can find a subset $\{\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}\}$ 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 0.8 (Existence of a basis)
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 0.7, $\mathcal{V}$ 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 definition a finite spanning set. □

Theorem 0.9 (Dimension of a vector space)
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 0.5 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 $\mathbb{C}^n$.

Theorem 0.10 (Enlarging vectors to a basis)
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. □

0.2.2 Subspaces

Definition 0.11 (Subspace)
A nonempty subset $\mathcal{S}$ of a real or complex vector space $\mathcal{V}$ is called a **subspace** *of $\mathcal{V}$ if*

(V1) *The sum $\boldsymbol{u} + \boldsymbol{v}$ is in $\mathcal{S}$ for any $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}$.*

(S1) *The scalar multiple $c\boldsymbol{u}$ is in $\mathcal{S}$ for any scalar c and any $\boldsymbol{u} \in \mathcal{S}$.*

Using the operations in $\mathcal{V}$, any subspace $\mathcal{S}$ of $\mathcal{V}$ is a vector space, i. e., all 10 axioms $V1 - V5$ and $S1 - S5$ are satisfied for $\mathcal{S}$. In particular, $\mathcal{S}$ must contain the zero element in $\mathcal{V}$. This follows since the operations of vector addition and scalar multiplication are inherited from $\mathcal{V}$.

Example 0.12 (Examples of subspaces)

1. $\{\mathbf{0}\}$, *where* $\mathbf{0}$ *is the zero vector is a subspace, the* **trivial subspace**. *The dimension of the trivial subspace is defined to be zero. All other subspaces are* **nontrivial**.

2. $\mathcal{V}$ *is a subspace of itself.*

3. $\text{span}(\mathcal{X})$ *is a subspace of* $\mathcal{V}$ *for any* $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathcal{V}$. *Indeed, it is easy to see that* **(V1)** *and* **(S1)** *hold.*

4. *The* **sum** *of two subspaces* $\mathcal{R}$ *and* $\mathcal{S}$ *of a vector space* $\mathcal{V}$ *is defined by*

$$\mathcal{R} + \mathcal{S} := \{\boldsymbol{r} + \boldsymbol{s} : \boldsymbol{r} \in \mathcal{R} \text{ and } \boldsymbol{s} \in \mathcal{S}\}. \tag{4}$$

   Clearly **(V1)** *and* **(S1)** *hold and it is a subspace of* $\mathcal{V}$.

5. *The* **intersection** *of two subspaces* $\mathcal{R}$ *and* $\mathcal{S}$ *of a vector space* $\mathcal{V}$ *is defined by*

$$\mathcal{R} \cap \mathcal{S} := \{\boldsymbol{x} : \boldsymbol{x} \in \mathcal{R} \text{ and } \boldsymbol{x} \in \mathcal{S}\}. \tag{5}$$

   It is a subspace of $\mathcal{V}$.

6. *The* **union** *of two subspaces* $\mathcal{R}$ *and* $\mathcal{S}$ *of a vector space* $\mathcal{V}$ *is defined by*

$$\mathcal{R} \cup \mathcal{S} := \{\boldsymbol{x} : \boldsymbol{x} \in \mathcal{R} \text{ or } \boldsymbol{x} \in \mathcal{S}\}. \tag{6}$$

   In general it is not a subspace of $\mathcal{V}$.

7. *A sum of two subspaces* $\mathcal{R}$ *and* $\mathcal{S}$ *of a vector space* $\mathcal{V}$ *is called a* **direct sum** *and denoted* $\mathcal{R} \oplus \mathcal{S}$ *if* $\mathcal{R} \cap \mathcal{S} = \{\mathbf{0}\}$. *The subspaces* $\mathcal{R}$ *and* $\mathcal{S}$ *are called* **complementary** *in the subspace* $\mathcal{R} \oplus \mathcal{S}$.

Theorem 0.13 (Dimension formula for sums of subspaces)
Let $\mathcal{R}$ *and* $\mathcal{S}$ *be two finite dimensional subspaces of a vector space* $\mathcal{V}$. *Then*

$$\dim(\mathcal{R} + \mathcal{S}) = \dim(\mathcal{R}) + \dim(\mathcal{S}) - \dim(\mathcal{R} \cap \mathcal{S}). \tag{7}$$

In particular, for a direct sum

$$\dim(\mathcal{R} \oplus \mathcal{S}) = \dim(\mathcal{R}) + \dim(\mathcal{S}). \tag{8}$$

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 $\mathcal{R} \cap \mathcal{S} = \{\mathbf{0}\}$. We use Theorem 0.10 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 $\boldsymbol{u} + \boldsymbol{r} + \boldsymbol{s} = \mathbf{0}$, 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 $\boldsymbol{r} := \sum_{j=1}^{p} \beta_j \boldsymbol{u}_j$. 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 then 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 $\mathcal{R} + \mathcal{S}$. But then

$$\dim(\mathcal{R} + \mathcal{S}) = p + q + t = (p + q) + (p + t) - p = \dim(\mathcal{R}) + \dim(\mathcal{S}) - \dim(\mathcal{R} \cap \mathcal{S})$$

and (7) follows. (7) implies (8) since $\dim\{\boldsymbol{0}\} = 0$. □

It is convenient to introduce a matrix transforming a basis in a subspace into a basis for the space itself.

Lemma 0.14 (Change of basis matrix)
Suppose $\mathcal{S}$ is a subspace of a finite dimensional vector space $\mathcal{V}$ 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

$$\boldsymbol{s}_j = \sum_{i=1}^{m} a_{ij} \boldsymbol{v}_i \text{ for } j = 1, \ldots, n. \tag{9}$$

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 $\boldsymbol{A} = [a_{ij}] \in \mathbb{C}^{m \times n}$ is given by (9). The matrix $\boldsymbol{A}$ has linearly independent columns.

Proof. (9) holds for some a_{ij} since $\boldsymbol{s}_j \in \mathcal{V}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ 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

$$\sum_{i=1}^{m} b_i \boldsymbol{v}_i = \boldsymbol{x} = \sum_{j=1}^{n} c_j \boldsymbol{s}_j \overset{(9)}{=} \sum_{j=1}^{n} c_j \big(\sum_{i=1}^{m} a_{ij} \boldsymbol{v}_i \big) = \sum_{i=1}^{m} \big(\sum_{j=1}^{n} a_{ij} c_j \big) \boldsymbol{v}_i.$$

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 it follows that $\boldsymbol{c} = \boldsymbol{0}$. □

The matrix $\boldsymbol{A}$ in Lemma 0.14 is called a **change of basis matrix**.

0.2.3 The vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$

When $\mathcal{V} = \mathbb{R}^m$ we can think of n vectors in $\mathbb{R}^m$, say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, as a set $\mathcal{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ or as the columns of a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m \times n}$. A linear combination can then be written as a matrix times vector $\boldsymbol{X}\boldsymbol{c}$, where $\boldsymbol{c} = [c_1, \ldots, c_n]^T$ is the vector of scalars. Thus

$$\operatorname{span}(\mathcal{X}) = \operatorname{span}(\boldsymbol{X}) = \{\boldsymbol{X}\boldsymbol{c} : \boldsymbol{c} \in \mathbb{R}^n\}.$$

Of course the same holds for $\mathbb{C}^m$.

In $\mathbb{R}^m$ and $\mathbb{C}^m$ each of the following statements is equivalent to linear independence of $\mathcal{X}$.

(i) $\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0} \Rightarrow \boldsymbol{c} = \boldsymbol{0}$,

(ii) $\boldsymbol{X}$ has linearly independent columns,

Definition 0.15 (Column space and null space)
Associated with a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m \times n}$ are the following subspaces

1. *The subspace* $\operatorname{span}(\boldsymbol{X})$ *is called the* **column space** *of $\boldsymbol{X}$. It is the smallest subspace containing $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.*
2. $\operatorname{span}(\boldsymbol{X}^T)$ *is called the* **row space** *of $\boldsymbol{X}$. It is generated by the rows of $\boldsymbol{X}$ written as column vectors.*
3. *The subspace* $\ker(\boldsymbol{X}) := \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{X}\boldsymbol{y} = \boldsymbol{0}\}$ *is called the* **null space** *or* **kernel space** *of $\boldsymbol{X}$.*

Note that the subspace $\ker(\boldsymbol{X})$ is nontrivial if and only if $\mathcal{X}$ is linearly dependent.

0.3 Vector Norms

To measure the size of a vector we use norms.

Definition 0.16 (Vector norm)
A **(vector) norm** *in a real (resp. complex) vector space $\mathcal{V}$ is a function $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ that satisfies for all $\boldsymbol{x}, \boldsymbol{y}$ in $\mathcal{V}$ and all a in $\mathbb{R}$ (resp. $\mathbb{C}$)*

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 triple $(\mathcal{V}, \mathbb{R}, \|\cdot\|)$ *(resp.* $(\mathcal{V}, \mathbb{C}, \|\cdot\|)$*) is called a* **normed vector space** *and the inequality 3. is called the* **triangle inequality**.

In this book we will use the following family of vector norms on $\mathcal{V} = \mathbb{C}^n$ and $\mathcal{V} = \mathbb{R}^n$.


Otto Ludwig Hölder, 1859-1937 (left), Hermann Minkowski, 1864-1909 (right).

Definition 0.17 (Vector p-norms)
We define for $p \geq 1$ *and* $\boldsymbol{x} \in \mathbb{R}^n$ *or* $\boldsymbol{x} \in \mathbb{C}^n$ *the* p**-norms** *by*

$$\|\boldsymbol{x}\|_p := \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}, \tag{10}$$

$$\|\boldsymbol{x}\|_\infty := \max_{1 \leq j \leq n} |x_j|. \tag{11}$$

The most important cases are $p = 1, 2, \infty$*:*

1. $\|\boldsymbol{x}\|_1 = \sum_{j=1}^{n} |x_j|$, *(***the one-norm** *or* l_1**-norm***)*
2. $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^{n} |x_j|^2}$, *(***the two-norm**, l_2**-norm**, *or* **Euclidian norm***)*
3. $\|\boldsymbol{x}\|_\infty = \max_{1 \leq j \leq n} |x_j|$, *(***the infinity-norm**, l_∞**-norm**, *or* **max norm***)*

Some remarks are in order.

1. That the Euclidian norm is a vector norm follows from Theorem 0.23. In Section 7.3, we show that the p-norms are vector norms for $1 \leq p \leq \infty$.
2. The triangle inequality $\|\boldsymbol{x} + \boldsymbol{y}\|_p \leq \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$ is called **Minkowski's inequality**.

3. To prove it one first establishes **Hölder's inequality**

$$\sum_{j=1}^{n}|x_j y_j| \le \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q, \quad \frac{1}{p}+\frac{1}{q}=1, \quad \boldsymbol{x},\boldsymbol{y}\in\mathbb{C}^n. \tag{12}$$

The relation $\frac{1}{p}+\frac{1}{q}=1$ means that if $p=1$ then $q=\infty$ and if $p=2$ then $q=2$ and Hölder's inequality is the same as the Cauchy-Schwarz inequality (cf. Theorem 0.22) for the Euclidian norm.

4. The infinity norm is related to the other p-norms by

$$\lim_{p\to\infty}\|\boldsymbol{x}\|_p = \|\boldsymbol{x}\|_\infty \text{ for all } \boldsymbol{x}\in\mathbb{C}^n. \tag{13}$$

5. The equation (13) 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

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_p \le n^{1/p}\|\boldsymbol{x}\|_\infty, \quad p\ge 1. \tag{14}$$

Since $\lim_{p\to\infty} n^{1/p}=1$ for any $n\in\mathbb{N}$ we see that (13) follows.

We return now to the general case.

Definition 0.18 (Equivalent norms)
We say that two norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathcal{V}$ are **equivalent** *if there are positive constants m and M such that for all vectors $\boldsymbol{x}\in\mathcal{V}$ we have*

$$m\|\boldsymbol{x}\|' \le \|\boldsymbol{x}\| \le M\|\boldsymbol{x}\|'. \tag{15}$$

By (14) the p- and ∞-norms are equivalent for any $p\ge 1$. This result is generalized in the following theorem.

Theorem 0.19 (Basic properties of vector norms)
The following holds for a normed vector space $(\mathcal{V},\mathbb{C},\|\cdot\|)$.

1. $\|\boldsymbol{x}-\boldsymbol{y}\| \ge |\,\|\boldsymbol{x}\|-\|\boldsymbol{y}\|\,|$, *for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{C}^n$ (inverse triangle inequality).*
2. *The vector norm is a continuous function $\mathcal{V}\to\mathbb{R}$.*
3. *All vector norms on $\mathcal{V}$ are equivalent provided $\mathcal{V}$ is finite dimensional.*

Proof.

1. 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.

2. This follows from the inverse triangle inequality.

3. The following proof can be skipped by those who do not have the necessary background in advanced calculus. Define the $\|\cdot\|'$ unit sphere

$$\mathcal{S} := \{\boldsymbol{y} \in \mathcal{V} : \|\boldsymbol{y}\|' = 1\}.$$

The set $\mathcal{S}$ is a closed and bounded set and the function $f : \mathcal{S} \to \mathbb{R}$ given by $f(\boldsymbol{y}) = \|\boldsymbol{y}\|$ is continuous by what we just showed. Therefore f attains its minimum and maximum value on $\mathcal{S}$. Thus, there are positive constants m and M such that

$$m \leq \|\boldsymbol{y}\| \leq M, \quad \boldsymbol{y} \in \mathcal{S}. \tag{16}$$

For any $\boldsymbol{x} \in \mathcal{V}$ one has $\boldsymbol{y} := \boldsymbol{x}/\|\boldsymbol{x}\|' \in \mathcal{S}$, and (15) follows if we apply (16) to these $\boldsymbol{y}$.

□

0.4 Inner Products

An **inner product** or **scalar product** in a vector space is a function mapping pairs of vectors into a scalar.

0.4.1 Real and complex inner products

We consider first the real case.

Definition 0.20 (Real inner product)
An **inner product** *in a real vector space $\mathcal{V}$ is a function $\langle\cdot,\cdot\rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ 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 pair $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ is called a **real inner product space**. *The function*

$$\|\cdot\| : \mathcal{V} \longrightarrow \mathbb{R}, \qquad \boldsymbol{x} \longmapsto \|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} \tag{17}$$

is called the **inner product norm**.

The **standard inner product in** $\mathcal{V} = \mathbb{R}^n$ is given by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^T \boldsymbol{y} = \boldsymbol{y}^T \boldsymbol{x}$. It is clearly an inner product in $\mathbb{R}^n$. The corresponding inner product norm is the Euclidian norm $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \|\boldsymbol{x}\|_2$.

Consider next inner products in a complex vector space. Property 2. in the definition of a real inner product is altered from symmetry to skew symmetry.

Definition 0.21 (Complex inner product)
An **inner product** *in a complex vector space* $\mathcal{V}$ *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 pair $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ *is called a* **complex inner product space** *The function*

$$\|\cdot\| : \mathcal{V} \to \mathbb{R}, \qquad \boldsymbol{x} \longmapsto \|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \tag{18}$$

is called the **inner product norm**.

Note the complex conjugate in 2. We find

$$\langle \boldsymbol{x}, a\boldsymbol{y} + b\boldsymbol{z} \rangle = \overline{a}\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \overline{b}\langle \boldsymbol{x}, \boldsymbol{z} \rangle, \quad \langle a\boldsymbol{x}, a\boldsymbol{y} \rangle = |a|^2 \langle \boldsymbol{x}, \boldsymbol{y} \rangle. \tag{19}$$

The **standard inner product in** $\mathbb{C}^n$ is given by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{y}^* \boldsymbol{x} = \boldsymbol{x}^T \overline{\boldsymbol{y}} = \sum_{j=1}^{n} x_j \overline{y_j}.$$

It is clearly an inner product in $\mathbb{C}^n$. The corresponding inner product norm is the Euclidian norm $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^* \boldsymbol{x}}$.



Viktor Yakovlevich Bunyakovsky, 1804-1889 (left), Augustin-Louis Cauchy, 1789-1857 (center), Karl Hermann Amandus Schwarz,1843-1921 (right). The name Bunyakovsky is also associated with the Cauchy-Schwarz inequality.

The following inequality holds for any inner product.

Theorem 0.22 (Cauchy-Schwarz inequality)
For any $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|, \tag{20}$$

with equality if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent.

Proof. If $\boldsymbol{y} = \boldsymbol{0}$ then $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, 0\boldsymbol{y} \rangle = 0\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ and $\|\boldsymbol{y}\| = 0$. Thus the inequality holds with equality, and $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent. So assume $\boldsymbol{y} \neq \boldsymbol{0}$. Define

$$\boldsymbol{z} := \boldsymbol{x} - a\boldsymbol{y}, \quad a := \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle}.$$

Then $\langle \boldsymbol{z}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle - a\langle \boldsymbol{y}, \boldsymbol{y} \rangle = 0$ so that by 2. and (19)

$$\langle a\boldsymbol{y}, \boldsymbol{z} \rangle + \langle \boldsymbol{z}, a\boldsymbol{y} \rangle = a\overline{\langle \boldsymbol{z}, \boldsymbol{y} \rangle} + \overline{a}\langle \boldsymbol{z}, \boldsymbol{y} \rangle = 0. \tag{21}$$

But then

$$\begin{aligned} \|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle &= \langle \boldsymbol{z} + a\boldsymbol{y}, \boldsymbol{z} + a\boldsymbol{y} \rangle \\ &\overset{(21)}{=} \langle \boldsymbol{z}, \boldsymbol{z} \rangle + \langle a\boldsymbol{y}, a\boldsymbol{y} \rangle \overset{(19)}{=} \|\boldsymbol{z}\|^2 + |a|^2 \|\boldsymbol{y}\|^2 \\ &\geq |a|^2 \|\boldsymbol{y}\|^2 = \frac{|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2}{\|\boldsymbol{y}\|^2}. \end{aligned}$$

Multiplying by $\|\boldsymbol{y}\|^2$ gives (20). We have equality if and only if $\boldsymbol{z} = \boldsymbol{0}$, which means that $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent. □

Theorem 0.23 (Inner product norm)
The inner product norm is a vector norm.

Proof. For all $\boldsymbol{x}, \boldsymbol{y}$ in an inner product space and all a in $\mathbb{C}$ we need to show

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 first statement is an immediate consequence of positivity, while the second one follows from (19). Expanding $\|\boldsymbol{x} + a\boldsymbol{y}\|^2 = \langle \boldsymbol{x} + a\boldsymbol{y}, \boldsymbol{x} + a\boldsymbol{y} \rangle$ using (19) we obtain

$$\|\boldsymbol{x} + a\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + a\langle \boldsymbol{y}, \boldsymbol{x} \rangle + \overline{a}\langle \boldsymbol{x}, \boldsymbol{y} \rangle + |a|^2 \|\boldsymbol{y}\|^2, \quad a \in \mathbb{C}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}. \tag{22}$$

Now (22) with $a = 1$ and the Cauchy-Schwarz inequality implies

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq \|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2 = (\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2.$$

Taking square roots completes the proof. □

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

$$\cos\theta = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}. \tag{23}$$

This defines the **angle** between vectors in a real inner product space.

Exercise 0.24 (The $\boldsymbol{A}^T\boldsymbol{A}$ inner product)
Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ 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 0.25 (Angle between vectors in complex case)
Show that in the complex case there is a unique angle θ in $[0, \pi/2]$ such that

$$\cos\theta = \frac{|\langle \boldsymbol{x}, \boldsymbol{y} \rangle|}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}. \tag{24}$$

0.4.2 Orthogonality

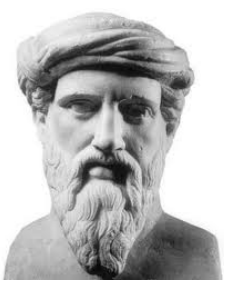
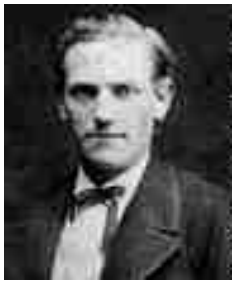
Definition 0.26 (Orthogonality)
Two vectors $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space are **orthogonal** *or* **perpendicular**, *denoted as $\boldsymbol{x} \perp \boldsymbol{y}$, if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$. The vectors are* **orthonormal** *if in addition $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$.*

From the definitions (23), (24) of angle θ between two vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$ it follows that $\boldsymbol{x} \perp \boldsymbol{y}$ if and only if $\theta = \pi/2$.

Theorem 0.27 (Pythagoras)
For a real or complex inner product space

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2, \quad \textit{if} \quad \boldsymbol{x} \perp \boldsymbol{y}. \tag{25}$$

Proof. We set $a = 1$ in (22) and use the orthogonality. □


Pythagoras of Samos, BC 570-BC 495 (left), Jørgen Pedersen Gram, 1850-1916 (center), Erhard Schmidt,1876-1959 (right).

Definition 0.28 (Orthogonal- and orthonormal bases)
A set of nonzero vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ *in a subspace* $\mathcal{S}$ *of a real or complex inner product space is 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 a subspace of an inner product space can be turned into an orthogonal- or orthonormal basis for the subspace by the following construction.

Theorem 0.29 (Gram-Schmidt)
Let $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ *be a basis for a real or complex inner product space* $(\mathcal{S}, \langle \cdot, \cdot \rangle)$*. Define*

$$\boldsymbol{v}_1 := \boldsymbol{s}_1, \quad \boldsymbol{v}_j := \boldsymbol{s}_j - \sum_{i=1}^{j-1} \frac{\langle \boldsymbol{s}_j, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i, \quad j = 2, \ldots, k. \tag{26}$$

Then $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ *is an orthogonal basis for* $\mathcal{S}$ *and the normalized vectors*

$$\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\} := \left\{ \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}, \ldots, \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|} \right\}$$

form 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. Define subspaces $S_j := \text{span}\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_j\}$ 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 (26) as a linear combination of $\boldsymbol{s}_j$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{j-1}$. Now each of these $\boldsymbol{v}_i$ is a linear combination of $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_i$, and 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 $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_j$

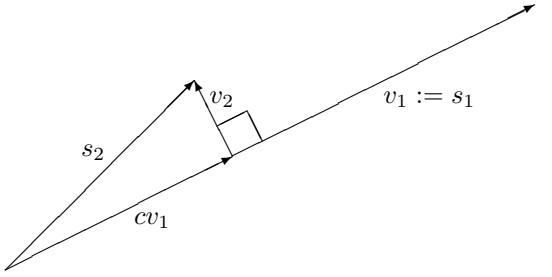


Figure 1. *The construction of* $\boldsymbol{v}_1$ *and* $\boldsymbol{v}_2$ *in Gram-Schmidt. The constant* c *is given by* $c := \langle \boldsymbol{s}_2, \boldsymbol{v}_1 \rangle / \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle$.

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 $\mathcal{S}$. □

Sometimes we want to extend an orthogonal basis for a subspace to an orthogonal basis for a larger space.

Theorem 0.30 (Orthogonal Extension of basis)
Suppose $\mathcal{S} \subset \mathcal{T}$ *are finite dimensional subspaces of a vector space* $\mathcal{V}$. *An orthogonal basis for* $\mathcal{S}$ *can always be extended to an orthogonal basis for* $\mathcal{T}$.

Proof. Suppose $\dim \mathcal{S} := k < \dim \mathcal{T} = n$. Using Theorem 0.10 we first extend an orthogonal basis $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$ for $\mathcal{S}$ to a basis $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k, \boldsymbol{s}_{k+1}, \ldots, \boldsymbol{s}_n$ for $\mathcal{T}$, and then apply the Gram-Schmidt process to this basis obtaining an orthogonal basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for $\mathcal{T}$. This is an extension of the basis for $\mathcal{S}$ since $\boldsymbol{v}_i = \boldsymbol{s}_i$ for $i = 1, \ldots, k$. We show this by induction. Clearly $\boldsymbol{v}_1 = \boldsymbol{s}_1$. Suppose for some $2 \leq r < k$ that $\boldsymbol{v}_j = \boldsymbol{s}_j$ for $j = 1, \ldots, r-1$. Consider (26) for $j = r$. Since $\langle \boldsymbol{s}_r, \boldsymbol{v}_i \rangle = \langle \boldsymbol{s}_r, \boldsymbol{s}_i \rangle = 0$ for $i < r$ we obtain $\boldsymbol{v}_r = \boldsymbol{s}_r$. □

Letting $\mathcal{S} = \operatorname{span}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ and $\mathcal{T}$ be $\mathbb{R}^n$ or $\mathbb{C}^n$ we obtain

Corollary 0.31 (Extending orthogonal vectors to a basis)
For $1 \leq k < n$ a set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ of nonzero orthogonal vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$ can be extended to an orthogonal basis for the whole space.

0.5 Linear Systems

Consider a linear system

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

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 $\boldsymbol{a}_j = \begin{bmatrix} a_{1j}, \ldots, \boldsymbol{a}_{mj} \end{bmatrix}^T \in \mathbb{C}^m$ for $j = 1, \ldots, n$ and $\boldsymbol{b} = \begin{bmatrix} b_1, \ldots, b_m \end{bmatrix}^T \in \mathbb{C}^m$. It can also be written as a matrix equation

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \boldsymbol{b}.$$

The system is **homogeneous** if $\boldsymbol{b} = \boldsymbol{0}$ and it is said to be **underdetermined**, **square**, or **overdetermined** if $m < n$, $m = n$, or $m > n$, respectively.

0.5.1 Basic properties

A linear system has a unique solution, infinitely many solutions, or no solution. To discuss this we first consider the real case, and a homogeneous underdetermined system.

Lemma 0.32 (Underdetermined system)
Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m < n$. Then there is a nonzero $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$.

Proof. Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m < n$. The n columns of $\boldsymbol{A}$ span a subspace of $\mathbb{R}^m$. Since $\mathbb{R}^m$ has dimension m the dimension of this subspace is at most m. By

Lemma 0.5 the columns of $\boldsymbol{A}$ must be linearly dependent. It follows that there is a nonzero $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{Ax} = \boldsymbol{0}$. ∎

A square matrix is either **nonsingular** or **singular**.

Definition 0.33 (Real nonsingular or singular matrix)
A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is said to be **nonsingular** *if the only real solution of the homogeneous system $\boldsymbol{Ax} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$. The matrix is* **singular** *if there is a nonzero $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{Ax} = \boldsymbol{0}$.*

Theorem 0.34 (Linear systems; existence and uniqueness)
Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. The linear system $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} \in \mathbb{R}^n$ for any $\boldsymbol{b} \in \mathbb{R}^n$ if and only if the matrix $\boldsymbol{A}$ is nonsingular.

Proof. Suppose $\boldsymbol{A}$ is nonsingular. We define $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{A}\ \boldsymbol{b}\end{bmatrix} \in \mathbb{R}^{n\times(n+1)}$ by adding a column to $\boldsymbol{A}$. By Lemma 0.32 there is a nonzero $\boldsymbol{z} \in \mathbb{R}^{n+1}$ such that $\boldsymbol{Bz} = \boldsymbol{0}$. If we write $\boldsymbol{z} = \begin{bmatrix}\tilde{\boldsymbol{z}}\\ z_{n+1}\end{bmatrix}$ where $\tilde{\boldsymbol{z}} = \begin{bmatrix}z_1, \ldots, z_n\end{bmatrix}^T \in \mathbb{R}^n$ and $z_{n+1} \in \mathbb{R}$, 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 nonsingularity of $\boldsymbol{A}$. Define $\boldsymbol{x} := -\tilde{\boldsymbol{z}}/z_{n+1}$. Then

$$\boldsymbol{Ax} = -\boldsymbol{A}\left(\frac{\tilde{\boldsymbol{z}}}{z_{n+1}}\right) = -\frac{1}{z_{n+1}}\boldsymbol{A}\tilde{\boldsymbol{z}} = -\frac{1}{z_{n+1}}\big(-z_{n+1}\boldsymbol{b}\big) = \boldsymbol{b},$$

so $\boldsymbol{x}$ is a solution.

Suppose $\boldsymbol{Ax} = \boldsymbol{b}$ and $\boldsymbol{Ay} = \boldsymbol{b}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{0}$ and since $\boldsymbol{A}$ is nonsingular we conclude that $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0}$ or $\boldsymbol{x} = \boldsymbol{y}$. Thus the solution is unique.

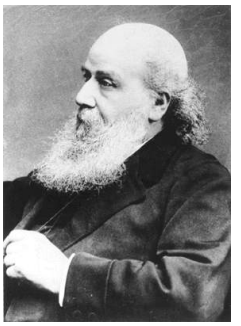
Conversely, if $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution for any $\boldsymbol{b} \in \mathbb{R}^n$ then $\boldsymbol{Ax} = \boldsymbol{0}$ has a unique solution which must be $\boldsymbol{x} = \boldsymbol{0}$. Thus $\boldsymbol{A}$ is nonsingular. ∎

For the complex case we have

Lemma 0.35 (Complex underdetermined system)
Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ with $m < n$. Then there is a nonzero $\boldsymbol{x} \in \mathbb{C}^n$ such that $\boldsymbol{Ax} = \boldsymbol{0}$.

Definition 0.36 (Complex nonsingular matrix)
A square matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is said to be **nonsingular** *if the only complex solution of the homogeneous system $\boldsymbol{Ax} = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$. The matrix is* **singular** *if it is not nonsingular.*

Theorem 0.37 (Complex linear system; existence and uniqueness)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. The linear system $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} \in \mathbb{C}^n$ for any $\boldsymbol{b} \in \mathbb{C}^n$ if and only if the matrix $\boldsymbol{A}$ is nonsingular.



James Joseph Sylvester, 1814-1897. The word matrix to denote a rectangular array of numbers, was first used by Sylvester in 1850.

0.5.2 The inverse matrix

Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is a square matrix. A matrix $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ is called a **right inverse** of $\boldsymbol{A}$ if $\boldsymbol{AB} = \boldsymbol{I}$. A matrix $\boldsymbol{C} \in \mathbb{C}^{n\times n}$ 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 by $\boldsymbol{A}^{-1}$. 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.

Theorem 0.38 (Product of nonsingular matrices)
If $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^{n\times n}$ with $\boldsymbol{AB} = \boldsymbol{C}$ then $\boldsymbol{C}$ is nonsingular if and only if both $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular. In particular, if $\boldsymbol{AB} = \boldsymbol{I}$ or $\boldsymbol{BA} = \boldsymbol{I}$ then $\boldsymbol{A}$ is nonsingular and $\boldsymbol{A}^{-1} = \boldsymbol{B}$.

Proof. Suppose both $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular and let $\boldsymbol{Cx} = \boldsymbol{0}$. Then $\boldsymbol{ABx} = \boldsymbol{0}$ and since $\boldsymbol{A}$ is nonsingular we see that $\boldsymbol{Bx} = \boldsymbol{0}$. Since $\boldsymbol{B}$ is nonsingular we have $\boldsymbol{x} = \boldsymbol{0}$. We conclude that $\boldsymbol{C}$ is nonsingular.

For the converse suppose first that $\boldsymbol{B}$ is singular and let $\boldsymbol{x} \in \mathbb{C}^n$ 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 nonsingular, but $\boldsymbol{A}$ is singular. Let $\tilde{\boldsymbol{x}}$ be a nonzero

vector such that $\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{0}$. By Theorem 0.37 there is a vector $\boldsymbol{x}$ such that $\boldsymbol{B}\boldsymbol{x} = \tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ is nonzero since $\tilde{\boldsymbol{x}}$ is nonzero. But then $\boldsymbol{C}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{B})\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{x}) = \boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{0}$ for a nonzero vector $\boldsymbol{x}$ and $\boldsymbol{C}$ is singular. □

Theorem 0.39 (When is a square matrix invertible?)
A square matrix is invertible if and only if it is nonsingular.

Proof. Suppose first $\boldsymbol{A}$ is a nonsingular matrix. By Theorem 0.37 each of the linear systems $\boldsymbol{A}\boldsymbol{b}_i = \boldsymbol{e}_i$ has a unique solution $\boldsymbol{b}_i$ for $i = 1, \ldots, n$. Let $\boldsymbol{B} = \big[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\big]$. Then $\boldsymbol{A}\boldsymbol{B} = \big[\boldsymbol{A}\boldsymbol{b}_1, \ldots, \boldsymbol{A}\boldsymbol{b}_n\big] = \big[\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\big] = \boldsymbol{I}$ so that $\boldsymbol{A}$ has a right inverse $\boldsymbol{B}$. By Theorem 0.38 $\boldsymbol{B}$ is nonsingular since $\boldsymbol{I}$ is nonsingular and $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}$. Since $\boldsymbol{B}$ is nonsingular we can use what we have shown for $\boldsymbol{A}$ to conclude that $\boldsymbol{B}$ has a right inverse $\boldsymbol{C}$, i.e. $\boldsymbol{B}\boldsymbol{C} = \boldsymbol{I}$. But then $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{C} = \boldsymbol{I}$ so $\boldsymbol{B}$ has both a right inverse and a left inverse which must be equal so $\boldsymbol{A} = \boldsymbol{C}$. Since $\boldsymbol{B}\boldsymbol{C} = \boldsymbol{I}$ we have $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}$, 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 $\boldsymbol{B}$. Since $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}$ and $\boldsymbol{I}$ is nonsingular, we again use Theorem 0.38 to conclude that $\boldsymbol{A}$ is nonsingular. □

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 or a right inverse of $\boldsymbol{A}$. This calculation also proves that $\boldsymbol{A}$ is nonsingular. We use this observation to give simple proofs of the following results.

Corollary 0.40 (Basic properties of the inverse matrix)
Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ are nonsingular and c is a nonzero constant.

1. $\boldsymbol{A}^{-1}$ *is nonsingular and* $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
2. $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$ *is nonsingular and* $\boldsymbol{C}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
3. $\boldsymbol{A}^T$ *is nonsingular and* $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T =: \boldsymbol{A}^{-T}$.
4. $\boldsymbol{A}^*$ *is nonsingular and* $(\boldsymbol{A}^*)^{-1} = (\boldsymbol{A}^{-1})^* =: \boldsymbol{A}^{-*}$.
5. $c\boldsymbol{A}$ *is nonsingular 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 nonsingular and $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
2. We note that $(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1})(\boldsymbol{A}\boldsymbol{B}) = \boldsymbol{B}^{-1}(\boldsymbol{A}^{-1}\boldsymbol{A})\boldsymbol{B} = \boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}$. Thus $\boldsymbol{A}\boldsymbol{B}$ 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 $\boldsymbol{A}^T$. The proof of part 4 is similar.

4. The matrix $\frac{1}{c}\boldsymbol{A}^{-1}$ is a one sided inverse of $c\boldsymbol{A}$.

□

Exercise 0.41 (The inverse of a general 2×2 matrix)
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 0.42 (The inverse of a special 2×2 matrix)
Find the inverse of

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Exercise 0.43 (Sherman-Morrison formula)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, and $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n\times m}$ for some $n, m \in \mathbb{N}$. If $(\boldsymbol{I} + \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}$ exists then

$$(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I} + \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}^T\boldsymbol{A}^{-1}.$$

0.6 Determinants

Determinants, denoted by $\det(\cdot)$ or $|\cdot|$,are useful for studying eigenvalues. Recall that if $\boldsymbol{A}, \boldsymbol{B}$ are square matrices of order n with real or complex elements, then (see Appendix A for proofs)

1. $\det(\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
2. If $\boldsymbol{A}$ is triangular then $\det(\boldsymbol{A}) = a_{11}a_{22}\cdots a_{nn}$. In particular, $\det(\boldsymbol{I}) = 1$.
3. $\det(\boldsymbol{A}^T) = \det(\boldsymbol{A})$, and $\det(\boldsymbol{A}^*) = \overline{\det(\boldsymbol{A})}$, (complex conjugate).
4. $\det(a\boldsymbol{A}) = a^n \det(\boldsymbol{A})$, for $a \in \mathbb{C}$.
5. $\boldsymbol{A}$ is singular if and only if $\det(\boldsymbol{A}) = 0$.
6. If $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{D} \\ \boldsymbol{0} & \boldsymbol{E} \end{bmatrix}$ for some square matrices $\boldsymbol{C}, \boldsymbol{E}$ then $\det(\boldsymbol{A}) = \det(\boldsymbol{C})\det(\boldsymbol{E})$.

7. **Cramer's rule** Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is nonsingular and $\boldsymbol{b} \in \mathbb{C}^n$. Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ be the unique solution of $\boldsymbol{Ax} = \boldsymbol{b}$. Then

$$x_j = \frac{\det(\boldsymbol{A}_j(\boldsymbol{b}))}{\det(\boldsymbol{A})}, \quad j = 1, 2, \ldots, n,$$

where $\boldsymbol{A}_j(\boldsymbol{b})$ denote the matrix obtained from $\boldsymbol{A}$ by replacing the jth column of $\boldsymbol{A}$ by $\boldsymbol{b}$.

8. **Adjoint**. Let $\boldsymbol{A}_{i,j}$ denote the submatrix of $\boldsymbol{A}$ obtained by deleting the ith row and jth column of $\boldsymbol{A}$. For $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $1 \le i, j \le n$ the determinant $\det(\boldsymbol{A}_{ij})$ is called the **cofactor** of a_{ij}. The matrix $\operatorname{adj}(\boldsymbol{A}) \in \mathbb{C}^{n\times n}$ with elements $\operatorname{adj}(\boldsymbol{A})_{i,j} = (-1)^{i+j} \det(\boldsymbol{A}_{j,i})$ is called the **adjoint** of $\boldsymbol{A}$.

9. **Adjoint formula for the inverse**. If $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is nonsingular then

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})} \operatorname{adj}(\boldsymbol{A}).$$

10. **Cofactor expansion**. For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have

$$\det(\boldsymbol{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij}) \text{ for } i = 1, \ldots, n, \tag{27}$$

$$\det(\boldsymbol{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij}) \text{ for } j = 1, \ldots, n. \tag{28}$$

Arthur Cayley, 1821-1895 (left), Gabriel Cramer 1704-1752 (center), Alexandre-Thophile Vandermonde,1735-1796 (right). The notation | | for determinants is due to Cayley 1841.

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.

Exercise 0.44 (Cramer's rule; special case)
Solve the following system by Cramers rule:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

Exercise 0.45 (Adjoint matrix; special case)
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,

$$adj(\boldsymbol{A})\boldsymbol{A} = \begin{bmatrix} 343 & 0 & 0 \\ 0 & 343 & 0 \\ 0 & 0 & 343 \end{bmatrix} = \det(\boldsymbol{A})\boldsymbol{I}.$$

Example 0.46 (Determinant equation for a straight line)
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.

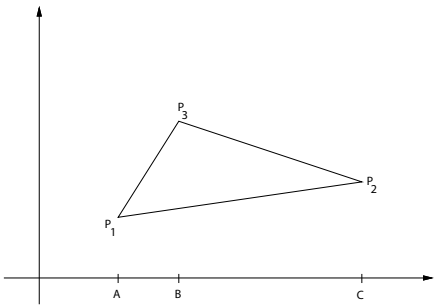


Figure 2. *The triangle T defined by the three points P_1, P_2 and P_3.*

Exercise 0.47 (Determinant equation for a plane)
Show that

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0.$$

is the equation for a plane through three points (x_1, y_1, z_1), (x_2, y_2, z_2) and (x_3, y_3, z_3) in space.

Exercise 0.48 (Signed area of a triangle)
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[1]

$$A(T) = \frac{1}{2} \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix}.$$

The area is positive if we traverse the vertices in counterclockwise order.

Exercise 0.49 (Vandermonde matrix)
Show that

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{i>j} (x_i - x_j),$$

[1] Hint: $A(T) = A(ABP_3P_1) + A(P_3BCP_2) - A(P_1ACP_2)$, c.f. Figure 2

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* Vandermonde determinant. [2]

Exercise 0.50 (Cauchy determinant (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* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *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 i*th 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

$$A_k(x) = \prod_{s\neq k}\left(\frac{\alpha_s - x}{\alpha_s - \alpha_k}\right), \qquad B_k(x) = \prod_{s\neq k}\left(\frac{\beta_s - x}{\beta_s - \beta_k}\right).$$

[2] Hint: subtract x_n^k times column k from column $k+1$ for $k = n-1, n-2, \ldots, 1$.

Exercise 0.51 (Inverse of the Hilbert matrix)
Let $\boldsymbol{H}_n = (h_{i,j})$ be the $n \times n$ matrix with elements $h_{i,j} = 1/(i+j-1)$. Use Exercise 0.50 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

$$f(i+1) = \left(\frac{i^2-n^2}{i^2}\right) f(i), \quad i = 1, 2, \ldots, \quad f(1) = -n.$$

0.7 Eigenpairs

Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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**.[3] The set of eigenvalues is called the **spectrum** of $\boldsymbol{A}$ and is denoted by $\sigma(\boldsymbol{A})$. For example, $\sigma(\boldsymbol{I}) = \{1, \ldots, 1\} = \{1\}$.

Lemma 0.52 (Characteristic equation)
For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have $\lambda \in \sigma(\boldsymbol{A}) \iff \det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$.

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 $\boldsymbol{A}$. □

The expression $\det(\boldsymbol{A} - \lambda\boldsymbol{I})$ is a polynomial of exact degree n in λ. 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

$$\begin{aligned} \det(\boldsymbol{A} - \lambda\boldsymbol{I}) &= (a_{11}-\lambda)\begin{vmatrix} a_{22}-\lambda & a_{23} \\ a_{32} & a_{33}-\lambda \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33}-\lambda \end{vmatrix} \\ &\quad + a_{31}\begin{vmatrix} a_{12} & a_{13} \\ a_{22}-\lambda & a_{23} \end{vmatrix} = (a_{11}-\lambda)(a_{22}-\lambda)(a_{33}-\lambda) + r(\lambda) \end{aligned}$$

[3] The word "eigen" is derived from German and means "own"

for some polynomial r of degree at most one. In general

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda) + r(\lambda), \tag{29}$$

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.

We observe that $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (-1)^n \det(\lambda \boldsymbol{I} - \boldsymbol{A})$ so $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$ if and only if $\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0$.

Definition 0.53 (Characteristic polynomial of a matrix)
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* $\boldsymbol{A}$. *The equation* $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$ *is called the* **characteristic equation** *of* $\boldsymbol{A}$.

By the fundamental theorem of algebra an $n \times n$ matrix has, counting multiplicities, 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 $\boldsymbol{A}\overline{\boldsymbol{x}} = \overline{\lambda}\overline{\boldsymbol{x}}$.

Using Property 6. of determinants we have an additional characterization of a singular matrix.

Theorem 0.54 (Zero eigenvalue)
The matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ *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. □

In general it is not easy to find all eigenvalues of a matrix. One notable exception is a triangular matrix. By Property 2. of determinants we obtain

Theorem 0.55 (Eigenvalues of a triangular matrix)
The eigenvalues of a triangular matrix are given by its diagonal elements.

0.8 Algorithms and Numerical Stability

In this text we consider mathematical problems (i. e., linear algebra problems) and many detailed numerical algorithms to solve them. Complexity is discussed briefly in Section 2.2.2. As for programming issues we often vectorize the algorithms leading to shorter and more efficient programs. Stability is important both for the mathematical problems and for the numerical algorithms. Stability can be studied in terms of perturbation theory leading to condition numbers, see

Chapters 7, 11, 12. We will often use phrases like "the algorithm is numerically stable" or "the algorithm is not numerically stable" without saying precisely what we mean by this. Loosely speaking, an algorithm is numerically stable if the solution, computed in floating point arithmetic, is the exact solution of a slightly perturbed problem. To determine upper bounds for these perturbations is the topic of **backward error analysis**. We give a rather limited introduction to floating point arithmetic and backward error analysis in Appendix B, but in the text we will not discuss this. This does not mean that numerical stability is not an important issue. In fact, numerical stability is crucial for a good algorithm. For thorough treatments of numerical stability issues we refer to the books [12] and [26, 27].

A list of freely available software for solving linear algebra problems can be found at

`http://www.netlib.org/utk/people/JackDongarra/la-sw.html`

Part I

Direct Methods for Linear Systems

Chapter 1

A Special Linear System

Consider a system of n linear equations in n unknowns. In component form the system can be written

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \quad \cdots \quad + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \quad \cdots \quad + a_{2n}x_n &= b_2, \\ \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \quad \cdots \quad + a_{nn}x_n &= b_n, \end{aligned}$$

and in matrix form

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \boldsymbol{b}.$$

The elements of $\boldsymbol{A}$ and $\boldsymbol{b}$ can be either real or complex numbers.

We recall (see Theorem 0.34) 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 also recall (cf. Theorem 0.39) that a square matrix has an inverse 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 $\boldsymbol{A}$.


Carl Friedrich Gauss, 1777-1855 (left), Myrick Hascall Doolittle, 1830-1911 (right).

Gaussian elimination with row interchanges is the classical method for solving n linear equations in n unknowns[4]. After a introductory and elementary discussion of Gaussian elimination we consider a problem leading to a linear system where the coefficient matrix is tridiagonal. This special matrix will occur repeatedly throughout this text. We then give an introduction to block multiplication which is an indispensable tool in matrix analysis. We end the chapter with some basic properties of triangular matrices.

1.1 Gaussian Elimination Example

We illustrate how Gaussian elimination works on a 3×3 system. For a general discussion see Section 2.9.

Example 1.1 (Gaussian elimination on a 3×3 system)
Consider a nonsingular system of three equations in three unknowns:

$$\begin{aligned} a_{11}^1 x_1 + a_{12}^1 x_2 + a_{13}^1 x_3 &= b_1^1, \quad \text{I} \\ a_{21}^1 x_1 + a_{22}^1 x_2 + a_{23}^1 x_3 &= b_2^1, \quad \text{II} \\ a_{31}^1 x_1 + a_{32}^1 x_2 + a_{33}^1 x_3 &= b_3^1. \quad \text{III.} \end{aligned}$$

To solve this system by Gaussian elimination suppose $a_{11}^1 \neq 0$. We subtract $m_{21} := a_{21}^1/a_{11}^1$ times equation I from equation II and $m_{31} := a_{31}^1/a_{11}^1$ times equation I

[4]The method was known long before Gauss used it in 1809. It was further developed by Doolittle in 1881, see [6].

from equation III. The result is

$$\begin{aligned} a_{11}^1 x_1 + a_{12}^1 x_2 + a_{13}^1 x_3 &= b_1^1, \quad \text{I} \\ a_{22}^2 x_2 + a_{23}^2 x_3 &= b_2^2, \quad \text{II}' \\ a_{32}^2 x_2 + a_{33}^2 x_3 &= b_3^2, \quad \text{III}', \end{aligned}$$

where $b_i^2 = b_i^1 - m_{i1} b_i^1$ *for* $i = 2, 3$ *and* $a_{ij}^2 = a_{ij}^1 - m_{i,1} a_{1j}^1$ *for* $i, j = 2, 3$. *If* $a_{11}^1 = 0$ *and* $a_{21}^1 \neq 0$ *we first interchange equation I and equation II. If* $a_{11}^1 = a_{21}^1 = 0$ *we interchange equation I and III. Since the system is nonsingular the first column cannot be zero and an interchange is always possible.*

If $a_{22}^2 \neq 0$ *we subtract* $m_{32} := a_{32}^2 / a_{22}^2$ *times equation II' from equation III' to obtain*

$$\begin{aligned} a_{11}^1 x_1 + a_{12}^1 x_2 + a_{13}^1 x_3 &= b_1^1, \quad \text{I} \\ a_{22}^2 x_2 + a_{23}^2 x_3 &= b_2^2, \quad \text{II}' \\ a_{33}^3 x_3 &= b_3^3, \quad \text{III}'', \end{aligned}$$

where $b_3^3 = b_3^2 - m_{32} b_2^2$ *and* $a_{33}^3 = a_{33}^2 - m_{32} a_{23}^2$. *If* $a_{22}^2 = 0$ *then* $a_{32}^2 \neq 0$ *(cf. Theorem 2.59) and we first interchange equation II' and equation III'. The reduced system is easy to solve since it is upper triangular. Starting from the bottom and moving upwards we find*

$$\begin{aligned} x_3 &= b_3^3 / a_{33}^3 \\ x_2 &= (b_2^2 - a_{23}^2 x_3) / a_{22}^2 \\ x_1 &= (b_1^1 - a_{12}^1 x_2 - a_{13}^1 x_3) / a_{11}^1. \end{aligned}$$

This is known as **back substituion**.

Exercise 1.2 (Gaussian elimination example)
Solve the linear system $\boldsymbol{Ax} := \begin{bmatrix} 1 & 1 & -1 \\ -1 & 1 & 3 \\ 2 & 8 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ *using Gaussian elimination.*

Gaussian elimination with row interchanges can in principle be used to solve any nonsingular linear system (cf. Theorem 2.59). However, for many systems occuring in applications this method in its general form is not necessarily the method of choice. Some of the issues are:

1. **Computing time**. Solving a dense system of order n by Gaussian elimination requires $O(n^3)$ arithmetic operations and solving large linear systems can require more time than we are willing to spend. For example, if $n = 10^6$ and one arithmetic operation takes 10^{-12} seconds then the computing time could be a staggering $10^{-12} n^3 \approx 278$ hours.

2. **Row interchanges** is another issue in Gaussian elimination. For example, if we interchange two rows in a tridiagonal matrix then the tridiagonal structure is lost in general.

3. **Stability**. For a well conditioned problem[5] Gaussian elimination using floating point arithmetic will in most cases give an accurate solution. However there is no guarantee for this, see the book [12] for a thorough discussion.

In this chapter we present a problem leading to a $n \times n$ tridiagonal linear system. We show that row interchanges are not necessary for the two problems we consider and derive stable algorithms that only requires $O(n)$ arithmetic operations.

1.2 The Tridiagonal Second Derivative Matrix

Consider the simple **two point boundary value problem**

$$-u''(x) = f(x), \quad x \in [0,1], \quad u(0) = 0, \ u(1) = 0, \tag{1.1}$$

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 (1.1) 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.

Suppose f cannot be integrated exactly. Problem (1.1) can then be solved approximately using the **finite difference method**. We need a difference approximation to the second derivative. If g is a function differentiable at x then

$$g'(x) = \lim_{h \to 0} \frac{g(x + \frac{h}{2}) - g(x - \frac{h}{2})}{h}$$

and applying this to a function u that is twice differentiable at x

$$\begin{aligned} u''(x) &= \lim_{h \to 0} \frac{u'(x + \frac{h}{2}) - u'(x - \frac{h}{2})}{h} = \lim_{h \to 0} \frac{\frac{u(x+h)-u(x)}{h} - \frac{u(x)-u(x-h)}{h}}{h} \\ &= \lim_{h \to 0} \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \end{aligned}$$

To define the points where this difference approximation is used we choose a positive integer m, let $h := 1/(m+1)$ be the discretization parameter, and replace the interval $[0,1]$ by grid points $x_j := jh$ for $j = 0, 1, \ldots, m+1$. We then 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

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

[5] see Chapter 7 for an introduction to condition numbers.

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

$$\boldsymbol{T}\boldsymbol{v} = \begin{bmatrix} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & & & \\ 0 & \ddots & \ddots & \ddots & & \\ & & & & & 0 \\ & & & -1 & 2 & -1 \\ & & & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{m-1} \\ v_m \end{bmatrix} = h^2 \begin{bmatrix} f(h) \\ f(2h) \\ \vdots \\ f\big((m-1)h\big) \\ f(mh) \end{bmatrix} =: \boldsymbol{b}. \tag{1.2}$$

The matrix $\boldsymbol{T}$ is called the **second derivative matrix** and will occur frequently in this book. It is our first example of a tridiagonal matrix, $\boldsymbol{T} = \text{tridiag}(a_i, d_i, c_i) \in \mathbb{R}^{m\times m}$, where in this case $a_i = c_i = -1$ and $d_i = 2$ for all i.

1.3 LU Factorization of a Tridiagonal System

Consider solving the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{A} = \text{tridiag}(a_i, d_i, c_i) \in \mathbb{C}^{n\times n}$ is tridiagonal. Instead of using Gaussian elimination we can try to construct triangular matrices $\boldsymbol{L}$ and $\boldsymbol{U}$ such that the product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ has the form

$$\begin{bmatrix} d_1 & c_1 & & & \\ a_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix} = \begin{bmatrix} 1 & & & \\ l_1 & 1 & & \\ & \ddots & \ddots & \\ & & l_{n-1} & 1 \end{bmatrix} \begin{bmatrix} u_1 & c_1 & & \\ & \ddots & \ddots & \\ & & u_{n-1} & c_{n-1} \\ & & & u_n \end{bmatrix}. \tag{1.3}$$

If $\boldsymbol{L}$ and $\boldsymbol{U}$ can be determined we can find $\boldsymbol{x}$ by solving two simpler systems $\boldsymbol{L}\boldsymbol{y} = \boldsymbol{b}$ and $\boldsymbol{U}\boldsymbol{x} = \boldsymbol{y}$.

1.3.1 Algorithms

For $n = 3$ equation (1.3) takes the form

$$\begin{bmatrix} d_1 & c_1 & 0 \\ a_1 & d_2 & c_2 \\ 0 & a_2 & d_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_1 & 1 & 0 \\ 0 & l_2 & 1 \end{bmatrix} \begin{bmatrix} u_1 & c_1 & 0 \\ 0 & u_2 & c_2 \\ 0 & 0 & u_3 \end{bmatrix} = \begin{bmatrix} u_1 & c_1 & 0 \\ l_1u_1 & l_1c_1 + u_2 & c_2 \\ 0 & l_2u_2 & l_2c_2 + u_3 \end{bmatrix},$$

and if

$$u_1 = d_1, \quad l_1 = a_1/u_1, \quad l_2 = a_2/u_2, \quad u_2 = d_2 - l_1c_1, \quad u_3 = d_3 - l_2c_2,$$

then $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. For general n, if

$$u_1 = d_1, \quad l_k = a_k/u_k, \quad u_{k+1} = d_{k+1} - l_kc_k, \quad k = 1, 2, \ldots, n-1, \tag{1.4}$$

then $\boldsymbol{A} = \boldsymbol{LU}$, and if $u_1, u_2, \ldots, u_{n-1}$ are nonzero then (1.4) is well defined. If in addition $u_n \neq 0$ then we can solve $\boldsymbol{Ly} = \boldsymbol{b}$ and $\boldsymbol{Ux} = \boldsymbol{y}$ for $\boldsymbol{y}$ and $\boldsymbol{x}$.

$$\begin{aligned} y_1 &= b_1, & y_k &= b_k - l_{k-1}y_{k-1}, & k &= 2, 3, \ldots, n, \\ x_n &= y_n/u_n, & x_k &= (y_k - c_k x_{k+1})/u_k, & k &= n-1, \ldots, 2, 1. \end{aligned} \tag{1.5}$$

We formulate this as two algorithms. Since division by zero can occur, the algorithms will not work in general. We give sufficient conditions for success in Theorem 1.7 below.

Algorithm 1.3 (trifactor)
Vectors $\boldsymbol{l} \in \mathbb{C}^{n-1}$, $\boldsymbol{u} \in \mathbb{C}^n$ are computed from $\boldsymbol{a}, \boldsymbol{c} \in \mathbb{C}^{n-1}$, $\boldsymbol{d} \in \mathbb{C}^n$. This implements the LU factorization of a tridiagonal matrix.

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

Algorithm 1.4 (trisolve)
The solution x of the tridiagonal system $\boldsymbol{LUx} = \boldsymbol{b}$ is found from (1.5). Here $\boldsymbol{l}, \boldsymbol{c} \in \mathbb{C}^{n-1}$, $\boldsymbol{u} \in \mathbb{C}^n$ and $\boldsymbol{b} \in \mathbb{C}^{n,r}$ for some $r \in \mathbb{N}$. Thus we can solve a system with several righthand sides. The vectors $\boldsymbol{l}, \boldsymbol{u}$ can be output from `trifactor`.

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

The number of arithmetic operations to compute the LU factorization of a tridiagonal matrix using Algorithm 1.3 is $3n - 3$, while the number of arithmetic operations for Algorithm 1.4 is $5n - 4$. This means that the complexity to solve a tridiagonal system is $O(n)$, or more precisely $8n - 7$, and this number only grows linearly with n, while Gaussian elimination on a full $n \times n$ system is an $O(n^3)$ process.

1.3.2 Diagonal dominance

We show that Algorithms 1.3, 1.4 are well defined for a class of tridiagonal linear systems. Moreover, these linear systems have unique solutions.

Definition 1.5 (Diagonal dominance)
The matrix $\boldsymbol{A} = [a_{ij}] \in \mathbb{C}^{n\times n}$ *is* **weakly diagonally dominant** *if*

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|, \ \ i = 1, \ldots, n. \tag{1.6}$$

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

The following holds for strictly diagonally dominant matrices.

Theorem 1.6 (Strict diagonal dominance)
A strictly diagonally dominant matrix is nonsingular. Moreover, the solution $\boldsymbol{x}$ *of* $\boldsymbol{Ax} = \boldsymbol{b}$ *is bounded as follows:*

$$\max_{1\leq i\leq n} |x_i| \leq \max_{1\leq i\leq n} \Big(\frac{|b_i|}{\sigma_i}\Big), \ \ \textit{where } \sigma_i := |a_{ii}| - \sum_{j\neq i} |a_{ij}|. \tag{1.7}$$

Proof. We first show that the bound (1.7) holds for any solution $\boldsymbol{x}$. Choose k so that $|x_k| = \max_i |x_i|$. Then

$$|b_k| = |a_{kk}x_k + \sum_{j\neq k} a_{kj}x_j| \geq |a_{kk}||x_k| - \sum_{j\neq k} |a_{kj}||x_j| \geq |x_k|\big(|a_{kk}| - \sum_{j\neq k} |a_{kj}|\big),$$

and this implies $\max_{1\leq i\leq n}|x_i| = |x_k| \leq \frac{|b_k|}{\sigma_k} \leq \max_{1\leq i\leq n}\big(\frac{|b_i|}{\sigma_i}\big)$. For nonsingularity, if $\boldsymbol{Ax} = \boldsymbol{0}$, then $\max_{1\leq i\leq n}|x_i| \leq 0$ by (1.7), and so $\boldsymbol{x} = \boldsymbol{0}$. ∎

The zero matrix is weakly diagonally dominant and we need additional condition to guarantee nonsingularity. Consider the 3 matrices

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

They are all weakly diagonally dominant, but $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are singular, while $\boldsymbol{A}_3$ is nonsingular. Indeed, for $\boldsymbol{A}_1$ column two is the sum of columns one and three, $\boldsymbol{A}_2$ has a zero row, and $\det(\boldsymbol{A}_3) = 4 \neq 0$.

In the literature diagonal dominance is therefore most often defined by including some additional sufficient conditions. We also need conditions guaranteeing that the LU factorization (1.3) of a tridiagonal matrix is well defined.

Theorem 1.7 (Weak diagonal dominance)
Suppose $\boldsymbol{A} = tridiag(a_i, d_i, c_i) \in \mathbb{C}^{n\times n}$ *is tridiagonal and weakly diagonally dominant. If* $|d_1| > |c_1|$ *and* $a_i \neq 0$ *for* $i = 1, \ldots, n-2$*, then* $\boldsymbol{A}$ *has a unique LU factorization* (1.3)*. If in addition* $d_n \neq 0$*, then* $\boldsymbol{A}$ *is nonsingular.*

Proof. The matrix $\boldsymbol{A}$ has an LU factorization if the u_k's in (1.4) are nonzero for $k = 1, \ldots, n-1$. For this it is sufficient to show by induction that $|u_k| > |c_k|$ for $k = 1, \ldots, n-1$. By assumption $|u_1| = |d_1| > |c_1|$. Suppose $|u_k| > |c_k|$ for some $1 \leq k \leq n-2$. Then $|c_k|/|u_k| < 1$ and by (1.4) and since $a_k \neq 0$

$$|u_{k+1}| = |d_{k+1} - l_k c_k| = |d_{k+1} - \frac{a_k c_k}{u_k}| \geq |d_{k+1}| - \frac{|a_k||c_k|}{|u_k|} > |d_{k+1}| - |a_k|. \quad (1.8)$$

This also holds for $k = n-1$ if $a_{n-1} \neq 0$. By (1.8) and weak diagonal dominance $|u_{k+1}| > |d_{k+1}| - |a_k| \geq |c_{k+1}|$ and it follows by induction that an LU factorization exists. It is unique since any LU factorization must satisfy (1.4). For nonsingularity we need to show that $u_n \neq 0$. For then by Lemma 1.22, both $\boldsymbol{L}$ and $\boldsymbol{U}$ are nonsingular, and this is equivalent to $\boldsymbol{A} = \boldsymbol{LU}$ being nonsingular. If $a_{n-1} \neq 0$ then by (1.4) $|u_n| > |d_n| - |a_{n-1}| \geq 0$ by weak diagonal dominance, while if $a_{n-1} = 0$ then again by (1.8) $|u_n| \geq |d_n| > 0$. □

Exercise 1.8 (Strict diagonal dominance)
Show that a strictly diagonally dominant and tridiagonal matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *has a unique LU factorization*[6]*.*

Consider now the special system $\boldsymbol{Tv} = \boldsymbol{b}$ given by (1.2). The matrix $\boldsymbol{T}$ is weakly diagonally dominant and satisfies the additional conditions in Theorem 1.7. Thus it is nonsingular and we can solve the system in $O(n)$ arithmetic operations using Algorithms 1.3,1.4.

We could use the explicit inverse of $\boldsymbol{T}$, given in Exercise 1.10, to compute the solution of $\boldsymbol{Tv} = \boldsymbol{b}$ as $\boldsymbol{v} = \boldsymbol{T}^{-1}\boldsymbol{b}$. However this is not a good idea. In fact, all elements in $\boldsymbol{T}^{-1}$ are nonzero and the calculation of $\boldsymbol{T}^{-1}\boldsymbol{b}$ requires $O(n^2)$ operations.

Exercise 1.9 (LU factorization of 2. derivative matrix)

[6] Hint, argue as in (1.8)

Show that $\boldsymbol{T} = \boldsymbol{LU}$*, where*

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\frac{1}{2} & 1 & \ddots & & \vdots \\ 0 & -\frac{2}{3} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{m-1}{m} & 1 \end{bmatrix}, \boldsymbol{U} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ 0 & \frac{3}{2} & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \frac{m}{m-1} & -1 \\ 0 & \cdots & \cdots & 0 & \frac{m+1}{m} \end{bmatrix} \tag{1.9}$$

is the LU factorization of $\boldsymbol{T}$*.*

Exercise 1.10 (Inverse of 2. derivative matrix)
Let $\boldsymbol{S} \in \mathbb{R}^{m\times m}$ *have elements* s_{ij} *given by*

$$s_{i,j} = s_{j,i} = \frac{1}{m+1} j\big(m+1-i\big), \quad 1 \le j \le i \le m. \tag{1.10}$$

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

Exercise 1.11 (Central difference approximation of 2. derivative)
Consider

$$\delta^2 f(x) := \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}, \quad h > 0, \quad f : [x-h, x+h] \to \mathbb{R}.$$

1. Show using Taylor expansion that if $f \in C^2[x-h, x+h]$ then for some η_2

$$\delta^2 f(x) = f''(\eta_2), \quad x - h < \eta_2 < x + h.$$

2. If $f \in C^4[x-h, x+h]$ then for some η_4

$$\delta^2 f(x) = f''(x) + \frac{h^2}{12} f^{(4)}(\eta_4), \quad x - h < \eta_4 < x + h.$$

$\delta^2 f(x)$ is known as the **central difference approximation** to the second derivative at x.

Exercise 1.12 (Two point boundary value problem)
We consider a finite difference method for the two point boundary value problem

$$\begin{aligned} -u''(x) + r(x)u'(x) + q(x)u(x) &= f(x), \ \textit{for } x \in [a,b], \\ u(a) = g_0, \quad u(b) &= g_1. \end{aligned} \tag{1.11}$$

We assume that the given functions f, q and r are continuous on $[a, b]$ and that $q(x) \geq 0$ for $x \in [a, b]$. It can then be shown that (1.11) has a unique solution u.

To solve (1.11) 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

$$\frac{-v_{j-1} + 2v_j - v_{j+1}}{h^2} + r(x_j)\frac{v_{j+1} - v_{j-1}}{2h} + q(x_j)v_j = f(x_j), \quad j = 1, \ldots, m, \tag{1.12}$$

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

(a) *Show that (1.12) leads to a tridiagonal linear system $\boldsymbol{Av} = \boldsymbol{b}$, where $\boldsymbol{A} = tridiag(a_j, d_j.c_j) \in \mathbb{R}^{m\times m}$ 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

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

(b) *Show that the linear system satisfies the conditions in Theorem 1.7 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 1.13 (Two point boundary value problem; computation)

(a) *Consider the problem (1.11) 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 (1.12) for $h = 0.1, 0.05, 0.025, 0.0125$, and compute the "error" $\max_{1\leq j\leq m}|u(x_j) - v_j|$ 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 $h = 0.1, 0.05, 0.025, 0.0125$.*

1.4 Block Multiplication

Block multiplication is a powerful and essentiasl tool for dealing with matrices. It will be used extensively in this book.

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 are often referred to as **blocks** and a partitioned matrix is sometimes called a **block matrix**.

In the following we assume that $\boldsymbol{A} \in \mathbb{C}^{m\times p}$ and $\boldsymbol{B} \in \mathbb{C}^{p\times n}$. 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 $i = 1, \ldots, m$.

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 $\boldsymbol{B}_1 \in \mathbb{C}^{p\times r}$ and $\boldsymbol{B}_2 \in \mathbb{C}^{p\times(n-r)}$ 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 $\boldsymbol{A}_1 \in \mathbb{C}^{k\times p}$ and $\boldsymbol{A}_2 \in \mathbb{C}^{(m-k)\times p}$ 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 $\boldsymbol{A}_1 \in \mathbb{C}^{m\times s}$, $\boldsymbol{A}_2 \in \mathbb{C}^{m\times(p-s)}$, $\boldsymbol{B}_1 \in \mathbb{C}^{s\times n}$ and $\boldsymbol{B}_2 \in \mathbb{C}^{(p-s)\times n}$ 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, $(\boldsymbol{AB})_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj} = \sum_{k=1}^{s} a_{ik}b_{kj} + \sum_{k=s+1}^{p} a_{ik}b_{kj} = (\boldsymbol{A}_1\boldsymbol{B}_1)_{ij} + (\boldsymbol{A}_2\boldsymbol{B}_2)_{ij} = (\boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2)_{ij}$.

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 the number of columns in $\boldsymbol{A}$ equals the number of rows in $\boldsymbol{B}$. 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 $\boldsymbol{A}_{ik}\boldsymbol{B}_{kj}$ 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

$$\begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1s} \\ \vdots & & \vdots \\ \boldsymbol{A}_{p1} & \cdots & \boldsymbol{A}_{ps} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \cdots & \boldsymbol{B}_{1q} \\ \vdots & & \vdots \\ \boldsymbol{B}_{s1} & \cdots & \boldsymbol{B}_{sq} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{11} & \cdots & \boldsymbol{C}_{1q} \\ \vdots & & \vdots \\ \boldsymbol{C}_{p1} & \cdots & \boldsymbol{C}_{pq} \end{bmatrix}.$$

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.

Exercise 1.14 (Matrix element as a quadratic form)
For any matrix $\boldsymbol{A}$ show that $a_{ij} = \boldsymbol{e}_i^T \boldsymbol{A} \boldsymbol{e}_j$ for all i, j.

Exercise 1.15 (Outer product expansion of a matrix)
For any matrix $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ show that $\boldsymbol{A} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \boldsymbol{e}_i \boldsymbol{e}_j^T$.

Exercise 1.16 (The product $\boldsymbol{A}^T \boldsymbol{A}$)
Let $\boldsymbol{B} = \boldsymbol{A}^T \boldsymbol{A}$. Explain why this product is defined for any matrix $\boldsymbol{A}$. Show that $b_{ij} = \boldsymbol{a}_{:i}^T \boldsymbol{a}_{:j}$ for all i, j.

Exercise 1.17 (Outer product expansion)
For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times n}$ show that

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

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

Exercise 1.18 (System with many right hand sides; compact form)
Suppose $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{m \times p}$, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Show that

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

Exercise 1.19 (Block multiplication example)
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 $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{A}_1\boldsymbol{B}_1$?

Exercise 1.20 (Another block multiplication example)
Suppose $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n\times n}$ 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 $\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1 \in \mathbb{R}^{(n-1)\times(n-1)}$. Show that

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

1.5 Triangular Matrices; Basic facts

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

Lemma 1.21 (Inverse of a block triangular matrix)
Suppose

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

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

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} & \boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix}, \tag{1.13}$$

for some matrix $\boldsymbol{C}$.

Proof. Suppose $\boldsymbol{A}$ is nonsingular. We partition $\boldsymbol{B} := \boldsymbol{A}^{-1}$ conformally with $\boldsymbol{A}$ and have

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

Using block-multiplication we find

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

The first equation implies that $\boldsymbol{A}_{11}$ is nonsingular, this in turn implies that $\boldsymbol{B}_{21} = \boldsymbol{0}\boldsymbol{A}_{11}^{-1} = \boldsymbol{0}$ 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. From the fourth equation we find

$$\boldsymbol{B}_{12} = \boldsymbol{C} = -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}.$$

Conversely, if $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are nonsingular then

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

and $\boldsymbol{A}$ is nonsingular with the indicated inverse. □

Consider now a triangular matrix.

Lemma 1.22 (Inverse of a triangular matrix)
An upper (lower) triangular matrix $\boldsymbol{A} = [a_{ij}] \in \mathbb{C}^{n\times n}$ 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 $a_{11} \neq \boldsymbol{0}$ and in that case $\boldsymbol{A}^{-1} = [a_{11}^{-1}]$. Suppose the result holds for $n = k$ and let $\boldsymbol{A} \in \mathbb{C}^{(k+1)\times(k+1)}$ be upper triangular. We partition $\boldsymbol{A}$ in the form

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

and note that $\boldsymbol{A}_k \in \mathbb{C}^{k\times k}$ is upper triangular. By Lemma 1.21 $\boldsymbol{A}$ is nonsingular if and only if $\boldsymbol{A}_k$ and $(a_{k+1,k+1})$ are nonsingular and in that case

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

for some $\boldsymbol{c} \in \mathbb{C}^n$. 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 1.23 (Product of triangular matrices)
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 1.24 (Unit triangular matrices)
For a unit upper (lower) triangular matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$:

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 1.22, while Lemma 1.23 implies 2. □

1.6 Review Questions

1.6.1 Define the second derivative matrix $\boldsymbol{T}$. Why is it nonsingular?

1.6.2 Is a weakly diagonally dominant matrix nonsingular?

1.6.3 Why do we not use the explicit inverse of $\boldsymbol{T}$ to solve the linear system $\boldsymbol{Tx} = \boldsymbol{b}$

1.6.4 Show that a strictly diagonally dominant matrix is nonsingular.

1.6.5 Does a tridiagonal matrix always have an LU factorization?

Chapter 2

LU Factorizations

Numerical methods for solving systems of linear equations are often based on writing a matrix as a product of simpler matrices. Such a **factorization** is useful if the corresponding matrix problem for each of the factors is simple to solve, and extra numerical stability issues are not introduced. Examples of factorizations were encountered in Chapter 1 and we saw how an LU factorization can be used to solve certain tridiagonal systems efficiently. Other factorizations based on unitary matrices will be considered later in this book.

In this chapter we consider the general theory of LU factorizations. We consider some related factorizations called symmetric LU or LDLT, and Cholesky. The latter can be used for symmetric positive matrices, and we give an introduction to positive definite and positive semindefinite matrices. We consider a matrix formulation of Gaussian elimination using Gauss transformations and permutation matrices leading to the PLU factorization of a matrix.

2.1 Algorithms for triangular systems

Recall that a matrix $\boldsymbol{U}$ is upper triangular if $u_{ij} = 0$ for $i > j$, and a matrix $\boldsymbol{L}$ is lower triangular if $l_{ij} = 0$ for $i < j$. If $\boldsymbol{U}$ is upper triangular then $\boldsymbol{U}^T$ is lower triangular.

A nonsingular triangular linear system $\boldsymbol{Ax} = \boldsymbol{b}$ is easy to solve. By Lemma 1.22 $\boldsymbol{A}$ has nonzero diagonal elements. Consider first the lower triangular case. For $n = 3$ the system is

$$\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} & 0 \\ 0 & 0 & 0 & a_{54} & a_{55} \end{bmatrix}, \quad \begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 \\ 0 & a_{42} & a_{43} & a_{44} & 0 \\ 0 & 0 & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

Figure 2.1. *Lower triangular* 5×5 *band matrices:* $d = 1$ *(left) and* $d = 2$ *(right).*

From the first equation we find $x_1 = b_1/a_{11}$. Solving the second equation for x_2 we obtain $x_2 = (b_2 - a_{21}x_1)/a_{22}$. 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**. In general

$$x_k = \big(b_k - \sum_{j=1}^{k-1} a_{k,j}x_j\big)/a_{kk}, \quad k = 1, 2, \ldots, n. \tag{2.1}$$

When $\boldsymbol{A}$ is a lower triangular band matrix the number of arithmetic operations necessary to find $\boldsymbol{x}$ can be reduced. Suppose $\boldsymbol{A}$ is a lower triangular d-banded, so that $a_{k,j} = 0$ for $j \notin \{l_k, l_k + 1, \ldots, k\}$ for $k = 1, 2, \ldots, n$, and where $l_k := \max(1, k-d)$, see Figure 2.1. For a lower triangular d-band matrix the calculation in (2.1) can be simplified as follows

$$x_k = \big(b_k - \sum_{j=l_k}^{k-1} a_{k,j}x_j\big)/a_{kk}, \quad k = 1, 2, \ldots, n. \tag{2.2}$$

Note that (2.2) reduces to (2.1) if $d = n$. Letting $A(k, l_k : k-1) * x(l_k : k-1)$ denote the sum $\sum_{j=l_k}^{k-1} a_{kj}x_j$ we arrive at the following algorithm.

Algorithm 2.1 (forwardsolve (row oriented))
Given a nonsingular lower triangular d-banded matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{b} \in \mathbb{C}^n$. An $\boldsymbol{x} \in \mathbb{C}^n$ is computed so that $\boldsymbol{Ax} = \boldsymbol{b}$.

```
function x=rforwardsolve(A,b,d)
n=length(b); x=b;
x(1)=b(1)/A(1,1);
for k=2:n
    lk=max(1,k-d);
    x(k)=(b(k)-A(k,lk:k-1)*x(lk:k-1))/A(k,k);
end
```

A system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is upper triangular must be solved by **back substitution** or 'bottom-up'. We first find x_n from the last equation and then

move upwards for the remaining unknowns. For an upper triangular d-banded matrix this leads to the following algorithm.

Algorithm 2.2 (backsolve (row oriented))
Given a nonsingular upper triangular d-banded matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{b} \in \mathbb{C}^n$. An $\boldsymbol{x} \in \mathbb{C}^n$ is computed so that $\boldsymbol{Ax} = \boldsymbol{b}$.

```
function x=rbacksolve(A,b,d)
n=length(b); x=b;
x(n)=b(n)/A(n,n);
for k=n-1:-1:1
    uk=min(n,k+d);
    x(k)=(b(k)-A(k,k+1:uk)*x(k+1:uk))/A(k,k);
end
```

Exercise 2.3 (Column oriented forward- and backsolve)
The intial "r" in the names of Algorithms 2.1,2.2 signals that these algorithms are row oriented. For each k we take the inner product of a part of a row with the already computed unknowns. In this exercise we develop column oriented vectorized versions of forward and backward substitution. Consider the system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is lower triangular. Suppose after $k-1$ steps of the algorithm we have a reduced system in the form

$$\begin{bmatrix} a_{k,k} & 0 & \cdots & 0 \\ a_{k+1,k} & a_{k+1,k+1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ a_{n,k} & & \cdots & a_{n\times n} \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_k \\ b_{k+1} \\ \vdots \\ b_n \end{bmatrix}.$$

This system is of order $n-k+1$. The unknowns are $x_k, \ldots, x_n$.

a) We see that $x_k = b_k/a_{k,k}$ and eliminating x_k from the remaining equations we obtain a system of order $n-k$ with unknowns $x_{k+1}, \ldots, x_n$

$$\begin{bmatrix} a_{k+1,k+1} & 0 & \cdots & 0 \\ a_{k+2,k+1} & a_{k+2,k+2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ a_{n,k+1} & & \cdots & a_{n\times n} \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_{k+1} \\ \vdots \\ b_n \end{bmatrix} - x_k \begin{bmatrix} a_{k+1,k} \\ \vdots \\ a_{n,k} \end{bmatrix}.$$

Thus at the kth step, $k = 1, 2, \ldots n$ we set $x_k = b_k/A(k,k)$ and update b as follows:

$$b(k+1:n) = b(k+1:n) - x(k) * A(k+1:n, k).$$

This leads to the following algorithm.

Algorithm 2.4 (Forward solve (column oriented))
Given a nonsingular lower triangular d-banded matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{b} \in \mathbb{C}^n$. An $\boldsymbol{x} \in \mathbb{C}^n$ is computed so that $\boldsymbol{Ax} = \boldsymbol{b}$.

```
function x=cforwardsolve(A,b,d)
x=b; n=length(b);
for   k=1:n-1
    x(k)=b(k)/A(k,k); uk=min(n,k+d);
   b(k+1:uk)=b(k+1:uk)-A(k+1:uk,k)*x(k);
end
x(n)=b(n)/A(n,n);
end
```

b) Suppose now $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is nonsingular, upper triangular, d-banded, and $\boldsymbol{b} \in \mathbb{C}^n$. Justify the following column oriented vectorized algorithms for solving $\boldsymbol{Ax} = \boldsymbol{b}$.

Algorithm 2.5 (Backsolve (column oriented))
Given a nonsingular upper triangular d-banded matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{b} \in \mathbb{C}^n$. An $\boldsymbol{x} \in \mathbb{C}^n$ is computed so that $\boldsymbol{Ax} = \boldsymbol{b}$.

```
function x=cbacksolve(A,b,d)
x=b; n=length(b);
for   k=n:-1:2
    x(k)=b(k)/A(k,k); lk=max(1,k-d);
    b(lk:k-1)=b(lk:k-1)-A(lk:k-1,k)*x(k);
end
x(1)=b(1)/A(1,1);
end
```

Exercise 2.6 (Computing the inverse of a triangular matrix)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is a nonsingular triangular matrix with inverse $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n]$. The kth column $\boldsymbol{b}_k$ of $\boldsymbol{B}$ is the solution of the linear systems $\boldsymbol{Ab}_k = \boldsymbol{e}_k$. Write this system as a 2×2 triangular block system and explain why we can find $\boldsymbol{b}_k$ by solving the linear systems

$$\boldsymbol{A}(k{:}n, k{:}n)\boldsymbol{b}_k(k{:}n) = \boldsymbol{I}(k{:}n, k), \quad k = 1, \ldots, n \quad \text{lower triangular}, \tag{2.3}$$

$$\boldsymbol{A}(1{:}k, 1{:}k)\boldsymbol{b}_k(1{:}k) = \boldsymbol{I}(1{:}k, k), \quad k = n, n-1, \ldots, 1, \quad \text{upper triangular} \tag{2.4}$$

Is it possible to store the interesting part of $\boldsymbol{b}_k$ in $\boldsymbol{A}$ as soon as it is computed?

2.2 The LU Factorization

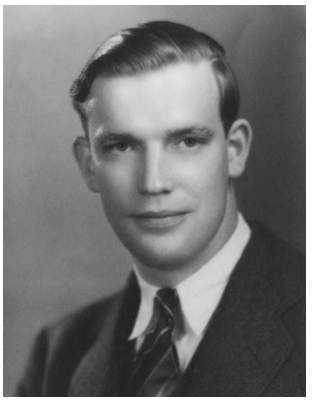
We say that $\boldsymbol{A} = \boldsymbol{LU}$ is an **LU factorization** of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ if $\boldsymbol{L} \in \mathbb{C}^{n\times n}$ is lower triangular (**left triangular**) and $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ is upper triangular (**right triangular**).

2.2.1 The LU theorem

Consider finding $\boldsymbol{L}$ and $\boldsymbol{U}$. Equating the i, j element in $\boldsymbol{A}$ and the product $\boldsymbol{LU}$, and noting that $l_{i,j} = 0$ for $j > i$ and $u_{i,j} = 0$ for $i > j$, we obtain an equation

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik} u_{kj}, , \quad i, j = 1, 2, \ldots, n \tag{2.5}$$

involving the unknown elements in $\boldsymbol{L}$ and $\boldsymbol{U}$. This is an underdetermined system of n^2 equations in $n^2 + n$ unknowns. One way to reduce the number of unknowns is to require that one of the triangular matrices should be **unit triangular**, i. e., have ones on the diagonal. Other scalings of the diagonals are also possible, see Section 2.6. Choosing $\boldsymbol{U}$ to be unit triangular is sometimes known as a **Crout factorization**.

Henry Jensen, 1915-1974 (left), Prescott Durand Crout, 1907-1984. Jensen worked on LU factorizations. His name is also associated with a very useful inequality (cf. Theorem 7.37).

For our discussion we will assume that $\boldsymbol{L}$ is unit triangular. Three things can happen. An LU factorization exists and is unique, it exists, but it is not unique, or it does not exist. The following 2×2 example illustrates this.

Example 2.7 (LU of 2×2 matrix)
Let $a, b, c, d \in \mathbb{C}$. An LU factorization of $\boldsymbol{A} = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ must satisfy the equations

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_1 & 1 \end{bmatrix} \begin{bmatrix} u_1 & u_3 \\ 0 & u_2 \end{bmatrix} = \begin{bmatrix} u_1 & u_3 \\ l_1 u_1 & l_1 u_3 + u_2 \end{bmatrix}$$

for the unknowns l_1 in $\boldsymbol{L}$ and u_1, u_2, u_3 in $\boldsymbol{U}$. The equations are

$$u_1 = a. \quad u_3 = b, \quad l_1 a = c, \quad u_2 = d - l_1 b.$$

These equations do not always have a solution. Indeed, the main problem is the nonlinear equation $l_1 a = c$. There are three cases

1. *$a \neq 0$: The matrix has a unique LU factorization with $l_1 = c/a$.*
2. *$a = 0, c \neq 0$: No LU factorization exists.*
3. *$a = c = 0$: The LU factorization exists, but it is not unique. Any value for l_1 can be used.*

Of the four matrices

$$\boldsymbol{A}_1 := \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \boldsymbol{A}_2 := \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{A}_3 := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{A}_4 := \begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix}.$$

$\boldsymbol{A}_1$ has a unique LU factorization, $\boldsymbol{A}_2$ has no LU factorization, $\boldsymbol{A}_3$ has a unique LU factorization even if it is singular, and $\boldsymbol{A}_4$ has an LU factorization, but it is not unique.

Example 2.8 (LU of 3×3 matrices)
The matrix

$$\boldsymbol{A} := \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \end{bmatrix}$$

has an LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$, with

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & y & 1 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 2-y \end{bmatrix}.$$

It is not unique since $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ for any $y \in \mathbb{C}$.

To characterize matrices with a unique LU factorization we first give a definition.

Definition 2.9 (Principal submatrix)
For $k = 1, \ldots, n$ the matrices $\boldsymbol{A}_{[k]} \in \mathbb{C}^{k \times k}$ given by

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

are called the **leading principal submatrices** *of $\boldsymbol{A} \in \mathbb{C}^{n \times n}$. More generally, a matrix $\boldsymbol{B} \in \mathbb{C}^{k \times k}$ 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 $1 \leq r_1 < \cdots < r_k \leq n$. 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 $k = 1$ The only principal submatrices of order $k = 1$ are the diagonal elements of $\boldsymbol{A}$.

Example 2.10 (Principal submatrices)
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

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

In preparation for the main theorem about LU factorization we prove a simple lemma.

Lemma 2.11 (LU of leading principal sub matrices)
Suppose $\boldsymbol{A} = \boldsymbol{LU}$ *is an LU factorization of* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. *For* $k = 1, \ldots, n$ *let* $\boldsymbol{A}_{[k]}, \boldsymbol{L}_{[k]}, \boldsymbol{U}_{[k]}$ *be the leading principal submatrices of* $\boldsymbol{A}, \boldsymbol{L}, \boldsymbol{U}$, *respectively. Then* $\boldsymbol{A}_{[k]} = \boldsymbol{L}_{[k]}\boldsymbol{U}_{[k]}$ *is an LU factorization of* $\boldsymbol{A}_{[k]}$ *for* $k = 1, \ldots, n$.

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

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{[k]} & \boldsymbol{B}_k \\ \boldsymbol{C}_k & \boldsymbol{F}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{[k]} & \boldsymbol{0} \\ \boldsymbol{M}_k & \boldsymbol{N}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{[k]} & \boldsymbol{S}_k \\ \boldsymbol{0} & \boldsymbol{T}_k \end{bmatrix} = \boldsymbol{LU}, \tag{2.6}$$

where $\boldsymbol{F}_k, \boldsymbol{N}_k, \boldsymbol{T}_k \in \mathbb{C}^{(n-k)\times(n-k)}$. Using block multiplication we find $\boldsymbol{A}_{[k]} = \boldsymbol{L}_{[k]}\boldsymbol{U}_{[k]}$. Since $\boldsymbol{L}_{[k]}$ is unit lower triangular and $\boldsymbol{U}_{[k]}$ is upper triangular this is an LU factorization of $\boldsymbol{A}_{[k]}$. □

The following theorem give a necessary and sufficient condition for existence of a unique LU factorization.

Theorem 2.12 (LU theorem)
A square matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *has a unique LU factorization if and only if the leading principal submatrices* $\boldsymbol{A}_{[k]}$ *of* $\boldsymbol{A}$ *are nonsingular for* $k = 1, \ldots, n-1$.

Proof. Suppose $\boldsymbol{A}_{[k]}$ is nonsingular for $k = 1, \ldots, n-1$. 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 $\boldsymbol{A}_{[n-1]} = \boldsymbol{L}_{n-1}\boldsymbol{U}_{n-1}$, and that $\boldsymbol{A}_{[1]}, \ldots, \boldsymbol{A}_{[n-1]}$ are nonsingular. By block multiplication

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{[n-1]} & \boldsymbol{b} \\ \boldsymbol{c}^T & a_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{n-1} & \boldsymbol{0} \\ \boldsymbol{m}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{n-1} & \boldsymbol{s} \\ 0 & t \end{bmatrix} = \boldsymbol{L}\boldsymbol{U}, \tag{2.7}$$

if and only if $\boldsymbol{m}, \boldsymbol{s} \in \mathbb{C}^{n-1}$ and $t \in \mathbb{C}$ satisfy $\boldsymbol{b} = \boldsymbol{L}_{n-1}\boldsymbol{s}$, $\boldsymbol{c}^T = \boldsymbol{m}^T\boldsymbol{U}_{n-1}$, and $a_{nn} = \boldsymbol{m}^T\boldsymbol{s} + t$. Since $\boldsymbol{A}_{[n-1]}$ is nonsingular it follows that $\boldsymbol{L}_{n-1}$ and $\boldsymbol{U}_{n-1}$ are nonsingular and therefore $\boldsymbol{m}, \boldsymbol{s}$, and t are uniquely given. Thus (2.7) gives a unique LU factorization of $\boldsymbol{A}$.

Conversely, suppose $\boldsymbol{A}$ has a unique LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. By Lemma 2.11 $\boldsymbol{A}_{[k]} = \boldsymbol{L}_{[k]}\boldsymbol{U}_{[k]}$ is an LU factorization of $\boldsymbol{A}_{[k]}$ for $k = 1, \ldots, n-1$. Suppose $\boldsymbol{A}_{[k]}$ is singular for some $k \leq n-1$. We will show that this leads to a contradiction. Let k be the smallest integer so that $\boldsymbol{A}_{[k]}$ is singular. Since $\boldsymbol{A}_{[j]}$ is nonsingular for $j \leq k-1$ it follows from what we have already shown that $\boldsymbol{A}_{[k]} = \boldsymbol{L}_{[k]}\boldsymbol{U}_{[k]}$ is the unique LU factorization of $\boldsymbol{A}_{[k]}$. The matrix $\boldsymbol{U}_{[k]}$ is singular since $\boldsymbol{A}_{[k]}$ is singular and $\boldsymbol{L}_{[k]}$ is nonsingular. By block multiplication in (2.6) we have $\boldsymbol{C}_k = \boldsymbol{M}_k\boldsymbol{U}_{[k]}$ or $\boldsymbol{U}_{[k]}^T\boldsymbol{M}_k^T = \boldsymbol{C}_k^T$. This can be written as $n-k$ linear systems for the columns of $\boldsymbol{M}_k^T$. By assumption $\boldsymbol{M}_k^T$ exists, but since $\boldsymbol{U}_{[k]}^T$ is singular $\boldsymbol{M}_k$ is not unique, a contradiction. ∎

A matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ can have an LU factorization even if $\boldsymbol{A}_{[k]}$ is singular for some $k < n$. By Theorem 2.12 such an LU factorization cannot be unique.

Remark 2.13 (LU of upper triangular matrix)
An LU factorization of an upper triangular matrix $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{I}\boldsymbol{A}$ so it always exists even if $\boldsymbol{A}$ has zeros somewhere on the diagonal. By Lemma 1.22, if some a_{kk} is zero then $\boldsymbol{A}_{[k]}$ is singular and the LU factorization cannot be unique. In particular, for the zero matrix any unit lower triangular matrix can be used as $\boldsymbol{L}$ in an LU factorization.

Remark 2.14 (PLU factorization)
We have shown that a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ has a unique 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 $\boldsymbol{P}^T\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ or equivalently $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$, where $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{L}$ is unit lower triangular, and $\boldsymbol{U}$ is upper triangular. We call this a **PLU factorization** *of $\boldsymbol{A}$. (Cf. Theorem 2.60).*

Exercise 2.15 (Row interchange)
Show that $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ has a unique LU factorization. Note that we have only interchanged rows in Example 2.7.

Exercise 2.16 (LU of singular matrix)
Find an LU factorization of the singular matrix $\left[\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right]$. *Is it unique?*

Exercise 2.17 (LU and determinant)
Suppose $\boldsymbol{A}$ *has an LU factorization* $\boldsymbol{A} = \boldsymbol{LU}$. *Show that* $\det(\boldsymbol{A}_{[k]}) = u_{11}u_{22}\cdots u_{kk}$ *for* $k = 1, \ldots, n$.

Exercise 2.18 (Diagonal elements in U)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *and* $\boldsymbol{A}_{[k]}$ *is nonsingular for* $k = 1, \ldots, n-1$. *Use Exercise 2.17 to show that the diagonal elements* u_{kk} *in the LU factorization are*

$$u_{11} = a_{11}, \quad u_{kk} = \frac{\det(\boldsymbol{A}_{[k]})}{\det(\boldsymbol{A}_{[k-1]})}, \text{ for } k = 2, \ldots, n. \tag{2.8}$$

2.2.2 Operation count

It is useful to have a number which indicates the amount of work an algorithm requires. In this book we measure this by estimating the total number of arithmetic operations. We count both additions, subtractions, multiplications and divisions, but not work on indices. As an example it is shown below that the calculation to find the LU factorization of a full matrix of order n is exactly

$$N_{LU} := \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n. \tag{2.9}$$

We are only interested in this number when n is large and for such n the term $\frac{2}{3}n^3$ dominates. We therefore regularly ignore lower order terms and use **number of arithmetic operations** both for the exact count and for the highest order term. We also say more loosely that the the number of operations is $O(n^3)$. We will use the number of operations counted in one of these ways as a measure of the **complexity of an algorithm** and say that the complexity of LU factorization of a full matrix is $O(n^3)$ or more precisely $\frac{2}{3}n^3$.

We will compare the number of arithmetic operations of many algorithms with the number of arithmetic operations of LU factorization and define for $n \in \mathbb{N}$ the number G_n[7] as follows:

Definition 2.19 ($G_n := \frac{2}{3}n^3$)
We define $G_n := \frac{2}{3}n^3$.

The complexity of solving a system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a full upper or lower triangular matrix is easily shown to be exactly n^2. This number is reduced to $n^2 - n$ if $\boldsymbol{A}$ has ones on the diagonal.

[7] It can be shown that the complexity of Gaussian elimination is also equal to G_n.

Consider now finding the total number of arithmetic operations, N_{LU} for LU factorization. Suppose in (2.7) that $k := n-1$ and $\boldsymbol{L}_k$ and $\boldsymbol{U}_k$ are already computed. To find $\boldsymbol{L}$ and $\boldsymbol{U}$ we have to solve the triangular systems $\boldsymbol{L}_k\boldsymbol{s} = \boldsymbol{b}$, $\boldsymbol{U}_k^T\boldsymbol{m} = \boldsymbol{c}$, and then $t = a_{k+1,k+1} - \boldsymbol{m}^T\boldsymbol{s}$. Since $\boldsymbol{L}_k$ is unit lower triangular and $\boldsymbol{U}_k^T$ is lower triangular of order k this requires $k(k-1)$, k^2, and $2k$ operations for $\boldsymbol{s}$, $\boldsymbol{m}$, and t, respecively, a total of $k(2k+1)$ operations. Taking also $\boldsymbol{L}_k$ and $\boldsymbol{U}_k$ into consideration we obtain

$$N_{LU} = \sum_{k=1}^{n-1} k(2k+1) = 2\sum_{k=1}^{n-1} k^2 + \sum_{k=1}^{n-1} k = \frac{2}{3}n(n-1)(n-\frac{1}{2}) + \frac{1}{2}n(n-1)$$

which equals the number in (2.9).

There is a quick way to arrive at the estimate $2n^3/3$. We only consider the arithmetic operations contributing to the leading term (the inner loops). 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. We find

$$N_{LU} = \sum_{k=1}^{n-1} k(2k+1) \approx \sum_{k=1}^{n-1} 2k^2 \approx \int_1^{n-1} 2k^2 dk \approx \int_0^n 2k^2 dk = G_n.$$

We see that LU factorization is an $O(n^3)$ process while solving a triangular system requires $O(n^2)$ arithmetic operations. Thus, if $n = 10^6$ and one arithmetic operation requires $c = 10^{-12}$ seconds of computing time then $cn^3 = 10^6$ seconds $\approx$ 278 hours and $cn^2 = 1$ second, giving dramatic differences in computing time.

Exercise 2.20 (Finite sums of integers)
Use induction on m, or some other method, to show that

$$1 + 2 + \cdots + m = \frac{1}{2}m(m+1), \tag{2.10}$$

$$1^2 + 2^2 + \cdots + m^2 = \frac{1}{3}m(m+\frac{1}{2})(m+1), \tag{2.11}$$

$$1 + 3 + 5 + \cdots + 2m - 1 = m^2, \tag{2.12}$$

$$1*2 + 2*3 + 3*4 + \cdots + (m-1)m = \frac{1}{3}(m-1)m(m+1). \tag{2.13}$$

Exercise 2.21 (Operations)
To solve an upper triangular linear system by back substitution takes n^2 arithmetic operations. Show that the number of arithmetic operations in (2.4) *is $\frac{1}{3}n(n+\frac{1}{2})(n+1) \approx \frac{1}{2}G_n$.*

Exercise 2.22 (Multiplying triangular matrices)
Show that the matrix multiplication $\boldsymbol{AB}$ can be done in $\frac{1}{3}n(2n^2+1) \approx G_n$ arithmetic operations when $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is lower triangular and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is upper triangular. What about $\boldsymbol{BA}$?

2.3 The Symmetric LU Factorization

We consider next LU factorization of a real symmetric matrix.

Definition 2.23 (Symmetric LU)
Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. 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** *or an* **LDLT factorization** *of $\boldsymbol{A}$.*

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

Example 2.24 (2×2 symmetric LU)
Let $a, b, c \in \mathbb{R}$. A symmetric LU factorization of $\boldsymbol{A} := \left[\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right]$ must satisfy the equations

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_1 & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \begin{bmatrix} 1 & l_1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} d_1 & d_1 l_1 \\ l_1 d_1 & l_1^2 d_1 + d_2 \end{bmatrix}$$

for the unknowns l_1 in $\boldsymbol{L}$ and d_1, d_2 in $\boldsymbol{D}$. The equations are

$$d_1 = a. \quad l_1 a = b, \quad d_2 = c - a l_1^2.$$

As in the nonsymmetric case the main problem is the nonlinear equation. Again there are three cases

1. *$a \neq 0$: The matrix has a unique symmetric LU factorization with $l_1 = b/a$.*
2. *$a = 0, b \neq 0$: No symmetric LU factorization exists.*
3. *$a = b = 0$: The LU factorization exists, but it is not unique. Any value for l_1 can be used.*

Consider the four matrices

$$\boldsymbol{A}_1 := \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \boldsymbol{A}_2 := \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{A}_3 := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{A}_4 := \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}.$$

Then the symmetric LU factorization is unique for $\boldsymbol{A}_1$ and $\boldsymbol{A}_3$, is not unique for $\boldsymbol{A}_4$ and does not exist for $\boldsymbol{A}_2$.

In view of this example it might come as no surprise that Theorem 2.12 carries over to the symmetric case. Again we start with an lemma.

Lemma 2.25 (Symmetric LU of leading principal sub matrices)
Suppose $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ is a symmetric LU factorization of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. For $k = 1,\ldots,n$ let $\boldsymbol{A}_{[k]}, \boldsymbol{L}_{[k]}, \boldsymbol{D}_{[k]}$ be the leading principal submatrices of $\boldsymbol{A}, \boldsymbol{L}, \boldsymbol{D}$, respectively. Then $\boldsymbol{A}_{[k]} = \boldsymbol{L}_{[k]}\boldsymbol{D}_{[k]}\boldsymbol{L}_{[k]}^T$ is a symmetric LU factorization of $\boldsymbol{A}_{[k]}$ for $k = 1,\ldots,n$.

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

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{[k]} & \boldsymbol{B}_k \\ \boldsymbol{C}_k & \boldsymbol{F}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{[k]} & \boldsymbol{0} \\ \boldsymbol{M}_k & \boldsymbol{N}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_{[k]} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{L}_{[k]}^T & \boldsymbol{M}_k^T \\ \boldsymbol{0} & \boldsymbol{N}_k^T \end{bmatrix} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T, \quad (2.14)$$

where $\boldsymbol{F}_k, \boldsymbol{N}_k, \boldsymbol{E}_k \in \mathbb{R}^{n-k,n-k}$. Block multiplication gives $\boldsymbol{A}_{[k]} = \boldsymbol{L}_{[k]}\boldsymbol{D}_{[k]}\boldsymbol{L}_{[k]}^T$. Since $\boldsymbol{L}_{[k]}$ is unit lower triangular and $\boldsymbol{D}_{[k]}$ is diagonal this is a symmetric LU factorization of $\boldsymbol{A}_{[k]}$. ∎

Theorem 2.26 (Symmetric LU theorem)
The matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has a unique symmetric LU factorization if and only if $\boldsymbol{A} = \boldsymbol{A}^T$ and $\boldsymbol{A}_{[k]}$ is nonsingular for $k = 1,\ldots,n-1$.

Proof. If $\boldsymbol{A}$ is nonsingular then $\boldsymbol{D}$ is nonsingular and it can be shown that the theorem is a simple corollary of the LU theorem. To prove the general case we repeat the proof of Theorem 2.12 incorporating the necessary changes. Suppose $\boldsymbol{A}^T = \boldsymbol{A}$ and that $\boldsymbol{A}_{[k]}$ is nonsingular for $k = 1,\ldots,n-1$. Note that $\boldsymbol{A}_{[k]}^T = \boldsymbol{A}_{[k]}$ for $k = 1,\ldots,n$. We use induction on n to show that $\boldsymbol{A}$ has a unique symmetric LU factorization. The result is clearly true for $n = 1$, since the unique symmetric LU factorization of a 1-by-1 matrix is $[a_{11}] = [1][a_{11}][1]$. Suppose that $\boldsymbol{A}_{[n-1]}$ has a unique symmetric LU factorization $\boldsymbol{A}_{[n-1]} = \boldsymbol{L}_{n-1}\boldsymbol{D}_{n-1}\boldsymbol{L}_{n-1}^T$, and that $\boldsymbol{A}_{[1]},\ldots,\boldsymbol{A}_{[n-1]}$ are nonsingular. By block multiplication

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{[n-1]} & \boldsymbol{b} \\ \boldsymbol{b}^T & a_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{n-1} & \boldsymbol{0} \\ \boldsymbol{x}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_{n-1} & \boldsymbol{0} \\ 0 & d_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{L}_{n-1}^T & \boldsymbol{x} \\ \boldsymbol{0}^T & 1 \end{bmatrix}, \quad (2.15)$$

if and only if $\boldsymbol{b} = \boldsymbol{L}_{n-1}\boldsymbol{D}_{n-1}\boldsymbol{x}$ and $a_{nn} = d_{nn} + \boldsymbol{x}^T\boldsymbol{D}_{n-1}\boldsymbol{x}$. Thus we obtain a symmetric LU factorization of $\boldsymbol{A}$ that is unique since $\boldsymbol{L}_{n-1}$ and $\boldsymbol{D}_{n-1}$ are nonsingular.

For the converse we use Lemma 2.25 in the same way as Lemma 2.11 was used to prove Theorem 2.12. ∎

The number of arithmetic operations for the symmetric LU fartorization is approximately $\frac{1}{2}G_n$, half the number of operations needed for the LU factorization.

For in the LU factorization we needed to solve two triangular systems to find the vectors $\boldsymbol{s}$ and $\boldsymbol{m}$, while only one such system is needed to find $\boldsymbol{x}$ in the symmetric case (2.15). The work to find d_{nn} is $O(n)$ and does not contribute to the highest order term.

2.4 Block LU factorization

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

$$\boldsymbol{A} := \begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1m} \\ \vdots & & \vdots \\ \boldsymbol{A}_{m1} & \cdots & \boldsymbol{A}_{mm} \end{bmatrix}, \tag{2.16}$$

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

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{I} & & & \\ \boldsymbol{L}_{21} & \boldsymbol{I} & & \\ \vdots & & \ddots & \\ \boldsymbol{L}_{m1} & \cdots & \boldsymbol{L}_{m,m-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{11} & & \cdots & \boldsymbol{U}_{1m} \\ & \boldsymbol{U}_{21} & \cdots & \boldsymbol{U}_{2m} \\ & & \ddots & \vdots \\ & & & \boldsymbol{U}_{mm} \end{bmatrix} \tag{2.17}$$

a **block LU factorization of $\boldsymbol{A}$**. Here the ith diagonal blocks $\boldsymbol{I}$ and $\boldsymbol{U}_{ii}$ in $\boldsymbol{L}$ and $\boldsymbol{U}$ have the same size 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 2.27 (Block LU theorem)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is a block matrix of the form (2.16)*. Then $\boldsymbol{A}$ has a unique block LU factorization* (2.17) *if and only if 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 $k = 1, \ldots, m-1$.

Proof. Suppose $\boldsymbol{A}_{\{k\}}$ is nonsingular for $k = 1, \ldots, m-1$. Following the proof in Theorem 2.12 suppose $\boldsymbol{A}_{\{m-1\}}$ has a unique block LU factorization $\boldsymbol{A}_{\{m-1\}} = \boldsymbol{L}_{\{m-1\}}\boldsymbol{U}_{\{m-1\}}$, and that $\boldsymbol{A}_{\{1\}}, \ldots, \boldsymbol{A}_{\{m-1\}}$ are nonsingular. Then $\boldsymbol{L}_{\{m-1\}}$ and $\boldsymbol{U}_{\{m-1\}}$ are nonsingular and

$$\begin{aligned} \boldsymbol{A} &= \begin{bmatrix} \boldsymbol{A}_{\{m-1\}} & \boldsymbol{B} \\ \boldsymbol{C}^T & \boldsymbol{A}_{mm} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{L}_{\{m-1\}} & \boldsymbol{0} \\ \boldsymbol{C}^T\boldsymbol{U}_{\{m-1\}}^{-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{\{m-1\}} & \boldsymbol{L}_{\{m-1\}}^{-1}\boldsymbol{B} \\ 0 & \boldsymbol{A}_{mm} - \boldsymbol{C}^T\boldsymbol{U}_{\{m-1\}}^{-1}\boldsymbol{L}_{\{m-1\}}^{-1}\boldsymbol{B} \end{bmatrix}, \end{aligned} \tag{2.18}$$

is a block LU factorization of $\boldsymbol{A}$. It is unique by derivation. Conversely, suppose $\boldsymbol{A}$ has a unique block LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. Then as in Lemma 2.11 it is easily seen that $\boldsymbol{A}_{\{k\}} = \boldsymbol{L}_{\{k\}}\boldsymbol{U}_{\{k\}}$ is the unique block LU factorization of $\boldsymbol{A}_{[k]}$ for $k = 1, \ldots, m$. The rest of the proof is similar to the proof of Theorem 2.12. □

Remark 2.28 (Comparing LU and block LU)
The number of arithmetic operations 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 2.29 (A block LU is not an LU)
Note that (2.17) *is not an LU factorization of* $\boldsymbol{A}$ *since the* $\boldsymbol{U}_{ii}$*'s are not upper triangular in general. To relate the block LU factorization to the usual LU factorization we assume that each* $\boldsymbol{U}_{ii}$ *has an LU factorization* $\boldsymbol{U}_{ii} = \tilde{\boldsymbol{L}}_{ii}\tilde{\boldsymbol{U}}_{ii}$*. Then* $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{U}}$*, where* $\hat{\boldsymbol{L}} := \boldsymbol{L}\,\mathrm{diag}(\tilde{\boldsymbol{L}}_{ii})$ *and* $\hat{\boldsymbol{U}} := \mathrm{diag}(\tilde{\boldsymbol{L}}_{ii}^{-1})\boldsymbol{U}$*, and this is an ordinary LU factorization of* $\boldsymbol{A}$*.*

Exercise 2.30 (Making block LU into LU)
Show that $\hat{\boldsymbol{L}}$ *is unit lower triangular and* $\hat{\boldsymbol{U}}$ *is upper triangular.*

2.5 Positive Definite and Semidefinite Matrices

Symmetric positive definite matrices occur often in scientific computing. In this section we study some properties of positive definite matrices. We study only real matrices, but consider both the symmetric and nonsymmetric case.

2.5.1 Definitions and examples

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

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

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

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0. \tag{2.19}$$

- 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 (2.19) implies that $\boldsymbol{x} = \boldsymbol{0}$.

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

We considered only real valued vectors $\boldsymbol{x}$ above. For symmetric matrices we have:

Lemma 2.31 (Quadratic form with $\boldsymbol{x} \in \mathbb{C}^n$)
If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite then $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{C}^n$.

Proof. Suppose $\boldsymbol{x} := \boldsymbol{y} + i\boldsymbol{z}$ is nonzero with $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $i := \sqrt{-1}$. Since $\boldsymbol{A}$ is symmetric we find $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{y}-i\boldsymbol{z})^T\boldsymbol{A}(\boldsymbol{y}+i\boldsymbol{z}) = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}+\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}$. This is positive since $\boldsymbol{A}$ is positive definite and at least one of the vectors $\boldsymbol{y}, \boldsymbol{z}$ is nonzero. □

Example 2.32 (2×2 positive definite)
The family of matrices

$$\boldsymbol{A}[a] := \begin{bmatrix} 2 & 2-a \\ a & 1 \end{bmatrix}, \quad a \in \mathbb{R}$$

is positive definite for any $a \in \mathbb{R}$. Indeed for any nonzero $\boldsymbol{x} = [x_1, x_2]^T \in \mathbb{R}^2$

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 2x_1^2 + (2-a)x_1x_2 + ax_2x_1 + x_2^2 = x_1^2 + (x_1+x_2)^2 > 0.$$

Lemma 2.33 ($\boldsymbol{T}$ is symmetric positive definite)
The second derivative matrix $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ 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. ∎

Symmetric positive definite matrices is important in nonlinear optimization.

Example 2.34 (Gradient and hessian)
Consider (cf. (C.1)*) the gradient ∇f and hessian Hf of a function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$*

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

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 $Hf(\boldsymbol{x})$ is symmetric positive definite then $\boldsymbol{x}$ is a local minimum for f. This can be shown using the second-order Taylor expansion (C.2)*. Moreover, $\boldsymbol{x}$ is a local maximum if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ and $Hf(\boldsymbol{x})$ is negative definite.*

Theorem 2.35 (A general criterium)
Let m, n be positive integers. If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is (symmetric) positive semidefinite then $\boldsymbol{B} := \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \in \mathbb{R}^{m\times m}$ is (symmetric) positive semidefinite for any $\boldsymbol{X} \in \mathbb{R}^{n\times m}$. If in addition $\boldsymbol{A}$ is (symmetric) positive definite and $\boldsymbol{X}$ has linearly independent columns then $\boldsymbol{B}$ is (symmetric) positive definite.

Proof. Let $\boldsymbol{y} \in \mathbb{R}^m$ and set $\boldsymbol{x} := \boldsymbol{X}\boldsymbol{y} \in \mathbb{R}^n$. Then $\boldsymbol{y}^T\boldsymbol{B}\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$. This is nonnegative if $\boldsymbol{A}$ is positive semidefinite and positive if $\boldsymbol{A}$ is positive definite and $\boldsymbol{X}$ has linearly independent columns. For then $\boldsymbol{x}$ is nonzero if $\boldsymbol{y}$ is nonzero. If $\boldsymbol{A}$ is symmetric then $\boldsymbol{B}$ is symmetric and the statements about symmetry follows. ∎

Corollary 2.36 ($\boldsymbol{A}^T\boldsymbol{A}$ is symmetric positive semidefinite)
Let m, n be positive integers. If $\boldsymbol{A} \in \mathbb{R}^{m,n}$ then $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric positive semidefinite. It is symmetric positive definite if and only if $\boldsymbol{A}$ has linearly independent columns.

Proof. If $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric positive definite then $\boldsymbol{x}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 > 0$ for all nonzero $\boldsymbol{x}$ and $\boldsymbol{A}$ has linearly independent columns. Taking $\boldsymbol{A} := \boldsymbol{I}$ and $\boldsymbol{X} := \boldsymbol{A}$ in Theorem 2.35 gives $\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}^T\boldsymbol{A}$, and we obtain the remaining statements of the corollary. □

2.5.2 The nonsymmetric case

A positive definite matrix has the following properties:

Theorem 2.37 (The nonsymmetric case)
Suppose $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is a positive definite matrix and let $\boldsymbol{B}$ be a principal submatrix. Then

1. *$\boldsymbol{B}$ is positive definite,*
2. *$\boldsymbol{A}$ has a unique LU factorization,*
3. *the real eigenvalues of $\boldsymbol{B}$ are positive,*
4. $\det(\boldsymbol{B}) > 0$,
5. $a_{ii}a_{jj} > a_{ij}a_{ji}$, *for* $i \neq j$.

Proof.

1. Suppose the submatrix $\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{r}, \boldsymbol{r})$ is defined by the rows and columns $\boldsymbol{r} = [r_1, \ldots, r_k]^T$ of $\boldsymbol{A}$. Let $\boldsymbol{X} = [\boldsymbol{e}_{r_1}, \ldots, \boldsymbol{e}_{r_k}] \in \mathbb{R}^{n\times k}$. Then $\boldsymbol{B} := \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$, and $\boldsymbol{B}$ is positive definite by Theorem 2.35.

2. Since all leading submatrices are nonsingular this follows from the LU Theorem 2.12.

3. Suppose $(\lambda, \boldsymbol{x})$ is an eigenpair of $\boldsymbol{A}$ and that λ is real. Since $\boldsymbol{A}$ is real we can choose $\boldsymbol{x}$ to be real. Multiplying $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ by $\boldsymbol{x}^T$ and solving for λ we find $\lambda = \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} > 0$.

4. The determinant of $\boldsymbol{B}$ equals the product of its eigenvalues. The eigenvalues are either real and positive or occur in complex conjugate pairs. The product of two nonzero complex conjugate numbers is positive.

5. The principal submatrix $\left[\begin{smallmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{smallmatrix}\right]$ has a positive determinant.

□

We note that:

1. Part 5 of Theorem 2.37 implies that all diagonal elements of a positive definite matrix are positive. Moreover, the greatest element in absolute value is a diagonal element. This can be used to decide by inspection that a given matrix cannot be positive definite.

2. A nonsymmetric positive definite matrix can have complex eigenvalues. For example, the eigenvalues of $\boldsymbol{A}[a]$ in Example 2.32 are positive for $a \in [1 - \frac{\sqrt{5}}{2}, 1 + \frac{\sqrt{5}}{2}]$ and complex for other values of $\boldsymbol{A}$.

2.5.3 The symmetric case

Theorem 2.37 can be strengthened considerably when $\boldsymbol{A}$ is symmetric positive definite.

Theorem 2.38 (Symmetric positive definite characterization)
The following statements are equivalent for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

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

Proof. $1 \Leftrightarrow 2$ is shown in Lemma 2.41 below. We show that $1 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$.
1 $\Rightarrow$ **3**: This follows from Theorem 2.37
3 $\Rightarrow$ **4**: By Lemma 2.42 below $\boldsymbol{A}$ has a unique symmetric LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ with positive diagonal elements in $\boldsymbol{D}$. But then $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^T$, where $\boldsymbol{B} := \boldsymbol{L}\boldsymbol{D}^{1/2}$, with $\boldsymbol{D}^{1/2} := \operatorname{diag}(d_{1,1}^{1/2}, \ldots, d_{n\times n}^{1/2})$.
4 $\Rightarrow$ **1**: This follows from Corollary 2.36. □

Exercise 2.39 (Positive definite characterizations)
Show directly that all 4 characterizations in Theorem 2.38 hold for the matrix $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

Consider the eigenvalues of a real symmetric positive definite matrix. Note that such a matrix is Hermitian.

Lemma 2.40 (Eigenvalues of a Hermitian matrix)
All eigenvalues of a Hermitian matrix are real.

Proof. Suppose $\boldsymbol{A}^* = \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\boldsymbol{x} \neq 0$. We multiply both sides of $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ from the left 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. □

Lemma 2.41 (Symmetry and positive eigenvalues)
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite if and only if it is symmetric and all eigenvalues are positive.

Proof. By Lemma 2.40 all eigenvalues of $\boldsymbol{A}$ are real, and by Theorem 2.37 all eigenvalues are positive. Suppose conversely that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric with positive eigenvalues $\lambda_1, \ldots, \lambda_n$. By the spectral theorem (cf. Corollary 5.23) we have $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}$, where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. 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 $\boldsymbol{c}^T\boldsymbol{c} = \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x}$ so $\boldsymbol{c}$ is nonzero. Since $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{c}$ we find

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{U}\boldsymbol{c})^T\boldsymbol{A}\boldsymbol{U}\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{D}\boldsymbol{c} = \sum_{j=1}^{n} \lambda_j c_j^2 > 0$$

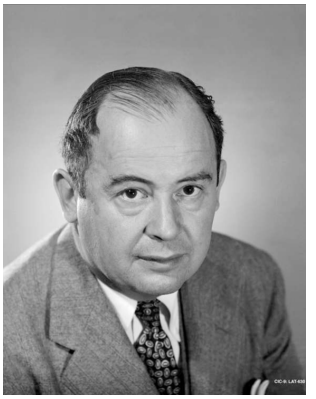
and it follows that $\boldsymbol{A}$ is positive definite. □

Lemma 2.42 (Symmetric positve definite and symmetric LU)
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric positive definite if and only if it has a symmetric LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ with positive diagonal elements in $\boldsymbol{D}$.

Proof. Suppose $\boldsymbol{A}$ is symmetric positive definite. By Theorem 2.37 the leading principal submatrices $\boldsymbol{A}_{[k]} \in \mathbb{R}^{k\times k}$ are nonsingular for $k = 1, \ldots, n-1$, and $\boldsymbol{A}$ has a unique symmetric LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ by Theorem 2.26. The ith diagonal element in $\boldsymbol{D}$ is positive, since $d_{ii} = \boldsymbol{e}_i^T\boldsymbol{D}\boldsymbol{e}_i = \boldsymbol{e}_i^T\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{L}^{-T}\boldsymbol{e}_i = \boldsymbol{x}_i^T\boldsymbol{A}\boldsymbol{x}_i > 0$. Indeed, $\boldsymbol{x}_i := \boldsymbol{L}^{-T}\boldsymbol{e}_i$ is nonzero since $\boldsymbol{L}^{-T}$ is nonsingular.

Conversely, suppose $\boldsymbol{A}$ has a symmetric LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ with positive diagonal elements in $\boldsymbol{D}$. Then $\boldsymbol{A}$ is symmetric, and for any nonzero $\boldsymbol{y} \in \mathbb{R}^n$ we have $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y} > 0$, since $\boldsymbol{y} := \boldsymbol{L}^T\boldsymbol{x} \neq \boldsymbol{0}$. □

2.6 The Cholesky Factorization

André-Louis Cholesky, 1875-1918 (left), John von Neumann, 1903-1957 (right).

Lemma 2.42 implies that $\boldsymbol{A}$ is symmetric positive definite if and only if it has a symmetric LU factorization, and from the proof of 3. implies 4 in that theorem we can write this in the form $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^T$ where $\boldsymbol{B}$ is lower triangular matrix with positive diagonal elements. Such a factorization has a special name.

Definition 2.43 (Cholesky)
A factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is lower triangular with positive diagonal elements is called a **Cholesky factorization** *of $\boldsymbol{A}$. The matrix $\boldsymbol{L}$ is called a* **Cholesky factor**.

From the discussion before the definition we have

Theorem 2.44 (Cholesky)
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has a Cholesky factorization if and only if it is symmetric positive definite. Moreover, the Cholesky factorization is unique.

Proof. We still need to show uniqueness. Suppose $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{S}\boldsymbol{S}^T$ are two Cholesky factorizations of the symmetric positive definite matrix $\boldsymbol{A}$. Since $\boldsymbol{A}$ is nonsingular both $\boldsymbol{L}$ and $\boldsymbol{S}$ are nonsingular. Then $\boldsymbol{S}^{-1}\boldsymbol{L} = \boldsymbol{S}^T\boldsymbol{L}^{-T}$, where by Lemma 1.22 $\boldsymbol{S}^{-1}\boldsymbol{L}$ is lower triangular and $\boldsymbol{S}^T\boldsymbol{L}^{-T}$ is upper triangular, with diagonal elements ℓ_{ii}/s_{ii} and s_{ii}/ℓ_{ii}, respectively. But then both matrices must be equal to the same diagonal matrix and $\ell_{ii}^2 = s_{ii}^2$. By positivity $\ell_{ii} = s_{ii}$ and we conclude that $\boldsymbol{S}^{-1}\boldsymbol{L} = \boldsymbol{I} = \boldsymbol{S}^T\boldsymbol{L}^{-T}$ which means that $\boldsymbol{L} = \boldsymbol{S}$. □

A Cholesky factorization can also be written in the equivalent form $\boldsymbol{A} =$

$\boldsymbol{R}^T\boldsymbol{R}$, where $\boldsymbol{R} = \boldsymbol{L}^T$ is upper triangular with positive diagonal elements. The matrix $\boldsymbol{A}$ must be symmetric since $\boldsymbol{L}\boldsymbol{L}^T$ is symmetric.

Example 2.45 (2×2)
The matrix $\boldsymbol{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ *has a symmetric LU- and a Cholesky-factorization given by*

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 \\ -1/\sqrt{2} & \sqrt{3/2} \end{bmatrix} \begin{bmatrix} \sqrt{2} & -1/\sqrt{2} \\ 0 & \sqrt{3/2} \end{bmatrix}.$$

Consider computing the Cholesky factorization directly. The equation $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ implies that

$$a_{ik} = \sum_{j=1}^{n} \ell_{ij}\ell_{kj} = \sum_{j=1}^{\min(i,k)} \ell_{ij}\ell_{kj}, \quad i,k = 1,\ldots,n. \tag{2.20}$$

The unknown elements in $\boldsymbol{L}$ can be computed row by row or column by column. Consider the column case. Suppose we have computed the $k-1$ first columns of $\boldsymbol{L}$. The kth column can then be computed from (2.20). Indeed, letting $i = k$ and solving for ℓ_{kk} we find

$$\ell_{kk} = \big(a_{kk} - \sum_{j=1}^{k-1} \ell_{kj}^2\big)^{1/2}, \tag{2.21}$$

and similarly for $i > k$

$$\ell_{ik} = \big(a_{ik} - \sum_{j=1}^{k-1} \ell_{ij}\ell_{kj}\big)/\ell_{kk} \quad i = k+1,\ldots,n. \tag{2.22}$$

Since $\boldsymbol{A}$ is symmetric positive definite the Cholesky factor $\boldsymbol{L}$ exists, is unique, and real, and therefore the term under the square root in (2.21) must be positive. We note however that we can encounter problems in floating point computation if the term is very small.

It is easily seen that the Cholesky-factorization of an n-by-n matrix based on (2.21) and (2.22) requires $\frac{1}{2}G_n = n^3/3$ arithmetic operations. This is essentially the same as for the symmetric LU factorization. The halving of the count compared to LU factorization is due to the symmetry of $\boldsymbol{A}$.

If $\boldsymbol{A}$ is d-banded then the same is true for the Cholesky factor.

Lemma 2.46 (Banded Cholesky factor)
Suppose $\boldsymbol{A}$ *is symmetric positive definite with Cholesky-factor* $\boldsymbol{L}$. *If* $a_{ik} = 0$ *for* $i > k + d$, *then* $\ell_{ik} = 0$ *for* $i > k + d$.

Proof. We show that if $\boldsymbol{L}$ has bandwidth d in its first $k-1$ columns then column k also has bandwidth d. The proof then follows by induction on k. Now, if $i > k+d$, then $a_{ik} = 0$, and if $\boldsymbol{L}$ has bandwidth d in its first $k-1$ columns then $\ell_{ij} = 0$ for $j = 1, \ldots, k-1$. By (2.22) $\ell_{ik} = 0$. □

We obtain formulas for the Cholesky factorization of a symmetric positive definite band matrix by simply replacing the lower bound $j = 1$ by $j = \max(1, k-d)$ in (2.21) and (2.22) and letting i run from $k+1$ to $\min(n, k+d)$ in (2.22). The lower triangular matrix $\boldsymbol{L}$ is computed in sparse form. Only the lower triangular part of $\boldsymbol{A}$ is used. This leads to the following algorithm. For a different algorithm based on outer products, which can also be used for symmetric positive semidefinite matrices, see Algorithm 2.53.

Algorithm 2.47 (bandcholesky)

```
function L=bandcholesky(A,d)
%L=bandcholesky(A,d)
n=length(A);
L=sparse(zeros(n,n));
for k=1:n
    km=max(1,k-d); kp=min(n,k+d); s=L(k,km:k-1);
    L(k,k)=sqrt(A(k,k)-s*s');
    L(k+1:kp,k)=(A(k+1:kp,k) -...
      L(k+1:kp,km:k-1)*s')/L(k,k);
end
```

The leading term in an operation count for a band matrix is $O(d^2 n)$. When d is small this is a considerable saving compared to the count $\frac{1}{2}G_n = n^3/3$ for a full matrix.

There is also a banded version of the symmetric LU factorization which requires approximately the same number of arithmetic operations as the Cholesky factorization. The choice between using a symmetric LU factorization or an $\boldsymbol{LL}^T$ factorization depends on several factors. Usually an LU or a symmetric LU factorization is preferred for matrices with small bandwidth (tridiagonal, pentadiagonal), while the $\boldsymbol{LL}^T$ factorization is restricted to symmetric positive definite matrices and is often used when the bandwidth is larger.

2.7 The Symmetric Positive Semidefinite Case

We start with the following necessary conditions for a matrix to be symmetric positive semidefinite. It shows that if a diagonal element a_{ii} is zero then all elements in row i and column i are zero.

Lemma 2.48 (Criteria symmetric semidefinite)
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. Part 3 follows from part 1 and part 4 from part 2. Now

$$0 \leq (\alpha \boldsymbol{e}_i + \beta \boldsymbol{e}_j)^T \boldsymbol{A} (\alpha \boldsymbol{e}_i + \beta \boldsymbol{e}_j) = \alpha^2 a_{ii} + \beta^2 a_{jj} + 2\alpha\beta a_{ij}, \text{ all } i, j, \ \alpha, \beta \in \mathbb{R}, \quad (2.23)$$

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 part 1. Part 2 follows trivially from part 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 (2.23) 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 part 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 part 1 nor part 2 hold.

Theorem 2.49 (Positive symmetric semidefinite characterization)
The following is equivalent for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.

1. $\boldsymbol{A}$ *is positive semidefinite.*
2. $\boldsymbol{A}$ *has only nonnegative eigenvalues.*
3. $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^T$ *for some* $\boldsymbol{B} \in \mathbb{R}^{n \times n}$.
4. *All principal submatrices have a nonnegative determinant.*

Proof. The proof of 1 $\Leftrightarrow$ 2 follows as in the proof of Theorem 2.38. 1. $\Leftrightarrow$ 3. follows from Theorem 2.51 while 1. $\Rightarrow$ 4. is a consequence of Theorem 2.37. 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 [24]). 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 part 4 of Theorem 2.49 we require nonnegativity of all principal minors, while only positivity of leading principal minors was required for positive definite matrices (cf. Theorem 2.38). 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.

2.8 Semi-Cholesky factorization of a banded matrix

A symmetric positive semidefinite matrix has a factorization that is similar to the Cholesky factorization.

Definition 2.50 (Semi-Cholesky factorization)
A factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is lower 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 $\boldsymbol{L}$ must be nonsingular. Thus the diagonal elements of $\boldsymbol{L}$ cannot be zero.

Theorem 2.51 (Characterization, semi-Cholesky factorization)
A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has a semi-Cholesky factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ if and only if it is symmetric positive semidefinite.

Proof. If $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ is a semi-Cholesky factorization then $\boldsymbol{A}$ is symmetric positive semidefinite by Corollary 2.36. For the converse we use induction on n. A positive semidefinite matrix of order one has a semi-Cholesky factorization since the one and only element in $\boldsymbol{A}$ is nonnegative. Suppose any symmetric positive semidefinite matrix of order $n-1$ has a semi-Cholesky factorization and suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric positive semidefinite. We partition $\boldsymbol{A}$ as follows

$$\boldsymbol{A} = \begin{bmatrix} \alpha & \boldsymbol{v}^T \\ \boldsymbol{v} & \boldsymbol{B} \end{bmatrix}, \quad \alpha \in \mathbb{R},\ \boldsymbol{v} \in \mathbb{R}^{n-1},\ \boldsymbol{B} \in \mathbb{R}^{(n-1)\times(n-1)}. \tag{2.24}$$

There are two cases. Suppose first $\alpha = \boldsymbol{e}_1^T\boldsymbol{A}\boldsymbol{e}_1 > 0$. We claim that $\boldsymbol{C} := \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T/\alpha$ is symmetric positive semidefinite. $\boldsymbol{C}$ is symmetric. To show that $\boldsymbol{C}$ is positive semidefinite we consider any $\boldsymbol{y} \in \mathbb{R}^{n-1}$ and define $\boldsymbol{x}^T := [-\boldsymbol{v}^T\boldsymbol{y}/\alpha, \boldsymbol{y}^T] \in \mathbb{R}^n$. Then

$$\begin{aligned} 0 \le \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= [-\boldsymbol{v}^T\boldsymbol{y}/\alpha, \boldsymbol{y}^T] \begin{bmatrix} \alpha & \boldsymbol{v}^T \\ \boldsymbol{v} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} -\boldsymbol{v}^T\boldsymbol{y}/\alpha \\ \boldsymbol{y} \end{bmatrix} \\ &= [0, -(\boldsymbol{v}^T\boldsymbol{y})\boldsymbol{v}^T/\alpha + \boldsymbol{y}^T\boldsymbol{B}] \begin{bmatrix} -\boldsymbol{v}^T\boldsymbol{y}/\alpha \\ \boldsymbol{y} \end{bmatrix} \\ &= -(\boldsymbol{v}^T\boldsymbol{y})(\boldsymbol{v}^T\boldsymbol{y})/\alpha + \boldsymbol{y}^T\boldsymbol{B}\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}, \end{aligned} \tag{2.25}$$

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 $\boldsymbol{C} \in \mathbb{R}^{(n-1)\times(n-1)}$ is symmetric positive semidefinite and by the induction hypothesis it has a semi-Cholesky factorization $\boldsymbol{C} = \boldsymbol{L}_1\boldsymbol{L}_1^T$. The matrix

$$\boldsymbol{L}^T := \begin{bmatrix} \beta & \boldsymbol{v}^T/\beta \\ \boldsymbol{0} & \boldsymbol{L}_1^T \end{bmatrix}, \quad \beta := \sqrt{\alpha}, \tag{2.26}$$

is upper triangular with nonnegtive diagonal elements and

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

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

If $\alpha = 0$ then it follows from 4. in Lemma 2.48 that $\boldsymbol{v} = \boldsymbol{0}$. Moreover, $\boldsymbol{B} \in \mathbb{R}^{(n-1)\times(n-1)}$ in (2.24) is positive semidefinite and therefore has a semi-Cholesky factorization $\boldsymbol{B} = \boldsymbol{L}_1\boldsymbol{L}_1^T$. But then $\boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L} = \begin{bmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{L}_1 \end{bmatrix}$ is a semi-Cholesky factorization of $\boldsymbol{A}$. Indeed, $\boldsymbol{L}$ is lower triangular and

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

□

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

Theorem 2.52 (Bandwidth semi-Cholesy factor)
The semi-Cholesky factor $\boldsymbol{L}$ given by (2.26) *has the same bandwidth as $\boldsymbol{A}$.*

Proof. Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is d-banded. Then $\boldsymbol{v}^T = [\boldsymbol{u}^T, \boldsymbol{0}^T]$ in (2.24), 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, $\boldsymbol{C} = \boldsymbol{L}_1\boldsymbol{L}_1^T$, where $\boldsymbol{L}_1^T$ has the same bandwidth as $\boldsymbol{C}$. But then $\boldsymbol{L}$ in (2.26) has the same bandwidth as $\boldsymbol{A}$. □

Algorithm 2.53 (bandsemi-cholesky)
Suppose $\boldsymbol{A}$ is symmetric positive semidefinite. A lower triangular matrix $\boldsymbol{L}$ is computed so that $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$. 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 `tril`.

```
function L=bandsemicholeskyL(A,d)
%L=bandsemicholeskyL(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+1:kp,k)=A(k+1:kp,k)/A(k,k);
        for j=k+1:kp
            A(j:kp,j)=A(j:kp,j)-A(j,k)*A(j:kp,k);
        end
    else
        A(k:kp,k)=zeros(kp-k+1,1);
    end
end
L=tril(A);
```

Consider now implementing an algorithm based on the previous discussion. Since $\boldsymbol{A}$ is symmetric we only need to use the lower part of $\boldsymbol{A}$. The first column of $\boldsymbol{L}$ is $[\beta, \boldsymbol{v}^T/\beta]^T$ if $\alpha > 0$. If $\alpha = 0$ then by 4 in Lemma 2.48 the first column of $\boldsymbol{A}$ is zero and this is also the first column of $\boldsymbol{L}$. We obtain

$$\boxed{\begin{array}{l}\text{if } A(1,1) > 0 \\ \quad A(1,1) = \sqrt{A(1,1)} \\ \quad A(2:n,1) = A(2:n,1)/A(1,1) \\ \quad \text{for } j = 2:n \\ \quad\quad A(j:n,j) = A(j:n,j) - A(j,1) * A(j:n,1)\end{array}} \tag{2.27}$$

Here we store the first column of $\boldsymbol{L}$ in the first column of $\boldsymbol{A}$ and the lower part of $\boldsymbol{C} = \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T/\alpha$ in the lower part of $A(2:n, 2:n)$.

The code can be made more efficient when $\boldsymbol{A}$ is a d-banded matrix. We simply replace all occurrences of n by $\min(i+d, n)$. Continuing the reduction we arrive at Algorithm 2.53.

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

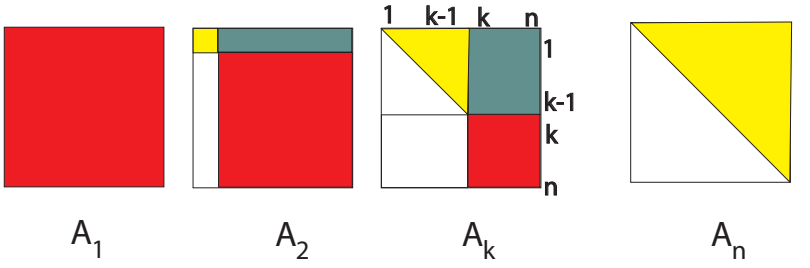


Figure 2.2. *Gaussian elimination*

There are many versions of Cholesky factorizations, see [5]. Algorithm 2.47 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. However deciding when a diagonal element is zero is a problem in floating point arithmetic.

2.9 Gaussian Elimination

In this section we take a closer look at Gaussian elimination. We show that if the conditions of the LU theorem is satisfied then Gaussian elimination without row interchanges is just a way of computing the LU factorization of the coefficient matrix. If row interchanges are incorporated then we need to introduce a matrix permutation matrix $\boldsymbol{P}$, and obtain a factorization of the form $\boldsymbol{A} = \boldsymbol{PLU}$. We obtain this **PLU factorization** by using a matrix formulation of Gaussian elimination.

2.9.1 Reduction to upper triangular form

Consider the general $n \times n$ case (see Example 1.1 for the 3 by 3 case.). We start with a nonsingular 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 $\boldsymbol{A}_n\boldsymbol{x} = \boldsymbol{b}_n$ can be solved using one of Algorithms 2.2 or 2.5. The process is illustrated in Figure 2.2.

The matrix $\boldsymbol{A}_k$ has the form

$$\boldsymbol{A}_k = \left[\begin{array}{ccc|ccccc} a_{1,1}^1 & \cdots & a_{1,k-1}^1 & a_{1,k}^1 & \cdots & a_{1,j}^1 & \cdots & a_{1,n}^1 \\ & \ddots & \vdots & \vdots & & \vdots & & \vdots \\ & & a_{k-1,k-1}^{k-1} & a_{k-1,k}^{k-1} & \cdots & a_{k-1,j}^{k-1} & \cdots & a_{k-1,n}^{k-1} \\ \hline & & & a_{k,k}^k & \cdots & a_{k,j}^k & \cdots & a_{k,n}^k \\ & & & \vdots & & \vdots & & \vdots \\ & & & a_{i,k}^k & \cdots & a_{i,j}^k & \cdots & a_{i,n}^k \\ & & & \vdots & & \vdots & & \vdots \\ & & & a_{n,k}^k & \cdots & a_{n,j}^k & \cdots & a_{n\times n}^k \end{array}\right]. \tag{2.28}$$

The process transforming $\boldsymbol{A}_k$ into $\boldsymbol{A}_{k+1}$ for $k = 1, \ldots, n-1$ can be described as follows:

$$\boxed{\begin{array}{l} \text{for } k = 1 : n-1 \\ \quad \text{Find } r_k \geq k \text{ such that } a_{r_k,k} \neq 0; \\ \quad \text{Interchange row } k \text{ and } r_k \text{ of } \boldsymbol{A}_k; \\ \quad \text{for } i = k+1 : n \\ \qquad m_{ik} = a_{ik}^k / a_{kk}^k \\ \qquad \text{for } j = k : n \\ \qquad\quad a_{ij}^{k+1} = a_{ij}^k - m_{ik} a_{kj}^k \end{array}} \tag{2.29}$$

Since $a_{ik}^{k+1} = a_{ik}^k - m_{ik} a_{kk}^k = 0$ for $i = k+1, \ldots, n$ it follows that $\boldsymbol{A}_{k+1}$ will have zeros under the diagonal in its first k columns and the elimination is carried one step further. The numbers m_{ik} in (2.29) are called **multipliers**. Interchanging two rows (and/or two columns) during Gaussian elimination is known as **pivoting**. The element which is moved to the diagonal position (k, k) is called the **pivot element** or **pivot** for short, and the row containing the pivot is called the **pivot row**.

2.9.2 Pivot strategies

The most common pivoting strategy used in the Gaussian elimination process (2.29) is

$$|a_{r_k,k}^k| := \max\{|a_{i,k}^k| : k \leq i \leq n\}$$

with r_k the smallest such index in case of a tie. This is known as **partial pivoting**. The following example illustrating that small pivots should be avoided.

Example 2.54 (Row pivoting)
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.

Related to partial pivoting is **scaled partial pivoting**. Here r_k is the smallest index such that

$$\frac{|a^k_{r_k,k}|}{s_k} := \max\{\frac{|a^k_{i,k}|}{s_k} : k \le i \le n\}, \quad s_k := \max_{1 \le j \le n} |a_{kj}|.$$

This can sometimes give more accurate results if the coefficient matrix have coefficients of wildly different sizes. Note that the scaling factors s_k are computed using the initial matrix.

It also 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 numerically stable than partial pivoting, but requires a lot of search and is seldom used in practice.

2.9.3 Permutation matrices

Pivoting can be described in terms of permutation matrices.

Definition 2.55 *Let the components of $\boldsymbol{p} = [k_1, \ldots, k_n]^T$ be a permutation of the components of $[1, 2, \ldots, n]^T$. We call $\boldsymbol{P} := \boldsymbol{I}(:, \boldsymbol{p}) = [\boldsymbol{e}_{k_1}, \boldsymbol{e}_{k_2}, \ldots, \boldsymbol{e}_{k_n}] \in \mathbb{R}^{n\times n}$ a* **permutation matrix**. *When discussing Gaussian elimination a permutation $\boldsymbol{p}$ is sometimes called a* **pivot vector**.

Since $\boldsymbol{P}^T = \boldsymbol{I}(\boldsymbol{p}, :)$ it follows that $(\boldsymbol{P}^T\boldsymbol{P})_{i,j} = \boldsymbol{e}^T_{k_i}\boldsymbol{e}_{k_j} = \delta_{ij}$. Thus $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$, the inverse of $\boldsymbol{P}$ is equal to its transpose, and $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}$ as well. If $\boldsymbol{p}$ and $\boldsymbol{P}$ are as in Definition 2.55 and $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ then

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A}(:, \boldsymbol{p}), \quad \boldsymbol{P}^T\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{p}, :), \quad \boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{p}, \boldsymbol{p}). \tag{2.30}$$

Thus, post-multiplying a matrix $\boldsymbol{A}$ by a permutation matrix results in a permutation of the columns, pre-multiplying by the transpose of a permutation matrix gives a permutation of the rows, while the transformation $\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$ permutes both the rows and columns using the same permutation $\boldsymbol{p}$. In particular, the diagonal of $\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$ is a permutation of the diagonal of $\boldsymbol{A}$:

$$\operatorname{diag}(\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}) = \operatorname{diag}\big(\boldsymbol{A}\big)(\boldsymbol{p}). \tag{2.31}$$

We will use a particularly simple permutation matrix.

Definition 2.56 (Interchange matrix)
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 $\boldsymbol{I}^2_{jk} = \boldsymbol{I}$ 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, post-multiplication interchanges two columns. By (2.31) the diagonal of $\boldsymbol{I}_{j,k}\boldsymbol{A}\boldsymbol{I}_{j,k}$ is almost conserved; only the diagonal elements $a_{j,j}$ and $a_{k,k}$ are interchanged.

2.9.4 Gauss transformations

The elimination process in (2.29) can be interpreted in matrix terms using interchange matrices and Gauss transformations.

Definition 2.57 (Gauss transformation)
Suppose for some $1 \leq k < n$ *that* $\boldsymbol{g}_k = [0, \ldots, 0, g_{k+1,k}, \ldots, g_{n,k}]^T \in \mathbb{R}^n$ *has its first* k *components equal to zero and let* $\boldsymbol{e}_k$ *be the* k*'th unit vector in* $\mathbb{R}^n$*. The matrix*

$$\boldsymbol{G}_k := \boldsymbol{I} - \boldsymbol{g}_k \boldsymbol{e}_k^T$$

is called a **Gauss transformation**. *The name* **elementary lower triangular matrix** *is also used.*

In the 3 by 3 case the Gauss transforms takes the form

$$\boldsymbol{G}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -g_{2,1} & 1 & 0 \\ -g_{3,1} & 0 & 1 \end{bmatrix}, \quad \boldsymbol{G}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -g_{3,2} & 1 \end{bmatrix},$$

where $\boldsymbol{g}_1 = \begin{bmatrix} 0 & g_{21} & g_{31} \end{bmatrix}^T$ and $\boldsymbol{g}_3 = \begin{bmatrix} 0 & 0 & g_{32} \end{bmatrix}^T$.

For general n and $1 \leq k < n$ a Gauss transformation can be used to zero out column k under the diagonal in the matrix $\boldsymbol{A}_k$. Column k in $\boldsymbol{A}_k$ is transformed into column k of $\boldsymbol{A}_{k+1}$ using a matrix $\boldsymbol{M}_k^-$ as follows:

$$\begin{bmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & -m_{k+1,k} & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & -m_{n,k} & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_{1,k-1}^1 \\ \vdots \\ \\ a_{k,k}^k \\ a_{k+1,k}^k \\ \vdots \\ a_{n,k}^k \end{bmatrix} = \begin{bmatrix} a_{1,k-1}^1 \\ \vdots \\ \\ a_{k,k}^k \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.32}$$

The transformation matrix $\boldsymbol{M}_k^-$ in (2.32) is a Gauss transformation:

$$\boldsymbol{M}_k^- := \boldsymbol{I} - \boldsymbol{m}_k \boldsymbol{e}_k^T \quad , \quad \boldsymbol{m}_k = [0, \ldots, 0, m_{k+1,k}, \ldots, m_{n,k}]^T. \tag{2.33}$$

We collect some properties of Gauss transformations that we will need.

Lemma 2.58 (Gausstransformations)
For $1 \leq k < n$ *let* $\boldsymbol{g}_k = [0, \ldots, 0, g_{k+1,k}, \ldots, g_{n,k}]^T \in \mathbb{R}^n$*. Then*

1. $(\boldsymbol{I} - \boldsymbol{g}_k \boldsymbol{e}_k^T)^{-1} = \boldsymbol{I} + \boldsymbol{g}_k \boldsymbol{e}_k^T$,
2. $(\boldsymbol{I} - \boldsymbol{g}_1 \boldsymbol{e}_1^T) \cdots (\boldsymbol{I} - \boldsymbol{g}_k \boldsymbol{e}_k^T) = \boldsymbol{I} - \sum_{j=1}^k \boldsymbol{g}_j \boldsymbol{e}_j^T$,

3. $\boldsymbol{I}_{i,j}(\boldsymbol{I} - \boldsymbol{g}_k\boldsymbol{e}_k^T)\boldsymbol{I}_{i,j} = \boldsymbol{I} - (\boldsymbol{I}_{i,j}\boldsymbol{g}_k)\boldsymbol{e}_k^T$, *for* $k < i, j \le n$.

Proof. We note that

$$\boldsymbol{e}_j^T\boldsymbol{g}_k = 0 \text{ for } j = 1, 2, \ldots, k. \tag{2.34}$$

1. By direct multiplication using (2.34)

$$(\boldsymbol{I} - \boldsymbol{g}_k\boldsymbol{e}_k^T)(\boldsymbol{I} + \boldsymbol{g}_k\boldsymbol{e}_k^T) = \boldsymbol{I} - \boldsymbol{g}_k\boldsymbol{e}_k^T + \boldsymbol{g}_k\boldsymbol{e}_k^T - \boldsymbol{g}_k(\boldsymbol{e}_k^T\boldsymbol{g}_k)\boldsymbol{e}_k^T = \boldsymbol{I}.$$

2. Part 2. clearly holds for $k = 1$. Assuming by induction the result for $k-1$ we obtain by (2.34)

$$(\boldsymbol{I} - \boldsymbol{g}_1\boldsymbol{e}_j^T)\cdots(\boldsymbol{I} - \boldsymbol{g}_k\boldsymbol{e}_k^T) = (\boldsymbol{I} - \sum_{j=1}^{k-1}\boldsymbol{g}_j\boldsymbol{e}_j^T)(\boldsymbol{I} - \boldsymbol{g}_k\boldsymbol{e}_k^T)$$
$$= \boldsymbol{I} - \sum_{j=1}^{k-1}\boldsymbol{g}_j\boldsymbol{e}_j^T - \boldsymbol{g}_k\boldsymbol{e}_k^T + \sum_{j=1}^{k-1}\boldsymbol{g}_j(\boldsymbol{e}_j^T\boldsymbol{g}_k)\boldsymbol{e}_k^T = \boldsymbol{I} - \sum_{j=1}^{k}\boldsymbol{g}_j\boldsymbol{e}_j^T.$$

3. We find $\boldsymbol{I}_{i,j}(\boldsymbol{I} - \boldsymbol{g}_k\boldsymbol{e}_k^T)\boldsymbol{I}_{i,j} = \boldsymbol{I}_{i,j}^2 - (\boldsymbol{I}_{i,j}\boldsymbol{g}_k)(\boldsymbol{e}_k^T\boldsymbol{I}_{i,j})$. Now $\boldsymbol{I}_{i,j}^2 = \boldsymbol{I}$ and $\boldsymbol{e}_k^T\boldsymbol{I}_{i,j} = \boldsymbol{e}_k^T$ in view of $i, j > k$. Thus Part 3 follows.

□

It should be noted that the order of the factors in the product in Part 2 of Lemma 2.58 is important. For example for $n = 3$

$$(\boldsymbol{I} + \boldsymbol{g}_1\boldsymbol{e}_1^T)(\boldsymbol{I} + \boldsymbol{g}_2\boldsymbol{e}_2^T) = \begin{bmatrix} 1 & 0 & 0 \\ g_{21} & 1 & 0 \\ g_{31} & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & g_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ g_{21} & 1 & 0 \\ g_{31} & g_{3,2} & 1 \end{bmatrix}.$$

In general,

$$(\boldsymbol{I} + \boldsymbol{g}_1\boldsymbol{e}_1^T)\cdots(\boldsymbol{I} + \boldsymbol{g}_{n-1}\boldsymbol{e}_{n-1}^T) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ g_{2,1} & 1 & 0 & \cdots & 0 \\ g_{3,1} & g_{3,2} & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ g_{n,1} & g_{n,2} & \cdots & g_{n,n-1} & 1 \end{bmatrix}. \tag{2.35}$$

Multiplying the factors in any other order does not give such a nice result.

2.9.5 PLU factorization

We can reformulate the Gaussian elimination process (2.29) in matrix terms. This leads to a **factorization** of the coefficient matrix $\boldsymbol{A} = \boldsymbol{PLU}$, where $\boldsymbol{P}$ is a

permutation matrix, $\boldsymbol{L}$ is a lower triangular, and $\boldsymbol{U}$ is upper triangular. Without interchanges the matrix $\boldsymbol{P}$ is the identity and we obtain an LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$.

The transformation $\boldsymbol{A}_k \to \boldsymbol{A}_{k+1}$ in (2.29) can be written in terms of an interchange matrix and a Gauss transformation as follows:

$$\boldsymbol{A}_{k+1} = \boldsymbol{M}_k^- \boldsymbol{J}_k \boldsymbol{A}_k, \quad \boldsymbol{J}_k := \boldsymbol{I}_{k,r_k},$$

where we first interchanged row k and $r_k \geq k$ in $\boldsymbol{A}_k$ and then introduced zeros under the diagonal in column k using

$$\boldsymbol{M}_k^- := \boldsymbol{I} - \boldsymbol{m}_k \boldsymbol{e}_k^T \quad , \quad \boldsymbol{m}_k = [0, \ldots, 0, m_{k+1,k}, \ldots, m_{n,k}]^T, \quad m_{ik} := a_{ik}^k / a_{kk}^k.$$

Applying this repeatedly we obtain

$$\boldsymbol{A}_k = \boldsymbol{M}_{k-1}^- \boldsymbol{J}_{k-1} \cdots \boldsymbol{M}_2^- \boldsymbol{J}_2 \boldsymbol{M}_1^- \boldsymbol{J}_1 \boldsymbol{A}, \quad k = 2, \ldots, n. \tag{2.36}$$

Since $\boldsymbol{J}_k^{-1} = \boldsymbol{J}_k$ and $(\boldsymbol{M}_k^-)^{-1} = \boldsymbol{M}_k^+$, we find the factorizations

$$\boldsymbol{A} = \boldsymbol{J}_1 \boldsymbol{M}_1^+ \boldsymbol{J}_2 \boldsymbol{M}_2^+ \cdots \boldsymbol{J}_{k-1} \boldsymbol{M}_{k-1}^+ \boldsymbol{A}_k, \quad k = 2, \ldots, n. \tag{2.37}$$

Gaussian elimination with row pivoting is mathematically well defined on a nonsingular matrix.

Theorem 2.59 (Gaussian elimination is well defined)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n,n}$ is nonsingular. Then for $k = 1, 2, \ldots, n-1$ we can in (2.29) find $r_k \geq k$ such that $a_{r_k,k} \neq 0$.

Proof. The result holds for $k = 1$ since $\boldsymbol{A}$ is nonsingular and therefore cannot have a zero column. Thus $\boldsymbol{A}_2$ is well defined. Suppose for some $k \geq 2$ that $a_{i,i}^i \neq 0$ for $i = 1, 2, \ldots, k-1$. We partition $\boldsymbol{A}_k$ given by (2.28) in upper block triangular form

$$\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{0} & \boldsymbol{E}_k \end{bmatrix}.$$

The matrix $\boldsymbol{B}_k$ is upper triangular of order $k-1$ with diagonal elements $a_{1,1}^1 \cdots a_{k-1,k-1}^{k-1}$. Therefore $\boldsymbol{B}_k$ is nonsingular, and $\boldsymbol{A}_k$ is nonsingular, since by (2.36) it is a product of nonsingular matrices. By Lemma 1.21 $\boldsymbol{E}_k$ is nonsingular and cannot have a zero first column. Thus $\boldsymbol{A}_{k+1}$ is well defined and the result follows by induction. □

Theorem 2.60 (PLU theorem)
Gaussian elimination on a nonsingular matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$, with row pivoting as described in (2.29), leads to a factorization $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$, where $\boldsymbol{P}$ is a permutation

matrix, $\boldsymbol{L}$ is lower triangular with ones on the diagonal, and $\boldsymbol{U}$ is upper triangular. More explicitly,

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{J}_1\boldsymbol{J}_2\cdots\boldsymbol{J}_{n-1}, \quad \boldsymbol{J}_k := \boldsymbol{I}_{r_k,k}, \\ \boldsymbol{L} &= \boldsymbol{L}_1\boldsymbol{L}_2\cdots\boldsymbol{L}_{n-1}, \\ \boldsymbol{L}_k &= \boldsymbol{I} + \tilde{\boldsymbol{m}}_k\boldsymbol{e}_k^T, \quad \tilde{\boldsymbol{m}}_k := \boldsymbol{J}_{n-1}\cdots\boldsymbol{J}_{k+1}\boldsymbol{m}_k, \quad k = 1, 2, \ldots, n-1, \\ \boldsymbol{U} &= \boldsymbol{A}_n. \end{aligned} \tag{2.38}$$

Proof. By repeated use of Part 3 of Lemma 2.58 we have

$$\boldsymbol{L}_k = \boldsymbol{J}_{n-1}\cdots\boldsymbol{J}_{k+1}\boldsymbol{M}_k^+\boldsymbol{J}_{k+1}\cdots\boldsymbol{J}_{n-1}, \quad k = 1, \ldots, n-1.$$

Using $\boldsymbol{J}_k^2 = \boldsymbol{I}$ repeatedly gives for $n = 4$

$$\begin{aligned} \boldsymbol{PLU} &= (\boldsymbol{J}_1\boldsymbol{J}_2\boldsymbol{J}_3)(\boldsymbol{L}_1\boldsymbol{L}_2\boldsymbol{L}_3)\boldsymbol{A}_4 \\ &= (\boldsymbol{J}_1\boldsymbol{J}_2\boldsymbol{J}_3)(\boldsymbol{J}_3\boldsymbol{J}_2\boldsymbol{M}_1^+\boldsymbol{J}_2\boldsymbol{J}_3)(\boldsymbol{J}_3\boldsymbol{M}_2^+\boldsymbol{J}_3)(\boldsymbol{M}_3^+)\boldsymbol{A}_4 \\ &= \boldsymbol{J}_1\boldsymbol{M}_1^+\boldsymbol{J}_2\boldsymbol{M}_2^+\boldsymbol{J}_3\boldsymbol{M}_3^+\boldsymbol{A}_4. \end{aligned}$$

But then $\boldsymbol{PLU} = \boldsymbol{A}$ by (2.37). Using the same cancellation effect for general n proves the theorem. ∎

Using Part 2 of Lemma 2.58 we see that the matrix $\boldsymbol{L}$ in Theorem 2.60 has the form (2.35) with $\boldsymbol{g}_k = \tilde{\boldsymbol{m}}_k$, $k = 1, \ldots, n$.

Once we have a PLU factorization of $\boldsymbol{A}$ the system $\boldsymbol{Ax} = \boldsymbol{b}$ is solved easily in three steps. Since $\boldsymbol{PLUx} = \boldsymbol{b}$ we have $\boldsymbol{Pz} = \boldsymbol{b}$, $\boldsymbol{Ly} = \boldsymbol{z}$, and $\boldsymbol{Ux} = \boldsymbol{y}$. With $\boldsymbol{P} = \boldsymbol{I}(:, \boldsymbol{p})$ the solution $\boldsymbol{x}$ can be found from Algorithms 2.1 and 2.2 in two steps.

1. y=rforwardsolve(L,b(p),n);
2. x=rbacksolve(U,y,n);

Exercise 2.61 (Using PLU of $\boldsymbol{A}$ to solve $\boldsymbol{A}^T\boldsymbol{x} = \boldsymbol{b}$)
Suppose we know the PLU factors $\boldsymbol{P}, \boldsymbol{L}, \boldsymbol{U}$ in a PLU factorization $\boldsymbol{A} = \boldsymbol{PLU}$ of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Explain how we can solve the system $\boldsymbol{A}^T\boldsymbol{x} = \boldsymbol{b}$ economically.

Exercise 2.62 (Using PLU to compute the determinant)
Suppose we know the PLU factors $\boldsymbol{P}, \boldsymbol{L}, \boldsymbol{U}$ in a PLU factorization $\boldsymbol{A} = \boldsymbol{PLU}$ of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Explain how we can use this to compute the determinant of $\boldsymbol{A}$.

Exercise 2.63 (Using PLU to compute the inverse)
Suppose the factors $\boldsymbol{P}, \boldsymbol{L}, \boldsymbol{U}$ in a PLU factorization of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are known. Use Exercises 2.21,2.22 to show that it takes approximately $2G_n$ arithmetic operations to compute $\boldsymbol{A}^{-1} = \boldsymbol{U}^{-1}\boldsymbol{L}^{-1}\boldsymbol{P}^T$. Here we have not counted the final multiplication with $\boldsymbol{P}^T$ which amounts to n row interchanges.

2.9.6 The LU factorization

Consider now the lucky situation where no row interchanges is necessary in Gaussian elimination. In this case (2.29) simplifies to

$$\boxed{\begin{array}{l}\text{for } k = 1 : n-1\\ \quad \text{for } i = k+1 : n\\ \quad\quad m_{ik} = a_{ik}^k / a_{kk}^k\\ \quad\quad \text{for } j = k : n\\ \quad\quad\quad a_{ij}^{k+1} = a_{ij}^k - m_{ik} a_{kj}^k\end{array}} \tag{2.39}$$

Gaussian elimination without row interchanges is sometimes refered to as **naive Gaussian elimination**, and we then have $\boldsymbol{P} = \boldsymbol{I}$ in the PLU factorization. The PLU theorem then gives:

Theorem 2.64 (LU factorization)
If Gaussian elimination without row interchanges is well defined then we obtain in (2.39) *the LU factorization*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{2,1} & 1 & 0 & \cdots & 0 \\ m_{3,1} & m_{3,2} & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & m_{n,n-1} & 1 \end{bmatrix} \begin{bmatrix} a_{1,1}^1 & a_{1,2}^1 & a_{1,3}^1 & \cdots & a_{1,n}^1 \\ 0 & a_{2,2}^2 & a_{2,3}^2 & \cdots & a_{2,n}^2 \\ 0 & 0 & a_{3,3}^3 & \cdots & a_{3,n}^3 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,n}^n \end{bmatrix} = \boldsymbol{LU}. \tag{2.40}$$

Since we get division by zero in (2.39) if $a_{kk}^k = 0$ for some $k \leq n-1$ it is important to know when this can happen. We first show:

Theorem 2.65 (Nonzero pivots)
Let $\boldsymbol{A} \in \mathbb{C}^{n,n}$ and let $m_{i,k}$ and $a_{i,j}^k$ be defined by (2.39).

1. *If $a_{r,r}^r \neq 0$, $r = 1, 2, \ldots, k-1$, then*

$$\begin{vmatrix} a_{11} & \cdots & a_{k1} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{vmatrix} = a_{1,1}^1 a_{2,2}^2 \cdots a_{k,k}^k, \quad k = 1, 2, \ldots, n.$$

2. *Let $1 \leq k \leq n$. Then $a_{r,r}^r \neq 0$, $r = 1, 2, \ldots, k$, if and only if the leading principal submatrix $\boldsymbol{A}_{[r]}$ is nonsingular for $r = 1, 2, \ldots, k$.*

Proof.

1. Since $a_{r,r}^r \neq 0$, $r = 1, 2, \ldots, k-1$ and $\boldsymbol{J}_r = \boldsymbol{I}$ in (2.36) we obtain $\boldsymbol{A}_k = \boldsymbol{M}\boldsymbol{A}$, where $\boldsymbol{M} := \boldsymbol{M}_{k-1}^- \cdots \boldsymbol{M}_1^-$. By (2.28) we have

$$\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{F}_k & \boldsymbol{G}_k \\ \boldsymbol{0} & \boldsymbol{H}_k \end{bmatrix}, \quad \boldsymbol{F}_k = \begin{bmatrix} a_{1,1}^1 & \cdots & a_{1,k}^1 \\ & \ddots & \vdots \\ & & a_{k,k}^k \end{bmatrix} \in \mathbb{C}^{k,k}.$$

Since $\boldsymbol{A}_k = \boldsymbol{M}\boldsymbol{A}$ and $\boldsymbol{M}$ is lower triangular we obtain $\boldsymbol{F}_k = \boldsymbol{M}_{[k]}\boldsymbol{A}_{[k]}$ and since $\boldsymbol{M}_{[k]}$ has ones on the diagonal

$$a_{1,1}^1 a_{2,2}^2 \cdots a_{k,k}^k = \det(\boldsymbol{F}_k) = \det(\boldsymbol{M}_{[k]}) \det(\boldsymbol{A}_{[k]}) = \det(\boldsymbol{A}_{[k]}.$$

2. Suppose $a_{r,r}^r \neq 0$, $r = 1, 2, \ldots, k$. By Part 1 $\det(\boldsymbol{A}_{[r]}) = a_{1,1}^1 \cdots a_{r,r}^r$, $r = 1, 2, \ldots, k$, so that $\boldsymbol{A}_{[r]}$ is nonsingular for $r = 1, 2, \ldots, k$. Conversely, suppose $a_{i,i}^i = 0$ for some $i \leq k$. Let i be the smallest integer such that $a_{i,i}^i = 0$. We can then do Gaussian elimination without row interchanges on $\boldsymbol{A}$ to obtain $\boldsymbol{A}_i$. By Part 1 $\det(\boldsymbol{A}_{[i]}) = a_{1,1}^1 \cdots a_{i,i}^i = 0$ so that $(\boldsymbol{A})_{[i]}$ is singular. □

□

The theorem implies:

Corollary 2.66 (When is naive Gaussian elimination possible?)
In (2.39) *we have* $a_{kk}^k \neq 0$ *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* $k = 1, \ldots, n-1$.

We note that

1. The PLU factorization can alternatively be written $\boldsymbol{P}^T\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. Thus, if $\boldsymbol{A}$ is nonsingular then there exists a permutation of the rows of $\boldsymbol{A}$ so that the matrix with the rows permuted has an LU factorization. This means that if we knew the row pivots in advance then we can carry out Gaussian elimination without row pivoting on the matrix $\boldsymbol{P}^T\boldsymbol{A}$, where $\boldsymbol{P}^T = \boldsymbol{I}_{r_{n-1},n-1} \cdots \boldsymbol{I}_{r_1,1}$.

2. If the leading principal submatrices $\boldsymbol{A}_{[k]}$ are nonsingular for $k = 1, \ldots, n-1$ then the LU factorization is unique and Gaussian elimination is just one particular way of computing the LU factorization.

3. The calculation in (2.39) reqires $\sum_{k=1}^{n-1}(n-k)^2$ multiplications, the same number of subtaractions, and $\sum_{k=1}^{n-1} k$ divisions. So the complexity of Gaussian elimintiation is $\frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n$. This is exactly the complexity of LU factorization (cf. (2.9)).

4. Corolary 2.66 holds even if $\boldsymbol{A}$ is singular. Since $\boldsymbol{L}$ is nonsingular the matrix $\boldsymbol{U}$ is then singular, and since $a_{kk}^k \neq 0$ for $k = 1, \ldots, n-1$ we must have $a_{nn}^n = 0$ when $\boldsymbol{A}$ is singular.

5. To verify the nonsingularity of the leading principal submatrices can be difficult in practice. We have show that this condition holds for a class of diagonally dominant matrices and for positive definite matrices.

Exercise 2.67 (Direct proof of Theorem 2.64)
Equation (2.39) *implies that* $m_{ik} = (a_{ij}^k - a_{ij}^{k+1})/a_{kj}^k$ *for* $k \leq \min(i-1, j-1)$. *Use this to give a proof of Theorem 2.64 by directly showing that* $(\boldsymbol{LU})_{ij} = \sum_{k=1}^n l_{ik}u_{kj} = a_{ij}$. *Consider separately the two cases* $i \leq j$ *and* $i > j$.

2.10 Review Questions

2.9.1 When is a triangular matrix nonsingular?

2.9.2 Approximately how many arithmetic operatios are needed for

- the multiplication of two square matrices?
- The LU factorization of a matrix?
- the solution of $\boldsymbol{Ax} = \boldsymbol{b}$, when $\boldsymbol{A}$ is triangular?

2.9.3 What is the content of

- the LU theorem?
- the symmetric LU theorem?

2.9.4 Is $\boldsymbol{A}^T\boldsymbol{A}$ symmetric positive definite?

2.9.5
- What class of matrices has a Cholesky factorization?
- What is the bandwidth of the Cholesky factor of a band matrix?

2.9.6 For a symmetric matrix give 3 conditions that are equivalent to positive definiteness.

2.9.7 What class of matrices has a semi-Cholesky factorization?

2.9.8 What is the general condition for Gaussian elimination without row interchanges to be well defined?

2.9.9 What is a PLU factorization? When does it existt?

2.9.10 What is complete pivoting?

Chapter 3

The Kronecker Product

Leopold Kronecker, 1823-1891 (left), Siméon Denis Poisson, 1781-1840 (right).

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.

3.1 Test Matrices

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

3.1.1 The 2D Poisson problem

Let $\Omega := (0,1)^2 = \{(x,y) : 0 < x, y < 1\}$ be the open unit square with boundary $\partial\Omega$. Consider the problem

$$-\Delta u := -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \text{ on } \Omega, \tag{3.1}$$

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

Here the function f is given and continuous on Ω, and we seek a function $u = u(x,y)$ such that (3.1) 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 $\overline{\Omega}_h \backslash \Omega_h$ are the boundary points. The solution is zero at the boundary points. Using the difference approximation from Chapter 1 for the second derivative we obtain the following approximations for the partial derivatives

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

Inserting this in (3.1) we get the following discrete analog of (3.1)

$$\begin{aligned} -\Delta_h v_{j,k} &= f_{j,k}, \quad (jh,kh) \in \Omega_h, \\ v_{j,k} &= 0, \quad (jh,kh) \in \partial\Omega_h, \end{aligned} \tag{3.2}$$

where $f_{j,k} := f(jh,kh)$ and

$$-\Delta_h v_{j,k} := \frac{-v_{j-1,k} + 2v_{j,k} - v_{j+1,k}}{h^2} + \frac{-v_{j,k-1} + 2v_{j,k} - v_{j,k+1}}{h^2}. \tag{3.3}$$

Multiplying both sides of (3.2) by h^2 we obtain

$$\begin{aligned} 4v_{j,k} - v_{j-1,k} - v_{j+1,k} - v_{j,k-1} - v_{j,k+1} &= h^2 f_{jk}, \quad (jh,kh) \in \Omega_h, \\ v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1} &= 0, \qquad j,k = 0, 1, \ldots, m+1. \end{aligned} \tag{3.4}$$

The equations in (3.4) define a set of linear equations for the unknowns $\boldsymbol{V} := [v_{jk}] \in \mathbb{R}^{m \times m}$.

Observe that (3.4) can be written as a matrix equation in the form

$$\boldsymbol{T}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T} = h^2 \boldsymbol{F} \quad \text{with} \quad h = 1/(m+1), \tag{3.5}$$

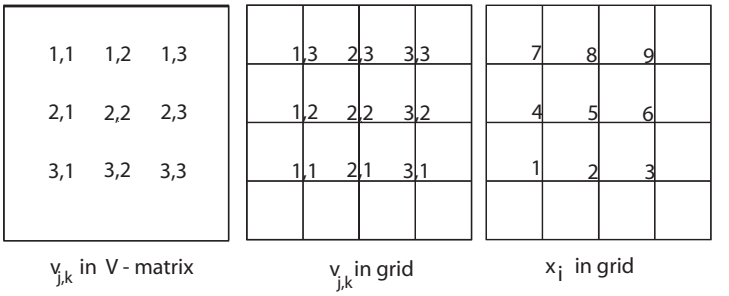


Figure 3.1. *Numbering of grid points*

where $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m\times m}$ is the second derivative matrix given by (1.2) and $\boldsymbol{F} = (f_{jk}) = (f(jh, kh)) \in \mathbb{R}^{m\times m}$. 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 (3.4).

To write (3.4) 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 3.1 (vec operation)
For any $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ 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 3.1. We call this the **natural ordering**. The elements in (3.4) form a 5-point stencil, as shown in Figure 3.2.

To find the matrix $\boldsymbol{A}$ we note that for values of j, k where the 5-point stencil does not touch the boundary, (3.4) 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

$$\boldsymbol{Ax} = \boldsymbol{b}, \quad \boldsymbol{A} \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n, \quad n = m^2, \tag{3.6}$$

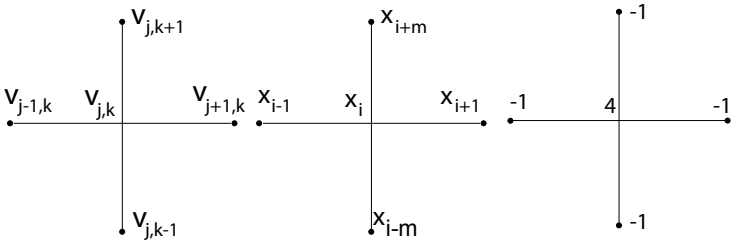


Figure 3.2. *The 5-point stencil*

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

$$\begin{array}{rll} a_{ii} = 4, & i = 1,\ldots,n, & \\ a_{i+1,i} = a_{i,i+1} = -1, & i = 1,\ldots,n-1, & i \neq m, 2m, \ldots, (m-1)m, \\ a_{i+m,i} = a_{i,i+m} = -1, & i = 1,\ldots,n-m, & \\ a_{ij} = 0, & \text{otherwise.} & \end{array} \tag{3.7}$$

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 3.2 (4×4 **Poisson matrix)**
Write down the Poisson matrix for $m = 2$ and show that it is strictly diagonally dominant.

3.1.2 The test matrices

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

$$\boldsymbol{T}_1 := \text{tridiag}(a, d, a) \tag{3.8}$$

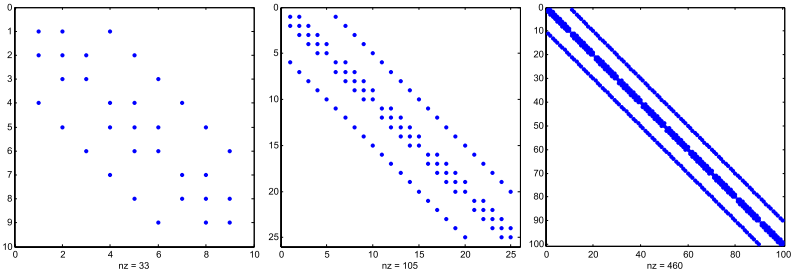


Figure 3.3. *Band structure of the 2D test matrix,* $n = 9$*,* $n = 25$*,* $n = 100$

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

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

$$\begin{array}{rcll} a_{ii} & = & 2d, & i = 1, \ldots, n, \\ a_{i,i+1} = a_{i+1,i} & = & a, & i = 1, \ldots, n-1, \quad i \neq m, 2m, \ldots, (m-1)m, \\ a_{i,i+m} = a_{i+m,i} & = & a, & i = 1, \ldots, n-m, \\ a_{ij} & = & 0, & \text{otherwise}, \end{array} \tag{3.9}$$

and where a, d are real numbers. We will refer to this matrix as simply the **2D test matrix**. For $m = 3$ the 2D test matrix looks as follows

$$\boldsymbol{T}_2 = \left[\begin{array}{ccc|ccc|ccc} 2d & a & 0 & a & 0 & 0 & 0 & 0 & 0 \\ a & 2d & a & 0 & a & 0 & 0 & 0 & 0 \\ 0 & a & 2d & 0 & 0 & a & 0 & 0 & 0 \\ \hline a & 0 & 0 & 2d & a & 0 & a & 0 & 0 \\ 0 & a & 0 & a & 2d & a & 0 & a & 0 \\ 0 & 0 & a & 0 & a & 2d & 0 & 0 & a \\ \hline 0 & 0 & 0 & a & 0 & 0 & 2d & a & 0 \\ 0 & 0 & 0 & 0 & a & 0 & a & 2d & a \\ 0 & 0 & 0 & 0 & 0 & a & 0 & a & 2d \end{array}\right]. \tag{3.10}$$

The partition into 3×3 sub matrices shows that $\boldsymbol{T}_2$ is block tridiagonal.

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

3.2 The Kronecker Product

Definition 3.3 (Kronecker product)
For any positive integers p, q, r, s we define the **Kronecker product** *of two matrices $\boldsymbol{A} \in \mathbb{R}^{p\times q}$ and $\boldsymbol{B} \in \mathbb{R}^{r\times s}$ as a matrix $\boldsymbol{C} \in \mathbb{R}^{pr\times qs}$ given in block form as*

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

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 $\boldsymbol{B} \otimes \boldsymbol{A}$.

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 $p \times r$.

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

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

then

$$\boldsymbol{T}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}_1 = \begin{bmatrix} \boldsymbol{T}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{T}_1 \end{bmatrix} + \begin{bmatrix} d\boldsymbol{I} & a\boldsymbol{I} & \boldsymbol{0} \\ a\boldsymbol{I} & d\boldsymbol{I} & a\boldsymbol{I} \\ \boldsymbol{0} & a\boldsymbol{I} & d\boldsymbol{I} \end{bmatrix} = \boldsymbol{T}_2$$

given by (3.10). The same equation holds for any integer $m \geq 2$

$$\boldsymbol{T}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}_1 = \boldsymbol{T}_2, \quad \boldsymbol{T}_1, \boldsymbol{I} \in \mathbb{R}^{m\times m}, \quad \boldsymbol{T}_2 \in \mathbb{R}^{(m^2)\times(m^2)}. \tag{3.11}$$

The sum of two Kronecker products involving the identity matrix is worthy of a special name.

Definition 3.4 (Kronecker sum)
For positive integers r, s, k, let $\boldsymbol{A} \in \mathbb{R}^{r\times r}$, $\boldsymbol{B} \in \mathbb{R}^{s\times s}$, 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 $\boldsymbol{T}_2$ is the Kronecker sum involving the 1D test matrix $\boldsymbol{T}_1$.

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

$$\begin{aligned} \big(\lambda \boldsymbol{A}\big) \otimes \big(\mu \boldsymbol{B}\big) &= \lambda\mu\big(\boldsymbol{A} \otimes \boldsymbol{B}\big), \\ \big(\boldsymbol{A}_1 + \boldsymbol{A}_2\big) \otimes \boldsymbol{B} &= \boldsymbol{A}_1 \otimes \boldsymbol{B} + \boldsymbol{A}_2 \otimes \boldsymbol{B}, \\ \boldsymbol{A} \otimes \big(\boldsymbol{B}_1 + \boldsymbol{B}_2\big) &= \boldsymbol{A} \otimes \boldsymbol{B}_1 + \boldsymbol{A} \otimes \boldsymbol{B}_2, \\ \big(\boldsymbol{A} \otimes \boldsymbol{B}\big) \otimes \boldsymbol{C} &= \boldsymbol{A} \otimes \big(\boldsymbol{B} \otimes \boldsymbol{C}\big). \end{aligned} \tag{3.12}$$

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 $\boldsymbol{P}, \boldsymbol{Q}$ such that $\boldsymbol{B} \otimes \boldsymbol{A} = \boldsymbol{P}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{Q}$, see [14].

Exercise 3.5 (Properties of Kronecker products)
Prove (3.12).

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

Lemma 3.6 (Mixed product rule)
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

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD}). \tag{3.13}$$

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

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

□

Using the mixed product rule we obtain the following properties of Kronecker products and sums.

Theorem 3.7 (Properties of Kronecker products)
Suppose for $r, s \in \mathbb{N}$ that $\boldsymbol{A} \in \mathbb{R}^{r,r}$ and $\boldsymbol{B} \in \mathbb{R}^{s,s}$ are square matrices with eigenpairs $(\lambda_i, \boldsymbol{u}_i)$ $i = 1, \ldots, r$ and $(\mu_j, \boldsymbol{v}_j)$, $j = 1, \ldots, s$. Moreover, let $\boldsymbol{F}, \boldsymbol{V} \in \mathbb{R}^{r \times s}$. Then

1. $(\boldsymbol{A} \otimes \boldsymbol{B})^T = \boldsymbol{A}^T \otimes \boldsymbol{B}^T$, *(this also holds for rectangular matrices).*
2. *If* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are nonsingular then* $\boldsymbol{A} \otimes \boldsymbol{B}$ *is nonsingular. with* $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$.
3. *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.*
4. $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = \lambda_i \mu_j (\boldsymbol{u}_i \otimes \boldsymbol{v}_j), \quad i = 1, \ldots, r, \quad j = 1, \ldots, s,$
5. $(\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B})(\boldsymbol{u}_i \otimes \boldsymbol{v}_j) = (\lambda_i + \mu_j)(\boldsymbol{u}_i \otimes \boldsymbol{v}_j), \quad i = 1, \ldots, r, \quad j = 1, \ldots, s.,$
6. *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.*
7. $\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}^T = \boldsymbol{F} \quad \Leftrightarrow \quad (\boldsymbol{A} \otimes \boldsymbol{B})\,\mathrm{vec}(\boldsymbol{V}) = \mathrm{vec}(\boldsymbol{F}),$
8. $\boldsymbol{A}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{B}^T = \boldsymbol{F} \quad \Leftrightarrow \quad (\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B})\,\mathrm{vec}(\boldsymbol{V}) = \mathrm{vec}(\boldsymbol{F}).$

Before giving the simple proofs of this theorem we present some comments.

1. The transpose (or the inverse) of an ordinary matrix product equals the transpose (or the inverse) of the matrices in reverse order. For Kronecker products the order is kept.
2. The eigenvalues of the Kronecker product (or sum) are the producet (or sum) of the eigenvalues of the factors. The eigenvectors are the Kronecker products of the eigenvectors of the factors. In particular, the eigenvalues of the test matrix $\boldsymbol{T}_2$ are sums of eigenvalues of $\boldsymbol{T}_1$. We will find these eigenvalues in the next section.
3. Since we already know that $\boldsymbol{T} = \mathrm{tridiag}(-1, 2, -1)$ is positive definite the 2D Poisson matrix $\boldsymbol{A} = \boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}$ is also positive definite.
4. The system $\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}^T = \boldsymbol{F}$ in part 7 can be solved by first finding $\boldsymbol{W}$ from $\boldsymbol{A}\boldsymbol{W} = \boldsymbol{F}$, and then finding $\boldsymbol{V}$ from $\boldsymbol{B}\boldsymbol{V}^T = \boldsymbol{W}^T$. This is preferable to solving the much larger linear system $(\boldsymbol{A} \otimes \boldsymbol{B})\,\mathrm{vec}(\boldsymbol{V}) = \mathrm{vec}(\boldsymbol{F})$.
5. A fast way to solve the 2D Poisson problem in the form $\boldsymbol{T}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T} = \boldsymbol{F}$ will be considered in the next chapter.

Proof.

1. Exercise.
2. By the mixed product rule $\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)\big(\boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}\big) = \big(\boldsymbol{A}\boldsymbol{A}^{-1}\big) \otimes \big(\boldsymbol{B}\boldsymbol{B}^{-1}\big) = \boldsymbol{I}_r \otimes \boldsymbol{I}_s = \boldsymbol{I}_{rs}$.Thus $\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)$ is nonsingular with the indicated inverse.

3. By 1, $(\boldsymbol{A}\otimes\boldsymbol{B})^T = \boldsymbol{A}^T\otimes\boldsymbol{B}^T = \boldsymbol{A}\otimes\boldsymbol{B}$. Moreover, since then $\boldsymbol{A}\otimes\boldsymbol{I}$ and $\boldsymbol{I}\otimes\boldsymbol{B}$ are symmetric, their sum is symmetric.

4. $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{u}_i\otimes\boldsymbol{v}_j) = (\boldsymbol{A}\boldsymbol{u}_i)\otimes(\boldsymbol{B}\boldsymbol{v}_j) = (\lambda_i\boldsymbol{u}_i)\otimes(\mu_j\boldsymbol{v}_j) = (\lambda_i\mu_j)(\boldsymbol{u}_i\otimes\boldsymbol{v}_j)$, for all i,j, where we used the mixed product rule.

5. $(\boldsymbol{A}\otimes\boldsymbol{I}_s)(\boldsymbol{u}_i\otimes\boldsymbol{v}_j) = \lambda_i(\boldsymbol{u}_i\otimes\boldsymbol{v}_j)$, and $(\boldsymbol{I}_r\otimes\boldsymbol{B})(\boldsymbol{u}_i\otimes\boldsymbol{v}_j) = \mu_j(\boldsymbol{u}_i\otimes\boldsymbol{v}_j)$.The result now follows by summing these relations.

6. By 1, $\boldsymbol{A}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{B}$ is symmetric. Moreover, the eigenvalues $\lambda_i+\mu_j$ are positive since for all i,j, both λ_i and μ_j are nonnegative and one of them is positive. It follows that $\boldsymbol{A}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{B}$ is symmetric positive definite.

7. 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

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

8. This follows immediately from (7) as follows

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

□

For more on Kronecker products see [14].

3.3 Properties of the 1D and 2D Test Matrices

Using Theorem 3.7 we can derive properties of the 2D test matrix $\boldsymbol{T}_2$ from those of $\boldsymbol{T}_1$. We need to determine the eigenpairs of $\boldsymbol{T}_1$. We show that the eigenvectors are the columns of the **sine matrix** defined by

$$
\boldsymbol{S} = \left[\sin\frac{jk\pi}{m+1}\right]_{j,k=1}^{m} \in \mathbb{R}^{m\times m}. \tag{3.14}
$$

For $m = 3$,

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

Lemma 3.8 (Eigenpairs of 1D test matrix)
Suppose $\boldsymbol{T}_1 = (t_{kj})_{k,j} = \text{tridiag}(a, d, a) \in \mathbb{R}^{m\times m}$ *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*

$$\boldsymbol{s}_j = [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^T, \tag{3.15}$$
$$\lambda_j = d + 2a\cos(j\pi h). \tag{3.16}$$

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

$$\boldsymbol{s}_j^T\boldsymbol{s}_k = \frac{m+1}{2}\delta_{j,k} = \frac{1}{2h}\delta_{j,k}, \quad j, k = 1, \ldots, m. \tag{3.17}$$

Proof. We find for $1 < k < m$

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

This also holds for $k = 1, m$, and part 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 3.9 below that the eigenvectors $\boldsymbol{s}_j$ are orthogonal. To finish the proof of (3.17) we compute the square of the Euclidian norm of each $\boldsymbol{s}_j$ as follows:

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

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

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

and (3.17) follows. ∎

Lemma 3.9 (Eigenpairs of a Hermitian matrix)
The eigenvalues of a Hermitian matrix are real. Moreover, eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof. The first part was shown in Lemma 2.40. Suppose that $(\lambda, \boldsymbol{x})$ and $(\mu, \boldsymbol{y})$ are two eigenpairs 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 3.10 (Eigenpairs of 2D test matrix)
For fixed $m \geq 2$ let $\boldsymbol{T}_2$ be the matrix given by (3.9) 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*

$$\boldsymbol{x}_{j,k} = \boldsymbol{s}_j \otimes \boldsymbol{s}_k, \tag{3.18}$$

$$\boldsymbol{s}_j = [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^T, \tag{3.19}$$

$$\lambda_{j,k} = 2d + 2a\cos(j\pi h) + 2a\cos(k\pi h). \tag{3.20}$$

2. *The eigenvectors are orthogonal*

$$\boldsymbol{x}_{j,k}^T\boldsymbol{x}_{p,q} = \frac{1}{4h^2}\delta_{j,p}\delta_{k,q}, \quad j,k,p,q = 1,\ldots,m. \tag{3.21}$$

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

Proof. By Theorem 3.7 the eigenvalues of $\boldsymbol{T}_2 = \boldsymbol{T}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}_1$ are sums of eigenvalues of $\boldsymbol{T}_1$ and the eigenvectors are Kronecker producets of the eigenvectors of $\boldsymbol{T}_1$. Part 1 now follows from Lemma 3.8. Using the transpose rule, the mixed product rule and (3.17) 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 part 2 follows. Since $\boldsymbol{T}_2$ is symmetric, part 3 will follow if the eigenvalues are positive. But this is true if $d > 0$ and $d \geq 2|a|$. Thus $\boldsymbol{T}_2$ is positive definite. ∎

Exercise 3.11 (2. derivative matrix is positive definite)
Write down the eigenvalues of $\boldsymbol{T} = \text{tridiag}(-1, 2, -1)$ using Lemma 3.8 and conclude that $\boldsymbol{T}$ is symmetric positive definite.

Exercise 3.12 (1D test matrix is positive definite?)
Use Lemma 3.8 to show that the matrix $\boldsymbol{T}_1 := tridiag(a, d, a) \in \mathbb{R}^{n\times n}$ *is symmetric positive definite if* $d > 0$ *and* $d \geq 2|a|$.

Exercise 3.13 (Eigenvalues for 2D test matrix of order 4)
For $m = 2$ *the matrix* (3.9) *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* $\boldsymbol{x} = [1, 1, 1, 1]^T$. *Verify that apart from a scaling of the eigenvector this agrees with* (3.20) *and* (3.19) *for* $j = k = 1$ *and* $m = 2$.

Exercise 3.14 (Nine point scheme for Poisson problem)
Consider the following 9 point difference approximation to the Poisson problem $-\Delta u = f$, $u = 0$ *on the boundary of the unit square (cf.* (3.1)*)*

$$\begin{array}{lrcll} \text{(a)} & -(\square_h v)_{j,k} & = & (\mu f)_{j,k} & j,k = 1,\ldots,m \\ \text{(b)} & 0 & = & v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1}, & j,k = 0,1,\ldots,m+1, \\ \text{(c)} & -(\square_h v)_{j,k} & = & [20v_{j,k} - 4v_{j-1,k} - 4v_{j,k-1} - 4v_{j+1,k} - 4v_{j,k+1} & \\ & & & \quad - v_{j-1,k-1} - v_{j+1,k-1} - v_{j-1,k+1} - v_{j+1,k+1}]/(6h^2), & \\ \text{(d)} & (\mu f)_{j,k} & = & [8f_{j,k} + f_{j-1,k} + f_{j,k-1} + f_{j+1,k} + f_{j,k+1}]/12. & \end{array} \tag{3.22}$$

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 (3.22) *defines an* $O(h^4)$ *approximation to* (3.1).

Exercise 3.15 (Matrix equation for nine point scheme)
Consider the nine point difference approximation to (3.1) *given by* (3.22) *in Problem 3.14.*

a) *Show that* (3.22) *is equivalent to the matrix equation*

$$\boldsymbol{TV} + \boldsymbol{VT} - \frac{1}{6}\boldsymbol{TVT} = h^2\mu\boldsymbol{F}. \tag{3.23}$$

Here $\mu\boldsymbol{F}$ *has elements* $(\mu f)_{j,k}$ *given by (3.22d).*

b) *Show that the standard form of the matrix equation* (3.23) *is* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, *where* $\boldsymbol{A} = \boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T} - \frac{1}{6}\boldsymbol{T} \otimes \boldsymbol{T}$, $\boldsymbol{x} = vec(\boldsymbol{V})$, *and* $\boldsymbol{b} = h^2 vec(\mu \boldsymbol{F})$.

Exercise 3.16 (Biharmonic equation)
Consider the biharmonic equation

$$\begin{array}{ll} \Delta^2 u(s,t) := \Delta\big(\Delta u(s,t)\big) = f(s,t) & (s,t) \in \Omega, \\ u(s,t) = 0, \quad \Delta u(s,t) = 0 & (s,t) \in \partial\Omega. \end{array} \tag{3.24}$$

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

a) *Let* $v = -\Delta u$. *Show that* (3.24) *can be written as a system*

$$\begin{array}{rcll} -\Delta v(s,t) & = & f(s,t) & (s,t) \in \Omega \\ -\Delta u(s,t) & = & v(s,t) & (s,t) \in \Omega \\ u(s,t) & = & v(s,t) = 0 & (s,t) \in \partial\Omega. \end{array} \tag{3.25}$$

b) *Discretizing, using* (3.3), *with* $\boldsymbol{T} = tridiag(-1, 2, -1) \in \mathbb{R}^{m\times m}$, $h = 1/(m+1)$, *and* $\boldsymbol{F} = \big(f(jh, kh)\big)_{j,k=1}^{m}$ *we get two matrix equations*

$$\boldsymbol{T}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T} = h^2\boldsymbol{F}, \qquad \boldsymbol{T}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{T} = h^2\boldsymbol{V}.$$

Show that

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

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*

$$\boldsymbol{T}^2\boldsymbol{U} + 2\boldsymbol{T}\boldsymbol{U}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{T}^2 = h^4\boldsymbol{F} \quad \textit{and} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{3.26}$$

where $\boldsymbol{A} = \boldsymbol{T}^2 \otimes \boldsymbol{I} + 2\boldsymbol{T} \otimes \boldsymbol{T} + \boldsymbol{I} \otimes \boldsymbol{T}^2 \in \mathbb{R}^{n\times n}$, $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{U})$, *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* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. *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 the basis of an algorithm? Why?*

3.4 Review Questions

3.4.1 Consider the Poisson matrix.

- Write this matrix as a Kronecker sum,
- how are its eigenvalues and eigenvectors related to the second derivative matrix?
- is it symmetric? positive definite?

3.4.2 What are the eigenpairs of $\boldsymbol{T}_1 := \text{tridiagonal}(a, d, a)$?

3.4.3 What are the inverse and transpose of a Kronecker product?

3.4.4

- give an economical general way to solve the linear system $(\boldsymbol{A} \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{V}) = \text{vec}(\boldsymbol{F})$?
- Same for $(\boldsymbol{A} \otimes \boldsymbol{I}_s + \boldsymbol{I}_r \otimes \boldsymbol{B})\,\text{vec}(\boldsymbol{V}) = \text{vec}(\boldsymbol{F})$.

Chapter 4

Fast Direct Solution of a Large Linear System

4.1 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 (3.7). Thus, for $n = 9$

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

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

1. It is symmetric positive definite.
2. It is banded.
3. It is block-tridiagonal.

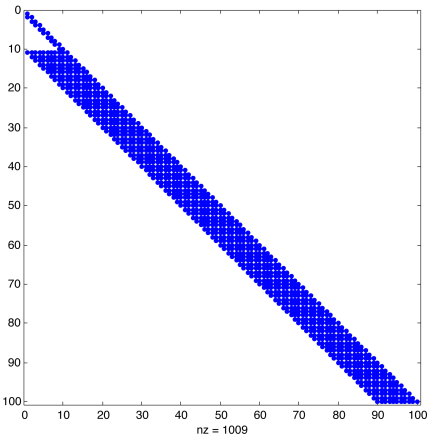


Figure 4.1. *Fill-inn in the Cholesky factor of the Poisson matrix ($n = 100$).*

4. We know the eigenvalues and eigenvectors of $\boldsymbol{A}$.

5. The eigenvectors are orthogonal.

4.1.1 Cholesky factorization

Since $\boldsymbol{A}$ is symmetric positive definite we can use the Cholesky factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$, with $\boldsymbol{L}$ lower triangular, to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Since $\boldsymbol{A}$ is banded with bandwidth $d = \sqrt{n}$ the matrix $\boldsymbol{L}$ has bandwidth $d = \sqrt{n}$ (cf. Lemma 2.46) and the complexity of this factorization is $O(nd^2) = O(n^2)$. We need to store $\boldsymbol{A}$, and this can be done in sparse form.

The nonzero elements in $\boldsymbol{L}$ are shown in Figure 4.1 for $n = 100$. Note that most of the zeros between the diagonals in $\boldsymbol{A}$ have become nonzero in $\boldsymbol{L}$. This is known as **fill-inn**.

4.1.2 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

$$\begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{C}_1 & & & \\ \boldsymbol{A}_1 & \boldsymbol{D}_2 & \boldsymbol{C}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{A}_{m-2} & \boldsymbol{D}_{m-1} & \boldsymbol{C}_{m-1} \\ & & & \boldsymbol{A}_{m-1} & \boldsymbol{D}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & & & \\ \boldsymbol{L}_1 & \boldsymbol{I} & & \\ & \ddots & \ddots & \\ & & \boldsymbol{L}_{m-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{C}_1 & & \\ & \ddots & \ddots & \\ & & \boldsymbol{U}_{m-1} & \boldsymbol{C}_{m-1} \\ & & & \boldsymbol{U}_m \end{bmatrix}. \tag{4.1}$$

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

Using block multiplication the formulas (1.4) generalize to

$$\boldsymbol{U}_1 = \boldsymbol{D}_1, \quad \boldsymbol{L}_k = \boldsymbol{A}_k \boldsymbol{U}_k^{-1}, \quad \boldsymbol{U}_{k+1} = \boldsymbol{D}_{k+1} - \boldsymbol{L}_k \boldsymbol{C}_k, \quad k = 1, 2, \ldots, m-1. \tag{4.2}$$

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 $\boldsymbol{U}\boldsymbol{x} = \boldsymbol{y}$ are as follows:

$$\begin{aligned} &\boldsymbol{y}_1 = \boldsymbol{b}_1, && \boldsymbol{y}_k = \boldsymbol{b}_k - \boldsymbol{L}_{k-1}\boldsymbol{y}_{k-1}, && k = 2, 3, \ldots, m, \\ &\boldsymbol{x}_m = \boldsymbol{U}_m^{-1}\boldsymbol{y}_m, && \boldsymbol{x}_k = \boldsymbol{U}_k^{-1}(\boldsymbol{y}_k - \boldsymbol{C}_k\boldsymbol{x}_{k+1}), && k = m-1, \ldots, 2, 1. \end{aligned} \tag{4.3}$$

The solution is then $\boldsymbol{x}^T = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T]$. To find $\boldsymbol{L}_k$ in (4.2) we solve the linear systems $\boldsymbol{L}_k\boldsymbol{U}_k = \boldsymbol{A}_k$. Similarly we need to solve a linear system to find $\boldsymbol{x}_k$ in (4.3).

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.

4.1.3 Other methods

Other methods include

- Iterative methods, (we study this in Chapters 8 and 9),
- multigrid. See [8],
- fast solvers based on diagonalization and the fast Fourier transform. See Sections 4.2, 4.3.

4.2 A Fast Poisson Solver based on Diagonalization

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

To start we recall that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ can be written as a matrix equation in the form (cf. (3.5))

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

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

$$\boldsymbol{S} := [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m] = \left[\sin(jk\pi h)\right]_{j,k=1}^{m} \in \mathbb{R}^{m\times m}, \quad \boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m). \tag{4.4}$$

Then

$$\boldsymbol{T}\boldsymbol{S} = [\boldsymbol{T}\boldsymbol{s}_1, \ldots, \boldsymbol{T}\boldsymbol{s}_m] = [\lambda_1 \boldsymbol{s}_1, \ldots, \lambda_m \boldsymbol{s}_m] = \boldsymbol{S}\boldsymbol{D}, \quad \boldsymbol{S}^2 = \boldsymbol{S}^T\boldsymbol{S} = \frac{1}{2h}\boldsymbol{I}.$$

Define $\boldsymbol{X} \in \mathbb{R}^{m\times m}$ by $\boldsymbol{V} = \boldsymbol{S}\boldsymbol{X}\boldsymbol{S}$, where $\boldsymbol{V}$ is the solution of $\boldsymbol{T}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T} = h^2\boldsymbol{F}$. Then

$$\begin{aligned}&\boldsymbol{T}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T} = h^2\boldsymbol{F}\\ \overset{\boldsymbol{V}=\boldsymbol{S}\boldsymbol{X}\boldsymbol{S}}{\Longleftrightarrow}\ &\boldsymbol{T}\boldsymbol{S}\boldsymbol{X}\boldsymbol{S} + \boldsymbol{S}\boldsymbol{X}\boldsymbol{S}\boldsymbol{T} = h^2\boldsymbol{F}\\ \overset{\boldsymbol{S}(\ )\boldsymbol{S}}{\Longleftrightarrow}\ &\boldsymbol{S}\boldsymbol{T}\boldsymbol{S}\boldsymbol{X}\boldsymbol{S}^2 + \boldsymbol{S}^2\boldsymbol{X}\boldsymbol{S}\boldsymbol{T}\boldsymbol{S} = h^2\boldsymbol{S}\boldsymbol{F}\boldsymbol{S} = h^2\boldsymbol{G}\\ \overset{\boldsymbol{T}\boldsymbol{S}=\boldsymbol{S}\boldsymbol{D}}{\Longleftrightarrow}\ &\boldsymbol{S}^2\boldsymbol{D}\boldsymbol{X}\boldsymbol{S}^2 + \boldsymbol{S}^2\boldsymbol{X}\boldsymbol{S}^2\boldsymbol{D} = h^2\boldsymbol{G}\\ \overset{\boldsymbol{S}^2=\boldsymbol{I}/(2h)}{\Longleftrightarrow}\ &\boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{D} = 4h^4\boldsymbol{G}.\end{aligned}$$

Since $\boldsymbol{D}$ is diagonal, the equaltion $\boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{D} = 4h^4\boldsymbol{G}$, is easy to solve. For the j, k element we find

$$(\boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{D})_{j,k} = \sum_{\ell=1}^{m} d_{j,\ell}x_{\ell,k} + \sum_{\ell=1}^{m} x_{j,\ell}d_{\ell,k} = \lambda_j x_{j,k} + \lambda_k x_{j,k}$$

so that for all j, k

$$x_{jk} = 4h^4 g_{jk}/(\lambda_j + \lambda_k) = h^4 g_{jk}/(\sigma_j + \sigma_k), \quad \sigma_j := \lambda_j/4 = \sin^2(j\pi h/2).$$

Thus to find $\boldsymbol{V}$ we compute

1. $\boldsymbol{G} = \boldsymbol{S}\boldsymbol{F}\boldsymbol{S}$,
2. $x_{j,k} = h^4 g_{j,k}/(\sigma_j + \sigma_k), \quad j, k = 1, \ldots, m,$
3. $\boldsymbol{V} = \boldsymbol{S}\boldsymbol{X}\boldsymbol{S}$.

We can compute $mX, \boldsymbol{S}$ and the σ's without using loops. Using outer products, element by element division, and raising a matrix element by element to a power we find

$$\boldsymbol{X} = h^4\boldsymbol{G}/\boldsymbol{M}, \text{ where } \boldsymbol{M} := \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_m \end{bmatrix} [\,1, \ldots, 1\,] + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [\,\sigma_1 \ldots \sigma_m\,],$$

$$\boldsymbol{S} = \sin\big(\pi h \begin{bmatrix} 1 \\ 2 \\ \vdots \\ m \end{bmatrix} [\,1\ 2 \ldots m\,]\big), \quad \boldsymbol{\sigma} = \sin\big(\frac{\pi h}{2} \begin{bmatrix} 1 \\ 2 \\ \vdots \\ m \end{bmatrix}\big) \wedge 2.$$

We now get the following algorithm to solve numerically the Poisson problem $-\Delta u = f$ 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 $\boldsymbol{F} = (f(jh, kh)) \in \mathbb{R}^{m\times m}$. We compute $\boldsymbol{V} \in \mathbb{R}^{(m+2)\times(m+2)}$ using diagonalization of $\boldsymbol{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m\times m}$.

Algorithm 4.1 (Fast Poisson solver)

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

The formulas are fully vectorized. Since the 6th line in Algorithm 4.1 only requires $O(m^2)$ arithmetic operations the complexity of this algorithm is for large m determined by the 4 m-by-m matrix multiplications and is given by $O(4 \times 2m^3) = O(8n^{3/2})$. [8] The method is very fast and will be used as a preconditioner for a more complicated problem in Chapter 9. In 2012 it took about 0.2 seconds on a laptop to find the 10^6 unknowns $v_{j,k}$ on a 1000×1000 grid.

4.3 A Fast Poisson Solver based on the discrete sine and Fourier transforms

In Algorithm 4.1 we need to compute the product of the sine matrix $\boldsymbol{S} \in \mathbb{R}^{m\times m}$ given by (4.4) and a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$. 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.

[8] 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 4.8.

We need to discuss certain transforms known as the **discrete sine transform**, the **discrete Fourier transform** and the **fast Fourier transform**. In addition we have the **discrete cosine transform** which will not be discussed here. 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.

4.3.1 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

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

is the **discrete sine transform** (DST) of $\boldsymbol{v}$. In matrix form we can write the DST as the matrix times vector $\boldsymbol{w} = \boldsymbol{S}\boldsymbol{v}$, where $\boldsymbol{S}$ is the sine matrix given by (4.4). We 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 4.1 by carrying out discrete sine transforms on 4 m-by-m matrices in addition to the computation of $\boldsymbol{X}$.

4.3.2 The discrete Fourier transform (DFT)

Jean Baptiste Joseph Fourier, 1768 - 1830.

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 $N \in \mathbb{N}$

$$\omega_N = \exp^{-2\pi i/N} = \cos(2\pi/N) - i\sin(2\pi/N), \tag{4.5}$$

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

$$\boldsymbol{z} = \boldsymbol{F}_N \boldsymbol{y}, \quad z_{j+1} = \sum_{k=0}^{N-1} \omega_N^{jk} y_{k+1}, \quad j = 0, \ldots, N-1$$

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 **Fourier matrix** $\boldsymbol{F}_N \in \mathbb{C}^{N \times N}$ has elements ω_N^{jk}, $j, k = 0, 1, \ldots, N-1$. For a matrix 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 $\omega_4^2 = (-i)^2 = -1$, $\omega_4^3 = (-i)(-1) = i$, $\omega_4^4 = (-1)^2 = 1$, $\omega_4^6 = i^2 = -1$, $\omega_4^9 = i^3 = -i$, and so

$$\boldsymbol{F}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega_4 & \omega_4^2 & \omega_4^3 \\ 1 & \omega_4^2 & \omega_4^4 & \omega_4^6 \\ 1 & \omega_4^3 & \omega_4^6 & \omega_4^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}. \tag{4.6}$$

The following lemma shows how the discrete sine transform of order m can be computed from the discrete Fourier transform of order $2m+2$. We recall that for any complex number w

$$\sin w = \frac{e^{iw} - e^{-iw}}{2i}.$$

Lemma 4.2 (Sine transform as Fourier transform)
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

$$\boldsymbol{z}^T = [0, \boldsymbol{x}^T, 0, -\boldsymbol{x}_B^T] \in \mathbb{R}^{2m+2}, \quad \boldsymbol{x}_B^T := [x_m, \ldots, x_2, x_1].$$

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 $\omega = \omega_{2m+2} = e^{-2\pi i/(2m+2)} = e^{-\pi i/(m+1)}$. We note that

$$\omega^{jk} = e^{-\pi i jk/(m+1)}, \quad \omega^{(2m+2-j)k} = e^{-2\pi i} e^{\pi i jk/(m+1)} = e^{\pi i jk/(m+1)}.$$

Component $k+1$ of $\boldsymbol{F}_{2m+2}\boldsymbol{z}$ is then given by

$$\begin{aligned}(\boldsymbol{F}_{2m+2}\boldsymbol{z})_{k+1} &= \sum_{j=0}^{2m-1} \omega^{jk} z_{j+1} = \sum_{j=1}^{m} x_j \omega^{jk} - \sum_{j=1}^{m} x_j \omega^{(2m+2-j)k} \\ &= \sum_{j=1}^{m} x_j \big(e^{-\pi i jk/(m+1)} - e^{\pi i jk/(m+1)}\big) \\ &= -2i \sum_{j=1}^{m} x_j \sin\left(\frac{jk\pi}{m+1}\right) = -2i(\boldsymbol{S}_m \boldsymbol{x})_k.\end{aligned}$$

Dividing both sides by $-2i$ and noting $-1/(2i) = -i/(2i^2) = i/2$, proves the lemma. □

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

4.3.3 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 arithmetic operations 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 $\boldsymbol{P}_N \in \mathbb{R}^{N\times N}$ 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 (4.6)

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

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

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

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 4.3 (Fast Fourier transform)
If $N = 2m$ is even then

$$\boldsymbol{F}_{2m}\boldsymbol{P}_{2m} = \left[\begin{array}{c|c} \boldsymbol{F}_m & \boldsymbol{D}_m\boldsymbol{F}_m \\ \hline \boldsymbol{F}_m & -\boldsymbol{D}_m\boldsymbol{F}_m \end{array}\right], \tag{4.7}$$

where

$$\boldsymbol{D}_m = \operatorname{diag}(1, \omega_N, \omega_N^2, \ldots, \omega_N^{m-1}). \tag{4.8}$$

Proof. Fix integers p, q with $1 \le p, q \le m$ and set $j := p - 1$ and $k := q - 1$. Since

$$\omega_m^m = 1, \quad \omega_{2m}^{2k} = \omega_m^k, \quad \omega_{2m}^m = -1, \quad (\boldsymbol{F}_m)_{p,q} = \omega_m^{jk}, \quad (\boldsymbol{D}_m\boldsymbol{F}_m)_{p,q} = \omega_{2m}^j\omega_m^{jk},$$

we find by considering elements in the four sub-blocks in turn

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

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

Using Theorem 4.3 we can carry out the DFT as a block multiplication. Let $\boldsymbol{y} \in \mathbb{R}^{2m}$ and set $\boldsymbol{w} = \boldsymbol{P}_{2m}^T\boldsymbol{y} = [\boldsymbol{w}_1^T, \boldsymbol{w}_2^T]^T$, where

$$\boldsymbol{w}_1^T = [y_1, y_3, \ldots, y_{2m-1}], \quad \boldsymbol{w}_2^T = [y_2, y_4, \ldots, y_{2m}].$$

Then

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

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 $\boldsymbol{F}_m\boldsymbol{w}_2$. Thus, 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 4.4 (Recursive FFT)

```
1  function z=fftrec(y)
2  %function z=fftrec(y)
3  y=y(:);
4  n=length(y);
5  if n==1
6      z=y;
7  else
8      q1=fftrec(y(1:2:n-1))
9      q2=exp(-2*pi*i/n).^(0:n/2-1)'.*fftrec(y(2:2:n))
10     z=[q1+q2; q1-q2];
11 end
```

Statement 3 is included so that the input $\boldsymbol{y} \in \mathbb{R}^n$ can be either a row or column vector, while the output $\boldsymbol{z}$ is a column vector.

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 [32] 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 arithmetic operations) 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 [32].

The efficiency improvement using the FFT to compute the DFT is spectacular for large N. The direct multiplication $\boldsymbol{F}_N\boldsymbol{y}$ requires $O(8n^2)$ arithmetic

operations since complex arithmetic is involved. Assuming that the FFT uses $5N \log_2 N$ arithmetic operations 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 arithmetic operations then the direct multiplication would take something like 84000 seconds or 23 hours.

4.3.4 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 4.1. We first compute $\boldsymbol{H} = \boldsymbol{SF}$ using Lemma 4.2 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 $O(\gamma(2m + 2) \log_2(2m + 2))$ arithmetic operations 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 $O(8n^{3/2})$ arithmetic operations used in Algorithm 4.1 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 [32] for other efficient ways to implement the DST.

Exercise 4.5 (Fourier matrix)
Show that the Fourier matrix $\boldsymbol{F}_4$ is symmetric, but not Hermitian.

Exercise 4.6 (Sine transform as Fourier transform)
Verify Lemma 4.2 directly when $m = 1$.

Exercise 4.7 (Explicit solution of the discrete Poisson equation)
Show that the exact solution of the discrete Poisson equation (3.4) 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 4.8 (Improved version of Algorithm 4.1)
Algorithm 4.1 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 $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$. Let $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, where $\lambda_j = 4\sin^2(j\pi h/2)$, $j = 1, \ldots, m$.

(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

$$(\boldsymbol{T} + \lambda_j \boldsymbol{I})\boldsymbol{x}_j = \boldsymbol{c}_j \quad j = 1, \ldots, m, \tag{4.9}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$ and $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m]$. 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 (1.4) and (1.5) in $8m - 7$ arithmetic operations. Give an algorithm to find $\boldsymbol{X}$ which only requires $O(\delta m^2)$ arithmetic operations for some constant δ independent of m.

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

(d) Describe a method based on the fast Fourier transform which requires $O(2\gamma n \log_2 n)$ where γ is the same constant as mentioned at the end of the last section.

Exercise 4.9 (Fast solution of 9 point scheme)
Consider the equation

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

that was derived in Exercise 3.15 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 (3.23). 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 3.15 b) is symmetric positive definite and that (3.22) always has a solution $\boldsymbol{V}$.

Exercise 4.10 (Algorithm for fast solution of 9 point scheme)
Derive an algorithm for solving (3.22) *which for large m requires essentially the same number of operations as in Algorithm 4.1. (We assume that $\mu\boldsymbol{F}$ already has been formed).*

Exercise 4.11 (Fast solution of biharmonic equation)
For the biharmonic problem we derived in Exercise 3.16 the equation

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

Define the matrix $\boldsymbol{X} = (x_{j,k})$ by $\boldsymbol{U} = \boldsymbol{S}\boldsymbol{X}\boldsymbol{S}$ where $\boldsymbol{U}$ is the solution of (3.26). *Show that*

$$\boldsymbol{D}^2\boldsymbol{X} + 2\boldsymbol{D}\boldsymbol{X}\boldsymbol{D} + \boldsymbol{X}\boldsymbol{D}^2 = 4h^6\boldsymbol{G}, \text{ where } \boldsymbol{G} = \boldsymbol{S}\boldsymbol{F}\boldsymbol{S},$$

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 4.12 (Algorithm for fast solution of biharmonic equation)
Use Exercise 4.11 to derive an algorithm

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

which requires only $O(\delta n^{3/2})$ operations to find $\boldsymbol{U}$ in Problem 3.16. Here δ is some constant independent of n.

Exercise 4.13 (Check algorithm for fast solution of biharmonic equation)
In Exercise 4.12 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 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ in* (3.26). *You can use $\boldsymbol{x} = \boldsymbol{A}\backslash\boldsymbol{b}$ for solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. 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. Use the Matlab command* `surf(U)` *for plotting U for, say, $m = 50$. Compare the result with Exercise 4.12 by plotting the difference between both matrices.*

Exercise 4.14 (Fast solution of biharmonic equation using 9 point rule)
Repeat Exercises 3.16, 4.12 and 4.13 using the nine point rule (3.22) *to solve the system* (3.25).

4.4 Review Questions

4.4.1 Consider the Poisson matrix.

- What is the bandwidth of its Cholesky factor?
- approximately how many arithmetic operations does it take to find the Cholesky factor?
- same question for block LU,
- same question for the fast Poisson solver with and without FFT.

4.4.2 What is the discrete sine transform and discrete Fourier transform of a vector?

Part II

Some Matrix Theory

Chapter 5

Matrix Reduction by Similarity Transformations

A basic problem in numerical linear algebra is to compute eigenvalues and eigenvectors of a matrix $\boldsymbol{A}$. Before attempting to find eigenvalues and eigenvectors of $\boldsymbol{A}$ (exceptions are made for certain sparse matrices), it should be reduced by similarity transformations to a simpler form. The contents of this chapter is mainly theoretical, but the results are useful in numerical analysis.

5.1 Some Properties of Eigenpairs

We recall (cf. Section 0.7) that $(\lambda, \boldsymbol{x})$ is an **eigenpair** for $\boldsymbol{A}$ if $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ and $\boldsymbol{x}$ is nonzero. The scalar λ is called an **eigenvalue** and $\boldsymbol{x}$ is said to be an **eigenvector**. The set of eigenvalues is called the **spectrum** of $\boldsymbol{A}$ and is denoted by $\sigma(\boldsymbol{A})$. For example, $\sigma(\boldsymbol{I}) = \{1, \ldots, 1\} = \{1\}$. The eigenvalues are the roots of the **characteristic polynomial** given by $\pi_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda\boldsymbol{I})$ for $\lambda \in \mathbb{C}$.

5.1.1 Transformations of eigenpairs and trace

The following results will be useful.

Theorem 5.1 (Transformations of eigenpairs)
Suppose $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then

1. *If $\boldsymbol{A}$ is nonsingular then $(\lambda^{-1}, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}^{-1}$.*
2. *$(\lambda^k, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{A}^k$ for $k \in \mathbb{N}$.*
3. *If p is a polynomial given by $p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k$ then $(p(\lambda), \boldsymbol{x})$ 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. *λ is an eigenvalue for $\boldsymbol{A}^T$, in fact $\pi_{\boldsymbol{A}^T} = \pi_{\boldsymbol{A}}$.*

5. $\overline{\lambda}$ *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. $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \implies \boldsymbol{A}^{-1}\boldsymbol{x} = \lambda^{-1}\boldsymbol{x}$.

2. We use induction on k. The case $k = 1$ is trivial, and if $\boldsymbol{A}^{k-1}\boldsymbol{x} = \lambda^{k-1}\boldsymbol{x}$ then $\boldsymbol{A}^k\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}^{k-1}\boldsymbol{x} = \lambda^{k-1}\boldsymbol{A}\boldsymbol{x} = \lambda^k\boldsymbol{x}$.

3. $p(\boldsymbol{A})\boldsymbol{x} = \sum_{j=0}^{k} a_j\boldsymbol{A}^j\boldsymbol{x} \overset{2.}{=} \sum_{j=0}^{k} a_j\lambda^j\boldsymbol{x} = p(\lambda)\boldsymbol{x}$.

4. Using Property 3. of determinants 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 $\pi_{\boldsymbol{A}^*}(\overline{\lambda}) \overset{4.}{=} \pi_{\overline{\boldsymbol{A}}}(\overline{\lambda}) = \det(\overline{\boldsymbol{A}} - \overline{\lambda}\boldsymbol{I}) = \overline{\det(\boldsymbol{A} - \lambda\boldsymbol{I})} = \overline{\pi_{\boldsymbol{A}}(\lambda)}$. Thus $\pi_{\boldsymbol{A}}(\lambda) = 0 \Leftrightarrow \pi_{\boldsymbol{A}^*}(\overline{\lambda}) = 0$ and the result follows.

6. By Property 6. of determinants

$$\pi_{\boldsymbol{A}}(\lambda) = \begin{vmatrix} \boldsymbol{B} - \lambda\boldsymbol{I} & \boldsymbol{C} \\ 0 & \boldsymbol{D} - \lambda\boldsymbol{I} \end{vmatrix} = \det(\boldsymbol{B} - \lambda\boldsymbol{I})\det(\boldsymbol{D} - \lambda\boldsymbol{I}) = \pi_{\boldsymbol{B}}(\lambda) \cdot \pi_{\boldsymbol{D}}(\lambda).$$

□

There are two important relations between the elements of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and its eigenvalues $\lambda_1, \ldots, \lambda_n$.

Theorem 5.2 (Sums and products of eigenvalues; trace)
For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$

$$\operatorname{trace}(\boldsymbol{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n, \quad \det(\boldsymbol{A}) = \lambda_1\lambda_2\cdots\lambda_n, \tag{5.1}$$

where the **trace** *of* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *is the sum of its diagonal elements*

$$\operatorname{trace}(\boldsymbol{A}) := a_{11} + a_{22} + \cdots + a_{nn}. \tag{5.2}$$

Proof. We compare two different expansion of $\pi_{\boldsymbol{A}}$. On the one hand from (29) we find

$$\pi_{\boldsymbol{A}}(\lambda) = (-1)^n\lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_0,$$

where $c_{n-1} = (-1)^{n-1}\,\mathrm{trace}(\boldsymbol{A})$ 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 (5.1). □

For a 2×2 matrix the characteristic equation takes the convenient form

$$\lambda^2 - \mathrm{trace}(\boldsymbol{A})\lambda + \det(\boldsymbol{A}) = 0. \tag{5.3}$$

Thus, if $\boldsymbol{A} = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right]$ then $\mathrm{trace}(\boldsymbol{A}) = 4$, $\det(\boldsymbol{A}) = 3$ so that $\pi_{\boldsymbol{A}}(\lambda) = \lambda^2 - 4\lambda + 3$.

Exercise 5.3 (Eigenvalues of an idempotent matrix)
Let $\lambda \in \sigma(\boldsymbol{A})$ where $\boldsymbol{A}^2 = \boldsymbol{A} \in \mathbb{C}^{n\times n}$. Show that $\lambda = 0$ or $\lambda = 1$. (A matrix is called **idempotent** *if $\boldsymbol{A}^2 = \boldsymbol{A}$).*

Exercise 5.4 (Eigenvalues of an nilpotent matrix)
Let $\lambda \in \sigma(\boldsymbol{A})$ where $\boldsymbol{A}^k = 0$ for some $k \in \mathbb{N}$. Show that $\lambda = 0$. (A matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ such that $\boldsymbol{A}^k = 0$ for some $k \in \mathbb{N}$ is called **nilpotent***).*

Exercise 5.5 (Eigenvalues of a unitary matrix)
Let $\lambda \in \sigma(\boldsymbol{A})$, where $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I}$. Show that $|\lambda| = 1$.

Exercise 5.6 (Nonsigular approximation of a singular matrix)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is singular. Then we can find $\epsilon_0 > 0$ such that $\boldsymbol{A} + \epsilon\boldsymbol{I}$ is nonsingular for all $\epsilon \in \mathbb{C}$ with $|\epsilon| < \epsilon_0$. Hint: $\det(\boldsymbol{A}) = \lambda_1\lambda_2\cdots\lambda_n$, *where λ_i are the eigenvalues of $\boldsymbol{A}$.*

Exercise 5.7 (Companion matrix)
For $q_0, \ldots, q_{n-1} \in \mathbb{C}$ let $p(\lambda) = \lambda^n + q_{n-1}\lambda^{n-1} + \cdots + q_0$ be a polynomial of degree n in λ. We derive two matrices that have $(-1)^n p$ as its characteristic polynomial.

a) *Show that $p = (-1)^n\pi_{\boldsymbol{A}}$ where*

$$\boldsymbol{A} = \begin{bmatrix} -q_{n-1} & -q_{n-2} & \cdots & -q_1 & -q_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

$\boldsymbol{A}$ is called the **companion matrix** *of f.*

b) *Show that* $p = (-1)^n \pi_{\boldsymbol{B}}$ *where*

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -q_0 \\ 1 & 0 & \cdots & 0 & -q_1 \\ 0 & 1 & \cdots & 0 & -q_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -q_{n-1} \end{bmatrix}.$$

Thus $\boldsymbol{B}$ *can also be regarded as a companion matrix for* p.

5.1.2 Similarity transformations

Row operations are used in Gaussian elimination 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 5.8 (Similar matrices)
Two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ *are said to be* **similar** *if there is a nonsingular matrix* $\boldsymbol{S} \in \mathbb{C}^{n\times n}$ *such that* $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$. *The transformation* $\boldsymbol{A} \to \boldsymbol{B}$ *is called a* **similarity transformation**. *It is called a* **unitary similarity transformation** *if* $\boldsymbol{S}^*\boldsymbol{S} = \boldsymbol{I}$ *and an* **orthonormal similarity transformation** *if* $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{I}$.

Theorem 5.9 (Eigenpairs of similar matrices)
Let $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$, *where* $\boldsymbol{S} \in \mathbb{C}^{n\times n}$ *is nonsingular with columns* $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. *Then* $\boldsymbol{B}$ *and* $\boldsymbol{A}$ *have the same characteristic polynomial. Moreover,* $(\lambda, \boldsymbol{v})$ *is an eigenpair for* $\boldsymbol{B}$ *if and only if* $(\lambda, \boldsymbol{S}\boldsymbol{v})$ *is an eigenpair for* $\boldsymbol{A}$.

Proof. 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. Moreover, $(\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S})\boldsymbol{v} = \lambda\boldsymbol{v}$ if and only if $\boldsymbol{A}(\boldsymbol{S}\boldsymbol{v}) = \lambda(\boldsymbol{S}\boldsymbol{v})$. ∎

As a corollary we have the following useful result.

Corollary 5.10 (Spectra of $\boldsymbol{AB}$ and $\boldsymbol{BA}$)
For any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *and* $\boldsymbol{B} \in \mathbb{C}^{n\times m}$ *the matrices* $\boldsymbol{A}\boldsymbol{B}$ *and* $\boldsymbol{B}\boldsymbol{A}$ *have the same spectrum apart from some extra zero eigenvalues. More precisely,*

$$\lambda^n \pi_{\boldsymbol{A}\boldsymbol{B}}(\lambda) = \lambda^m \pi_{\boldsymbol{B}\boldsymbol{A}}(\lambda), \quad \lambda \in \mathbb{C}.$$

If $m = n$ *then* $\pi_{\boldsymbol{A}\boldsymbol{B}} = \pi_{\boldsymbol{B}\boldsymbol{A}}$.

Proof. Define block matrices of order $n+m$ by

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{AB} & \boldsymbol{0} \\ \boldsymbol{B} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B} & \boldsymbol{BA} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

The matrix $\boldsymbol{S}$ is nonsingular with $\boldsymbol{S}^{-1} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}$. Moreover, $\boldsymbol{ES} = \boldsymbol{SF}$ so $\boldsymbol{E}$ and $\boldsymbol{F}$ are similar and have the same characteristic polynomial by Theorem 5.9. By Property 6. of Theorem 5.1 we have $\pi_{\boldsymbol{E}}(\lambda) = \lambda^n \pi_{\boldsymbol{AB}}(\lambda) = \pi_{\boldsymbol{F}}(\lambda) = \lambda^m \pi_{\boldsymbol{BA}}(\lambda)$. If $m = n$ we can cancel the λ factors. □

5.2 Unitary Similarity Transformations

In this section we consider the reduction of a matrix to triangular, or almost triangular form using unitary similarity transformations, and characterize matrices with orthonormal eigenvectors.

We start by reviewing some basic facts about matrices with orthonormal columns.

5.2.1 Unitary and orthonormal and matrices

labelsec:orthmat

Definition 5.11 (Unitary matrix)
A matrix $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ is said to be **unitary** *if $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$. A real unitary matrix is called* **orthonormal**.

Warning: An orthonormal matrix is often called an "orthogonal matrix" in the literature.

In the following we consider only the complex case. The real case follows by replacing conjugate transpose "*" by transpose "T" and $\mathbb{C}$ by $\mathbb{R}$. We use the **standard inner product in** $\mathbb{C}^n$ given by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{y}^*\boldsymbol{x}$. In the real case we have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}$. Orthogonality and orthonormality is with respect to the standard inner product.

Since $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ the matrix $\boldsymbol{U}$ is nonsingular, $\boldsymbol{U}^{-1} = \boldsymbol{U}^*$ and $\boldsymbol{U}\boldsymbol{U}^* = \boldsymbol{I}$ as well. Moreover, both the columns and rows of a unitary matrix of order n form orthonormal bases for $\mathbb{C}^n$. We also note that the product of two unitary matrices is unitary. Indeed if $\boldsymbol{U}_1^*\boldsymbol{U}_1 = \boldsymbol{I}$ and $\boldsymbol{U}_2^*\boldsymbol{U}_2 = \boldsymbol{I}$ then $(\boldsymbol{U}_1\boldsymbol{U}_2)^*(\boldsymbol{U}_1\boldsymbol{U}_2) = \boldsymbol{U}_2^*\boldsymbol{U}_1^*\boldsymbol{U}_1\boldsymbol{U}_2 = \boldsymbol{I}$.

Theorem 5.12 (Unitary matrix)
The matrix $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ is unitary if and only if $\langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{U}\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$. In particular, if $\boldsymbol{U}$ is unitary then $\|\boldsymbol{U}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \mathbb{C}^n$.

Proof. If $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ then

$$\langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{U}\boldsymbol{y} \rangle = (\boldsymbol{U}\boldsymbol{y})^*(\boldsymbol{U}\boldsymbol{x}) = \boldsymbol{y}^*\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{x} = \boldsymbol{y}^*\boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle.$$

Conversely, if $\langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{U}\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ then $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ since for $i, j = 1, \ldots, n$

$$(\boldsymbol{U}^*\boldsymbol{U})_{i,j} = \boldsymbol{e}_i^T\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{e}_j = (\boldsymbol{U}\boldsymbol{e}_i)^*(\boldsymbol{U}\boldsymbol{e}_j) = \langle \boldsymbol{U}\boldsymbol{e}_j, \boldsymbol{U}\boldsymbol{e}_i \rangle = \langle \boldsymbol{e}_j, \boldsymbol{e}_i \rangle = \boldsymbol{e}_i^*\boldsymbol{e}_j = \delta_{i,j}.$$

The last part of the theorem follows immediately by taking $\boldsymbol{y} = \boldsymbol{x}$: ∎


Issai Schur, 1875-1941 (left),John William Strutt (Lord Rayleigh), 1842-1919 (right).

5.2.2 The Schur decomposition

Although not every matrix can be diagonalized it can be brought into triangular form by a *unitary* similarity transformation.

Theorem 5.13 (Schur decomposition)
For each $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ there exists a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ such that $\boldsymbol{R} := \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ is upper triangular.

The matrices $\boldsymbol{U}$ and $\boldsymbol{R}$ in the Schur decomposition are called **Schur factors**.

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 $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, where $n := k + 1$. Let $(\lambda_1, \boldsymbol{v}_1)$ be an eigenpair for $\boldsymbol{A}$ with $\|\boldsymbol{v}_1\|_2 = 1$. By Theorem 0.30 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 $\boldsymbol{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{C}^{n\times n}$ is unitary, and for the first column of the product $\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V}$ we find

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

It follows that

$$\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V} = \left[\begin{array}{c|c} \lambda_1 & \boldsymbol{x}^* \\ \hline \boldsymbol{0} & \boldsymbol{M} \end{array}\right], \text{ for some } \boldsymbol{M} \in \mathbb{C}^{k\times k} \text{ and } \boldsymbol{x} \in \mathbb{C}^k. \tag{5.4}$$

By the induction hypothesis there is a unitary matrix $\boldsymbol{W}_1 \in \mathbb{C}^{k\times k}$ 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 $\boldsymbol{U}$ 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. □

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 5.13.

Theorem 5.14 (Schur form, real eigenvalues)
For each $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with real eigenvalues there exists a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$, such that $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$ is upper triangular.

Proof. Consider the proof of Theorem 5.13. Since $\boldsymbol{A}$ and λ_1 are real the eigenvector $\boldsymbol{v}_1$ is real and the matrix $\boldsymbol{W}$ is real and $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$. By the induction hypothesis $\boldsymbol{V}$ is real and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. But then also $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{W}$ is real and $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$. □

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 $\boldsymbol{M}$. This is an example of a **deflation** technique which is very useful in numerical work.

Example 5.15 (Deflation example)
The matrix $\boldsymbol{T} := \left[\begin{smallmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{smallmatrix}\right]$ has an eigenpair $(2, \boldsymbol{x}_1)$, where $\boldsymbol{x}_1 = [-1, 0, 1]^T$. We can extend $\boldsymbol{x}_1$ to a basis $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ for $\mathbb{R}^3$ by defining $\boldsymbol{x}_2 = [0, 1, 0]^T$, $\boldsymbol{x}_3 = [1, 0, 1]^T$. This is already an orthogonal basis and normalizing we obtain the orthonormal matrix

$$\boldsymbol{V} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

We obtain (5.4) *with* $\lambda = 2$ *and*

$$\boldsymbol{M} = \begin{bmatrix} 2 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix}.$$

We can now find the remaining eigenvalues of $\boldsymbol{A}$ *from the* 2×2 *matrix* $\boldsymbol{M}$.

Exercise 5.16 (Schur decomposition example)
Show that a Schur decomposition of $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 2 \end{smallmatrix}\right]$ *is* $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \left[\begin{smallmatrix} -1 & -1 \\ 0 & 4 \end{smallmatrix}\right]$, *where* $\boldsymbol{U} = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right]$.

How far can we reduce a real matrix $\boldsymbol{A}$ with some complex eigenvalues by a real unitary similarity transformation? To study this we note that the complex eigenvalues of a real matrix occur in conjugate pairs, $\lambda = \mu + i\nu$, $\overline{\lambda} = \mu - i\nu$, where μ, ν are real. The real 2×2 matrix

$$\boldsymbol{M} = \begin{bmatrix} \mu & \nu \\ -\nu & \mu \end{bmatrix} \tag{5.5}$$

has eigenvalues $\lambda = \mu + i\nu$ and $\overline{\lambda} = \mu - i\nu$.

Definition 5.17 (Quasi-triangular matrix)
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 5.1 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.

Any $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ can be reduced to quasi-triangular form by a real orthonormal similarity transformation. A proof is given in Section 5.7.

5.2.3 Normal matrices

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

Definition 5.18 (Normal matrix)
A matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *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)

The 2. derivative matrix $\boldsymbol{T}$ in (1.2) and the discrete Poisson matrix (cf. Lemma 3.10) are examples of normal matrices.

Exercise 5.19 (Skew-Hermitian matrix)
Suppose $\boldsymbol{C} = \boldsymbol{A} + i\boldsymbol{B}$, *where* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$. *Show that* $\boldsymbol{C}$ *is skew-Hermitian if and only if* $\boldsymbol{A}^T = -\boldsymbol{A}$ *and* $\boldsymbol{B}^T = \boldsymbol{B}$.

Exercise 5.20 (Eigenvalues of a skew-Hermitian matrix)
Show that any eigenvalue of a skew-Hermitian matrix is purely imaginary.

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

Theorem 5.21 (Orthonormal eigenpairs characterization)
A matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *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

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

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 5.13 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. □

The special cases where $\boldsymbol{A}$ is Hermitian, or real and symmetric, occur often in applicationals and deserve special attention.

Corollary 5.22 (Spectral theorem, complex form)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is Hermitian. Then $\boldsymbol{A}$ has real eigenvalues $\lambda_1, \ldots, \lambda_n$. Moreover, there is a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ such that $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. For any such $\boldsymbol{U}$ the columns $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for $j = 1, \ldots, n$.

Proof. That the eigenvalues are real was shown in Lemma 3.9. The rest follows from Theorem 5.21. □

There is also a real version.

Corollary 5.23 (Spectral theorem (real form))
Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{A}^T = \boldsymbol{A}$. Then $\boldsymbol{A}$ has real eigenvalues $\lambda_1, \ldots, \lambda_n$. Moreover, there is an orthonormal matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. For any such $\boldsymbol{U}$ the columns $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ for $j = 1, \ldots, n$.

Proof. Since a real symmetric matrix has real eigenvalues and eigenvectors this follows from Theorem 5.22. □

Example 5.24 *The orthonormal 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* $\boldsymbol{U} = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right]$.

5.3 Minmax theorems for Hermitian Matrices

There are some useful characterizations of the eigenvalues of a Hermitian matrix. They are based on the Rayleigh quotient that is a useful tool when studying eigenpairs.

5.3.1 The Rayleigh quotient

Definition 5.25 (Rayleigh quotient)
For $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and a nonzero $\boldsymbol{x} \in \mathbb{C}^n$ the number

$$R(\boldsymbol{x}) = R_{\boldsymbol{A}}(\boldsymbol{x}) := \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}$$

is called a **Rayleigh quotient**.

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

Equation (5.6) in the following lemma shows that the Rayleigh quotient of a normal matrix is a **convex combination** of its eigenvalues.

Lemma 5.26 (Convex combination of the eigenvalues)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is normal with orthonormal eigenpairs $(\lambda_j, \boldsymbol{u}_j)$, $j = 1, 2, \ldots, n$ and let $\boldsymbol{x} \in \mathbb{C}^n$. Then

$$R_{\boldsymbol{A}}(\boldsymbol{x}) = \sum_{j=1}^{n} b_j\lambda_j, \quad b_j \geq 0, \quad \sum_{j=1}^{n} b_j = 1, \tag{5.6}$$

where $b_j = |c_j|^2/\sum_{i=1}^n |c_i|^2$, $j = 1, \ldots, n$, and $\boldsymbol{x} = \sum_{j=1}^n c_j\boldsymbol{u}_j$ is the eigenvector expansion of $\boldsymbol{x}$.

Proof. By orthonormality of the eigenvectors $\boldsymbol{x}^*\boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n \overline{c}_i\overline{\boldsymbol{u}}_i c_j\boldsymbol{u}_j = \sum_{j=1}^n |c_j|^2$. Similarly, $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n \overline{c}_i\overline{\boldsymbol{u}}_i c_j\lambda_j\boldsymbol{u}_j = \sum_{j=1}^n \lambda_j|c_j|^2$. and (5.6) follows with $b_j = |c_j|^2/\sum_{i=1}^n |c_i|^2$, $j = 1, \ldots, n$. This is clearly a combination of nonnegative quantities and a convex combination since $\sum_{j=1}^n |c_j|^2/\sum_{i=1}^n |c_i|^2 = 1$. ∎

5.3.2 Minmax and maxmin

First we show

Theorem 5.27 (Minmax)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is Hermitian with eigenvalues $\lambda_1, \ldots, \lambda_n$, ordered so that $\lambda_1 \geq \cdots \geq \lambda_n$. Let $1 \leq k \leq n$. For any subspace $\mathcal{S}$ of $\mathbb{C}^n$ of dimension $n - k + 1$

$$\lambda_k \leq \max_{\substack{\boldsymbol{x}\in\mathcal{S} \\ \boldsymbol{x}\neq\boldsymbol{0}}} R(\boldsymbol{x}), \tag{5.7}$$

with equality for $\mathcal{S} = \tilde{\mathcal{S}} := \operatorname{span}(\boldsymbol{u}_k, \ldots, \boldsymbol{u}_n)$ *and* $\boldsymbol{x} = \boldsymbol{u}_k$. *Here* $(\lambda_j, \boldsymbol{u}_j)$, $1 \le j \le n$ *are orthonormal eigenpairs for* $\boldsymbol{A}$.

Proof. Let $\mathcal{S}$ be any subspace of $\mathbb{C}^n$ of dimension $n-k+1$ and define $\mathcal{S}' := \operatorname{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$. We need to find $\boldsymbol{y} \in \mathcal{S}$ so that $R(\boldsymbol{y}) \ge \lambda_k$. Now $\mathcal{S} + \mathcal{S}' := \{\boldsymbol{s} + \boldsymbol{s}' : \boldsymbol{s} \in \mathcal{S}, \boldsymbol{s}' \in \mathcal{S}'\}$ is a subspace of $\mathbb{C}^n$ and by (7)

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

It follows that $\mathcal{S} \cap \mathcal{S}'$ is nonempty. Let $\boldsymbol{y} \in \mathcal{S} \cap \mathcal{S}' = \sum_{j=1}^{k} c_j \boldsymbol{u}_j$ with $\sum_{j=1}^{k} |c_j|^2 = 1$. Defining $c_j = 0$ for $k+1 \le j \le n$, we obtain by Lemma 5.26

$$\max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) \ge R(\boldsymbol{y}) = \sum_{j=1}^{n} \lambda_j |c_j|^2 = \sum_{j=1}^{k} \lambda_j |c_j|^2 \ge \sum_{j=1}^{k} \lambda_k |c_j|^2 = \lambda_k,$$

and (5.7) follows. If $\boldsymbol{y} \in \tilde{\mathcal{S}}$, say $\boldsymbol{y} = \sum_{j=k}^{n} d_j \boldsymbol{u}_j$ with $\sum_{j=k}^{n} |d_j|^2 = 1$ then again by Lemma 5.26 $R(\boldsymbol{y}) = \sum_{j=k}^{n} \lambda_j |d_j|^2 \le \lambda_k$, and since $\boldsymbol{y} \in \tilde{\mathcal{S}}$ is arbitray we have $\max_{\substack{\boldsymbol{x} \in \tilde{\mathcal{S}} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}) \le \lambda_k$ and equality in (5.7) follows for $\mathcal{S} = \tilde{\mathcal{S}}$. Moreover, $R(\boldsymbol{u}_k) = \lambda_k$. □

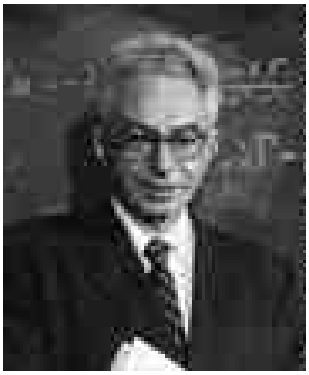
There is also a maxmin version of this result.

Theorem 5.28 (Maxmin)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ *is Hermitian with eigenvalues* $\lambda_1, \ldots, \lambda_n$, *ordered so that* $\lambda_1 \ge \cdots \ge \lambda_n$. *Let* $1 \le k \le n$. *For any subspace* $\mathcal{S}$ *of* $\mathbb{C}^n$ *of dimension* k

$$\lambda_k \ge \min_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} R(\boldsymbol{x}), \tag{5.8}$$

with equality for $\mathcal{S} = \tilde{\mathcal{S}} := \operatorname{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$ *and* $\boldsymbol{x} = \boldsymbol{u}_k$. *Here* $(\lambda_j, \boldsymbol{u}_j)$, $1 \le j \le n$ *are orthonormal eigenpairs for* $\boldsymbol{A}$.

Proof. The proof is very similar to the proof of Theorem 5.27. We define $\mathcal{S}' := \operatorname{span}(\boldsymbol{u}_k, \ldots, \boldsymbol{u}_n)$ and show that $R(\boldsymbol{y}) \le \lambda_k$ for some $\boldsymbol{y} \in \mathcal{S} \cap \mathcal{S}'$. It is easy to see that $R(\boldsymbol{y}) \ge \lambda_k$ for any $\boldsymbol{y} \in \tilde{\mathcal{S}}$. □




Richard Courant, 1888-1972 (left), Ernst Sigismund Fischer, 1875-1954 (right).

These theorems immediately lead to classical minmax and maxmin characterizations.

Corollary 5.29 (The Courant-Fischer theorem)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *is Hermitian with eigenvalues* $\lambda_1, \ldots, \lambda_n$*, ordered so that* $\lambda_1 \geq \cdots \geq \lambda_n$*. Then*

$$\lambda_k = \min_{\dim(\mathcal{S})=n-k+1} \max_{\substack{\boldsymbol{x}\in\mathcal{S}\\ \boldsymbol{x}\neq\boldsymbol{0}}} R(\boldsymbol{x}) = \max_{\dim(\mathcal{S})=k} \min_{\substack{\boldsymbol{x}\in\mathcal{S}\\ \boldsymbol{x}\neq\boldsymbol{0}}} R(\boldsymbol{x}), \quad k = 1, \ldots, n. \tag{5.9}$$

Using the maxmin Theorem 5.27 we can prove inequalities of eigenvalues without knowing the eigenvectors and we can get both upper and lower bounds.

Theorem 5.30 (Eigenvalue perturbation for Hermitian matrices)
Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ *be Hermitian with eigenvalues* $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ *and* $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$*. Then*

$$\alpha_k + \varepsilon_n \leq \beta_k \leq \alpha_k + \varepsilon_1, \text{ for } k = 1, \ldots, n, \tag{5.10}$$

where $\varepsilon_1 \geq \varepsilon_2 \geq \cdots \geq \varepsilon_n$ *are the eigenvalues of* $\boldsymbol{E} := \boldsymbol{B} - \boldsymbol{A}$*.*

Proof. Since $\boldsymbol{E}$ is a sum of Hermitian matrices it is Hermitian and the eigenvalues are real. Let $(\alpha_j, \boldsymbol{u}_j)$, $j = 1, \ldots, n$ be orthonormal eigenpairs for $\boldsymbol{A}$ and let $\mathcal{S} := \operatorname{span}\{\boldsymbol{u}_k, \ldots, \boldsymbol{u}_n\}$. By Theorem 5.27 we obtain

$$\beta_k \leq \max_{\substack{\boldsymbol{x}\in\mathcal{S}\\ \boldsymbol{x}\neq\boldsymbol{0}}} R_{\boldsymbol{B}}(\boldsymbol{x}) \leq \max_{\substack{\boldsymbol{x}\in\mathcal{S}\\ \boldsymbol{x}\neq\boldsymbol{0}}} R_{\boldsymbol{A}}(\boldsymbol{x}) + \max_{\substack{\boldsymbol{x}\in\mathcal{S}\\ \boldsymbol{x}\neq\boldsymbol{0}}} R_{\boldsymbol{E}}(\boldsymbol{x}) \leq \max_{\substack{\boldsymbol{x}\in\mathcal{S}\\ \boldsymbol{x}\neq\boldsymbol{0}}} R_{\boldsymbol{A}}(\boldsymbol{x}) + \max_{\substack{\boldsymbol{x}\in\mathbb{C}^n\\ \boldsymbol{x}\neq\boldsymbol{0}}} R_{\boldsymbol{E}}(\boldsymbol{x}) = \alpha_k + \varepsilon_1,$$

and this proves the upper inequality. For the lower one we define $\boldsymbol{D} := -\boldsymbol{E}$ and observe that $-\varepsilon_n$ is the largest eigenvalue of $\boldsymbol{D}$. Since $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{D}$ it follows from

the result just proved that $\alpha_k \leq \beta_k - \varepsilon_n$, which is the same as the lower inequality. □

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

Exercise 5.31 (Eigenvalue perturbation for Hermitian matrices)
Show that in Theorem 5.30, if $\boldsymbol{E}$ is symmetric positive semidefinite then $\beta_i \geq \alpha_i$.


Alan Jerome Hoffman, 1924- (left), Helmut Wielandt, 1910-2001 (right).

5.3.3 The Hoffman-Wielandt theorem

We can also give a bound involving all eigenvalues. The following theorem shows that the eigenvalue problem for a normal matrix is well conditioned.

Theorem 5.32 (Hoffman-Wielandt theorem)
Suppose $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ 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

$$\sum_{j=1}^{n} |\mu_{i_j} - \lambda_j|^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij} - b_{ij}|^2. \tag{5.11}$$

For a proof of this theorem see [[28], 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 5.33 (Hoffman-Wielandt)
Show that (5.11) *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?*

5.4 The Jordan Form

For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ there is a unitary matrix $\boldsymbol{U}$ such that $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{R}$ is upper triangular. Moreover $\boldsymbol{R}$ is diagonal if $\boldsymbol{A}$ is normal. The following question arises. How close to a diagonal matrix can we reduce a general matrix by a similarity transformation? The main result is the Jordan form in Theorem 5.44. For a proof, see for example [13].

5.4.1 Diagonalizable matrices and linear independence of eigenvectors

We start by giving a characterization of matrices that are similar to a diagonal matrix.

Definition 5.34 (Diagonalizable matrix)
A matrix of order n is **diagonalizable** *if it is similar to a diagonal matrix, and* **defective** *if this is not the case.*

We have $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ if and only if $\boldsymbol{S}^*\boldsymbol{A}^*\boldsymbol{S}^{-*} = \operatorname{diag}(\overline{\lambda}_1, \ldots, \overline{\lambda}_n)$, where $\boldsymbol{S}^{-*} := (\boldsymbol{S}^*)^{-1} = (\boldsymbol{S}^{-1})^*$. Thus $\boldsymbol{A}$ is is diagonalizable if and only if $\boldsymbol{A}^*$ is diagonalizable.

Theorem 5.35 (Eigenvectors of diagonalizable matrices)
A matrix of order n is diagonalizable if and only if its eigenvectors form a basis for $\mathbb{C}^n$.

Proof. Let $\boldsymbol{S} \in \mathbb{C}^{n\times n}$ be nonsingular with columns $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. Then

$$\begin{aligned}\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) &\Leftrightarrow \boldsymbol{A}\boldsymbol{S} = \boldsymbol{S}\operatorname{diag}(\lambda_1, \ldots, \lambda_n)\\ &\Leftrightarrow \boldsymbol{A}\boldsymbol{s}_i = \lambda_i\boldsymbol{s}_i, \quad i = 1, \ldots, n.\end{aligned}$$

Since $\boldsymbol{S}$ is nonsingular the n columns of are linearly independent and therefore constitute a basis for $\mathbb{C}^n$. ∎

If the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of a matrix of order n are linearly independent then any $\boldsymbol{x} \in \mathbb{C}^n$ can be written $\boldsymbol{x} = \sum_{j=1}^n c_j\boldsymbol{v}_j$ for some scalars $c_1, \ldots, c_n$. We call this an **eigenvector expansion** of $\boldsymbol{x}$.

For distinct eigenvalues we have the following result.

Theorem 5.36 (Distinct eigenvalues)
Eigenvectors corresponding to distinct eigenvalues are linearly independent.

Proof. Suppose $(\lambda_k, \boldsymbol{x}_k)$, $k = 1, \ldots, m$ are eigenpairs of $\boldsymbol{A}$ and that $\lambda_1, \ldots, \lambda_m$ are distinct, but $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are linearly dependent. With m the smallest such positive

integer we will obtain a contradiction. For some nonzero vector $\boldsymbol{c} := [c_1, \ldots, c_m]$ we have

$$\sum_{j=1}^{m} c_j \boldsymbol{x}_j = \boldsymbol{0}. \tag{5.12}$$

Clearly $m \geq 2$ since the eigenvectors and $\boldsymbol{c}$ are nonzero. Applying $\boldsymbol{A}$ to (5.12) 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 (5.12) 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 5.37 (Diagonalizable matrix)
A matrix with distinct eigenvalues is diagonalizable.

Proof. By the previous theorem the eigenvectors are linearly independent. □

5.4.2 Algebraic and geometric multiplicity of eigenvalues

A defective matrix must necessarily have one or more multiple eigenvalues, but as the following example shows this is not sufficient.

Example 5.38 (Two upper triangular matrices)
Consider the 2 matrices of order 3

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

Both matrices are upper triangular and have an eigenvalue $\lambda = 1$ of multiplicity 3.

1. *The eigenvectors of $\boldsymbol{A}_1$ are the linearly independent unit vectors $\boldsymbol{x}_i = \boldsymbol{e}_i$, $i = 1, 2, 3$. Thus $\boldsymbol{A}_1$ is diagonalizable.*

2. *An eigenvector $\boldsymbol{x} = [x_1, x_2, x_3]^T$ of $\boldsymbol{A}_2$ must be a solution of the homogenous triangular linear system*

$$(\boldsymbol{A} - \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \text{ or } \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

But then $x_2 = x_3 = 0$ and any eigenvector must be a multiple of $\boldsymbol{e}_1$. We conclude that $\boldsymbol{A}_2$ is defective.

Linear independence of eigenvectors depends on the multiplicity of the eigenvalues in a nontrivial way. For multiple eigenvalues we need to distinguish between two kinds of multiplicities.

Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and

$$\pi_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = (\mu_1 - \lambda)^{a_1} \cdots (\mu_r - \lambda)^{a_r}, \quad \mu_i \neq \mu_j,\ i \neq j, \quad \sum_{i=1}^{r} a_i = n. \tag{5.13}$$

The positive integer $a_i = a(\mu_i) = a_{\boldsymbol{A}}(\mu_i)$ is called the **multiplicity**, or more precisely the **algebraic multiplicity** of the eigenvalue μ_i. The multiplicity of an eigenvalue is simple (double, triple) if a_i is equal to one (two, three).

To define a second kind of multiplicity we consider for each $\lambda \in \sigma(\boldsymbol{A})$ the nullspace

$$\ker(\boldsymbol{A} - \lambda \boldsymbol{I}) := \{\boldsymbol{x} \in \mathbb{C}^n : (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}\} \tag{5.14}$$

of $\boldsymbol{A} - \lambda \boldsymbol{I}$. The nullspace is a subspace of $\mathbb{C}^n$ consisting of all eigenvectors of $\boldsymbol{A}$ corresponding to the eigenvalue λ. The dimension of the subspace must be at least one since $\boldsymbol{A} - \lambda \boldsymbol{I}$ is singular.

Definition 5.39 (Geometric multiplicity)
The **geometric multiplicity** $g = g(\lambda) = g_{\boldsymbol{A}}(\lambda)$ *of an eigenvalue* λ *of* $\boldsymbol{A}$ *is the dimension of the nullspace* $\ker(\boldsymbol{A} - \lambda \boldsymbol{I})$.

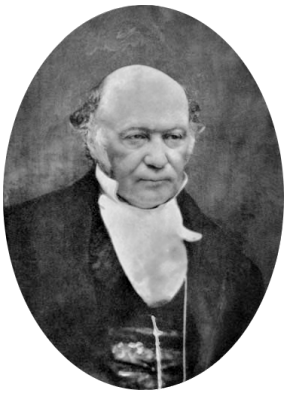
Example 5.40 (Geometric multiplicity)
The $n \times n$ *identity matrix* $\boldsymbol{I}$ *has the eigenvalue* $\lambda = 1$ *with* $\pi_{\boldsymbol{I}}(\lambda) = (1-\lambda)^n$. *Since* $\boldsymbol{I} - \lambda \boldsymbol{I}$ *is the zero matrix when* $\lambda = 1$, *the nullspace of* $\boldsymbol{I} - \lambda \boldsymbol{I}$ *is all of* n*-space and it follows that* $a = g = n$. *On the other hand we saw in Example 5.38 that 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 only one eigenvector. Thus* $g = 1$.

Theorem 5.41 (Geometric multiplicity of similar matrices)
Similar matrices have the same eigenvalues with the same algebraic and geometric multiplicities.

Proof. Similar matrices have the same characteristic polynomials and only the invariance of geometric multiplicity needs to be shown. Suppose $\lambda \in \sigma(\boldsymbol{A})$, $\dim\ker(\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} - \lambda \boldsymbol{I}) = k$, and $\dim\ker(\boldsymbol{A} - \lambda \boldsymbol{I}) = \ell$. We need to show that $k = \ell$. Suppose $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ is a basis for $\ker(\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} - \lambda \boldsymbol{I})$. Then $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}\boldsymbol{v}_i = \lambda \boldsymbol{v}_i$ or $\boldsymbol{A}\boldsymbol{S}\boldsymbol{v}_i = \lambda \boldsymbol{S}\boldsymbol{v}_i$, $i = 1, \ldots, k$. But then $\{\boldsymbol{S}\boldsymbol{v}_1, \ldots, \boldsymbol{S}\boldsymbol{v}_k\} \subset \ker(\boldsymbol{A} - \lambda \boldsymbol{I})$, which implies that $k \leq \ell$. Simiarly, if $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_\ell$ is a basis for $\ker(\boldsymbol{A} - \lambda \boldsymbol{I})$ then $\{\boldsymbol{S}^{-1}\boldsymbol{w}_1, \ldots, \boldsymbol{S}^{-1}\boldsymbol{w}_\ell\} \subset \ker(\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} - \lambda \boldsymbol{I})$. which implies that $k \geq \ell$. We conclude that $k = \ell$. □

Exercise 5.42 (Find eigenpair example)
Find eigenvalues and eigenvectors of $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix}$.

5.4.3 The Jordan form

Marie Ennemond Camille Jordan, 1838-1922 (left), William Rowan Hamilton, 1805-1865 (right).

Definition 5.43 (Jordan block)
A **Jordan block** *of order* m*, 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 $\boldsymbol{J}_3(\lambda) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$. Since a Jordan block is upper triangular λ is an eigenvalue of $\boldsymbol{J}_m(\lambda)$ and any eigenvector must be a multiple of $\boldsymbol{e}_1$. Indeed, if $\boldsymbol{J}_m(\lambda)\boldsymbol{v} = \lambda\boldsymbol{v}$ for some $\boldsymbol{v} = [v_1, \ldots, v_m]$ then $v_2 = \cdots = v_m = 0$. Thus, the eigenvectors of $\boldsymbol{J}_m(\lambda)$ have algebraic multiplicity m and geometric multiplicity one.

The Jordan form is a decomposition of a matrix into Jordan blocks.

Theorem 5.44 (The Jordan form of a matrix)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ *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*

$\boldsymbol{S} \in \mathbb{C}^{n\times n}$ *such that*

$$\boldsymbol{J} := \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} = \operatorname{diag}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k), \text{ with } \boldsymbol{U}_i \in \mathbb{C}^{a_i \times a_i}, \tag{5.15}$$

where each $\boldsymbol{U}_i$ *is block diagonal having* g_i *Jordan blocks along the diagonal*

$$\boldsymbol{U}_i = \operatorname{diag}(\boldsymbol{J}_{m_{i,1}}(\lambda_i), \ldots, \boldsymbol{J}_{m_{i,g_i}}(\lambda_i)). \tag{5.16}$$

Here $m_{i,1}, \ldots, m_{i,g_i}$ *are unique integers ordered 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 matrices $\boldsymbol{S}$ and $\boldsymbol{J}$ in (5.15) are called **Jordan factors**. We also call $\boldsymbol{J}$ the **Jordan form** of $\boldsymbol{A}$. As an example consider the Jordan form

$$\boldsymbol{J} := \operatorname{diag}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \begin{bmatrix} 2 & 1 & 0 & & & & & \\ 0 & 2 & 1 & & & & & \\ 0 & 0 & 2 & & & & & \\ & & & 2 & 1 & & & \\ & & & 0 & 2 & & & \\ & & & & & 2 & & \\ & & & & & & 3 & 1 \\ & & & & & & 0 & 3 \end{bmatrix} \in \mathbb{R}^{8,8}. \tag{5.17}$$

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 $\boldsymbol{A}\boldsymbol{S} = \boldsymbol{S}\boldsymbol{J}$. As an example, in (5.17) we have $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_8]$ and we find

$$\begin{aligned} &\boldsymbol{A}\boldsymbol{s}_1 = 2\boldsymbol{s}_1, \quad \boldsymbol{A}\boldsymbol{s}_2 = 2\boldsymbol{s}_2 + \boldsymbol{s}_1, \quad \boldsymbol{A}\boldsymbol{s}_3 = 2\boldsymbol{s}_3 + \boldsymbol{s}_2, \\ &\boldsymbol{A}\boldsymbol{s}_4 = 2\boldsymbol{s}_4, \quad \boldsymbol{A}\boldsymbol{s}_5 = 2\boldsymbol{s}_5 + \boldsymbol{s}_4, \\ &\boldsymbol{A}\boldsymbol{s}_6 = 2\boldsymbol{s}_6, \\ &\boldsymbol{A}\boldsymbol{s}_7 = 3\boldsymbol{s}_7, \quad \boldsymbol{A}\boldsymbol{s}_8 = 3\boldsymbol{s}_8 + \boldsymbol{s}_7, \end{aligned}$$

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 5.45 (Jordan example)
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 5.46 (Big Jordan example)
Find the Jordan form of the matrix

$$A = \frac{1}{9} \begin{bmatrix} 10 & 16 & -8 & -5 & 6 & 1 & -3 & 4 \\ -7 & 32 & -7 & -10 & 12 & 2 & -6 & 8 \\ -6 & 12 & 12 & -15 & 18 & 3 & -9 & 12 \\ -5 & 10 & -5 & -2 & 24 & 4 & -12 & 16 \\ -4 & 8 & -4 & -16 & 30 & 14 & -15 & 20 \\ -3 & 6 & -3 & -12 & 9 & 24 & -9 & 24 \\ -2 & 4 & -2 & -8 & 6 & -2 & 15 & 28 \\ -1 & 2 & -1 & -4 & 3 & -1 & -6 & 41 \end{bmatrix}. \tag{5.18}$$

The Jordan form implies

Corollary 5.47 (Geometric multiplicity)
We have

1. *The geometric multiplicity of an eigenvalue is always bounded above by the algebraic multiplicity of the eigenvalue.*
2. *The number of linearly independent eigenvectors of a matrix equals the sum of the geometric multiplicities of the eigenvalues.*
3. *A matrix* $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ *has* n *linearly independent eigenvectors if and only if the algebraic and geometric multiplicity of all eigenvalues are the same.*

Proof.

1. The algebraic multiplicity a_i of an eigenvalue λ_i is equal to the size of the corresponding $\boldsymbol{U}_i$. Moreover each $\boldsymbol{U}_i$ contains g_i Jordan blocks of size $m_{i,j} \geq 1$. Thus $g_i \leq a_i$.
2. Since $\boldsymbol{A}$ and $\boldsymbol{J}$ are similar the geometric multiplicities of the eigenvalues of these matrices are the same, and it is enough to prove statement 2 for the Jordan factor $\boldsymbol{J}$. We show this only for the matrix $\boldsymbol{J}$ given by (5.17). The general case should then be clear. There are only 4 eigenvectors of $\boldsymbol{J}$, namly $\boldsymbol{e}_1, \boldsymbol{e}_4, \boldsymbol{e}_6, \boldsymbol{e}_7$ corresponding to the 4 Jordan blocks. These 4 vectors are clearly linearly independent. Moreover there are $k = 2$ distinct eigenvalues and $g_1 + g_2 = 3 + 1 = 4$.
3. Since $g_i \leq a_i$ for all i and $\sum_i a_i = n$ we have $\sum_i g_i = n$ if and only if $a_i = g_i$ for $i = 1, \ldots, k$.

□

The following lemma and the following exercise is useful when studying powers of matrices.

Lemma 5.48 (A nilpotent matrix)
We have

$$\big(\boldsymbol{J}_m(\lambda) - \lambda \boldsymbol{I}\big)^m = \boldsymbol{0}, \quad \lambda \in \mathbb{C}; \quad m \in \mathbb{N}.$$

Proof. Let $\boldsymbol{E}_m := \boldsymbol{J}_m(\lambda) - \lambda \boldsymbol{I}$. For $m = 4$ we find

$$\boldsymbol{E}_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \ \boldsymbol{E}_4^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \ \boldsymbol{E}_4^3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{E}_4^4 = \boldsymbol{0}.$$

In general, $\boldsymbol{E}_m^r = \left[\begin{smallmatrix} \boldsymbol{0} & \boldsymbol{I}_{m-r} \\ \boldsymbol{0} & \boldsymbol{0} \end{smallmatrix}\right]$ for $1 \leq r \leq m-1$, and it follows that $\boldsymbol{E}_m^m = \boldsymbol{0}$. □

Exercise 5.49 (Properties of the Jordan form)
Let $\boldsymbol{J}$ be the Jordan form of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ as given in Theorem 5.44. 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. $\boldsymbol{J}_m(\lambda)^r = (\boldsymbol{E}_m + \lambda \boldsymbol{I}_m)^r = \sum_{k=0}^{\min\{r,m-1\}} \binom{r}{k} \lambda^{r-k} \boldsymbol{E}_m^k$.

Exercise 5.50 (Powers of a Jordan block)
Find $\boldsymbol{J}^{100}$ and $\boldsymbol{A}^{100}$ for the matrix in Exercise 5.45.

5.5 The Minimal Polynomial

Let $\boldsymbol{J}$ be the Jordan form of $\boldsymbol{A}$ given in Theorem 5.44. 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

$$\pi_{\boldsymbol{A}}(\lambda) = \pi_{\boldsymbol{J}}(\lambda) = \prod_{i=1}^{k} \prod_{j=1}^{g_i} (\lambda_i - \lambda)^{m_{i,j}}.$$

The polynomials $p_{ij}(\lambda) := (\lambda_i - \lambda)^{m_{i,j}}$ are called the **elementary divisors** of $\boldsymbol{A}$. They divide the characteristic polynomial.

Definition 5.51 (Minimal polynomial of a matrix)
Suppose $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$ *is the Jordan canonical form of* $\boldsymbol{A}$*. The polynomial*

$$\mu(\lambda) := \prod_{i=1}^{k} (\lambda_i - \lambda)^{m_i} \text{ where } m_i := \max_{1 \le j \le g_i} m_{i,j},$$

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 5.52 (Minimal polynomial example)
What is the characteristic polynomial and the minimal polynomial of the matrix $\boldsymbol{J}$ *in* (5.17)*?*

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. We substitute a matrix for the independent variable in these polynomials and define

$$\pi_{\boldsymbol{A}}(\boldsymbol{A}) := \prod_{i=1}^{k} \prod_{j=1}^{g_i} (\lambda_i \boldsymbol{I} - \boldsymbol{A})^{m_{i,j}}, \quad \mu(\boldsymbol{A}) := \prod_{i=1}^{k} (\lambda_i \boldsymbol{I} - \boldsymbol{A})^{m_i}. \tag{5.19}$$

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 $\mu(\boldsymbol{J}) = \boldsymbol{0}$. Since $\boldsymbol{J}$ and $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k$ are block diagonal we find

$$\begin{aligned} \mu(\boldsymbol{J}) &= \prod_{i=1}^{k} (\lambda_i \boldsymbol{I} - \boldsymbol{J})^{m_i} = \prod_{i=1}^{k} \operatorname{diag}\left((\lambda_i \boldsymbol{I} - \boldsymbol{U}_1)^{m_i}, \ldots, (\lambda_i \boldsymbol{I} - \boldsymbol{U}_k)^{m_i} \right) \\ &= \operatorname{diag}\Big(\prod_{i=1}^{k} (\lambda_i \boldsymbol{I} - \boldsymbol{U}_1)^{m_i}, \ldots, \prod_{i=1}^{k} (\lambda_i \boldsymbol{I} - \boldsymbol{U}_k)^{m_i} \Big) = \boldsymbol{0}, \end{aligned}$$

since by Lemma 5.48 and the maximality of m_r

$$(\lambda_r \boldsymbol{I} - \boldsymbol{U}_r)^{m_r} = \operatorname{diag}\left((\lambda_r \boldsymbol{I} - \boldsymbol{J}_{m_{r,1}})^{m_r}, \ldots, (\lambda_r \boldsymbol{I} - \boldsymbol{J}_{m_{r,g_r}})^{m_r} \right) = \boldsymbol{0}, \ r = 1, \ldots, k.$$

We have shown that a matrix satisfies its minimal polynomial equation $\mu(\boldsymbol{A}) = \boldsymbol{0}$. Moreover, the degree of any polynomial p such that $p(\boldsymbol{A}) = \boldsymbol{0}$ 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 $\pi_{\boldsymbol{A}}(\boldsymbol{A}) = \boldsymbol{0}$.

Exercise 5.53 (Similar matrix polynomials)
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 5.54 (Minimal polynomial of a diagonalizable matrix)
What is the minimal polynomial of the unit matrix and more generally of a diagonalizable matrix?

5.6 Left Eigenvectors

Definition 5.55 (Left eigenpair)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *is a square matrix,* $\lambda \in \mathbb{C}$ *and* $\boldsymbol{y} \in \mathbb{C}^n$. *We say that* $(\lambda, \boldsymbol{y})$ *is a* **left eigenpair** *for* $\boldsymbol{A}$ *if* $\boldsymbol{y}^*\boldsymbol{A} = \lambda\boldsymbol{y}^*$ *and* $\boldsymbol{y}$ *is nonzero.*

Since $\boldsymbol{A}^*\boldsymbol{y} = \overline{\lambda}\boldsymbol{y}$ Theorem 5.1 implies that λ is an eigenvalue of $\boldsymbol{A}$, while a **left eigenvector** is an eigenvector of $\boldsymbol{A}^*$. Thus left and right eigenvalues are identical, but left and right eigenvectors are in general different. For an Hermitian matrix the right and left eigenpairs are the same

Left- and right eigenvectors corresponding to distinct eigenvalues are orthogonal.

Theorem 5.56 (Biorthogonality)
Suppose $(\mu, \boldsymbol{y})$ *and* $(\lambda, \boldsymbol{x})$ *are left and right eigenpairs of* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. *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 $\boldsymbol{y}^*\boldsymbol{x} = 0$. □

Right and left eigenvectors corresponding to the same eigenvalue are sometimes orthogonal, sometimes not.

Theorem 5.57 (Simple eigenvalue)
Suppose $(\lambda, \boldsymbol{x})$ *and* $(\lambda, \boldsymbol{y})$ *are right and left eigenpairs of* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. *If* λ *has algebraic multiplicity one then* $\boldsymbol{y}^*\boldsymbol{x} \neq 0$.

Proof. Assume that $\|\boldsymbol{x}\|_2 = 1$. We have (cf. (5.4))

$$\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V} = \left[\begin{array}{c|c} \lambda & \boldsymbol{z}^* \\ \hline \boldsymbol{0} & \boldsymbol{M} \end{array}\right],$$

where $\boldsymbol{V}$ is unitary and $\boldsymbol{V}\boldsymbol{e}_1 = \boldsymbol{x}$. We show that if $\boldsymbol{y}^*\boldsymbol{x} = 0$ then λ is also an eigenvalue of $\boldsymbol{M}$ contradicting the multiplicity assumption of λ. 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 5.1 it follows that λ is an eigenvalue of $\boldsymbol{M}$. ∎

The case with multiple eigenvalues 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal.

Theorem 5.58 (Biorthogonal eigenvector expansion)
If $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ has linearly independent right eigenvectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ then there exists a set of left eigenvectors $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ with $\boldsymbol{y}_i^*\boldsymbol{x}_j = \delta_{i,j}$. Conversely, if $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ has linearly independent left eigenvectors $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ then there exists a set of right eigenvectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with $\boldsymbol{y}_i^*\boldsymbol{x}_j = \delta_{i,j}$. For any scaling of these sets we have the eigenvector expansions

$$\boldsymbol{v} = \sum_{j=1}^{n} \frac{\boldsymbol{y}_j^*\boldsymbol{v}}{\boldsymbol{y}_j^*\boldsymbol{x}_j}\boldsymbol{x}_j = \sum_{k=1}^{n} \frac{\boldsymbol{x}_k^*\boldsymbol{v}}{\boldsymbol{y}_k^*\boldsymbol{x}_k}\boldsymbol{y}_k. \tag{5.20}$$

Proof. For any right eigenpairs $(\lambda_1, \boldsymbol{x}_1), \ldots, (\lambda_n, \boldsymbol{x}_n)$ and left eigenpairs $(\lambda_1, \boldsymbol{y}_1), \ldots, (\lambda_n, \boldsymbol{y}_n)$ of $\boldsymbol{A}$ we have $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{D}$, $\boldsymbol{Y}^*\boldsymbol{A} = \boldsymbol{D}\boldsymbol{Y}^*$, where

$$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n], \quad \boldsymbol{Y} := [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n], \quad \boldsymbol{D} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

If $\boldsymbol{X}$ is nonsingular then $\boldsymbol{X}^{-1}\boldsymbol{A} = \boldsymbol{D}\boldsymbol{X}^{-1}$ and it follows that $\boldsymbol{Y}^* := \boldsymbol{X}^{-1}$ contains a collection of left eigenvectors such that $\boldsymbol{Y}^*\boldsymbol{X} = \boldsymbol{I}$. Thus the columns of $\boldsymbol{Y}$ are linearly independent and $\boldsymbol{y}_i^*\boldsymbol{x}_j = \delta_{i,j}$. Similarly, if $\boldsymbol{Y}$ is nonsingular then $\boldsymbol{A}\boldsymbol{Y}^{-*} = \boldsymbol{Y}^{-*}\boldsymbol{D}$ and it follows that $\boldsymbol{X} := \boldsymbol{Y}^{-*}$ contains a collection of linearly independent right eigenvectors such that $\boldsymbol{Y}^*\boldsymbol{X} = \boldsymbol{I}$. If $\boldsymbol{v} = \sum_{j=1}^{n} c_j\boldsymbol{x}_j$ then $\boldsymbol{y}_i^*\boldsymbol{v} = \sum_{j=1}^{n} c_j\boldsymbol{y}_i^*\boldsymbol{x}_j = c_i\boldsymbol{y}_i^*\boldsymbol{x}_i$, so $c_i = \boldsymbol{y}_i^*\boldsymbol{v}/\boldsymbol{y}_i^*\boldsymbol{x}_i$ for $i = 1, \ldots, n$ and the first expansion in (5.20) follows. The second expansion follows similarly. ∎

For an Hermitian matrix the right eigenvectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are also left eigenvectors and (5.20) takes the form

$$\boldsymbol{v} = \sum_{j=1}^{n} \frac{\boldsymbol{x}_j^*\boldsymbol{v}}{\boldsymbol{x}_j^*\boldsymbol{x}_j}\boldsymbol{x}_j. \tag{5.21}$$

Exercise 5.59 (Biorthogonal expansion)
Determine right and left eigenpairs for the matrix $\boldsymbol{A} := \left[\begin{smallmatrix} 3 & 1 \\ 2 & 2 \end{smallmatrix}\right]$ and the two expansions in (5.20) *for any $\boldsymbol{v} \in \mathbb{R}^2$.*

Exercise 5.60 (Generalized Rayleigh quotient)
For $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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 called a **generalized 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 $\boldsymbol{y}^*\boldsymbol{x} \neq 0$.*

5.7 Proof of the Real Schur Form

In this section we prove the following theorem.

Theorem 5.61 (The real Schur form)
Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Then we can find $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ 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 5.14 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

$$\boldsymbol{A}\boldsymbol{x} = \mu\boldsymbol{x} - \nu\boldsymbol{y}, \quad \boldsymbol{A}\boldsymbol{y} = \nu\boldsymbol{x} + \mu\boldsymbol{y}. \tag{5.22}$$

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 (5.22) 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 (5.22), 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.

(5.22) can be written in matrix form as

$$\boldsymbol{A}\boldsymbol{X}_1 = \boldsymbol{X}_1\boldsymbol{M}, \quad \boldsymbol{X}_1 = [\boldsymbol{x}, \boldsymbol{y}] \in \mathbb{R}^{n,2}, \quad \boldsymbol{M} = \begin{bmatrix} \mu & \nu \\ -\nu & \mu \end{bmatrix}. \tag{5.23}$$

By Theorem 10.12 we can find an orthonormal matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ 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

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

Since $\boldsymbol{R}$ is nonsingular, $\boldsymbol{QX}$ and $\boldsymbol{X}$ are nonsingular. Moreover, using (5.23)

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

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

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

or

$$\boldsymbol{QAQ}^T = \begin{bmatrix} \boldsymbol{RMR}^{-1} & \boldsymbol{RB} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}. \tag{5.24}$$

By Theorem 5.9 the 2×2 matrix $\boldsymbol{RMR}^{-1}$ 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 $\boldsymbol{U} = \boldsymbol{QV}$. Then $\boldsymbol{U} \in \mathbb{R}^{n\times n}$, $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{U}^T\boldsymbol{AU}$ is quasi-triangular. ∎

5.8 Conclusions

Consider the eigenpair problem for some classes of matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$.

Diagonal Matrices. The eigenpairs are easily determined. Since $\boldsymbol{Ae}_i = a_{ii}\boldsymbol{e}_i$ the eigenpairs are $(\lambda_i, \boldsymbol{e}_i)$, where $\lambda_i = a_{ii}$ for $i = 1, \ldots, n$. Moreover, $a(\lambda_i) = g(\lambda_i)$ for all i, since the eigenvectors of $\boldsymbol{A}$ are linearly independent.

Triangular Matrices Suppose $\boldsymbol{A}$ is upper or lower triangular. Since $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \prod_{i=1}^{n}(a_{ii} - \lambda)$ the eigenvalues are $\lambda_i = a_{ii}$ for $i = 1, \ldots, n$, the diagonal elements of $\boldsymbol{A}$. To determine the eigenvectors can be challenging as Example 5.40 indicates.

Block Diagonal Matrices Suppose

$$\boldsymbol{A} = \operatorname{diag}(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_r), \quad \boldsymbol{A}_i \in \mathbb{C}^{m_i\times m_i}.$$

Here the eigenpair problem reduces to r smaller problems. Let $\boldsymbol{A}_i\boldsymbol{X}_i = \boldsymbol{X}_i\boldsymbol{D}_i$ define the eigenpairs of $\boldsymbol{A}_i$ for $i = 1, \ldots, r$ and let $\boldsymbol{X} := \operatorname{diag}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r)$, $\boldsymbol{D} := \operatorname{diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_r)$. Then the eigenpairs for $\boldsymbol{A}$ are given by

$$\begin{aligned}\boldsymbol{A}\boldsymbol{D} &= \operatorname{diag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r)\operatorname{diag}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r) = \operatorname{diag}(\boldsymbol{A}_1\boldsymbol{X}_1, \ldots, \boldsymbol{A}_r\boldsymbol{X}_r) \\ &= \operatorname{diag}(\boldsymbol{X}_1\boldsymbol{D}_1, \ldots, \boldsymbol{X}_r\boldsymbol{D}_r) = \boldsymbol{X}\boldsymbol{D}.\end{aligned}$$

Block Triangular matrices Matrices Let $\boldsymbol{A}_{11}, \boldsymbol{A}_{22}, \ldots, \boldsymbol{A}_{rr}$ be the diagonal blocks of $\boldsymbol{A}$. By Property 8. of determinants

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \prod_{i=1}^{r} \det(\boldsymbol{A}_{ii} - \lambda\boldsymbol{I})$$

and the eigenvalues are found from the eigenvalues of the diagonal blocks.

5.9 Review Questions

5.9.1 Does $\boldsymbol{A}$ and $\boldsymbol{A}^T$, $\boldsymbol{A}$ and $\boldsymbol{A}^*$ have the same eigenvalues? What about $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$?

5.9.2 Can the geometric multiplicity of an eigenvalue be bigger than the algebraic multiplicity?

5.9.3 What are the eigenvalues of a diagonal matrix?

5.9.4 What are the Schur factors of a matrix?

5.9.5 What is a quasi-triangular matrix?

5.9.6 Give some classes of normal matrices. Why are normal matrices important?.

5.9.7 State the Courant-Fischer theorem.

5.9.8 State the Hoffman-Wieland theorem for Hermtian matrices.

5.9.9 What is a left eigenvector of a matrix.

Chapter 6

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.

We know from Theorem 5.21 that a square matrix $\boldsymbol{A}$ can be diagonalized by a unitary similarity transformation if and only if it is normal, that is $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^*$. In particular, if $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is normal then it has a set of orthonormal eigenpairs $(\lambda_1, \boldsymbol{u}_1), \ldots, (\lambda_n, \boldsymbol{u}_n)$. Letting $\boldsymbol{U} := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathbb{C}^{n\times n}$ and $\boldsymbol{D} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ we have the spectral decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*, \text{ where } \boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}. \tag{6.1}$$

6.1 SVD and SVF

The singular value decomposition (SVD) is a generalization of the spectral decomposition, to any matrix, even a rectangular one. For any $m, n \in \mathbb{N}$ we say that $\boldsymbol{D} \in \mathbb{C}^{m\times n}$ is a diagonal matrix if $d_{i,j} = 0$ for all $i \neq j$. A diagonal matrix is a **nonnegative (positive) diagonal matrix** if all the diagonal elements $d_{i,i}$, $i = 1, \ldots, \min(m, n)$ are nonnegative (positive).

6.1.1 Definition and examples

Definition 6.1 (SVD)
A decomposition of $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ of the form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, where $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{C}^{n\times n}$ are unitary, and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is a nonnegative diagonal matrix, is called a **singular value decomposition (SVD)** *of $\boldsymbol{A}$. If $\boldsymbol{A}$ is real, then $\boldsymbol{U}$ and $\boldsymbol{V}$*

are real and orthogonal and an SVD takes the form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. *The diagonal elements of* Σ *are denoted* $\sigma_1, \ldots, \sigma_{\min(m,n)}$, *and are called* **singular values**. *The columns* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ *of* $\boldsymbol{U}$ *and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ *of* $\boldsymbol{V}$ *are called* **left singular vectors** *and* **right singular vectors**, *respectively. The SVD is* **ordered** *if* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$. *For a fixed* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *we define* $\sigma_j := 0$ *for all integers* $j > \min(m, n)$.

Example 6.2 (SVD1)
The decomposition

$$\boldsymbol{A} := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \tag{6.2}$$

is both a spectral decomposition and a singular value decomposition. Indeed, $\boldsymbol{A}$ *has eigenpairs* $(2, [1, 1]^T)$ *and* $(0, [1, -1]^T)$ *and normalizing the eigenvectors we obtain a spectral decomposition. Since the elements of the diagonal matrix are nonnegative this is also a singular value decomposition* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ *with* $\Sigma = \boldsymbol{D}$ *and* $\boldsymbol{V} = \boldsymbol{U}$.

Example 6.3 (SVD2)
The matrix $\boldsymbol{A} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$ *is not normal and therefore does not have a spectral decomposition. Since it has an eigenvalue zero of algebraic multiplicity 2, but only one eigenvector* $[1, 1]^T$ *it is defective and cannot be diagonalized by any similarity transformtation. But*

$$\boldsymbol{A} := \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} =: \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{6.3}$$

is a singular value decomposition.

Example 6.4 (SVD3)
The matrix $\boldsymbol{A} := \frac{1}{25}\begin{bmatrix} 11 & 48 \\ 48 & 39 \end{bmatrix}$ *is symmetric with the eigenpairs* $(3, [3, 4]^T)$ *and* $(-1, [-4, 3]^T)$. *Normalizing the eigenvectors we obtain the spectral decomposition*

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

This is not a singular value decomposition since one of the elements of the diagonal matrix is negative. A singular value decomposition is given by

$$\boldsymbol{A} = \frac{1}{25}\begin{bmatrix} 11 & 48 \\ 48 & 39 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \frac{1}{5}\begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{6.4}$$

6.1.2 Existence

Every matrix has a singular value decomposition. To show this we first consider the matrices $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$.

Theorem 6.5 (The matrices $\boldsymbol{A}^*\boldsymbol{A}$, $\boldsymbol{A}\boldsymbol{A}^*$)
Suppose $m, n \in \mathbb{N}$ and $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. Then

1. *The matrices $\boldsymbol{A}^*\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{A}\boldsymbol{A}^* \in \mathbb{C}^{m\times m}$ are Hermitian with nonnegative eigenvalues.*

2. *The characteristic polynomials of these matrices are closely related:*

$$\lambda^m \pi_{\boldsymbol{A}^*\boldsymbol{A}}(\lambda) = \lambda^n \pi_{\boldsymbol{A}\boldsymbol{A}^*}(\lambda), \quad \lambda \in \mathbb{C}.$$

3. *Let $(\lambda_j, \boldsymbol{v}_j)$ be orthonormal eigenpairs for $\boldsymbol{A}^*\boldsymbol{A}$. If $\lambda_j > 0$, $j = 1, \ldots, r$ and $\lambda_j = 0$, $j = r+1, \ldots, n$ then $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ is an orthogonal basis for the column space* $\operatorname{span}(\boldsymbol{A}) := \{\boldsymbol{A}\boldsymbol{y} \in \mathbb{C}^m : \boldsymbol{y} \in \mathbb{C}^n\}$ *and $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$ is an orthonormal basis for the nullspace* $\ker(\boldsymbol{A}) := \{\boldsymbol{y} \in \mathbb{C}^n : \boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}\}$.

4. *Let $(\lambda_j, \boldsymbol{u}_j)$ be orthonormal eigenpairs for $\boldsymbol{A}\boldsymbol{A}^*$. If $\lambda_j > 0$, $j = 1, \ldots, r$ and $\lambda_j = 0$, $j = r+1, \ldots, m$ then $\{\boldsymbol{A}^*\boldsymbol{u}_1, \ldots, \boldsymbol{A}^*\boldsymbol{u}_r\}$ is an orthogonal basis for the column space* $\operatorname{span}(\boldsymbol{A}^*)$ *and $\{\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m\}$ is an orthonormal basis for the nullspace* $\ker(\boldsymbol{A}^*)$.

5. *The rank of $\boldsymbol{A}$ equals the number of positive eigenvalues of $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$.*

Proof.

1. Clearly $\boldsymbol{B}_1 := \boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{B}_2 := \boldsymbol{A}\boldsymbol{A}^*$ are Hermitian. If $\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$ with $\boldsymbol{v} \neq \boldsymbol{0}$, then
$$\lambda = \frac{\boldsymbol{v}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} = \frac{\|\boldsymbol{A}\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \geq 0 \tag{6.5}$$
and the eigenvalues of $\boldsymbol{B}_1$ are nonnegative. Similarly, $\boldsymbol{B}_2$ has nonnegative eigenvalues.

2. This follows from Corollary 5.10.

3. By orthonormality of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ we have $(\boldsymbol{A}\boldsymbol{v}_j)^*\boldsymbol{A}\boldsymbol{v}_k = \boldsymbol{v}_j^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}_k = \lambda_k\boldsymbol{v}_j^*\boldsymbol{v}_k = 0, j \neq k$, showing that $\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_n$ are orthogonal vectors. Moreover, (6.5) implies that $\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r$ are nonzero and $\boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{0}$ for $j = r+1, \ldots, n$. In particular, the elements of $\{\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r\}$ and $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n\}$ are linearly independent vectors in $\operatorname{span}(\boldsymbol{A})$ and $\ker(\boldsymbol{A})$, respectively. The proof will be complete once it is shown that $\operatorname{span}(\boldsymbol{A}) \subset \operatorname{span}(\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r)$ and $\ker(\boldsymbol{A}) \subset \operatorname{span}(\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n)$. Suppose $\boldsymbol{x} \in \operatorname{span}(\boldsymbol{A})$. Then $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y}$

for some $\boldsymbol{y} \in \mathbb{C}^n$, Let $\boldsymbol{y} = \sum_{j=1}^n c_j \boldsymbol{v}_j$ be an eigenvector expansion of $\boldsymbol{y}$. Since $\boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{0}$ for $j = r+1, \ldots, n$ we obtain $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} = \sum_{j=1}^n c_j \boldsymbol{A}\boldsymbol{v}_j = \sum_{j=1}^r c_j \boldsymbol{A}\boldsymbol{v}_j \in \operatorname{span}(\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r)$. Finally, if $\boldsymbol{y} = \sum_{j=1}^n c_j \boldsymbol{v}_j \in \ker(\boldsymbol{A})$, then we have $\boldsymbol{A}\boldsymbol{y} = \sum_{j=1}^r c_j \boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{0}$, and $c_1 = \cdots = c_r = 0$ since $\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r$ are linearly independent. But then $\boldsymbol{y} = \sum_{j=r+1}^n c_j \boldsymbol{v}_j \in \operatorname{span}(\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n)$.

4. Since $\boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{B}^*\boldsymbol{B}$ with $\boldsymbol{B} := \boldsymbol{A}^*$ this follows from part 3 with $\boldsymbol{A} = \boldsymbol{B}$.

5. By part 1 and 2 $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$ have the same number r of positive eigenvalues and by part 3 and 4 r is the rank of $\boldsymbol{A}$.

□

Theorem 6.6 (The matrices $\boldsymbol{A}^*\boldsymbol{A}$, $\boldsymbol{A}\boldsymbol{A}^*$ and SVD)
If $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$ is a singular value decomposition of $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ and $\sigma_j := 0$ for $j > \min(m,n)$ then

1. *$\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{V}\operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)\boldsymbol{V}^*$ is a spectral decomposition of $\boldsymbol{A}^*\boldsymbol{A}$.*

2. *$\boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{U}\operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2)\boldsymbol{U}^*$ is a spectral decomposition of $\boldsymbol{A}\boldsymbol{A}^*$.*

3. *The columns of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}^*$.*

4. *The columns of $\boldsymbol{V}$ are orthonormal eigenvectors of $\boldsymbol{A}^*\boldsymbol{A}$.*

5. *The rank of $\boldsymbol{A}$ is equal to the number of positive singular values.*

Proof. We assume $m \geq n$. The case $m < n$ is similar. If $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]\boldsymbol{\Sigma}[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]^*$ is a singular value factorization of $\boldsymbol{A}$ then $\boldsymbol{A}^*\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*)^*(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*) = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^*$ and part 1 follows. Part 2 is similar. Since these are spectral decompositions part 3 and 4 follow. Part 5 follows from part 5 of Theorem 6.5. □

Theorem 6.7 (Existence of SVD)
Suppose for $m, n, r \in \mathbb{N}$ that $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ has rank r, and that $(\lambda_j, \boldsymbol{v}_j)$ are orthonormal eigenpairs for $\boldsymbol{A}^*\boldsymbol{A}$ with $\lambda_1 \geq \cdots \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$. Define

1. *$\boldsymbol{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{C}^{n\times n}$,*

2. *$\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with diagonal elements $\sigma_j := \sqrt{\lambda_j}$ for $j = 1, \ldots, \min(m,n)$,*

3. $\boldsymbol{U} := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m] \in \mathbb{C}^{m\times m}$, *where* $\boldsymbol{u}_j = \sigma_j^{-1}\boldsymbol{A}\boldsymbol{v}_j$ *for* $j = 1, \ldots, r$ *and* $\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m$ *is an extension of* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ *to an orthonormal basis* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ *for* $\mathbb{C}^m$.

Then $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ *is an ordered singular value decomposition of* $\boldsymbol{A}$.

Proof. Let $\boldsymbol{\Sigma}, \boldsymbol{U}, \boldsymbol{V}$ be as in the theorem. The vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ are orthonormal since $\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r$ are orthogonal and $\sigma_j = \|\boldsymbol{A}\boldsymbol{v}_j\|_2 > 0$, $j = 1, \ldots, r$ by (6.5). But then $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary and $\boldsymbol{\Sigma}$ is a nonnegative diagonal matrix. Moreover,

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

Thus $\boldsymbol{U}\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{V}$ and since $\boldsymbol{V}$ is unitary we find $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^* = \boldsymbol{A}$ and we have an ordered SVD of $\boldsymbol{A}$. □

6.1.3 The singular value factorization

Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ is an ordered singular value decomposition of $\boldsymbol{A}$ of rank r. The matrix Σ can be partitioned in the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0}_{r,n-r} \\ \boldsymbol{0}_{m-r,r} & \boldsymbol{0}_{m-r,n-r} \end{bmatrix} \in \mathbb{R}^{m\times n}, \text{ where } \boldsymbol{\Sigma}_1 := \operatorname{diag}(\sigma_1, \ldots, \sigma_r), \tag{6.6}$$

Thus Σ_1 contains the r ordered positive singular values on the diagonal. Here, for $k, l \geq 0$ the symbol $\boldsymbol{0}_{k,l} = [\,]$ denotes the empty matrix if $k = 0$ or $l = 0$, and the zero matrix with k rows and l columns otherwise.

Using the block partitions

$$\begin{aligned} \boldsymbol{U} &= [\boldsymbol{U}_1, \boldsymbol{U}_2] \in \mathbb{C}^{m\times m}, \quad \boldsymbol{U}_1 := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r], \quad \boldsymbol{U}_2 := [\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m], \\ \boldsymbol{V} &= [\boldsymbol{V}_1, \boldsymbol{V}_2] \in \mathbb{C}^{n\times n}, \quad \boldsymbol{V}_1 := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r], \quad \boldsymbol{V}_2 := [\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n], \end{aligned} \tag{6.7}$$

we obtain by block multiplication

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*. \tag{6.8}$$

As an example:

$$\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}[2]\,\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \end{bmatrix}.$$

Definition 6.8 (SVF)
Let $m, n, r \in \mathbb{N}$ *with* $1 \leq r \leq \min(m, n)$. *A* **singular value factorization (SVF)** *is a factorization of* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *of the form* $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$, *where* $\boldsymbol{U}_1 \in \mathbb{C}^{m\times r}$ *and* $\boldsymbol{V}_1 \in \mathbb{C}^{n\times r}$ *have orthonormal columns, and* $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{r\times r}$ *is a diagonal matrix with positive diagonal elements. We say that the SVF is* **ordered** *if the diagonal elements of* Σ_1 *are ordered.*

An SVD and an SVF of a matrix $\boldsymbol{A}$ are closely related.

1. Let $\boldsymbol{A}$ have rank r and let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be an ordered SVD of $\boldsymbol{A}$. Then $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$ is an ordered SVF of $\boldsymbol{A}$. Moreover, $\boldsymbol{U}_1$, $\boldsymbol{V}_1$ contain the first r columns of $\boldsymbol{U}$, $\boldsymbol{V}$ and Σ_1 is a diagonal matrix with the positive singular values on the diagonal.

2. Conversely, suppose $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$ is a singular value factorization of $\boldsymbol{A}$ with $\Sigma_1 \in \mathbb{R}^{r\times r}$. Extend $\boldsymbol{U}_1$ and $\boldsymbol{V}_1$ in any way to unitary matrices $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{C}^{n\times n}$, and let $\boldsymbol{\Sigma}$ be given by (6.6). Then $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ is an SVD of $\boldsymbol{A}$. Moreover, r is uniquely given as the rank of $\boldsymbol{A}$.

3. If $\boldsymbol{A} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]^*$ is a singular value factorization of $\boldsymbol{A}$ then
$$\boldsymbol{A} = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^*. \tag{6.9}$$
This is known as the **outer product form** of the SVF.

4. We note that a nonsingular square matrix has full rank and only positive singular values. Thus the SVD and SVF are the same for a nonsingular matrix.

Theorem 6.9 (Singular values of a normal matrix)
The singular values of a symmetric positive semidefinite matrix are its eigenvalues. The singular values of a normal matrix are the absolute values of its eigenvalues.

Proof. If $\boldsymbol{A}$ is normal then by Theorem 5.21, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$, where $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ contains the eigenvalues of $\boldsymbol{A}$. Now $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}^*\boldsymbol{D}\boldsymbol{U}^*$, and $\boldsymbol{D}^*\boldsymbol{D} = \mathrm{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$ and by Theorem 6.6 $\sigma_j = \sqrt{|\lambda_j|^2} = |\lambda_j|$ for $j = 1, \ldots, n$. If $\boldsymbol{A}$ is symmetric positive semidefinite then the eigenvalues are nonnegative. □

6.1.4 Examples

We use Theorem 6.7 to derive some singular value factorizations and decompositions.

Example 6.10 (Nonsingular matrix)
Derive the SVF and SVD of the matrix in (6.4)*. Discussion: Eigenpairs of* $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \left[\begin{smallmatrix} 97 & 96 \\ 96 & 153 \end{smallmatrix}\right]/25$ *are given by*

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

Taking square roots and normalizing we find $\sigma_1 = 3$, $\sigma_2 = 1$, $\boldsymbol{v}_1 = \left[\begin{smallmatrix}3\\4\end{smallmatrix}\right]/5$, *and* $\boldsymbol{v}_2 = \left[\begin{smallmatrix}4\\-3\end{smallmatrix}\right]/5$. *Thus* $\boldsymbol{u}_1 := \boldsymbol{A}\boldsymbol{v}_1/\sigma_1 = \left[\begin{smallmatrix}3\\4\end{smallmatrix}\right]/5$, $\boldsymbol{u}_2 := \boldsymbol{A}\boldsymbol{v}_2/\sigma_2 = \left[\begin{smallmatrix}-4\\3\end{smallmatrix}\right]/5$, *and* (6.4) *follows. Since* $\boldsymbol{A}$ *is nonsingular this is both an SVF and an SVD of* $\boldsymbol{A}$.

Example 6.11 (Full column rank)
Find the SVF and SVD of

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

Discussion: Eigenpairs of

$$\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \frac{1}{25}\begin{bmatrix}52 & 36\\ 36 & 73\end{bmatrix}$$

are found from

$$\boldsymbol{B}\begin{bmatrix}3\\4\end{bmatrix} = 4\begin{bmatrix}3\\4\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}4\\-3\end{bmatrix} = 1\begin{bmatrix}4\\-3\end{bmatrix}.$$

Thus $\sigma_1 = 2$, $\sigma_2 = 1$, *and* $\boldsymbol{V} = \frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}$. *Now* $\boldsymbol{u}_1 = \boldsymbol{A}\boldsymbol{u}/\sigma_1 = [1,2,2]^T/3$, $\boldsymbol{u}_2 = \boldsymbol{A}\boldsymbol{v}_2/\sigma_2 = [2,-2,1]^T/3$ *giving the singular value factorization*

$$\boldsymbol{A} = \frac{1}{3}\begin{bmatrix}1 & 2\\ 2 & -2\\ 2 & 1\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 1\end{bmatrix}\frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}.$$

For an SVD 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 normalizing* $\boldsymbol{u}_3$ *we obtain the singular value decomposition*

$$\boldsymbol{A} = \frac{1}{3}\begin{bmatrix}1 & 2 & 2\\ 2 & -2 & 1\\ 2 & 1 & -2\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}\frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}. \tag{6.10}$$

Example 6.12 (Full row rank)
Find the SVF and SVD of

$$\boldsymbol{A}_1 := \frac{1}{15}\begin{bmatrix}14 & 4 & 16\\ 2 & 22 & 13\end{bmatrix} \in \mathbb{R}^{2\times 3}.$$

Discussion: Since $\boldsymbol{A}_1 = \boldsymbol{A}^T$, *where* $\boldsymbol{A}$ *is the matrix in Example 6.11 we can find an SVF and SVD of* $\boldsymbol{A}_1$ *by simply transposing the corresponding factorization of*

$\boldsymbol{A}$. *Thus*

$$\begin{aligned}\boldsymbol{A}_1 &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T = \frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}\begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\frac{1}{3}\begin{bmatrix}1 & 2 & 2\\ 2 & -2 & 1\\ 2 & 1 & -2\end{bmatrix}\\ &= (\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T)^T = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^T\boldsymbol{U}_1^T = \frac{1}{5}\begin{bmatrix}3 & 4\\ 4 & -3\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 1\end{bmatrix}\frac{1}{3}\begin{bmatrix}1 & 2 & 2\\ 2 & -2 & 1\end{bmatrix}.\end{aligned} \tag{6.11}$$

Example 6.13 ($r < n < m$)
Find the SVD of

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

Discussion: Eigenpairs of

$$\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}$$

are derived from

$$\boldsymbol{B}\begin{bmatrix}1\\ 1\end{bmatrix} = 4\begin{bmatrix}1\\ 1\end{bmatrix}, \quad \boldsymbol{B}\begin{bmatrix}1\\ -1\end{bmatrix} = 0\begin{bmatrix}1\\ -1\end{bmatrix}$$

and we find $\sigma_1 = 2$, $\sigma_2 = 0$, *Thus* $r = 1$, $m = 3$, $n = 2$ *and*

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_1 & 0\\ 0 & 0\\ 0 & 0\end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = [2], \quad \boldsymbol{V} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}.$$

We find $\boldsymbol{u}_1 = \boldsymbol{A}\boldsymbol{v}_1/\sigma_1 = \boldsymbol{s}_1/\sqrt{2}$, *where* $\boldsymbol{s}_1 = [1, 1, 0]^T$, *and the SVF of* $\boldsymbol{A}$ *is given by*

$$\boldsymbol{A} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\ 1\\ 0\end{bmatrix}[2]\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\end{bmatrix}.$$

To find an SVD we need to extend $\boldsymbol{u}_1$ *to an orthonormal basis for* $\mathbb{R}^3$. *We first 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 (26)

$$\boldsymbol{w}_1 = \boldsymbol{s}_1,\ \boldsymbol{w}_2 = \boldsymbol{s}_2 - \frac{\boldsymbol{s}_2^T\boldsymbol{w}_1}{\boldsymbol{w}_1^T\boldsymbol{w}_1}\boldsymbol{w}_1 = \begin{bmatrix}-1/2\\ 1/2\\ 0\end{bmatrix},\ \boldsymbol{w}_3 = \boldsymbol{s}_3 - \frac{\boldsymbol{s}_3^T\boldsymbol{w}_1}{\boldsymbol{w}_1^T\boldsymbol{w}_1}\boldsymbol{w}_1 - \frac{\boldsymbol{s}_3^T\boldsymbol{w}_2}{\boldsymbol{w}_2^T\boldsymbol{w}_2}\boldsymbol{w}_2 = \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}.$$

Normalizing the $\boldsymbol{w}_i$*'s we obtain* $\boldsymbol{u}_1 = \boldsymbol{w}_1/\|\boldsymbol{w}_1\|_2 = [1/\sqrt{2}, 1/\sqrt{2}, 0]^T$, $\boldsymbol{u}_2 = \boldsymbol{w}_2/\|\boldsymbol{w}_2\|_2 = [-1/\sqrt{2}, 1/\sqrt{2}, 0]^T$, *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*

$$\boldsymbol{U} := \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3,3}, \quad \boldsymbol{\Sigma} := \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{3,2}, \quad \boldsymbol{V} := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \in \mathbb{R}^{2,2}.$$

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. Standard computer implementations of the singular value decomposition ([29]) first reduces $\boldsymbol{A}$ to bidiagonal form and then 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 13.

Exercise 6.14 (SVD examples)
Find the singular value decomposition of the following matrices

(a) $\boldsymbol{A} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$.

(b) $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}$.

Exercise 6.15 (More SVD examples)
Find the singular value decomposition of the following matrices

(a) $\boldsymbol{A} = \boldsymbol{e}_1$ *the first unit vector in* $\mathbb{R}^m$.

(b) $\boldsymbol{A} = \boldsymbol{e}_n^T$ *the last unit vector in* $\mathbb{R}^n$.

(c) $\boldsymbol{A} = \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}$.

6.2 SVD and the Four Fundamental Subspaces

The singular vectors form orthonormal bases for the four fundamental subspaces $\text{span}(\boldsymbol{A})$, $\ker(\boldsymbol{A})$, $\text{span}(\boldsymbol{A}^*)$, and $\ker(\boldsymbol{A}^*)$.

Theorem 6.16 (Singular vectors and orthonormal bases)
For positive integers m, n *let* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *have rank* r *and a singular value decomposition* $\boldsymbol{A} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]\boldsymbol{\Sigma}[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. *Then the singular vectors satisfy*

$$\begin{aligned} \boldsymbol{A}\boldsymbol{v}_i &= \sigma_i \boldsymbol{u}_i, \ i = 1, \ldots, r, \quad \boldsymbol{A}\boldsymbol{v}_i = 0, \ i = r+1, \ldots, n, \\ \boldsymbol{A}^*\boldsymbol{u}_i &= \sigma_i \boldsymbol{v}_i, \ i = 1, \ldots, r, \quad \boldsymbol{A}^*\boldsymbol{u}_i = 0, \ i = r+1, \ldots, m. \end{aligned} \tag{6.12}$$

Moreover,

$$\begin{aligned}
&1.\ \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_r\} \text{ is an orthonormal basis for } \operatorname{span}(\boldsymbol{A}),\\
&2.\ \{\boldsymbol{u}_{r+1},\ldots,\boldsymbol{u}_m\} \text{ is an orthonormal basis for } \ker(\boldsymbol{A}^*),\\
&3.\ \{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r\} \text{ is an orthonormal basis for } \operatorname{span}(\boldsymbol{A}^*),\\
&4.\ \{\boldsymbol{v}_{r+1},\ldots,\boldsymbol{v}_n\} \text{ is an orthonormal basis for } \ker(\boldsymbol{A}).
\end{aligned} \tag{6.13}$$

Proof. If $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ then $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}$, or in terms of the block partition (6.7) $\boldsymbol{A}[\boldsymbol{V}_1, \boldsymbol{V}_2] = [\boldsymbol{U}_1, \boldsymbol{U}_2]\left[\begin{smallmatrix}\Sigma_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{smallmatrix}\right]$. But then $\boldsymbol{A}\boldsymbol{V}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1$, $\boldsymbol{A}\boldsymbol{V}_2 = \boldsymbol{0}$, and this implies the first part of (6.12). Taking conjugate transpose of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ gives $\boldsymbol{A}^* = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^*$ or $\boldsymbol{A}^*\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^T$. Using the block partition as before we obtain the last part of (6.12).

It follows from Theorem 6.5 that $\{\boldsymbol{A}\boldsymbol{v}_1,\ldots,\boldsymbol{A}\boldsymbol{v}_r\}$ is an orthogonal basis for $\operatorname{span}(\boldsymbol{A})$ and $\{\boldsymbol{v}_{r+1},\ldots,\boldsymbol{v}_m\}$ is an orthonormal basis for $\ker(\boldsymbol{A})$. Applying this theorem to $\boldsymbol{A}\boldsymbol{A}^*$ it also follows that $\{\boldsymbol{A}^*\boldsymbol{u}_1,\ldots,\boldsymbol{A}^*\boldsymbol{u}_r\}$ is an orthogonal basis for $\operatorname{span}(\boldsymbol{A}^*)$ and $\{\boldsymbol{u}_{r+1},\ldots,\boldsymbol{u}_m\}$ is an orthonormal basis for $\ker(\boldsymbol{A}^*)$. By (6.12) $\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_r\}$ is an orthonormal basis for $\operatorname{span}(\boldsymbol{A})$ and $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r\}$ is an orthonormal basis for $\operatorname{span}(\boldsymbol{A}^*)$. □

Exercise 6.17 (Counting dimensions of fundamental subspaces)
Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. Show using SVD that

1. $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}^*)$.
2. $\operatorname{rank}(\boldsymbol{A}) + \operatorname{null}(\boldsymbol{A}) = n$,
3. $\operatorname{rank}(\boldsymbol{A}) + \operatorname{null}(\boldsymbol{A}^*) = m$,

where $\operatorname{null}(\boldsymbol{A})$ *is defined as the dimension of* $\ker(\boldsymbol{A})$.

Exercise 6.18 (Rank and nullity relations)
Use Theorem 6.5 to show that for any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$

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* $\operatorname{null}(\boldsymbol{A}\boldsymbol{A}^*) = \operatorname{null}(\boldsymbol{A}^*)$.

Exercise 6.19 (Orthonormal bases example)
Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be as in Example 6.11. Give orthonormal bases for $\operatorname{span}(\boldsymbol{B})$ *and* $\ker(\boldsymbol{B})$.

Exercise 6.20 (Some spanning sets)
Show for any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ that $\operatorname{span}(\boldsymbol{A}^*\boldsymbol{A}) = \operatorname{span}(\boldsymbol{V}_1) = \operatorname{span}(\boldsymbol{A}^*)$

Exercise 6.21 (Singular values and eigenpair of composite matrix)
Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ 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

$$\boldsymbol{p}_i = \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix}, \quad \boldsymbol{q}_i = \begin{bmatrix} \boldsymbol{u}_i \\ -\boldsymbol{v}_i \end{bmatrix}, \quad \boldsymbol{r}_j = \begin{bmatrix} \boldsymbol{u}_j \\ \boldsymbol{0} \end{bmatrix}, \text{ for } i = 1, \ldots, n \text{ and } j = n+1, \ldots, m.$$

6.3 A Geometric Interpretation

The singular value decomposition and factorization give insight into the geometry of a linear transformation. Consider the linear transformation $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^m$ given by $\boldsymbol{T}\boldsymbol{z} := \boldsymbol{A}\boldsymbol{z}$ where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Assume that $\operatorname{rank}(\boldsymbol{A}) = n$. The function $\boldsymbol{T}$ maps the unit sphere $\mathcal{S} := \{\boldsymbol{z} \in \mathbb{R}^n : \|\boldsymbol{z}\|_2 = 1\}$ onto an ellipsoid $\mathcal{E} := \boldsymbol{A}\mathcal{S} = \{\boldsymbol{A}\boldsymbol{z} : \boldsymbol{z} \in \mathcal{S}\}$ in $\mathbb{R}^m$.

Theorem 6.22 (SVF ellipse)
Suppose $\boldsymbol{A} \in \mathbb{R}^{m,n}$ has rank $r = n$, and let $\boldsymbol{A} = \boldsymbol{U}_1\Sigma_1\boldsymbol{V}_1^T$ be a singular value factorization of $\boldsymbol{A}$. Then

$$\mathcal{E} = \boldsymbol{U}_1\tilde{\mathcal{E}} \text{ where } \tilde{\mathcal{E}} := \{\boldsymbol{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n : \frac{y_1^2}{\sigma_1^2} + \cdots + \frac{y_n^2}{\sigma_n^2} = 1\}.$$

Proof. Supose $\boldsymbol{z} \in \mathcal{S}$. Now $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{U}_1\Sigma_1\boldsymbol{V}_1^T\boldsymbol{z} = \boldsymbol{U}_1\boldsymbol{y}$, where $\boldsymbol{y} := \Sigma_1\boldsymbol{V}_1^T\boldsymbol{z}$. Since $\operatorname{rank}(\boldsymbol{A}) = n$ it follows that $\boldsymbol{V}_1 = \boldsymbol{V}$ is square so that $\boldsymbol{V}_1\boldsymbol{V}_1^T = \boldsymbol{I}$. But then $\boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{y} = \boldsymbol{z}$ and we obtain

$$1 = \|\boldsymbol{z}\|_2^2 = \|\boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{y}\|_2^2 = \|\boldsymbol{\Sigma}_1^{-1}\boldsymbol{y}\|_2^2 = \frac{y_1^2}{\sigma_1^2} + \cdots + \frac{y_n^2}{\sigma_n^2}.$$

This implies that $\boldsymbol{y} \in \tilde{\mathcal{E}}$. Finally, $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} = \boldsymbol{U}_1\Sigma_1\boldsymbol{V}_1^T\boldsymbol{z} = \boldsymbol{U}_1\boldsymbol{y}$, where $\boldsymbol{y} \in \tilde{\mathcal{E}}$ implies that $\mathcal{E} = \boldsymbol{U}_1\tilde{\mathcal{E}}$. ∎

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 orthonormal transformation $\boldsymbol{U}_1\boldsymbol{y} \to \boldsymbol{x}$ preserves length, the image $\mathcal{E} = \boldsymbol{A}\mathcal{S}$ is a rotated ellipsoid

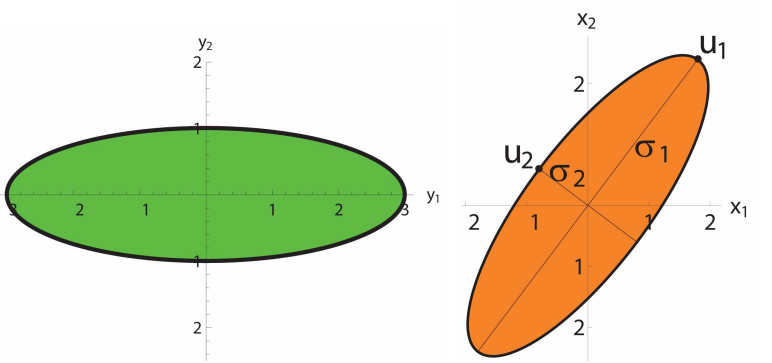


Figure 6.1. *The ellipse $y_1^2/9 + y_2^2 = 1$ (left) and the rotated ellipse $\boldsymbol{A}\mathcal{S}$ (right).*

with semiaxes along the left singular vectors $\boldsymbol{u}_j = \boldsymbol{U}\boldsymbol{e}_j$, of length σ_j, $j = 1, \ldots, n$. Since $\boldsymbol{A}\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j$, for $j = 1, \ldots, n$ the right singular vectors defines points in $\mathcal{S}$ that are mapped onto the semiaxes of $\mathcal{E}$.

Example 6.23 (Ellipse)
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 6.10. Recall that $\sigma_1 = 3$, $\sigma_2 = 1$, $\boldsymbol{u}_1 = [3, 4]^T/5$ and $\boldsymbol{u}_2 = [-4, 3]^T/5$. The ellipses $y_1^2/\sigma_1^2 + y_2^2/\sigma_2^2 = 1$ and $\mathcal{E} = \boldsymbol{A}\mathcal{S} = \boldsymbol{U}_1\tilde{\mathcal{E}}$ are shown in Figure 6.1. Since $\boldsymbol{y} = \boldsymbol{U}_1^T\boldsymbol{x} = [3/5x_1 + 4/5x_2, -4/5x_1 + 3/5x_2]^T$, the equation for the ellipse on the right is

$$\frac{(\frac{3}{5}x_1 + \frac{4}{5}x_2)^2}{9} + \frac{(-\frac{4}{5}x_1 + \frac{3}{5}x_2)^2}{1} = 1,$$

6.4 Determining the Rank of a Matrix Numerically

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 6.7 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.

6.4.1 The Frobenius norm

This commonly occuring matrix norm will be used here in a discussion of how many of the computed singular values can possibly be considered to be zero. The **Frobenius norm**, of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is defined by

$$\|\boldsymbol{A}\|_F := \big(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\big)^{1/2}. \tag{6.14}$$

There is a relation between the Frobenius norm of a matrix and its singular values. First we derive some elementary properties of this norm. A systematic study of matrix norms is given in the next chapter.

Lemma 6.24 (Frobenius norm properties)
For any $m, n \in \mathbb{N}$ and any matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$

1. $\|\boldsymbol{A}^*\|_F = \|\boldsymbol{A}\|_F$,
2. $\|\boldsymbol{A}\|_F^2 = \sum_{j=1}^{n}\|\boldsymbol{a}_{:j}\|_2^2$,
3. $\|\boldsymbol{UA}\|_F = \|\boldsymbol{AV}\|_F = \|\boldsymbol{A}\|_F$ *for any unitary matrices* $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ *and* $\boldsymbol{V} \in \mathbb{C}^{n\times n}$,
4. $\|\boldsymbol{AB}\|_F \le \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F$ *for any* $\boldsymbol{B} \in \mathbb{C}^{n,k}, \quad k \in \mathbb{N}$,
5. $\|\boldsymbol{Ax}\|_2 \le \|\boldsymbol{A}\|_F\|\boldsymbol{x}\|_2$, *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. This follows since the Frobenius norm is the Euclidian norm of a vector, $\|\boldsymbol{A}\|_F := \|\mathrm{vec}(\boldsymbol{A})\|_2$, 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.
3. Recall that if $\boldsymbol{U}^*\boldsymbol{U} = I$ then $\|\boldsymbol{Ux}\|_2 = \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \mathbb{C}^n$. Applying this to each column $\boldsymbol{a}_{:j}$ of $\boldsymbol{A}$ we find $\|\boldsymbol{UA}\|_F^2 \overset{2.}{=} \sum_{j=1}^{n}\|\boldsymbol{Ua}_{:j}\|_2^2 = \sum_{j=1}^{n}\|\boldsymbol{a}_{:j}\|_2^2 \overset{2.}{=} \|\boldsymbol{A}\|_F^2$. Similarly, since $\boldsymbol{VV}^* = I$ we find $\|\boldsymbol{AV}\|_F \overset{1.}{=} \|\boldsymbol{V}^*\boldsymbol{A}^*\|_F = \|\boldsymbol{A}^*\|_F \overset{1.}{=} \|\boldsymbol{A}\|_F$.

4. Using the Cauchy-Schwarz inequality and 2. we obtain

$$\|\boldsymbol{AB}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}|\boldsymbol{a}_{i:}^T\boldsymbol{b}_{:j}|^2 \leq \sum_{i=1}^{m}\sum_{j=1}^{k}\|\boldsymbol{a}_{i:}\|_2^2\|\boldsymbol{b}_{:j}\|_2^2 = \|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2.$$

5. Since $\|\boldsymbol{v}\|_F = \|\boldsymbol{v}\|_2$ for a vector this follows by taking $k = 1$ and $\boldsymbol{B} = \boldsymbol{x}$ in 4.

□

Theorem 6.25 (Frobenius norm and singular values)
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 6.24 we find $\|\boldsymbol{A}\|_F \overset{3.}{=} \|\boldsymbol{U}^*\boldsymbol{AV}\|_F = \|\boldsymbol{\Sigma}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$.
□

6.4.2 Low rank approximation

Suppose $m \geq n \geq 1$ and $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ has an ordered singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\left[\begin{smallmatrix}\boldsymbol{D}\\ \boldsymbol{0}\end{smallmatrix}\right]\boldsymbol{V}^*$, where $\boldsymbol{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. 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 $\boldsymbol{A'} := \boldsymbol{U}\left[\begin{smallmatrix}\boldsymbol{D'}\\ \boldsymbol{0}\end{smallmatrix}\right]\boldsymbol{V}^*$, where $\boldsymbol{D'} := \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \in \mathbb{R}^{n\times n}$. By Lemma 6.24

$$\|\boldsymbol{A} - \boldsymbol{A'}\|_F = \|\boldsymbol{U}\left[\begin{smallmatrix}\boldsymbol{D}-\boldsymbol{D'}\\ \boldsymbol{0}\end{smallmatrix}\right]\boldsymbol{V}^*\|_F = \|\left[\begin{smallmatrix}\boldsymbol{D}-\boldsymbol{D'}\\ \boldsymbol{0}\end{smallmatrix}\right]\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2} < \epsilon.$$

Thus, if ϵ is small then $\boldsymbol{A}$ 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 6.26 (Best low rank approximation)
Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *has singular values* $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. *For any* $r \leq \operatorname{rank}(\boldsymbol{A})$ *we have*

$$\|\boldsymbol{A} - \boldsymbol{A'}\|_F = \min_{\substack{\boldsymbol{B}\in\mathbb{R}^{m\times n}\\ rank(\boldsymbol{B})=r}} \|\boldsymbol{A} - \boldsymbol{B}\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2}.$$

For the proof of this theorem we refer to p. 322 of [29].

Exercise 6.27 (Rank example)
Consider the singular value decomposition

$$\boldsymbol{A} := \begin{bmatrix} 0 & 3 & 3 \\ 4 & 1 & -1 \\ 4 & 1 & -1 \\ 0 & 3 & 3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \end{bmatrix}$$

(a) *Give orthonormal bases for* $\operatorname{span}(\boldsymbol{A}), \operatorname{span}(\boldsymbol{A}^T), \ker(\boldsymbol{A}), \ker(\boldsymbol{A}^T)$ *and* $\operatorname{span}(\boldsymbol{A})^\perp$.

(b) *Explain why for all matrices* $\boldsymbol{B} \in \mathbb{R}^{4,3}$ *of rank one we have* $\|\boldsymbol{A} - \boldsymbol{B}\|_F \geq 6$.

(c) *Give a matrix* $\boldsymbol{A}_1$ *of rank one such that* $\|\boldsymbol{A} - \boldsymbol{A}_1\|_F = 6$.

Exercise 6.28 (Another rank example)
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 Frobenius norm of* $\boldsymbol{A} - \boldsymbol{B}$ *is small for large* n, *and the matrix* $\boldsymbol{A}$ *is very close to being singular for large* n.

(b) *Use Theorem 6.26 to show that the smallest singular vale* σ_n *of* $\boldsymbol{A}$ *is bounded above by* 2^{2-n}.

6.5 The Minmax Theorem for Singular Values and the Hoffman-Wielandt Theorem

We have a minmax and maxmin characterization for singular values.

Theorem 6.29 (The Courant-Fischer theorem for singular values)
Suppose $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ *has singular values* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *ordered so that* $\sigma_1 \geq \cdots \geq \sigma_n$. *Then for* $k = 1, \ldots, n$

$$\sigma_k = \min_{\dim(\mathcal{S})=n-k+1} \max_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\dim(\mathcal{S})=k} \min_{\substack{\boldsymbol{x} \in \mathcal{S} \\ \boldsymbol{x} \neq \boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}. \tag{6.15}$$

Proof. Since

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = \frac{(\boldsymbol{A}\boldsymbol{x})^*(\boldsymbol{A}\boldsymbol{x})}{\boldsymbol{x}^*\boldsymbol{x}} = \frac{\boldsymbol{x}^*(\boldsymbol{A}^*\boldsymbol{A})\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}$$

is 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 5.29. ∎

By taking $k = 1$ and $k = n$ in (6.15) we obtain for any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$

$$\sigma_1 = \max_{\substack{\boldsymbol{x}\in\mathbb{C}^n \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}, \quad \sigma_n = \min_{\substack{\boldsymbol{x}\in\mathbb{C}^n \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}. \tag{6.16}$$

This follows since the only subspace of $\mathbb{C}^n$ of dimension n is $\mathbb{C}^n$ itself.

The Hoffman-Wielandt Theorem, see Theorem 5.32, for eigenvalues of Hermitian matrices can be written

$$\sum_{j=1}^{n} |\mu_j - \lambda_j|^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2 := \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij} - b_{ij}|^2, \tag{6.17}$$

where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ 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, see also Section 11.6.

Theorem 6.30 (Hoffman-Wielandt theorem for singular values)
For any $m, n \in \mathbb{N}$ and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m\times n}$ we have

$$\sum_{j=1}^{n} |\beta_j - \alpha_j|^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2. \tag{6.18}$$

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.

6.6 Proof of the Hoffman-Wielandt Theorem for Singular Values

We apply the Hoffman-Wielandt Theorem for eigenvalues to the Hermitian matrices

$$\boldsymbol{C} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^* & \boldsymbol{0} \end{bmatrix} \text{ and } \boldsymbol{D} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{0} \end{bmatrix} \in \mathbb{C}^{m+n\times m+n}.$$

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

$$\sum_{j=1}^{m+n} |\lambda_j - \mu_j|^2 \leq \|\boldsymbol{C} - \boldsymbol{D}\|_F^2. \tag{6.19}$$

Suppose $\boldsymbol{A}$ has rank r and singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. We use (6.12) 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 $\sum_{j=1}^{m+n} |\lambda_j - \mu_j|^2 = 2\sum_{i=1}^{t} |\alpha_i - \beta_i|^2$ 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 (6.19) 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 (6.18).

6.7 Review Questions

6.7.1 Consider an SVD and an SVF of a matrix $\boldsymbol{A}$.

- What are the singular values of $\boldsymbol{A}$?
- how is the SVD defined?
- how can we find an SVF if we know an SVD?
- how can we find an SVD if we know an SVF?
- what are the relations between the singular vectors?
- which singular vectors form bases for $\operatorname{span}(\boldsymbol{A})$ and $\ker(\boldsymbol{A}^*)$?

6.7.2 How are the Frobenius norm and singular values related?

6.7.3 State the Courant-Fischer theorem for singular values.

6.7.4 State the Hoffman-Wieland theorem for singular values.

Chapter 7

Matrix Norms

To measure the size of a matrix we can use a matrix norm. In this chapter we initiate a systematic study of matrix norms.

7.1 Matrix Norms

For simplicity we consider only matrix norms on the vector space $(\mathbb{C}^{m\times n}, \mathbb{C})$. All results also holds for $(\mathbb{R}^{m\times n}, \mathbb{R})$.

Definition 7.1 (Matrix norms)
Suppose m, n are positive integers. A function $\|\cdot\|\colon \mathbb{C}^{m\times n} \to \mathbb{R}$ is called a **matrix norm** *on $\mathbb{C}^{m\times n}$ if for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m\times n}$ 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 $(\mathbb{C}^{m\times n}, \mathbb{C})$ of $m \times n$ matrices. Adapting Theorem 0.19 to this situation gives

Theorem 7.2 (Matrix norm equivalence)
All matrix norms are equivalent. Thus, if $\|\cdot\|$ and $\|\cdot\|'$ are two matrix norms on $\mathbb{C}^{m\times n}$ then there are positive constants μ and M such that

$$\mu\|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|' \leq M\|\boldsymbol{A}\|$$

holds for all $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. Moreover, a matrix norm is a continuous function.

Any vector norm $\|\cdot\|_V$ on $\mathbb{C}^{mn}$ defines a matrix norm on $\mathbb{C}^{m\times n}$ given 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

$$\|\boldsymbol{A}\|_S := \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|, \quad \|\boldsymbol{A}\|_F := \Big(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\Big)^{1/2}, \quad \|\boldsymbol{A}\|_M := \max_{i,j}|a_{ij}|. \tag{7.1}$$

Ferdinand Georg Frobenius, (1849-1917).

Of these norms the Frobenius norm is the most useful. Some of its properties were derived in Lemma 6.24 and Theorem 6.25.

7.1.1 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 $\boldsymbol{AB}$ is defined in a fixed space $\mathbb{C}^{n\times n}$, while in the rectangular case matrix multiplication combines matrices in different spaces. The following definition captures this distinction.

Definition 7.3 (Consistent matrix norms)
A matrix norm is called **consistent on** $\mathbb{C}^{n\times n}$ *if*

4. $\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$ *(submultiplicativity)*

holds for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$. *A matrix norm is* **consistent** *if it is defined on* $\mathbb{C}^{m\times n}$ *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 (7.1) are defined for all $m, n \in \mathbb{N}$. From Lemma 6.24 it follows that the Frobenius norm is consistent.

Exercise 7.4 (Consitency of sum norm?)
Show that the sum norm is consistent.

Exercise 7.5 (Consitency of max norm?)
Show that the max norm is not consistent by considering $\left[\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right]$.

Exercise 7.6 (Consitency of modified max norm)

(a) *Show that the norm*

$$\|\boldsymbol{A}\| := \sqrt{mn}\|\boldsymbol{A}\|_M, \quad \boldsymbol{A} \in \mathbb{C}^{m\times n}$$

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 $\mathbb{C}^{n\times n}$ we have the inequality

$$\|\boldsymbol{A}^k\| \leq \|\boldsymbol{A}\|^k \text{ for } k \in \mathbb{N}. \tag{7.2}$$

When working with norms one often has to bound the vector norm of a matrix times a vector by the norm of the matrix times the norm of the vector. This leads to the following definition.

Definition 7.7 (Subordinate matrix norms)
Suppose $m, n \in \mathbb{N}$ *are given, let* $\|\ \|$ *on* $\mathbb{C}^m$ *and* $\|\ \|_\beta$ *on* $\mathbb{C}^n$ *be vector norms, and let* $\|\ \|$ *be a matrix norm on* $\mathbb{C}^{m\times n}$. *We say that the matrix norm* $\|\ \|$ *is* **subordinate** *to the vector norms* $\|\ \|$ *and* $\|\ \|_\beta$ *if* $\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{A}\boldsymbol{x}\|$ *for all* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *and all* $\boldsymbol{x} \in \mathbb{C}^n$. *If* $\|\ \| = \|\ \|_\beta$ *then we say that* $\|\ \|$ *is subordinate to* $\|\ \|$.

By Lemma 6.24 we have $\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A}\|_F\|\boldsymbol{x}\|_2$, for all $\boldsymbol{x} \in \mathbb{C}^n$. Thus the Frobenius norm is subordinate to the Euclidian vector norm.

Exercise 7.8 (What is the sum norm subordinate to?)
Show that the sum norm is subordinate to the l_1*-norm.*

Exercise 7.9 (What is the max norm subordinate to?)

(a) *Show that the max norm is subordinate to the* ∞ *and* 1 *norm, i. e.,* $\|\boldsymbol{A}\boldsymbol{x}\|_\infty \leq \|\boldsymbol{A}\|_M\|\boldsymbol{x}\|_1$ *holds for all* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *and all* $\boldsymbol{x} \in \mathbb{C}^n$.

(b) *Show that if* $\|\boldsymbol{A}\|_M = |a_{kl}|$, *then* $\|\boldsymbol{A}\boldsymbol{e}_l\|_\infty = \|\boldsymbol{A}\|_M\|\boldsymbol{e}_l\|_1$.

(c) *Show that* $\|\boldsymbol{A}\|_M = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_1}$.

7.1.2 Operator norms

Corresponding to vector norms on $\mathbb{C}^n$ and $\mathbb{C}^m$ there is an induced matrix norm on $\mathbb{C}^{m\times n}$ which we call the **operator norm**. It is possible to consider one vector norm on $\mathbb{C}^m$ and another vector norm on $\mathbb{C}^n$, but we treat only the case of one vector norm defined on $\mathbb{C}^n$ for all $n \in \mathbb{N}$[9].

Definition 7.10 (Operator norm)
Let $\|\ \|$ be a vector norm defined on $\mathbb{C}^n$ for all $n \in \mathbb{N}$. For given $m, n \in \mathbb{N}$ and $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ we define

$$\|\boldsymbol{A}\| := \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}. \tag{7.3}$$

We call this the **operator norm** *corresponding to the vector norm $\|\ \|$.*

With a risk of confusion we use the same symbol for the operator norm and the corresponding vector norm. Before we show that the operator 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

$$\|\boldsymbol{A}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|. \tag{7.4}$$

   The set

$$\mathcal{S} := \{x \in \mathbb{C}^n : \|\boldsymbol{x}\| = 1\} \tag{7.5}$$

   is the unit sphere in $\mathbb{C}^n$ with respect to the vector norm $\|\ \|$. It is enough to take the max over this unit sphere since

$$\max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x}\neq 0} \Big\|\boldsymbol{A}\big(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\big)\Big\| = \max_{\|\boldsymbol{y}\|=1} \|\boldsymbol{A}\boldsymbol{y}\|.$$

2. The operator norm is subordinate to the corresponding vector norm. Thus,

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\| \text{ for all } \boldsymbol{A} \in \mathbb{C}^{m\times n} \text{ and } \boldsymbol{x} \in \mathbb{C}^n. \tag{7.6}$$

3. We can use max instead of sup in (7.3). This follows by the following compactness argument. The unit sphere $\mathcal{S}$ given by (7.5) is bounded. It is also finite dimensional and closed, and hence compact. Moreover, since the vector norm $\|\ \| : \mathcal{S} \to \mathbb{R}$ is a continuous function, it follows that the function $f : \mathcal{S} \to \mathbb{R}$ given by $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|$ is continuous. But then f attains its max and min and we have

$$\|\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{x}^*\| \text{ for some } \boldsymbol{x}^* \in \mathcal{S}. \tag{7.7}$$

[9] In the case of one vector norm $\|\ \|$ on $\mathbb{C}^m$ and another vector norm $\|\ \|_\beta$ on $\mathbb{C}^n$ we would define $\|\boldsymbol{A}\| := \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|_\beta}$.

Lemma 7.11 (The operator norm is a matrix norm)
For any vector norm the operator norm given by (7.3) *is a consistent matrix norm. Moreover,* $\|\boldsymbol{I}\| = 1$.

Proof. We use (7.4). In 2. and 3. below we take the max over the unit sphere $\mathcal{S}$ given by (7.5).

1. Nonnegativity is obvious. If $\|\boldsymbol{A}\| = 0$ then $\|\boldsymbol{A}\boldsymbol{y}\| = 0$ 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. $\|c\boldsymbol{A}\| = \max_{\boldsymbol{x}} \|c\boldsymbol{A}\boldsymbol{x}\| = \max_{\boldsymbol{x}} |c| \, \|\boldsymbol{A}\boldsymbol{x}\| = |c| \, \|\boldsymbol{A}\|$.
3. $\|\boldsymbol{A} + \boldsymbol{B}\| = \max_{\boldsymbol{x}} \|(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x}\| \leq \max_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x}\| + \max_{\boldsymbol{x}} \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$.
4. $\|\boldsymbol{A}\boldsymbol{B}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{B}\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{B}\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{B}\boldsymbol{x}\|} \frac{\|\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$
$\leq \max_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{y}\|}{\|\boldsymbol{y}\|} \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{A}\| \, \|\boldsymbol{B}\|$.

That $\|\boldsymbol{I}\| = 1$ for any operator norm follows immediately from the definition. ∎

Since $\|\boldsymbol{I}\|_F = \sqrt{n}$, we see that the Frobenius norm is not an operator norm for $n > 1$.

7.1.3 The operator p-norms

Recall that the p or ℓ_p vector norms (10) 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 $1 \leq p \leq \infty$

$$\|\boldsymbol{A}\|_p := \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} = \max_{\|\boldsymbol{y}\|_p=1} \|\boldsymbol{A}\boldsymbol{y}\|_p. \tag{7.8}$$

For $p = 1, 2, \infty$ we have explicit expressions for these norms.

Theorem 7.12 (onetwoinfnorms)
For $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *we have*

$$\begin{aligned}
&\|\boldsymbol{A}\|_1 := \max_{1\leq j\leq n} \|\boldsymbol{A}\boldsymbol{e}_j\|_1 = \max_{1\leq j\leq n} \sum_{k=1}^{m} |a_{k,j}|, && \textit{(max column sum)} \\
&\|\boldsymbol{A}\|_2 := \sigma_1, && \textit{(largest singular value of } \boldsymbol{A}\textit{)} \\
&\|\boldsymbol{A}\|_\infty = \max_{1\leq k\leq m} \|\boldsymbol{e}_k^T\boldsymbol{A}\|_1 = \max_{1\leq k\leq m} \sum_{j=1}^{n} |a_{k,j}|. && \textit{(max row sum)}
\end{aligned} \tag{7.9}$$

The **two-norm** $\|\boldsymbol{A}\|_2$ is also called the **spectral norm** of $\boldsymbol{A}$.

Proof. The result for $p = 2$ follows from the minmax theorem for singular values. Indeed, by (6.16) we have $\sigma_1 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$. For $p = 1, \infty$ we proceed as follows:

(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 (7.8) that $\|\boldsymbol{A}\|_p = \|\boldsymbol{A}\boldsymbol{y}^*\|_p = K_p$.

1-norm: Define K_1, c and $\boldsymbol{y}^*$ by $K_1 := \|\boldsymbol{A}\boldsymbol{e}_c\|_1 = \max_{1\leq j\leq n}\|\boldsymbol{A}\boldsymbol{e}_j\|_1$ and $\boldsymbol{y}^* := \boldsymbol{e}_c$, a unit vector. Then $\|\boldsymbol{y}^*\|_1 = 1$ and we obtain

(a)

$$\|\boldsymbol{A}\boldsymbol{x}\|_1 = \sum_{k=1}^{m} \Big| \sum_{j=1}^{n} a_{kj}x_j \Big| \leq \sum_{k=1}^{m}\sum_{j=1}^{n} |a_{kj}||x_j| = \sum_{j=1}^{n} \Big(\sum_{k=1}^{m} |a_{kj}| \Big) |x_j| \leq K_1.$$

(b) $\|\boldsymbol{A}\boldsymbol{y}^*\|_1 = K_1$.

∞-norm: Define K_∞, r and $\boldsymbol{y}^*$ by $K_\infty := \|\boldsymbol{e}_r^T\boldsymbol{A}\|_1 = \max_{1\leq k\leq m}\|\boldsymbol{e}_k^T\boldsymbol{A}\|_1$ and $\boldsymbol{y}^* := [e^{-i\theta_1}, \ldots, e^{-i\theta_n}]^T$, where $a_{rj} = |a_{rj}|e^{i\theta_j}$ for $j = 1, \ldots, n$.

(a) $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{1\leq k\leq m} \Big| \sum_{j=1}^{n} a_{kj}x_j \Big| \leq \max_{1\leq k\leq m} \sum_{j=1}^{n} |a_{kj}||x_j| \leq K_\infty$.

(b) $\|\boldsymbol{A}\boldsymbol{y}^*\|_\infty = \max_{1\leq k\leq m} \Big| \sum_{j=1}^{n} a_{kj}e^{-i\theta_j} \Big| = K_\infty$.

The last equality is correct because $\Big| \sum_{j=1}^{n} a_{kj}e^{-i\theta_j} \Big| \leq \sum_{j=1}^{n} |a_{kj}| \leq K_\infty$ with equality for $k = r$.

□

Example 7.13 (Compare onetwoinfnorms)
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* $\sigma_1 = 2$ *(cf. Example 6.12). 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 7.14 (Spectral norm)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *has singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ *and eigenvalues* $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. *Then*

$$\|\boldsymbol{A}\|_2 = \sigma_1 \text{ and } \|\boldsymbol{A}^{-1}\|_2 = \frac{1}{\sigma_n}, \tag{7.10}$$

$$\|\boldsymbol{A}\|_2 = \lambda_1 \text{ and } \|\boldsymbol{A}^{-1}\|_2 = \frac{1}{\lambda_n}, \quad \text{if } \boldsymbol{A} \text{ is symmetric positive definite}, \tag{7.11}$$

$$\|\boldsymbol{A}\|_2 = |\lambda_1| \text{ and } \|\boldsymbol{A}^{-1}\|_2 = \frac{1}{|\lambda_n|}, \text{ if } \boldsymbol{A} \text{ is normal.} \tag{7.12}$$

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 $\boldsymbol{A}^{-1}$, (7.10) follows. By Theorem 6.9 the singular values of a symmetric positive definite matrix (normal matrix) are equal to the eigenvalues (absolute value of the eigenvalues). This implies (7.11) and (7.12). □

The following result is sometimes useful.

Theorem 7.15 (Spectral norm bound)
For any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *we have* $\|\boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty$.

Proof. Let $(\sigma^2, \boldsymbol{v})$ be an eigenpair for $\boldsymbol{A}^*\boldsymbol{A}$ corresponding to the largest singular value σ of $\boldsymbol{A}$. Then

$$\|\boldsymbol{A}\|_2^2 \|\boldsymbol{v}\|_1 = \sigma^2 \|\boldsymbol{v}\|_1 = \|\sigma^2 \boldsymbol{v}\|_1 = \|\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v}\|_1 \leq \|\boldsymbol{A}^*\|_1 \|\boldsymbol{A}\|_1 \|\boldsymbol{v}\|_1.$$

Observing that $\|\boldsymbol{A}^*\|_1 = \|\boldsymbol{A}\|_\infty$ by Theorem 7.12 and canceling $\|\boldsymbol{v}\|_1$ proves the result. □

Exercise 7.16 (Spectral norm)
Let $m, n \in \mathbb{N}$ *and* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. *Show that*

$$\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1} |\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{x}|.$$

Exercise 7.17 (Spectral norm of the inverse)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *is nonsingular. Use* (7.10) *and* (6.16) *to show that*

$$\|\boldsymbol{A}^{-1}\|_2 = \max_{\boldsymbol{x}\neq \boldsymbol{0}} \frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{A}\boldsymbol{x}\|_2}.$$

Exercise 7.18 (p-norm example)
Let

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Compute $\|\boldsymbol{A}\|_p$ *and* $\|\boldsymbol{A}^{-1}\|_p$ *for* $p = 1, 2, \infty$.

7.1.4 Unitary invariant matrix norms

Definition 7.19 (Unitary invariant norm)
A matrix norm $\|\ \|$ *on* $\mathbb{C}^{m\times n}$ *is called* **unitary invariant** *if* $\|\boldsymbol{UAV}\| = \|\boldsymbol{A}\|$ *for any* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *and any unitary matrices* $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ *and* $\boldsymbol{V} \in \mathbb{C}^{n\times n}$.

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 $\boldsymbol{U}(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{V} = \boldsymbol{UAV} + \boldsymbol{F}$, where $\|\boldsymbol{F}\| = \|\boldsymbol{E}\|$.

It follows from Lemma 6.24 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 [13] p. 308.

Theorem 7.20 (Unitary invariant norms)
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 6.24. Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ and let $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{C}^{n\times n}$ be unitary. Since the 2-vector norm is unitary invariant we obtain

$$\|\boldsymbol{UA}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{UAx}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{Ax}\|_2 = \|\boldsymbol{A}\|_2.$$

Now $\boldsymbol{A}$ and $\boldsymbol{A}^*$ have the same nonzero singular values, and it follows from Theorem 7.12 that $\|\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}\|_2$. Moreover $\boldsymbol{V}^*$ is unitary. Using these facts we find

$$\|\boldsymbol{AV}\|_2 = \|(\boldsymbol{AV})^*\|_2 = \|\boldsymbol{V}^*\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}^*\|_2 = \|\boldsymbol{A}\|_2.$$

□

Exercise 7.21 (Unitary invariance of the spectral norm)
Show that $\|\boldsymbol{VA}\|_2 = \|\boldsymbol{A}\|_2$ *holds even for a rectangular* $\boldsymbol{V}$ *as long as* $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$.

Exercise 7.22 ($\|\boldsymbol{AU}\|_2$ rectangular $\boldsymbol{A}$)
Find $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ *and* $\boldsymbol{U} \in \mathbb{R}^{2\times 1}$ *with* $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ *such that* $\|\boldsymbol{AU}\|_2 < \|\boldsymbol{A}\|_2$. *Thus, in general,* $\|\boldsymbol{AU}\|_2 = \|\boldsymbol{A}\|_2$ *does not hold for a rectangular* $\boldsymbol{U}$ *even if* $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$.

Exercise 7.23 (p-norm of diagonal matrix)
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 7.24 (spectral norm of a column vector)
A vector $\boldsymbol{a} \in \mathbb{C}^m$ *can also be considered as a matrix* $\boldsymbol{A} \in \mathbb{C}^{m,1}$*.*

(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* $1 \le p \le \infty$*.*

7.1.5 Absolute and monotone norms

A vector norm on $\mathbb{C}^n$ is 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 [13]) 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 7.25 (Norm of absolute value matrix)
If $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *has elements* a_{ij}*, let* $|\boldsymbol{A}| \in \mathbb{R}^{m\times n}$ *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* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ $\|\boldsymbol{A}\|_F = \|\,|\boldsymbol{A}|\,\|_F$, $\|\boldsymbol{A}\|_p = \|\,|\boldsymbol{A}|\,\|_p$ *for* $p = 1, \infty$*.*

(c) *Show that for any* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ $\|\boldsymbol{A}\|_2 \le \|\,|\boldsymbol{A}|\,\|_2$*.*

(d) *Find a real symmetric* 2×2 *matrix* $\boldsymbol{A}$ *such that* $\|\boldsymbol{A}\|_2 < \|\,|\boldsymbol{A}|\,\|_2$*.*

The study of matrix norms will be continued in Chapter 8.

7.2 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 can be 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 matrix norm on $\mathbb{C}^{n\times n}$ which for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ and any $\boldsymbol{x} \in \mathbb{C}^n$ satisfy

$$\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\| \text{ and } \|\boldsymbol{Ax}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|.$$

This holds 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 6.24.

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 7.26 (Perturbation in the right-hand side)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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

$$\frac{1}{K(\boldsymbol{A})}\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{b}\|} \le \frac{\|\boldsymbol{y} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le K(\boldsymbol{A})\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{b}\|}, \quad K(\boldsymbol{A}) = \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|. \tag{7.13}$$

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 (7.13). 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 (7.13). $\|\boldsymbol{e}\|/\|\boldsymbol{b}\|$ 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 $K(\boldsymbol{A}) \geq 1$. For since $\|\boldsymbol{x}\| = \|\boldsymbol{I}\boldsymbol{x}\| \leq \|\boldsymbol{I}\|\|\boldsymbol{x}\|$ for any $\boldsymbol{x}$ we have $\|\boldsymbol{I}\| \geq 1$ and therefore $\|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\| \geq \|\boldsymbol{A}\boldsymbol{A}^{-1}\| = \|\boldsymbol{I}\| \geq 1$.

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.

The following explicit expressions for the 2-norm condition number follow from Theorem 7.14.

Theorem 7.27 (Spectral condition number)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is nonsingular with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ and eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| > 0$. Then $K_2(\boldsymbol{A}) := \|\boldsymbol{A}\|_2\|\boldsymbol{A}^{-1}\|_2 = \sigma_1/\sigma_n$. Moreover,

$$K_2(\boldsymbol{A}) = \begin{cases} \lambda_1/\lambda_n, & \text{if } \boldsymbol{A} \text{ is symmetric positive definite,} \\ |\lambda_1|/|\lambda_n|, & \text{if } \boldsymbol{A} \text{ is normal.} \end{cases} \tag{7.14}$$

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.

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 7.28 (Perturbation and residual)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{b} \in \mathbb{C}^n$, $\boldsymbol{A}$ is nonsingular and $\boldsymbol{b} \neq \boldsymbol{0}$. Let $\boldsymbol{r}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}$ for any $\boldsymbol{y} \in \mathbb{C}^n$. If $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ then

$$\frac{1}{K(\boldsymbol{A})}\frac{\|\boldsymbol{r}(\boldsymbol{y})\|}{\|\boldsymbol{b}\|} \leq \frac{\|\boldsymbol{y} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq K(\boldsymbol{A})\frac{\|\boldsymbol{r}(\boldsymbol{y})\|}{\|\boldsymbol{b}\|}. \tag{7.15}$$

Proof. We simply take $\boldsymbol{e} = \boldsymbol{r}(\boldsymbol{y})$ in Theorem 7.26. □

If $\boldsymbol{A}$ is well-conditioned, (7.15) 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.

Consider next the effect of a perturbation in the coefficient matrix. Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n\times n}$ with $\boldsymbol{A}$ nonsingular. We like to compare the solution $\boldsymbol{x}$ and $\boldsymbol{y}$ of the systems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $(\boldsymbol{A} + \boldsymbol{E})\boldsymbol{y} = \boldsymbol{b}$. We expect $\boldsymbol{A} + \boldsymbol{E}$ to be nonsingular if the elements of $\boldsymbol{E}$ are sufficiently small and we need to address this question. Consider first the case where $\boldsymbol{A} = \boldsymbol{I}$.

Theorem 7.29 (Nonsingularity of perturbation of identity)
Suppose $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ and $\|\boldsymbol{B}\| < 1$ for some consistent matrix norm on $\mathbb{C}^{n\times n}$. Then $\boldsymbol{I} - \boldsymbol{B}$ is nonsingular and

$$\frac{1}{1+\|\boldsymbol{B}\|} \le \|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| \le \frac{1}{1-\|\boldsymbol{B}\|}. \tag{7.16}$$

Proof. Suppose $\boldsymbol{I} - \boldsymbol{B}$ is singular. Then $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{x} = \boldsymbol{0}$ for some nonzero $\boldsymbol{x} \in \mathbb{C}^n$, and $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x}$ so that $\|\boldsymbol{x}\| = \|\boldsymbol{B}\boldsymbol{x}\| \le \|\boldsymbol{B}\|\|\boldsymbol{x}\|$. But then $\|\boldsymbol{B}\| \ge 1$. It follows that $\boldsymbol{I} - \boldsymbol{B}$ is nonsingular if $\|\boldsymbol{B}\| < 1$. Next, since

$$\begin{aligned}\|\boldsymbol{I}\| &= \|(\boldsymbol{I}-\boldsymbol{B})(\boldsymbol{I}-\boldsymbol{B})^{-1}\| \le \|\boldsymbol{I}-\boldsymbol{B}\|\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| \\ &\le \big(\|\boldsymbol{I}\| + \|\boldsymbol{B}\|\big)\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\|,\end{aligned}$$

and since $\|\boldsymbol{I}\| \ge 1$, we obtain the lower bound in (7.16):

$$\frac{1}{1+\|\boldsymbol{B}\|} \le \frac{\|\boldsymbol{I}\|}{\|\boldsymbol{I}\|+\|\boldsymbol{B}\|} \le \|(\boldsymbol{I}-\boldsymbol{B})^{-1}\|. \tag{7.17}$$

Taking norms and using the inverse triangle inequality in

$$\boldsymbol{I} = (\boldsymbol{I}-\boldsymbol{B})(\boldsymbol{I}-\boldsymbol{B})^{-1} = (\boldsymbol{I}-\boldsymbol{B})^{-1} - \boldsymbol{B}(\boldsymbol{I}-\boldsymbol{B})^{-1}$$

implies

$$\|\boldsymbol{I}\| \ge \|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| - \|\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{B})^{-1}\| \ge \big(1-\|\boldsymbol{B}\|\big)\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\|.$$

If the matrix norm is an operator norm then $\|\boldsymbol{I}\| = 1$ and the upper bound follows. We show in Section 8.4 that the upper bound also holds for the Frobenius norm, and more generally for any consistent matrix norm on $\mathbb{C}^{n\times n}$. ∎

Theorem 7.30 (Nonsingularity of perturbation)
Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n\times n}$, $\boldsymbol{b} \in \mathbb{C}^n$ with $\boldsymbol{A}$ invertible and $\boldsymbol{b} \ne \boldsymbol{0}$. If $r := \|\boldsymbol{A}^{-1}\boldsymbol{E}\| < 1$ for some matrix norm consistent on $\mathbb{C}^{n\times n}$ then $\boldsymbol{A}+\boldsymbol{E}$ is nonsingular. If $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{y} = \boldsymbol{b}$ then

$$\frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le \|\boldsymbol{A}^{-1}\boldsymbol{E}\| \le K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}, \tag{7.18}$$

$$\frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le 2K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}. \tag{7.19}$$

In (7.19) we have assumed that $r \le 1/2$.

Proof. Since $r < 1$ Theorem 7.29 implies that the matrix $\boldsymbol{I} - \boldsymbol{B} := \boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{E}$ is nonsingular and then $\boldsymbol{A} + \boldsymbol{E} = \boldsymbol{A}\big(\boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{E}\big)$ is nonsingular. Subtracting $(\boldsymbol{A} + \boldsymbol{E})\boldsymbol{y} = \boldsymbol{b}$ from $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ gives $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{E}\boldsymbol{y}$ or $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{y}$. Taking norms and dividing by $\|\boldsymbol{y}\|$ proves (7.18). Solving $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{y}$ for $\boldsymbol{y}$ we obtain $\boldsymbol{y} = (\boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{E})^{-1}\boldsymbol{x}$. By (7.16)

$$\|\boldsymbol{y}\| \le \|(\boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{E})^{-1}\|\|\boldsymbol{x}\| \le \frac{\|\boldsymbol{x}\|}{1 - \|\boldsymbol{A}^{-1}\boldsymbol{E}\|} \le 2\|\boldsymbol{x}\|.$$

But then (7.19) follows from (7.18). □

In Theorem 7.30 we gave bounds for the relative error in $\boldsymbol{x}$ as an approximation to $\boldsymbol{y}$ and the relative error in $\boldsymbol{y}$ as an approximation to $\boldsymbol{x}$. $\|\boldsymbol{E}\|/\|\boldsymbol{A}\|$ 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 $\|\boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{y}\| \le \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{E}\|\,\|\boldsymbol{y}\|$. 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 (7.18) can give a fairly realistic estimate for $\|\boldsymbol{y} - \boldsymbol{x}\|/\|\boldsymbol{y}\|$.

We end this section with a perturbation result for the inverse matrix. Again the condition number plays an important role.

Theorem 7.31 (Perturbation of inverse matrix)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is nonsingular and let $\|\cdot\|$ be a consistent matrix norm on $\mathbb{C}^{n\times n}$. If $\boldsymbol{E} \in \mathbb{C}^{n\times n}$ is so small that $r := \|\boldsymbol{A}^{-1}\boldsymbol{E}\| < 1$ then $\boldsymbol{A} + \boldsymbol{E}$ is nonsingular and

$$\|(\boldsymbol{A} + \boldsymbol{E})^{-1}\| \le \frac{\|\boldsymbol{A}^{-1}\|}{1 - r}. \tag{7.20}$$

If $r < 1/2$ then

$$\frac{\|(\boldsymbol{A} + \boldsymbol{E})^{-1} - \boldsymbol{A}^{-1}\|}{\|\boldsymbol{A}^{-1}\|} \le 2K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}. \tag{7.21}$$

Proof. We showed in Theorem 7.30 that $\boldsymbol{A} + \boldsymbol{E}$ is nonsingular and since $(\boldsymbol{A} + \boldsymbol{E})^{-1} = (\boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{E})^{-1}\boldsymbol{A}^{-1}$ we obtain

$$\|(\boldsymbol{A} + \boldsymbol{E})^{-1}\| \le \|(\boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{E})^{-1}\|\|\boldsymbol{A}^{-1}\| \le \frac{\|\boldsymbol{A}^{-1}\|}{1 - \|\boldsymbol{A}^{-1}\boldsymbol{E}\|}$$

and (7.20) follows. Since

$$(\boldsymbol{A} + \boldsymbol{E})^{-1} - \boldsymbol{A}^{-1} = \big(\boldsymbol{I} - \boldsymbol{A}^{-1}(\boldsymbol{A} + \boldsymbol{E})\big)(\boldsymbol{A} + \boldsymbol{E})^{-1} = -\boldsymbol{A}^{-1}\boldsymbol{E}(\boldsymbol{A} + \boldsymbol{E})^{-1}$$

we obtain by (7.20)

$$\|(\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1}\| \le \|\boldsymbol{A}^{-1}\|\|\boldsymbol{E}\|\|(\boldsymbol{A}+\boldsymbol{E})^{-1}\| \le K(\boldsymbol{A})\frac{\|\boldsymbol{E}\|}{\|\boldsymbol{A}\|}\frac{\|\boldsymbol{A}^{-1}\|}{1-r}.$$

Dividing by $\|\boldsymbol{A}^{-1}\|$ and setting $r = 1/2$ proves (7.21). ∎

Exercise 7.32 (Sharpness of perturbation bounds)
The upper and lower bounds for $\|\boldsymbol{y}-\boldsymbol{x}\|/\|\boldsymbol{x}\|$ given by (7.13) 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 $\|\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{y}_{\boldsymbol{A}}\|$ and $\|\boldsymbol{A}^{-1}\| = \|\boldsymbol{A}^{-1}\boldsymbol{y}_{\boldsymbol{A}^{-1}}\|$.

(a) *Show that the upper bound in (7.13) is attained if $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{y}_{\boldsymbol{A}}$ 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 $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{y}_{\boldsymbol{A}}$.*

Exercise 7.33 (Condition number of 2. derivative matrix)
In this exercise we will show that for $m \ge 1$

$$\frac{4}{\pi^2}(m+1)^2 - 2/3 < cond_p(\boldsymbol{T}) \le \frac{1}{2}(m+1)^2, \quad p = 1, 2, \infty, \tag{7.22}$$

where $\boldsymbol{T} := tridiag(-1, 2, -1) \in \mathbb{R}^{m\times m}$ and $cond_p(\boldsymbol{T}) := \|\boldsymbol{T}\|_p\|\boldsymbol{T}^{-1}\|_p$ is the p-norm condition number of $\boldsymbol{T}$. The p matrix norm is given by (7.8). You will need the explicit inverse of $\boldsymbol{T}$ given by (1.10) and the eigenvalues given in Lemma 3.8. As usual we define $h := 1/(m+1)$.

a) *Show that for $m \ge 3$*

$$cond_1(\boldsymbol{T}) = cond_\infty(\boldsymbol{T}) = \frac{1}{2}\begin{cases} h^{-2}, & m \text{ odd}, \\ h^{-2}-1, & m \text{ even}. \end{cases} \tag{7.23}$$

and that $cond_1(\boldsymbol{T}) = cond_\infty(\boldsymbol{T}) = 3$ for $m = 2$.

b) *Show that for $p = 2$ and $m \ge 1$ 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*

$$\frac{4}{\pi^2}h^{-2} - \frac{2}{3} < cond_2(\boldsymbol{T}) < \frac{4}{\pi^2}h^{-2}. \tag{7.24}$$

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.

d) *Show (7.22).*

7.3 Proof that the p-Norms are Norms

We want to show

Theorem 7.34 (The p vector norms are norms)
Let for $1 \le p \le \infty$ and $\boldsymbol{x} \in \mathbb{C}^n$

$$\|\boldsymbol{x}\|_p := \big(\sum_{j=1}^{n} |x_j|^p\big)^{1/p}, \quad \|\boldsymbol{x}\|_\infty := \max_{1\le j\le n} |x_j|.$$

Then for all $1 \le p \le \infty$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ and all $a \in \mathbb{C}$

1. $\|\boldsymbol{x}\|_p \ge 0$ *with equality if and only if* $\boldsymbol{x} = \boldsymbol{0}$. *(positivity)*
2. $\|a\boldsymbol{x}\|_p = |a|\,\|\boldsymbol{x}\|_p$. *(homogeneity)*
3. $\|\boldsymbol{x} + \boldsymbol{y}\|_p \le \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$. *(subadditivity)*

Positivity and homogeneity follows immediately. To show the subadditivity we need some elementary properties of convex functions.

Definition 7.35 (Convex function)
Let $I \subset \mathbb{R}$ be an interval. A function $f : I \to \mathbb{R}$ is convex if

$$f\big((1-\lambda)x_1 + \lambda x_2\big) \le (1-\lambda)f(x_1) + \lambda f(x_2) \tag{7.25}$$

for all $x_1, x_2 \in I$ with $x_1 < x_2$ and all $\lambda \in [0,1]$. The sum $\sum_{j=1}^n \lambda_j x_j$ is called a **convex combination** *of $x_1, \ldots, x_n$ if $\lambda_j \ge 0$ for $j = 1, \ldots, n$ and $\sum_{j=1}^n \lambda_j = 1$.*

The convexity condition is illustrated in Figure 7.1.

Lemma 7.36 (A sufficient condition for convexity)
If $f \in C^2[a,b]$ and $f''(x) \ge 0$ for $x \in [a,b]$ then f is convex.

Proof. We recall the formula for linear interpolation with remainder, (cf a book on numerical methods) For any $a \le x_1 \le x \le x_2 \le b$ there is a $c \in [x_1, x_2]$ such that

$$\begin{aligned} f(x) &= \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2) + (x - x_1)(x - x_2) f''(c)/2 \\ &= (1-\lambda)f(x_1) + \lambda f(x_2) + (x_2 - x_1)^2 \lambda(\lambda - 1) f''(c)/2, \quad \lambda := \frac{x - x_1}{x_2 - x_1}. \end{aligned}$$

Since $\lambda \in [0,1]$ the remainder term is not positive. Moreover,

$$x = \frac{x_2 - x}{x_2 - x_1} x_1 + \frac{x - x_1}{x_2 - x_1} x_2 = (1-\lambda)x_1 + \lambda x_2$$

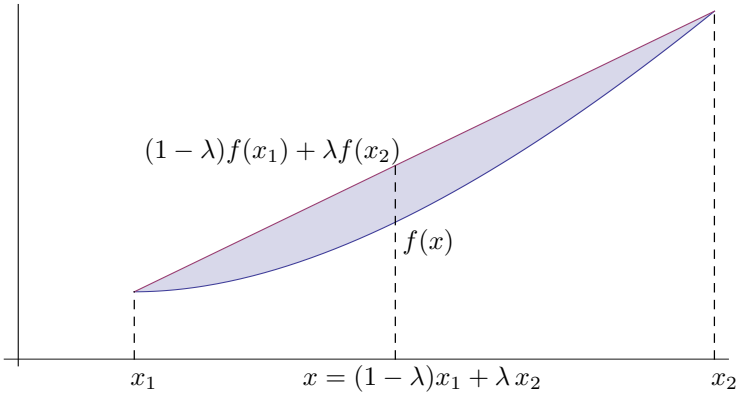


Figure 7.1. *A convex function.*

so that (7.25) holds, and f is convex. ∎

The following inequality is elementary, but can be used to prove many non-trivial inequalities.

Theorem 7.37 (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 $z_1, \ldots, z_n \in I$

$$f(\sum_{j=1}^n \lambda_j z_j) \leq \sum_{j=1}^n \lambda_j f(z_j).$$

Proof. We use induction on n. The result is trivial for $n = 1$. Let $n \geq 2$, assume the inequality holds for $n-1$, 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 least one i so assume without loss of generality that $\lambda_1 < 1$, and define $u := \sum_{j=2}^n \frac{\lambda_j}{1-\lambda_1} z_j$. Since $\sum_{j=2}^n \lambda_j = 1 - \lambda_1$ this is a convex combination of $n-1$ 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

$$f(\sum_{j=1}^n \lambda_j z_j) = f(\lambda_1 z_1 + (1-\lambda_1)u) \leq \lambda_1 f(z_1) + (1-\lambda_1) f(u) \leq \sum_{j=1}^n \lambda_j f(z_j)$$

and the inequality holds for n. ∎

Corollary 7.38 (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*

$$a_1^{\lambda_1} a_2^{\lambda_2} \cdots a_n^{\lambda_n} \leq \sum_{j=1}^n \lambda_j a_j, \tag{7.26}$$

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. Consider the function $f : (0, \infty)$ given by $f(x) = -\log x$. Since $f''(x) = 1/x^2 > 0$ for $x \in (0, \infty)$, this function is convex. By Jensen's inequality

$$-\log\big(\sum_{j=1}^n \lambda_j a_j\big) \leq -\sum_{j=1}^n \lambda_j \log(a_j) = -\log\big(a_1^{\lambda_1} \cdots a_n^{\lambda_n}\big)$$

or $\log\big(a_1^{\lambda_1} \cdots a_n^{\lambda_n}\big) \leq \log\big(\sum_{j=1}^n \lambda_j a_j\big)$. The inequality follows since $\exp(\log x) = x$ for $x > 0$ and the exponential function is monotone increasing. ∎

Taking $\lambda_j = \frac{1}{n}$ for all j in (7.26) we obtain the classical **geometric/arithmetic mean inequality**

$$(a_1 a_2 \cdots a_n)^{\frac{1}{n}} \leq \frac{1}{n} \sum_{j=1}^n a_j. \tag{7.27}$$

Corollary 7.39 (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

$$a^{\frac{1}{p}} b^{\frac{1}{q}} \leq \frac{1}{p} a + \frac{1}{q} b, \text{ where } \frac{1}{p} + \frac{1}{q} = 1. \tag{7.28}$$

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 7.28 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 7.40 (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

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p^p \leq \|\boldsymbol{x}\|_p \|\boldsymbol{x} + \boldsymbol{y}\|_p^{p/q} + \|\boldsymbol{y}\|_p \|\boldsymbol{x} + \boldsymbol{y}\|_p^{p/q} = (\|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p)\|\boldsymbol{x} + \boldsymbol{y}\|_p^{p-1},$$

and canceling the common factor, the inequality follows. ∎

It is possible to characterize the p-noms that are derived from an inner product. We start with the following identity.

Theorem 7.41 (Parallelogram identity)
For all $\boldsymbol{x}, \boldsymbol{y}$ in a real or complex inner product space

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 + \|\boldsymbol{x} - \boldsymbol{y}\|^2 = 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2. \tag{7.29}$$

Proof. We set $a = \pm 1$ in (22) and add the two equations. ∎

Theorem 7.42 (When is a norm an inner product norm?)
To a given norm on a real or complex vector space $\mathcal{V}$ there exists an inner product on $\mathcal{V}$ such that $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|^2$ if and only if the parallelogram identity (7.29) *holds for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$.*

Proof. If $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|^2$ then Theorem 7.41 shows that the parallelogram identity holds. For the converse we prove the real case and leave the complex case as an exercise. Suppose (7.29) holds for all $\boldsymbol{x}, \boldsymbol{y}$ in the real vector space $\mathcal{V}$. We show that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \frac{1}{4}\big(\|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2\big), \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V} \tag{7.30}$$

defines an inner product on $\mathcal{V}$. Clearly 1. and 2. in Definition 0.20 hold. The hard part is to show 3. We need to show that

$$\langle \boldsymbol{x}, \boldsymbol{z}\rangle + \langle \boldsymbol{y}, \boldsymbol{z}\rangle = \langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{z}\rangle, \quad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}, \tag{7.31}$$

$$\langle a\boldsymbol{x}, \boldsymbol{y}\rangle = a\langle \boldsymbol{x}, \boldsymbol{y}\rangle, \quad a \in \mathbb{R}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}. \tag{7.32}$$

Now

$$\begin{aligned}
4\langle \boldsymbol{x}, \boldsymbol{z}\rangle + 4\langle \boldsymbol{y}, \boldsymbol{z}\rangle &\overset{(7.30)}{=} \|\boldsymbol{x}+\boldsymbol{z}\|^2 - \|\boldsymbol{x}-\boldsymbol{z}\|^2 + \|\boldsymbol{y}+\boldsymbol{z}\|^2 - \|\boldsymbol{y}-\boldsymbol{z}\|^2 \\
&= \|\big(\boldsymbol{z} + \frac{\boldsymbol{x}+\boldsymbol{y}}{2}\big) + \frac{\boldsymbol{x}-\boldsymbol{y}}{2}\|^2 - \|\big(\boldsymbol{z} - \frac{\boldsymbol{x}+\boldsymbol{y}}{2}\big) + \frac{\boldsymbol{y}-\boldsymbol{x}}{2}\|^2 \\
&+ \|\big(\boldsymbol{z} + \frac{\boldsymbol{x}+\boldsymbol{y}}{2}\big) - \frac{\boldsymbol{x}-\boldsymbol{y}}{2}\|^2 - \|\big(\boldsymbol{z} - \frac{\boldsymbol{x}+\boldsymbol{y}}{2}\big) - \frac{\boldsymbol{y}-\boldsymbol{x}}{2}\|^2 \\
&\overset{(7.29)}{=} 2\|\boldsymbol{z} + \frac{\boldsymbol{x}+\boldsymbol{y}}{2}\|^2 + 2\|\frac{\boldsymbol{x}-\boldsymbol{y}}{2}\|^2 - 2\|\boldsymbol{z} - \frac{\boldsymbol{x}+\boldsymbol{y}}{2}\|^2 - 2\|\frac{\boldsymbol{y}-\boldsymbol{x}}{2}\|^2 \\
&\overset{(7.30)}{=} 8\langle \frac{\boldsymbol{x}+\boldsymbol{y}}{2}, \boldsymbol{z}\rangle,
\end{aligned}$$

or

$$\langle \boldsymbol{x}, \boldsymbol{z}\rangle + \langle \boldsymbol{y}, \boldsymbol{z}\rangle = 2\langle \frac{\boldsymbol{x}+\boldsymbol{y}}{2}, \boldsymbol{z}\rangle, \quad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}.$$

In particular, since $\boldsymbol{y} = \boldsymbol{0}$ implies $\langle \boldsymbol{y}, \boldsymbol{z}\rangle = 0$ we obtain $\langle \boldsymbol{x}, \boldsymbol{z}\rangle = 2\langle \frac{\boldsymbol{x}}{2}, \boldsymbol{z}\rangle$ for all $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{V}$. This means that $2\langle \frac{\boldsymbol{x}+\boldsymbol{y}}{2}, \boldsymbol{z}\rangle = \langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{z}\rangle$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$ and (7.31) follows.

We first show (7.32) when $a = n$ is a positive integer. By induction

$$\langle n\boldsymbol{x}, \boldsymbol{y}\rangle = \langle (n-1)\boldsymbol{x} + \boldsymbol{x}, \boldsymbol{y}\rangle \overset{(7.31)}{=} \langle (n-1)\boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{x}, \boldsymbol{y}\rangle = n\langle \boldsymbol{x}, \boldsymbol{y}\rangle. \tag{7.33}$$

If $m, n \in \mathbb{N}$ then

$$m^2\langle \frac{n}{m}\boldsymbol{x}, \boldsymbol{y}\rangle \overset{(7.33)}{=} m\langle n\boldsymbol{x}, \boldsymbol{y}\rangle \overset{(7.33)}{=} mn\langle \boldsymbol{x}, \boldsymbol{y}\rangle,$$

implying that (7.32) holds for positive rational numbers

$$\langle \frac{n}{m}\boldsymbol{x}, \boldsymbol{y}\rangle = \frac{n}{m}\langle \boldsymbol{x}, \boldsymbol{y}\rangle.$$

Nov if $a > 0$ there is a sequence $\{a_n\}$ of positive rational numbers converging to a. For each n

$$a_n\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle a_n\boldsymbol{x}, \boldsymbol{y}\rangle \overset{(7.30)}{=} \frac{1}{4}\big(\|a_n\boldsymbol{x} + \boldsymbol{y}\|^2 - \|a_n\boldsymbol{x} - \boldsymbol{y}\|^2\big).$$

Taking limits and using continuity of norms we obtain $a\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle a\boldsymbol{x}, \boldsymbol{y}\rangle$. This also holds for $a = 0$. Finally, if $a < 0$ then $(-a) > 0$ and from what we just showed

$$(-a)\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle (-a)\boldsymbol{x}, \boldsymbol{y}\rangle \overset{(7.30)}{=} \frac{1}{4}\big(\|-a\boldsymbol{x} + \boldsymbol{y}\|^2 - \|-a\boldsymbol{x} - \boldsymbol{y}\|^2\big) = -\langle a\boldsymbol{x}, \boldsymbol{y}\rangle,$$

so (7.32) also holds for negative a. ∎

Corollary 7.43 (Are the p-norms inner product norms?)
For the p vector norms on $\mathcal{V} = \mathbb{R}^n$ or $\mathcal{V} = \mathbb{C}^n$, $1 \le p \le \infty$, $n \ge 2$, there is an inner product on $\mathcal{V}$ such that $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|_p^2$ for all $\boldsymbol{x} \in \mathcal{V}$ if and only if $p = 2$.

Proof. For $p = 2$ the p-norm is the Euclidian norm which corresponds to the standard inner product. If $p \neq 2$ then the parallelogram identity (7.29) does not hold for say $\boldsymbol{x} := \boldsymbol{e}_1$ and $\boldsymbol{y} := \boldsymbol{e}_2$. ∎

Exercise 7.44 (When is a complex norm an inner product norm?)
Given a vector norm in a complex vector space $\mathcal{V}$, and suppose (7.29) *holds for all $\boldsymbol{x}, \boldsymbol{y}$ Show that*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \frac{1}{4}\big(\|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2 + i\|\boldsymbol{x} + i\boldsymbol{y}\|^2 - i\|\boldsymbol{x} - i\boldsymbol{y}\|^2\big), \tag{7.34}$$

defines an inner product on $\mathcal{V}$, where $i = \sqrt{-1}$. The identity (7.34) *is called the* **polarization identity**.[10]

Exercise 7.45 (p norm for $p = 1$ and $p = \infty$)
Show that $\|\cdot\|p$ is a vector norm in $\mathbb{R}^n$ for $p = 1$, $p = \infty$.

Exercise 7.46 (The p- norm unit sphere)
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 7.47 (Sharpness of p-norm inequalitiy)
For $p \ge 1$, and any $\boldsymbol{x} \in \mathbb{C}^n$ we have $\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_p \le n^{1/p}\|\boldsymbol{x}\|_\infty$ (cf. (14)*).*

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, these inequalities are sharp.

Exercise 7.48 (p-norm inequalitiies for arbitrary p)
If $1 \le q \le p \le \infty$ then

$$\|\boldsymbol{x}\|_p \le \|\boldsymbol{x}\|_q \le n^{1/q-1/p}\|\boldsymbol{x}\|_p, \quad \boldsymbol{x} \in \mathbb{C}^n.$$

Hint: For the rightmost inequality use Jensen's inequality Cf. Theorem 7.37 with $f(z) = z^{p/q}$ and $z_i = |x_i|^q$. For the left inequality consider first $y_i = x_i/\|\boldsymbol{x}\|_\infty$, $i = 1, 2, \ldots, n$.

[10] Hint: We have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = s(\boldsymbol{x}, \boldsymbol{y}) + is(\boldsymbol{x}, i\boldsymbol{y})$, where $s(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{4}\big(\|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2\big)$.

7.4 Review Questions

7.4.1]

- What is a consistent matrix norm?
- what is a subordinate matrix norm?
- is an operator norm consistent?
- why is the Frobenius norm not an operator norm?
- what is the spectral norm of a matrix?
- how do we compute $\|\boldsymbol{A}\|_\infty$?
- what is the spectral condition number of a symmetric positive definte matrix?

7.4.2 Why is $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$ for any matrix $\boldsymbol{A}$?

7.4.3 What is the spectral norm of the inverse of a normal matrix?

Part III

Iterative Methods for Large Linear Systems

Chapter 8

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.

In this chapter we consider the classical iterative methods of Richardson, 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 [35].

We give the main points of this theory specialized to the 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 (cf. Algorithm 4.1). 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 an acceleration parameter.

For convergence we study convergence of powers of matrices.

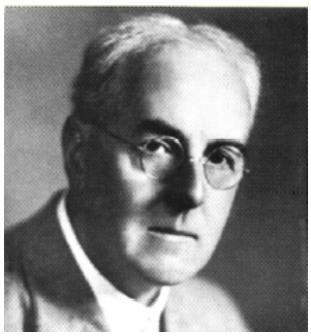
8.1 Classical Iterative Methods; Component Form

We start with an example showing how a linear system can be solved using an iterative method.

Example 8.1 (Iterative methods on a special 2×2 matrix)
Solving for the diagonal elements the linear system $\left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right]\left[\begin{smallmatrix} y \\ z \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ *can be written in component form as* $y = (z+1)/2$ *and* $z = (y+1)/2$. *Starting with* y_0, z_0 *we generate two sequences* $\{y_k\}$ *and* $\{z_k\}$ *using the difference equations* $y_{k+1} = (z_k + 1)/2$ *and* $z_{k+1} = (y_k + 1)/2$. *This is known as Jacobi's method. If* $y_0 = z_0 = 0$ *then we find* $y_1 = z_1 = 1/2$ *and in general* $y_k = z_k = 1 - 2^{-k}$ *for* $k = 0, 1, 2, 3, \ldots$. *The iteration converges to the exact solution* $[1, 1]^T$, *and the error is halved in each iteration.*

We can improve the convergence rate by using the most current approximation in each iteration. This leads to Gauss-Seidel's method: $y_{k+1} = (z_k + 1)/2$ *and* $z_{k+1} = (y_{k+1} + 1)/2$. *If* $y_0 = z_0 = 0$ *then we find* $y_1 = 1/2$, $z_1 = 3/4$, $y_2 = 7/8$, $z_2 = 15/16$, *and in general* $y_k = 1 - 2 \cdot 4^{-k}$ *and* $z_k = 1 - 4^{-k}$ *for* $k = 1, 2, 3, \ldots$. *The error is now reduced by a factor* 4 *in each iteration.*

Consider the general case. Suppose $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ is nonsingular and $\boldsymbol{b} \in \mathbb{C}^n$. Suppose we know an approximation $\boldsymbol{x}_k = [\boldsymbol{x}_k(1), \ldots, \boldsymbol{x}_k(n)]^T$ to the exact solution $\boldsymbol{x}$ of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.




Lewis Fry Richardson, 1881-1953 (left),Carl Gustav Jacob Jacobi, 1804-1851 (right).


Philipp Ludwig von Seidel, 1821-1896 (left), David M. Young Jr., 1923-2008 (right)

In **Richardson's method (R method)** we pick a positive parameter α and compute a new approximation by adding a multiple of the residual vector $\boldsymbol{r}_k := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$:

$$\boldsymbol{x}_{k+1}(i) = x_k(i) + \alpha \boldsymbol{r}_k(i),\ \boldsymbol{r}_k(i) := b_i - \sum_{j=1}^{n} a_{ij}\boldsymbol{x}_k(j),\ \text{for } i = 1, 2, \ldots, n. \quad (8.1)$$

Richardson considered the simplest case $\alpha = 1$. The parameter α is added to get faster convergence.

For the other methods we need to assume that $\boldsymbol{A}$ has nonzero diagonal elements. Solving the ith equation of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ for $\boldsymbol{x}(i)$, we obtain a **fixed-point form** of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

$$\boldsymbol{x}(i) = \big(-\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}(j) - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}(j) + b_i\big)/a_{ii}, \quad i = 1, 2, \ldots, n. \quad (8.2)$$

1. In **Jacobi's method (J method)** we substitute $\boldsymbol{x}_k$ into the right hand side of (8.2) and compute a new approximation by

$$\boldsymbol{x}_{k+1}(i) = \big(-\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_k(j) - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_k(j) + b_i\big)/a_{ii},\ \text{for } i = 1, 2, \ldots, n. \quad (8.3)$$

2. **Gauss-Seidel's method (GS method)** is a modification of Jacobi's method, where we use the new $\boldsymbol{x}_{k+1}(i)$ immediately after it has been computed.

$$\boldsymbol{x}_{k+1}(i) = \big(-\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_{k+1}(j) - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_k(j) + b_i\big)/a_{ii},\ \text{for } i = 1, 2, \ldots, n. \quad (8.4)$$

3. The **Successive overrelaxation method (SOR method)** is obtained by introducing an acceleration parameter $0 < \omega < 2$ in the GS method. We write $\boldsymbol{x}(i) = \omega\boldsymbol{x}(i) + (1-\omega)\boldsymbol{x}(i)$ and this leads to the method

$$\boldsymbol{x}_{k+1}(i) = \omega\big(-\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_{k+1}(j) - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_k(j) + b_i\big)/a_{ii} + (1-\omega)\boldsymbol{x}_k(i). \quad (8.5)$$

   The SOR method reduces to the Gauss-Seidel method for $\omega = 1$. Denoting the right hand side of (8.4) by $\boldsymbol{x}_{k+1}^{gs}$ we can write (8.5) as $\boldsymbol{x}_{k+1} = \omega\boldsymbol{x}_{k+1}^{gs} + (1-\omega)\boldsymbol{x}_k$, and we see that $\boldsymbol{x}_{k+1}$ is located on the straight line passing through the two points $\boldsymbol{x}_{k+1}^{gs}$ and $\boldsymbol{x}_k$. The restriction $0 < \omega < 2$ is necessary for convergence (cf. Theorem 8.14). Normally, the best results are obtained for the relaxation parameter ω in the range $1 \leq \omega < 2$ and then $\boldsymbol{x}_{k+1}$ is computed by linear extrapolation, i. e., it is not located between $\boldsymbol{x}_{k+1}^{gs}$ and $\boldsymbol{x}_k$.
4. We mention also briefly the symmetric successive overrelaxation method **SSOR**. One iteration in SSOR consists of two SOR sweeps. A forward SOR sweep (8.5), computing an approximation denoted $\boldsymbol{x}_{k+1/2}$ instead of $\boldsymbol{x}_{k+1}$, is followed by a back SOR sweep computing

$$\boldsymbol{x}_{k+1}(i) = \omega\big(-\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_{k+1/2}(j) - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_{k+1}(j) + b_i\big)/a_{ii} + (1-\omega)\boldsymbol{x}_{k+1/2}(i) \quad (8.6)$$

   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 9.6 for an alternative example of a preconditioner.

We will refer to the R, J, GS and SOR methods as the **classical (iteration) methods**. The R method will be discussed later, see Section 8.3.1.

8.1.1 The discrete Poisson system

Consider the classical methods applied to the discrete Poisson matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ given by (3.7). Let $n = m^2$ and set $h = 1/(m+1)$. In component form the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ can be written (cf. (3.4))

$$4\boldsymbol{v}(i,j) - \boldsymbol{v}(i-1,j) - \boldsymbol{v}(i+1,j) - \boldsymbol{v}(i,j-1) - \boldsymbol{v}(i,j+1) = h^2 f_{i,j}, \quad i,j = 1,\ldots,m,$$

with homogenous boundary conditions also given in (3.4). Solving for $\boldsymbol{v}(i,j)$ we obtain the **fixed point form**

$$\boldsymbol{v}(i,j) = \big(\boldsymbol{v}(i-1,j)+\boldsymbol{v}(i+1,j)+\boldsymbol{v}(i,j-1)+v(i,j+1)+e_{i,j}\big)/4, \quad e_{i,j} := f_{i,j}/(m+1)^2. \tag{8.7}$$

The R, J, GS , and SOR methods takes the form

$$\begin{aligned}
R : \boldsymbol{v}_{k+1}(i,j) &= \boldsymbol{v}_k(i,j) + \alpha\big(\boldsymbol{v}_k(i-1,j) + \boldsymbol{v}_k(i,j-1) + \boldsymbol{v}_k(i+1,j) \\
&\qquad + \boldsymbol{v}_k(i,j+1) - 4\boldsymbol{v}_k(i,j) + \boldsymbol{e}(i,j)\big) \\
J : \boldsymbol{v}_{k+1}(i,j) &= \big(\boldsymbol{v}_k(i-1,j) + \boldsymbol{v}_k(i,j-1) + \boldsymbol{v}_k(i+1,j) + \boldsymbol{v}_k(i,j+1) \\
&\qquad + \boldsymbol{e}(i,j)\big)/4 \\
GS : \boldsymbol{v}_{k+1}(i,j) &= \big(\boldsymbol{v}_{k+1}(i-1,j) + \boldsymbol{v}_{k+1}(i,j-1) + \boldsymbol{v}_k(i+1,j) + \boldsymbol{v}_k(i,j+1) \\
&\qquad + \boldsymbol{e}(i,j)\big)/4 \\
SOR : \boldsymbol{v}_{k+1}(i,j) &= \omega\big(\boldsymbol{v}_{k+1}(i-1,j) + \boldsymbol{v}_{k+1}(i,j-1) + \boldsymbol{v}_k(i+1,j) + \boldsymbol{v}_k(i,j+1) \\
&\qquad + \boldsymbol{e}(i,j)\big)/4 + (1-\omega)\boldsymbol{v}_k(i,j).
\end{aligned} \tag{8.8}$$

We note that

- For $\alpha = 1/4$ the R and J methods in (8.8) are identical.
- For a general system the R and J methods are identical if $\boldsymbol{A}$ is constant and nonzero on the diagonal and α is chosen as the inverse of this constant. See Exercise 8.2.
- For the discrete Poisson problem the choice $\alpha = 1/4$ in the J method is optimal in a way to become clear in Section 8.3.1.
- For GS and SOR we have used the **natural ordering**, i. e., $(i_1, j_1) < (i_2, j_2)$ if and only if $j_1 \leq j_2$ and $i_1 < i_2$ if $j_1 = j_2$. For the J method any ordering can be used.

Exercise 8.2 (Richardson and Jacobi)
Show that if $a_{ii} = d \neq 0$ for all i then Richardson's method with $\alpha := 1/d$ is the same as Jacobi's method.

In Algorithm 8.3 we give a Matlab program to test the convergence of Jacobi's method on the discrete Poisson problem. We carry out Jacobi iterations on the linear system (8.7) with $\boldsymbol{F} = (f_{ij}) \in \mathbb{R}^{m\times m}$, starting with $\boldsymbol{V}_0 = \boldsymbol{0} \in \mathbb{R}^{(m+2)\times(m+2)}$. The output is the number of iterations k, to obtain $\|\boldsymbol{V}^{(k)} - \boldsymbol{U}\|_M := \max_{i,j}|v_{ij} - u_{ij}| < tol$. Here $[u_{ij}] \in \mathbb{R}^{(m+2)\times(m+2)}$ is the "exact" solution of (8.7) computed using the fast Poisson solver in Algorithm 4.1. We set $k = K+1$

	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 8.1. *The number of iterations k_n to solve the discrete Poisson problem with n unknowns using the methods of Jacobi, Gauss-Seidel, and SOR (see text) with a tolerance 10^{-8}.*

if convergence is not obtained in K iterations. In Table 8.1 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 $\omega^* := 2/\big(1+\sin(\frac{\pi}{m+1})\big)$. We will derive this value later.

Algorithm 8.3 (Jacobi)

```
function k=jdp(F,K,tol)
m=length(F); U=fastpoisson(F); V=zeros(m+2,m+2); E=F/(m+1)^2;
for k=1:K
  V(2:m+1,2:m+1)=(V(1:m,2:m+1)+V(3:m+2,2:m+1)...
        +V(2:m+1,1:m)+V(2:m+1,3:m+2)+E)/4;
  if max(max(abs(V-U)))<tol,   return
  end
end
k=K+1;
```

For the GS and SOR methods we have used Algorithm 8.4. This is the analog of Algorithm 8.3 using 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.

Algorithm 8.4 (SOR)

```
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 j=2:m+1
      for i=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 8.1.

1. The rate (speed) of convergence is quite different for the four methods. The R, J and GS methods converge, but rather slowly. The R method with $\alpha = 1/4$ and the J method needs about twice as many iterations as the GS method. The improvement using the SOR method with optimal ω is spectacular.

2. We show in Section 8.3.4 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 (8.8) it follows that each iteration requires $O(n)$ arithmetic operations. Thus the number of arithmetic operations to achieve a given tolerance is $O(k_n \times n)$. Therefore the number of arithmetic operations 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.

5. Jacobi's method has the advantage that it can be easily parallelized.

8.2 Classical Iterative Methods; Matrix Form

To study convergence we need matrix formulations of the classical methods.

8.2.1 Fixed-point form

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 $\boldsymbol{M}^{-1}\boldsymbol{Ax} = \boldsymbol{M}^{-1}\boldsymbol{b}$. This system can be written $\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{M}^{-1}\boldsymbol{Ax} + \boldsymbol{M}^{-1}\boldsymbol{b}$, and we obtain $\boldsymbol{Ax} = \boldsymbol{b}$ in a **fixed-point form**

$$\boldsymbol{x} = \boldsymbol{Gx} + \boldsymbol{c}, \quad \boldsymbol{G} = \boldsymbol{I} - \boldsymbol{M}^{-1}\boldsymbol{A}, \quad \boldsymbol{c} = \boldsymbol{M}^{-1}\boldsymbol{b}. \tag{8.9}$$

For a general $\boldsymbol{G} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{c} \in \mathbb{C}^n$ a solution of $\boldsymbol{x} = \boldsymbol{Gx} + \boldsymbol{c}$ is called a **fixed-point**. The fixed-point is unique if $\boldsymbol{I} - \boldsymbol{G}$ is nonsingular.

The corresponding iterative method is given by

$$\boldsymbol{x}_{k+1} := \boldsymbol{Gx}_k + \boldsymbol{c}. \tag{8.10}$$

This is known as a **fixed-point iteration**. 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 fixed point since

$$\boldsymbol{x} = \lim_{k\to\infty} \boldsymbol{x}_{k+1} = \lim_{k\to\infty} (\boldsymbol{Gx}_k + \boldsymbol{c}) = \boldsymbol{G} \lim_{k\to\infty} \boldsymbol{x}_k + \boldsymbol{c} = \boldsymbol{Gx} + \boldsymbol{c}.$$

8.2.2 The preconditioning and splitting matrix

Different choices of $\boldsymbol{M}$ in (8.9) lead to different iterative methods. The matrix $\boldsymbol{M}$ can be interpreted in two ways. It is a **preconditioning matrix** since a good choice of $\boldsymbol{M}$ can lead to a preconditioned system $\boldsymbol{M}^{-1}\boldsymbol{Ax} = \boldsymbol{M}^{-1}\boldsymbol{b}$ 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

$$\boldsymbol{Mx}_{k+1} = (\boldsymbol{M} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{b} \tag{8.11}$$

which is equivalent to (8.9).

8.2.3 The splitting matrices for the classical methods

We now derive $\boldsymbol{M}$ for the classical methods. For J, GS and SOR it is convenient to 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 $\boldsymbol{D} := \operatorname{diag}(a_{11}, \ldots, a_{nn})$,

$$\boldsymbol{A_L} := \begin{bmatrix} 0 & & & \\ -a_{2,1} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n,1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix}, \quad \boldsymbol{A_R} := \begin{bmatrix} 0 & -a_{1,2} & \cdots & -a_{1,n} \\ & \ddots & \ddots & \vdots \\ & & 0 & -a_{n-1,n} \\ & & & 0 \end{bmatrix}. \tag{8.12}$$

Theorem 8.5 (Splitting matrices for R, J, and SOR)
The splitting and iteration matrices for the R, J and SOR methods are given by

$$\begin{aligned} &\boldsymbol{M}_R = \alpha^{-1}\boldsymbol{I},\ \boldsymbol{M}_J = \boldsymbol{D},\ \boldsymbol{M}_\omega = \omega^{-1}\boldsymbol{D} - \boldsymbol{A}_L, \\ &\boldsymbol{G}_R = \boldsymbol{I} - \alpha\boldsymbol{A},\ \boldsymbol{G}_J = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A},\ \boldsymbol{G}_\omega = \boldsymbol{I} - (\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_L)^{-1}\boldsymbol{A}. \end{aligned} \tag{8.13}$$

We obtain the matrices $\boldsymbol{M}_1$ and $\boldsymbol{G}_1$ for the GS method by letting $\omega = 1$ in $\boldsymbol{M}_\omega$ and $\boldsymbol{G}_\omega$

Proof. To find $\boldsymbol{M}$ we write the methods in the form (8.11). The formulas for $\boldsymbol{G}$ then follows immediately from (8.9).

The matrix form of the R method is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)$ or $\alpha^{-1}\boldsymbol{x}_{k+1} = \alpha^{-1}(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{b}$, and the formulas for $\boldsymbol{M}_R$ and $\boldsymbol{G}_R$ follows. The equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ can be written $\boldsymbol{D}\boldsymbol{x} - \boldsymbol{A}_L\boldsymbol{x} - \boldsymbol{A}_R\boldsymbol{x} = \boldsymbol{b}$ or $\boldsymbol{D}\boldsymbol{x} = \boldsymbol{A}_L\boldsymbol{x} + \boldsymbol{A}_R\boldsymbol{x} + \boldsymbol{b}$. This leads to

$$\begin{aligned} J:\ \boldsymbol{D}\boldsymbol{x}_{k+1} &= \boldsymbol{A}_L\boldsymbol{x}_k + \boldsymbol{A}_R\boldsymbol{x}_k + \boldsymbol{b},\ \text{or} \\ \boldsymbol{M}_J\boldsymbol{x}_{k+1} &= (\boldsymbol{A}_L + \boldsymbol{A}_R)\boldsymbol{x}_k + \boldsymbol{b}, \\ SOR:\ \boldsymbol{D}\boldsymbol{x}_{k+1} &= \omega\big(\boldsymbol{A}_L\boldsymbol{x}_{k+1} + \boldsymbol{A}_R\boldsymbol{x}_k + \boldsymbol{b}\big) + (1-\omega)\boldsymbol{D}\boldsymbol{x}_k,\ \text{or} \\ \boldsymbol{M}_\omega\boldsymbol{x}_{k+1} &= \big(\boldsymbol{A}_R + (\omega^{-1} - 1)\boldsymbol{D}\big)\boldsymbol{x}_k + \boldsymbol{b}. \end{aligned} \tag{8.14}$$

□

Example 8.6 (Splitting matrices)
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

$$\boldsymbol{G}_\omega = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \omega/2 & 0 \\ \omega^2/4 & \omega/2 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1-\omega & \omega/2 \\ \omega(1-\omega)/2 & 1-\omega+\omega^2/4 \end{bmatrix}. \tag{8.15}$$

For the J and GS method we have

$$\boldsymbol{G}_J = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix},\quad \boldsymbol{G}_1 = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix}. \tag{8.16}$$

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

$$\boldsymbol{x}_{k+1}(1) = \frac{1}{2}\boldsymbol{x}_k(2) + \frac{1}{2}, \quad \boldsymbol{x}_{k+1}(2) = \frac{1}{2}\boldsymbol{x}_{k+1}(1) + \frac{1}{2}.$$

Substituting the value of $\boldsymbol{x}_{k+1}(1)$ *from the first equation into the second equation we find*

$$\boldsymbol{x}_{k+1}(2) = \frac{1}{2}(\frac{1}{2}\boldsymbol{x}_k(2) + \frac{1}{2}) + \frac{1}{2} = \frac{1}{4}\boldsymbol{x}_k(2) + \frac{3}{4}.$$

Thus

$$\boldsymbol{x}_{k+1} = \begin{bmatrix} \boldsymbol{x}_{k+1}(1) \\ \boldsymbol{x}_{k+1}(2) \end{bmatrix} = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k(1) \\ \boldsymbol{x}_k(2) \end{bmatrix} + \begin{bmatrix} 1/2 \\ 3/4 \end{bmatrix} = \boldsymbol{G}_1\boldsymbol{x}_k + \boldsymbol{c}.$$

8.3 Convergence

Definition 8.7 (Convergence of $\boldsymbol{x}_{k+1} := \boldsymbol{G}\boldsymbol{x}_k + \boldsymbol{c}$)
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* $\boldsymbol{x}_0$.

We have the following necessary and sufficient condition for convergence:

Theorem 8.8 (Convergence of an iterative method)
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. We subtract $\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{c}$ from $\boldsymbol{x}_{k+1} = \boldsymbol{G}\boldsymbol{x}_k + \boldsymbol{c}$. The vector $\boldsymbol{c}$ cancels and we obtain $\boldsymbol{x}_{k+1} - \boldsymbol{x} = \boldsymbol{G}(\boldsymbol{x}_k - \boldsymbol{x})$. By induction on k

$$\boldsymbol{x}_k - \boldsymbol{x} = \boldsymbol{G}^k(\boldsymbol{x}_0 - \boldsymbol{x}), \quad k = 0, 1, 2, \ldots \tag{8.17}$$

Clearly $\boldsymbol{x}_k - \boldsymbol{x} \to \boldsymbol{0}$ if $\boldsymbol{G}^k \to \boldsymbol{0}$. The converse follows by choosing $\boldsymbol{x}_0 - \boldsymbol{x} = \boldsymbol{e}_j$, the jth unit vector for $j = 1, \ldots, n$. □

Theorem 8.9 (Sufficient condition for convergence)
If $\|\boldsymbol{G}\| < 1$ *for some consistent matrix norm on* $\mathbb{C}^{n\times n}$, *then the iteration* $\boldsymbol{x}_{k+1} = \boldsymbol{G}\boldsymbol{x}_k + \boldsymbol{c}$ *converges.*

Proof. We have

$$\|\boldsymbol{x}_k - \boldsymbol{x}\| = \|\boldsymbol{G}^k(\boldsymbol{x}_0 - \boldsymbol{x})\| \le \|\boldsymbol{G}^k\|\|\boldsymbol{x}_0 - \boldsymbol{x}\| \le \|\boldsymbol{G}\|^k\|\boldsymbol{x}_0 - \boldsymbol{x}\| \to \boldsymbol{0}, \quad k \to \infty.$$

□

A necessary and sufficient condition for convergence involves the eigenvalues of $\boldsymbol{G}$. We define the **spectral radius** of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ as the maximum absolute value of its eigenvalues.

$$\rho(\boldsymbol{A}) := \max_{\lambda\in\sigma(\boldsymbol{A})} |\lambda|. \tag{8.18}$$

Theorem 8.10 (When does an iterative method converge?)
Suppose $\boldsymbol{G} \in \mathbb{C}^{n\times n}$ 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 $\rho(\boldsymbol{G}) < 1$.

We will prove this theorem using Theorem 8.27 in Section 8.4.

8.3.1 Convergence of Richardson's method.

Recall that Richardson's method can be written in the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha\boldsymbol{r}_k, \quad \boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k. \tag{8.19}$$

We will assume that α is real. If all eigenvalues of $\boldsymbol{A}$ have positive real parts then then the R method converges provided α is positive and sufficiently small. We show this result for positive eigenvalues and leave the more general case to Exercise 8.13.

Theorem 8.11 (Convergence of Richardson's method)
If $\boldsymbol{A}$ has positive eigenvalues then the R method converges if and only if $0 < \alpha < 2/\rho(\boldsymbol{A})$. Moreover,

$$\min_{\alpha} \boldsymbol{G}(\alpha) = \boldsymbol{G}(\alpha^*), \quad \alpha^* := \frac{2}{\lambda_{max} + \lambda_{min}}, \text{ and } \rho\big(\boldsymbol{G}(\alpha^*)\big) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\kappa - 1}{\kappa + 1}, \tag{8.20}$$

where λ_{max} and λ_{min} are the largest and smallest eigenvalue of $\boldsymbol{A}$ and $\kappa := \lambda_{max}/\lambda_{min}$.

Proof. The eigenvalues of $\boldsymbol{G}(\alpha) = \boldsymbol{I} - \alpha\boldsymbol{A}$ are $\mu_j(\alpha) = 1 - \alpha\lambda_j, \quad j = 1, \ldots, n$. We have $\max_j \mu_j < 1$ if and only if $\alpha > 0$ and $\min_j \mu_j = 1 - \alpha\rho(\boldsymbol{A}) > -1$ if and only if $\alpha < 2/\rho(\boldsymbol{A})$. The method converges if and only if $\rho(\boldsymbol{G}(\alpha)) < 1$ which from what we have shown is equivalent to $0 < \alpha < 2/\rho(\boldsymbol{A})$. Since $1 - \alpha^*\lambda_{min} = \alpha^*\lambda_{max} - 1$ we have we have

$$\rho\big(\boldsymbol{G}(\alpha^*)\big) = 1 - \alpha^*\lambda_{min} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\kappa - 1}{\kappa + 1}.$$

Now $\boldsymbol{G}(\alpha) \geq \boldsymbol{G}(\alpha^*)$, for if $0 < \alpha < \alpha^*$ then from what we showed $\rho\big(\boldsymbol{G}(\alpha)\big) \geq 1 - \alpha\lambda_{min} > 1 - \alpha^*\lambda_{min} = \rho\big(\boldsymbol{G}(\alpha^*)\big)$, and if $\alpha^* < \alpha < 2/\rho(\boldsymbol{A})$ then $-\rho\big(\boldsymbol{G}(\alpha)\big) \leq 1 - \alpha\lambda_{max} < 1 - \alpha^*\lambda_{max} = -\rho\big(\boldsymbol{G}(\alpha^*)\big)$, and again $\rho\big(\boldsymbol{G}(\alpha)\big) > \rho\big(\boldsymbol{G}(\alpha^*)\big)$. ∎

Corollary 8.12 (Rate of convergence for the R method)
Suppose $\boldsymbol{A}$ is symmetric positive definite with largest and smallest eigenvalue λ_{max} and λ_{min}, respectively. Richardson's method with acceleration parameter $\alpha^* := \frac{2}{\lambda_{max}+\lambda_{min}}$ converges. More precisely

$$\|\boldsymbol{x}_k - \boldsymbol{x}\|_2 \leq \big(\frac{\kappa-1}{\kappa+1}\big)^k \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2, \quad k = 0, 1, 2, \ldots \tag{8.21}$$

where $\kappa := \lambda_{max}/\lambda_{min}$ is the spectral condition number of $\boldsymbol{A}$.

Proof. The spectral norm $\| \ \|_2$ is consistent and therefore $\|\boldsymbol{x}_k - \boldsymbol{x}\|_2 \leq \|\boldsymbol{G}(\alpha^*)\|_2^k \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2$. But for a symmetric positive definite matrix the spectral norm is equal to the spectral radius and the result follows form (8.20). □

Exercise 8.13 (Convergence of the R-method when eigenvalues have positive real p
Suppose all eigenvalues λ_j of $\boldsymbol{A}$ have positive real parts u_j for $j = 1, \ldots, n$ and that α is real. Show that the R method converges if and only if $0 < \alpha < \min_j(2u_j/|\lambda_j|^2)$.

8.3.2 Convergence of SOR

The condition $\omega \in (0, 2)$ is necessary for convergence of the SOR method.

Theorem 8.14 (Necessay condition for convergence of SOR)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is nonsingular with nonzero diagonal elements. If the SOR method applied to $\boldsymbol{A}$ converges then $\omega \in (0, 2)$.

Proof. We have (cf. (8.14)) $\boldsymbol{D}\boldsymbol{x}_{k+1} = \omega\big(\boldsymbol{A}_L\boldsymbol{x}_{k+1} + \boldsymbol{A}_R\boldsymbol{x}_k + \boldsymbol{b}\big) + (1-\omega)\boldsymbol{D}\boldsymbol{x}_k$ or $\boldsymbol{x}_{k+1} = \omega\big(\boldsymbol{L}\boldsymbol{x}_{k+1} + \boldsymbol{R}\boldsymbol{x}_k + \boldsymbol{D}^{-1}\boldsymbol{b}\big) + (1-\omega)\boldsymbol{x}_k$, where $\boldsymbol{L} := \boldsymbol{D}^{-1}\boldsymbol{A_L}$ and $\boldsymbol{R} := \boldsymbol{D}^{-1}\boldsymbol{A_R}$. Thus $(\boldsymbol{I} - \omega\boldsymbol{L})\boldsymbol{x}_{k+1} = (\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I})\boldsymbol{x}_k + \boldsymbol{D}^{-1}\boldsymbol{b}$ so the following form of the iteration matrix is obtained

$$\boldsymbol{G}_\omega = (\boldsymbol{I} - \omega\boldsymbol{L})^{-1}\big(\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}\big). \tag{8.22}$$

We next compute the determinant of $\boldsymbol{G}_\omega$. Since $\boldsymbol{I} - \omega\boldsymbol{L}$ is lower triangular with ones on the diagonal, the same holds for the inverse by Lemma 1.22, and therefore the determinant of this matrix is equal to 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 $\boldsymbol{G}_\omega$ and 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 8.10 SOR diverges. □

The SOR method always converges for a symmetric positive definite matrix.

Theorem 8.15 (SOR on positive definite matrix)
SOR converges for a symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ if and only if $0 < \omega < 2$. In particular, Gauss-Seidel's method converges for a symmetric positive definite matrix.

Proof. By Theorem 8.14 convergence implies $0 < \omega < 2$. Suppose now $0 < \omega < 2$ and let $(\lambda, \boldsymbol{x})$ be an eigenpair for $\boldsymbol{G}_\omega$. Note that λ and $\boldsymbol{x}$ can be complex. We need to show that $|\lambda| < 1$. The following identity will be shown below:

$$\omega^{-1}(2-\omega)|1-\lambda|^2\boldsymbol{x}^*\boldsymbol{D}\boldsymbol{x} = (1-|\lambda|^2)\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}, \tag{8.23}$$

where $\boldsymbol{D} := \operatorname{diag}(a_{11}, \ldots, a_{nn})$. Now $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{x}^*\boldsymbol{D}\boldsymbol{x}$ are positive for all nonzero $x \in \mathbb{C}^n$ since a positive definite matrix has positive diagonal elements $a_{ii} = \boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{e}_i > 0$. It follows that the left hand side of (8.23) is nonnegative and then the right hand side must be nonnegative as well. This implies $|\lambda| \le 1$. It remains to show that we cannot have $\lambda = 1$. By (8.13) the eigenpair equation $\boldsymbol{G}_\omega\boldsymbol{x} = \lambda\boldsymbol{x}$ can be written $\boldsymbol{x} - (\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{L}})^{-1}\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ or

$$\boldsymbol{A}\boldsymbol{x} = (\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{L}})\boldsymbol{y}, \quad \boldsymbol{y} := (1-\lambda)\boldsymbol{x}. \tag{8.24}$$

Now $\boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{0}$ implies that $\lambda \neq 1$.

To prove equation (8.23) consider the matrix $\boldsymbol{E} := \omega^{-1}\boldsymbol{D} + \boldsymbol{A}_{\boldsymbol{R}} - \boldsymbol{D}$. Since $\boldsymbol{A}_{\boldsymbol{R}} - \boldsymbol{D} = -\boldsymbol{A}_{\boldsymbol{L}} - \boldsymbol{A}$ we find $\boldsymbol{E}\boldsymbol{y} = (\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{L}} - \boldsymbol{A})\boldsymbol{y} \overset{(8.24)}{=} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y} = \lambda\boldsymbol{A}\boldsymbol{x}$. Observe that $(\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{L}})^* = \omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{R}}$ so that by (8.24)

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{x})^*\boldsymbol{y} + \boldsymbol{y}^*(\lambda\boldsymbol{A}\boldsymbol{x}) &= \boldsymbol{y}^*(\omega^{-1}\boldsymbol{D} - \boldsymbol{A}_{\boldsymbol{R}})\boldsymbol{y} + \boldsymbol{y}^*\boldsymbol{E}\boldsymbol{y} = \boldsymbol{y}^*(2\omega^{-1}-1)\boldsymbol{D}\boldsymbol{y}\\ &= \omega^{-1}(2-\omega)|1-\lambda|^2\boldsymbol{x}^*\boldsymbol{D}\boldsymbol{x}.\end{aligned}$$

Since $(\boldsymbol{A}\boldsymbol{x})^* = \boldsymbol{x}^*\boldsymbol{A}$, $\boldsymbol{y} := (1-\lambda)\boldsymbol{x}$ and $\boldsymbol{y}^* = (1-\overline{\lambda})\boldsymbol{x}^*$ this also equals

$$(\boldsymbol{A}\boldsymbol{x})^*\boldsymbol{y} + \boldsymbol{y}^*(\lambda\boldsymbol{A}\boldsymbol{x}) = (1-\lambda)\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \lambda(1-\overline{\lambda})\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} = (1-|\lambda|^2)\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x},$$

and (8.23) follows. ∎

Exercise 8.16 (Examle: GS converges, J diverges)
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 $-1/2 < a < 1$. Thus, GS converges for these values of a. Show that the J method does not converge for $1/2 < a < 1$.

Exercise 8.17 (Divergence example for J and GS)
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 8.18 (Strictly diagonally dominance; The J method)
Show that the J method converges if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ *for* $i = 1, \ldots, n$.

Exercise 8.19 (Strictly diagonally dominance; The GS method)
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* $|\boldsymbol{\epsilon}_{k+1}(j)| \leq r \|\boldsymbol{\epsilon}_k\|_\infty$ *for* $j = 1, \ldots, i$. *Conclude that Gauss-Seidel's method is convergent when* $\boldsymbol{A}$ *is strictly diagonally dominant.*

8.3.3 Convergence of the classical methods for the discrete Poisson matrix

We know the eigenvalues of the discrete Poisson matrix $\boldsymbol{A}$ given by (3.7) and we can use this to estimate the number of iterations necessary to achieve a given accuracy for the various methods.

Recall that by (3.20) the eigenvalues $\lambda_{j,k}$ of $\boldsymbol{A}$ are

$$\lambda_{j,k} = 4 - 2\cos(j\pi h) - 2\cos(k\pi h), \quad j,k = 1, \ldots, m, h = 1/(m+1).$$

It follows that the largest and smallest eigenvalue of $\boldsymbol{A}$, and the spectral condition number κ of $\boldsymbol{A}$, are given by

$$\lambda_{max} = 8\cos^2 w, \quad \lambda_{min} = 8\sin^2 w, \quad \kappa := \frac{\cos^2 w}{\sin^2 w}, \quad w := \frac{\pi}{2(m+1)}. \tag{8.25}$$

Consider first the J method. The matrix $\boldsymbol{G}_J = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{A}/4$ has eigenvalues

$$\mu_{j,k} = 1 - \frac{1}{4}\lambda_{j,k} = \frac{1}{2}\cos(j\pi h) + \frac{1}{2}\cos(k\pi h), \quad j,k = 1, \ldots, m. \tag{8.26}$$

It follows that $\rho(\boldsymbol{G}_J) = \cos(\pi h) < 1$. Since $\boldsymbol{G}_J$ is symmetric it is normal, and the spectral norm is equal to the spectral radios (cf. Theorem 7.14). We obtain

$$\|\boldsymbol{x}_k - \boldsymbol{x}\|_2 \leq \|\boldsymbol{G}_J)\|_2^k \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 = \cos^k(\pi h)\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2, \quad k = 0, 1, 2, \ldots \tag{8.27}$$

The R method given by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha \boldsymbol{r}_k$ with $\alpha = 2/(\lambda_{max} + \lambda_{min}) = 1/4$ is the same as the J-method so (8.27) holds in this case as well. This also follows from Corollary 8.12 with κ given by (8.25).

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 8.3.2.

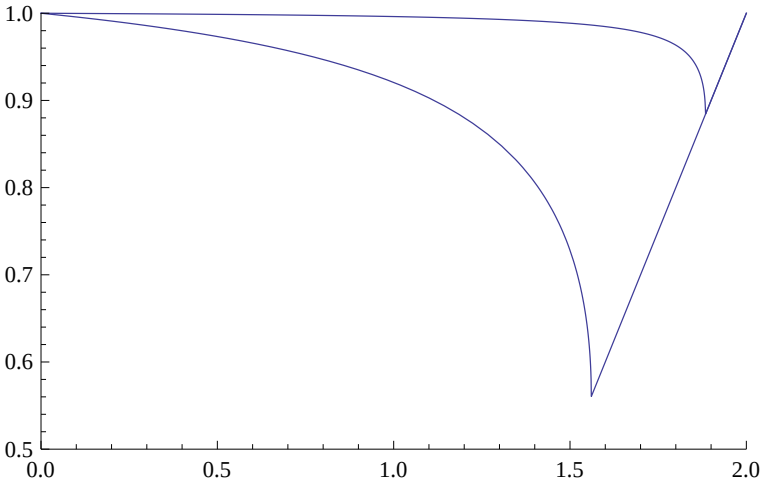


Figure 8.1. $\rho(\boldsymbol{G}_\omega)$ *with* $\omega \in [0,2]$ *for* $n = 100$*, (lower curve) and* $n = 2500$ *(upper curve).*

Theorem 8.20 (The spectral radius of SOR matrix)
Consider the SOR iteration (8.8)*, with the natural ordering. The spectral radius of* $\boldsymbol{G}_\omega$ *is*

$$\rho(\boldsymbol{G}_\omega) = \begin{cases} \frac{1}{4}\left(\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega-1)}\right)^2, & \text{for } 0 < \omega \le \omega^*, \\ \omega - 1, & \text{for } \omega^* < \omega < 2, \end{cases} \tag{8.28}$$

where $\beta := \rho(\boldsymbol{G}_J) = \cos(\pi h)$ *and*

$$\omega^* := \frac{2}{1 + \sqrt{1-\beta^2}} > 1. \tag{8.29}$$

Moreover,

$$\rho(\boldsymbol{G}_\omega) > \rho(\boldsymbol{G}_{\omega^*}) \text{ for } \omega \in (0,2) \setminus \{\omega^*\}. \tag{8.30}$$

A plot of $\rho(\boldsymbol{G}_\omega)$ as a function of $\omega \in (0,2)$ is shown in Figure 8.1 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 (8.28),(8.29) that

$$\omega^* = \frac{2}{1+\sin(\pi h)}, \quad \rho(\boldsymbol{G}_{\omega^*}) = \omega^* - 1 = \frac{1-\sin(\pi h)}{1+\sin(\pi h)}, \quad h = \frac{1}{m+1}. \tag{8.31}$$

	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 8.2. *Spectral radia for $\boldsymbol{G}_J$, $\boldsymbol{G}_1$, $\boldsymbol{G}_{\omega^*}$ and the smallest integer k_n such that $\rho(\boldsymbol{G})^{k_n} \leq 10^{-8}$.*

Letting $\omega = 1$ in (8.28) we find $\rho(\boldsymbol{G}_1) = \beta^2 = \rho(\boldsymbol{G}_J)^2 = \cos^2(\pi h)$ for the GS method. 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 8.2 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 8.1.

8.3.4 Number of iterations

Consider next the **rate of convergence** of the iteration $\boldsymbol{x}_{k+1} = \boldsymbol{G}\boldsymbol{x}_k + \boldsymbol{c}$. We like to know how fast the iterative method converges. Suppose $\| \ \|$ is a matrix norm that is subordinate to a vector norm also denoted by $\| \ \|$. Recall that $\boldsymbol{x}_k - \boldsymbol{x} = \boldsymbol{G}^k(\boldsymbol{x}_0 - \boldsymbol{x})$. For k sufficiently large

$$\|\boldsymbol{x}_k - \boldsymbol{x}\| \leq \|\boldsymbol{G}^k\|\|\boldsymbol{x}_0 - \boldsymbol{x}\| \approx \rho(\boldsymbol{G})^k\|\boldsymbol{x}_0 - \boldsymbol{x}\|.$$

For the last formula we apply Theorem 8.30 which says that $\lim_{k\to\infty}\|\boldsymbol{G}^k\|^{1/k} = \rho(\boldsymbol{G})$. For Jacobi's method and the spectral norm we have $\|\boldsymbol{G}_J^k\|_2 = \rho(\boldsymbol{G}_J)^k$ (cf. (8.27)).

For fast convergence we should use a $\boldsymbol{G}$ with small spectral radius.

Lemma 8.22 (Number of iterations)
Suppose $\rho(\boldsymbol{G}) = 1-\eta$ for some $0 < \eta < 1$, $\| \ \|$ a consistent matrix norm on $\mathbb{C}^{n\times n}$, and let $s \in \mathbb{N}$. Then

$$\tilde{k} := \frac{s\log(10)}{\eta} \tag{8.32}$$

is an estimate for the smallest number of iterations k so that $\rho(\boldsymbol{G})^k \leq 10^{-s}$.

Proof. The estimate $\tilde{k}$ is an approximate solution of the equation $\rho(\boldsymbol{G})^k = 10^{-s}$. Thus, since $-\log(1-\eta) \approx \eta$ when η is small

$$k = -\frac{s\log(10)}{\log(1-\eta)} \approx \frac{s\log(10)}{\eta} = \tilde{k}.$$

□

The following estimates are obtained. They agree with those we found numerically in Section 8.1.1.

- R and 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,

$$\tilde{k}_n = \frac{\log(10)s}{2\pi} \sqrt{n} + O(n^{-1/2}) = O(\sqrt{n}).$$

Exercise 8.23 (Convergence example for fix point iteration)
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

$$\boldsymbol{x}_k(1) = \boldsymbol{x}_k(2) = 1 - a^k, \quad k \geq 0,$$

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 (8.32) for the rate of convergence for $a = 0.9$ and $s = 16$ and compare with the true number of iterations determined from $|a|^k \leq 10^{-16}$.

Exercise 8.24 (Estimate in Lemma 8.22 can be exact)
Consider the iteration in Example 8.6. Show that $\rho(\boldsymbol{G}_J) = 1/2$. Then show that $\boldsymbol{x}_k(1) = \boldsymbol{x}_k(2) = 1 - 2^{-k}$ for $k \geq 0$. Thus the estimate in Lemma 8.22 is exact in this case.

We note that

1. The convergence depends on the behavior of the powers $\boldsymbol{G}^k$ as k increases. The matrix $\boldsymbol{M}$ should be chosen so that all elements in $\boldsymbol{G}^k$ converge quickly to zero and such that the linear system (8.11) 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 $\boldsymbol{G}$ with small elements, but then (8.11) might not be easy to solve for $\boldsymbol{x}_{k+1}$.

2. The convergence $\lim_{k\to\infty}\|\boldsymbol{G}^k\|^{1/k} = \rho(\boldsymbol{G})$ can be quite slow (cf. Exercise 8.25).

Exercise 8.25 (Slow spectral radius convergence)
The convergence $\lim_{k\to\infty}\|\boldsymbol{A}^k\|^{1/k} = \rho(\boldsymbol{A})$ *can be quite slow. Consider*

$$\boldsymbol{A} := \begin{bmatrix} \lambda & a & 0 & \cdots & 0 & 0 \\ 0 & \lambda & a & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & a \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

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* $k \geq n-1$.

(a) *Pick an* n, *e.g.* $n = 5$, *and make a plot of* $f(k)$ *for* $\lambda = 0.9$, $a = 10$, *and* $n-1 \leq k \leq 200$. *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* $\boldsymbol{A}^k = (a\boldsymbol{E} + \lambda\boldsymbol{I})^k = \sum_{j=0}^{\min\{k,n-1\}}\binom{k}{j}a^j\lambda^{k-j}\boldsymbol{E}^j$ *and conclude that the* $(1,n)$ *element is given by* $f(k)$ *for* $k \geq n-1$.

8.3.5 Stopping the iteration

In Algorithms 8.3 and 8.4 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. The following result indicates that $\|\boldsymbol{x}_k - \boldsymbol{x}\|$ can be quite large if $\|\boldsymbol{G}\|$ is close to one.

Lemma 8.26 (Be careful when stopping)
Suppose $\|\boldsymbol{G}\| < 1$ *for some consistent matrix norm on* $\mathbb{C}^{n\times n}$ *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*

$$\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\| \geq \frac{1 - \|\boldsymbol{G}\|}{\|\boldsymbol{G}\|}\|\boldsymbol{x}_k - \boldsymbol{x}\|, \quad k \geq 1. \tag{8.33}$$

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 (8.33). ∎

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 (8.10) in an alternative form. If $\boldsymbol{M}$ is the splitting matrix of the method then by (8.11) we have $\boldsymbol{M}\boldsymbol{x}_{k+1} = \boldsymbol{M}\boldsymbol{x}_k - \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{b}$. This leads to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{M}^{-1}\boldsymbol{r}_k, \quad \boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k. \tag{8.34}$$

Testing on $\boldsymbol{r}_k$ works fine if $\boldsymbol{A}$ is well conditioned, but Theorem 7.28 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.

8.4 Powers of a matrix

Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ be a square matrix. In this section we consider the special matrix sequence $\{\boldsymbol{A}^k\}$ of powers of $\boldsymbol{A}$. We want to know when this sequence converges to the zero matrix. Such a sequence occurs in iterative methods (cf. (8.17)), in Markov processes in statistics, in the converge of geometric series of matrices (Neumann series cf. Section 8.4.2) and in many other applications.

8.4.1 The spectral radius

In this section we show the following theorem.

Theorem 8.27 (When is $\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{0}$?)
For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *we have*

$$\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{0} \Longleftrightarrow \rho(\boldsymbol{A}) < 1,$$

where $\rho(\boldsymbol{A})$ *is the spectral radius of* $\boldsymbol{A}$ *given by* (8.18).

Clearly $\rho(\boldsymbol{A}) < 1$ is a necessary condition for $\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{0}$. For if $(\lambda, \boldsymbol{x})$ is an eigenpair of $\boldsymbol{A}$ with $|\lambda| \geq 1$ and $\|\boldsymbol{x}\|_2 = 1$ then $\boldsymbol{A}^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$, and this implies $\|\boldsymbol{A}^k\|_2 \geq \|\boldsymbol{A}^k\boldsymbol{x}\|_2 = \|\lambda^k\boldsymbol{x}\|_2 = |\lambda|^k$, and it follows that $\boldsymbol{A}^k$ does not tend to zero.

The sufficiency condition is harder to show. We construct a consistent matrix norm on $\mathbb{C}^{n\times n}$ such that $\|\boldsymbol{A}\| < 1$ and then use Theorems 8.8 and 8.9.

We start with

Theorem 8.28 (Any consistent norm majorizes the spectral radius) *For any matrix norm $\|\cdot\|$ that is consistent on $\mathbb{C}^{n\times n}$ and any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$.*

Proof. Let $(\lambda, \boldsymbol{x})$ be an eigenpair for $\boldsymbol{A}$ and define $\boldsymbol{X} := [\boldsymbol{x}, \ldots, \boldsymbol{x}] \in \mathbb{C}^{n\times n}$. 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 next theorem shows that if $\rho(\boldsymbol{A}) < 1$ then $\|\boldsymbol{A}\| < 1$ for some consistent matrix norm on $\mathbb{C}^{n\times n}$, thus completing the proof of Theorem 8.27.

Theorem 8.29 (The spectral radius can be approximated by a norm) *Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\epsilon > 0$ be given. There is a consistent matrix norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ 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 5.13 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 $\boldsymbol{D}_t := \operatorname{diag}(t, t^2, \ldots, t^n) \in \mathbb{R}^{n\times n}$, 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 $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ and $t > 0$ we use the one norm to define the matrix norm $\|\boldsymbol{B}\|_t := \|\boldsymbol{D}_t\boldsymbol{U}^*\boldsymbol{B}\boldsymbol{U}\boldsymbol{D}_t^{-1}\|_1$. We leave it as an exercise to show that $\|\ \|_t$ is a consistent matrix norm on $\mathbb{C}^{n\times n}$. 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

$$\begin{aligned}\|\boldsymbol{A}\| = \|\boldsymbol{D}_t\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{D}_t^{-1}\|_1 = \|\boldsymbol{D}_t\boldsymbol{R}\boldsymbol{D}_t^{-1}\|_1 &= \max_{1\leq j\leq n}\sum_{i=1}^{n}|\left(\boldsymbol{D}_t\boldsymbol{R}\boldsymbol{D}_t^{-1}\right)_{ij}| \\ &\leq \max_{1\leq j\leq n}(|\lambda_j| + \epsilon) = \rho(\boldsymbol{A}) + \epsilon.\end{aligned}$$

□

A consistent matrix norm of a matrix can be much larger than the spectral radius. However the following result holds.

Theorem 8.30 (Spectral radius convergence)
For any consistent matrix norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ and any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have

$$\lim_{k\to\infty} \|\boldsymbol{A}^k\|^{1/k} = \rho(\boldsymbol{A}). \tag{8.35}$$

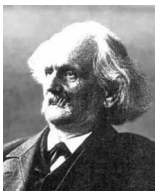
Proof. By Theorems 5.1 and 8.28 we obtain $\rho(\boldsymbol{A})^k = \rho(\boldsymbol{A}^k) \leq \|\boldsymbol{A}^k\|$ for any $k \in \mathbb{N}$ so that $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}^k\|^{1/k}$. Let $\epsilon > 0$ and consider the matrix $\boldsymbol{B} := (\rho(\boldsymbol{A})+\epsilon)^{-1}\boldsymbol{A}$. Then $\rho(\boldsymbol{B}) = \rho(\boldsymbol{A})/(\rho(\boldsymbol{A}) + \epsilon) < 1$ and $\|\boldsymbol{B}^k\| \to 0$ by Theorem 8.27 as $k \to \infty$. Choose $N \in \mathbb{N}$ such that $\|\boldsymbol{B}^k\| < 1$ for all $k \geq N$. Then for $k \geq N$

$$\|\boldsymbol{A}^k\| = \|(\rho(\boldsymbol{A}) + \epsilon)^k \boldsymbol{B}^k\| = \big(\rho(\boldsymbol{A}) + \epsilon\big)^k \|\boldsymbol{B}^k\| < \big(\rho(\boldsymbol{A}) + \epsilon\big)^k.$$

We have shown that $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}^k\|^{1/k} \leq \rho(\boldsymbol{A}) + \epsilon$ for $k \geq N$. Since ϵ is arbitrary the result follows. □

Exercise 8.31 (A special norm)
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 8.29 is a consistent matrix norm on $\mathbb{C}^{n\times n}$.

8.4.2 Neumann series



Carl Neumann., 1832–1925. He studied potential theory. The Neumann boundary conditions are named after him.

Let $\boldsymbol{B}$ be a square matrix. In this section we consider the **Neumann series** $\sum_{k=0}^{\infty} \boldsymbol{B}^k$ which is a matrix analogue of a geometric series of numbers.

Consider an infinite series $\sum_{k=0}^{\infty} \boldsymbol{A}_k$ of matrices in $\mathbb{C}^{n\times n}$. We say that the series converges if the sequence of partial sums $\{\boldsymbol{S}_m\}$ given by $\boldsymbol{S}_m = \sum_{k=0}^{m} \boldsymbol{A}_k$ converges. The series converges if and only if $\{\boldsymbol{S}_m\}$ is a Cauchy sequence, i.e. to each $\epsilon > 0$ there exists an integer N so that $\|\boldsymbol{S}_l - \boldsymbol{S}_m\| < \epsilon$ for all $l > m \geq N$.

Theorem 8.32 (Neumann series)
Suppose $\boldsymbol{B} \in \mathbb{C}^{n\times n}$. 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* $\mathbb{C}^{n\times n}$ *then*

$$\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| \leq \frac{1}{1-\|\boldsymbol{B}\|}. \tag{8.36}$$

Proof.

1. Suppose $\rho(\boldsymbol{B}) < 1$. We show that $\boldsymbol{S}_m := \sum_{k=0}^{m} \boldsymbol{B}^k$ is a Cauchy sequence and hence convergent. Let $\epsilon > 0$. By Theorem 8.29 there is a consistent matrix norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ such that $\|\boldsymbol{B}\| < 1$. Then for $l > m$

$$\|\boldsymbol{S}_l - \boldsymbol{S}_m\| = \|\sum_{k=m+1}^{l} \boldsymbol{B}^k\| \leq \sum_{k=m+1}^{l} \|\boldsymbol{B}\|^k \leq \|\boldsymbol{B}\|^{m+1} \sum_{k=0}^{\infty} \|\boldsymbol{B}\|^k = \frac{\|\boldsymbol{B}\|^{m+1}}{1-\|\boldsymbol{B}\|}.$$

   But then $\{\boldsymbol{S}_m\}$ is a Cauchy sequence provided N is such that $\frac{\|\boldsymbol{B}\|^{N+1}}{1-\|\boldsymbol{B}\|} < \epsilon$.

   Conversely, suppose $(\lambda, \boldsymbol{x})$ is an eigenpair for $\boldsymbol{B}$ with $|\lambda| \geq 1$. We find $\boldsymbol{S}_m\boldsymbol{x} = \sum_{k=0}^{m} \boldsymbol{B}^k\boldsymbol{x} = \big(\sum_{k=0}^{m} \lambda^k\big)\boldsymbol{x}$. Since λ^k does not tend to zero the series $\sum_{k=0}^{\infty} \lambda^k$ is not convergent and therefore $\{\boldsymbol{S}_m\boldsymbol{x}\}$ and hence $\{\boldsymbol{S}_m\}$ does not converge.

2. We have

$$\big(\sum_{k=0}^{m} \boldsymbol{B}^k\big)(\boldsymbol{I}-\boldsymbol{B}) = \boldsymbol{I}+\boldsymbol{B}+\cdots+\boldsymbol{B}^m-(\boldsymbol{B}+\cdots+\boldsymbol{B}^{m+1}) = \boldsymbol{I}-\boldsymbol{B}^{m+1}. \tag{8.37}$$

   Since $\rho(\boldsymbol{B}) < 1$ we conclude that $\boldsymbol{B}^{m+1} \to 0$ and hence taking limits in (8.37) 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 2: $\|(\boldsymbol{I}-\boldsymbol{B})^{-1}\| = \|\sum_{k=0}^{\infty} \boldsymbol{B}^k\| \leq \sum_{k=0}^{\infty} \|\boldsymbol{B}\|^k = \frac{1}{1-\|\boldsymbol{B}\|}$.

□

Exercise 8.33 (When is $\boldsymbol{A}+\boldsymbol{E}$ nonsingular?)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *is nonsingular and* $\boldsymbol{E} \in \mathbb{C}^{n\times n}$. *Show that* $\boldsymbol{A}+\boldsymbol{E}$ *is nonsingular if and only if* $\rho(\boldsymbol{A}^{-1}\boldsymbol{E}) < 1$.

8.5 The Optimal SOR Parameter ω

The following analysis is only carried out for the discrete Poisson matrix. It also holds for the averaging matrix given by (3.9). A more general theory is presented

in [35]. We will compare the eigenpair equations for $\boldsymbol{G}_J$ and $\boldsymbol{G}_\omega$. It is convenient to write these equations using the matrix formulation $\boldsymbol{TV} + \boldsymbol{VT} = h^2\boldsymbol{F}$. If $\boldsymbol{G}_J\boldsymbol{v} = \mu\boldsymbol{v}$ is an eigenpair of $\boldsymbol{G}_J$ then

$$\frac{1}{4}(v_{i-1,j} + v_{i,j-1} + v_{i+1,j} + v_{i,j+1}) = \mu v_{i,j}, \quad i,j = 1,\ldots,m, \tag{8.38}$$

where $\boldsymbol{V} := \mathrm{vec}(\boldsymbol{v}) \in \mathbb{R}^{m\times m}$ and $v_{i,j} = 0$ if $i \in \{0, m+1\}$ or $j \in \{0, m+1\}$.

Suppose $(\lambda, \boldsymbol{w})$ is an eigenpair for $\boldsymbol{G}_\omega$. By (8.22) $(\boldsymbol{I} - \omega\boldsymbol{L})^{-1}\big(\omega\boldsymbol{R} + (1-\omega)\boldsymbol{I}\big)\boldsymbol{w} = \lambda\boldsymbol{w}$ or

$$(\omega\boldsymbol{R} + \lambda\omega\boldsymbol{L})\boldsymbol{w} = (\lambda + \omega - 1)\boldsymbol{w}, \tag{8.39}$$

where $l_{i,i-m} = r_{i,i+m} = 1/4$ for all i, and all other elements in $\boldsymbol{L}$ and $\boldsymbol{R}$ are equal to zero. Let $\boldsymbol{w} = \mathrm{vec}(\boldsymbol{W})$, where $\boldsymbol{W} \in \mathbb{C}^{m\times m}$. Then (8.39) can be written

$$\frac{\omega}{4}(\lambda w_{i-1,j} + \lambda w_{i,j-1} + w_{i+1,j} + w_{i,j+1}) = (\lambda + \omega - 1)w_{i,j}, \tag{8.40}$$

where $w_{i,j} = 0$ if $i \in \{0, m+1\}$ or $j \in \{0, m+1\}$.

Theorem 8.34 (The optimal ω)
Consider the SOR method applied to the discrete Poisson matrix (3.9)*, 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*

$$\mu := \frac{\lambda + \omega - 1}{\omega\lambda^{1/2}} \tag{8.41}$$

   is an eigenvalue of $\boldsymbol{G}_J$.

2. *If μ is an eigenvalue of $\boldsymbol{G}_J$ and λ satisfies the equation*

$$\mu\omega\lambda^{1/2} = \lambda + \omega - 1 \tag{8.42}$$

   then λ is an eigenvalue of $\boldsymbol{G}_\omega$.

Proof. Suppose $(\lambda, \boldsymbol{w})$ is an eigenpair for $\boldsymbol{G}_\omega$. We claim that $(\mu, \boldsymbol{v})$ is an eigenpair for $\boldsymbol{G}_J$, where μ is given by (8.41) and $\boldsymbol{v} = \vec{(\boldsymbol{V})}$ with $v_{i,j} := \lambda^{-(i+j)/2}w_{i,j}$. Indeed, replacing $w_{i,j}$ by $\lambda^{(i+j)/2}v_{i,j}$ in (8.40) 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

$$\boldsymbol{G}_J\boldsymbol{v} = (\boldsymbol{L} + \boldsymbol{R})\boldsymbol{v} = \frac{\lambda + \omega - 1}{\omega\lambda^{1/2}}\boldsymbol{v} = \mu\boldsymbol{v}.$$

For the converse let $(\mu, \boldsymbol{v})$ be an eigenpair for $\boldsymbol{G}_J$ and let et λ be a solution of (8.42). We define as before $\boldsymbol{v} =: \mathrm{vec}(\boldsymbol{V})$, $\boldsymbol{W} = \mathrm{vec}(\boldsymbol{W})$ with $w_{i,j} := \lambda^{(i+j)/2} v_{i,j}$. Inserting this in (8.38) and canceling $\lambda^{-(i+j)/2}$ we obtain

$$\frac{1}{4}(\lambda^{1/2} w_{i-1,j} + \lambda^{1/2} w_{i,j-1} + \lambda^{-1/2} w_{i+1,j} + \lambda^{-1/2} w_{i,j+1}) = \mu w_{i,j}.$$

Multiplying by $\omega\lambda^{1/2}$ we obtain

$$\frac{\omega}{4}(\lambda w_{i-1,j} + \lambda w_{i,j-1} + w_{i+1,j} + w_{i,j+1}) = \omega\mu\lambda^{1/2} w_{i,j},$$

Thus, if $\omega\mu^{1/2}\lambda^{1/2} = \lambda + \omega - 1$ then by (8.40) $(\lambda, \boldsymbol{w})$ is an eigenpair for $\boldsymbol{G}_\omega$. ∎

Proof of Theorem 8.20
Combining statement 1 and 2 in Theorem 8.34 we see that $\rho(\boldsymbol{G}_\omega) = |\lambda(\mu)|$, where $\lambda(\mu)$ is an eigenvalue of $\boldsymbol{G}_\omega$ satisfying (8.42) for some eigenvalue μ of $\boldsymbol{G}_J$. The eigenvalues of $\boldsymbol{G}_J$ are $\frac{1}{2}\cos(j\pi h) + \frac{1}{2}\cos(k\pi h)$, $j, k = 1, \ldots, m$, so μ is real and both μ and $-\mu$ are eigenvalues. Thus, to compute $\rho(\boldsymbol{G}_\omega)$ it is enough to consider (8.42) for a positive eigenvalue μ of $\boldsymbol{G}_J$. Solving (8.42) for $\lambda = \lambda(\mu)$ gives

$$\lambda(\mu) := \frac{1}{4}\Big(\omega\mu \pm \sqrt{(\omega\mu)^2 - 4(\omega - 1)}\Big)^2. \tag{8.43}$$

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 when $d(\omega) = 0$. The root in $(0, 2)$ is

$$\omega = \tilde{\omega}(\mu) := 2\frac{1 - \sqrt{1 - \mu^2}}{\mu^2} = \frac{2}{1 + \sqrt{1 - \mu^2}}. \tag{8.44}$$

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 (8.43) are positive and the larger one is

$$\lambda(\mu) = \frac{1}{4}\Big(\omega\mu + \sqrt{(\omega\mu)^2 - 4(\omega - 1)}\Big)^2, \quad 0 < \omega \le \tilde{\omega}(\mu). \tag{8.45}$$

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 (8.28) for $0 < \omega \leq \tilde{\omega}(\beta) = \omega^*$.

Evidently $\rho(\boldsymbol{G}_\omega) = \omega - 1$ is strictly increasing in $\omega^* < \omega < 2$. Equation (8.30) will follow if we can show that $\rho(\boldsymbol{G}_\omega)$ is strictly decreasing in $0 < \omega < \omega^*$. By differentiation

$$\frac{d}{d\omega}\left(\omega\beta + \sqrt{(\omega\beta)^2 - 4(\omega - 1)}\right) = \frac{\beta\sqrt{(\omega\beta)^2 - 4(\omega - 1)} + \omega\beta^2 - 2}{\sqrt{(\omega\beta)^2 - 4(\omega - 1)}}.$$

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.

8.6 Review Questions

8.6.1 Consider a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ with nonzero diagonal elements.

- Define the J and GS method in component form,
- Do they always converge?
- Give a necessary and sufficient condition that $\boldsymbol{A}^n \to \boldsymbol{0}$.
- Is there a matrix norm $\|\ \|$ consistent on $\mathbb{C}^{n\times n}$ such that $\|\boldsymbol{A}\| < \rho(\boldsymbol{A})$?

8.6.2 What is a Neumann series? when does it converge?

8.6.3 How do we define convergence of a fixed point iteration $\boldsymbol{x}_{k+1} = \boldsymbol{G}\boldsymbol{x}_k + \boldsymbol{c}$? When does it converge?

8.6.4 Define Richardson's method.

Chapter 9

The Conjugate Gradient Method


Magnus Rudolph Hestenes, 1906-1991 (left), Eduard L. Stiefel, 1909-1978 (right).

The **conjugate gradient method** was published by Hestenes and Stiefel in 1952, [11] as a direct method for solving linear systems. Today its main use is as an iterative method for solving large sparse linear systems. On a test problem we show that it performs as well as the SOR method with optimal acceleration parameter, and we do not have to estimate any such parameter. However the conjugate gradient method is restricted to symmetric positive definite systems. We also consider the mathematical formulation of the **preconditioned conjugate gradient method**. It is used to speed up convergence of the conjugate gradient method and we study this on a partial differential equation example.

The conjugate gradient method can also be used for minimization and is related to a method known as **steepest descent**. This method and the conjugate gradient method are both minimization methods and iterative methods for solving linear equations.

Throughout this chapter $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ will be a symmetric positive definite matrix. Thus, $\boldsymbol{A}^T = \boldsymbol{A}$ and $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y} > 0$ for all nonzero $\boldsymbol{y} \in \mathbb{R}^n$. We recall that $\boldsymbol{A}$ has positive eigenvalues and that the spectral (2-norm) condition number of $\boldsymbol{A}$ is given by $\kappa := \frac{\lambda_{max}}{\lambda_{min}}$, where λ_{max} and λ_{min} are the largest and smallest eigenvalue

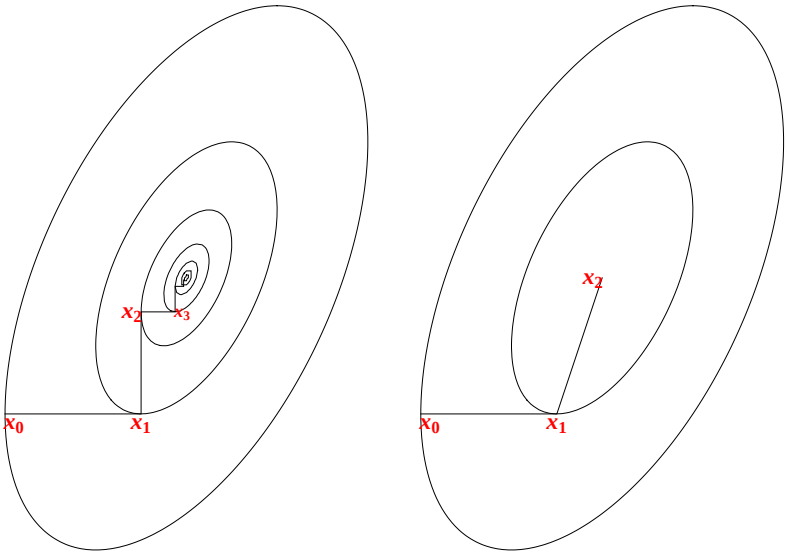


Figure 9.1. *Level curves for $Q(x, y)$ given by (9.2). Also shown is a steepest descent iteration (left) and a conjugate gradient iteration (right) to find the minimum of Q. (cf Examples 9.3,9.6)*
.

of $\boldsymbol{A}$.

9.1 Quadratic Minimization and Steepest Descent

We start by discussing some aspect of quadratic minimization and its relation to solving linear systems.

Consider for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$ the quadratic function $Q : \mathbb{R}^n \to \mathbb{R}$ given by

$$Q(\boldsymbol{y}) := \frac{1}{2}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{y}. \tag{9.1}$$

As an example, some level curves of

$$Q(x, y) := \frac{1}{2}\begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = x^2 - xy + y^2 \tag{9.2}$$

are shown in Figure 9.1. The level curves are ellipses and the graph of Q is a paraboloid (cf. Exercise 9.1).

Exercise 9.1 (Paraboloid)
Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ be the spectral decomposition of $\boldsymbol{A}$, i. e., $\boldsymbol{U}$ is orthonormal and $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal. Define new varables $\boldsymbol{v} = [v_1, \ldots, v_n]^T := \boldsymbol{U}^T\boldsymbol{y}$, and set $\boldsymbol{c} := \boldsymbol{U}^T\boldsymbol{b} = [c_1, \ldots, c_n]^T$. Show that

$$Q(\boldsymbol{y}) = \frac{1}{2}\sum_{j=1}^{n} \lambda_j v_j^2 - \sum_{j=1}^{n} c_j v_j.$$

Minimizing a quadratic function is equivalent to solving a linear system.

Lemma 9.2 (Quadratic function)
A vector $\boldsymbol{x} \in \mathbb{R}^n$ minimizes Q given by (9.1) if and only if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Moreover, the residual $\boldsymbol{r}(\boldsymbol{y}) := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}$ at any $\boldsymbol{y} \in \mathbb{R}^n$ is equal to the negative gradient, i. e., $\boldsymbol{r}(\boldsymbol{y}) = -\nabla Q(\boldsymbol{y})$, where $\nabla := \left[\frac{\partial}{\partial y_1}, \ldots, \frac{\partial}{\partial y_n}\right]^T$.

Proof. Expanding $Q(\boldsymbol{y} + \varepsilon\boldsymbol{h}) := \frac{1}{2}(\boldsymbol{y} + \varepsilon\boldsymbol{h})^T\boldsymbol{A}(\boldsymbol{y} + \varepsilon\boldsymbol{h}) - \boldsymbol{b}^T(\boldsymbol{y} + \varepsilon\boldsymbol{h})$ we find for any $\boldsymbol{y}, \boldsymbol{h} \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}$

$$Q(\boldsymbol{y} + \varepsilon\boldsymbol{h}) = Q(\boldsymbol{y}) - \varepsilon\boldsymbol{h}^T r(\boldsymbol{y}) + \frac{1}{2}\varepsilon^2\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}, \text{ where } r(\boldsymbol{y}) := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}. \tag{9.3}$$

If $\boldsymbol{y} = \boldsymbol{x}$, $\varepsilon = 1$, and $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ then (9.3) simplifies to $Q(\boldsymbol{x} + \boldsymbol{h}) = Q(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}$, and since $\boldsymbol{A}$ is symmetric positive definite $Q(\boldsymbol{x} + \boldsymbol{h}) > Q(\boldsymbol{x})$ for all nonzero $\boldsymbol{h} \in \mathbb{R}^n$. It follows that $\boldsymbol{x}$ is the unique minimum of Q. Conversely, if $\boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{b}$ and $\boldsymbol{h} := \boldsymbol{r}(\boldsymbol{x})$, then by (9.3), $Q(\boldsymbol{x} + \varepsilon\boldsymbol{h}) - Q(\boldsymbol{x}) = -\varepsilon(\boldsymbol{h}^T\boldsymbol{r}(x) - \frac{1}{2}\varepsilon\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}) < 0$ for $\varepsilon > 0$ sufficiently small. Thus $\boldsymbol{x}$ does not minimize Q. By (9.3) for $\boldsymbol{y} \in \mathbb{R}^n$

$$\begin{aligned}\frac{\partial}{\partial y_i}Q(\boldsymbol{y}) &:= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(Q(\boldsymbol{y} + \varepsilon\boldsymbol{e}_i) - Q(\boldsymbol{y})\right)\\ &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(-\varepsilon\boldsymbol{e}_i^T\boldsymbol{r}(\boldsymbol{y})) + \frac{1}{2}\varepsilon^2\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{e}_i\right) = -\boldsymbol{e}_i^T\boldsymbol{r}(\boldsymbol{y}), \quad i = 1, \ldots, n,\end{aligned}$$

showing that $\boldsymbol{r}(\boldsymbol{y}) = -\nabla Q(\boldsymbol{y})$. □

A general class of minimization algorithms for Q and solution algorithms for a linear system is given as follows:

1. Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$.
2. For $k = 0, 1, 2, \ldots$

Choose a "search direction" $\boldsymbol{p}_k$, Choose a "step length" α_k, Compute $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k$.

(9.4)

We would like to generate a sequence $\{\boldsymbol{x}_k\}$ that converges quickly to the minimum $\boldsymbol{x}$ of Q.

For a fixed direction $\boldsymbol{p}_k$ we say that α_k is **optimal** if $Q(\boldsymbol{x}_{k+1})$ is as small as possible, i.e.

$$Q(\boldsymbol{x}_{k+1}) = Q(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) = \min_{\alpha \in \mathbb{R}} Q(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k).$$

By (9.3) we have $Q(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) = Q(\boldsymbol{x}_k) - \alpha \boldsymbol{p}_k^T \boldsymbol{r}_k + \frac{1}{2}\alpha^2 \boldsymbol{p}_k^T \boldsymbol{A}\boldsymbol{p}_k$, where $\boldsymbol{r}_k := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$. Since $\boldsymbol{p}_k^T \boldsymbol{A}\boldsymbol{p}_k \geq 0$ we find a minimum α_k by solving $\frac{\partial}{\partial \alpha} Q(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) = 0$. It follows that the optimal α_k is uniquely given by

$$\alpha_k := \frac{\boldsymbol{p}_k^T \boldsymbol{r}_k}{\boldsymbol{p}_k^T \boldsymbol{A}\boldsymbol{p}_k}. \tag{9.5}$$

In the method of **Steepest descent**, also known as the **Gradient method** we choose $\boldsymbol{p}_k = \boldsymbol{r}_k$ the negative gradient, and the optimal α_k. Starting from $\boldsymbol{x}_0$ we compute for $k = 0, 1, 2 \ldots$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \big(\frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{r}_k^T \boldsymbol{A}\boldsymbol{r}_k}\big)\boldsymbol{r}_k. \tag{9.6}$$

This is similer to Richardson's method (8.19), but in that method we used a constant step length. Computationally, a step in the steepest descent iteration canbe organizedd as follows

$$\boxed{\begin{aligned} \boldsymbol{t}_k &= \boldsymbol{A}\boldsymbol{r}_k, \\ \alpha_k &= (\boldsymbol{r}_k^T \boldsymbol{r}_k)/(\boldsymbol{r}_k^T \boldsymbol{t}_k), \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k, \\ \boldsymbol{r}_{k+1} &= \boldsymbol{r}_k - \alpha_k \boldsymbol{t}_k. \end{aligned}} \tag{9.7}$$

Here, and in general, the following updating of the residual is used:

$$\boldsymbol{r}_{k+1} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{k+1} = \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) = \boldsymbol{r}_k - \alpha_k \boldsymbol{t}_k, \quad \boldsymbol{t}_k := \boldsymbol{A}\boldsymbol{p}_k. \tag{9.8}$$

Example 9.3 (Steepest descent iteration)
Suppose $Q(x, y)$ is given by (9.2). *Starting with $\boldsymbol{x}_0 = [-1, -1/2]^T$ and $\boldsymbol{r}_0 = -\boldsymbol{A}\boldsymbol{x}_0 = [3/2, 0]^T$ we find*

$$\boldsymbol{t}_0 = 3 \begin{bmatrix} 1 \\ -1/2 \end{bmatrix}, \quad \alpha_0 = \frac{1}{2}, \quad \boldsymbol{x}_1 = -4^{-1} \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \boldsymbol{r}_1 = 3 * 4^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\boldsymbol{t}_1 = 3 * 4^{-1} \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \quad \alpha_1 = \frac{1}{2}, \quad \boldsymbol{x}_2 = -4^{-1} \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}, \quad \boldsymbol{r}_2 = 3 * 4^{-1} \begin{bmatrix} 1/2 \\ 0 \end{bmatrix},$$

and in general for $k \geq 1$

$$\boldsymbol{t}_{2k-2} = 3 * 4^{1-k} \begin{bmatrix} 1 \\ -1/2 \end{bmatrix}, \quad \boldsymbol{x}_{2k-1} = -4^{-k} \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \boldsymbol{r}_{2k-1} = 3 * 4^{-k} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$
$$\boldsymbol{t}_{2k-1} = 3 * 4^{-k} \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \quad \boldsymbol{x}_{2k} = -4^{-k} \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}, \quad \boldsymbol{r}_{2k} = 3 * 4^{-k} \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}.$$

Since $\alpha_k = 1/2$ *is constant for all* k *the methods of Richardson, Jacobi and steepest descent are the same on this simple problem. See the left part of Figure 9.1. The rate of convergence is determined from* $\|\boldsymbol{x}_{j+1}\|_2/\|\boldsymbol{x}_j\| = \|\boldsymbol{r}_{j+1}\|_2/\|\boldsymbol{r}_j\|_2 = 1/2$ *for all* j*.*

Exercise 9.4 (Steepest descent iteration)
Verify the numbers in Example 9.3.

9.2 The Conjugate Gradient Method

In the steepest descent method the choice $\boldsymbol{p}_k = \boldsymbol{r}_k$ implies that the last two gradients are orthogonal. Indeed, by (9.8), $\boldsymbol{r}_{k+1}^T \boldsymbol{r}_k = (\boldsymbol{r}_k - \alpha_k \boldsymbol{A}\boldsymbol{r}_k)^T \boldsymbol{p}_k = 0$ since $\alpha_k = \frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{p}_k^T \boldsymbol{A}\boldsymbol{p}_k}$ and $\boldsymbol{A}$ is symmetric. In the conjugate gradient method all gradients are orthogonal[11]. We achieve this by using **$\boldsymbol{A}$-orthogonal search directions** i. e., $\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_j = 0$ for all $i \neq j$.

9.2.1 Derivation of the method

As in the steepest descent method we choose a starting vector $\boldsymbol{x}_0 \in \mathbb{R}^n$. If $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{0}$ then $\boldsymbol{x}_0$ is the exact solution and we are finished, otherwise we initially make a steepest descent step. It follows that $\boldsymbol{r}_1^T \boldsymbol{r}_0 = 0$ and $\boldsymbol{p}_0 := \boldsymbol{r}_0$.

For the general case we define for $j \geq 0$

$$\boldsymbol{p}_j := \boldsymbol{r}_j - \sum_{i=0}^{j-1} \left(\frac{\boldsymbol{r}_j^T \boldsymbol{A}\boldsymbol{p}_i}{\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_i}\right) \boldsymbol{p}_i, \tag{9.9}$$

$$\boldsymbol{x}_{j+1} := \boldsymbol{x}_j + \alpha_j \boldsymbol{p}_j \quad \alpha_j := \frac{\boldsymbol{r}_j^T \boldsymbol{r}_j}{\boldsymbol{p}_j^T \boldsymbol{A}\boldsymbol{p}_j}, \tag{9.10}$$

$$\boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \alpha_j \boldsymbol{A}\boldsymbol{p}_j. \tag{9.11}$$

We note that

1. $\boldsymbol{p}_j$ is computed by the Gram-Schmidt orthogonalization process applied to the residuals $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_j$ using the $\boldsymbol{A}$-inner product. The search directions are therefore $\boldsymbol{A}$-orthogonal and nonzero as long as the residuals are linearly independent.

[11] It is this property that has given the method its name.

2. Equation (9.11) follows from (9.8).

3. It can be shown that the step length α_j is optimal for all j (cf. Exercise 9.9)).

Lemma 9.5 (The residuals are orthogonal)
Suppose that for some $k \geq 0$ that $\boldsymbol{x}_j$ is well defined, $\boldsymbol{r}_j \neq 0$, and $\boldsymbol{r}_i^T \boldsymbol{r}_j = 0$ for $i, j = 0, 1, \ldots, k$, $i \neq j$. Then $\boldsymbol{x}_{k+1}$ is well defined and $\boldsymbol{r}_{k+1}^T \boldsymbol{r}_j = 0$ for $j = 0, 1, \ldots, k$.

Proof. Since the residuals $\boldsymbol{r}_j$ are orthogonal and nonzero for $j \leq k$, they are linearly independent, and it follows form the Gram-Schmidt Theorem 0.29 that $\boldsymbol{p}_k$ is nonzero and $\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_i = 0$ for $i < k$. But then $\boldsymbol{x}_{k+1}$ and $\boldsymbol{r}_{k+1}$ are well defined. Now

$$\begin{aligned}
\boldsymbol{r}_{k+1}^T \boldsymbol{r}_j &\overset{(9.11)}{=} (\boldsymbol{r}_k - \alpha_k \boldsymbol{A} \boldsymbol{p}_k)^T \boldsymbol{r}_j \\
&\overset{(9.9)}{=} \boldsymbol{r}_k^T \boldsymbol{r}_j - \alpha_k \boldsymbol{p}_k^T \boldsymbol{A} \big(\boldsymbol{p}_j + \sum_{i=0}^{j-1} (\frac{\boldsymbol{r}_j^T \boldsymbol{A} \boldsymbol{p}_i}{\boldsymbol{p}_i^T \boldsymbol{A} \boldsymbol{p}_i}) \boldsymbol{p}_i \big) \\
&\overset{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_i = 0}{=} \boldsymbol{r}_k^T \boldsymbol{r}_j - \alpha_k \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_j = 0, \quad j = 0, 1, \ldots, k.
\end{aligned}$$

That the final expression is equal to zero follows by orthogonality and $\boldsymbol{A}$-ortogonality for $j < k$ and by the definition of α_k for $j = k$. This completes the proof. ∎

The expression (9.9) for $\boldsymbol{p}_k$ can be greatly simplified. All terms except the last one vanish, since by orthogonality of the residuals

$$\boldsymbol{r}_j^T \boldsymbol{A} \boldsymbol{p}_i \overset{(9.11)}{=} \boldsymbol{r}_j^T \big(\frac{\boldsymbol{r}_i - \boldsymbol{r}_{i+1}}{\alpha_i} \big) = 0, \quad i = 0, 1, \ldots, j-2.$$

For the last term with $k = j - 1$

$$\beta_k := -\frac{\boldsymbol{r}_{k+1}^T \boldsymbol{A} \boldsymbol{p}_k}{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k} \overset{(9.11)}{=} \frac{\boldsymbol{r}_{k+1}^T (\boldsymbol{r}_{k+1} - \boldsymbol{r}_k)}{\alpha_k \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k} \overset{(9.10)}{=} \frac{\boldsymbol{r}_{k+1}^T \boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^T \boldsymbol{r}_k}. \tag{9.12}$$

To summarize, in the **conjugate gradient method** we start with $\boldsymbol{x}_0$, $\boldsymbol{p}_0 = \boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$ and then generate a sequence of vectors $\{\boldsymbol{x}_k\}$ as follows:

For $k = 0, 1, 2, \ldots$

$$\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \quad \alpha_k := \frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k}, \tag{9.13}$$

$$\boldsymbol{r}_{k+1} := \boldsymbol{r}_k - \alpha_k \boldsymbol{A} \boldsymbol{p}_k, \tag{9.14}$$

$$\boldsymbol{p}_{k+1} := \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k, \quad \beta_k := \frac{\boldsymbol{r}_{k+1}^T \boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^T \boldsymbol{r}_k}. \tag{9.15}$$

The residuals and search directions are orthogonal and $\boldsymbol{A}$-orthogonal, respectively.

The conjugate gradient method is also a direct method. Since $\dim \mathbb{R}^n = n$ the $n+1$ residuals $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_n$ cannot all be nonzero and for orthogonal residuals we find the exact solution in at most n iterations.

For computation we organize the iterations as follows for $k = 0, 1, 2, \ldots$

$$\boxed{\begin{aligned} \boldsymbol{t}_k &= \boldsymbol{A}\boldsymbol{p}_k, \\ \alpha_k &= (\boldsymbol{r}_k^T \boldsymbol{r}_k)/(\boldsymbol{p}_k^T \boldsymbol{t}_k), \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \\ \boldsymbol{r}_{k+1} &= \boldsymbol{r}_k - \alpha_k \boldsymbol{t}_k, \\ \beta_k &= (\boldsymbol{r}_{k+1}^T \boldsymbol{r}_{k+1})/(\boldsymbol{r}_k^T \boldsymbol{r}_k), \\ \boldsymbol{p}_{k+1} &:= \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k. \end{aligned}} \tag{9.16}$$

Example 9.6 (Conjugate gradient iteration)
Consider (9.16) *applied to the positive definite linear system* $\left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right]\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$. *Starting as in Example 9.3 with* $\boldsymbol{x}_0 = \begin{bmatrix} -1 \\ -1/2 \end{bmatrix}$ *we find* $\boldsymbol{p}_0 = \boldsymbol{r}_0 = \left[\begin{smallmatrix} 3/2 \\ 0 \end{smallmatrix}\right]$ *and then*

$$\boldsymbol{t}_0 = \left[\begin{smallmatrix} 3 \\ -3/2 \end{smallmatrix}\right], \quad \alpha_0 = 1/2, \quad \boldsymbol{x}_1 = \begin{bmatrix} -1/4 \\ -1/2 \end{bmatrix}, \quad \boldsymbol{r}_1 = \left[\begin{smallmatrix} 0 \\ 3/4 \end{smallmatrix}\right], \quad \beta_0 = 1/4, \quad \boldsymbol{p}_1 = \begin{bmatrix} 3/8 \\ 3/4 \end{bmatrix},$$
$$\boldsymbol{t}_1 = \left[\begin{smallmatrix} 0 \\ 9/8 \end{smallmatrix}\right], \quad \alpha_1 = 2/3, \quad \boldsymbol{x}_2 = \boldsymbol{0}, \quad \boldsymbol{r}_2 = \boldsymbol{0}.$$

Thus $\boldsymbol{x}_2$ *is the exact solution as illustrated in the right part of Figure 9.1.*

Exercise 9.7 (Conjugate gradient iteration, II)
Do one iteration with the conjugate gradient method when $\boldsymbol{x}_0 = \boldsymbol{0}$. *(Answer:* $\boldsymbol{x}_1 = \left(\frac{\boldsymbol{b}^T\boldsymbol{b}}{\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}}\right)\boldsymbol{b}$.)

Exercise 9.8 (Conjugate gradient iteration, III)
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 $\boldsymbol{x}_0 = \boldsymbol{0}$.

Exercise 9.9 (The cg step length is optimal)
Show that the step length α_k *in the conjugate gradient method is optimal*[12].

[12]Hint: use induction on k to show that $\boldsymbol{p}_k = \boldsymbol{r}_k + \sum_{j=0}^{k-1} a_{k,j}\boldsymbol{r}_j$ for some constants $a_{k,j}$.

Exercise 9.10 (Starting value in cg)
Show that the conjugate gradient method (9.16) *for* $\boldsymbol{Ax} = \boldsymbol{b}$ *starting with* $\boldsymbol{x}_0$ *is the same as applying the method to the system* $\boldsymbol{Ay} = \boldsymbol{r}_0 := \boldsymbol{b} - \boldsymbol{Ax}_0$ *starting with* $\boldsymbol{y}_0 = \boldsymbol{0}$.[13]

9.2.2 The conjugate gradient algorithm

In this section we give numerical examples and discuss implementation.

The formulas in (9.16) form a basis for an algorithm.

Algorithm 9.11 (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 $||\boldsymbol{r}_k||_2/||\boldsymbol{b}||_2 \leq$ tol or $k >$ itmax. K is the number of iterations used.

```
function [x,K]=cg(A,b,x,tol,itmax)
r=b-A*x; p=r;   rho0=b'*b; rho=r'*r;
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 (($\boldsymbol{p}^T\boldsymbol{t}$ and $\boldsymbol{r}^T\boldsymbol{r}$),
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 $\boldsymbol{t} = \boldsymbol{Ap}$.

9.2.3 Numerical example

We test the conjugate gradient method on two examples. For a similar test for the steepest descent method see Exercise 9.17. Consider the matrix given by the

[13] Hint: The conjugate gradient method for $\boldsymbol{Ay} = \boldsymbol{r}_0$ can be written $\boldsymbol{y}_{k+1} := \boldsymbol{y}_k + \gamma_k\boldsymbol{q}_k$, $\gamma_k := \frac{\boldsymbol{s}_k^T\boldsymbol{s}_k}{\boldsymbol{q}_k^T\boldsymbol{Aq}_k}$, $\boldsymbol{s}_{k+1} := \boldsymbol{s}_k - \gamma_k\boldsymbol{Aq}_k$, $\boldsymbol{q}_{k+1} := \boldsymbol{s}_{k+1} + \delta_k\boldsymbol{q}_k$, $\delta_k := \frac{\boldsymbol{s}_{k+1}^T\boldsymbol{s}_{k+1}}{\boldsymbol{s}_k^T\boldsymbol{s}_k}$. Show that $\boldsymbol{y}_k = \boldsymbol{x}_k - \boldsymbol{x}_0$, $\boldsymbol{s}_k = \boldsymbol{r}_k$, and $\boldsymbol{q}_k = \boldsymbol{p}_k$, for $k = 0, 1, 2 \ldots$.

n	2 500	10 000	40 000	1 000 000	4 000 000
K	19	18	18	16	15

Table 9.12. *The number of iterations K for the averaging problem on a $\sqrt{n} \times \sqrt{n}$ grid for various n*

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 $\boldsymbol{f} = [1, \ldots, 1]^T \in \mathbb{R}^n$.

We consider two problems.

1. $a = 1/9$, $d = 5/18$, the Averaging matrix.

2. $a = -1$, $d = 2$, the Poisson matrix.

9.2.4 Implementation issues

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 9.11 in $O(n)$ arithmetic operations. With such an implementation the total number of arithmetic operations 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 $\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{P}, \boldsymbol{B}, \boldsymbol{T} \in \mathbb{R}^{m \times m}$ 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 $\boldsymbol{T}_2\boldsymbol{x} = h^2\boldsymbol{f} \iff \boldsymbol{T}_1\boldsymbol{V} + \boldsymbol{V}\boldsymbol{T}_1 = \boldsymbol{B}$, and $\boldsymbol{t} = \boldsymbol{T}_2\boldsymbol{p} \iff \boldsymbol{T} = \boldsymbol{T}_1\boldsymbol{P} + \boldsymbol{P}\boldsymbol{T}_1$.

This leads to the following algorithm for testing the conjugate gradient algorithm on the matrix

$$\boldsymbol{A} = \text{tridiag}_m(a, d, a) \otimes \boldsymbol{I}_m + \boldsymbol{I}_m \otimes \text{tridiag}_m(a, d, a) \in \mathbb{R}^{(m^2) \times (m^2)}.$$

n	2 500	10 000	40 000	160 000
K	94	188	370	735
$K/\sqrt{n}$	1.88	1.88	1.85	1.84

Table 9.14. *The number of iterations K for the Poisson problem on a $\sqrt{n} \times \sqrt{n}$ grid for various n*

Algorithm 9.13 (Testing conjugate gradient)

```
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;
```

For both the averaging- and Poison matrix we use $tol = 10^{-8}$.

For the averaging matrix we obtain the values in Table 9.12.

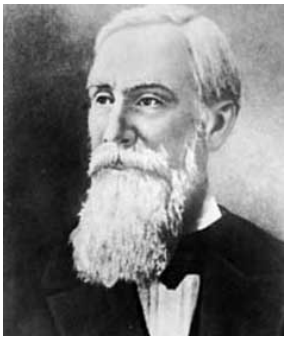
The convergence is quite rapid. It appears that the number of iterations can be bounded independently of n, and therefore we solve the problem in $O(n)$ operations. This is the best we can do for a problem with n unknowns.

Consider next the Poisson problem. In in Table 9.14 we list K, the required number of iterations, and $K/\sqrt{n}$.

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.

9.3 Convergence





Leonid Vitaliyevich Kantorovich, 1912-1986 (left), Aleksey Nikolaevich Krylov, 1863-1945 (center), Pafnuty Lvovich Chebyshev, 1821-1894 (right)

9.3.1 The A-norm

The convergence analysis for both steepest descent and conjugate gradients is in terms of a special inner product. We define the $\boldsymbol{A}$-inner product and the corresponding $\boldsymbol{A}$-norm by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{y}, \quad \|\boldsymbol{y}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \tag{9.17}$$

Exercise 9.15 (The A-inner product)
Show that if $\boldsymbol{A}$ is symmetric positive definite then the $\boldsymbol{A}$-inner product is indeed an inner product.

9.3.2 The Main Theorem

The following theorem gives upper bounds for the $\boldsymbol{A}$-norm of the error in both methods.

Theorem 9.16 (Error bound for steepest descent and conjugate gradients)

Suppose $\boldsymbol{A}$ is symmetric positive definite. For the $\boldsymbol{A}$-norms of the errors in the steepest descent method (9.6) *the following upper bounds hold*

$$\frac{||\boldsymbol{x} - \boldsymbol{x}_k||_{\boldsymbol{A}}}{||\boldsymbol{x} - \boldsymbol{x}_0||_{\boldsymbol{A}}} \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k < e^{-\frac{2}{\kappa} k}, \quad , k > 0, \tag{9.18}$$

while for the conjugate gradient method we have

$$\frac{||\boldsymbol{x} - \boldsymbol{x}_k||_{\boldsymbol{A}}}{||\boldsymbol{x} - \boldsymbol{x}_0||_{\boldsymbol{A}}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k < 2e^{-\frac{2}{\sqrt{\kappa}} k}, \quad k \geq 0. \tag{9.19}$$

Here $\kappa = cond_2(\boldsymbol{A}) := \lambda_{max}/\lambda_{min}$ *is the spectral condition number of* $\boldsymbol{A}$*, and* λ_{max} *and* λ_{min} *are the largest and smallest eigenvalue of* $\boldsymbol{A}$*, respectively.*

Theorem 9.16 implies

1. Since $\frac{\kappa-1}{\kappa+1} < 1$ the steepest descent method always converges for a symmetric positive definite matrix. The convergence can be slow when $\frac{\kappa-1}{\kappa+1}$ is close to one, and this happens even for a moderately ill-conditioned $\boldsymbol{A}$.

2. The rate of convergence for the conjugate gradient method appears to be determined by the square root of the spectral condition number. This is is much better than the estimate for the steepest descent method. Especially for problems with large condition numbers.

3. The proofs of the estimates in (9.18) and (9.19) are quite different. This is in spite of their similar appearance.

9.3.3 The number of iterations for the model problems

Consider the test matrix

$$\boldsymbol{T}_2 := \text{tridiag}_m(a, d, a) \otimes \boldsymbol{I}_m + \boldsymbol{I}_m \otimes \text{tridiag}_m(a, d, a) \in \mathbb{R}^{(m^2)\times(m^2)}.$$

The eigenvalues were given in (3.20) as

$$\lambda_{j,k} = 2d + 2a\cos(j\pi h) + 2a\cos(k\pi h), \quad j, k = 1, \ldots, m. \tag{9.20}$$

For the averaging problem given by $d = 5/18$, $a = 1/9$, the largest and smallest eigenvalue of $\boldsymbol{T}_2$ are given by $\lambda_{max} = \frac{5}{9} + \frac{4}{9}\cos(\pi h)$ and $\lambda_{min} = \frac{5}{9} - \frac{4}{9}\cos(\pi h)$. Thus

$$\kappa_A = \frac{5 + 4\cos(\pi h)}{5 - 4\cos(\pi h)} \leq 9,$$

and the condition number is bounded independently of n. It follows from (9.19) that the number of iterations can be bounded independently of the size n of the problem, and this is in agreement with what we observed in Table 9.12.

For the Poisson problem we have by (8.25) the condition number

$$\kappa_P = \frac{\lambda_{max}}{\lambda_{min}} = \frac{\cos^2(\pi h/2)}{\sin^2(\pi h/2)} \text{ and } \sqrt{\kappa_P} = \frac{\cos(\pi h/2)}{\sin(\pi h/2)} \approx \frac{2}{\pi h} \approx \frac{2}{\pi}\sqrt{n}.$$

Thus, (see also Exercise 7.33) we solve the discrete Poisson problem in $O(n^{3/2})$ arithmetic operations using the conjugate gradient method. This is the same as for the SOR method and for the fast method without the FFT. In comparison the Cholesky Algorithm requires $O(n^2)$ arithmetic operations both for the averaging and the Poisson problem.

Exercise 9.17 (Program code for testing steepest descent)
Write a function K=sdtest(m,a,d,tol,itmax) to test the Steepest descent method on the matrix $\boldsymbol{T}_2$. Make the analogues of Table 9.12 and Table 9.14. For Table 9.14 it is enough to test for say $n = 100, 400, 1600, 2500$, and tabulate K/n instead of $K/\sqrt{n}$ in the last row. Conclude that the upper bound (9.18) *is realistic. Compare also with the number of iterations for the J and GS method in Table 8.1.*

Exercise 9.18 (Using cg to solve normal equations)
Consider solving the linear system $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$ by using the conjugate gradient method. Here $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{A}^T\boldsymbol{A}$ is positive definite[14]*. Explain why only the following modifications in Algorithm 9.11 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 $\boldsymbol{A}$.

9.4 Proof of the Convergence Estimates

9.4.1 Convergence proof for steepest descent

For the proof of (9.18) the following inequality will be used.

Theorem 9.19 (Kantorovich inequality)
For any symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$

$$1 \le \frac{(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y})(\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{y})}{(\boldsymbol{y}^T\boldsymbol{y})^2} \le \frac{(M+m)^2}{4Mm} \quad \boldsymbol{y} \neq \boldsymbol{0},\ \boldsymbol{y} \in \mathbb{R}^n, \tag{9.21}$$

where $M := \lambda_{max}$ and $m := \lambda_{min}$ are the largest and smallest eigenvalue of $\boldsymbol{A}$, respectively.

Proof. For $j = 1, \ldots, n$ let $(\lambda_j, \boldsymbol{u}_j)$ be orthonormal eigenpairs of $\boldsymbol{A}$ and $\boldsymbol{y} \in \mathbb{R}^n$. By Theorem 5.1 $(\lambda_j^{-1}, \boldsymbol{u}_j)$ are eigenpairs for $\boldsymbol{A}^{-1}$. Let $\boldsymbol{y} = \sum_{j=1}^n c_j\boldsymbol{u}_j$ be the corresponding eigenvector expansion of $\boldsymbol{y}$. By orthonormality, (cf. (5.6))

$$a := \frac{\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \sum_{i=1}^n t_i\lambda_i, \quad b := \frac{\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} = \sum_{i=1}^n \frac{t_i}{\lambda_i}, \tag{9.22}$$

[14]This system known as the **normal equations** appears in linear least squares problems and will be considered in this context in Chapter 11.

where

$$t_i = \frac{c_i^2}{\sum_{j=1}^n c_j^2} \geq 0, \quad i = 1, \ldots, n \text{ and } \sum_{i=1}^n t_i = 1. \tag{9.23}$$

Thus a and b are **convex combinations** of the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$, respectively. Let c be a positive constant to be chosen later. By the geometric/arithmetic mean inequality (7.27) and (9.22)

$$\sqrt{ab} = \sqrt{(ac)(b/c)} \leq (ac + b/c)/2 = \frac{1}{2}\sum_{i=1}^n t_i\big(\lambda_i c + 1/(\lambda_i c)\big) = \frac{1}{2}\sum_{i=1}^n t_i f(\lambda_i c),$$

where $f : [mc, Mc] \to \mathbb{R}$ is given by $f(x) := x + 1/x$. By (9.23)

$$\sqrt{ab} \leq \frac{1}{2}\max_{mc \leq x \leq Mc} f(x).$$

Since $f \in C^2$ and f'' is positive it follows from Lemma 7.36 that f is a convex function. But a convex function takes it maximum at one of the endpoints of the range (cf. Exercise 9.20) and we obtain

$$\sqrt{ab} \leq \frac{1}{2}\max\{f(mc), f(Mc)\}. \tag{9.24}$$

Choosing $c := 1/\sqrt{mM}$ we find $f(mc) = f(Mc) = \sqrt{\frac{M}{m}} + \sqrt{\frac{m}{M}} = \frac{M+m}{\sqrt{mM}}$. By (9.24) we obtain

$$\frac{(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y})(\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{y})}{(\boldsymbol{y}^T\boldsymbol{y})^2} = ab \leq \frac{(M+m)^2}{4Mm},$$

the upper bound in (9.21). For the lower bound we use the Cauchy-Schwarz inequality as follows

$$1 = \big(\sum_{i=1}^n t_i\big)^2 = \left(\sum_{i=1}^n (t_i\lambda_i)^{1/2}(t_i/\lambda_i)^{1/2}\right)^2 \leq \big(\sum_{i=1}^n t_i\lambda_i\big)\big(\sum_{i=1}^n t_i/\lambda_i\big) = ab.$$

□

Exercise 9.20 (Maximum of a convex function)
Show that if $f : [a, b] \to \mathbb{R}$ *is convex then* $\max_{a \leq x \leq b} f(x) \leq \max\{f(a), f(b)\}$.

Proof of (9.18)
Let $\boldsymbol{\epsilon}_j := \boldsymbol{x} - \boldsymbol{x}_j$, $j = 0, 1, \ldots$, where $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. It is enough to show that

$$\frac{\|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2}{\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2, \quad k = 0, 1, 2, \ldots, \tag{9.25}$$

for then $\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}} \le \left(\frac{\kappa-1}{\kappa+1}\right)\|\boldsymbol{\epsilon}_{k-1}\| \le \cdots \le \left(\frac{\kappa-1}{\kappa+1}\right)^k \|\boldsymbol{\epsilon}_0\|$. It follows from (9.6) that

$$\boldsymbol{\epsilon}_{k+1} = \boldsymbol{\epsilon}_k - \alpha_k \boldsymbol{r}_k, \quad \alpha_k := \frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{r}_k^T \boldsymbol{A}\boldsymbol{r}_k}.$$

We find

$$\begin{aligned}
\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2 &= \boldsymbol{\epsilon}_k^T \boldsymbol{A}\boldsymbol{\epsilon}_k = \boldsymbol{r}_k^T \boldsymbol{A}^{-1}\boldsymbol{r}_k, \\
\|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2 &= (\boldsymbol{\epsilon}_k - \alpha_k \boldsymbol{r}_k)^T \boldsymbol{A}(\boldsymbol{\epsilon}_k - \alpha_k \boldsymbol{r}_k) \\
&= \boldsymbol{\epsilon}_k^T \boldsymbol{A}\boldsymbol{\epsilon}_k - 2\alpha_k \boldsymbol{r}_k^T \boldsymbol{A}\boldsymbol{\epsilon}_k + \alpha_k^2 \boldsymbol{r}_k^T \boldsymbol{A}\boldsymbol{r}_k = \|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2 - \frac{(\boldsymbol{r}_k^T \boldsymbol{r}_k)^2}{\boldsymbol{r}_k^T \boldsymbol{A}\boldsymbol{r}_k}.
\end{aligned}$$

Combining these and using Kantorovich inequality

$$\frac{\|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2}{\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2} = 1 - \frac{(\boldsymbol{r}_k^T \boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T \boldsymbol{A}\boldsymbol{r}_k)(\boldsymbol{r}_k^T \boldsymbol{A}^{-1}\boldsymbol{r}_k)} \le 1 - \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{min}+\lambda_{max})^2} = \left(\frac{\kappa-1}{\kappa+1}\right)^2$$

and (9.25) is proved.

The inequality

$$\frac{x-1}{x+1} < e^{-2/x} \quad \text{for} \quad x > 1 \tag{9.26}$$

follows from the familiar series expansion of the exponential function. Indeed, with $y = 1/x$, using $2^k/k! = 2$, $k = 1, 2$, and $2^k/k! < 2$ for $k > 2$, 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 (9.26) follows. □

9.4.2 Krylov spaces and the best approximatetion property

For the convergence analysis of the conjugate gradient method certain subspaces of of $\mathbb{R}^n$ called **Krylov spaces** play a central role. In fact the iterates in the conjugate gradient method are best approximation of the solution from these subspaces using the $\boldsymbol{A}$-norm to measure the error.

The Krylov spaces are defined by $\mathbb{W}_0 = \{\boldsymbol{0}\}$ and

$$\mathbb{W}_k = \operatorname{span}(\boldsymbol{r}_0, \boldsymbol{A}\boldsymbol{r}_0, \boldsymbol{A}^2\boldsymbol{r}_0, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{r}_0), \quad k = 1, 2, 3, \cdots.$$

They are nested subspaces

$$\mathbb{W}_0 \subset \mathbb{W}_1 \subset \mathbb{W}_2 \subset \cdots \subset \mathbb{W}_n \subset \mathbb{R}^n$$

with $\dim(\mathbb{W}_k) \le k$ for all $k \ge 0$. Moreover, If $\boldsymbol{v} \in \mathbb{W}_k$ then $\boldsymbol{A}\boldsymbol{v} \in \mathbb{W}_{k+1}$.

Lemma 9.21 (Krylov space)
For the iterates in the conjugate gradient method we have

$$\boldsymbol{x}_k - \boldsymbol{x}_0 \in \mathbb{W}_k, \quad \boldsymbol{r}_k, \boldsymbol{p}_k \in \mathbb{W}_{k+1}, \quad k = 0, 1, \ldots, \tag{9.27}$$

and

$$\boldsymbol{r}_k^T \boldsymbol{w} = \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{w} = 0, \quad \boldsymbol{w} \in \mathbb{W}_k. \tag{9.28}$$

Proof. (9.27) clearly holds for $k = 0$ since $\boldsymbol{p}_0 = \boldsymbol{r}_0$. Suppose it holds for some $k \geq 0$. Then $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k \boldsymbol{A}\boldsymbol{p}_k \in \mathbb{W}_{k+2}$ and by $\boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k \in \mathbb{W}_{k+2}$ and $\boldsymbol{x}_{k+1} - \boldsymbol{x}_0 \overset{(9.10)}{=} \boldsymbol{x}_k - \boldsymbol{x}_0 + \alpha_k \boldsymbol{p}_k \in \mathbb{W}_{k+1}$. Thus (9.27) follows by induction. 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 $\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{k-1}\}$, (9.28) follows. ∎

Theorem 9.22 (Best approximation property)
Suppose $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric positive definite and $\{\boldsymbol{x}_k\}$ is generated by the conjugate gradient method (cf. (9.13)*). Then*

$$\|\boldsymbol{x} - \boldsymbol{x}_k\|_{\boldsymbol{A}} = \min_{\boldsymbol{w} \in \mathbb{W}_k} \|\boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{w}\|_{\boldsymbol{A}}. \tag{9.29}$$

Proof. Fix k, let $\boldsymbol{w} \in \mathbb{W}_k$ and $\boldsymbol{u} := \boldsymbol{x}_k - \boldsymbol{x}_0 - \boldsymbol{w},$. By (9.27) $\boldsymbol{u} \in \mathbb{W}_k$ and then (9.28) implies that $\boldsymbol{r}_k^T \boldsymbol{u} = 0$. Since $(\boldsymbol{x} - \boldsymbol{x}_k)^T \boldsymbol{A}\boldsymbol{u} = \boldsymbol{r}_k^T \boldsymbol{u}$ we find

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{w}\|_{\boldsymbol{A}}^2 &= (\boldsymbol{x} - \boldsymbol{x}_k + \boldsymbol{u})^T \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k + \boldsymbol{u}) \\ &= (\boldsymbol{x} - \boldsymbol{x}_k)\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_k) + 2\boldsymbol{r}_k^T \boldsymbol{u} + \boldsymbol{u}^T \boldsymbol{A}\boldsymbol{u} \\ &= \|\boldsymbol{x} - \boldsymbol{x}_k\|_{\boldsymbol{A}}^2 + \|\boldsymbol{u}\|_{\boldsymbol{A}}^2 \geq \|\boldsymbol{x} - \boldsymbol{x}_k\|_{\boldsymbol{A}}^2. \end{aligned}$$

Taking square roots the result follows. ∎

If $\boldsymbol{x}_0 = \boldsymbol{0}$ then (9.29) says that $\boldsymbol{x}_k$ is the element in $\mathbb{W}_k$ that is closest to the solution $\boldsymbol{x}$ in the $\boldsymbol{A}$-norm. More generally, if $\boldsymbol{x}_0 \neq \boldsymbol{0}$ then $\boldsymbol{x} - \boldsymbol{x}_k = (\boldsymbol{x} - \boldsymbol{x}_0) - (\boldsymbol{x}_k - \boldsymbol{x}_0)$ and $\boldsymbol{x}_k - \boldsymbol{x}_0$ is the element in $\mathbb{W}_k$ that is closest to $\boldsymbol{x} - \boldsymbol{x}_0$ in the $\boldsymbol{A}$-norm. This is the orthogonal projection of $\boldsymbol{x} - \boldsymbol{x}_0$ into $\mathbb{W}_k$, see Figure 9.2.

Recall that to each polynomial $p(t) := \sum_{j=0}^m a_j t^m$ there corresponds a matrix polynomial $p(\boldsymbol{A}) := a_0 \boldsymbol{I} + a_1 \boldsymbol{A} + \cdots + a_m \boldsymbol{A}^m$. Moreover, if $(\lambda_j, \boldsymbol{u}_j)$ are eigenpairs of $\boldsymbol{A}$ then $(p(\lambda_j), \boldsymbol{u}_j)$ are eigenpairs of $p(\boldsymbol{A})$ for $j = 1, \ldots, n$.

Lemma 9.23 (Krylov space and polynomials)
Suppose $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric positive definite with orthonormal eigenpairs $(\lambda_j, \boldsymbol{u}_j)$, $j = 1, 2, \ldots, n$, and let $\boldsymbol{r}_0 := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$ for some $\boldsymbol{x}_0 \in \mathbb{R}^n$.

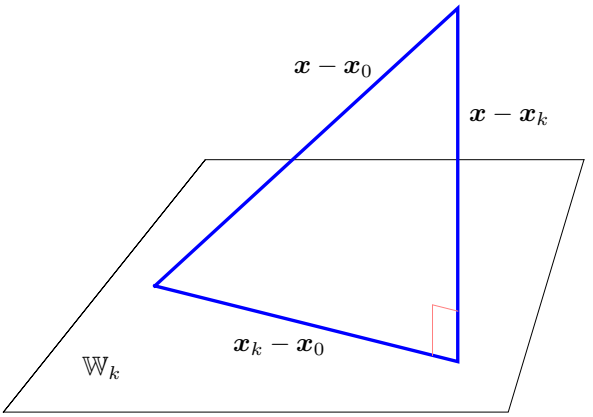


Figure 9.2. *The orthogonal projection of* $\boldsymbol{x} - \boldsymbol{x}_0$ *into* $\mathbb{W}_k$.

To each $\boldsymbol{w} \in \mathbb{W}_k$ *there corresponds a polynomial* $P(t) := \sum_{j=0}^{k-1} a_j t^{k-1}$ *such that* $\boldsymbol{w} = P(\boldsymbol{A})\boldsymbol{r}_0$. *Moreover, if* $\boldsymbol{r}_0 = \sum_{j=1}^{n} \sigma_j \boldsymbol{u}_j$ *then*

$$||\boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{w}||_{\boldsymbol{A}}^2 = \sum_{j=1}^{n} \frac{\sigma_j^2}{\lambda_j} Q(\lambda_j)^2, \quad Q(t) := 1 - tP(t). \tag{9.30}$$

Proof. If $\boldsymbol{w} \in \mathbb{W}_k$ then $\boldsymbol{w} = a_0\boldsymbol{r}_0 + a_1\boldsymbol{A}\boldsymbol{r}_0 + \cdots + a_{k-1}\boldsymbol{A}^{k-1}\boldsymbol{r}_0$ for some scalars $a_0, \ldots, a_{k-1}$. But then $\boldsymbol{w} = P(\boldsymbol{A})\boldsymbol{r}_0$. We find

$$\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{w}) = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0 - P(\boldsymbol{A})\boldsymbol{r}_0) = \boldsymbol{r}_0 - \boldsymbol{A}P(\boldsymbol{A})\boldsymbol{r}_0 = Q(\boldsymbol{A})\boldsymbol{r}_0,$$

and so $\|\boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{w}\|_{\boldsymbol{A}}^2 = \boldsymbol{c}^T\boldsymbol{A}^{-1}\boldsymbol{c}$, where $\boldsymbol{c} = Q(\boldsymbol{A})\boldsymbol{r}_0$. Using the eigenvector expansion for $\boldsymbol{r}_0$ we obtain

$$\boldsymbol{c} = \sum_{j=1}^{n} \sigma_j Q(\lambda_j)\boldsymbol{u}_j, \quad \boldsymbol{A}^{-1}\boldsymbol{c} = \sum_{i=1}^{n} \sigma_i \frac{Q(\lambda_i)}{\lambda_i}\boldsymbol{u}_i. \tag{9.31}$$

Now (9.30) follows by the orthonormality of the eigenvectors. □

We will use the following theorem to estimate the rate of convergence.

Theorem 9.24 (cg and best polynomial approximation)
Suppose $[a, b]$ *with* $0 < a < b$ *is an interval containing all the eigenvalues of* $\boldsymbol{A}$.

Then in the conjugate gradient method

$$\frac{||\boldsymbol{x}-\boldsymbol{x}_k||_{\boldsymbol{A}}}{||\boldsymbol{x}-\boldsymbol{x}_0||_{\boldsymbol{A}}} = \min_{\substack{Q\in\Pi_k\\ Q(0)=1}} \max_{a\le x\le b} |Q(x)|, \tag{9.32}$$

where Π_k denotes the class of univariate polynomials of degree $\le k$ with real coefficients.

Proof. With the notation in Lemma 9.23 we find $||\boldsymbol{x}-\boldsymbol{x}_0||_{\boldsymbol{A}}^2 = \boldsymbol{r}_0\boldsymbol{A}^{-1}\boldsymbol{r}_0 = \sum_{j=1}^n \frac{\sigma_j^2}{\lambda_j}$. Therefore, by the best approximation property and (9.30), for any $\boldsymbol{w}\in\mathbb{W}_k$

$$||\boldsymbol{x}-\boldsymbol{x}_k||_{\boldsymbol{A}}^2 \le ||\boldsymbol{x}-\boldsymbol{x}_0-\boldsymbol{w}||_{\boldsymbol{A}}^2 \le \max_{a\le x\le b}|Q(x)|^2 \sum_{j=1}^n \frac{\sigma_j^2}{\lambda_j} = \max_{a\le x\le b}|Q(x)|^2 ||\boldsymbol{x}-\boldsymbol{x}_0||_{\boldsymbol{A}}^2,$$

where $Q\in\Pi_k$ and $Q(0)=1$. Minimizing over such polynomials Q and taking square roots the result follows. □

In the next section we use properties of the Chebyshev polynomials to show that

$$\frac{||\boldsymbol{x}-\boldsymbol{x}_k||_{\boldsymbol{A}}}{||\boldsymbol{x}-\boldsymbol{x}_0||_{\boldsymbol{A}}} \le \min_{\substack{Q\in\Pi_k\\ Q(0)=1}} \max_{\lambda_{min}\le x\le\lambda_{max}} |Q(x)| = \frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k}, \tag{9.33}$$

where $\kappa = \lambda_{max}/\lambda_{min}$ is the spectral condition number of $\boldsymbol{A}$. Ignoring the second term in the denominator this implies the first inequality in (9.19). The second inequality follows from (9.26).

Exercise 9.25 (Krylov space and cg iterations)
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 $k\le 3$ 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* $k = 1, 2, 3$.
- $\{|\boldsymbol{r}_k\boldsymbol{\epsilon}\}$ *is monotonically decreasing.*
- $\{|\boldsymbol{x}_k - \boldsymbol{x}\boldsymbol{\epsilon}\}$ *is monotonically decreasing.*

9.4.3 Chebyshev polynomials

The proof of the estimate (9.33) for the error in the conjugate gradient method is based on an extremal property of the Chebyshev polynomials. Suppose $a < b$, $c \notin [a, b]$ and $k \in \mathbb{N}$. Consider the set $\mathcal{S}_k$ of all polynomials Q of degree $\leq k$ such that $Q(c) = 1$. For any continuous function f on $[a, b]$ we define

$$\|f\|_\infty = \max_{a \leq x \leq b} |f(x)|.$$

We want to find a polynomial $Q^* \in \mathcal{S}_k$ such that

$$\|Q^*\|_\infty = \min_{Q \in \mathcal{S}_k} \|Q\|_\infty.$$

We will show that Q^* is uniquely given as a suitably shifted and normalized version of the **Chebyshev poynomial**. The Chebyshev polynomial T_n of degree n can be defined recursively by

$$T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t), \quad n \geq 1, \quad t \in \mathbb{R},$$

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 9.26 (Closed forms of Chebyshev polynomials)
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

$$x_{n+1} - 2tx_n + x_{n-1} = 0 \text{ for } n \geq 1, \text{ with } x_0 = 1, x_1 = t. \tag{9.34}$$

To solve this difference equation we insert $x_n = z^n$ into (9.34) and obtain $z^{n+1} - 2tz^n + z^{n-1} = 0$ or $z^2 - 2tz + 1 = 0$. The roots of this equation are

$$z_1 = t + \sqrt{t^2-1}, \quad z_2 = t - \sqrt{t^2-1} = \left(t + \sqrt{t^2-1}\right)^{-1}.$$

Now z_1^n, z_2^n and more generally $c_1 z_1^n + c_2 z_2^n$ are solutions of (9.34) 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 $c_1 = c_2 = \frac{1}{2}$. ∎

We show that the unique solution to our minimization problem is

$$Q^*(x) = \frac{T_k(u(x))}{T_k(u(c))}, \qquad u(x) = \frac{b + a - 2x}{b - a}. \tag{9.35}$$

Clearly $Q^* \in \mathcal{S}_k$.

Theorem 9.27 (A minimal norm problem)
Suppose $a < b$, $c \notin [a, b]$ and $k \in \mathbb{N}$. If $Q \in \mathcal{S}_k$ and $Q \neq Q^*$ then $\|Q\|_\infty > \|Q^*\|_\infty$.

Proof. Recall that a nonzero polynomial p of degree k can have at most k zeros. If $p(z) = p'(z) = 0$, we say that p has a double zero at z. Counting such a zero as two zeros it is still true that a nonzero polynomial of degree k has at most k zeros.

$|Q^*|$ takes on its maximum $1/|T_k(u(c))|$ at the $k+1$ points $\mu_0, \ldots, \mu_k$ in $[a, b]$ such that $u(\mu_i) = \cos(i\pi/k)$ for $i = 0, 1, \ldots, k$. Suppose $Q \in S_k$ and that $\|Q\| \leq \|Q^*\|$. We have to show that $Q \equiv Q^*$. Let $f \equiv Q - Q^*$. We show that f

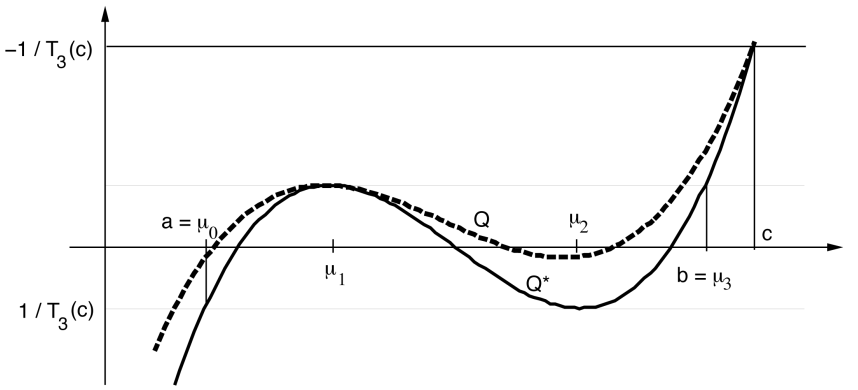


Figure 9.3. *This is an illustration of the proof of Theorem 9.27 for* $k = 3$. $f \equiv Q - Q^*$ *has a double zero at* μ_1 *and one zero between* μ_2 *and* μ_3.

has at least k zeros in $[a, b]$. Since f is a polynomial of degree $\leq k$ and $f(c) = 0$, this means that $f \equiv 0$ or equivalently $Q \equiv Q^*$.

Consider $I_j = [\mu_{j-1}, \mu_j]$ for a fixed j. Let

$$\sigma_j = f(\mu_{j-1}) f(\mu_j).$$

We have $\sigma_j \leq 0$. For if say $Q^*(\mu_j) > 0$ then

$$Q(\mu_j) \leq \|Q\|_\infty \leq \|Q^*\|_\infty = Q^*(\mu_j)$$

so that $f(\mu_j) \leq 0$. Moreover,

$$-Q(\mu_{j-1}) \leq \|Q\|_\infty \leq \|Q^*\|_\infty = -Q^*(\mu_{j-1}).$$

Thus $f(\mu_{j-1}) \geq 0$ and It follows that $\sigma_j \leq 0$. Similarly, $\sigma_j \leq 0$ if $Q^*(\mu_j) < 0$.

If $\sigma_j < 0$, f must have a zero in I_j since it is continuous. Suppose $\sigma_j = 0$. Then $f(\mu_{j-1}) = 0$ or $f(\mu_j) = 0$. If $f(\mu_j) = 0$ then $Q(\mu_j) = Q^*(\mu_j)$. But then μ_j is a maximum or minimum both for Q and Q^*. If $\mu_j \in (a, b)$ then $Q'(\mu_j) = Q^{*\prime}(\mu_j) = 0$. Thus $f(\mu_j) = f'(\mu_j) = 0$, and f has a double zero at μ_j. We can count this as one zero for I_j and one for I_{j+1}. If $\mu_j = b$, we still have a zero in I_j. Similarly, if $f(\mu_{j-1}) = 0$, a double zero of f at μ_{j-1} appears if $\mu_{j-1} \in (a, b)$. We count this as one zero for I_{j-1} and one for I_j.

In this way we associate one zero of f for each of the k intervals I_j, $j = 1, 2, \ldots, k$. We conclude that f has at least k zeros in $[a, b]$. □

Exercise 9.28 (Another explicit formula for the Chebyshev polynomial)

Show that

$$T_n(t) = \cosh(n \operatorname{arccosh} t) \text{ for } |t| \geq 1,$$

where arccosh is the inverse function of $\cosh x := (e^x + e^{-x})/2$.

Theorem 9.27 with a, and b, the smallest and largest eigenvalue of $\boldsymbol{A}$, and $c = 0$ implies that the minimizing polynomial in (9.33) is given by

$$Q^*(x) = T_k\left(\frac{b+a-2x}{b-a}\right) / T_k\left(\frac{b+a}{b-a}\right). \tag{9.36}$$

By Lemma 9.26

$$\max_{a \leq x \leq b} \left| T_k\left(\frac{b+a-2x}{b-a}\right)\right| = \max_{-1 \leq t \leq 1} \left|T_k(t)\right| = 1. \tag{9.37}$$

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 9.26 we find

$$T_k\left(\frac{b+a}{b-a}\right) = T_k\left(\frac{\kappa+1}{\kappa-1}\right) = \frac{1}{2}\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right] \tag{9.38}$$

and (9.33) follows.

9.4.4 Monotonicity of the error

The error analysis for the conjugate gradient method is based on the $\boldsymbol{A}$-norm. We end this chapter by considering the Euclidian norm of the error, and show that it is strictly decreasing.

Theorem 9.29 (The error in cg is strictly decreasing)
Let in the conjugate gradient method m *be the smallest integer such that* $\boldsymbol{r}_{m+1} = \boldsymbol{0}$. *For* $k \leq m$ *we have* $\|\boldsymbol{\epsilon}_{k+1}\|_2 < \|\boldsymbol{\epsilon}_k\|_2$. *More precisely,*

$$\|\boldsymbol{\epsilon}_k\|_2^2 - \|\boldsymbol{\epsilon}_{k+1}\|_2^2 = \frac{\|\boldsymbol{p}_k\|_2^2}{\|\boldsymbol{p}_k\|_{\boldsymbol{A}}^2}(\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2 + \|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2)$$

where $\boldsymbol{\epsilon}_j = \boldsymbol{x} - \boldsymbol{x}_j$ *and* $\boldsymbol{A}\boldsymbol{x} = b$..

Proof. For $j \le m$

$$\boldsymbol{\epsilon}_j = \boldsymbol{x}_{m+1} - \boldsymbol{x}_j = \boldsymbol{x}_m - \boldsymbol{x}_j + \alpha_m \boldsymbol{p}_m = \boldsymbol{x}_{m-1} - \boldsymbol{x}_j + \alpha_{m-1}\boldsymbol{p}_{m-1} + \alpha_m \boldsymbol{p}_m = ...$$

so that

$$\boldsymbol{\epsilon}_j = \sum_{i=j}^{m} \alpha_i \boldsymbol{p}_i, \quad \alpha_i = \frac{\boldsymbol{r}_i^T \boldsymbol{r}_i}{\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_i}. \tag{9.39}$$

By (9.39) and $\boldsymbol{A}$-orthogonality

$$\|\boldsymbol{\epsilon}_j\|_{\boldsymbol{A}}^2 = \boldsymbol{\epsilon}_j \boldsymbol{A} \boldsymbol{\epsilon}_j = \sum_{i=j}^{m} \alpha_i^2 \boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_i = \sum_{i=j}^{m} \frac{(\boldsymbol{r}_i^T \boldsymbol{r}_i)^2}{\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_i}. \tag{9.40}$$

By (9.15) and Lemma 9.21

$$\boldsymbol{p}_i^T \boldsymbol{p}_k = (\boldsymbol{r}_i + \beta_{i-1}\boldsymbol{p}_{i-1})^T \boldsymbol{p}_k = \beta_{i-1}\boldsymbol{p}_{i-1}^T \boldsymbol{p}_k = \cdots = \beta_{i-1} \cdots \beta_k (\boldsymbol{p}_k^T \boldsymbol{p}_k),$$

and since $\beta_{i-1} \cdots \beta_k = (\boldsymbol{r}_i^T \boldsymbol{r}_i)/(\boldsymbol{r}_k^T \boldsymbol{r}_k)$ we obtain

$$\boldsymbol{p}_i^T \boldsymbol{p}_k = \frac{\boldsymbol{r}_i^T \boldsymbol{r}_i}{\boldsymbol{r}_k^T \boldsymbol{r}_k} \boldsymbol{p}_k^T \boldsymbol{p}_k, \quad i \ge k. \tag{9.41}$$

Since

$$\|\boldsymbol{\epsilon}_k\|_2^2 = \|\boldsymbol{\epsilon}_{k+1} + \boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 = \|\boldsymbol{\epsilon}_{k+1} + \alpha_k \boldsymbol{p}_k\|_2^2,$$

we obtain

$$\begin{aligned}
\|\boldsymbol{\epsilon}_k\|_2^2 - \|\boldsymbol{\epsilon}_{k+1}\|_2^2 =& \alpha_k \big(2\boldsymbol{p}_k^T \boldsymbol{\epsilon}_{k+1} + \alpha_k \boldsymbol{p}_k^T \boldsymbol{p}_k\big) \\
\overset{(9.39)}{=}& \alpha_k \big(2 \sum_{i=k+1}^{m} \alpha_i \boldsymbol{p}_i^T \boldsymbol{p}_k + \alpha_k \boldsymbol{p}_k^T \boldsymbol{p}_k\big) = \big(\sum_{i=k}^{m} + \sum_{i=k+1}^{m}\big) \alpha_k \alpha_i \boldsymbol{p}_i^T \boldsymbol{p}_k \\
\overset{(9.41)}{=}& \big(\sum_{i=k}^{m} + \sum_{i=k+1}^{m}\big) \frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{p}_k^T \boldsymbol{A}\boldsymbol{p}_k} \frac{\boldsymbol{r}_i^T \boldsymbol{r}_i}{\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_i} \frac{\boldsymbol{r}_i^T \boldsymbol{r}_i}{\boldsymbol{r}_k^T \boldsymbol{r}_k} \boldsymbol{p}_k^T \boldsymbol{p}_k \\
\overset{(9.40)}{=}& \frac{\|\boldsymbol{p}_k\|_2^2}{\|\boldsymbol{p}_k\|_{\boldsymbol{A}}^2} \big(\|\boldsymbol{\epsilon}_k\|_{\boldsymbol{A}}^2 + \|\boldsymbol{\epsilon}_{k+1}\|_{\boldsymbol{A}}^2\big).
\end{aligned}$$

and the Theorem is proved. □

9.5 Preconditioning

For problems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ of size n, where both n and $\mathrm{cond}_2(\boldsymbol{A})$ are large, it is often possible to improve the performance of the conjugate gradient method by using a

technique known as **preconditioning**. Instead of $\boldsymbol{Ax} = \boldsymbol{b}$ we consider an equivalent system $\boldsymbol{BAx} = \boldsymbol{Bb}$, where $\boldsymbol{B}$ is nonsingular and $\text{cond}_2(\boldsymbol{BA})$ is smaller than $\text{cond}_2(\boldsymbol{A})$. The matrix $\boldsymbol{B}$ will in many cases be the inverse of another matrix, $\boldsymbol{B} = \boldsymbol{M}^{-1}$. We cannot use CG on $\boldsymbol{BAx} = \boldsymbol{Bb}$ directly since $\boldsymbol{BA}$ in general is not symmetric even if both $\boldsymbol{A}$ and $\boldsymbol{B}$ are. But if $\boldsymbol{B}$ (and hence $\boldsymbol{M}$) is symmetric positive definite then we can apply CG to a symmetrized system and then transform the recurrence formulas to an iterative method for the original system $\boldsymbol{Ax} = \boldsymbol{b}$. This iterative method is known as the **preconditioned 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 2.38 there is a nonsingular matrix $\boldsymbol{C}$ such that $\boldsymbol{B} = \boldsymbol{C}^T\boldsymbol{C}$. ($\boldsymbol{C}$ is only needed for the derivation and will not appear in the final formulas). 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

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{9.42}$$

$$\boldsymbol{BAx} = \boldsymbol{Bb} \tag{9.43}$$

$$(\boldsymbol{CAC}^T)\boldsymbol{y} = \boldsymbol{Cb},\ \&\ \boldsymbol{x} = \boldsymbol{C}^T\boldsymbol{y}. \tag{9.44}$$

Note that (9.42) and (9.44) 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 (9.44) 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 $\boldsymbol{z}_k = \boldsymbol{Cb} - \boldsymbol{CAC}^T\boldsymbol{y}_k$ we obtain the following from (9.13), (9.14), and (9.15).

$$\begin{aligned}
\boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \alpha_k\boldsymbol{q}_k, \quad \alpha_k = \boldsymbol{z}_k^T\boldsymbol{z}_k/\boldsymbol{q}_k^T(\boldsymbol{CAC}^T)\boldsymbol{q}_k,\\
\boldsymbol{z}_{k+1} &= \boldsymbol{z}_k - \alpha_k(\boldsymbol{CAC}^T)\boldsymbol{q}_k,\\
\boldsymbol{q}_{k+1} &= \boldsymbol{z}_{k+1} + \beta_k\boldsymbol{q}_k, \quad \beta_k = \boldsymbol{z}_{k+1}^T\boldsymbol{z}_{k+1}/\boldsymbol{z}_k^T\boldsymbol{z}_k.
\end{aligned}$$

With

$$\boldsymbol{x}_k := \boldsymbol{C}^T\boldsymbol{y}_k, \quad \boldsymbol{p}_k := \boldsymbol{C}^T\boldsymbol{q}_k, \quad \boldsymbol{s}_k := \boldsymbol{C}^T\boldsymbol{z}_k, \quad \boldsymbol{r}_k := \boldsymbol{C}^{-1}\boldsymbol{z}_k \tag{9.45}$$

this can be transformed into

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \quad \alpha_k = \frac{\boldsymbol{s}_k^T \boldsymbol{r}_k}{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k}, \tag{9.46}$$

$$\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k \boldsymbol{A} \boldsymbol{p}_k, \tag{9.47}$$

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \alpha_k \boldsymbol{B} \boldsymbol{A} \boldsymbol{p}_k, \tag{9.48}$$

$$\boldsymbol{p}_{k+1} = \boldsymbol{s}_{k+1} + \beta_k \boldsymbol{p}_k, \quad \beta_k = \frac{\boldsymbol{s}_{k+1}^T \boldsymbol{r}_{k+1}}{\boldsymbol{s}_k^T \boldsymbol{r}_k}. \tag{9.49}$$

Here $\boldsymbol{x}_k$ will be an approximation to the solution $\boldsymbol{x}$ of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$ is the residual in the original system, and $\boldsymbol{s}_k = \boldsymbol{B}\boldsymbol{b} - \boldsymbol{B}\boldsymbol{A}\boldsymbol{x}_k$ is the residual in the preconditioned system. This follows since by (9.45)

$$\boldsymbol{r}_k = \boldsymbol{C}^{-1}\boldsymbol{z}_k = \boldsymbol{b} - \boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^T\boldsymbol{y}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$$

and $\boldsymbol{s}_k = \boldsymbol{C}^T\boldsymbol{z}_k = \boldsymbol{C}^T\boldsymbol{C}\boldsymbol{r}_k = \boldsymbol{B}\boldsymbol{r}_k$. We start with $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$, $\boldsymbol{p}_0 = \boldsymbol{s}_0 = \boldsymbol{B}\boldsymbol{r}_0$ and obtain the following preconditioned conjugate gradient algorithm for determining approximations $\boldsymbol{x}_k$ to the solution of a symmetric positive definite system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

Algorithm 9.30 (Preconditioned conjugate gradient)
The symmetric positive definite linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is solved by the preconditioned conjugate gradient method on the system $\boldsymbol{B}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{b}$, where $\boldsymbol{B}$ is symmetric positive definite . $\boldsymbol{x}$ is a starting vector for the iteration. The iteration is stopped when $||\boldsymbol{r}_k||_2/||\boldsymbol{b}||_2 \leq$ tol or $k >$ itmax. K is the number of iterations used.

```
function [x,K]=pcg(A,B,b,x,tol,itmax)
r=b-A*x; p=B*r; s=p; rho=s'*r; rho0=b'*b;
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;
    w=B*t; s=s-a*w;
    rhos=rho; rho=s'*r;
    p=r+(rho/rhos)*p;
end
K=itmax+1;
```

Appart from the calculation of ρ this algorithm is quite similar to Algorithm 9.11. The main additional work is contained in $w = B * t$. We'll discuss this further in connection with an example. There the inverse of $\boldsymbol{B}$ is known and we have to solve a linear system to find $\boldsymbol{w}$.

We have the following convergence result for this algorithm.

Theorem 9.31 (Error bound preconditioned cg)
Suppose we apply a symmetric positive definite preconditioner $\boldsymbol{B}$ to the symmetric positive definite system $\boldsymbol{Ax} = \boldsymbol{b}$. Then the quantities $\boldsymbol{x}_k$ computed in Algorithm 9.30 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 $\boldsymbol{BA}$.

Proof. Since Algorithm 9.30 is equivalent to solving (9.44) by the conjugate gradient method Theorem 9.16 implies that

$$\frac{||\boldsymbol{y} - \boldsymbol{y}_k||_{\boldsymbol{CAC}^T}}{||\boldsymbol{y} - \boldsymbol{y}_0||_{\boldsymbol{CAC}^T}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \quad \text{for} \quad k \geq 0,$$

where $\boldsymbol{y}_k$ is the conjugate gradient approximation to the solution $\boldsymbol{y}$ of (9.44) and κ is the ratio of the largest and smallest eigenvalue of $\boldsymbol{CAC}^T$. Since $\boldsymbol{BA}$ and $\boldsymbol{CAC}^T$ are similar this is the same as the κ in the theorem. By (9.45) we have

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{y}_k\|^2_{\boldsymbol{CAC}^T} &= (\boldsymbol{y} - \boldsymbol{y}_k)^T(\boldsymbol{CAC}^T)(\boldsymbol{y} - \boldsymbol{y}_k) \\ &= (\boldsymbol{C}^T(\boldsymbol{y} - \boldsymbol{y}_k))^T\boldsymbol{A}(\boldsymbol{C}^T(\boldsymbol{y} - \boldsymbol{y}_k)) = \|\boldsymbol{x} - \boldsymbol{x}_k\|^2_{\boldsymbol{A}}\end{aligned}$$

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 arithmetic operations, ideally $O(n)$ for both.

9.6 Preconditioning Example

9.6.1 A variable coefficient problem

Consider the problem

$$\begin{aligned} -\tfrac{\partial}{\partial x}\big(c(x,y)\tfrac{\partial u}{\partial x}\big) - \tfrac{\partial}{\partial y}\big(c(x,y)\tfrac{\partial u}{\partial y}\big) &= f(x,y) \quad (x,y) \in \Omega = (0,1)^2 \\ u(x,y) &= 0 \quad (x,y) \in \partial\Omega. \end{aligned} \tag{9.50}$$

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 (9.50) 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 (9.50) 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 (9.50) 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 (9.50) turns out to be

$$\begin{array}{rcll} -(\boldsymbol{P}_h v)_{j,k} &=& h^2 f_{j,k} & j,k = 1,\ldots,m \\ v_{j,k} &=& 0 & j = 0, m+1 \text{ all } k \text{ or } k = 0, m+1 \text{ all j,} \end{array} \tag{9.51}$$

where

$$\begin{aligned} -(\boldsymbol{P}_h v)_{j,k} &= (c_{j,k-\frac{1}{2}} + c_{j-\frac{1}{2},k} + c_{j+\frac{1}{2},k} + c_{j,k+\frac{1}{2}})v_{j,k} \\ &\quad - c_{j,k-\frac{1}{2}}v_{j,k-1} - c_{j-\frac{1}{2},k}v_{j-1,k} - c_{j+\frac{1}{2},k}v_{j+1,k} - c_{j,k+\frac{1}{2}}v_{j,k+1} \end{aligned} \tag{9.52}$$

and $f_{j,k} = f(jh,kh)$.

As before we let $\boldsymbol{V} = (v_{j,k}) \in \mathbb{R}^{m\times m}$ and $\boldsymbol{F} = (f_{j,k}) \in \mathbb{R}^{m\times m}$. 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

$$\begin{array}{rcll} a_{i,i} &=& c_{j_i,k_i-\frac{1}{2}} + c_{j_i-\frac{1}{2},k_i} + c_{j_i+\frac{1}{2},k_i} + c_{j_i,k_i+\frac{1}{2}}, & i = 1,2,\ldots,n \\ a_{i+1,i} &=& a_{i,i+1} = -c_{j_i+\frac{1}{2},k_i}, & i \bmod m \neq 0 \\ a_{i+m,i} &=& a_{i,i+m} = -c_{j_i,k_i+\frac{1}{2}}, & i = 1,2,\ldots,n-m \\ a_{i,j} &=& 0 & \text{otherwise,} \end{array} \tag{9.53}$$

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 $i = 1, \ldots, n$. If $c(x, y) = 1$ for all $(x, y) \in \Omega$ we recover the Poisson matrix.

In general we cannot write $\boldsymbol{A}$ as a Kronecker sum. But we can show that $\boldsymbol{A}$ is symmetric and it is positive definite as long as the function c is positive on Ω.

Theorem 9.32 (Positive definite matrix)
If $c(x, y) > 0$ for $(x, y) \in \Omega$ then the matrix $\boldsymbol{A}$ given by (9.53) is symmetric positive definite.

Proof.

To each $x \in \mathbb{R}^n$ there corresponds a matrix $\boldsymbol{V} \in \mathbb{R}^{m \times m}$ such that $x = \text{vec}(\boldsymbol{V})$. We claim that

$$x^T \boldsymbol{A} \boldsymbol{x} = \sum_{j=1}^{m} \sum_{k=0}^{m} c_{j,k+\frac{1}{2}} \big(v_{j,k+1} - v_{j,k}\big)^2 + \sum_{k=1}^{m} \sum_{j=0}^{m} c_{j+\frac{1}{2},k} \big(v_{j+1,k} - v_{j,k}\big)^2, \quad (9.54)$$

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 (9.54) we see that $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \ge 0$ 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 (9.54). From the connection between (9.52) and (9.53) 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 obtain

$$\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j,k-\frac{1}{2}} v_{j,k}^2 = \sum_{j=1}^{m}\sum_{k=0}^{m} c_{j,k+\frac{1}{2}} v_{j,k+1}^2,$$
$$\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j,k-\frac{1}{2}} v_{j,k-1} v_{j,k} = \sum_{j=1}^{m}\sum_{k=0}^{m} c_{j,k+\frac{1}{2}} v_{j,k+1} v_{j,k},$$
$$\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j-\frac{1}{2},k} v_{j,k}^2 = \sum_{k=1}^{m}\sum_{j=0}^{m} c_{j+\frac{1}{2},k} v_{j+1,k}^2,$$
$$\sum_{j=1}^{m}\sum_{k=1}^{m} c_{j-\frac{1}{2},k} v_{j-,k} v_{j,k} = \sum_{k=1}^{m}\sum_{j=0}^{m} c_{j+\frac{1}{2},k} v_{j+1,k} v_{j,k}.$$

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 (9.54) follows. □

9.6.2 Applying preconditioning

Consider solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is given by (9.53) 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 $O(n^2)$ arithmetic operations for large n.

If we choose $c(x, y) \equiv 1$ in (9.50), we get the Poisson problem. 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 (9.53). This suggests using $\boldsymbol{B} = \boldsymbol{A}_p^{-1}$, the inverse of the discrete Poisson matrix as a preconditioner for the system (9.51).

Consider Algorithm 9.30. With this preconditioner the calculation $\boldsymbol{w} = \boldsymbol{B}\boldsymbol{t}$ takes the form $\boldsymbol{A}_p \boldsymbol{w}_k = \boldsymbol{t}_k$.

In Section 4.2 we developed a Simple fast Poisson Solver, Cf. Algorithm 4.1. This method can be utilized to solve $\boldsymbol{A}_p \boldsymbol{w} = \boldsymbol{t}$.

Consider the specific problem where

$$c(x, y) = e^{-x+y} \text{ and } f(x, y) = 1.$$

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 9.33. *The number of iterations K (no preconditioning) and K_{pre} (with preconditioning) for the problem (9.50) using the discrete Poisson problem as a preconditioner.*

We have used Algorithm 9.11 (conjugate gradient without preconditioning), and Algorithm 9.30 (conjugate gradient with preconditioning) to solve the problem (9.50). We used $\boldsymbol{x}_0 = 0$ and $\epsilon = 10^{-8}$. The results are shown in Table 9.33.

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.

Using a preconditioner increases the work in each iteration. For the present example the number of arithmetic operations 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 significantly.

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 (9.50) can be bounded independently of n.

Theorem 9.34 (Eigevalues of preconditioned matrix)
Suppose $0 < c_0 \leq c(x, y) \leq c_1$ for all $(x, y) \in [0, 1]^2$. For the eigenvalues of the matrix $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A}_p^{-1}\boldsymbol{A}$ just described we have

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \leq \frac{c_1}{c_0}.$$

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 (9.54) 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 (9.54) from below by c_0 and above by c_1 we find

$$c_0(x^T \boldsymbol{A}_p x) \leq x^T \boldsymbol{A}\boldsymbol{x} \leq c_1(x^T \boldsymbol{A}_p x)$$

which implies that $c_0 \leq \lambda \leq c_1$ for all eigenvalues λ of $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A}_p^{-1}\boldsymbol{A}$. □

Using $c(x, y) = e^{-x+y}$ as above, we find $c_0 = e^{-2}$ and $c_1 = 1$. Thus $\kappa \leq e^2 \approx 7.4$, a quite acceptable matrix condition number which explains the convergence results from our numerical experiment.

9.7 Review Questions

9.7.1 Does the steepest descent and conjugate gradient method always converge?

9.7.2 What kind of orthogonalities occur in the conjugate gradient method?

9.7.3 What is a Krylow space?

9.7.4 What is a convex function?

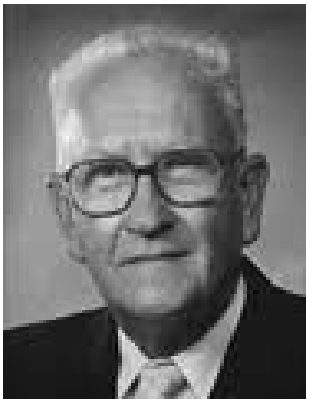
9.7.5 How do SOR and conjugate gradient compare?

Part IV

Orthonormal Transformations and Least Squares

Chapter 10

Orthonormal and Unitary Transformations




Alston Scott Householder, 1904-1993 (left), James Hardy Wilkinson, 1919-1986 (right). Householder and Wilkinson are two of the founders of modern numerical analysis and scientific computing.

Gauss transformations, (cf. Theorem 2.60, the PLU theorem) are used in Gaussian elimination to reduce a matrix to triangular form. These are not the only kind of transformations that can be used for such a task. In this chapter we study how transformations by orthonormal and unitary matrices can be used to reduce a square matrix to upper triangular form and more generally 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 reduced form which we refer to as a **QR factorization**. The QR decomposition and factorization will be used in later chapters to solve least squares- and eigenvalue problems.

It cannot be repeated too often that orthonormal transformations have the advantage that they preserve the Euclidian norm of a vector, and the spectral norm and Frobenius norm of a matrix, see Lemma 6.24 and Theorem 7.20. This means that when an orthonormal transformation is applied to an inaccurate vector or matrix then the error will not grow. Thus in general an orthonormal transfor-

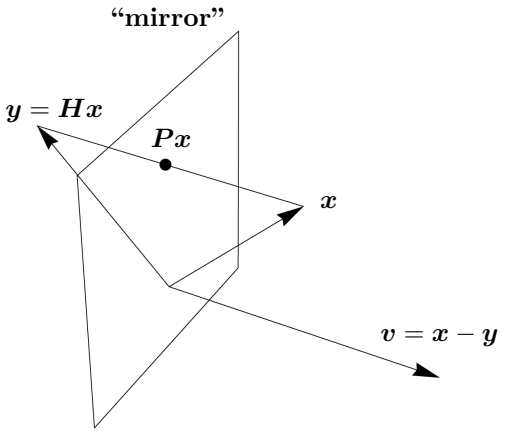


Figure 10.1. *The Householder transformation in Exercise 10.2*

mation is numerically stable.

10.1 The Householder Transformation

Definition 10.1 (Householder transformation)
A matrix $\boldsymbol{H} \in \mathbb{C}^{n\times n}$ *of the form*

$$\boldsymbol{H} := \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*, \text{ where } \boldsymbol{u} \in \mathbb{C}^n \text{ and } \boldsymbol{u}^*\boldsymbol{u} = 2$$

is called a **Householder transformation**. *The name* **elementary reflector** *is also used.*

In the real case and 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 Hermitian and unitary. Indeed, $\boldsymbol{H}^* = (\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*)^* = \boldsymbol{H}$ and

$$\boldsymbol{H}^*\boldsymbol{H} = \boldsymbol{H}^2 = (\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*)(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*) = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^* + \boldsymbol{u}(\boldsymbol{u}^*\boldsymbol{u})\boldsymbol{u}^* = \boldsymbol{I}.$$

In the real case $\boldsymbol{H}$ is symmetric and orthonormal.

There are several ways to represent a Householder transformation. Householder used $\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^*$, where $\boldsymbol{u}^*\boldsymbol{u} = 1$. For any nonzero $\boldsymbol{v} \in \mathbb{R}^n$ the matrix

$$\boldsymbol{H} := \boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^*}{\boldsymbol{v}^*\boldsymbol{v}} \tag{10.1}$$

is a Householder transformation. Indeed, $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*$, where $\boldsymbol{u} := \sqrt{2}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}$ has length $\sqrt{2}$. Moreover, if $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$ and $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{y} \neq \boldsymbol{0}$, then $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{y}$ (Cf. Exercise 10.2).

Exercise 10.2 (Reflector)
Suppose $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ *with* $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$ *and* $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{y} \neq \boldsymbol{0}$.

(a) *Show that* $\boldsymbol{H}\boldsymbol{x} := \big(\boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}\big)\boldsymbol{x} = \boldsymbol{y}$. [15].

(b) *Let* $\mathcal{M} := \{\boldsymbol{w} \in \mathbb{R}^n : \boldsymbol{w}^T\boldsymbol{v} = 0\}$ *and* $\boldsymbol{P} := \boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}$. *Show that* $\boldsymbol{P}\boldsymbol{x} = (\boldsymbol{x}+\boldsymbol{y})/2 \in \mathcal{M}$. *Thus* $\boldsymbol{y}$ *is the reflected image of* $\boldsymbol{x}$, *where* $\mathcal{M}$ *is the "mirror". See Figure 10.1.*

(c) *Determine the matrices* $\boldsymbol{H}, \boldsymbol{P}$ *and the "mirror"* $\mathcal{M}$ *when* $\boldsymbol{x} := [1, 0, 1]^T$ *and* $\boldsymbol{y} := [-1, 0, 1]^T$.

A main use of Householder transformations is to produce zeros in vectors.

Theorem 10.3 (Zeros in vectors)
Suppose $\boldsymbol{x} \in \mathbb{C}^n$ *is nonzero and define* $\rho \in \mathbb{C}$ *and* $\boldsymbol{z}, \boldsymbol{u} \in \mathbb{C}^n$ *by*

$$\rho := \begin{cases} x_1/|x_1|, & \text{if } x_1 \neq 0, \\ 1, & \text{otherwise.} \end{cases}, \quad , \boldsymbol{z} := \overline{\rho}\boldsymbol{x}/\|\boldsymbol{x}\|_2 \quad \boldsymbol{u} := \frac{\boldsymbol{z} + \boldsymbol{e}_1}{\sqrt{1 + z_1}}. \tag{10.2}$$

Then $\boldsymbol{u}^*\boldsymbol{u} = 2$, $\boldsymbol{x} = \rho\|\boldsymbol{x}\|_2\boldsymbol{z}$ *and*

$$\boldsymbol{H}\boldsymbol{x} := (\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*)\boldsymbol{x} = a\boldsymbol{e}_1, \quad a := -\rho\|\boldsymbol{x}\|_2. \tag{10.3}$$

Proof. Since $|\rho| = 1$ we have $\rho\|\boldsymbol{x}\|_2\boldsymbol{z} = |\rho|^2\boldsymbol{x} = \boldsymbol{x}$. Moreover, $\|\boldsymbol{z}\|_2 = 1$ and $z_1 = |x_1|/\|\boldsymbol{x}\|_2$ is real so that $\boldsymbol{u}^*\boldsymbol{u} = \frac{(\boldsymbol{z}+\boldsymbol{e}_1)^*(\boldsymbol{z}+\boldsymbol{e}_1)}{1+z_1} = \frac{2+2z_1}{1+z_1} = 2$. Finally,

$$\begin{aligned}\boldsymbol{H}\boldsymbol{x} &= \boldsymbol{x} - (\boldsymbol{u}^*\boldsymbol{x})\boldsymbol{u} = \rho\|\boldsymbol{x}\|_2(\boldsymbol{z} - (\boldsymbol{u}^*\boldsymbol{z})\boldsymbol{u}) = \rho\|\boldsymbol{x}\|_2(\boldsymbol{z} - \frac{(\boldsymbol{z}^* + \boldsymbol{e}_1^*)\boldsymbol{z}}{1 + z_1}(\boldsymbol{z} + \boldsymbol{e}_1)) \\ &= \rho\|\boldsymbol{x}\|_2(\boldsymbol{z} - (\boldsymbol{z} + \boldsymbol{e}_1)) = -\rho\|\boldsymbol{x}\|_2\boldsymbol{e}_1 = a\boldsymbol{e}_1.\end{aligned}$$

□

The formulas in Theorem 10.3 are implemented in the following algorithm adapted from [26]. To any given $\boldsymbol{x} \in \mathbb{C}^n$ a number a and a vector $\boldsymbol{u}$ with $\boldsymbol{u}^*\boldsymbol{u} = 2$ is computed so that $(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^*)\boldsymbol{x} = a\boldsymbol{e}_1$.

[15] Hint: Show first that $\boldsymbol{v}^T\boldsymbol{v} = 2\boldsymbol{v}^T\boldsymbol{x}$

Algorithm 10.4 (Generate a Householder transformation)

```
function [u,a]=housegen(x)
a=norm(x);
if  a==0
   u=x; u(1)=sqrt(2); return;
end
if x(1)== 0
    r=1;
else
    r=x(1)/abs(x(1));
end
u=conj(r)*x/a;
u(1)=u(1)+1;
u=u/sqrt(u(1));
a=-r*a;
end
```

Note that

- 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.
- In Theorem 10.3 the first component of $\boldsymbol{z}$ is $z_1 = |x_1|/\|\boldsymbol{x}\|_2 \geq 0$. Since $\|\boldsymbol{z}\|_2 = 1$ we have $1 \leq 1 + z_1 \leq 2$. It follows that $\boldsymbol{u}$ is well defined and we avoid cancelation error when computing $1 + z_1$.

Exercise 10.5 (What does algorithm housegen do when $\boldsymbol{x} = \boldsymbol{e}_1$?)
Determine $\boldsymbol{H}$ in Algorithm 10.4 when $\boldsymbol{x} = \boldsymbol{e}_1$.

Householder transformations can also be used to zero out only the lower part of a vector. Suppose $\boldsymbol{x}^T := [\boldsymbol{y}, \boldsymbol{z}]^T$, where $\boldsymbol{y} \in \mathbb{C}^k$, $\boldsymbol{z} \in \mathbb{C}^{n-k}$ for some $1 \leq k < n$. The command $[\hat{\boldsymbol{u}}, a] :=$ housegen$(\boldsymbol{z})$ defines a Householder transformation $\hat{\boldsymbol{H}} = \boldsymbol{I} - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^*$ so that $\hat{\boldsymbol{H}}\boldsymbol{z} = a\boldsymbol{e}_1$. With $\boldsymbol{u}^T := [\boldsymbol{0}, \hat{\boldsymbol{u}}]^T \in \mathbb{C}^n$ we see that $\boldsymbol{u}^*\boldsymbol{u} = \hat{\boldsymbol{u}}^*\hat{\boldsymbol{u}} = 2$, and

$$\boldsymbol{H}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ a\boldsymbol{e}_1 \end{bmatrix}, \text{ where } \boldsymbol{H} := \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^* = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} - \begin{bmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{u}} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} & \hat{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{H}} \end{bmatrix},$$

defines a Householder transformation that produces zeros in the lower part of $\boldsymbol{x}$.

Exercise 10.6 (Examples of Householder transformations)
If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$ and $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{y} \neq \boldsymbol{0}$ then it follows from Exercise 10.2 that $\left(\boldsymbol{I} - 2\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}\right)\boldsymbol{x} = \boldsymbol{y}$. Use this to construct a Householder transformation $\boldsymbol{H}$ such that $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{y}$ 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 10.7 (2×2 Householder transformation)
Show that a real 2×2 Householder transformation can be written in the form

$$\boldsymbol{H} = \begin{bmatrix} -\cos\phi & \sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

Find $\boldsymbol{Hx}$ if $\boldsymbol{x} = [\cos\phi, \sin\phi]^T$.

10.2 Householder Triangulation

We say that a matrix $\boldsymbol{R} \in \mathbb{C}^{m\times n}$ is **upper trapezoidal**, if $r_{i,j} = 0$ for $j < i$ and $i = 1, 2, \ldots, m$. Upper trapezoidal matrices corresponding to $m < n$, $m = n$, and $m > n$ look as follows:

$$\begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix}, \quad \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix}, \quad \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & 0 \end{bmatrix}.$$

In this section we consider a method for bringing a matrix to upper trapezoidal form using Householder transformations. 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 Householder transformations such that

$$\boldsymbol{A}_{n+1} := \boldsymbol{H}_n \boldsymbol{H}_{n-1} \cdots \boldsymbol{H}_1 \boldsymbol{A} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{R},$$

and where $\boldsymbol{R}_1$ is upper triangular. We define

$$\boldsymbol{A}_1 := \boldsymbol{A}, \quad \boldsymbol{A}_{k+1} = \boldsymbol{H}_k \boldsymbol{A}_k, \quad k = 1, 2, \ldots, n.$$

Suppose $\boldsymbol{A}_k$ is upper trapezoidal in its first $k-1$ columns (which is true for $k = 1$)

$$\boldsymbol{A}_k = \left[\begin{array}{ccc|ccccc} a_{1,1}^1 & \cdots & a_{1,k-1}^1 & a_{1,k}^1 & \cdots & a_{1,j}^1 & \cdots & a_{1,n}^1 \\ & \ddots & \vdots & \vdots & & \vdots & & \vdots \\ & & a_{k-1,k-1}^{k-1} & a_{k-1,k}^{k-1} & \cdots & a_{k-1,j}^{k-1} & \cdots & a_{k-1,n}^{k-1} \\ \hline & & & a_{k,k}^k & \cdots & a_{k,j}^k & \cdots & a_{k,n}^k \\ & & & \vdots & & \vdots & & \vdots \\ & & & a_{i,k}^k & \cdots & a_{i,j}^k & \cdots & a_{i,n}^k \\ & & & \vdots & & \vdots & & \vdots \\ & & & a_{m,k}^k & \cdots & a_{m,j}^k & \cdots & a_{m,n}^k \end{array}\right] = \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{0} & \boldsymbol{D}_k \end{bmatrix}. \tag{10.4}$$

Let $\hat{\boldsymbol{H}}_k := \boldsymbol{I} - \hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^*$ be a Householder transformation that maps the first column $[a_{k,k}^k, \ldots, a_{m,k}^k]^T$ of $\boldsymbol{D}_k$ to a multiple of $\boldsymbol{e}_1$, $\hat{\boldsymbol{H}}_k(\boldsymbol{D}_k \boldsymbol{e}_1) = a_k \boldsymbol{e}_1$. Using Algorithm 10.4 we have $[\hat{\boldsymbol{u}}_k, a_k] = \texttt{housegen}(\boldsymbol{D}_k \boldsymbol{e}_1)$. Then $\boldsymbol{H}_k := \left[\begin{smallmatrix} \boldsymbol{I}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{H}}_k \end{smallmatrix}\right]$ is a Householder transformation and

$$\boldsymbol{A}_{k+1} := \boldsymbol{H}_k \boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{0} & \hat{\boldsymbol{H}}_k \boldsymbol{D}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{k+1} & \boldsymbol{C}_{k+1} \\ \boldsymbol{0} & \boldsymbol{D}_{k+1} \end{bmatrix},$$

where $\boldsymbol{B}_{k+1} \in \mathbb{C}^{k \times k}$ is upper triangular and $\boldsymbol{D}_{k+1} \in \mathbb{C}^{(m-k)\times(n-k)}$. Thus $\boldsymbol{A}_{k+1}$ is upper trapezoidal in its first k columns and the reduction has been carried one step further. At the end $\boldsymbol{R} := \boldsymbol{A}_{n+1} = \left[\begin{smallmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{smallmatrix}\right]$, where $\boldsymbol{R}_1$ is upper triangular.

The process can also be applied to $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ if $m \leq n$. In this case $m - 1$ Householder transformations will suffice and $\boldsymbol{H}_{m-1} \cdots \boldsymbol{H}_1 \boldsymbol{A}$ is upper trapezoidal.

In an algorithm we can store most of the vectors $\hat{\boldsymbol{u}}_k = [u_{kk}, \ldots, u_{mk}]^T$ and $\boldsymbol{A}_k$ in $\boldsymbol{A}$. However, the elements $u_{k,k}$ and $a_k = r_{k,k}$ 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, if the looser is needed, it has to be stored in a separate vector. In the following algorithm we store $a_k = r_{k,k}$ in $\boldsymbol{A}$. We also apply the Householder transformations to a second matrix $\boldsymbol{B}$. The algorithm can then be used to solve linear systems and least squares problems with one or more right hand sides, or to compute the product of the Householder transformations by choosing $\boldsymbol{B} = \boldsymbol{I}$.

Algorithm 10.8 (Householder triangulation)
Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $\boldsymbol{B} \in \mathbb{C}^{m\times r}$ and let $s := \min(n, m-1)$. The algorithm uses `housegen` to compute Householder transformations $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_s$ such that $\boldsymbol{R} = \boldsymbol{H}_s \cdots \boldsymbol{H}_1 \boldsymbol{A}$ is upper trapezoidal and $\boldsymbol{C} = \boldsymbol{H}_s \cdots \boldsymbol{H}_1 \boldsymbol{B}$. 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);
```

Here $v = \hat{\boldsymbol{u}}_k$ and we have used $\hat{\boldsymbol{H}}_k \boldsymbol{C} = (\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^*)\boldsymbol{C} = \boldsymbol{C} - \boldsymbol{v}(\boldsymbol{v}^*\boldsymbol{C})$ for the update. The Matlab command `triu` extracts the upper triangular part of $\boldsymbol{A}$ putting zeros in rows $n+1, \ldots, m$.

10.2.1 Solving linear systems using unitary transformations

Consider now the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is square. Using Algorithm 10.8 we obtain an upper triangular system $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{c}$ that is nonsingular if $\boldsymbol{A}$ is nonsingular. Thus, it can be solved by back substitution and we have a method for solving linear systems that is an alternative to Gaussian elimination. The two methods are similar since they both reduce $\boldsymbol{A}$ to upper triangular form using certain transformations and they both work for nonsingular systems.

Which method is better? Here is a short discussion.

- Advantages with Householder:
  - Row interchanges are not necessary, but see [5].
  - Numerically stable.
- Advantages with Gauss
  - Half the number of arithmetic operations compared to Householder.
  - Row interchanges are often not necessary.
  - Usually stable (but no guarantee).

Linear systems can be constructed where Gaussian elimination will fail numerically even if row interchanges are used, see [34]. On the other hand the transformations used in Householder triangulation are unitary so the method is quite stable. So why is Gaussian elimination more popular than Householder triangulation? One reason is that the number of arithmetic operations in (10.5)

when $m = n$ is $4n^3/3 = 2G_n$, which is twice the number for Gaussian elimination. We show this below. 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. Also Gaussian elimination often wins for banded and sparse problems.

10.2.2 The number of arithmetic operations

The bulk of the work in Algorithm 10.8 is the computation of $\boldsymbol{C} - \boldsymbol{v} * (\boldsymbol{v}^T * \boldsymbol{C})$ for each k. In the real case it can be determined from the following lemma.

Lemma 10.9 (Updating a Householder transformation)
Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^n$. The computation of $\boldsymbol{A} - \boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{A})$ and $\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{v})\boldsymbol{v}^T$ both cost approximately $4mn$ arithmetic operations.

Proof. It costs $2mn$ arithmetic operations to compute $\boldsymbol{w}^T := \boldsymbol{u}^T\boldsymbol{A}$, mn arithmetic operations to compute $\boldsymbol{W} = \boldsymbol{u}\boldsymbol{w}^T$ and mn arithmetic operations for the final subtraction $\boldsymbol{A} - \boldsymbol{W}$, a total of $4mn$ arithmetic operations. Taking the transpose we obtain the same count for $\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{v})\boldsymbol{v}^T$. □

Since in Algorithm 10.8, $\boldsymbol{C} \in \mathbb{C}^{(m-k+1)\times(n+r-k)}$ and $m \geq n$ the cost of computing the update $\boldsymbol{C} - \boldsymbol{v} * (\boldsymbol{v}^T * \boldsymbol{C})$ is $4(m-k)(n+r-k)$ arithmetic operations. This implies that the work in Algorithm 10.8 can be estimated as

$$\int_0^n 4(m-k)(n+r-k)dk = 2m(n+r)^2 - \frac{2}{3}(n+r)^3. \tag{10.5}$$

For $m = n$ and $r = 0$ this gives $4n^3/3$ for the number of arithmetic operations to bring a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ to upper triangular form using Householder transformtions.

10.3 The QR Decomposition and QR Factorization

Gaussian elimination without row interchanges results in an LU factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Consider Householder triangulation of $\boldsymbol{A}$. Applying Algorithm 10.8 gives $\boldsymbol{R} = \boldsymbol{H}_{n-1} \cdots \boldsymbol{H}_1\boldsymbol{A}$ implying the factorization $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} = \boldsymbol{H}_1 \cdots \boldsymbol{H}_{n-1}$ is orthonormal and $\boldsymbol{R}$ is upper triangular. This is known as a QR-factorization of $\boldsymbol{A}$.

10.3.1 Existence

For a rectangular matrix we define the following.

Definition 10.10 (QR decomposition)
Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ with $m, n \in \mathbb{N}$. We say that $\boldsymbol{A} = \boldsymbol{QR}$ is a **QR decomposition** *of $\boldsymbol{A}$ if $\boldsymbol{Q} \in \mathbb{C}^{m,m}$ is square and unitary and $\boldsymbol{R}$ is upper trapezoidal. If $m \geq n$ then $\boldsymbol{R}$ takes the form*

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0}_{m-n,n} \end{bmatrix}$$

where $\boldsymbol{R}_1 \in \mathbb{C}^{n\times n}$ is upper triangular and $\boldsymbol{0}_{m-n,n}$ is the zero matrix with $m-n$ rows and n columns. For $m \geq n$ we call $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ a **QR factorization** *of $\boldsymbol{A}$ if $\boldsymbol{Q}_1 \in \mathbb{C}^{m\times n}$ has orthonormal columns and $\boldsymbol{R}_1 \in \mathbb{C}^{n\times n}$ is upper triangular.*

Suppose $m \geq n$. A QR factorization is obtained from a QR decomposition $\boldsymbol{A} = \boldsymbol{QR}$ 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 $\boldsymbol{Q}_1 \in \mathbb{R}^{m\times n}$ and $\boldsymbol{R}_1 \in \mathbb{R}^{n\times n}$ 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 $\boldsymbol{A} = \boldsymbol{QR}$, 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 10.11 (QR decomposition and factorization)
An example of a QR decomposition is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 7 \\ 1 & -1 & -4 \\ 1 & -1 & 2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \times \begin{bmatrix} 2 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix} = \boldsymbol{QR},$$

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 10.12 (Existence of QR decomposition)
Any matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ with $m, n \in \mathbb{N}$ has a QR decomposition.

Proof. The function `housegen(x)` returns a Householder transformation for any $\boldsymbol{x} \in \mathbb{C}^n$. Thus with $\boldsymbol{B} = \boldsymbol{I}$ in Algorithm 10.8 we obtain a QR decomposition $\boldsymbol{A} = \boldsymbol{QR}$, where $\boldsymbol{Q} = \boldsymbol{C}^* = \boldsymbol{H}_1 \cdots \boldsymbol{H}_s$, is unitary. Thus a QR decomposition always exists. □

Theorem 10.13 (Uniqueness of QR factorization)
If $m \geq n$ and $\boldsymbol{A}$ is real then the QR factorization is unique if $\boldsymbol{A}$ has linearly independent columns and $\boldsymbol{R}$ has positive diagonal elements.

Proof. Let $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ be a QR factorization of $\boldsymbol{A}$. Note that $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 = \boldsymbol{R}_1^T\boldsymbol{R}_1$. Since $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric positive definite the matrix $\boldsymbol{R}_1$ is nonsingular, and if its diagonal elements are positive this is the Cholesky factorization of $\boldsymbol{A}^T\boldsymbol{A}$. Since the Cholesky factorization is unique it follows that $\boldsymbol{R}_1$ is unique and since necessarily $\boldsymbol{Q}_1 = \boldsymbol{A}\boldsymbol{R}_1^{-1}$, it must also be unique. ∎

Example 10.14 (QR decomposition and factorization)
Consider finding the QR decomposition and factorization of the matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right]$ using the method of the uniqueness proof of Theorem 10.12. We find $\boldsymbol{B} := \boldsymbol{A}^T\boldsymbol{A} = \left[\begin{smallmatrix} 5 & -4 \\ -4 & 5 \end{smallmatrix}\right]$. The Cholesky factorization of $\boldsymbol{B} = \boldsymbol{R}^T\boldsymbol{R}$ is given by $\boldsymbol{R} = \frac{1}{\sqrt{5}}\left[\begin{smallmatrix} 5 & -4 \\ 0 & 3 \end{smallmatrix}\right]$. Now $\boldsymbol{R}^{-1} = \frac{1}{3\sqrt{5}}\left[\begin{smallmatrix} 3 & 4 \\ 0 & 5 \end{smallmatrix}\right]$ so $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{R}^{-1} = \frac{1}{\sqrt{5}}\left[\begin{smallmatrix} 2 & 1 \\ -1 & 2 \end{smallmatrix}\right]$. Since $\boldsymbol{A}$ is square $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is both the QR decomposition and QR factorization of $\boldsymbol{A}$.

The QR factorization can be used to prove a classical determinant inequality.

Theorem 10.15 (Hadamard's inequality)
For any $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{C}^{n\times n}$ we have

$$|\det(\boldsymbol{A})| \leq \prod_{j=1}^{n} \|\boldsymbol{a}_j\|_2. \tag{10.6}$$

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

$$1 = \det(\boldsymbol{I}) = \det(\boldsymbol{Q}^*\boldsymbol{Q}) = \det(\boldsymbol{Q}^*)\det(\boldsymbol{Q}) = \det(\boldsymbol{Q})^*\det(\boldsymbol{Q}) = |\det(\boldsymbol{Q})|^2$$

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. We clearly have equality if $\boldsymbol{A}$ has a zero column, for then both sides of (10.6) are zero. Suppose the columns are nonzero. We have equality if and only if $r_{jj} = \|\boldsymbol{r}_j\|_2$ 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 10.16 (QR decomposition)

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{Q} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 2 & 2 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Show that $\boldsymbol{Q}$ is orthonormal and that $\boldsymbol{QR}$ is a QR decomposition of $\boldsymbol{A}$. Find a QR factorization of $\boldsymbol{A}$.

Exercise 10.17 (Householder triangulation)

a) *Let*

$$\boldsymbol{A} := \begin{bmatrix} 1 & 0 & 1 \\ -2 & -1 & 0 \\ 2 & 2 & 1 \end{bmatrix}.$$

Find Householder transformations $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{R}^{3\times 3}$ such that $\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}$ is upper triangular.

b) *Find the QR factorization of $\boldsymbol{A}$, when $\boldsymbol{R}$ has positive diagonal elements.*

10.3.2 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 10.18 (QR and Gram-Schmidt)
Suppose $\boldsymbol{A} =\in \mathbb{R}^{m\times n}$ has rank n and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be the result of applying Gram Schmidt to the columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ of $\boldsymbol{A}$, i. e.,

$$\boldsymbol{v}_1 = \boldsymbol{a}_1, \quad \boldsymbol{v}_j = \boldsymbol{a}_j - \sum_{i=1}^{j-1} \frac{\boldsymbol{a}_j^T \boldsymbol{v}_i}{\boldsymbol{v}_i^T \boldsymbol{v}_i}\boldsymbol{v}_i, \quad \text{for } j = 2, \ldots, n. \tag{10.7}$$

Let

$$\boldsymbol{Q}_1 := [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n], \quad \boldsymbol{q}_j := \frac{\boldsymbol{v}_j}{\|\boldsymbol{v}_j\|_2}, \quad j = 1, \ldots, n \text{ and}$$

$$\boldsymbol{R}_1 := \begin{bmatrix} \|\boldsymbol{v}_1\|_2 & \boldsymbol{a}_2^T\boldsymbol{q}_1 & \boldsymbol{a}_3^T\boldsymbol{q}_1 & \cdots & \boldsymbol{a}_{n-1}^T\boldsymbol{q}_1 & \boldsymbol{a}_n^T\boldsymbol{q}_1 \\ 0 & \|\boldsymbol{v}_2\|_2 & \boldsymbol{a}_3^T\boldsymbol{q}_2 & \cdots & \boldsymbol{a}_{n-1}^T\boldsymbol{q}_2 & \boldsymbol{a}_n^T\boldsymbol{q}_2 \\ & 0 & \|\boldsymbol{v}_3\|_2 & \cdots & \boldsymbol{a}_{n-1}^T\boldsymbol{q}_3 & \boldsymbol{a}_n^T\boldsymbol{q}_3 \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & \|\boldsymbol{v}_{n-1}\|_2 & \boldsymbol{a}_n^T\boldsymbol{q}_{n-1} \\ & & & & 0 & \|\boldsymbol{v}_n\|_2 \end{bmatrix}. \tag{10.8}$$

Then $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1$ is the unique QR factorization of $\boldsymbol{A}$.

Proof. Let $\boldsymbol{Q}_1$ and $\boldsymbol{R}_1$ be given by (10.8). The matrix $\boldsymbol{Q}_1$ is well defined and has orthonormal columns, since $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}$ is an orthonormal basis for span($\boldsymbol{A}$) by Theorem 0.29. By (10.7)

$$\boldsymbol{a}_j = \boldsymbol{v}_j + \sum_{i=1}^{j-1} \frac{\boldsymbol{a}_j^T \boldsymbol{v}_i}{\boldsymbol{v}_i^T \boldsymbol{v}_i} \boldsymbol{v}_i = r_{jj}\boldsymbol{q}_j + \sum_{i=1}^{j-1} \boldsymbol{q}_i r_{ij} = \boldsymbol{Q}_1\boldsymbol{R}_1\boldsymbol{e}_j, \ j = 1, \ldots, n.$$

Clearly $\boldsymbol{R}_1$ has positive diagonal elements and the factorization is unique. ∎

Example 10.19 (QR using Gram-Schmidt)
Consider finding the QR decomposition and factorization of the matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right] = [\boldsymbol{a}_1, \boldsymbol{a}_2]$ *using Gram-Schmidt. Using* (10.7) *we find* $\boldsymbol{v}_1 = \boldsymbol{a}_1$ *and* $\boldsymbol{v}_2 = \boldsymbol{a}_2 - \frac{\boldsymbol{a}_2^T\boldsymbol{v}_1}{\boldsymbol{v}_1^T\boldsymbol{v}_1}\boldsymbol{v}_1 = \frac{3}{5}\left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$. *Thus* $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2]$, *where* $\boldsymbol{q}_1 = \frac{1}{\sqrt{5}}\left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right]$ *and* $\boldsymbol{q}_2 = \frac{1}{\sqrt{5}}\left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$. *By* (10.8) *we find*

$$\boldsymbol{R}_1 = \boldsymbol{R} = \begin{bmatrix} \|\boldsymbol{v}_1\|_2 & \boldsymbol{a}_2^T\boldsymbol{q}_1 \\ 0 & \|\boldsymbol{v}_2\|_2 \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 5 & -4 \\ 0 & 3 \end{bmatrix}$$

and this agrees with what we found in Example 10.14.

Exercise 10.20 (QR using Gram-Schmidt, II)
Construct $\boldsymbol{Q}_1$ and $\boldsymbol{R}_1$ in Example 10.11 using Gram-Schmidt orthogonalization.

Warning. 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, see [2]. Here we only considered Householder transformations (cf. Algorithm 10.8).

10.4 Givens Rotations

In some applications, the matrix we want to triangulate has a special structure. Suppose for example that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is square and upper Hessenberg as illustrated by a **Wilkinson diagram** for $n = 4$

$$\boldsymbol{A} = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix}.$$

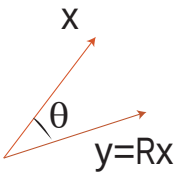


Figure 10.2. *A plane rotation.*

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 10.21 (Givens rotation, plane rotation)
A **plane rotation** *(also called a* **Given's rotation***) is a matrix* $\boldsymbol{P} \in \mathbb{R}^{2,2}$ *of the form*

$$\boldsymbol{P} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \text{ where } c^2 + s^2 = 1.$$

A plane rotation is orthonormal 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 10.22 (Plane rotation)
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 10.2.*

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 $\boldsymbol{P}_{ij} = (p_{kl}) \in \mathbb{R}^{n\times n}$ by $p_{kl} = \delta_{kl}$ except for positions ii, jj, ij, ji, which are given by

$$\begin{bmatrix} p_{ii} & p_{ij} \\ p_{ji} & p_{jj} \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \text{ where } c^2 + s^2 = 1.$$

Thus, for $n = 4$,

$$\boldsymbol{P}_{1,2} = \begin{bmatrix} c & s & 0 & 0 \\ -s & c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{P}_{13} = \begin{bmatrix} c & 0 & s & 0 \\ 0 & 1 & 0 & 0 \\ -s & 0 & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{P}_{23} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & s & c & 0 \\ 0 & -s & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$


Karl Adolf Hessenberg, 1904-1959 (left), James Wallace Givens, Jr, 1910-1993 (right)

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

$$\begin{bmatrix} \boldsymbol{B}(i,:) \\ \boldsymbol{B}(j,:) \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \boldsymbol{A}(i,:) \\ \boldsymbol{A}(j,:) \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{C}(:,i) & \boldsymbol{C}(:,j) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}(:,i) & \boldsymbol{A}(:,j) \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix}. \tag{10.9}$$

An upper Hessenberg matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ 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.

$$\boldsymbol{A} = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} \overset{\boldsymbol{P}_{12}}{\to} \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ \mathbf{0} & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} \overset{\boldsymbol{P}_{23}}{\to} \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ 0 & r_{22} & r_{23} & r_{24} \\ 0 & \mathbf{0} & x & x \\ 0 & 0 & x & x \end{bmatrix} \overset{\boldsymbol{P}_{34}}{\to} \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ 0 & r_{22} & r_{23} & r_{24} \\ 0 & 0 & r_{33} & r_{34} \\ 0 & 0 & \mathbf{0} & r_{44} \end{bmatrix}.$$

For an algorithm see Exercise 10.23.

Exercise 10.23 (Solving upper Hessenberg system using roations)
Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be upper Hessenberg and nonsingular, and let $\boldsymbol{b} \in \mathbb{R}^n$. The following algorithm solves the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ using rotations $\boldsymbol{P}_{k,k+1}$ for $k = 1, \ldots, n-1$. It uses the back solve algorithm 2.2. Determine the number of arithmetic operations of this algorithm.

Algorithm 10.24 (Upper Hessenberg linear system)

```
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),n);
```

10.5 Review Questions

10.5.1 What is a Householder transformation?

10.5.2 Why are they good for numerical work?

10.5.3 What are the main differences between solving a linear system by Gaussian elimination and Householder transformations?

10.5.4 What are the differences between a QR decomposition and a QR factorization?

10.5.5 Does any matrix have a QR decomposition?

10.5.6 What is a Givens transformation?

Chapter 11

Least Squares

Consider the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ of m equations in n unknowns. It is overdetermined, if $m > n$, square, if $m = n$, and underdetermined, if $m < n$. In either case the system can only be solved approximately if $\boldsymbol{b} \notin \text{span}(\boldsymbol{A})$. One way to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ approximately is to select a vector norm $\|\cdot\|$, say a p-norm, and look for $\boldsymbol{x} \in \mathbb{C}^n$ which minimizes $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|$. The use of the one and ∞ norm can be formulated as linear programming problems, while the Euclidian norm leads to a linear system. Only this norm is considered here.

Definition 11.1 (Least squares problem)
Suppose $m, n \in \mathbb{N}$, $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *and* $\boldsymbol{b} \in \mathbb{C}^m$. *To find* $\boldsymbol{x} \in \mathbb{C}^n$ *that minimizes* $E : \mathbb{C}^n \to \mathbb{R}$ *given by*

$$E(\boldsymbol{x}) := \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2,$$

is called the **least squares problem**. *A minimizer* $\boldsymbol{x}$ *is called a* **least squares solution**.

Since the square root function is monotone, minimizing $E(\boldsymbol{x})$ or $\sqrt{E(\boldsymbol{x})}$ is equivalent.

One way to solve the least squares problem is to write E as a quadratic function and set partial derivatives equal to zero. If $\boldsymbol{A}$ and $\boldsymbol{b}$ have real components we find

$$E(\boldsymbol{x}) := (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} - 2\boldsymbol{c}^T\boldsymbol{x} + \beta,$$

where

$$\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{A}, \quad \boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{b}, \quad \beta = \boldsymbol{b}^T\boldsymbol{b}.$$

By Lemma 9.2 all minimums are solutions of the linear system $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{c}$ or

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$$

known as the **normal equations**. The coefficient matrix is symmetric positive semidefinite. If it is positive definite then the least square problem has a unique solution. By Corollary 2.36 $\boldsymbol{A}^T\boldsymbol{A}$ is positive definite if and only if $\boldsymbol{A}$ has linearly independent columns. In particular, $m \geq n$ is necessary for a unique solution.

11.1 Numerical Examples

Example 11.2 (Average)
Consider the least squares problem defined by

$$\begin{matrix} x_1 = 1 \\ x_1 = 1, \\ x_1 = 2 \end{matrix} \quad \boldsymbol{A} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{x} = [x_1], \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix},$$

We find

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = (x_1 - 1)^2 + (x_1 - 1)^2 + (x_1 - 2)^2 = 3x_1^2 - 8x_1 + 6.$$

Setting the first derivative with respect to x_1 equal to zero we obtain $6x_1 - 8 = 0$ or $x_1 = 4/3$, the average of b_1, b_2, b_3. The second derivative is positive and $x_1 = 4/3$ is a global minimum. The normal equation is $3x_1 = 4$.

Example 11.3 (Input/output model)
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 $\min\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.

11.2 Curve Fitting

Given

- size: $1 \leq n \leq m$,
- sites: $\mathcal{S} := \{t_1, t_2, \ldots, t_m\} \subset [a, b]$,
- y-values: $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^T \in \mathbb{R}^m$,
- functions: $\phi_j : [a, b] \to \mathbb{R}$, $j = 1, \ldots, n$.

Find a function (curve fit) $p : [a, b] \to \mathbb{R}$ given by $p := \sum_{j=1}^{n} x_j \phi_j$ such that $p(t_k) \approx y_k$ for $k = 1, \ldots, m$.

An example is shown in Figure 11.1. Here $\phi_1(t) = 1$ and $\phi_2(t) = t$ and p is a straight line (linear regression).

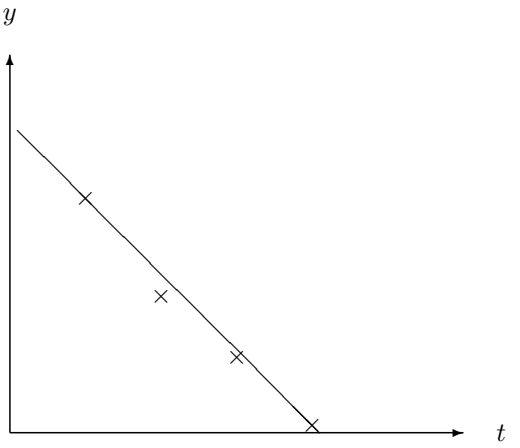


Figure 11.1. *A least squares fit to data.*

The curve fitting problem can be defined from an overdetermined linear system:

$$\begin{bmatrix} p(t_1) \\ \vdots \\ p(t_m) \end{bmatrix} = \boldsymbol{A}\boldsymbol{x} := \begin{bmatrix} \phi_1(t_1) & \cdots & \phi_n(t_1) \\ \vdots & & \vdots \\ \phi_1(t_m) & \cdots & \phi_n(t_m) \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} =: \boldsymbol{b}. \tag{11.1}$$

Then we find $\boldsymbol{x} \in \mathbb{R}^n$ as a solution of the corresponding least squares problem given by

$$E(\boldsymbol{x}) := \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \sum_{k=1}^{m} \big(\sum_{j=1}^{n} x_j \phi_j(t_k) - y_k \big)^2. \tag{11.2}$$

Typical examples of functions ϕ_j are polynomials, trigonometric functions, exponential functions, or splines.

In (11.2) one can also include **weights** $w_k > 0$ for $k = 1, \ldots, m$ and minimize

$$E(\boldsymbol{x}) := \sum_{k=1}^{m} w_k \big(\sum_{j=1}^{n} x_j \phi_j(t_k) - y_k \big)^2.$$

If y_k is an accurate observation, we can choose a large weight w_k. This will force $p(t_k) - y_k$ to be small. Similarly, a small w_k will allow $p(t_k) - y_k$ to be large. If an estimate for the standard deviation δy_k in y_k is known for each k, we can choose $w_k = 1/(\delta y_k)^2$, $k = 1, 2, \ldots, m$. For simplicity we will assume in the following that $w_k = 1$ for all k.

Lemma 11.4 (Curve fitting)
Let $\boldsymbol{A}$ be given by (11.1). *The matrix $\boldsymbol{A}^T\boldsymbol{A}$ is symmetric positive definite if and only if $\{\phi_1, \ldots, \phi_n\}$ is linearly independent on $\mathcal{S}$, i. e.,*

$$p(t_k) := \sum_{j=1}^{n} x_j \phi_j(t_k) = 0, \quad k = 1, \ldots, m \;\Rightarrow\; x_1 = \cdots = x_n = 0. \tag{11.3}$$

Proof. $\boldsymbol{A}$ is positive definite if and only if $\boldsymbol{A}$ has linearly independent columns. Since $(\boldsymbol{A}\boldsymbol{x})_k = \sum_{j=1}^{n} x_j \phi_j(t_k)$, $k = 1, \ldots, m$ this is equivalent to (11.3). □

Example 11.5 (Straight line fit)
Consider $m \geq n = 2, \phi_1(t) = 1$, and $\phi_2(t) = t$. The normal equations can be written

$$\begin{bmatrix} m & \sum t_k \\ \sum t_k & \sum t_k^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sum y_k \\ \sum t_k y_k \end{bmatrix}. \tag{11.4}$$

Here k ranges from 1 to m in the sums. Recall that a nonzero polynomial of degree at most n has at most n roots. Therefore, by the Lemma 11.4, this 2×2 system is symmetric positive definite if and only if there are at least 2 distinct sites t_k. With the data

t	1.0	2.0	3.0	4.0
y	3.1	1.8	1.0	0.1

the normal equations (11.4) *become* $\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 10.1 \end{bmatrix}$. *The data and the least squares polynomial $p(t) = x_1 + x_2 t = 3.95 - 0.98t$ are shown in Figure 11.1.*

Example 11.6 (Ill conditioning and the Hilbert matrix)
The normal equations can be extremely ill-conditioned. Consider the curve fitting problem using the polynomials $\phi_j(t) := t^{j-1}$, for $j = 1, \ldots, n$ and equidistant sites $t_k = (k-1)/(m-1)$ for $k = 1, \ldots, m$. The normal equations are $\boldsymbol{B}_n \boldsymbol{x} = \boldsymbol{c}_n$, where for $n = 3$

$$\boldsymbol{B}_3 \boldsymbol{x} := \begin{bmatrix} m & \sum t_k & \sum t_k^2 \\ \sum t_k & \sum t_k^2 & \sum t_k^3 \\ \sum t_k^2 & \sum t_k^3 & \sum t_k^4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \sum y_k \\ \sum t_k y_k \\ \sum t_k^2 y_k \end{bmatrix}.$$

$\boldsymbol{B}_n$ is symmetric positive definite if at least n of the t's are distinct. However $\boldsymbol{B}_n$ is extremely ill-conditioned even for moderate n. Indeed, $\frac{1}{m}\boldsymbol{B}_n \approx \boldsymbol{H}_n$, where $\boldsymbol{H}_n \in \mathbb{R}^{n\times n}$ is the **Hilbert Matrix** *with i, j element $1/(i+j-1)$. Thus for $n = 3$*

$$\boldsymbol{H}_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

The elements of $\frac{1}{m}\boldsymbol{B}_n$ are related to Riemann sums approximations to the elements of $\boldsymbol{H}_n$. In fact,

$$\frac{1}{m} b_{i,j} = \frac{1}{m} \sum_{k=1}^{m} t_k^{i+j-2} = \frac{1}{m} \sum_{k=1}^{m} \left(\frac{k-1}{m-1} \right)^{i+j-2} \approx \int_0^1 x^{i+j-2} dx = \frac{1}{i+j-1} = h_{i,j}.$$

The elements of $\boldsymbol{H}_n^{-1}$ are determined in Exercise 0.51. We find $K_1(\boldsymbol{H}_6) \approx 3 \cdot 10^7$. It appears that $\frac{1}{m}\boldsymbol{B}_n$ and hence $\boldsymbol{B}_n$ is ill-conditioned for moderate n at least if m is large. The cure for this problem is to use a different basis for polynomials. Orthogonal polynomials are an excellent choice. Another possibility is to use the shifted power basis $(t-\tilde{t})^{j-1}$, $j = 1, \ldots, n$, for a suitable $\tilde{t}$, see Exercise 11.8.

Exercise 11.7 (Straight line fit (linear regression))
Suppose $(t_i, y_i)_{i=1}^m$ are m points in the plane. We consider the over-determined systems

$$\textbf{(i)} \quad \begin{aligned} x_1 &= y_1 \\ x_1 &= y_2 \\ &\vdots \\ x_1 &= y_m \end{aligned} \qquad \textbf{(ii)} \quad \begin{aligned} x_1 + t_1 x_2 &= y_1 \\ x_1 + t_2 x_2 &= y_2 \\ &\vdots \\ x_1 + t_m x_2 &= y_m \end{aligned}$$

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 11.8 (Straight line fit using shifted power form)
Related to (**ii**) *in Exercise 11.7 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 11.9 (Fitting a circle to points)
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*

$$(t_i - c_1)^2 + (y_i - c_2)^2 = r^2, \ i = 1, \ldots, m, \tag{11.5}$$

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*

$$t_i x_1 + y_i x_2 + x_3 = t_i^2 + y_i^2, \ i = 1, \ldots, m, \tag{11.6}$$

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* (11.6) *from* (11.5)*. Explain how we can find* c_1, c_2, r *once* $[x_1, x_2, x_3]$ *is determined.*

b) *Formulate* (11.6) *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* (11.6) *to find the circle passing through the three points* $(1, 4), (3, 2), (1, 0)$.

11.3 Least Squares and Singular Value Decomposition and Factorization

The singular value decomposition and factorization can be used to characterize all solution of the least squares problems. We first consider orthogonal projections.

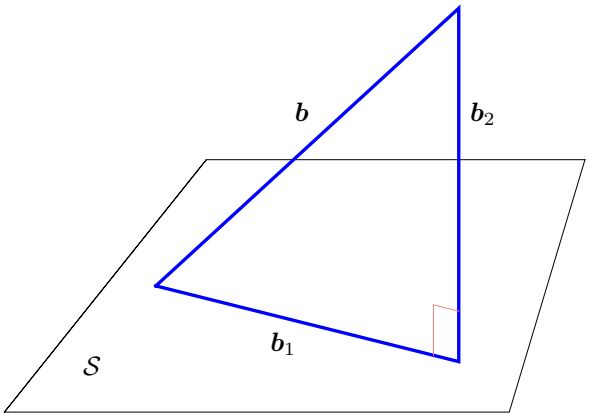


Figure 11.2. *The orthogonal projection of* $\boldsymbol{b}$ *into* $\mathcal{S}$.

11.3.1 Sum of subspaces and orthogonal projections

Suppose $\mathcal{S}$ and $\mathcal{T}$ are subspaces of $\mathbb{R}^n$ or $\mathbb{C}^n$ endowed with an inner product $\langle\ ,\ \rangle$. The following subsets are subspaces of $\mathbb{R}^n$ or $\mathbb{C}^n$.

- The set $\mathcal{S}+\mathcal{T} := \{\boldsymbol{s}+\boldsymbol{t} : \boldsymbol{s}\in\mathcal{S} \text{ and } \boldsymbol{t}\in\mathcal{T}\}$ is called the **sum** of $\mathcal{S}$ and $\mathcal{T}$.
- If $\mathcal{S}\cap\mathcal{T} = \{\boldsymbol{0}\}$ then $\mathcal{S}+\mathcal{T}$ is called a **direct sum** and denoted $\mathcal{S}\oplus\mathcal{T}$.
- If $\langle \boldsymbol{s},\boldsymbol{t}\rangle = 0$ for all $\boldsymbol{s}\in\mathcal{S}$ and $\boldsymbol{t}\in\mathcal{T}$ then $\mathcal{S}+\mathcal{T}$ is called an **orthogonal sum** and denoted $\mathcal{S}\overset{\perp}{\oplus}\mathcal{T}$.

Every $\boldsymbol{x}\in\mathcal{S}\oplus\mathcal{T}$ can be decomposed uniquely in the form $\boldsymbol{x}=\boldsymbol{s}+\boldsymbol{t}$, where $\boldsymbol{s}\in\mathcal{S}$ and $\boldsymbol{t}\in\mathcal{T}$. For if $\boldsymbol{x}=\boldsymbol{s}_1+\boldsymbol{t}_1=\boldsymbol{s}_2+\boldsymbol{t}_2$ for $\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{S}$ and $\boldsymbol{t}_1,\boldsymbol{t}_2\in\mathcal{T}$, then $\boldsymbol{s}_1-\boldsymbol{s}_2=\boldsymbol{t}_2-\boldsymbol{t}_1$ and it follows that $\boldsymbol{s}_1-\boldsymbol{s}_2$ and $\boldsymbol{t}_2-\boldsymbol{t}_1$ belong to both $\mathcal{S}$ and $\mathcal{T}$ and hence to $\mathcal{S}\cap\mathcal{T}$. But then $\boldsymbol{s}_1-\boldsymbol{s}_2=\boldsymbol{t}_2-\boldsymbol{t}_1=\boldsymbol{0}$ so $\boldsymbol{s}_1=\boldsymbol{s}_2$ and $\boldsymbol{t}_2=\boldsymbol{t}_1$.

An orthogonal sum is a direct sum. For if $\boldsymbol{b}\in\mathcal{S}\cap\mathcal{T}$ then $\boldsymbol{b}$ is orthogonal to itself, $\langle\boldsymbol{b},\boldsymbol{b}\rangle=0$, which implies that $\boldsymbol{b}=0$. Thus, every $\boldsymbol{b}\in\mathcal{S}\overset{\perp}{\oplus}\mathcal{T}$ can be written uniquely as $\boldsymbol{b}=\boldsymbol{b}_1+\boldsymbol{b}_2$, where $\boldsymbol{b}_1\in\mathcal{S}$ and $\boldsymbol{b}_2\in\mathcal{T}$. The vectors $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are called the **orthogonal projections** of $\boldsymbol{b}$ into $\mathcal{S}$ and $\mathcal{T}$. For any $\boldsymbol{s}\in\mathcal{S}$ we have $\langle\boldsymbol{b}-\boldsymbol{b}_1,\boldsymbol{s}\rangle=\langle\boldsymbol{b}_2,\boldsymbol{s}\rangle=0$, see Figure 11.2.

Consider a singular value decomposition of $\boldsymbol{A}$ and the corresponding singular value factorization:

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*=[\boldsymbol{U}_1,\boldsymbol{U}_2]\begin{bmatrix}\Sigma_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_1^*\\ \boldsymbol{V}_2^*\end{bmatrix}=\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*,\quad \Sigma_1=\operatorname{diag}(\sigma_1,\ldots,\sigma_r),$$

where $\boldsymbol{A}$ has rank r so that $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\boldsymbol{U}_1 = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]$, $\boldsymbol{V}_1 = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]$. We recall (cf. Theorem 6.16)

- the set of columns of $\boldsymbol{U}_1$ is an orthonormal basis for the column space $\operatorname{span}(\boldsymbol{A})$,
- the set of columns of $\boldsymbol{U}_2$ is an orthonormal basis for the null space $\ker(\boldsymbol{A}^*)$,

Theorem 11.10 (Orthogonal projection and least squares solution) *Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $\boldsymbol{b} \in \mathbb{C}^m$, $\mathcal{S} := \operatorname{span}(\boldsymbol{A})$ and $\mathcal{T} := \ker(\boldsymbol{A}^*)$. If $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$ is a singular value factorization of $\boldsymbol{A}$ then:*

1. *$\mathbb{C}^m = \mathcal{S} \overset{\perp}{\oplus} \mathcal{T}$ is an* **orthogonal decomposition** *of $\mathbb{C}^m$ with respect to the usual inner product $\langle \boldsymbol{s}, \boldsymbol{t}\rangle = \boldsymbol{t}^*\boldsymbol{s}$.*
2. *The orthogonal projection $\boldsymbol{b}_1$ of $\boldsymbol{b}$ into $\mathcal{S}$ is*
$$\boldsymbol{b}_1 = \boldsymbol{U}_1\boldsymbol{U}_1^*\boldsymbol{b} = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b}, \text{ where } \boldsymbol{A}^\dagger := \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^* \in \mathbb{C}^{n\times m}. \tag{11.7}$$
3. *$\boldsymbol{x} \in \mathbb{C}^n$ is a solution of the least squares problem if and only if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$. In particular the least squares problem always has a solution.*

Proof. By block multiplication $\boldsymbol{b} = \boldsymbol{U}\boldsymbol{U}^*\boldsymbol{b} = [\boldsymbol{U}_1, \boldsymbol{U}_2]\begin{bmatrix}\boldsymbol{U}_1^*\\ \boldsymbol{U}_2^*\end{bmatrix}\boldsymbol{b} = \boldsymbol{b}_1 + \boldsymbol{b}_2$, where $\boldsymbol{b}_1 = \boldsymbol{U}_1\boldsymbol{U}_1^*\boldsymbol{b} \in \mathcal{S}$ and $\boldsymbol{b}_2 = \boldsymbol{U}_2\boldsymbol{U}_2^*\boldsymbol{b} \in \mathcal{T}$. Since $\boldsymbol{U}_2^*\boldsymbol{U}_1 = \boldsymbol{0}$ it follows that $\langle \boldsymbol{b}_1, \boldsymbol{b}_2\rangle = \boldsymbol{b}_2^*\boldsymbol{b}_1 = 0$ implying Part 1. Moreover, $\boldsymbol{b}_1$ is the orthogonal projection into $\mathcal{S}$. Since $\boldsymbol{V}_1^*\boldsymbol{V}_1 = \boldsymbol{I}$ we find

$$\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b} = (\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*)(\boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^*)\boldsymbol{b} = \boldsymbol{U}_1\boldsymbol{U}_1^*\boldsymbol{b} = \boldsymbol{b}_1$$

and Part 2 follows. For any $\boldsymbol{x} \in \mathbb{C}^n$ we have $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}_1 \in \mathcal{S}$ and $\boldsymbol{b}_2^*(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}_1) = 0$ so by Pythagoras

$$\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = \|(\boldsymbol{b}_1 - \boldsymbol{A}\boldsymbol{x}) + \boldsymbol{b}_2\|_2^2 = \|\boldsymbol{b}_1 - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\boldsymbol{b}_2\|_2^2 \geq \|\boldsymbol{b}_2\|_2^2.$$

We obtain the minimum value $\|\boldsymbol{b}_2\|_2$ of $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ 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. ∎

Example 11.11 (Projections and least squares solutions) *We have the singular value factorization (cf. Example 6.13)*

$$\boldsymbol{A} := \begin{bmatrix}1 & 1\\ 1 & 1\\ 0 & 0\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\ 1\\ 0\end{bmatrix}[2]\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\end{bmatrix}.$$

We then find

$$\boldsymbol{A}^\dagger = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}[\frac{1}{2}]\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1 & 0\end{bmatrix} = \frac{1}{4}\begin{bmatrix}1 & 1 & 0\\1 & 1 & 0\end{bmatrix} = \frac{1}{4}\boldsymbol{A}^T.$$

Thus,

$$\boldsymbol{b}_1 = \boldsymbol{U}_1\boldsymbol{U}_1^*\boldsymbol{b} = \frac{1}{2}\begin{bmatrix}1\\1\\0\end{bmatrix}[1\ 1\ 0]\boldsymbol{b} = \frac{1}{2}\begin{bmatrix}1 & 1 & 0\\1 & 1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}b_1\\b_2\\b_3\end{bmatrix} = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b} = \begin{bmatrix}(b_1+b_2)/2\\(b_1+b_2)/2\\0\end{bmatrix}.$$

Moreover, the set of all least quares solutions is

$$\{\boldsymbol{x}\in\mathbb{R}^2 : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1\} = \{x_1, x_2 \in \mathbb{R} : x_1 + x_2 = \frac{b_1+b_2}{2}\}. \tag{11.8}$$

Theorem 11.12 (Uniqueness)
The least squares solution is unique if and only if $\boldsymbol{A}$ *has linearly independent columns.*

Proof. By what we just showed any solution $\boldsymbol{x}$ of the least squares problem satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$. There is a unique such $\boldsymbol{x}$ if and only if $\operatorname{rank}(\boldsymbol{A}) = n$. □

It follows that only square or overdetermined systems can have unique solutions.

Theorem 11.13 (Characterization)
The following is equivalent:

1. $\boldsymbol{x}$ *is a least squares solution,*
2. $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}$ *for some* $\boldsymbol{z}\in\mathbb{C}^n$ *with* $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$,
3. $\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^*\boldsymbol{b}$ *(normal equations).*

Proof.

1. $\implies$ 2. Le $\boldsymbol{x}$ be a least squares solution, i. e., $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}_1$. If $\boldsymbol{z} := \boldsymbol{x} - \boldsymbol{A}^\dagger\boldsymbol{b}$ 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 $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}$.
2. $\implies$ 3. If $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}$ with $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$ then

$$\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^*\boldsymbol{A}(\boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z}) = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}^*\boldsymbol{b}_1 = \boldsymbol{A}^*\boldsymbol{b}.$$

The last equality follows since $\boldsymbol{b} = \boldsymbol{b}_1 + \boldsymbol{b}_2$ and $\boldsymbol{b}_2 \in \ker(\boldsymbol{A}^*)$.

3. $\Longrightarrow$ 1. If $\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^*\boldsymbol{b}$ then $\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}_1) = \boldsymbol{0}$. But then $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}_1 \in \text{span}(\boldsymbol{A}) \cap \ker(\boldsymbol{A}^*)$ and $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}_1 = \boldsymbol{0}$. It follows that $\boldsymbol{x}$ is a least squares solution.

□

Example 11.14 (Least squares solutions)
Consider the least squares solutions $\boldsymbol{x}$ in (11.8)*. Since* $\ker(\boldsymbol{A}) = \{[\begin{smallmatrix} z \\ -z \end{smallmatrix}] : z \in \mathbb{R}\}$ *it follows from Part 2 of Theorem 11.13 that*

$$\boldsymbol{x} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} z \\ -z \end{bmatrix} = \begin{bmatrix} \frac{b_1+b_2}{4} + z \\ \frac{b_1+b_2}{4} - z \end{bmatrix}.$$

The normal equations take the form

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^T\boldsymbol{b} = \begin{bmatrix} b_1 + b_2 \\ b_1 + b_2 \end{bmatrix}.$$

Thus we obtain (11.8)*.*

11.3.2 The generalized inverse

Consider the matrix $\boldsymbol{A}^\dagger := \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^*$ in (11.7). 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 $\boldsymbol{A}$. Thus $\boldsymbol{A}^\dagger$ is a generalization of the usual inverse. The matrix $\boldsymbol{A}^\dagger$ satisfies Properties (1)- (4) in Exercise 11.15 and these properties define $\boldsymbol{A}^\dagger$ uniquely (cf. Exercise 11.16). The unique matrix $\boldsymbol{B}$ satisfying Properties (1)- (4) in Exercise 11.15 is called the **generalized inverse** or **pseudo inverse** of $\boldsymbol{A}$ and denoted $\boldsymbol{A}^\dagger$. It follows that $\boldsymbol{A}^\dagger := \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^*$ for any singular value factorization $\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^*$ of $\boldsymbol{A}$. We show in Exercise 11.18 that if $\boldsymbol{A}$ has linearly independent columns then

$$\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*. \tag{11.9}$$

Exercise 11.15 (The generalized inverse)
Show that $\boldsymbol{B} := \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^*$ 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) $(\boldsymbol{A}\boldsymbol{B})^* = \boldsymbol{A}\boldsymbol{B}$.

Exercise 11.16 (Uniqueness of generalized inverse)
Given $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, and suppose $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^{n\times m}$ satisfy

$$\begin{array}{rclcrcl} \boldsymbol{A}\boldsymbol{B}\boldsymbol{A} & = & \boldsymbol{A} & (1) & \boldsymbol{A}\boldsymbol{C}\boldsymbol{A} & = & \boldsymbol{A}, \\ \boldsymbol{B}\boldsymbol{A}\boldsymbol{B} & = & \boldsymbol{B} & (2) & \boldsymbol{C}\boldsymbol{A}\boldsymbol{C} & = & \boldsymbol{C}, \\ (\boldsymbol{A}\boldsymbol{B})^* & = & \boldsymbol{A}\boldsymbol{B} & (3) & (\boldsymbol{A}\boldsymbol{C})^* & = & \boldsymbol{A}\boldsymbol{C}, \\ (\boldsymbol{B}\boldsymbol{A})^* & = & \boldsymbol{B}\boldsymbol{A} & (4) & (\boldsymbol{C}\boldsymbol{A})^* & = & \boldsymbol{C}\boldsymbol{A}. \end{array}$$

Verify the following proof that $\boldsymbol{B} = \boldsymbol{C}$.

$$\begin{aligned}\boldsymbol{B} &= (\boldsymbol{B}\boldsymbol{A})\boldsymbol{B} = (\boldsymbol{A}^*)\boldsymbol{B}^*\boldsymbol{B} = (\boldsymbol{A}^*\boldsymbol{C}^*)\boldsymbol{A}^*\boldsymbol{B}^*\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}(\boldsymbol{A}^*\boldsymbol{B}^*)\boldsymbol{B} \\ &= \boldsymbol{C}\boldsymbol{A}(\boldsymbol{B}\boldsymbol{A}\boldsymbol{B}) = (\boldsymbol{C})\boldsymbol{A}\boldsymbol{B} = \boldsymbol{C}(\boldsymbol{A}\boldsymbol{C})\boldsymbol{A}\boldsymbol{B} = \boldsymbol{C}\boldsymbol{C}^*\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{B}) \\ &= \boldsymbol{C}\boldsymbol{C}^*(\boldsymbol{A}^*\boldsymbol{B}^*\boldsymbol{A}^*) = \boldsymbol{C}(\boldsymbol{C}^*\boldsymbol{A}^*) = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C} = \boldsymbol{C}.\end{aligned}$$

Exercise 11.17 (Verify that a matrix is a generalized inverse)
Show that the matrices $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$ *and* $\boldsymbol{B} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$ *satisfy the axioms in Exercise 11.15. Thus we can conclude that* $\boldsymbol{B} = \boldsymbol{A}^\dagger$ *without computing the singular value decomposition of* $\boldsymbol{A}$.

Exercise 11.18 (Linearly independent columns and generalized inverse)

Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ *has linearly independent columns. Show that* $\boldsymbol{A}^*\boldsymbol{A}$ *is nonsingular and* $\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$. *If* $\boldsymbol{A}$ *has linearly independent rows, then show that* $\boldsymbol{A}\boldsymbol{A}^*$ *is nonsingular and* $\boldsymbol{A}^\dagger = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}$.

Exercise 11.19 (The generalized inverse of a vector)
Show that $\boldsymbol{u}^\dagger = (\boldsymbol{u}^*\boldsymbol{u})^{-1}\boldsymbol{u}^*$ *if* $\boldsymbol{u} \in \mathbb{C}^{n,1}$ *is nonzero.*

Exercise 11.20 (The generalized inverse of an outer product)
If $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^*$ *where* $\boldsymbol{u} \in \mathbb{C}^m$, $\boldsymbol{v} \in \mathbb{C}^n$ *are nonzero, show that*

$$\boldsymbol{A}^\dagger = \frac{1}{\alpha}\boldsymbol{A}^*, \quad \alpha = \|\boldsymbol{u}\|_2^2\|\boldsymbol{v}\|_2^2.$$

Exercise 11.21 (The generalized inverse of a diagonal matrix)
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 11.22 (Properties of the generalized inverse)
Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. *Show that*

a) $(\boldsymbol{A}^*)^\dagger = (\boldsymbol{A}^\dagger)^*$.

b) $(\boldsymbol{A}^\dagger)^\dagger = \boldsymbol{A}$.

c) $(\alpha\boldsymbol{A})^\dagger = \frac{1}{\alpha}\boldsymbol{A}^\dagger, \ \alpha \neq 0$.

Exercise 11.23 (The generalized inverse of a product)
Suppose $k, m, n \in \mathbb{N}$, $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $\boldsymbol{B} \in \mathbb{C}^{n\times k}$. Suppose $\boldsymbol{A}$ has linearly independent columns and $\boldsymbol{B}$ has linearly independent rows.

a) *Show that $(\boldsymbol{AB})^\dagger = \boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$. Hint: Let $\boldsymbol{E} = \boldsymbol{AF}$, $\boldsymbol{F} = \boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$. Show by using $\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{BB}^\dagger = \boldsymbol{I}$ that $\boldsymbol{F}$ is the generalized inverse of $\boldsymbol{E}$.*

b) *Find $\boldsymbol{A} \in \mathbb{R}^{1,2}$, $\boldsymbol{B} \in \mathbb{R}^{2,1}$ such that $(\boldsymbol{AB})^\dagger \neq \boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$.*

Exercise 11.24 (The generalized inverse of the conjugate transpose)
Show that $\boldsymbol{A}^* = \boldsymbol{A}^\dagger$ if and only if all singular values of $\boldsymbol{A}$ are either zero or one.

Exercise 11.25 (Linearly independent columns)
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 11.18.)*

Exercise 11.26 (Analysis of the general linear system)
Consider the linear system $\boldsymbol{Ax} = \boldsymbol{b}$ where $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ has rank $r > 0$ and $\boldsymbol{b} \in \mathbb{C}^n$. Let

$$\boldsymbol{U}^*\boldsymbol{AV} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

represent the singular value decomposition of $\boldsymbol{A}$.

a) *Let $\boldsymbol{c} = [c_1, \ldots, c_n]^T = \boldsymbol{U}^*\boldsymbol{b}$ and $\boldsymbol{y} = [y_1, \ldots, y_n]^T = \boldsymbol{V}^*\boldsymbol{x}$. Show that $\boldsymbol{Ax} = \boldsymbol{b}$ if and only if*

$$\begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{y} = \boldsymbol{c}.$$

b) *Show that $\boldsymbol{Ax} = \boldsymbol{b}$ has a solution $\boldsymbol{x}$ if and only if $c_{r+1} = \cdots = c_n = 0$.*

c) *Deduce that a linear system $\boldsymbol{Ax} = \boldsymbol{b}$ has either no solution, one solution or infinitely many solutions.*

Exercise 11.27 (Fredholm's alternative)
For any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $\boldsymbol{b} \in \mathbb{C}^n$ show that one and only one of the following systems has a solution

$$(1) \quad \boldsymbol{Ax} = \boldsymbol{b}, \qquad (2) \quad \boldsymbol{A}^*\boldsymbol{y} = \boldsymbol{0},\ \boldsymbol{y}^*\boldsymbol{b} \neq 0.$$

In other words either $\boldsymbol{b} \in \operatorname{span}(\boldsymbol{A})$, or we can find $\boldsymbol{y} \in \ker(\boldsymbol{A}^*)$ such that $\boldsymbol{y}^*\boldsymbol{b} \neq 0$. This is called **Fredholm's alternative**.

11.4 Numerical Solution

We assume that $m \geq n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Numerical methods can be based on normal equations, QR factorization, or Singular Value Factorization. We discuss each of these approaches in turn. Another possibility is to use an iterative method like the conjugate gradient method (cf. Exercise 9.18).

11.4.1 Normal equations

Suppose $\boldsymbol{A}$ 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 Cholesky factorization of $\boldsymbol{B}$. Consider forming the normal equations. We can use either a column oriented (inner product)- or a row oriented (outer product) approach.

1. inner product: $(\boldsymbol{A}^T\boldsymbol{A})_{i,j} = \sum_{k=1}^m a_{k,i}a_{k,j}$, $i,j = 1,\ldots,n$,
$(\boldsymbol{A}^T\boldsymbol{b})_i = \sum_{k=1}^m a_{k,i}b_k$, $i = 1,\ldots,n$,

2. outer product: $\boldsymbol{A}^T\boldsymbol{A} = \sum_{k=1}^m \begin{bmatrix} a_{k1} \\ \vdots \\ a_{kn} \end{bmatrix} [a_{k1} \ \cdots \ a_{kn}]$, $\boldsymbol{A}^T\boldsymbol{b} = \sum_{k=1}^m \begin{bmatrix} a_{k1} \\ \vdots \\ a_{kn} \end{bmatrix} b_k$.

The outer product form is suitable for large problems since it uses only one pass through the data importing one row of $\boldsymbol{A}$ at a time from some separate storage.

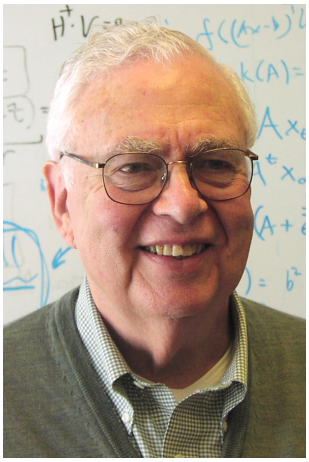
Consider the number of operations to find the least squares solution. We need $2m$ arithmetic operations for each inner product. Since $\boldsymbol{B}$ is symmetric we only need to compute $n(n+1)/2$ such inner products. It follows that $\boldsymbol{B}$ can be computed in approximately mn^2 arithmetic operations. In conclusion the number of operations are mn^2 to find $\boldsymbol{B}$, $2mn$ to find $\boldsymbol{A}^T\boldsymbol{b}$, $n^3/3$ to find $\boldsymbol{R}$, n^2 to solve $\boldsymbol{R}^T\boldsymbol{y} = \boldsymbol{c}$ and n^2 to solve $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$. If $m \approx n$ it takes $\frac{4}{3}n^3 = 2G_n$ arithmetic operations. If m is much bigger than n the number of operations is approximately mn^2, the work to compute $\boldsymbol{B}$.

A problem with the normal equations 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 eigenvalues of $\boldsymbol{B}$ are the square of the singular values of $\boldsymbol{A}$ so that

$$K_2(\boldsymbol{B}) = \frac{\sigma_1^2}{\sigma_n^2} = \left(\frac{\sigma_1}{\sigma_n}\right)^2 = K_2(\boldsymbol{A})^2.$$

If $\boldsymbol{A}$ is ill-conditioned, this could make the normal equations approach problematic. One difficulty which can be encountered is that the computed $\boldsymbol{A}^T\boldsymbol{A}$ might not be positive definite. See Problem 11.36 for an example.

11.4.2 QR factorization



Gene Golub, 1932-2007. He pioneered use of the QR factorization to solve least square problems.

Suppose $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has rank n and let $\boldsymbol{b} \in \mathbb{R}^m$. The QR factorization can be used to solve the least squares problem. 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 10.8. 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 rows of $\boldsymbol{R}$ and $\boldsymbol{c}$. Using also Algorithm 2.2 we have the following method to solve the full rank least squares problem.

1. `[R,c]=housetriang(A,b)`.

2. `x=rbacksolve(R(1:n,1:n),c(1:n),n)`.

Example 11.28 (Solution using QR factorization)

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 10.11. 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. It can even be extended to rank deficient problems, see [2]. 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 10.8 is quite stable.

But using Householder transformations requires more work. The leading term in the number of arithmetic operations in Algorithm 10.8 is approximately $2mn^2 - 2n^3/3$, (cf. (10.5) while the number of arithmetic operations needed to form the normal equations, taking advantage of symmetry is approximately mn^2. Thus for m much larger than n using Householder triangulation requires twice as many arithmetic operations as the approach based on the normal equations. Also, Householder triangulation have problems taking advantage of the structure in sparse problems.

11.4.3 Singular value factorization

This method can be used even if $\boldsymbol{A}$ does not have full rank. By Theorem 11.13

$$\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} + \boldsymbol{z},$$

where $\boldsymbol{A}^\dagger$ is the generalized inverse of $\boldsymbol{A}$, is a least squares solution for any $\boldsymbol{z} \in \ker(\boldsymbol{A})$. 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.

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 solution $\boldsymbol{A}^\dagger\boldsymbol{b}$.

Theorem 11.29 (Minimal solution)
The least squares solution with minimal Euclidian norm is $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b}$ *corresponding to* $\boldsymbol{z} = \boldsymbol{0}$.

Proof. Suppose $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{z}$, with $\boldsymbol{z} \in \ker(\boldsymbol{A})$. 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 $\boldsymbol{A}^\dagger = \boldsymbol{V}_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{U}_1^*$ 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 by Pythagoras $\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{z}\|_2^2 = \|\boldsymbol{A}^\dagger \boldsymbol{b}\|_2^2 + \|\boldsymbol{z}\|_2^2 \geq \|\boldsymbol{A}^\dagger \boldsymbol{b}\|_2^2$ with equality for $\boldsymbol{z} = \boldsymbol{0}$. ∎

Using MATLAB a least squares solution can be found using `x=A\b` if $\boldsymbol{A}$ has full rank.For rank deficient problems the function `x=lscov(A,b)` finds a least squares solution with a maximal number of zeros in $\boldsymbol{x}$.

Example 11.30 (Rank deficient least squares solution)
For $\boldsymbol{A}$ *as in Example 11.14 with* $\boldsymbol{b} = [1,1]^T$ `lscov` *gives the solution* $[1,0]^T$ *corresponding to* $\boldsymbol{z} = [1/2, -1/2]$. *The minimal norm solution is* $[1/2, 1/2]$.

11.5 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 $\boldsymbol{A}^* \boldsymbol{A} \boldsymbol{x} = \boldsymbol{A}^* \boldsymbol{b}$) as

$$\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b} = \boldsymbol{A}^\dagger \boldsymbol{b}_1, \quad \boldsymbol{A}^\dagger := (\boldsymbol{A}^* \boldsymbol{A})^{-1} \boldsymbol{A}^*,$$

where $\boldsymbol{b}_1$ is the orthogonal projection of $\boldsymbol{b}$ into the column space $\operatorname{span}(\boldsymbol{A})$.

11.5.1 Perturbing the right hand side

Let us now consider the effect of a perturbation in $\boldsymbol{b}$ on $\boldsymbol{x}$.

Theorem 11.31 (Perturbing the right hand side)
Suppose $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ *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 orthogonal projections of* $\boldsymbol{b}$ *and* $\boldsymbol{e}$ *into* $\operatorname{span}(\boldsymbol{A})$. *If* $\boldsymbol{b}_1 \neq \boldsymbol{0}$, *we have for any operator norm*

$$\frac{1}{K(\boldsymbol{A})} \frac{\|\boldsymbol{e}_1\|}{\|\boldsymbol{b}_1\|} \leq \frac{\|\boldsymbol{y} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq K(\boldsymbol{A}) \frac{\|\boldsymbol{e}_1\|}{\|\boldsymbol{b}_1\|}, \quad K(\boldsymbol{A}) = \|\boldsymbol{A}\| \|\boldsymbol{A}^\dagger\|. \tag{11.10}$$

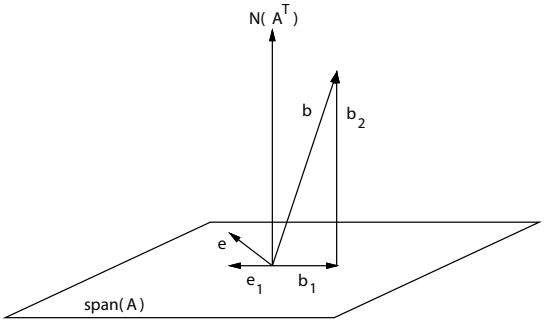


Figure 11.3. *Graphical interpretation of the bounds in Theorem 11.31.*

Proof. Subtracting $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b}_1$ from $\boldsymbol{y} = \boldsymbol{A}^\dagger \boldsymbol{b}_1 + \boldsymbol{A}^\dagger \boldsymbol{e}_1$ we have $\boldsymbol{y} - \boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{e}_1$. 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. ∎

(11.10) is analogous to the bound (7.13) 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 (11.10) and (7.13) is however that $\|\boldsymbol{e}\| / \|\boldsymbol{b}\|$ in (7.13) has been replaced by $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$, the orthogonal projections of $\boldsymbol{e}$ and $\boldsymbol{b}$ into span$(\boldsymbol{A})$. If $\boldsymbol{b}$ lies almost entirely in $\ker(\boldsymbol{A}^*)$, i.e. $\|\boldsymbol{b}\| / \|\boldsymbol{b}_1\|$ is large, then $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$ can be much larger than $\|\boldsymbol{e}\| / \|\boldsymbol{b}\|$. This is illustrated in Figure 11.3. If $\boldsymbol{b}$ is almost orthogonal to span$(\boldsymbol{A})$, $\|\boldsymbol{e}_1\| / \|\boldsymbol{b}_1\|$ will normally be much larger than $\|\boldsymbol{e}\| / \|\boldsymbol{b}\|$.

Example 11.32 (Perturbing the right hand side)
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,

$$\boldsymbol{A}^T \boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \ (\boldsymbol{A}^T \boldsymbol{A})^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \ \boldsymbol{A}^\dagger = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

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 $\boldsymbol{e}_1$. We find $\boldsymbol{A}\boldsymbol{A}^\dagger = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right]$. Hence

$$\boldsymbol{b}_1 = \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{b} = [10^{-4}, 0, 0]^T, \quad \text{and} \quad \boldsymbol{e}_1 = \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{e} = [10^{-6}, 0, 0]^T.$$

Thus $\|\boldsymbol{e}_1\|_\infty/\|\boldsymbol{b}_1\|_\infty = 10^{-2}$ *and* (11.10) *takes the form*

$$\frac{1}{4}10^{-2} \leq \frac{\|\boldsymbol{y}-\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq 4 \cdot 10^{-2}.$$

To verify the bounds we compute the solutions as $\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b} = [10^{-4}, 0]^T$ *and* $\boldsymbol{y} = \boldsymbol{A}^\dagger(\boldsymbol{b}+\boldsymbol{e}) = [10^{-4}+10^{-6}, 0]^T$. *Hence*

$$\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \frac{10^{-6}}{10^{-4}} = 10^{-2},$$

Exercise 11.33 (Condition number)
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* $\operatorname{span}(\boldsymbol{A})$ *and* $\ker(\boldsymbol{A}^T)$.

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 (11.10). The lower bound is best possible in a similar way.

Exercise 11.34 (Equality in perturbation bound)

a) *Let* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. *Show that we have equality to the right in* (11.10) *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 11.32. Find extremal* $\boldsymbol{b}$ *and* $\boldsymbol{e}$ *when the* l_∞ *norm is used.*

11.5.2 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 [20, p. 51]. For other estimates see [2] and [28].

Theorem 11.35 (Perturbing the matrix)
Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{m\times n}$, $m > n$, *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 orthogonal projections into* $\operatorname{span}(\boldsymbol{A})$

and $\ker(\boldsymbol{A}^*)$ *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*

$$\rho = \frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\|\boldsymbol{x}\|_2} \leq \frac{1}{\alpha}K(1+\beta K)\frac{\|\boldsymbol{E}\|_2}{\|\boldsymbol{A}\|_2}, \quad \beta = \frac{\|\boldsymbol{b}_2\|_2}{\|\boldsymbol{b}_1\|_2}, \ K = \|\boldsymbol{A}\|_2\|\boldsymbol{A}^\dagger\|_2. \quad (11.11)$$

(11.11) 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 $\boldsymbol{A}$. β will be small if $\boldsymbol{b}$ lies almost entirely in $\text{span}(\boldsymbol{A})$, and we have approximately $\rho \leq \frac{1}{\alpha}K\|\boldsymbol{E}\|_2/\|\boldsymbol{A}\|_2$. This corresponds to the estimate (7.19) 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 $\text{span}(\boldsymbol{A})$ and that $\boldsymbol{b}_2 = \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}$ is the residual of $\boldsymbol{x}$.

Exercise 11.36 (Problem using normal equations)
Consider the least squares problems where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1+\epsilon \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}, \ \epsilon \in \mathbb{R}.$$

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). For example, if* $u = 10^{-16}$ *then* $\sqrt{u} = 10^{-8}$. *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).*

11.6 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 6.30. If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ 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.

Theorem 11.37 (Perturbation of singular values)
Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ 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

$$|\alpha_j - \beta_j| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2, \text{ for } j = 1, 2, \ldots, n. \tag{11.12}$$

Proof. Fix j and let $\mathcal{S}$ be the $n - j + 1$ dimensional subspace for which the minimum in Theorem 6.29 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 7.31.

Theorem 11.38 (Generalized inverse when perturbing the matrix)
Let $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{m\times n}$ 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 11.37 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

$$\|(\boldsymbol{A} + \boldsymbol{E})^\dagger\|_2 \leq \frac{1}{\beta_r} \leq \frac{1}{\alpha_r - \epsilon_1} = \frac{1/\alpha_r}{1 - \epsilon_1/\alpha_r} = \frac{\|\boldsymbol{A}^\dagger\|_2}{1 - \|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{E}\|_2}.$$

□

11.7 Review Questions

11.7.1 Do the normal equations always have a solution?

11.7.2 When is the least squares solution unique?

11.7.3 Express the general least squares solution in terms of the generalized inverse.

11.7.4 Consider perturbing the right-hand side in a linear equation and a least squares problem. What is the main difference in the perturbation inequalities?

11.7.5 Why does one often prefer using QR factorization instead of normal equations for solving least squares problems.

11.7.6 What is an orthogonal sum?

11.7.7 How is an orthogonal projection defined?

Part V

Eigenvalues and Eigenvectors

Chapter 12

Numerical Eigenvalue Problems

12.1 Eigenpars

Eigenpairs have applications in quantum mechanics, differential equations, elasticity in mechanics, etc, etc. Typical computational problems involve

- Finding one or a few of the eigenvalues.
- Finding one or a few of the eigenpairs.
- Finding all eigenvalues.
- Finding all eigenpairs.

In this and the next chapter we consider some numerical methods for finding one or more of the eigenvalues and eigenvectors of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. 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.

12.2 Gerschgorin's Theorem

The following theorem is useful for locating eigenvalues of an arbitrary square matrix.

Theorem 12.1 (Gerschgorin's circle theorem)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Define for $i = 1, 2, \ldots, n$

$$R_i = \{z \in \mathbb{C} : |z - a_{ii}| \le r_i\}, \quad r_i := \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|,$$

$$C_j = \{z \in \mathbb{C} : |z - a_{jj}| \le c_j\}, \quad c_j := \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|.$$

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, $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ 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

$$|\lambda - a_{ii}| = |\sum_{j\neq i} a_{ij}x_j/x_i| \le \sum_{j\neq i} |a_{ij}||x_j/x_i| \le r_i$$

since $|x_j/x_i| \le 1$ 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 12.1. R_i is called a (Gerschgorin) row disk.

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 12.2 (Gerschgorin)
Let $\boldsymbol{T} = tridiag(-1, 2, -1) \in \mathbb{R}^{m\times m}$ 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

$R_1 = R_m = \{z \in \mathbb{R} : |z-2| \le 1\}$, *and* $R_i = \{z \in \mathbb{R} : |z-2| \le 2\}, \quad i = 2, 3, \ldots, m-1.$

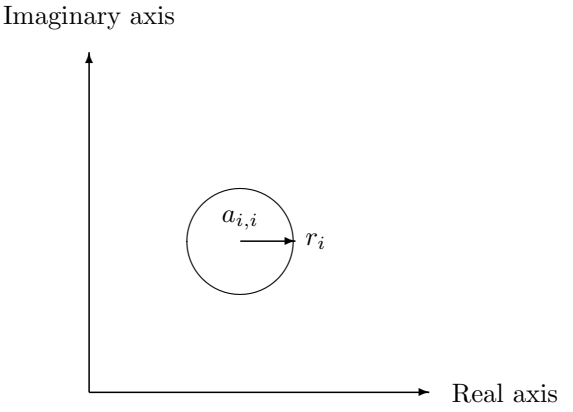


Figure 12.1. *The Gerschgorin disk R_i.*

We conclude that $\lambda \in [0,4]$ for any eigenvalue λ of $\boldsymbol{T}$. To check this, we recall that by Lemma 3.8 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 12.3 (Disjoint Gerschgorin disks)
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 12.8, see Exercise 12.5. The row disks $R_i(t)$ of $\boldsymbol{A}(t)$ have radius proportional to t, indeed

$$R_i(t) = \{z \in \mathbb{C} : |z - a_{ii}| \le t r_i\}, \quad r_i := \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}|.$$

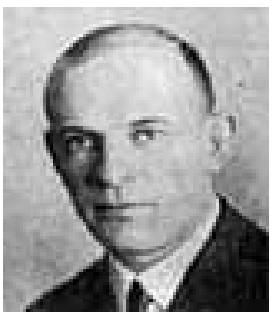
Clearly $0 \le t_1 < t_2 \le 1$ implies $R_i(t_1) \subset R_i(t_2)$ and $R_i(1)$ is a row disk of $\boldsymbol{A}$ for all i. Suppose $\bigcup_{k=1}^{p} R_{i_k}(1)$ are disjoint from the other disks of $\boldsymbol{A}$ and set $R^p(t) := \bigcup_{k=1}^{p} R_{i_k}(t)$ for $t \in [0,1]$. Now $R^p(0)$ contains 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 $\boldsymbol{A}$. □

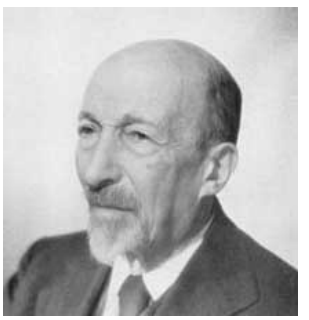
Example 12.4 *Consider the matrix* $\boldsymbol{A} = \begin{bmatrix} 1 & \epsilon_1 & \epsilon_2 \\ \epsilon_3 & 2 & \epsilon_4 \\ \epsilon_5 & \epsilon_6 & 3 \end{bmatrix}$, *where* $|\epsilon_i| \le 10^{-15}$ *all* i. *By Corollary 12.3 the eigenvalues* $\lambda_1, \lambda_2, \lambda_3$ *of* $\boldsymbol{A}$ *are distinct and satisfy* $|\lambda_j - j| \le 2 \times 10^{-15}$ *for* $j = 1, 2, 3$.

Exercise 12.5 (Continuity of eigenvalues)
Suppose $t_1, t_2 \in [0,1]$ *and that* μ *is an eigenvalue of* $\boldsymbol{A}(t_2)$. *Show, using Theorem 12.8 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.





Semyon Aranovich Gershgorin, 1901-1933 (left), Jacques Salomon Hadamard, 1865-1963 (right).

Exercise 12.6 (Nonsingularity using Gerschgorin)
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 12.7 (Gerschgorin, strictly diagonally dominant matrix)
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.*

12.3 Perturbation of Eigenvalues

In this section we study the following problem. Given matrices $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n\times n}$, 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 8.28, 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 12.8 (Elsner's theorem(1985))
Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n\times n}$*. To every* $\mu \in \sigma(\boldsymbol{A}+\boldsymbol{E})$ *there is a* $\lambda \in \sigma(\boldsymbol{A})$ *such that*

$$|\mu - \lambda| \leq K\|\boldsymbol{E}\|_2^{1/n}, \quad K = \big(\|\boldsymbol{A}\|_2 + \|\boldsymbol{A}+\boldsymbol{E}\|_2\big)^{1-1/n}. \tag{12.1}$$

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

$$\|(\mu\boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_1\|_2 = \|(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{u}_1 - \boldsymbol{A}\boldsymbol{u}_1\|_2 = \|\boldsymbol{E}\boldsymbol{u}_1\|_2 \leq \|\boldsymbol{E}\|_2,$$
$$\prod_{j=2}^{n} \|(\mu\boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_j\|_2 \leq \prod_{j=2}^{n} (|\mu| + \|\boldsymbol{A}\boldsymbol{u}_j\|_2) \leq \big(\|(\boldsymbol{A}+\boldsymbol{E})\|_2 + \|\boldsymbol{A}\|_2\big)^{n-1}.$$

Using this and Hadamard's inequality (10.6) we find

$$|\mu-\lambda_1|^n \le \prod_{j=1}^{n}|\mu-\lambda_j| = |\det(\mu\boldsymbol{I}-\boldsymbol{A})| = |\det\big((\mu\boldsymbol{I}-\boldsymbol{A})[\boldsymbol{u}_1,\dots,\boldsymbol{u}_n]\big)|$$
$$\le \|(\mu\boldsymbol{I}-\boldsymbol{A})\boldsymbol{u}_1\|_2\prod_{j=2}^{n}\|(\mu\boldsymbol{I}-\boldsymbol{A})\boldsymbol{u}_j\|_2 \le \|\boldsymbol{E}\|_2\big(\|(\boldsymbol{A}+\boldsymbol{E})\|_2+\|\boldsymbol{A}\|_2\big)^{n-1}.$$

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 $\|\boldsymbol{E}\|_2$. For a more general discussion than in the following theorem see [28].

Theorem 12.9 (Absolute errors)
Suppose $\boldsymbol{A}\in\mathbb{C}^{n\times n}$ has linearly independent eigenvectors $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$ and let $\boldsymbol{X}=[\boldsymbol{x}_1,\dots,\boldsymbol{x}_n]$ 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

$$|\lambda-\mu| \le K_p(\boldsymbol{X})\|\boldsymbol{r}\|_p, \quad 1\le p\le\infty, \tag{12.2}$$

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 $\boldsymbol{E}\in\mathbb{C}^{n\times n}$ 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

$$|\lambda-\mu| \le K_p(\boldsymbol{X})\|\boldsymbol{E}\|_p, \quad 1\le p\le\infty, \tag{12.3}$$

Proof. If $\mu\in\sigma(\boldsymbol{A})$ then we can take $\lambda=\mu$ and (12.2), (12.3) 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 $\boldsymbol{D}=\operatorname{diag}(\lambda_1,\dots,\lambda_n)$ and $(\lambda_j,\boldsymbol{x}_j)$ are the eigenpairs of $\boldsymbol{A}$ for $j=1,\dots,n$. Define $\boldsymbol{D}_1:=\boldsymbol{D}-\mu\boldsymbol{I}$. Then $\boldsymbol{D}_1^{-1}=\operatorname{diag}\big((\lambda_1-\mu)^{-1},\dots,(\lambda_n-\mu)^{-1}\big)$ 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 12.11 below we obtain

$$1=\|\boldsymbol{x}\|_p=\|\boldsymbol{X}\boldsymbol{D}_1^{-1}\boldsymbol{X}^{-1}\boldsymbol{r}\|_p\le\|\boldsymbol{D}_1^{-1}\|_pK_p(\boldsymbol{X})\|\boldsymbol{r}\|_p=\frac{K_p(\boldsymbol{X})\|\boldsymbol{r}\|_p}{\min_j|\lambda_j-\mu|}.$$

But then (12.2) 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 (12.2) proves (12.3). ∎

The equation (12.3) 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 exponent $1/n$, the equation (12.3) 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 (12.3) implies the following result.

Theorem 12.10 (Perturbations, normal matrix)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is normal and let μ be an eigenvalue of $\boldsymbol{A} + \boldsymbol{E}$ for some $\boldsymbol{E} \in \mathbb{C}^{n\times n}$. 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 5.30. 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.

In the proof of Theorem 12.9 we used that the p-norm of a diagonal matrix is equal to its spectral radius.

Lemma 12.11 (p-norm of a diagonal matrix)
If $\boldsymbol{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix then $\|\boldsymbol{A}\|_p = \rho(\boldsymbol{A})$ for $1 \leq p \leq \infty$.

Proof. For $p = \infty$ the proof is left as an exercise. For any $\boldsymbol{x} \in \mathbb{C}^n$ and $p < \infty$ we have

$$\|\boldsymbol{A}\boldsymbol{x}\|_p = \|[\lambda_1 x_1, \ldots, \lambda_n x_n]^T\|_p = \Big(\sum_{j=1}^{n} |\lambda_j|^p |x_j|^p\Big)^{1/p} \leq \rho(\boldsymbol{A})\|\boldsymbol{x}\|_p.$$

Thus $\|\boldsymbol{A}\|_p = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} \leq \rho(\boldsymbol{A})$. But from Theorem 8.28 we have $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|_p$ and the proof is complete. ∎

Exercise 12.12 (∞-norm of a diagonal matrix)
Give a direct proof that $\|\boldsymbol{A}\|_\infty = \rho(\boldsymbol{A})$ if $\boldsymbol{A}$ is diagonal.

For the accuracy of an eigenvalue of small magnitude we are interested in the size of the relative error.

Theorem 12.13 (Relative errors)
Suppose in Theorem 12.9 that $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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

$$\frac{|\lambda - \mu|}{|\lambda|} \leq K_p(\boldsymbol{X})K_p(\boldsymbol{A})\frac{\|\boldsymbol{r}\|_p}{\|\boldsymbol{A}\|_p}, \quad 1 \leq p \leq \infty, \tag{12.4}$$

where $\boldsymbol{r} := \boldsymbol{A}\boldsymbol{x} - \mu\boldsymbol{x}$. If for some $\boldsymbol{E} \in \mathbb{C}^{n\times n}$ 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

$$\frac{|\lambda - \mu|}{|\lambda|} \leq K_p(\boldsymbol{X})\|\boldsymbol{A}^{-1}\boldsymbol{E}\|_p \leq K_p(\boldsymbol{X})K_p(\boldsymbol{A})\frac{\|\boldsymbol{E}\|_p}{\|\boldsymbol{A}\|_p}, \quad 1 \leq p \leq \infty, \tag{12.5}$$

Proof. Applying Theorem 8.28 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 (12.4) follows from (12.2). To prove (12.5) 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 12.9 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 (12.5). The second inequality in (12.5) follows from the submultiplicativity of the p-norm. □

12.4 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 or unitary 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 carried out.

Recall that a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is upper Hessenberg if $a_{i,j} = 0$ for $j = 1, 2, \ldots, i-2$, $i = 3, 4, \ldots, n$. We will reduce $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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 Hermitian, the matrix $\boldsymbol{A}_{n-1}$ will be Hermitian and tridiagonal. For if $\boldsymbol{A}_k^* = \boldsymbol{A}_k$ then

$$\boldsymbol{A}_{k+1}^* = (\boldsymbol{H}_k \boldsymbol{A}_k \boldsymbol{H}_k)^* = \boldsymbol{H}_k \boldsymbol{A}_k^* \boldsymbol{H}_k = \boldsymbol{A}_{k+1}.$$

Since $\boldsymbol{A}_{n-1}$ is upper Hessenberg and Hermitian, 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 $\boldsymbol{B}_k \in \mathbb{C}^{k,k}$ is upper Hessenberg, and the first $k-1$ columns of $\boldsymbol{D}_k \in \mathbb{C}^{n-k,k}$ are zero, i.e. $\boldsymbol{D}_k = [\boldsymbol{0}, \boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{d}_k]$. Let $\boldsymbol{V}_k = \boldsymbol{I} - \boldsymbol{v}_k \boldsymbol{v}_k^* \in \mathbb{C}^{n-k,n-k}$ be a Householder transformation such that $\boldsymbol{V}_k \boldsymbol{d}_k = \alpha_k \boldsymbol{e}_1$. Define

$$\boldsymbol{H}_k = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_k \end{bmatrix} \in \mathbb{C}^{n \times n}.$$

The matrix $\boldsymbol{H}_k$ is a Householder transformation, and we find

$$\begin{aligned} \boldsymbol{A}_{k+1} = \boldsymbol{H}_k \boldsymbol{A}_k \boldsymbol{H}_k &= \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \\ \boldsymbol{D}_k & \boldsymbol{E}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_k \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{B}_k & \boldsymbol{C}_k \boldsymbol{V}_k \\ \boldsymbol{V}_k \boldsymbol{D}_k & \boldsymbol{V}_k \boldsymbol{E}_k \boldsymbol{V}_k \end{bmatrix}. \end{aligned}$$

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 10.4 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 12.14 (Householder reduction to Hessenberg form) This algorithm uses Householder similarity transformations to reduce a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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 a unitary matrix $\boldsymbol{Q}$ such that $\boldsymbol{B} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$. Algorithm 10.4 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 12.15 (Number of arithmetic operations, Hessenberg reduction)

Show that the number of arithmetic operations for Algorithm 12.14 is $\frac{10}{3}n^3 = 5G_n$.

We can use the output of Algorithm 12.14 to assemble the matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{Q}$ is orthonormal and $\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$ 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 12.16 (Assemble Householder transformations) Suppose `[L,B] = hesshousegen(A)` is the output of Algorithm 12.14. This algorithm assembles an orthonormal matrix $\boldsymbol{Q}$ from the columns of $\boldsymbol{L}$ such that $\boldsymbol{B} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$ 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 12.17 (Assemble Householder transformations)
Show that the number of arithmetic operations required by Algorithm 12.16 is $\frac{4}{3}n^3 = 2G_n$.

Exercise 12.18 (Tridiagonalize a symmetric matrix)
If $\boldsymbol{A}$ *is real and symmetric we can modify Algorithm 12.14 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 [27] for a detailed algorithm.*

12.5 Computing a Selected Eigenvalue of a Symmetric Matrix

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. 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.

$$\boldsymbol{A} = \begin{bmatrix} d_1 & c_1 & & & \\ c_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-2} & d_{n-1} & c_{n-1} \\ & & & c_{n-1} & d_n \end{bmatrix}. \tag{12.6}$$

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 5.1 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 12.19 (Distinct eigevalues of a tridiagonal matrix)
An irreducible, tridiagonal and symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has n real and distinct eigenvalues.

Proof. Let $\boldsymbol{A}$ be given by (12.6). By Theorem 5.23 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

$$p_{k+1}(x) = (x - d_{k+1})p_k(x) - c_k^2 p_{k-1}(x). \tag{12.7}$$

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. ∎

12.5.1 The inertia theorem

We say that two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ are **congruent** if $\boldsymbol{A} = \boldsymbol{E}^*\boldsymbol{B}\boldsymbol{E}$ for some nonsingular matrix $\boldsymbol{E} \in \mathbb{C}^{n\times n}$. By Theorem 5.21 a Hermitian matrix $\boldsymbol{A}$ is both congruent and similar to a diagonal matrix $\boldsymbol{D}$, $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}$ 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 12.20 (Sylvester's inertia theorem)
If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ 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 $\boldsymbol{A} = \boldsymbol{E}^*\boldsymbol{B}\boldsymbol{E}$, 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 0.32). 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

$$\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{y} = \sum_{i=1}^{n} \lambda_i |y_i|^2 = \sum_{i=1}^{k} \lambda_i |x_i|^2 > 0.$$

But

$$\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}^*\boldsymbol{E}^*\boldsymbol{B}\boldsymbol{E}\boldsymbol{y} = \boldsymbol{z}^*\boldsymbol{B}\boldsymbol{z} = \sum_{i=m+1}^{n} \mu_i |z_i|^2 \leq 0,$$

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 $\boldsymbol{U}_1^*\boldsymbol{A}\boldsymbol{U}_1 = \boldsymbol{D}_1$ and $\boldsymbol{U}_2^*\boldsymbol{B}\boldsymbol{U}_2 = \boldsymbol{D}_2$ where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices. Since $\boldsymbol{A} = \boldsymbol{E}^*\boldsymbol{B}\boldsymbol{E}$, we find $\boldsymbol{D}_1 = \boldsymbol{F}^*\boldsymbol{D}_2\boldsymbol{F}$ where $\boldsymbol{F} = \boldsymbol{U}_2^*\boldsymbol{E}\boldsymbol{U}_1$ 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 12.21 (Counting eigenvalues using the LDLT factorization)
Suppose $\boldsymbol{A} = \text{tridiag}(c_i, d_i, c_i) \in \mathbb{R}^{n\times n}$ is symmetric and that $\alpha \in \mathbb{R}$ is such that $\boldsymbol{A} - \alpha\boldsymbol{I}$ has an symmetric LU 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

$$d_1(\alpha) = d_1 - \alpha, \; d_k(\alpha) = d_k - \alpha - c_{k-1}^2/d_{k-1}(\alpha), \; k = 2, 3, \ldots, n. \tag{12.8}$$

Proof. Since the diagonal elements in $\boldsymbol{R}$ in an LU factorization equal the diagonal elements in $\boldsymbol{D}$ in an $\boldsymbol{LDL}^T$ factorization we see that the formulas in (12.8) follows immediately from (1.4). 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 $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ 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 12.22 (Counitng eigenvalues)
Consider the matrix in Exercise 12.6. Determine the number of eigenvalues greater than 4.5.

Exercise 12.23 (Overflow in LDLT factorization)
Let for $n \in \mathbb{N}$

$$\boldsymbol{A}_n = \begin{bmatrix} 10 & 1 & 0 & \cdots & 0 \\ 1 & 10 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 10 & 1 \\ 0 & \cdots & 0 & 1 & 10 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

a) *Let d_k be the diagonal elements of $\boldsymbol{D}$ in a symmetric 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 12.24 (Simultaneous diagonalization)
(Simultaneous diagonalization of two symmetric matrices by a congruence transformation). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$ where $\boldsymbol{A}^T = \boldsymbol{A}$ and $\boldsymbol{B}$ is symmetric positive definite. Let $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{DU}$ where $\boldsymbol{U}$ is orthonormal and $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_n)$. Let $\hat{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{UAU}^T\boldsymbol{D}^{-1/2}$ where

$$\boldsymbol{D}^{-1/2} := \operatorname{diag}\left(d_1^{-1/2}, \ldots, d_n^{-1/2}\right).$$

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* $\boldsymbol{E}^T \boldsymbol{A} \boldsymbol{E} = \hat{\boldsymbol{D}}$, $\boldsymbol{E}^T \boldsymbol{B} \boldsymbol{E} = \boldsymbol{I}$.

For a more general result see Theorem 10.1 in [17].

12.5.2 Approximating λ_m

Corollary 12.21 can be used to determine the mth eigenvalue of $\boldsymbol{A}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. 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 greater than x. Clearly $\rho(a) = n$, $\rho(b) = 0$. Choosing a tolerance ϵ and using bisection we proceed as follows:

```
h = b − a;
for j = 1 : itmax
    c = (a + b)/2;
    if b − a < eps ∗ h
        λ = (a + b)/2; return
    end
    k = ρ(c);
    if k ≥ m a = c else b = c;
end
```

(12.9)

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 (12.8) one of the $d_k(\alpha)$ is zero. One possibility is to replace such a $d_k(\alpha)$ by a suitable small number, say $\delta_k = c_k \epsilon_M$, where ϵ_M is the Machine epsilon, typically 2×10^{-16} for Matlab. This replacement is done if $|d_k(\alpha)| < |\delta_k|$.

Exercise 12.25 (Program code for one eigenvalue)
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* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ *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* (12.9).

a) *Write a* `function k=count(c,d,x)` *which for given x counts the number of eigenvalues of $\boldsymbol{A}$ strictly greater 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 $\{(a_j, b_j]\}$ of intervals each containing λ_m. Iterate until $b_j - a_j \leq (b-a)\epsilon_M$, 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 12.26 (Determinant of upper Hessenberg matrix)
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 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 $O(n^2)$ arithmetic operations.*

b) *Show that the number of arithmetic operations is the same if partial pivoting is used.*

c) *Estimate the number of arithmetic operations if Given's rotations are used.*

d) *Compare the two methods discussing advantages and disadvantages.*

12.6 Review Questions

12.6.1 Suppose $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{C}^{n\times n}$. To every $\mu \in \sigma(\boldsymbol{A} + \boldsymbol{E})$ there is a $\lambda \in \sigma(\boldsymbol{A})$ which is in some sense close to μ.

- What is the general result (Elsner's theorem)?
- what if $\boldsymbol{A}$ is non defective?
- what if $\boldsymbol{A}$ is normal?
- what if $\boldsymbol{A}$ is Hermitian?

12.6.2 Can Gerschgorin's theorem be used to check if a matrix is nonsingular?

12.6.3 How many arithmetic operation does it take to reduce a matrix by similarity transformations to upper Hessenberg form by Householder transformations?

12.6.4 Give a condition ensuring that a tridiagonal symmetric matrix has real and distinct eigenvalues:

12.6.5 What is the content of Sylvester's inertia theorem?

12.6.6 Give an application of this theorem.

Chapter 13
The QR Algorithm

The QR algorithm is a method to find all eigenvalues and eigenvectors of a matrix. It is related to a simpler method called the power method and we start studying this method and its variants.

13.1 The Power Method and its variants

These methods can be used to compute a single eigenpair of a matrix. They also play a role in the QR algorithm.

13.1.1 The power method

The **power method** in its basic form is a technique to compute the eigenvector corresponding to the largest (in absolute value) eigenvalue of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. 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

$$\boldsymbol{z}_k := \boldsymbol{A}^k \boldsymbol{z}_0 = \boldsymbol{A}\boldsymbol{z}_{k-1}, \quad k = 1, 2, \ldots. \tag{13.1}$$

Example 13.1 (Power method)
Let

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \boldsymbol{z}_0 := \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

We find

$$\boldsymbol{z}_1 = \boldsymbol{A}\boldsymbol{z}_0 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \boldsymbol{z}_2 = \boldsymbol{A}\boldsymbol{z}_1 = \begin{bmatrix} 5 \\ -4 \end{bmatrix}, \ \cdots, \boldsymbol{z}_k = \frac{1}{2}\begin{bmatrix} 1+3^k \\ 1-3^k \end{bmatrix}, \ \cdots.$$

It follows that $2\boldsymbol{z}_k/3^k$ *converges to the eigenvector* $[1,-1]$ *corresponding to the dominant eigenvalue* $\lambda = 3$. *The sequence of Rayleigh quotients* $\{\boldsymbol{z}_k^T\boldsymbol{A}\boldsymbol{z}_k/\boldsymbol{z}_k^T\boldsymbol{z}_k\}$ *will converge to the dominant eigenvalue* $\lambda = 3$.

To understand better what happens we expand $\boldsymbol{z}_0$ *in terms of the eigenvectors*

$$\boldsymbol{z}_0 = \frac{1}{2}\begin{bmatrix}1\\-1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2.$$

Since $\boldsymbol{A}^k$ *has eigenpairs* $(\lambda_j^k, \boldsymbol{v}_j)$, $j = 1, 2$ *we find*

$$\boldsymbol{z}_k = c_1\lambda_1^k\boldsymbol{v}_1 + c_2\lambda_2^k\boldsymbol{v}_2 = c_1 3^k\boldsymbol{v}_1 + c_2 1^k\boldsymbol{v}_2.$$

Thus $3^{-k}\boldsymbol{z}_k = c_1\boldsymbol{v}_1 + 3^{-k}c_2\boldsymbol{v}_2 \to c_1\boldsymbol{v}_1$. *Since* $c_1 \neq 0$ *the result is convergence to the dominant eigenvector.*

Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ have eigenpairs $(\lambda_j, \boldsymbol{v}_j)$, $j = 1, \ldots, n$ with $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$.

Given $\boldsymbol{z}_0 \in \mathbb{C}^n$ we assume that

$$\begin{aligned} &(i) && |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|,\\ &(ii) && \boldsymbol{z}_0^T\boldsymbol{v}_1 \neq 0\\ &(iii) && \boldsymbol{A} \text{ has linearly independent eigenvectors.} \end{aligned} \tag{13.2}$$

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 simplify the analysis.

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 (13.2) 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

$$\boldsymbol{z}_k = c_1\lambda_1^k\boldsymbol{v}_1 + c_2\lambda_2^k\boldsymbol{v}_2 + \cdots + c_n\lambda_n^k\boldsymbol{v}_n, \quad k = 0, 1, 2, \ldots. \tag{13.3}$$

Dividing by λ_1^k we find

$$\frac{\boldsymbol{z}_k}{\lambda_1^k} = c_1\boldsymbol{v}_1 + c_2\Big(\frac{\lambda_2}{\lambda_1}\Big)^k\boldsymbol{v}_2 + \cdots + c_n\Big(\frac{\lambda_n}{\lambda_1}\Big)^k\boldsymbol{v}_n, \quad k = 0, 1, 2, \ldots. \tag{13.4}$$

Assumption (i) of (13.2) implies that $(\lambda_j/\lambda_1)^k \to 0$ as $k \to \infty$ for all $j \geq 2$ and we obtain

$$\lim_{k\to\infty}\frac{\boldsymbol{z}_k}{\lambda_1^k} = c_1\boldsymbol{v}_1, \tag{13.5}$$

the dominant eigenvector of $\boldsymbol{A}$. It can be shown that this also holds for defective matrices as long as (i) and (ii) of (13.2) hold, see for example page 58 of [27].

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:

$$\begin{array}{ll} (i) & \boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} \\ (ii) & \boldsymbol{x}_k = \boldsymbol{y}_k/\|\boldsymbol{y}_k\|. \end{array} \tag{13.6}$$

Lemma 13.2 (Convergence of the power method)
Suppose (13.2) *holds. Then*

$$\lim_{k\to\infty} \Big(\frac{|\lambda_1|}{\lambda_1}\Big)^k \boldsymbol{x}_k = \frac{c_1}{|c_1|}\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}.$$

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

$$\boldsymbol{x}_k = \frac{\boldsymbol{z}_k}{\|\boldsymbol{z}_k\|} = \frac{c_1\lambda_1^k}{|c_1\lambda_1^k|}\frac{\boldsymbol{v}_1 + \frac{c_2}{c_1}\Big(\frac{\lambda_2}{\lambda_1}\Big)^k\boldsymbol{v}_2 + \cdots + \frac{c_n}{c_1}\Big(\frac{\lambda_n}{\lambda_1}\Big)^k\boldsymbol{v}_n}{\|\boldsymbol{v}_1 + \frac{c_2}{c_1}\Big(\frac{\lambda_2}{\lambda_1}\Big)^k\boldsymbol{v}_2 + \cdots + \frac{c_n}{c_1}\Big(\frac{\lambda_n}{\lambda_1}\Big)^k\boldsymbol{v}_n\|}, \quad k = 0, 1, 2, \ldots,$$

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 13.3 (The Rayleigh quotient minimizes the residual)
Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{u} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$, 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. Assume $\boldsymbol{u}^*\boldsymbol{u} = 1$ and extend $\boldsymbol{u}$ to an orthonormal basis $\{\boldsymbol{u}, \boldsymbol{U}\}$ for $\mathbb{C}^n$. Then $\boldsymbol{U}^*\boldsymbol{u} = \boldsymbol{0}$ and

$$\begin{bmatrix}\boldsymbol{u}^* \\ \boldsymbol{U}^*\end{bmatrix}(\boldsymbol{A}\boldsymbol{u} - \lambda\boldsymbol{u}) = \begin{bmatrix}\boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u} - \lambda\boldsymbol{u}^*\boldsymbol{u} \\ \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{u} - \lambda\boldsymbol{U}^*\boldsymbol{u}\end{bmatrix} = \begin{bmatrix}\boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u} - \lambda \\ \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{u}\end{bmatrix}.$$

By unitary invariance of the Euclidian norm

$$\rho(\lambda)^2 = |\boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u} - \lambda|^2 + \|\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{u}\|_2^2,$$

and ρ has a global minimum at $\lambda = \boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u}$. □

Exercise 13.4 (Orthogonal vectors)
Show that $\boldsymbol{u}$ *and* $\boldsymbol{A}\boldsymbol{u} - \lambda\boldsymbol{u}$ *are orthogonal when* $\lambda = \frac{\boldsymbol{u}^*\boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}^*\boldsymbol{u}}$.

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. But what does it mean to be sufficiently small? Recall that if $\boldsymbol{A}$ is nonsingular with a nonsingular eigenvector matrix $\boldsymbol{X}$ and $(\mu, \boldsymbol{u})$ is an approximate eigenpair with $\|\boldsymbol{u}\|_2 = 1$, then by (12.4) 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. Given $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, 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.

Algorithm 13.5 (The power method)

```
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 13.6 (Power method)
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* (13.2) *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* (13.4). *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 13.6 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.

13.1.2 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 (13.6) is replaced by

$$\begin{array}{ll} (i) & (\boldsymbol{A} - s\boldsymbol{I})\boldsymbol{y}_k = \boldsymbol{x}_{k-1} \\ (ii) & \boldsymbol{x}_k = \boldsymbol{y}_k / \|\boldsymbol{y}_k\|. \end{array} \tag{13.7}$$

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 computation.

13.1.3 Rayleigh quotient 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

$$s_k = \frac{\boldsymbol{y}_k^*\boldsymbol{A}\boldsymbol{y}_k}{\boldsymbol{y}_k^*\boldsymbol{y}_k} = s_{k-1} + \frac{\boldsymbol{y}_k^*(\boldsymbol{A} - s_{k-1}\boldsymbol{I})\boldsymbol{y}_k}{\boldsymbol{y}_k^*\boldsymbol{y}_k} = s_{k-1} + \frac{\boldsymbol{y}_k^*\boldsymbol{x}_{k-1}}{\boldsymbol{y}_k^*\boldsymbol{y}_k} = s_{k-1} + \rho_k,$$

$$\boldsymbol{r}_k = \boldsymbol{A}\boldsymbol{x}_k - s_k\boldsymbol{x}_k = \frac{\boldsymbol{A}\boldsymbol{y}_k - (s_{k-1} + \rho_k)\boldsymbol{y}_k}{\|\boldsymbol{y}_k\|_2} = \frac{\boldsymbol{x}_{k-1} - \rho_k\boldsymbol{y}_k}{\|\boldsymbol{y}_k\|_2} = \boldsymbol{w}_k - \rho_k\boldsymbol{x}_k.$$

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.

Given an approximation $(s, \boldsymbol{x})$ to an eigenpair $(\lambda, \boldsymbol{v})$ of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. The following 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

Algorithm 13.7 (Rayleigh quotient iteration)

```
1 function [x,s,nr]=rayleighit(A,x,s)
2 n=length(x);
3 y=(A-s*eye(n,n))\x;
4 yn=norm(y);
5 w=x/yn;
6 x=y/yn;
7 rho=x'*w;
8 s=s+rho;
9 nr=norm(w-rho*x);
```

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 13.9. *Quadratic convergence of Rayleigh quotient iteration.*

Since the shift changes from iteration to iteration the computation of $\boldsymbol{y}$ in `rayleighit` will require $O(n^3)$ arithmetic operations 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 13.8 (Rayleigh quotient iteration)
The smallest eigenvalue of the matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$ *is* $\lambda_1 = (5-\sqrt{33})/2 \approx -0.37$. *Starting with* $\boldsymbol{x} = [1,1]^T$ *and* $s = 0$ `rayleighit` *converges to this eigenvalue and corresponding eigenvector. In Table 13.9 we show the rate of convergence by iterating* `rayleighit` *5 times. The errors are approximately squared in each iteration indicating quadratic convergence.*

13.2 The basic QR Algorithm

The QR algorithm is an iterative method to compute all eigenvalues and eigenvectors of a matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. 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 $\boldsymbol{Q} \in \mathbb{C}^{n\times n}$ is unitary, $\boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{R} \in \mathbb{C}^{n\times n}$ is upper triangular.

The basic QR algorithm takes the following form:

$$\begin{array}{l} \boldsymbol{A}_1 = \boldsymbol{A} \\ \text{for } k = 1, 2, \ldots \\ \quad \boldsymbol{Q}_k\boldsymbol{R}_k = \boldsymbol{A}_k \qquad (\text{QR factorization of } \boldsymbol{A}_k) \\ \quad \boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k. \\ \text{end} \end{array} \tag{13.8}$$

The determination of the QR factorization of $\boldsymbol{A}_k$ and the computation of $\boldsymbol{R}_k\boldsymbol{Q}_k$ is called a QR step. It is not at all clear that a QR step does anything

useful. At this point, since $\boldsymbol{R}_k = \boldsymbol{Q}_k^* \boldsymbol{A}_k$ we find

$$\boldsymbol{A}_{k+1} = \boldsymbol{R}_k \boldsymbol{Q}_k = \boldsymbol{Q}_k^* \boldsymbol{A}_k \boldsymbol{Q}_k, \tag{13.9}$$

so $\boldsymbol{A}_{k+1}$ is unitary similar to $\boldsymbol{A}_k$. By induction $\boldsymbol{A}_{k+1}$ is unitary similar to $\boldsymbol{A}$. Thus, each $\boldsymbol{A}_k$ has the same eigenvalues as $\boldsymbol{A}$. We shall see that the basic QR algorithm is related to the power method.

Here are two examples to illustrate what happens.

Example 13.10 (QR iteration; real eigenvalues)
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

$$\boldsymbol{A}_2 = \boldsymbol{R}_1 \boldsymbol{Q}_1 = \frac{1}{5} \begin{bmatrix} 5 & 4 \\ 0 & 3 \end{bmatrix} * \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 14 & 3 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 2.8 & 0.6 \\ 0.6 & 1.2 \end{bmatrix}.$$

Continuing we find

$$\boldsymbol{A}_4 \approx \begin{bmatrix} 2.997 & -0.074 \\ -0.074 & 1.0027 \end{bmatrix}, \quad \boldsymbol{A}_{10} \approx \begin{bmatrix} 3.0000 & -0.0001 \\ -0.0001 & 1.0000 \end{bmatrix}$$

$\boldsymbol{A}_{10}$ *is almost diagonal and contains approximations to the eigenvalues* $\lambda_1 = 3$ *and* $\lambda_2 = 1$ *on the diagonal.*

Example 13.11 (QR iteration; complex eigenvalues)
Applying the QR iteration (13.8) *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 12.1 and Corollary 12.3 it follows that the approximations to the real eigenvalues are quite accurate. We would also expect the complex eigenvalues to have small absolute errors.*

These two examples illustrate what happens in general. The sequence $(\boldsymbol{A}_k)_k$ converges to the triangular Schur form (Cf. Theorem 5.13) if all the eigenvalues are real or the quasi-triangular Schur form (Cf. Definition 5.17) if some of the eigenvalues are complex.

13.2.1 Relation to the power method

Let us show that the basic QR algorithm is related to the power method. We obtain the QR factorization of the powers $\boldsymbol{A}^k$ as follows:

Theorem 13.12 (QR and power)
For $k = 1, 2, 3, \ldots$, *the QR factorization of* $\boldsymbol{A}^k$ *is* $\boldsymbol{A}^k = \tilde{\boldsymbol{Q}}_k \tilde{\boldsymbol{R}}_k$, *where*

$$\tilde{\boldsymbol{Q}}_k := \boldsymbol{Q}_1 \cdots \boldsymbol{Q}_k \text{ and } \tilde{\boldsymbol{R}}_k := \boldsymbol{R}_k \cdots \boldsymbol{R}_1, \tag{13.10}$$

and $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_k$, $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_k$ *are the matrices generated by the basic QR algorithm* (13.8).

Proof. By (13.9)

$$\boldsymbol{A}_k = \boldsymbol{Q}_{k-1}^* \boldsymbol{A}_{k-1} \boldsymbol{Q}_{k-1} = \boldsymbol{Q}_{k-1}^* \boldsymbol{Q}_{k-2}^* \boldsymbol{A}_{k-2} \boldsymbol{Q}_{k-2} \boldsymbol{Q}_{k-1} = \cdots = \tilde{\boldsymbol{Q}}_{k-1}^* \boldsymbol{A} \tilde{\boldsymbol{Q}}_{k-1}. \tag{13.11}$$

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 $\boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{A}_k$ and using (13.11)

$$\tilde{\boldsymbol{Q}}_k \tilde{\boldsymbol{R}}_k = \tilde{\boldsymbol{Q}}_{k-1} (\boldsymbol{Q}_k \boldsymbol{R}_k) \tilde{\boldsymbol{R}}_{k-1} = \tilde{\boldsymbol{Q}}_{k-1} \boldsymbol{A}_k \tilde{\boldsymbol{R}}_{k-1} = (\tilde{\boldsymbol{Q}}_{k-1} \tilde{\boldsymbol{Q}}_{k-1}^*) \boldsymbol{A} \tilde{\boldsymbol{Q}}_{k-1} \tilde{\boldsymbol{R}}_{k-1} = \boldsymbol{A}^k.$$

□

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 (13.6) 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 13.10 and 13.11. Indeed, what is observed in practice is that the sequence $(\tilde{\boldsymbol{Q}}_k^* \boldsymbol{A} \tilde{\boldsymbol{Q}}_k)_k$ converges to a (quasi-triangular) Schur form of $\boldsymbol{A}$.

$$\boldsymbol{A} = \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \overset{\boldsymbol{P}_{12}^*}{\rightarrow} \begin{bmatrix} x & x & x & x \\ \mathbf{x} & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \overset{\boldsymbol{P}_{23}^*}{\rightarrow} \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ 0 & \mathbf{x} & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \overset{\boldsymbol{P}_{34}^*}{\rightarrow} \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & \mathbf{x} & x \end{bmatrix}.$$

Figure 13.1. *Post multiplication in a QR step.*

13.2.2 Invariance of the Hessenberg form

One QR step requires $O(n^3)$ arithmetic operations for a matrix $\boldsymbol{A}$ of order n. By an initial reduction of $\boldsymbol{A}$ to upper Hessenberg form $\boldsymbol{H}_1$ using Algorithm 12.14, the cost of a QR step can be reduced to $O(n^2)$. Consider a QR step on $\boldsymbol{H}_1$. We first determine plane rotations $\boldsymbol{P}_{i,i+1}$, $i = 1, \ldots, n-1$ so that $\boldsymbol{P}_{n-1,n} \cdots \boldsymbol{P}_{1,2}\boldsymbol{H}_1 = \boldsymbol{R}_1$ is upper triangular. The details were described in Section 10.4. Thus $\boldsymbol{H}_1 = \boldsymbol{Q}_1\boldsymbol{R}_1$, where $\boldsymbol{Q}_1 = \boldsymbol{P}_{1,2}^* \cdots \boldsymbol{P}_{n-1,n}^*$ is a QR factorization of $\boldsymbol{H}_1$. To finish the QR step we compute $\boldsymbol{R}_1\boldsymbol{Q}_1 = \boldsymbol{R}_1\boldsymbol{P}_{1,2}^* \cdots \boldsymbol{P}_{n-1,n}^*$. This postmultiplication step is illustrated by the Wilkinson diagram in Figure 13.1.

The postmultiplication by $\boldsymbol{P}_{i,i+1}$ introduces a nonzero in position $(i+1, i)$ leaving the other elements marked by a zero in Figure 13.1 unchanged. Thus the final matrix $\boldsymbol{R}\boldsymbol{P}_{1,2}^* \cdots \boldsymbol{P}_{n-1,n}^*$ is upper Hessenberg and a QR step leaves the Hessenberg form invariant.

In conclusion, to compute $\boldsymbol{A}_{k+1}$ from $\boldsymbol{A}_k$ requires $O(n^2)$ arithmetic operations if $\boldsymbol{A}_k$ is upper Hessenberg and $O(n)$ arithmetic operations if $\boldsymbol{A}_k$ is tridiagonal.

13.2.3 Deflation

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.

To see what effect this can have on the eigenvalues of $\boldsymbol{A}$ suppose $|a_{i+1,i}^{(k)}| \leq \epsilon$. Let $\hat{\boldsymbol{A}}_k := \boldsymbol{A}_k - a_{i+1,i}^{(k)}\boldsymbol{e}_{i+1}\boldsymbol{e}_i^T$ be the matrix obtained from $\boldsymbol{A}_k$ by setting the $(i+1, i)$ element equal to zero. Since $\boldsymbol{A}_k = \tilde{\boldsymbol{Q}}_{k-1}^* \boldsymbol{A} \tilde{\boldsymbol{Q}}_{k-1}$ we have

$$\hat{\boldsymbol{A}}_k = \tilde{\boldsymbol{Q}}_{k-1}^*(\boldsymbol{A} + \boldsymbol{E})\tilde{\boldsymbol{Q}}_{k-1}, \quad \boldsymbol{E} = \tilde{\boldsymbol{Q}}_{k-1}(a_{i+1,i}^{(k)}\boldsymbol{e}_{i+1}\boldsymbol{e}_i^T)\tilde{\boldsymbol{Q}}_{k-1}^*.$$

Since $\tilde{\boldsymbol{Q}}_{k-1}$ is unitary, $\|\boldsymbol{E}\|_F = \|a_{i+1,i}^{(k)}\boldsymbol{e}_{i+1}\boldsymbol{e}_i^T\|_F = |a_{i+1,i}^{(k)}| \leq \epsilon$ and setting $a_{i+1,i}^{(k)} = 0$ amounts to a perturbation in the original $\boldsymbol{A}$ of at most ϵ. For how to chose ϵ see the discussion on page 94-95 in [27].

This deflation occurs often in practice and can with a proper implementation reduce the computation time considerably. It should be noted that to find the eigenvectors of the original matrix one has to continue with some care, see [27].

13.3 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:

$\boldsymbol{A}_1 = \boldsymbol{A}$
for $k = 1, 2, \ldots$
  Choose a shift s_k
  $\boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{A}_k - s_k \boldsymbol{I}$ (QR factorization of $\boldsymbol{A}_k - s\boldsymbol{I}$)
  $\boldsymbol{A}_{k+1} = \boldsymbol{R}_k \boldsymbol{Q}_k + s_k \boldsymbol{I}$.
end

Since $\boldsymbol{R}_k = \boldsymbol{Q}_k^*(\boldsymbol{A}_k - s_k \boldsymbol{I})$ we find

$$\boldsymbol{A}_{k+1} = \boldsymbol{Q}_k^*(\boldsymbol{A}_k - s_k \boldsymbol{I})\boldsymbol{Q}_k + s_k \boldsymbol{I} = \boldsymbol{Q}_k^* \boldsymbol{A}_k \boldsymbol{Q}_k$$

and $\boldsymbol{A}_{k+1}$ and $\boldsymbol{A}_k$ are unitary similar.

The shifted QR algorithm is related to the power method with shift, cf. Theorem 13.12 and also the inverse power method. In fact the last column of $\boldsymbol{Q}_k$ is the result of one iteration of the inverse power method to $\boldsymbol{A}^*$ with shift s_k. Indeed, since $\boldsymbol{A} - s_k \boldsymbol{I} = \boldsymbol{Q}_k \boldsymbol{R}_k$ we have $(\boldsymbol{A} - s_k \boldsymbol{I})^* = \boldsymbol{R}_k^* \boldsymbol{Q}_k^*$ and $(\boldsymbol{A} - s_k \boldsymbol{I})^* \boldsymbol{Q}_k = \boldsymbol{R}_k^*$. Thus, since $\boldsymbol{R}_k^*$ is lower triangular with n, n element $\overline{r}_{nn}^{(k)}$ we find $(\boldsymbol{A} - s_k \boldsymbol{I})^* \boldsymbol{Q}_k \boldsymbol{e}_n = \boldsymbol{R}_k^* \boldsymbol{e}_n = \overline{r}_{nn}^{(k)} \boldsymbol{e}_n$ from which the conclusion follows.

The shift $s_k := \boldsymbol{e}_n^T \boldsymbol{A}_k \boldsymbol{e}_n$ is called the **Rayleigh quotient shift**, while 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. The convergence is very fast and at least quadratic both for the Rayleigh quotient shift and the Wilkinson shift.

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**

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 $\boldsymbol{A}$.

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.

For further remarks and detailed algorithms see [27].

13.4 A Convergence Theorem

There is no theorem which proves convergence of the QR algorithm in general. The following theorem shows convergence of the basic QR algorithm under somewhat restrictive assumptions.

Theorem 13.13 (Convergence of basis QR)
Suppose in the basic QR algorithm (13.8) *that*

1. $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *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* $\boldsymbol{X}^{-1} = \boldsymbol{L}\boldsymbol{R}$.

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* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ *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 $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ be the QR factorization of $\boldsymbol{X}$. We observe that $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ is upper triangular. For since $\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{\Lambda}$ we have $\boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{\Lambda}$ so that $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{-1}$ 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

$$\boldsymbol{F}_k := \boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{L}\boldsymbol{\Lambda}^{-k}\boldsymbol{R}^{-1} = \hat{\boldsymbol{Q}}_k\hat{\boldsymbol{R}}_k, \quad \boldsymbol{G}_k := \hat{\boldsymbol{R}}_k\boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{R}, \quad \boldsymbol{D}_k := \operatorname{diag}\big(\frac{\delta_1}{|\delta_1|}, \ldots, \frac{\delta_n}{|\delta_n|}\big),$$

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

$$\begin{aligned}\boldsymbol{A}^k &= \boldsymbol{X}\boldsymbol{\Lambda}^k\boldsymbol{X}^{-1} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{L}\boldsymbol{R} = \boldsymbol{Q}(\boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{L}\boldsymbol{\Lambda}^{-k}\boldsymbol{R}^{-1})(\boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{R}) \\ &= \boldsymbol{Q}\boldsymbol{F}_k(\boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{R}) = \boldsymbol{Q}\hat{\boldsymbol{Q}}_k\hat{\boldsymbol{R}}_k(\boldsymbol{R}\boldsymbol{\Lambda}^k\boldsymbol{R}) = (\boldsymbol{Q}\hat{\boldsymbol{Q}}_k\boldsymbol{D}_k^{-1})(\boldsymbol{D}_k\boldsymbol{G}_k),\end{aligned}$$

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 13.12 $\boldsymbol{A}^k = \tilde{\boldsymbol{Q}}_k\tilde{\boldsymbol{R}}_k$ 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 $\boldsymbol{\Lambda}^k \boldsymbol{L} \boldsymbol{\Lambda}^{-k}$ is lower triangular with elements $(\frac{\lambda_i}{\lambda_j})^k l_{ij}$ on and under the diagonal. Thus for $n = 3$

$$\boldsymbol{\Lambda}^k \boldsymbol{L} \boldsymbol{\Lambda}^{-k} = \begin{bmatrix} 1 & 0 & 0 \\ (\frac{\lambda_2}{\lambda_1})^k l_{21} & 1 & 0 \\ (\frac{\lambda_3}{\lambda_1})^k l_{31} & (\frac{\lambda_3}{\lambda_2})^k l_{32} & 1 \end{bmatrix}.$$

By Assumption 2. it follows that $\boldsymbol{\Lambda}^k \boldsymbol{L} \boldsymbol{\Lambda}^{-k} \to \boldsymbol{I}$, 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 13.14 (QR convergence detail)
Use Theorem 7.31 to show that $\hat{\boldsymbol{R}}_k \to \boldsymbol{I}$ *implies* $\hat{\boldsymbol{R}}_k^{-1} \to \boldsymbol{I}$.

13.5 Review Questions

13.5.1 What is the main use of the power method?

13.5.2 Can the QR method be used to find all eigenvectors of a matrix?

13.5.3 Can the power method be used to find an eigenvalue?

13.5.4 Do the power method converge to an eigenvector corresponding to a complex eigenvalue?

13.5.5 What is the inverse power method?

13.5.6 Give a relation between the QR algorithm and the power method.

13.5.7 How can we make the basic QR algorithm converge faster?

Part VI

Appendix

Appendix A

Determinants

The first systematic treatment of determinants was given by Cauchy in 1812. He adopted the word "determinant". 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 prove the elementary properties of determinants that we need.

A.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.

Example A.1 (Properties of permutations)
It can be shown 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).*

A.2 Basic Properties of Determinants

For any $A \in \mathbb{C}^{n\times n}$ the determinant of A is defined the number

$$\det(A) = \sum_{\sigma\in S_n} \text{sign}(\sigma) a_{\sigma(1),1} a_{\sigma(2),2} \cdots a_{\sigma(n),n}. \tag{A.1}$$

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
$$\det\big(\big[\beta_1 \boldsymbol{a}_1, \beta_2 \boldsymbol{a}_2, \ldots, \beta_n \boldsymbol{a}_n)\big]\big) = \beta_1\beta_2\cdots\beta_n \det\big(\big[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)\big]\big).$$
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**
$$\det\big(\big[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k-1}, \boldsymbol{a}_k + \boldsymbol{a}'_k, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_n\big]\big) = \det\big(\big[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\big]\big) + \det\big(\big[\boldsymbol{a}_1, \ldots, \boldsymbol{a}'_k \ldots, \boldsymbol{a}_n\big]\big).$$
6. **Singular matrix** $\det(A) = 0$ if and only if A is singular.
7. **Product rule** If $A, B \in \mathbb{C}^{n\times n}$ 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

$$\begin{array}{rcl} \det(\boldsymbol{A}^T) & = & \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{\sigma^{-1}(1),1} a_{\sigma^{-1}(2),2} \cdots a_{\sigma^{-1}(n),n} \\ & = & \sum_{\sigma \in S_n} \operatorname{sign}(\sigma^{-1}) a_{\sigma^{-1}(1),1} a_{\sigma^{-1}(2),2} \cdots a_{\sigma^{-1}(n),n} \\ & = & \sum_{\sigma^{-1} \in S_n} \operatorname{sign}(\sigma^{-1}) a_{\sigma^{-1}(1),1} a_{\sigma^{-1}(2),2} \cdots a_{\sigma^{-1}(n),n} \\ & = & \det(\boldsymbol{A}). \end{array}$$

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 5.2, 0.54). 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

$$\begin{aligned}\det(\boldsymbol{AB}) &= \det(b_{1,1}\boldsymbol{a}_1, b_{1,2}\boldsymbol{a}_1) + \det(b_{2,1}\boldsymbol{a}_2, b_{1,2}\boldsymbol{a}_1) \\ &\quad + \det(b_{1,1}\boldsymbol{a}_1, b_{2,2}\boldsymbol{a}_2) + \det(b_{2,1}\boldsymbol{a}_2, b_{2,2}\boldsymbol{a}_2).\end{aligned}$$

Next we have by homogeneity

$$\begin{aligned}\det(\boldsymbol{AB}) &= b_{1,1}b_{1,2}\det(\boldsymbol{a}_1, \boldsymbol{a}_1) + b_{2,1}b_{1,2}\det(\boldsymbol{a}_2, \boldsymbol{a}_1) \\ &\quad + b_{1,1}b_{2,2}\det(\boldsymbol{a}_1, \boldsymbol{a}_2) + b_{2,1}b_{2,2}\det(\boldsymbol{a}_2, \boldsymbol{a}_2).\end{aligned}$$

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

$$S_{n,k} = \{\sigma \in S_n : \sigma(i) \leq k \text{ if } i \leq k, \text{ and } \sigma(i) \geq k{+}1 \text{ if } i \geq k{+}1\}.$$

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 $i \leq k$ then $a_{\sigma(i),i} = 0$ since it lies in the zero part of $\boldsymbol{A}$. If $\sigma(i) \leq k$ for some $i \geq k{+}1$, we must have $\sigma(j) > k$ for some $j \leq k$ 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

$$\begin{aligned} \det(\boldsymbol{A}) &= \textstyle\sum_{\hat{\rho} \in S_k} \sum_{\hat{\tau} \in S_{n-k}} \operatorname{sign}(\hat{\rho})\operatorname{sign}(\hat{\tau}) b_{\hat{\rho}(1),1} \cdots b_{\hat{\rho}(k),k} d_{\hat{\tau}(1),1} \cdots d_{\hat{\tau}(n-k),n-k} \\ &= \det(\boldsymbol{B}) \det(\boldsymbol{D}). \end{aligned}$$

□

A.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,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \boldsymbol{A}_1(\boldsymbol{b}) = \begin{bmatrix} 3 & 2 \\ 6 & 1 \end{bmatrix}, \quad \boldsymbol{A}_2(\boldsymbol{b}) = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix},$$
$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{I}_1(\boldsymbol{x}) = \begin{bmatrix} x_1 & 0 \\ x_2 & 1 \end{bmatrix} \quad \boldsymbol{I}_2(\boldsymbol{x}) = \begin{bmatrix} 1 & x_1 \\ 0 & x_2 \end{bmatrix}.$$

Theorem A.2 (Cramer's rule (1750))
Suppose $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *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* $\boldsymbol{Ax} = \boldsymbol{b}$. *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

$$\begin{aligned}\frac{\det(\boldsymbol{A}_j(\boldsymbol{b}))}{\det\big(\boldsymbol{A}\big)} &= \det(\boldsymbol{A}^{-1}\boldsymbol{A}_j(\boldsymbol{b})) \\ &= \det([\boldsymbol{A}^{-1}\boldsymbol{a}_1, \ldots, \boldsymbol{A}^{-1}\boldsymbol{a}_{j-1}, \boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1}\boldsymbol{a}_{j+1}, \ldots, \boldsymbol{A}^{-1}\boldsymbol{a}_n]\big) \\ &= \det([\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{j-1}, \boldsymbol{x}, \boldsymbol{e}_{j+1}, \ldots, \boldsymbol{e}_n]\big) = x_j,\end{aligned}$$

where we used Property 8 for the last equality. □

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,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad \boldsymbol{A}_{1,1} = \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix}, \quad \boldsymbol{A}_{1,2} = \begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix},$$
$$\boldsymbol{A}_{2,1} = \begin{bmatrix} 2 & 3 \\ 8 & 9 \end{bmatrix}, \quad \boldsymbol{A}_{2,2} = \begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix}, \qquad \text{etc.}$$

Definition A.3 (Cofactor and Adjoint)
For $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *and* $1 \leq i, j \leq n$ *the determinant* $\det(\boldsymbol{A}_{ij})$ *is called the* **cofactor** *of* a_{ij}. *The matrix* $adj(\boldsymbol{A}) \in \mathbb{C}^{n\times n}$ *with elements* $(-1)^{i+j}\det(\boldsymbol{A}_{j,i})$ *is called the* **adjoint** *of* $\boldsymbol{A}$.

Theorem A.4 (The inverse as an adjoint)
If $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *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 A.5 (The adjoint and the inverse)
For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have

$$\boldsymbol{A}\ adj(\boldsymbol{A}) = \ adj(\boldsymbol{A})\boldsymbol{A} = \det(\boldsymbol{A})\boldsymbol{I}. \tag{A.2}$$

Proof. If $\boldsymbol{A}$ is nonsingular then (A.2) follows from Theorem A.4. 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

$$(\boldsymbol{A} + \epsilon\boldsymbol{I})\ \mathrm{adj}(\boldsymbol{A} + \epsilon\boldsymbol{I}) = \ \mathrm{adj}(\boldsymbol{A} + \epsilon\boldsymbol{I})(\boldsymbol{A} + \epsilon\boldsymbol{I}) = \det(\boldsymbol{A} + \epsilon\boldsymbol{I})\boldsymbol{I}. \tag{A.3}$$

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 (A.3) to obtain (A.2). □

Corollary A.6 (Cofactor expansion)
For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have

$$\det(\boldsymbol{A}) = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(\boldsymbol{A}_{ij})\ \textit{for } i = 1, \ldots, n, \tag{A.4}$$

$$\det(\boldsymbol{A}) = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(\boldsymbol{A}_{ij})\ \textit{for } j = 1, \ldots, n. \tag{A.5}$$

Proof. By (A.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 (A.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 (A.5). □

A.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

$$\begin{aligned}\det(\boldsymbol{B}) &= \det\big(\,[\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{i-1},\boldsymbol{a}_i+\alpha\boldsymbol{a}_k,\boldsymbol{a}_{i+1},\ldots,\boldsymbol{a}_n]\,\big)\\ &\overset{5.}{=} \det(\boldsymbol{A}) + \det\big(\,[\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{i-1},\alpha\boldsymbol{a}_k,\boldsymbol{a}_{i+1},\ldots,\boldsymbol{a}_n]\,\big) \overset{7.}{=} \det(A)\end{aligned}$$

A.5 Some Useful Determinant Formulas

Suppose $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ 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 A.7 (Cauchy-Binet formula)
Let $\boldsymbol{A} \in \mathbb{C}^{m\times p}$, $\boldsymbol{B} \in \mathbb{C}^{p\times n}$ 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 $1 \le i_1 < i_2 < \cdots < i_r \le m$ and $1 \le j_1 < j_2 < \cdots < j_r \le n$. Then

$$\det\big(\boldsymbol{C}(\boldsymbol{i},\boldsymbol{j})\big) = \sum_{\boldsymbol{k}} \det\big(\boldsymbol{A}(\boldsymbol{i},\boldsymbol{k})\big) \det\big(\boldsymbol{B}(\boldsymbol{k},\boldsymbol{j})\big), \tag{A.6}$$

where we sum over all $\boldsymbol{k} = \{k_1, \ldots, k_r\}$ with $1 \le k_1 < k_2 < \cdots < k_r \le p$.

Appendix B

Computer Arithmetic

B.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 B.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 B.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 B.3 (Relative errors)
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.

B.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

$$x = \sum_{i=0}^{m} c_i \beta^i + \sum_{i=1}^{n} d_i \beta^{-i}, \quad 0 \leq c_i, d_i \leq \beta - 1.$$

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 [12],[26].

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

$$\epsilon_M := 2^{-t}, \qquad \text{machine epsilon,} \tag{B.1}$$

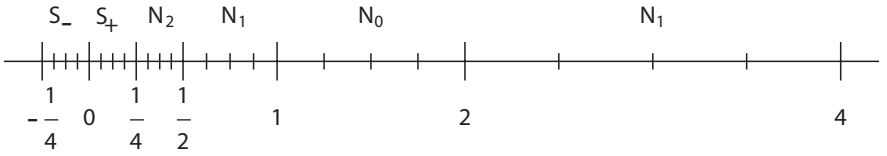


Figure B.1. *Distribution of some positive floating-point numbers*

and

$$\begin{aligned}
&\mathcal{F} := \{0\} \cup \mathcal{S} \cup \mathcal{N}, \text{where} \\
&\mathcal{N} := \mathcal{N}_+ \cup \mathcal{N}_-, \quad \mathcal{N}_+ := \cup_{e=\underline{e}}^{\overline{e}} \mathcal{N}_e, \quad \mathcal{N}_- := -\mathcal{N}_+, \\
&\mathcal{N}_e := \left\{(1.d_1 d_2 \cdots d_t)_2\right\} * 2^e = \{1, 1+\epsilon_M, 1+2\epsilon_M, \ldots, 2-\epsilon_M\} * 2^e, \\
&\mathcal{S} := \mathcal{S}_+ \cup \mathcal{S}_-, \quad \mathcal{S}_+ := \{\epsilon_M, 2\epsilon_M, 3\epsilon_M, \ldots, 1-\epsilon_M\} * 2^{\underline{e}}, \quad \mathcal{S}_- := -\mathcal{S}_+.
\end{aligned} \tag{B.2}$$

Example B.4 (Floating numbers)
Suppose $t := 2$, $\overline{e} = 3$ *and* $\underline{e} := -2$. *Then* $\epsilon_M = 1/4$ *and we find*

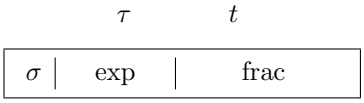
$$\begin{aligned}
&\mathcal{N}_{-2} = \{\frac{1}{4}, \frac{5}{16}, \frac{3}{8}, \frac{7}{16}\}, \quad \mathcal{N}_{-1} = \{\frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}\}, \quad \mathcal{N}_0 = \{1, \frac{5}{4}, \frac{3}{2}, \frac{3}{4}, \frac{7}{4}\}, \\
&\mathcal{N}_1 = \{2, \frac{5}{2}, 3, \frac{7}{2}\}, \quad \mathcal{N}_2 = \{4, 5, 6, 7\}, \quad \mathcal{N}_3 = \{8, 10, 12, 14\}, \\
&\mathcal{S}_+ = \{\frac{1}{16}, \frac{1}{8}, \frac{3}{16}\}, \quad \mathcal{S}_- = \{-\frac{3}{16}, -\frac{1}{8}, -\frac{1}{16}\}.
\end{aligned}$$

The position of some of these sets on the real line is shown in Figure B.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** $(1.d_1 d_2 \cdots d_t)_2$, 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

$$r_m := 2^{\underline{e}}, \quad r_M := (2 - \epsilon_M) * 2^{\overline{e}}. \tag{B.3}$$

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

$$x = (-1)^\sigma * (1.\text{frac})_2 * 2^{\text{exp}-b}, \text{ where } b := 2^{\tau-1} - 1. \tag{B.4}$$

The integer b is called the **bias**.

11. To explain the choice of b we note that the extreme values $\text{exp} = 0$ and $\text{exp} = 2^\tau - 1$ are used for special purposes. The value $\text{exp} = 0$ 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

$$\underline{e} = 1 - b, \quad \overline{e} = b := 2^{\tau-1} - 1. \tag{B.5}$$

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 (B.5) and r_M by (B.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**).

B.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 $\mathrm{fl}(x)$.

B.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 $|x| \le r_M$ 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 B.5 (Relative error in rounding)
If $r_m \le |x| \le r_M$ then

$$fl(x) = x(1+\delta), \quad |\delta| \le u_M := \frac{1}{2}\epsilon_M = 2^{-t-1}.$$

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 $|\mathrm{fl}(x) - x| \le \frac{1}{2}\epsilon_M 2^e \le \frac{1}{2}\epsilon_M * |x|$. The proof for a negative x is similar. □

The number u_M is called the **rounding unit**.

B.3.2 Arithmetic operations

Suppose $x, y \in \mathcal{N}$. In a standard system we have

$$\mathrm{fl}(x \circ y) = (x \circ y)(1+\delta), \quad |\delta| \le u_M, \quad \circ \in \{+, -, *, /, \sqrt{}\}, \tag{B.6}$$

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.

B.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 $|\delta| \le u_M$, 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, $|\delta| \le u_M$. 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.

B.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 $n = 3$

$$s_2 = \mathrm{fl}(\alpha_1 + \alpha_2) = \alpha_1(1+\delta_2) + \alpha_2(1+\delta_2),$$
$$s_3 = \mathrm{fl}(s_2 + \alpha_3) = s_2(1+\delta_3) + \alpha_3(1+\delta_3) = \alpha_1(1+\eta_1) + \alpha_2(1+\eta_2) + \alpha_3(1+\eta_3),$$
$$\eta_1 = \eta_2 = (1+\delta_2)(1+\delta_3), \quad \eta_3 = (1+\delta_3), \quad |\delta_i| \le u_M.$$

In general, with $\delta_1 := 0$,

$$\tilde{s} = \sum_{i=1}^{n} \alpha_i(1+\eta_i). \quad \eta_i = (1+\delta_i)\dots(1+\delta_n), \quad |\delta_i| \le u_M, \quad i = 1, \dots, n. \tag{B.7}$$

With $\phi(\alpha_1, \dots, \alpha_n) := \alpha_1 + \cdots + \alpha_n$ this shows that

$$\tilde{s} = \tilde{\phi}(\alpha_1, \dots, \alpha_n) = \phi(\beta_1, \dots, \beta_n), \quad \beta_i = \alpha_i(1+\eta_i). \tag{B.8}$$

The following lemma gives a convenient bound on the η factors.

Lemma B.6 (Bound on factors)
Suppose for integers k, m with $0 \le m \le k$ and $k \ge 1$ that

$$1 + \eta_k := \frac{(1+\delta_1)\cdots(1+\delta_m)}{(1+\delta_{m+1})\cdots(1+\delta_k)}, \quad |\delta_j| \le u_M, \quad j = 1, \dots, k.$$

If $ku_M \le \frac{1}{11}$ then

$$|\eta_k| \le ku'_M, \ \textit{where } u'_M := 1.1u_M. \tag{B.9}$$

Proof. We first show that

$$ku_M \le \alpha < 1 \Longrightarrow |\eta_k| \le k\frac{u_M}{1-\alpha}. \tag{B.10}$$

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 (B.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 (B.10) we obtain (B.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. [26] uses $\alpha = 1/10$ giving $u'_M = 1.12u_M$, while in [12] the value $\alpha = 0.01$ can be found. In the classical work [34] one finds $1/(1-\alpha) = 1.06$.

Let us return to the backward error (B.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

$$|\eta_i| \le (n-1)u'_M, \text{ for } i = 1, \ldots, n. \tag{B.11}$$

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

$$|\tilde{s} - s| = |\alpha_1\eta_1 + \cdots + \alpha_n\eta_n| \le (|\alpha_1| + \cdots + |\alpha_n|)(n-1)u'_M.$$

Thus the relative error in the computed sum of n numbers is bounded as follows

$$|\frac{\tilde{s}-s}{s}| \le \kappa(n-1)u'_M, \text{ where } \kappa := \frac{|\alpha_1| + \cdots + |\alpha_n|}{\alpha_1 + \cdots + \alpha_n}. \tag{B.12}$$

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-conditioned if the exact value of the sum is close to zero and some of the individual terms have large absolute values with opposite signs.

B.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

$$\boxed{\begin{array}{l} p_1 := \mathrm{fl}(\alpha_1\gamma_1) \\ \text{for } k = 2 : n \\ \quad p_k := \mathrm{fl}\big(p_{k-1} + \mathrm{fl}(\alpha_k\gamma_k)\big) \\ \text{end} \\ \tilde{p} := p_n \end{array}}$$

For a backward error analysis of this algorithm we only need to modify (B.7) slightly. All we have to do is to add terms $\mathrm{fl}(\alpha_k\gamma_k) = \alpha_k\gamma_k(1+\pi_k)$ 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

$$\left|\frac{\tilde{p}-p}{p}\right| \le \kappa n u_M, \quad \kappa := \frac{|\alpha_1\gamma_1| + \cdots + |\alpha_n\gamma_n|}{|\alpha_1\gamma_1 + \cdots + \alpha_n\gamma_n|}. \tag{B.13}$$

The computation can be ill conditioned if the exact value is close to zero and some of the components are large in absolute value.

B.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 (B.13)

$$\left|\frac{\tilde{c}_{ij} - c_{ij}}{c_{ij}}\right| \le \kappa_{ij} n u'_M, \quad \kappa_{ij} := \frac{|a_1b_1| + \cdots + |a_nb_n|}{|a_1b_1 + \cdots + a_nb_n|}, \text{ all } i, j. \tag{B.14}$$

We write this as $|\tilde{c}_{ij} - c_{ij}| \leq \kappa_{ij}|c_{ij}|nu'_M$. Using the infinity matrix norm we find

$$\sum_j |\tilde{c}_{ij} - c_{ij}| \leq nu'_M \sum_j \kappa_{ij}|c_{ij}| \leq \kappa nu'_M \sum_j |c_{ij}| \leq \kappa nu'_M \|\boldsymbol{C}\|_\infty, \text{ all } i,$$

where $\kappa := \max_{ij} \kappa_{ij}$. Maximizing over i we obtain

$$\frac{\|\tilde{\boldsymbol{C}} - \boldsymbol{C}\|_\infty}{\|\boldsymbol{C}\|_\infty} \leq \kappa nu'_M. \tag{B.15}$$

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 C

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** $\boldsymbol{H} f = \nabla \nabla^T f(\boldsymbol{x}) \in \mathbb{R}^{n,n}$ of f by

$$\nabla f := \begin{bmatrix} D_1 f \\ \vdots \\ D_n f \end{bmatrix}, \quad \boldsymbol{H} f := \nabla \nabla^T f := \begin{bmatrix} D_1 D_1 f & \cdots & D_1 D_n f \\ \vdots & & \vdots \\ D_n D_1 & \cdots & D_n D_n f \end{bmatrix}, \tag{C.1}$$

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.

Lemma C.1 (Product rules)
For $f, g : \mathbb{R}^n \to \mathbb{R}$ we have 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 $\boldsymbol{f} = [f_1, \dots f_m]^T : \mathbb{R}^n \to \mathbb{R}^m$ 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 C.2 (Second order Taylor expansion)
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

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \boldsymbol{h}^T \nabla f(\boldsymbol{x}) + \frac{1}{2} \boldsymbol{h}^T \nabla \nabla^T f(\boldsymbol{c}) \boldsymbol{h}, \text{ for some } \boldsymbol{c} \in L. \tag{C.2}$$

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 (C.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 C.3 (Functions involving matrices)
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.

□

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] E. Cohen, R. F. Riesenfeld, G. Elber, *Geometric Modeling with Splines: An Introduction*, A.K. Peters, Ltd., 2001,

[4] Demmel, J. W., Applied Numerical Linear Algebra, SIAM, Philadelphia, 1997.

[5] Golub, G. H., and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD, third edition, 1996.

[6] Grcar, Joseph F., Mathematicians of Gaussian elimination, Notices of the AMS, **58** (2011), 782–792.

[7] Greenbaum, Anne, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[8] Hackbush, Wolfgang, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, Berlin, 1994.

[9] Hall, C. A. and W. W. Meyer, Optimal error bounds for cubic spline interpolation. J. Approx. Theory, **16** (1976), 105122.

[10] Hestenes, Magnus, *Conjugate Direction Methods in Optimization*, Springer-Verlag, Berlin, 1980.

[11] 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.

[12] Higham, Nicloas J., *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[13] Horn, Roger A. and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[14] Horn, Roger A. and Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[15] Ipsen, I.C.F, Numerical Matrix Analysis: Linear Systems and Least Squares, SIAM, Philadelphia, 2009.

[16] Kato, *Perturbation Theory for Linear Operators*, Pringer.

[17] Lancaster, P., and Rodman, L., Canonical forms for hermitian matrix pairs under strict equivalence and congruence", SIAM Review, vol. 47, 2005, 407-443.

[18] Laub, A. J., Matrix Analysis for Scientists and Engineers, SIAM Philadelphia, 2005.

[19] Laub, A. J., Computational Matrix Analysis, SIAM Philadelphia, 2012.

[20] Lawson, C.L. and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N.J, 1974.

[21] Lax, Peter D., *Linear Algebra*, John Wiley & Sons, New York, 1997.

[22] Lay, D.C: Linear algebra and its applications, 2012. Addison Wesley / Pearson. Fourth edition.

[23] Leon, Steven J., *Linear Algebra with Applications*, Prentice Hall, NJ, Seventh Edition, 2006.

[24] Meyer, Carl D., *Matrix Analysis and Applied Linear Algebra* , Siam Philadelphia, 2000.

[25] Steel, J. Michael, *The Cauchy-Schwarz Master Class*, Cambridge University Press, Cambridge, UK, 2004.

[26] Stewart, G. W., *Matrix Algorithms Volume I: Basic Decompositions*, Siam Philadelphia, 1998.

[27] Stewart, G. W., *Matrix Algorithms Volume II: Eigensystems*, Siam Philadelphia, 2001.

[28] Stewart, G. W. and Ji-guang Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.

[29] Stewart, G. W., *Introduction to Matrix Computations*, Academic press, New York, 1973.

[30] Trefethen, Lloyd N., and David Bau III, *Numerical Linear Algebra*, Siam Philadelphia, 1997.

[31] Tveito, A., and R. Winther, *Partial Differential Equations*, Springer, Berlin.

[32] Van Loan, Charles, *Computational Frameworks for the Fast Fourier Transform,* Siam Philadelphia, 1992.

[33] Varga, R. S., *Matrix Iterative Analysis/ 2nd Edn.*, Springer Verlag, New York, 2000.

[34] Wilkinson, J. H., *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[35] Young, D. M.,*Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

[36] Zhang, F.,*Matrix Theory*, Springer, New York, 1999.

List of Figures

List of Tables

List of Algorithms

List of Exercises

List of Theorems

List of Corollaries

List of Lemmas

List of Definitions

Index