Aalto University

Spring 2018

# Linear Algebra

Antti Hannukainen, Nuutti Hyvönen, Vanni Noferini

May 3, 2023

# Contents

# Preface

These lecture notes cover basics of linear algebra from the standpoint of matrix computations: general vector spaces and linear maps are not in the focus of attention, but the primary objects of interest are matrices and (column) vectors that can directly be given as inputs to a computer. The students are expected to have basic knowledge about matrices; for example, the material presented during any of the courses MS-A00XX Matrix Algebra is sufficient. Many of the required fundamental tools are also reviewed in these notes.

The three Chapters 1, 2 and 3 correspond, respectively, to the three main themes of the lectures:

(1) Solvability and stability of linear systems $A\boldsymbol{x} = \boldsymbol{b}$.

(2) Eigenvalue problems $A\boldsymbol{v} = \lambda\boldsymbol{v}$ and their fundamental applications.

(3) Least squares problems, their geometric interpretation and related matrix decompositions.

For simplicity and notational convenience, Chapters 1 and 3 consider real vectors and matrices, even though the extension to the complex case would be straightforward. However, Chapter 2 touches upon the complex extension that cannot be avoided in the treatment of eigenvalue problems.

Matrices and subspaces are denoted by capital letters (e.g., $A$ or $E$), vectors by bolded lower case letters (e.g., $\boldsymbol{x}$ or $\boldsymbol{b}$) and scalars by standard lowercase letters (e.g., $\alpha$ or $a_{ij}$). In particular, the components of a vector $\boldsymbol{x} \in \mathbb{R}^n$ are $x_j$, $j = 1, \dots, n$. However, the zero scalar and the zero vector of $\mathbb{R}^n$ are both denoted simply as 0. The Cartesian basis vectors for $\mathbb{R}^n$ are $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n$ and the identity matrix is denoted as

$$I := \big[\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n\big] \in \mathbb{R}^{n \times n}.$$

We will frequently use such a notation to write matrices in terms of their column vectors. The positive integers $n$ and $m$ are reserved for spatial dimensions and $i, j, k, l \in \mathbb{N}_0$ are used as generic indices.

CHAPTER 1

# Linear systems

Let $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. In this chapter, we focus on the following simple problem: Find $\boldsymbol{x} \in \mathbb{R}^n$ such that

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

To be more precise, we tackle the following questions:

(1) Is the problem (1) solvable? Is the solution unique?

(2) How much does the solution $\boldsymbol{x}$ change if $A$ or $\boldsymbol{b}$ is perturbed slightly?

(3) How accurately can (1) be solved using a computer?

Before starting our analysis by considering the unique solvability of (1), let us make a couple of fundamental definitions that will be indispensable in what follows.

Definition 0.1. *The* linear span *of the vectors $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k \in \mathbb{R}^n$ is the set*

$$\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k) := \Big\{ \boldsymbol{z} \in \mathbb{R}^n \;\Big|\; \boldsymbol{z} = \sum_{j=1}^{k} \alpha_j \boldsymbol{q}_j \textit{ for some } \boldsymbol{\alpha} \in \mathbb{R}^k \Big\}.$$

*Moreover, $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k \in \mathbb{R}^n$ are called* linearly independent *if*

$$\sum_{j=1}^{k} \alpha_j \boldsymbol{q}_j = 0 \in \mathbb{R}^n$$

*is equivalent to $\boldsymbol{\alpha} = 0 \in \mathbb{R}^k$.*

If one stacks the vectors $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k \in \mathbb{R}^n$ as the columns of a matrix (as we will do frequently),

$$Q = \big[\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_k\big] \in \mathbb{R}^{n\times k}, \tag{2}$$

then obviously

$$\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k) = \big\{ \boldsymbol{z} \in \mathbb{R}^n \;\big|\; \boldsymbol{z} = Q\boldsymbol{\alpha} \text{ for some } \boldsymbol{\alpha} \in \mathbb{R}^k \big\}. \tag{3}$$

Moreover, $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k$ are linearly independent if and only if

$$Q\boldsymbol{\alpha} = 0 \quad \Longleftrightarrow \quad \boldsymbol{\alpha} = 0, \tag{4}$$

i.e., $Q\boldsymbol{\alpha} \in \mathbb{R}^n$ vanishes if and only if $\boldsymbol{\alpha} \in \mathbb{R}^k$ does.

## 1. Solvability of linear systems

The main aim of this section is to study the existence and uniqueness of a solution to (1). These properties can be characterized in terms of the *nullspace* and the *range* of the matrix $A$. In the following, we will denote the columns of $A$ by $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n \in \mathbb{R}^m$ (cf. (2)).

Definition 1.1. *Let $A \in \mathbb{R}^{m\times n}$. The* range *of $A$ is the set*

$$R(A) = \big\{ \boldsymbol{y} \in \mathbb{R}^m \mid \boldsymbol{y} = A\boldsymbol{x} \textit{ for some } \boldsymbol{x} \in \mathbb{R}^n \big\} = \operatorname{span}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n) \subset \mathbb{R}^m \tag{5}$$

*and the* nullspace *of $A$ is the set*

$$N(A) = \{ \boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} = 0 \} \subset \mathbb{R}^n.$$

The sets $R(A)$ and $N(A)$ are closed under addition and scalar multiplication. Indeed, if $\boldsymbol{x}, \boldsymbol{y} \in N(A)$, then

$$A(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha A\boldsymbol{x} + \beta A\boldsymbol{y} = 0,$$

meaning that also $\alpha\boldsymbol{x} + \beta\boldsymbol{y} \in N(A)$ for any $\alpha, \beta \in \mathbb{R}$. Similarly, if $\boldsymbol{x}, \boldsymbol{y} \in R(A)$, i.e. $\boldsymbol{x} = A\boldsymbol{z}$ and $\boldsymbol{y} = A\boldsymbol{w}$ for some $\boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^n$, then

$$A(\alpha\boldsymbol{z} + \beta\boldsymbol{w}) = \alpha A\boldsymbol{z} + \beta A\boldsymbol{w} = \alpha\boldsymbol{x} + \beta\boldsymbol{y},$$

i.e., $\alpha\boldsymbol{x} + \beta\boldsymbol{y} \in R(A)$ for any $\alpha, \beta \in \mathbb{R}$. This means that the subsets $R(A) \subset \mathbb{R}^m$ and $N(A) \subset \mathbb{R}^n$ are actually *subspaces.*

Definition 1.2. *Let $E \subset \mathbb{R}^n$ be nonempty and such that for any $\boldsymbol{x}, \boldsymbol{y} \in E$ and $\alpha \in \mathbb{R}$*

$$\boldsymbol{x} + \boldsymbol{y} \in E \quad \textit{and} \quad \alpha\boldsymbol{x} \in E.$$

*Then $E$ is called a* subspace *of $\mathbb{R}^n$. (In particular, $\mathbb{R}^n$ itself as well the* trivial subspace *$\{0\}$ are subspaces of $\mathbb{R}^n$.)*

The range of the matrix $A$ induces a condition for the existence of a solution to the equation (1). By definition, for any $\boldsymbol{b} \in R(A)$, there exists $\boldsymbol{z} \in \mathbb{R}^n$ such that

$$\boldsymbol{b} = A\boldsymbol{z}.$$

On the other hand, if (1) has a solution, then obviously $\boldsymbol{b} \in R(A)$. To sum up, there exists a solution to the linear system (1) if and only if $\boldsymbol{b} \in R(A) = \operatorname{span}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n)$.

The uniqueness of a solution is studied via the homogeneous problem: Find $\boldsymbol{y} \in \mathbb{R}^n$ such that

$$A\boldsymbol{y} = 0 \quad \Longleftrightarrow \quad \boldsymbol{y} \in N(A).$$

If the homogeneous equation only has the trivial solution $\boldsymbol{y} = 0$, a solution to the linear system (1) is unique (if one exists).

Lemma 1.1. *Assume that (1) has a solution. The solution is unique if and only if $N(A) = \{0\}$ is the trivial subspace.*

Proof. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a solution of (1) and assume first that $N(A)$ is nontrivial, i.e., there exists $0 \neq \boldsymbol{z} \in N(A)$. Then,

$$A(\boldsymbol{x} + \boldsymbol{z}) = A\boldsymbol{x} + A\boldsymbol{z} = A\boldsymbol{x} = \boldsymbol{b},$$

meaning that we have constructed a second, distinct solution $\boldsymbol{x} + \boldsymbol{z} \neq \boldsymbol{x}$. This proves the 'only if' part of the claim.

In order to prove the 'if part', assume that $N(A) = \{0\}$ and let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ both be solutions to (1), i.e.,

$$A\boldsymbol{x}_1 = \boldsymbol{b} \quad \text{and} \quad A\boldsymbol{x}_2 = \boldsymbol{b}.$$

Subtracting the two equations gives

$$A(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0.$$

Hence, $\boldsymbol{x}_1 - \boldsymbol{x}_2 \in N(A)$, which means by assumption that $\boldsymbol{x}_1 - \boldsymbol{x}_2 = 0$, and thus all solutions to (1) must be the same. □

To summarize the above observations, the problem (1) has a unique solution if and only if

$$\boldsymbol{b} \in R(A) \quad \text{and} \quad N(A) = \{0\}. \tag{6}$$

Moreover, the proof of Lemma 1.1 indicates that if $\boldsymbol{x} \in \mathbb{R}^n$ is a solution of (1), then $\boldsymbol{x} + \boldsymbol{z}$ is also a solution for any $\boldsymbol{z} \in N(A)$, which characterizes the possible nonuniqueness related to (1).

It turns out that the combined condition (6) *can be* true for all $\boldsymbol{b} \in \mathbb{R}^m$ only if $n = m$, i.e., only if $A$ is a square matrix. This claim can be proved by relating the *dimensions* of $N(A)$ and $R(A)$ to each other and to $n$. To this end, we need the concept of *basis.*

Definition 1.3. *Let $E \subset \mathbb{R}^n$ be a subspace. A* basis *of $E$ is a set of linearly independent vectors $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_k\} \subset \mathbb{R}^n$ such that*

$$E = \operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k).$$

With the notation of the above definition, for each $\boldsymbol{x} \in E$ there exists a *coordinate vector* $\boldsymbol{\alpha} \in \mathbb{R}^k$ such that

$$\boldsymbol{x} = \sum_{i=1}^{k} \boldsymbol{q}_i \alpha_i \quad \Longleftrightarrow \quad \boldsymbol{x} = Q\boldsymbol{\alpha}, \tag{7}$$

where $Q = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_k] \in \mathbb{R}^{n \times k}$ is a matrix with the basis vectors as its columns. As the basis vectors are by definition linearly independent, such a coordinate presentation is unique. Indeed, if (7) holds for two coordinate vectors $\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}} \in \mathbb{R}^k$, then subtracting the corresponding representations yields

$$Q(\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}) = 0 \quad \Longrightarrow \quad \boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}} = 0$$

by virtue of (4).

Take note that there exists infinitely many different bases for any nontrivial subspace $\{0\} \neq E \subset \mathbb{R}^n$. However, it can be proven that each basis of a subspace $E$ has the same number of basis vectors (the proof is omitted).

Definition 1.4. *Let $E \subset \mathbb{R}^n$ be a subspace and $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_k\} \subset \mathbb{R}^n$ its basis. Then the* dimension *of $E$ is defined to be*

$$\dim(E) = k = \#\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_k\}.$$

*For a matrix $A \in \mathbb{R}^{m \times n}$, $\dim(R(A))$ is called the* rank *of $A$. (In particular, the dimension of $\mathbb{R}^n$ itself is $n$ since, e.g., the Cartesian basis vectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ form its basis. No other subspace of $\mathbb{R}^n$ has the maximal dimension $n$.)*

In the following example two different bases are introduced for the same subspace.

Example 1.1. *Let*

$$A = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 2 & 2 \end{bmatrix}.$$

*The column vectors of $A$ are clearly linearly independent and so they form a basis for $R(A) \subset \mathbb{R}^3$; see (5). Let*

$$\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \qquad \boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

*be another pair of vectors in $\mathbb{R}^3$ and denote $W = [\boldsymbol{w}_1, \boldsymbol{w}_2] \in \mathbb{R}^{3 \times 2}$.*

*To study if $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ also form a basis for $R(A)$, one has to check two conditions:*

- *Are the vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ linearly independent? (This is obviously true in the considered setting.)*
- *Can any $\boldsymbol{y} \in R(A)$ be represented as a linear combination of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$? (This is not quite obvious.)*

*The latter question can be answered by first introducing a generic element of $R(A)$, i.e. $\boldsymbol{y} = A\boldsymbol{z}$ for some $\boldsymbol{z} \in \mathbb{R}^2$, and then trying to express it as a linear combination of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. In other words, one must check if for any $\boldsymbol{z} \in \mathbb{R}^2$, there exists $\boldsymbol{x} \in \mathbb{R}^2$ such that*

$$W\boldsymbol{x} = x_1\boldsymbol{w}_1 + x_2\boldsymbol{w}_2 = A\boldsymbol{z},$$

*that is,*

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 0 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 2 & 2 \end{bmatrix} \boldsymbol{z}.$$

*Gaussian elimination gives*

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} \boldsymbol{z},$$

*and thus* $x_1 = z_1 + z_2$ *and* $x_2 = -z_2$ *provides the needed solution. In consequence,* $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ *is a basis of* $R(A)$*. (Note that we did not need to separately check whether* $\boldsymbol{w}_1$ *and* $\boldsymbol{w}_2$ *belong to* $R(A)$ *because we knew to begin with that the dimension of* $R(A)$ *is two.)*

We start the actual discussion on the relationship between the dimensions of $R(A)$ and $N(A)$ with another example in $\mathbb{R}^3$.

Example 1.2. *Let*

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ -1 & 1 & 0 \end{bmatrix}.$$

*The nullspace of* $A$ *can be computed by solving the equation*

$$A\boldsymbol{x} = 0. \tag{8}$$

*Using Gaussian elimination, it is straightforward to deduce that the solutions to* (8) *are exactly the scalar multiples of*

$$\boldsymbol{x} = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \qquad \text{i.e.,} \qquad N(A) = \big\{\boldsymbol{x} \in \mathbb{R}^3 \mid \boldsymbol{x} = s[-1,-1,1]^T \textit{ for some } s \in \mathbb{R}\big\}.$$

*Hence, the dimension of* $N(A)$ *is one.*

*Recall that*

$$R(A) = \{\boldsymbol{y} \in \mathbb{R}^3 \mid \boldsymbol{y} = A\boldsymbol{x} \textit{ for some } \boldsymbol{x} \in \mathbb{R}^3\},$$

*which is the linear span of the column vectors of* $A$*. To deduce the dimension of* $R(A)$*, we introduce a special basis for* $\mathbb{R}^3$*, namely*

$$\boldsymbol{v}_1 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \qquad \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

*The first vector* $\boldsymbol{v}_1$ *is a basis for* $N(A)$*, and the main idea is to choose the other two so that the three vectors are linearly independent and thus they together form a basis for* $\mathbb{R}^3$*. Denote* $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$*. Since the columns of* $V \in \mathbb{R}^{3\times 3}$ *form a basis for* $\mathbb{R}^3$*, any vector* $\boldsymbol{x} \in \mathbb{R}^3$ *can be given as their linear combination, i.e., as* $\boldsymbol{x} = V\boldsymbol{\alpha}$ *for some* $\boldsymbol{\alpha} \in \mathbb{R}^3$*. Hence, the range of* $A$ *can be alternatively expressed as*

$$R(A) = \{\boldsymbol{y} \in \mathbb{R}^3 \mid \boldsymbol{y} = A(V\boldsymbol{\alpha}) \textit{ for some } \boldsymbol{\alpha} \in \mathbb{R}^3\}.$$

*By construction* $\boldsymbol{v}_1 \in N(A)$*, and so*

$$A(V\boldsymbol{\alpha}) = (AV)\boldsymbol{\alpha} = [A\boldsymbol{v}_1, A\boldsymbol{v}_2, A\boldsymbol{v}_3]\boldsymbol{\alpha} = [0, A\boldsymbol{v}_2, A\boldsymbol{v}_3]\boldsymbol{\alpha} = \alpha_2 A\boldsymbol{v}_2 + \alpha_3 A\boldsymbol{v}_3.$$

*Hence,* $R(A)$ *is the linear span of the vectors* $A\boldsymbol{v}_2$ *and* $A\boldsymbol{v}_3$*, which are linearly independent because they are by construction the (linearly independent) first and second columns of* $A$*. In particular, the dimension of* $R(A)$ *is two.*

In Example 1.2,

$$\dim(N(A)) + \dim(R(A)) = 3,$$

which is the dimension of the considered space $\mathbb{R}^3$. The basic idea of Example 1.2 is also valid for a general $A \in \mathbb{R}^{m\times n}$, although one needs to use more general arguments to demonstrate that the candidates for the basis vectors of $R(A)$ are linearly independent.

Theorem 1.1. *For any* $A \in \mathbb{R}^{m\times n}$*,*

$$\dim(R(A)) + \dim(N(A)) = n. \tag{9}$$

Proof. Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ be a basis for $N(A)$; in particular, it is assumed that the dimension of $N(A)$ is $k$. We introduce auxiliary vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-k}$ such that

$$\{\boldsymbol{v}_1, \ldots \boldsymbol{v}_k, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n-k}\}$$

is a basis for the whole $\mathbb{R}^n$.[1] Define $V \in \mathbb{R}^{n\times k}$ and $W \in \mathbb{R}^{n\times(n-k)}$ via

$$V = \big[\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\big] \quad \text{and} \quad W = \big[\boldsymbol{w}_1, \dots, \boldsymbol{w}_{n-k}\big]\,.$$

Because the columns of the composite matrix $[V,\, W] \in \mathbb{R}^{n\times n}$ form a basis for $\mathbb{R}^n$, any $\boldsymbol{x} \in \mathbb{R}^n$ can be written as $\boldsymbol{x} = [V,\, W]\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^n$. In particular,

$$R(A) = \big\{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y} = A[V,\, W]\boldsymbol{\alpha} \text{ for some } \boldsymbol{\alpha} \in \mathbb{R}^n\big\},$$

where

$$\begin{aligned} A[V, W]\boldsymbol{\alpha} &= \big[A\boldsymbol{v}_1, \dots, A\boldsymbol{v}_k, A\boldsymbol{w}_1, \dots, A\boldsymbol{w}_{n-k}\big]\boldsymbol{\alpha} \\ &= \big[0, \dots, 0, A\boldsymbol{w}_1, \dots, A\boldsymbol{w}_{n-k}\big]\boldsymbol{\alpha} \\ &= \sum_{i=1}^{n-k} \alpha_{i+k} A\boldsymbol{w}_i \end{aligned}$$

is a linear combination of $n-k$ vectors. Hence, $R(A)$ can be represented as a linear span of $n-k$ vectors, and so its dimension is at most $n-k$.

Completing the proof amounts to showing that the vectors $A\boldsymbol{w}_1, \dots, A\boldsymbol{w}_{n-k}$ are linearly independent, i.e., that they form a basis for $R(A)$. Let $\boldsymbol{\alpha} \in \mathbb{R}^{n-k}$ be such that

$$\sum_{i=1}^{n-k} \alpha_i(A\boldsymbol{w}_i) = \big[A\boldsymbol{w}_1, \dots, A\boldsymbol{w}_{n-k}\big]\boldsymbol{\alpha} = AW\boldsymbol{\alpha} = 0,$$

that is, $W\boldsymbol{\alpha} \in N(A)$. Since the columns of $V \in \mathbb{R}^{n\times k}$ form, by assumption, a basis for $N(A)$, there exists (a unique) $\boldsymbol{\beta} \in \mathbb{R}^k$ such that $W\boldsymbol{\alpha} = V\boldsymbol{\beta}$, or in an equivalent form,

$$[V,\;W]\begin{bmatrix}\boldsymbol{\beta} \\ -\boldsymbol{\alpha}\end{bmatrix} = 0.$$

Since the columns of $[V,\, W] \in \mathbb{R}^{n\times n}$ form a basis for $\mathbb{R}^n$ by our construction, it must hold that $\boldsymbol{\beta} = 0$ and $-\boldsymbol{\alpha} = 0$. In particular, since $\boldsymbol{\alpha}$ vanishes, $A\boldsymbol{w}_1, \dots, A\boldsymbol{w}_{n-k}$ are linearly independent. □

Let us then return to the unique solvability of (1). In order for the 'unique solvability conditions' (6) to hold for all $\boldsymbol{b} \in \mathbb{R}^m$, one must obviously have $R(A) = \mathbb{R}^m$ and $N(A) = \{0\}$. In particular, this yields

$$\dim\big(R(A)\big) + \dim\big(N(A)\big) = m + 0 = m.$$

By virtue of Theorem 1.1, this is possible only if $m = n$ and thus $A$ is a square matrix. In other words, if $m \neq n$, i.e., $A$ is not square, either (6) does not have any solution for some $\boldsymbol{b} \in \mathbb{R}^m$, or no solution for (6) is unique, or both.

Let us complete this section by considering the important special case of square matrices, i.e., $m = n$. According to the above considerations, (1) has a solution for all $\boldsymbol{b} \in \mathbb{R}^n$ if and only if $R(A) = \mathbb{R}^n$, i.e., $\operatorname{rank}(A) = \dim(R(A)) = n$. However, due to Theorem 1.1, this also guarantees that $\dim(N(A)) = 0$, i.e., $N(A) = \{0\}$, and the solution is also unique. Vice versa, the solution to (1) is unique if and only if $\dim(N(A)) = 0$, which also guarantees the existence of a solution by virtue of Theorem 1.1. To summarize, for square matrices the existence and the uniqueness of a solution to (1) are equivalent conditions.

There are also several other conditions that guarantee the unique solvability of (1) for all $\boldsymbol{b} \in \mathbb{R}^n$ when $m = n$, as indicated by the following corollary.

COROLLARY 1.1. *Let $m = n$ in (1). There exists a unique solution to (1) for any $\boldsymbol{b} \in \mathbb{R}^n$ if and only if one of the following equivalent conditions holds:*

1. *$R(A) = \mathbb{R}^n$,*
2. *$A$ has linearly independent columns (or rows),*
3. *$N(A) = \{0\}$,*
4. *$det(A) \neq 0$, or*

[1] Such $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{n-k}$ could be constructed, e.g., via a variant of the Gram–Schmidt orthogonalization process considered in Chapter 3.

(5) 0 *is not an eigenvalue of* $A$.

Proof. The conditions (1) and (3) were already covered in the above discussion. The condition (2) is equivalent to (1) due to (5). The condition (4) is equivalent to (2) by fundamental properties of the determinant (details are omitted). Finally, having 0 as an eigenvalue is equivalent to having a nontrivial nullspace, i.e., (5) and (3) are equivalent conditions (eigenvalues will be considered in detail in Chapter 2). □

To sum up, if any of the above conditions (1)–(5) holds for $A \in \mathbb{R}^{n\times n}$, then (1) is uniquely solvable for any $\boldsymbol{b} \in \mathbb{R}^n$. This unique solution depends linearly on the right-hand side of the equation: Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ be the solutions of (1) for the right-hand sides $\boldsymbol{b}_1, \boldsymbol{b}_1 \in \mathbb{R}^n$, respectively. Then for any $\alpha, \beta \in \mathbb{R}$,

$$A(\alpha \boldsymbol{x}_1 + \beta \boldsymbol{x}_2) = \alpha A\boldsymbol{x}_1 + \beta A \boldsymbol{x}_2 = \alpha \boldsymbol{b}_1 + \beta \boldsymbol{b}_2,$$

and thus $\alpha \boldsymbol{x}_1 + \beta \boldsymbol{x}_2$ is the solution of (1) corresponding to $\boldsymbol{b} = \alpha \boldsymbol{b}_1 + \beta \boldsymbol{b}_2$. Because any linear map can be represented as a matrix (in a given basis), it follows that the solution to (1) can be given as

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}, \tag{10}$$

for some $A^{-1} \in \mathbb{R}^{n\times n}$, if any of the (equivalent) conditions in Corollary 1.1 holds. The matrix $A^{-1}$ is called the *inverse* of $A$ and if such exists, $A$ is called *invertible*. In particular,

$$A^{-1}A = \left[A^{-1}\boldsymbol{a}_1, \dots, A^{-1}\boldsymbol{a}_n\right] = [\boldsymbol{e}_1, \dots, \boldsymbol{e}_n] = I$$

since the solution to $A\boldsymbol{x} = \boldsymbol{a}_j$, $j = 1, \dots, n$, is obviously the $j$th Cartesian basis vector $\boldsymbol{e}_j \in \mathbb{R}^n$. One can argue in the same way that also $AA^{-1} = I$, because $A$ is obviously the inverse of $A^{-1}$ (simply consider the symmetry between (1) and (10)).

## 2. Norms and inner products

To study the stability of the solution to (1) with respect to perturbations, inaccuracies or uncertainties in $\boldsymbol{b}$ and $A$, we need tools for measuring the 'size' of a vector or a matrix. This section will also introduce the concept or orthogonality via inner products.

**2.1. Vector norm.** The size of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is measured by a (vector) norm. The concept of norm is not unique, but there are many useful norms as we will see in what follows.

Definition 2.1. *A function* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ *is a* norm *if it satisfies*

(i) $\|\boldsymbol{x}\| \geq 0$ *for all* $\boldsymbol{x} \in \mathbb{R}^n$ *and* $\|\boldsymbol{x}\| = 0$ *if and only if* $\boldsymbol{x} = 0$,

(ii) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

(iii) $\|\alpha \boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|$ *for all* $\boldsymbol{x} \in \mathbb{R}^n$ *and* $\alpha \in \mathbb{R}$.

Definition 2.1 formalizes what one expects from a reasonable measure of length. As an example, consider the space $\mathbb{R}^2$ and associate to each $\boldsymbol{x} \in \mathbb{R}^2$ a geometric vector $x_1\boldsymbol{i} + x_2\boldsymbol{j}$ as in Figure 1. A natural way to measure the length of such a vector $\boldsymbol{x}$ is via the Pythagorean theorem, i.e., by introducing the *Euclidean norm* $\|\cdot\|_2 : \mathbb{R}^2 \to \mathbb{R}$ defined via

$$\|\boldsymbol{x}\|_2 := \left(x_1^2 + x_2^2\right)^{1/2}.$$

This function obviously has the following properties:

(i) **Positivity:** For all $\boldsymbol{x} \in \mathbb{R}^2$,

$$\|\boldsymbol{x}\|_2 \geq 0,$$

with the equality holding only if and only if $\boldsymbol{x} = 0$.

(ii) **Scaling:** The distance measured by $\|\cdot\|_2$ is scaling invariant, that is,

$$\|\alpha\boldsymbol{x}\|_2 = |\alpha|\|\boldsymbol{x}\|_2$$

for all $\alpha \in \mathbb{R}$.

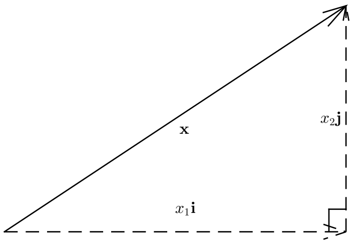

FIGURE 1. The geometric idea behind measuring the length of a vector in $\mathbb{R}^2$.

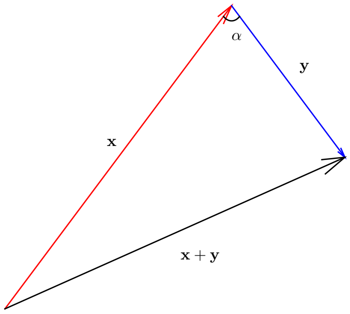

FIGURE 2. The geometric idea behind the triangle inequality.

(iii) **Triangle inequality:** The three vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y} \in \mathbb{R}^2$ form a triangle in the plane; see Figure 2. For every triangle, the sum of the lengths of two sides is larger than the length of the third one, that is,

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 \leq \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2.$$

Hence, $\|\cdot\|_2 : \mathbb{R}^2 \to \mathbb{R}$ is a norm in accordance with Definition 2.1. In fact, the geometric properties of the Euclidean norm are the motivation for the general norm of Definition 2.1. Notice also that the Euclidean norm, a.k.a. the 2-norm, generalizes to an arbitrary spatial dimension by defining $\|\cdot\|_2 : \mathbb{R}^n \to \mathbb{R}$ through

$$\|\boldsymbol{x}\|_2 = \Big(\sum_{i=1}^n |x_i|^2\Big)^{1/2}. \tag{11}$$

The norms used in linear algebra can be divided into generic norms and problem specific norms. Problem specific norms are used when vectors have some special interpretation; they can, e.g., represent coefficients in linear combinations of some elementary functions (cf., e.g., polynomials represented as linear combinations of monomials or a truncated Fourier series). Most theory in linear algebra is given in terms of generic norms that operate on generic vectors that have no special interpretation as such.

The most commonly used norm in $\mathbb{R}^n$ is the Euclidean norm introduced in (11). Other regularly used norms include the family of $p$-norms (most often the case $p = 1$)

$$\|\boldsymbol{x}\|_p = \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p}, \qquad 1 \leq p < \infty. \tag{12}$$

and the $\infty$-norm, $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$. The latter can be obtained as a limit of the $p$-norms when $p \to \infty$. The 1 and $\infty$-norms are often employed because the related *operator norms* are easy to compute, as we will learn in the following. Proving that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ (or more generally $\|\cdot\|_p$) really are norms is left as an exercise.

EXAMPLE 2.1. (a problem specific norm) *Consider a first order polynomial $p(x) = \alpha_0 + \alpha_1 x$ with real coefficients. The size of such a polynomial can be measured by the square integral*

$$\int_0^1 p^2(x)\,dx = \int_0^1 (\alpha_0^2 + 2\alpha_0\alpha_1 x + \alpha_1^2 x^2)\,dx = \alpha_0^2 + \alpha_0\alpha_1 + \frac{1}{3}\alpha_1^2.$$

*On the other hand, any vector in $\mathbb{R}^2$ can be identified with a first order polynomial via*

$$\begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix} \longleftrightarrow \alpha_0 + \alpha_1 x.$$

*Measuring the length of a vector defining the coefficients of a first order polynomial in, say, the 2-norm does not have any immediate interpretation. Instead it is more natural to measure the size of the coefficient vector using the function*

$$\left\| \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} \right\| := \left( \alpha_0^2 + \alpha_0 \alpha_1 + \frac{1}{3} \alpha_1^2 \right)^{1/2}, \tag{13}$$

*which is the square root of the square integral of the first order polynomial $\alpha_0 + \alpha_1 x$ introduced above. We will later prove that (13) really defines a norm, that is, it satisfies the conditions of Definition 2.1.*

**2.2. Inner product.** To begin with, let $n = 2$ and consider two nonzero geometric vectors $x_1 \boldsymbol{i} + x_2 \boldsymbol{j}$ and $y_1 \boldsymbol{i} + y_2 \boldsymbol{j}$. It follows from the *cosine theorem* that the angle between the two vectors, say $\theta$, satisfies

$$\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \cos\theta = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2(x_1 y_1 + x_2 y_2),$$

or, in other words,

$$\cos\theta = \frac{x_1 y_1 + x_2 y_2}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}.$$

Motivated by this formula, we define the *dot product* as

$$\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2.$$

The dot product introduces the concept of orthogonality in the space $\mathbb{R}^2$, that is, when $\boldsymbol{x} \cdot \boldsymbol{y} = 0$, then the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ form a right angle (or one/both of them vanish). We generalize the dot product to an arbitrary spatial dimension by defining

$$\boldsymbol{x} \cdot \boldsymbol{y} := \sum_{i=1}^{n} x_i y_i = \boldsymbol{x}^T \boldsymbol{y}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. In particular, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are said to be orthogonal (in the sense of the dot product or the Euclidean inner product) if $\boldsymbol{x} \cdot \boldsymbol{y} = 0$, which actually also matches the geometric intuition for $n = 3$.

The basic properties of the dot product motivate the definition of a general *inner product*; it is easy to check that the dot product is an inner product.

Definition 2.2. *A mapping $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an* inner product *if the following properties hold:*

(i) *Positive definiteness: $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = 0$.*

(ii) *Symmetry: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.*

(iii) *Bilinearity: $\langle \alpha\boldsymbol{x} + \beta\boldsymbol{y}, \boldsymbol{z} \rangle = \alpha\langle \boldsymbol{x}, \boldsymbol{z} \rangle + \beta\langle \boldsymbol{y}, \boldsymbol{z} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$.*

The conditions (ii) and (iii) together imply that also $\langle \boldsymbol{z}, \alpha\boldsymbol{x} + \beta\boldsymbol{y} \rangle = \alpha\langle \boldsymbol{z}, \boldsymbol{x} \rangle + \beta\langle \boldsymbol{z}, \boldsymbol{y} \rangle$. Take note that different inner products define different concepts of orthogonality in $\mathbb{R}^n$ via the relation $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Before moving on to vector norms that are induced by inner products, let us relate orthogonality and linear independence, starting with a definition.

Definition 2.3. *A set of vectors $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_k\} \subset \mathbb{R}^n$ is called* orthogonal *with respect to an inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ if*

$$\langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle = 0$$

*for all $i \neq j$.*

It turns out that an orthogonal set of vectors is always linearly independent, unless one of the vectors is the zero vector.

Lemma 2.1. *If the nonzero vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k \in \mathbb{R}^n$ compose a set of orthogonal vectors (with respect to any inner product), then they are linearly independent.*

PROOF. Let $\boldsymbol{\alpha} \in \mathbb{R}^k$ be such that

$$\sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i = 0. \tag{14}$$

Taking the inner product of the left-hand side of this identity with $\boldsymbol{q}_j$, $j = 1, \ldots, k$, and using (iii) of Definition 2.2 yields

$$\left\langle \sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i, \boldsymbol{q}_j \right\rangle = \sum_{i=1}^{k} \alpha_i \langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle = \alpha_j \langle \boldsymbol{q}_j, \boldsymbol{q}_j \rangle,$$

where the second step is just the orthogonality assumption. On the other hand, taking the inner product of the right-hand side of (14) with $\boldsymbol{q}_j$ results in

$$\langle 0, \boldsymbol{q}_j \rangle \; = \; 0 \, \langle 0, \boldsymbol{q}_j \rangle = 0$$

due to (iii) of Definition 2.2. Altogether we have deduced that

$$\alpha_j \langle \boldsymbol{q}_j, \boldsymbol{q}_j \rangle = \left\langle \sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i, \boldsymbol{q}_j \right\rangle = \langle 0, \boldsymbol{q}_j \rangle = 0. \tag{15}$$

for any $j = 1, \ldots, k$. Since $\langle \boldsymbol{q}_j, \boldsymbol{q}_j \rangle > 0$ by (i) of Definition 2.2 and our assumption on the orthogonal vectors, it must hold that $\alpha_j = 0$ for all $j = 1, \ldots, k$. Thus the vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$ are linearly independent. □

A byproduct of Lemma 2.1 is that any set of $n$ nonzero orthogonal vectors $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\} \subset \mathbb{R}^n$ forms a basis for the whole of $\mathbb{R}^n$. Indeed, it can be straightforwardly proven that any set of $n$ linearly independent vectors must be a basis for an $n$-dimensional (sub)space. Hence, for any $\boldsymbol{x} \in \mathbb{R}^n$, there exists a representation

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{q}_i \tag{16}$$

with some $\boldsymbol{\alpha} \in \mathbb{R}^n$. By taking the inner products of this identity with the orthogonal basis vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ one at a time, it can be deduced as in the proof of Lemma 2.1 that (cf. (15))

$$\alpha_j = \frac{\langle \boldsymbol{x}, \boldsymbol{q}_j \rangle}{\langle \boldsymbol{q}_j, \boldsymbol{q}_j \rangle}, \qquad j = 1, \ldots, n, \tag{17}$$

which provides a means to numerically compute the coefficients in the expansion (16). The (projection) formula (17) for the coefficients of (16) becomes even simpler if $\langle \boldsymbol{q}_j, \boldsymbol{q}_j \rangle = 1$ for all $j = 1, \ldots, n$. Such a basis is called *orthonormal*, that is, the basis vectors are orthogonal and *normalized*, i.e., of unit length. This statement makes more sense after the definition of norms induced by inner products. The (numerical) construction of orthonormal bases is considered in Chapter 3.

To motivate the connection between inner products and norms, observe that the dot product obviously has a special connection to the Euclidean norm:

$$\|\boldsymbol{x}\|_2 = (\boldsymbol{x} \cdot \boldsymbol{x})^{1/2}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$. This observation holds more generally: any inner product induces a norm.

LEMMA 2.2. *Let $\langle \cdot, \cdot \rangle$ be an inner product in $\mathbb{R}^n$. Then the function*

$$\|\boldsymbol{x}\| := \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2}, \qquad \boldsymbol{x} \in \mathbb{R}^n, \tag{18}$$

*is a norm. In addition, such a norm satisfies the Cauchy–Schwarz inequality*

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\| \tag{19}$$

*for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.*

Proof. We need to check that $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfies the conditions (i-iii) of Definition 2.1 and that (19) holds.

By the positive definiteness of an inner product,

$$\|\boldsymbol{x}\| := \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2} \geq 0,$$

where the equality holds only if $\boldsymbol{x} = 0$. This validates (i) of Definition 2.1. Furthermore, due to properties (ii) and (iii) of Definition 2.2,

$$\|\alpha \boldsymbol{x}\| = \langle \alpha \boldsymbol{x}, \alpha \boldsymbol{x} \rangle^{1/2} = \big(\alpha^2 \langle \boldsymbol{x}, \boldsymbol{x} \rangle\big)^{1/2} = |\alpha| \|\boldsymbol{x}\|,$$

which demonstrates that (iii) of Definition 2.1 is also satisfied.

Let us next prove the Cauchy–Schwarz inequality (19), which obviously holds if $\boldsymbol{x} = 0$ or $\boldsymbol{y} = 0$ since then both sides of (19) vanish due to (i) and (iii) of Definition 2.2 (note that, e.g., $\langle 0, \boldsymbol{y} \rangle = 0 \langle 0, \boldsymbol{y} \rangle = 0$). Let thus $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be nonzero vectors and decompose $\boldsymbol{x}$ as (cf. (17))

$$\boldsymbol{x} = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y} + \left( \boldsymbol{x} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y} \right).$$

Using (ii) and (iii) of Definition 2.2, we thus obtain that

$$\begin{aligned}
\|\boldsymbol{x}\|^2 &= \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2}{\|\boldsymbol{y}\|^4} \|\boldsymbol{y}\|^2 + 2 \left\langle \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y}, \boldsymbol{x} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y} \right\rangle + \left\| \boldsymbol{x} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y} \right\|^2 \\
&= \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2}{\|\boldsymbol{y}\|^2} + 2 \left( \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2}{\|\boldsymbol{y}\|^2} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2}{\|\boldsymbol{y}\|^4} \|\boldsymbol{y}\|^2 \right) + \left\| \boldsymbol{x} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y} \right\|^2 \\
&= \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2}{\|\boldsymbol{y}\|^2} + \left\| \boldsymbol{x} - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|^2} \boldsymbol{y} \right\|^2.
\end{aligned}$$

Dropping the second (positive) term on the right-hand side and multiplying the ensuing inequality by $\|\boldsymbol{y}\|^2$ results in

$$\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 \geq \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \qquad \text{or} \qquad |\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|,$$

which is the Cauchy–Schwarz inequality.

Finally, the triangle inequality, i.e. (ii) of Definition 2.1, follows by expanding,

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + 2 \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{y}\|^2 \leq \|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2 = (\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2$$

where the inequality is a trivial consequence of (19). □

We complete this section by pointing out a one-to-one correspondence between positive definite matrices and inner products.

Definition 2.4. *A matrix $A \in \mathbb{R}^{n \times n}$ is called* positive semidefinite *if*

$$\boldsymbol{x}^T A \boldsymbol{x} = \boldsymbol{x} \cdot A \boldsymbol{x} \geq 0 \tag{20}$$

*for all $\boldsymbol{x} \in \mathbb{R}^n$. If the second equality in (20) holds only for $\boldsymbol{x} = 0$, then $A$ is called* positive definite. *(Nota bene: Sometimes symmetry of $A$ is included in the definition of positive definiteness.)*

Take note that any positive definite matrix $A \in \mathbb{R}^{n \times n}$ is invertible: If $\boldsymbol{x} \in N(A)$, i.e., $A\boldsymbol{x} = 0$, then

$$0 = \boldsymbol{x}^T A \boldsymbol{x} \geq 0,$$

where the latter equality holds if and only if $\boldsymbol{x} = 0$. Hence, $N(A) = \{0\}$ and $A$ is invertible due to the condition (3) in Corollary 1.1. However, most invertible matrices are not positive definite.

The following lemma demonstrates that any inner product in $\mathbb{R}^n$ can be characterized with the help of the Euclidean inner product and a suitable symmetric positive definite matrix. In turn, any symmetric positive definite matrix defines an inner product.

LEMMA 2.3. *For any inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, there exists a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ such that*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^T A \boldsymbol{x}. \tag{21}$$

*for all $\boldsymbol{x} \in \mathbb{R}^n$. On the other hand, the formula (21) defines an inner product for any symmetric positive definite $A \in \mathbb{R}^{n \times n}$.*

PROOF. We start with the second part of the claim. Let $A \in \mathbb{R}^{n \times n}$ be an arbitrary symmetric positive definite matrix and set

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A := \boldsymbol{y}^T A \boldsymbol{x}.$$

We need to check that $\langle \cdot, \cdot \rangle_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an inner product, that is, we need to validate the conditions (i–iii) in Definition 2.2. (i) As $A$ is positive definite,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \boldsymbol{x}^T A \boldsymbol{x} \geq 0 \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n,$$

where the equality holds if and only if $\boldsymbol{x} = 0$. (ii) The symmetry requirement for an inner product is satisfied due to the assumed symmetry of $A$ and the symmetry of the Euclidean inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \boldsymbol{y}^T A \boldsymbol{x} = \boldsymbol{y}^T A^T \boldsymbol{x} = (A\boldsymbol{y})^T \boldsymbol{x} = \boldsymbol{x}^T A \boldsymbol{y} = \langle \boldsymbol{y}, \boldsymbol{x} \rangle_A.$$

(iii) The bilinearity follows from the (bi)linearity of matrix product (homework).

To complete the proof, let $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be an arbitrary inner product. Our aim is to construct a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ such that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^T A \boldsymbol{x}.$$

Expanding $\boldsymbol{x} = \sum_{i=1}^n x_i \boldsymbol{e}_i$ and $\boldsymbol{y} = \sum_{j=1}^n y_j \boldsymbol{e}_j$ in the Cartesian basis, it follows that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \left\langle \sum_{i=1}^n x_i \boldsymbol{e}_i, \sum_{j=1}^n y_j \boldsymbol{e}_j \right\rangle = \sum_{i,j=1}^n x_i y_j \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle,$$

where we repeatedly employed (iii) of Definition 2.2. We now define the matrix $A \in \mathbb{R}^{n \times n}$ elementwise as $a_{ji} = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle$, $i, j = 1, \dots, n$. Clearly, $a_{ij} = a_{ji}$, i.e., $A$ is symmetric. Moreover,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i,j=1}^n x_i y_j a_{ji} = \sum_{j=1}^n y_j \sum_{i=1}^n a_{ji} x_i = \boldsymbol{y}^T A \boldsymbol{x}$$

by the definition of matrix-vector product. Finally, the constructed $A$ is positive definite since

$$\boldsymbol{x}^T A \boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle > 0 \qquad \text{for all } 0 \neq \boldsymbol{x} \in \mathbb{R}^n$$

because $\langle \cdot, \cdot \rangle$ was assumed to be an inner product; see (i) of Definition 2.2. □

**2.3. Matrix norm.** In addition to measuring the length of a vector $\boldsymbol{x}$, we also need to be able to measure the 'size' of a given matrix $A \in \mathbb{R}^{m \times n}$. Let us begin the discussion with an example.

EXAMPLE 2.2. *Let*

$$A = \begin{bmatrix} 1 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}.$$

*The matrix $A$ associates every vector $\boldsymbol{x} \in \mathbb{R}^2$ with another one $A\boldsymbol{x} \in \mathbb{R}^2$. Such a mapping can be visualized (at least) in two ways:*

(1) *Draw some pixel image into the plane (e.g., a clown). Map the coordinates of the pixel corners as in Figure 3. Coloring the skewed pixels with the same color as their preimages results in a deformed image that describes the action of $A$. See Figure 3.*
(2) *Due to the linearity of matrix multiplication,*

$$A\boldsymbol{x} = \|\boldsymbol{x}\|_2 \, A \left( \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \right).$$

*Hence, all information in the matrix $A$ is actually contained in the set*

$$A(S) = \{ A\boldsymbol{v} \mid \|\boldsymbol{v}\|_2 = 1 \}$$

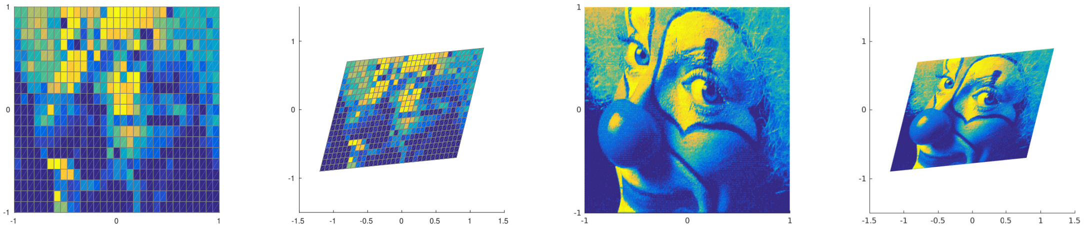

Figure 3. Two pixelwise discretizations of a clown and their images under the action of the matrix $A$ from Example 2.2.

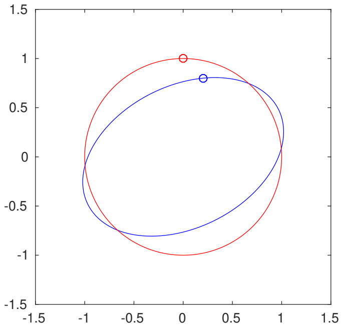

Figure 4. The image of the Euclidean unit sphere under the action of the matrix $A$ from Example 2.2. The red circle is the unit sphere and the blue ellipse its image.

*since any vector can be given as a vector of unit length times a suitable scalar. The set $A(S)$ is compared with its preimage, i.e. the unit circle $S$ (or, more generally, the unit sphere), in Figure 4.*

The second way to visualize the multiplication by a matrix suggest a way to measure its size.

Definition 2.5. *Let $\|\cdot\|$ be some vector norm. The corresponding* operator norm *for a matrix $A \in \mathbb{R}^{m\times n}$ is defined as*

$$\|A\|_{\mathrm{op}} := \max_{0\neq \boldsymbol{x}\in\mathbb{R}^n} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|.$$

The definition of an operator norm depends on the considered vector norm. In other words, different vector norms induce different operator norms. The operator norm measures the maximal stretching of the unit sphere under the action of a matrix $A$; note that even the unit sphere itself, i.e.,

$$S := \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$$

depends on the investigated vector norm. In Example 2.2, where the Euclidean norm was employed, the operator norm is simply the semi-major axis of the blue ellipse in Figure 4, i.e., $\|A\|_{\mathrm{op}} \approx 1.08$.

In linear algebra, it is customary to drop the subscript from $\|\cdot\|_{\mathrm{op}}$ and denote both a vector norm and the corresponding operator norm in the same way. We follow this convention during these lectures. In particular, the operator norms induced by 1, $\infty$ and 2-norms are denoted by $\|\cdot\|_1$, $\|\cdot\|_\infty$ and $\|\cdot\|_2$, respectively, or simply by $\|\cdot\|$ if the considered vector norm is clear from the context.

All operator norms share two useful properties, as indicated by the following lemma.

LEMMA 2.4. *Any operator norm satisfies,*

$$\|AB\| \leq \|A\|\|B\|\,, \qquad A \in \mathbb{R}^{l\times m},\ B \in \mathbb{R}^{m\times n} \tag{22}$$

*and*

$$\|A\boldsymbol{x}\| \leq \|A\|\|\boldsymbol{x}\|\,, \qquad A \in \mathbb{R}^{m\times n},\ \boldsymbol{x} \in \mathbb{R}^{n}. \tag{23}$$

PROOF. Homework. □

Computing the value of an operator norm requires one to find the maximum of $\|A\boldsymbol{x}\|\|\boldsymbol{x}\|^{-1}$ over $\mathbb{R}^n$ or equivalently that of $\|A\boldsymbol{x}\|$ over the unit sphere corresponding to the examined vector norm. For the operator norms induced by the vector norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$, this maximum can be found explicitly. This is one of the reasons why these three norms are used as generic norms in linear algebra.

The 2-norm of a matrix can be computed using the following lemma that utilizes eigenvalues of a symmetric matrix. Recall that an eigenvalue of a matrix $A \in \mathbb{R}^{n\times n}$ is such $\lambda \in \mathbb{C}$ that

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{24}$$

for some $0 \neq \boldsymbol{x} \in \mathbb{C}^n$ that is called an eigenvector corresponding to $\lambda$.[2] Eigenvalues and eigenvectors will be studied in detail in Chapter 2. Here, we only need the following result that should be familiar from the basic matrix algebra course: Any symmetric matrix $B \in \mathbb{R}^{n\times n}$ has $n$ orthonormal *real* eigenvectors (with respect to the dot product/Euclidean inner product), that is,

$$B\boldsymbol{q}_j = \lambda_j\boldsymbol{q}_j, \qquad j = 1,\dots,n,$$

with $\|\boldsymbol{q}_j\|_2 = 1$ for all $j = 1,\dots,n$ and $\boldsymbol{q}_j \cdot \boldsymbol{q}_k = \boldsymbol{q}_j^T\boldsymbol{q}_k = 0$ for all $j \neq k$. Moreover, $B$ can be decomposed as

$$B = Q\Lambda Q^T \tag{25}$$

where $Q = [\boldsymbol{q}_1,\dots,\boldsymbol{q}_n] \in \mathbb{R}^{n\times n}$ has the orthonormal eigenvectors as its columns and $\Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_n) \in \mathbb{R}^{n\times n}$ is a diagonal matrix composed of the *real* eigenvalues (some of which may be same). In particular, $Q^TQ = I$ because the columns of $Q$ are orthonormal, i.e., $Q$ is an *orthogonal matrix*.

LEMMA 2.5. *For any $A \in \mathbb{R}^{m\times n}$,*

$$\|A\|_2 = \big(\lambda_{\max}(A^TA)\big)^{1/2},$$

*where $\lambda_{\max}(A^TA)$ is the largest eigenvalue of the symmetric matrix $A^TA$.*

PROOF. By the monotonicity of the second power,

$$\|A\|_2^2 := \max_{0\neq\boldsymbol{x}\in\mathbb{R}^n} \frac{\|A\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} = \max_{0\neq\boldsymbol{x}\in\mathbb{R}^n} \frac{\boldsymbol{x}^TA^TA\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2} = \max_{0\neq\boldsymbol{x}\in\mathbb{R}^n} \frac{\boldsymbol{x}^TQ\Lambda Q^T\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2},$$

where $Q\Lambda Q^T$ is an eigendecomposition of the symmetric matrix $A^TA \in \mathbb{R}^{n\times n}$, with an orthogonal $Q \in \mathbb{R}^{n\times n}$ and a diagonal $\Lambda \in \mathbb{R}^{n\times n}$ as in (25). Defining $\boldsymbol{y} = Q^T\boldsymbol{x}$ and noting that $\boldsymbol{x} = Q\boldsymbol{y}$ runs through the whole of $\mathbb{R}^n$ when $\boldsymbol{y}$ does so, we get

$$\|A\|_2^2 = \max_{0\neq\boldsymbol{y}\in\mathbb{R}^n} \frac{\boldsymbol{y}^T\Lambda\boldsymbol{y}}{\|Q\boldsymbol{y}\|_2^2} = \max_{0\neq\boldsymbol{y}\in\mathbb{R}^n} \frac{\sum_{i=1}^n \lambda_i\boldsymbol{y}_i^2}{\boldsymbol{y}^TQ^TQ\boldsymbol{y}} = \max_{0\neq\boldsymbol{y}\in\mathbb{R}^n} \frac{\sum_{i=1}^n \lambda_i\boldsymbol{y}_i^2}{\|\boldsymbol{y}\|_2^2}, \tag{26}$$

where $\lambda_i$, $i = 1,\dots,n$, are the real eigenvalues of $A^TA \in \mathbb{R}^{n\times n}$ and we also employed the identity $Q^TQ = I$.

Notice that all $\lambda_i$, $i = 1,\dots,n$, are nonnegative (because $A^TA$ is positive semidefinite):

$$0 \leq \|A\boldsymbol{q}_i\|_2^2 = \boldsymbol{q}_i^TA^TA\boldsymbol{q}_i = \lambda_i\boldsymbol{q}_i^T\boldsymbol{q}_i = \lambda_i,$$

where we used the orthonormality of the (eigen)columns of $Q$ (cf. (25)). A straightforward estimate thus gives

$$\|A\|_2^2 \leq \max_{0\neq\boldsymbol{y}\in\mathbb{R}^n} \frac{\lambda_{\max}(A^TA)\sum_{i=1}^n \boldsymbol{y}_i^2}{\|\boldsymbol{y}\|_2^2} = \lambda_{\max}(A^TA) \max_{0\neq\boldsymbol{y}\in\mathbb{R}^n} \frac{\|\boldsymbol{y}\|_2^2}{\|\boldsymbol{y}\|_2^2} = \lambda_{\max}(A^TA).$$

[2] In particular, a real matrix may have complex eigenvalues and eigenvectors.

Letting $1 \leq k \leq n$ be such that $\lambda_k = \lambda_{\max}(A^T A)$ and choosing $\boldsymbol{y} = \boldsymbol{e}_k$ in (26) to be the $k$th Cartesian basis vector, we finally get

$$\|A\|_2^2 \geq \lambda_{\max}(A^T A),$$

which completes the proof as we have altogether established that $\lambda_{\max}(A^T A) \leq \|A\|_2^2 \leq \lambda_{\max}(A^T A)$. □

Corollary 2.1. *Lemma 2.5 remains valid if $\lambda_{\max}(A^T A)$ is replaced by $\lambda_{\max}(AA^T)$.*

Proof. It can be shown that the $\min\{n, m\}$ largest eigenvalues of the symmetric matrices $A^T A \in \mathbb{R}^{n \times n}$ and $AA^T \in \mathbb{R}^{m \times m}$ are the same[3] (when counted according to their algebraic multiplicity); a proof will be presented in connection to singular values in Chapter 3. The assertion is an immediate consequence of the aforementioned fact. □

It can also be shown that the operator norms of $A \in \mathbb{R}^{m \times n}$ corresponding to the 1 and $\infty$ vector norms can be computed simply as

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}| \quad \text{and} \quad \|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|,$$

that is, as the maximal absolute column and row sums of $A$, respectively.

The operator norms form a subclass of all *matrix norms*. In other words, every matrix norm is not an operator norm, but all operator norms are matrix norms as indicated by the lemma succeeding the following definition.

Definition 2.6. *A function $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a* matrix norm *if*

(i) $\|A\| \geq 0$ *and* $\|A\| = 0$ *if and only if* $A = 0$ *is the zero matrix,*

(ii) $\|A + B\| \leq \|A\| + \|B\|$,

(iii) $\|\alpha A\| = |\alpha| \|A\|$

*for all $A, B \in \mathbb{R}^{m \times n}$ and $\alpha \in \mathbb{R}$.*

Lemma 2.6. *Let $\|\cdot\|$ denote a vector norm and also the corresponding operator norm. Then $\|\cdot\|$ is also a matrix norm in accordance with Definition 2.6.*

Proof. Let us verify the conditions (i–iii) of Definition 2.6 one at a time.

(i) Obviously,

$$\|A\| := \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \geq 0.$$

Moreover, if any element of $A \in \mathbb{R}^{m \times n}$ is nonzero, it is easy to find $\boldsymbol{x} \in \mathbb{R}^n$ such that $A\boldsymbol{x}$ is also nonzero. Hence, $\|A\| > 0$ if $A$ is not the zero matrix.

(ii) Due to the definition of the operator norm and the linearity of matrix multiplication,

$$\|A + B\| = \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{\|A\boldsymbol{x} + B\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \left( \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} + \frac{\|B\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \right) \leq \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} + \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{\|B\boldsymbol{x}\|}{\|\boldsymbol{x}\|},$$

where the second step is the triangle inequality for the vector norm and the last step corresponds to a simple property of the max function: “The maximum of a sum is less than the sum of the maxima of the summands”.

(iii) Finally, it holds that

$$\|\alpha A\| = \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{\|\alpha A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{|\alpha| \|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = |\alpha| \max_{0 \neq \boldsymbol{x} \in \mathbb{R}^n} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = |\alpha| \|A\|$$

by the scaling property of the vector norm. □

[3] This is not very difficult to believe: $AA^T\boldsymbol{v} = \lambda\boldsymbol{v} \Rightarrow A^T A(A^T\boldsymbol{v}) = \lambda(A^T\boldsymbol{v})$ and $A^T A\boldsymbol{v} = \lambda\boldsymbol{v} \Rightarrow AA^T(A\boldsymbol{v}) = \lambda(A\boldsymbol{v})$.

As indicated by Definition 2.6, a matrix norm can simply be defined as a function from the space of matrices $\mathbb{R}^{m\times n}$ to nonnegative real numbers — as long as it satisfies the conditions of Definition 2.6. Examples of such directly defined matrix norms are the Frobenius norm,

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

and the max-norm,

$$\|A\|_{\max} = \max_{1\le i\le m} \max_{1\le j\le n} |a_{ij}|.$$

It is straightforward to verify that both of these really are matrix norms in accordance with Definition 2.6. (The proofs are analogous to showing that the 2 and $\infty$ vector norms are norms.)

## 3. Stability of the solution

In this section, we consider the equation (1) in the case that $m = n$ and the matrix $A$ is invertible, i.e., $A \in \mathbb{R}^{n\times n}$ has an inverse $A^{-1} \in \mathbb{R}^{n\times n}$. Recall that the existence of an inverse $A^{-1}$ is equivalent to the unique solvability of (1) for all $\boldsymbol{b} \in \mathbb{R}^n$, and Corollary 1.1 lists conditions that guarantee the latter. In this section, we employ the norms introduced in the previous section to study how perturbations of $A$ and $\boldsymbol{b}$ in (1) affect the corresponding solution.

Let $\boldsymbol{x}$ be the solution to the linear system (1), $\delta A \in \mathbb{R}^{n\times n}$ and $\delta\boldsymbol{b} \in \mathbb{R}^n$. The perturbations $\delta A$ and $\delta\boldsymbol{b}$ model (additive) uncertainties and inaccuracies in $A$ and $\boldsymbol{b}$, respectively. The general intuition should thus be that $\delta A$ and $\delta\boldsymbol{b}$ are small(ish) compared to $A$ and $\boldsymbol{b}$, respectively. To be more precise, we consider the perturbed linear equation

$$(A + \delta A)\tilde{\boldsymbol{x}} = \boldsymbol{b} + \delta\boldsymbol{b}. \tag{27}$$

and study the difference between the solution $\boldsymbol{x} \in \mathbb{R}^n$ of (1) and the 'perturbed solution' $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ of (27).

To make such considerations meaningful, we need to first pose a condition under which the unique solvability of (27) follows from that of (1). The topological interpretation of the following theorem is that the invertible matrices form an open subset in the space of all matrices equipped with some operator topology. In layman's terms, if a matrix is invertible, then so are all nearby matrices.

Theorem 3.1. *Let $\|\cdot\|$ be an operator norm. If $A \in \mathbb{R}^{n\times n}$ is invertible and $\delta A \in \mathbb{R}^{n\times n}$ satisfies*

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}, \tag{28}$$

*then $A + \delta A$ is also invertible.*

Proof. Our aim is to prove that $A + \delta A$ satisfies the condition (3) of Corollary 1.1, from which its invertibility follows. The reverse triangle inequality (homework) for $0 \neq \boldsymbol{z} \in \mathbb{R}^n$ gives,

$$\frac{\|(A+\delta A)\boldsymbol{z}\|}{\|\boldsymbol{z}\|} \ge \frac{\|A\boldsymbol{z}\| - \|\delta A\boldsymbol{z}\|}{\|\boldsymbol{z}\|} \ge \min_{0\neq \boldsymbol{z}\in\mathbb{R}^n} \frac{\|A\boldsymbol{z}\|}{\|\boldsymbol{z}\|} - \max_{0\neq \boldsymbol{z}\in\mathbb{R}^n} \frac{\|\delta A\boldsymbol{z}\|}{\|\boldsymbol{z}\|}. \tag{29}$$

We make the chance of variables $\boldsymbol{y} = A\boldsymbol{z}$, i.e. $\boldsymbol{z} = A^{-1}\boldsymbol{y}$, in the first term on the right-hand side of (29), which leads to

$$\min_{0\neq \boldsymbol{z}\in\mathbb{R}^n} \frac{\|A\boldsymbol{z}\|}{\|\boldsymbol{z}\|} = \min_{0\neq \boldsymbol{y}\in\mathbb{R}^n} \frac{\|\boldsymbol{y}\|}{\|A^{-1}\boldsymbol{y}\|} = \left( \max_{0\neq \boldsymbol{y}\in\mathbb{R}^n} \frac{\|A^{-1}\boldsymbol{y}\|}{\|\boldsymbol{y}\|} \right)^{-1} = \|A^{-1}\|^{-1}$$

due to the monotonicity of the inverse power on the positive real axis $\mathbb{R}_+$. Substituting this back in (29) and recalling the definition of an operator norm, yields

$$\frac{\|(A+\delta A)\boldsymbol{z}\|}{\|\boldsymbol{z}\|} \ge \frac{1}{\|A^{-1}\|} - \|\delta A\| > 0,$$

where the strict inequality is our assumption. Equivalently,

$$\|(A+\delta A)\boldsymbol{z}\| \geq \left(\frac{1}{\|A^{-1}\|} - \|\delta A\|\right)\|\boldsymbol{z}\| > 0$$

for all $0 \neq \boldsymbol{z} \in \mathbb{R}^n$. Hence, the nullspace $N(A+\delta A)$ cannot contain any nonzero vectors. □

The following corollary indicates that the norm of the inverse of $A+\delta A$ can, in fact, be controlled by the norms of the inverse $A^{-1}$ and the perturbation $\delta A$.

Corollary 3.1. *Under the assumptions of Theorem 3.1, it holds that*

$$\big\|(A+\delta A)^{-1}\big\| \leq \frac{\|A^{-1}\|}{1-\|\delta A\|\|A^{-1}\|}$$

*for any operator norm* $\|\cdot\|$.

Proof. Using (22), we first of all obtain

$$\big\|(A+\delta A)^{-1}\big\| = \big\|\big((I+\delta AA^{-1})A\big)^{-1}\big\| = \big\|A^{-1}(I+\delta AA^{-1})^{-1}\big\| \leq \|A^{-1}\|\big\|(I+\delta AA^{-1})^{-1}\big\|.$$

By the definition of an operator norm,

$$\|(I+\delta AA^{-1})^{-1}\| = \max_{0\neq\boldsymbol{x}\in\mathbb{R}^n}\frac{\|(I+\delta AA^{-1})^{-1}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{0\neq\boldsymbol{y}\in\mathbb{R}^n}\frac{\|\boldsymbol{y}\|}{\|(I+\delta AA^{-1})\boldsymbol{y}\|}, \tag{30}$$

where the second step corresponds to the change of variables $\boldsymbol{x} = (I+\delta AA^{-1})\boldsymbol{y}$. Furthermore, the denominator on the right-hand side of (30) can be estimated with the help of the reverse triangle inequality as follows:

$$\big\|(I+\delta AA^{-1})\boldsymbol{y}\big\| \geq \|\boldsymbol{y}\| - \|\delta AA^{-1}\boldsymbol{y}\| \geq \big(1-\|\delta A\|\|A^{-1}\|\big)\|\boldsymbol{y}\| > 0, \tag{31}$$

where we also twice used (23) as well as our assumption on the size of $\|A^{-1}\|$. Plugging (30) and (31) in turns in the first estimate of this proof, we finally get

$$\begin{aligned}\big\|(A+\delta A)^{-1}\big\| &\leq \|A^{-1}\|\max_{0\neq\boldsymbol{y}\in\mathbb{R}^n}\frac{\|\boldsymbol{y}\|}{\|(I+\delta AA^{-1})\boldsymbol{y}\|} \leq \|A^{-1}\|\max_{0\neq\boldsymbol{y}\in\mathbb{R}^n}\frac{\|\boldsymbol{y}\|}{\big(1-\|\delta A\|\|A^{-1}\|\big)\|\boldsymbol{y}\|}\\ &= \frac{\|A^{-1}\|}{1-\|\delta A\|\|A^{-1}\|},\end{aligned}$$

which completes the proof. □

By subtracting the equations (1) and (27), we obtain a linear system that determines the *error* $\tilde{\boldsymbol{x}} - \boldsymbol{x}$, that is,

$$(A+\delta A)(\tilde{\boldsymbol{x}}-\boldsymbol{x}) = \delta\boldsymbol{b} - \delta A\boldsymbol{x}. \tag{32}$$

The following theorem relates the (relative) error to the relative sizes of the perturbations, i.e. $\|\delta\boldsymbol{b}\|/\|\boldsymbol{b}\|$ and $\|\delta A\|/\|A\|$, and the condition number of $A$, defined as

$$\kappa(A) := \|A\|\|A^{-1}\|.$$

Take note that the condition number of a matrix depends on the considered (operator) norm.

Theorem 3.2. *Suppose the assumptions of Theorem 3.1 are valid. Then it holds that*

$$\frac{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\kappa(A)}{1-\frac{\|\delta A\|}{\|A\|}\kappa(A)}\left(\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}+\frac{\|\delta A\|}{\|A\|}\right) \tag{33}$$

*for any* $\|\cdot\|$ *operator norm and the corresponding condition number* $\kappa(A)$.

Proof. According to Theorem 3.1, the matrix $A+\delta A \in \mathbb{R}^{n\times n}$ is invertible and the error $\tilde{\boldsymbol{x}}-\boldsymbol{x}$ can thus be solved from (32) as

$$\tilde{\boldsymbol{x}}-\boldsymbol{x} = (A+\delta A)^{-1}(\delta\boldsymbol{b}-\delta A\boldsymbol{x}).$$

In particular,

$$\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\| \leq \|(A+\delta A)^{-1}\|\|\delta\boldsymbol{b}-\delta A\boldsymbol{x}\| \leq \|(A+\delta A)^{-1}\|\left(\|\delta\boldsymbol{b}\|+\|\delta A\boldsymbol{x}\|\right).$$

due to basic properties of an operator norm and the triangle inequality. The stability result of Corollary 3.1 yields

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq \frac{\|A^{-1}\|}{1 - \|\delta A\| \|A^{-1}\|} \left(\|\delta \boldsymbol{b}\| + \|\delta A \boldsymbol{x}\|\right).$$

Dividing by $\|\boldsymbol{x}\|$ and using the estimate $\|\delta A \boldsymbol{x}\| \leq \|\delta A\| \|\boldsymbol{x}\|$ gives

$$\frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A^{-1}\|}{1 - \|\delta A\| \|A^{-1}\|} \left( \frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{x}\|} + \|\delta A\| \right) = \frac{\|A\| \|A^{-1}\|}{1 - \frac{\|\delta A\|}{\|A\|} \|A^{-1}\| \|A\|} \left( \frac{\|\delta \boldsymbol{b}\|}{\|A\| \|\boldsymbol{x}\|} + \frac{\|\delta A\|}{\|A\|} \right), \tag{34}$$

where the latter step is mere algebraic manipulation. Since $\|\boldsymbol{b}\| = \|A\boldsymbol{x}\| \leq \|A\| \|\boldsymbol{x}\|$, we finally obtain

$$\frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A^{-1}\| \|A\|}{1 - \frac{\|\delta A\|}{\|A\|} \|A^{-1}\| \|A\|} \left( \frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|} + \frac{\|\delta A\|}{\|A\|} \right). \tag{35}$$

Substituting the definition of the condition number $\kappa(A)$ completes the proof. □

## 4. Solving $Ax = b$ on a computer

Computers operate using finite precision arithmetic: numbers are not stored exactly nor are operations performed exactly. A simple model for error in arithmetic operations is

$$fl(a \odot b) = (1 + \delta)(a \odot b), \tag{36}$$

where $a, b \in \mathbb{R}$, $\odot$ is one of the elementary operations, i.e. $\odot = +, -, *$ or $/$, and $|\delta| \leq u$, where $u$ is the machine unit or the machine epsilon. The result of a given operation in finite precision is denoted by $fl(a \odot b)$. The size of the machine epsilon depends on how numbers are represented, but it is typically of the size $u = 2.22 \cdot 10^{-16}$.

As an example, due to the finite precision, a numerically computed $LU$ decomposition[4] produced, say, by MATLAB is not exact, but only an approximation of an exact one. The triangular matrices $L \in \mathbb{R}^{n \times n}$ and $U \in \mathbb{R}^{n \times n}$ given by the computer do not exactly satisfy $A = LU$ for a given $A \in \mathbb{R}^{n \times n}$, but instead

$$LU = A + \delta A.$$

Hence, the utilization of the LU decomposition does not solve the equation (1), but a related linear system.

$$(A + \delta A)\tilde{\boldsymbol{x}} = \boldsymbol{b}.$$

An interesting and practical question is: How close is $\boldsymbol{x}$ to $\tilde{\boldsymbol{x}}$?

The answer can be given in two steps,

- Estimate the size of $\delta A$. For the $LU$ decomposition, it can be shown that

  $$\|\delta A\| \leq nu \|L\| \|U\|$$

  for the Frobenius, 1 and $\infty$ norms. In most cases, the ratio $(\|L\| \|U\|)/\|A\|$ is of moderate size, but one can also construct examples for which it tends to infinity. Fortunately, such examples are rarely encountered in practice.
- Use Theorem 3.2 to obtain an estimate for the relative change in the solution:

  $$\frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\kappa(A)}{1 - \frac{\|\delta A\|}{\|A\|} \kappa(A)} \frac{\|\delta A\|}{\|A\|}.$$

  Observe that the relative error $\|\delta A\|/\|A\|$ is amplified by the condition number $\kappa(A)$ both in the numerator and the denominator.

[4] Recall that the LU decomposition is a way of factoring a matrix $A$ as a product of a lower $L$ and an upper $U$ triangular matrix. It is commonly used by, e.g., mathematical software to solve systems of linear equations.

The heuristic of MATLAB is to assume that $\|\delta A\|/\|A\|$ behaves like the machine epsilon and to issue a warning if $\kappa(A)$ is sufficiently large.

To summarize, if $\kappa(A) \geq 10^{16}$, the given linear system cannot be solved on a computer using (mere) double precision arithmetic. Even small floating point errors are amplified by $\kappa(A)$ so that they become significant. To conclude: **Not all theoretically solvable linear systems can be solved on a computer!**

**4.1. A bad example: the Hilbert matrix.** A classical example of a linear system that cannot be solved accurately on a computer arises from approximation theory. Consider approximating a given function $f$ on the interval $(0, 1)$ by the $n$th degree polynomials $P^n$ in the sense of least squares. In other words, one is interested in the solution of the problem

$$\min_{p \in P^n} \frac{1}{2} \int_0^1 (f - p)^2 \, dx. \tag{37}$$

This minimization problem is quadratic and can, in principle, be solved rather easily.

The naive approach is to represent the $n$th degree polynomials in the monomial basis $\{x^{j-1}\}_{j=1}^{n+1}$ as

$$p(x) := \sum_{j=1}^{n+1} \alpha_j x^{j-1},$$

which means that each polynomial in $P^n$ has a one-to-one correspondence with a coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^{n+1}$. Analogously, the original problem (37) is equivalent to a certain minimization problem over $\mathbb{R}^{n+1}$:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{n+1}} \frac{1}{2} \int_0^1 \Big( f - \sum_{j=1}^{n+1} \alpha_j x^{j-1} \Big)^2 dx. \tag{38}$$

The solution of (38) can be characterized, e.g., by taking the gradient with respect to the coefficient vector and equating it to zero. For each $i = 1, \ldots, n+1$, this corresponds to

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} \frac{1}{2} \int_0^1 \Big( f - \sum_{j=1}^{n+1} \alpha_j x^{j-1} \Big)^2 \, dx \; &= \int_0^1 \Big( \sum_{j=1}^{n+1} \alpha_j x^{j-1} - f \Big) x^{i-1} \, dx \\ &= \sum_{j=1}^{n+1} \alpha_j \int_0^1 x^{i+j-2} \, dx - \int_0^1 f x^{i-1} \, dx \; = \; 0. \end{aligned}$$

These $n+1$ linear equations can be presented in a matrix form: Find $\boldsymbol{\alpha} \in \mathbb{R}^{n+1}$ such that

$$H\boldsymbol{\alpha} = \boldsymbol{b}, \tag{39}$$

where $H_{ij} = \int_0^1 x^{i+j-2} \, dx$ and $b_i = \int_0^1 f x^{i-1} \, dx$ for $i, j = 1, \ldots, n+1$. The integrals defining $H \in \mathbb{R}^{(n+1)\times(n+1)}$ can be computed by hand to obtain

$$H_{ij} = \frac{1}{i + j - 1}, \qquad i, j = 1, \ldots, n+1.$$

The condition number of this so-called *Hilbert matrix* grows extremely rapidly as a function of the dimension $n$. As an example, for $n = 15$ the condition number is already so large that the linear system (39) cannot be accurately solved using a computer.

Regardless of the above explained difficulties, the original polynomial approximation problem (37) can be solved efficiently and accurately. The trick is to use some other basis polynomials in place of the monomials. For example, the use of Legendre polynomials leads to significantly improved stability.

CHAPTER 2

# Eigenvalue problems

This second chapter discusses eigenvalues of matrices and their applications. The main application considered in these notes is the definition of matrix valued functions such as the matrix power or matrix exponential. Such functions are used, e.g., to solve recursive equations or to study the behavior of a system of ordinary differential equations close to an equilibrium point. In addition, eigenvalues and eigenvectors serve as a (computational) tool in many problems arising, e.g., from mathematics, statistics, physics or computer sciences.

We will discuss the following themes:

(1) Basic eigenvalue theory with emphasis on Hermitian matrices,

(2) Similarity transformations,

(3) Matrix exponential and its applications.

The eigenvalues and eigenvectors of a *square* matrix $A \in \mathbb{C}^{n\times n}$ are the solutions of the following problem: Find $(\lambda, \boldsymbol{v}) \in \mathbb{C} \times (\mathbb{C}^n \setminus \{0\})$ such that

$$A\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{40}$$

It should be emphasized that an eigenvector can never be the zero vector. When considering eigenvalues, we need to employ complex numbers $\mathbb{C}$ and the spaces of complex valued vectors $\mathbb{C}^n$ and matrices $\mathbb{C}^{n\times n}$ as even real matrices may have complex eigenvalues and vectors.

Almost all theory we have discussed thus far during these lectures directly applies to vectors in $\mathbb{C}^n$. The only notable exception is the definition of an inner product: The symmetry, i.e. (ii) of Definition 2.2 in Chapter 1, is replaced by the *conjugate symmetry* condition

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \overline{\langle \boldsymbol{y}, \boldsymbol{x} \rangle}.$$

This also means that the bilinearity condition (iii) of Definition 2.2 turns into *sesquilinearity*, i.e., the linearity with respect to the second variable turns into *antilinearity*: $\langle \boldsymbol{x}, \alpha\boldsymbol{y} \rangle = \overline{\alpha}\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. The Euclidean inner product and norm are defined in $\mathbb{C}^n$ as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x} \cdot \overline{\boldsymbol{y}} = \boldsymbol{y}^*\boldsymbol{x} \qquad \text{and} \qquad \|\boldsymbol{x}\|_2 := (\boldsymbol{x}^*\boldsymbol{x})^{1/2},$$

where the ‘overline’ marks componentwise complex conjugation and $*$ denotes the conjugate transpose,

$$\boldsymbol{x}^* = \overline{\boldsymbol{x}}^T.$$

In particular, if $\boldsymbol{x} = \boldsymbol{a} + \mathrm{i}\boldsymbol{b}$ for some $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, then

$$\overline{\boldsymbol{x}} = \boldsymbol{a} - \mathrm{i}\boldsymbol{b}.$$

## 1. Basic eigenvalue theory

The equation (40) can be reformulated as

$$(A - \lambda I)\,\boldsymbol{v} = 0.$$

In other words, if $\lambda \in \mathbb{C}$ is an eigenvalue of $A$, then the matrix $A - \lambda I$ has a nontrivial nullspace consisting of the corresponding eigenvectors. The eigenvalues can be determined by using the connection between the nullspace and the determinant indicated by Corollary 1.1 in Chapter 1: $N(A - \lambda)$ is nontrivial if and only if

$$\det\,(A - \lambda I) = 0.$$

When an eigenvalue has been found, the related eigenvector(s) are computed by finding a basis for $N(A - \lambda I)$. The equation (40) defines the eigenvalues $\lambda \in \mathbb{C}$ uniquely, but there exists infinitely many eigenvectors for each eigenvalue: the eigenvectors corresponding to an eigenvalue $\lambda$ form a subspace of $\mathbb{C}^n$ called the eigenspace,

$$E_\lambda := N(A - \lambda I).$$

The dimension of $E_\lambda$ can be any integer between 1 and $n$ and it is called the *geometric multiplicity* $\mu_G(\lambda)$ of the eigenvalue $\lambda$.

Example 1.1. *Let*

$$A = \begin{bmatrix} 3 & -1 \\ 1 & 1 \end{bmatrix}.$$

*The eigenvalues of $A$ are the solutions to the equation*

$$\det(A - \lambda I) = \det \begin{bmatrix} 3-\lambda & -1 \\ 1 & 1-\lambda \end{bmatrix} = (3-\lambda)(1-\lambda) + 1 = \lambda^2 - 4\lambda + 4 = (\lambda - 2)^2 = 0.$$

*Hence, there is only one eigenvalue $\lambda = 2$ that is a double root of the* characteristic polynomial*. The eigenspace corresponding to the eigenvalue $\lambda = 2$ is $E_2 = N(A - 2I)$. Because*

$$A - 2I = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix},$$

*it is easy to deduce that $E_2 = \text{span}\{[1,1]^T\}$ is a one-dimensional subspace of $\mathbb{R}^2$.*

Definition 1.1. *The* characteristic polynomial *of a matrix $A \in \mathbb{C}^{n\times n}$ is defined as*

$$p_A(\lambda) = \det(A - \lambda I),$$

*i.e., as the polynomial whose roots define the eigenvalues.*

Recall the subdeterminant rule that can be employed in analyzing determinants or computing them by hand.[1] The rule can be applied either column-wise,

$$\det A = (-1)^{j-1} \sum_{i=1}^{n} (-1)^{i-1} a_{ij} \det [A]_{i,j},$$

or row-wise,

$$\det A = (-1)^{j-1} \sum_{i=1}^{n} (-1)^{i-1} a_{ji} \det [A]_{j,i}.$$

In the above formulas, $j$ is the index of the column or the row with respect to which the determinant is expanded. Moreover, $[A]_{i,j} \in \mathbb{C}^{(n-1)\times(n-1)}$ is the square matrix that is obtained by removing the $i$th row and the $j$th column from the matrix $A$. A couple of examples are in order.

Example 1.2. *Let*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

*Then*

$$[A]_{1,1} = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}, \quad [A]_{1,2} = \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix}, \quad [A]_{1,3} = \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}.$$

[1] Mathematical software such as MATLAB do *not* numerically evaluate determinants using the subdeterminant rule. Typically, they first form a suitable factorization of the examined matrix involving only triangular matrices, e.g., the LU decomposition. Then two properties of the determinant are used: (i) The determinant of a product is the product of the determinants and (ii) the determinant of a triangular matrix is the product of its diagonal elements.

Example 1.3. *For*

$$A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 5 \\ 0 & 6 & 7 & 0 \\ 8 & 0 & 0 & 10 \end{bmatrix},$$

*the subdeterminant rule with respect to the first row gives*

$$\det A = \det \begin{bmatrix} 4 & 0 & 5 \\ 6 & 7 & 0 \\ 0 & 0 & 10 \end{bmatrix} - 2 \det \begin{bmatrix} 3 & 0 & 5 \\ 0 & 7 & 0 \\ 8 & 0 & 10 \end{bmatrix}.$$

*Applying the subdeterminant rule next to* $[A]_{1,1}$*, we obtain*

$$\det \begin{bmatrix} 4 & 0 & 5 \\ 6 & 7 & 0 \\ 0 & 0 & 10 \end{bmatrix} = 4 \det \begin{bmatrix} 7 & 0 \\ 0 & 10 \end{bmatrix} + 5 \det \begin{bmatrix} 6 & 7 \\ 0 & 0 \end{bmatrix} = 280.$$

*The determinant of* $[A]_{1,2}$ *is computed recursively according to the same procedure. This gives* $\det [A]_{1,2} = -70$ *and altogether* $\det A = 420$.

With the help of the subdeterminant rule, it is straightforward to prove that $p_A$ is always a polynomial of degree $n$.

Lemma 1.1. *If* $A \in \mathbb{C}^{n\times n}$*, then* $p_A(\lambda) = \det(A - \lambda I)$ *is a polynomial of degree* $n$*. Moreover, the coefficient of the* $n$*th order term is* $(-1)^n$.

Proof. The proof is based on induction with respect to the dimension of a square matrix.

Base case: For a general $A \in \mathbb{C}^{2\times 2}$,

$$\det(A - \lambda I) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21})$$

is clearly a second order polynomial and the coefficient of the term $\lambda^2$ is $1 = (-1)^2$.

Induction assumption: For any $B \in \mathbb{C}^{k\times k}$, $\det(B - \lambda I)$ is a polynomial of degree $k$ and the coefficient of the $k$th order term is $(-1)^k$.

Induction step: Let $A \in \mathbb{C}^{(k+1)\times(k+1)}$ be arbitrary. The subdeterminant rule yields

$$\det(A - \lambda I) = (a_{11} - \lambda)\det([A - \lambda I]_{1,1}) + \sum_{i=2}^{k} (-1)^{i-1} a_{i1} \det([A - \lambda I]_{i,1}). \tag{41}$$

By the induction assumption, the first subdeterminant appearing in (41) is a polynomial of degree $k$ having $(-1)^k$ as the coefficient of the $k$th order term. Moreover, each addend in the summation over $i$ on the right-hand side of (41) amounts (the proof of this claim is left as an exercise) to a polynomial of degree at most $k$ in $\lambda$, and the first term on the right-hand side of (41) is a polynomial of degree $k+1$ in $\lambda$ having $(-1)^{k+1}$ as the coefficient of the $(k+1)$th order term. Thus, $\det(A - \lambda I)$ is altogether a polynomial of degree $k+1$ having $(-1)^{k+1}$ as the coefficient of the $(k+1)$th order term, which completes the proof. □

As the eigenvalues of a matrix $A$ are the roots of its characteristic polynomial $p_A$, the eigenvalues can be studied by investigating the characteristic polynomial.

Lemma 1.2. *A matrix* $A \in \mathbb{C}^{n\times n}$ *has at most* $n$ *distinct eigenvalues.*

Proof. The eigenvalues of $A \in \mathbb{C}^{n\times n}$ are the roots of the characteristic polynomial $p_A$. Because the polynomial $p_A$ is exactly of the order $n$, it has at most $n$ distinct roots (and altogether $n$ complex roots if counted according to their multiplicity). □

Eigenvalues can be multiple roots of the characteristic polynomial, as in Example 1.1. Assume $A \in \mathbb{C}^{n\times n}$ has the distinct eigenvalues $\lambda_1, \dots, \lambda_k$, where $1 \le k \le n$. By virtue of the fundamental theorem of algebra, the characteristic polynomial can be factored as

$$p_A(\lambda) = (-1)^n \prod_{i=1}^{k} (\lambda - \lambda_i)^{\mu_A(\lambda_i)}, \tag{42}$$

where $\sum_{i=1}^{k} \mu_A(\lambda_i) = n$ since $p_A(\lambda)$ is of degree $n$. The positive integer power $\mu_A(\lambda_i)$ is the *algebraic multiplicity* of the eigenvalue $\lambda_i$, i.e., its multiplicity as a root of the characteristic polynomial. Recall that the dimension of the eigenspace $E_{\lambda_i} = N(A - \lambda_i I)$ was dubbed the geometric multiplicity of the eigenvalue $\lambda_i$ and denoted by $\mu_G(\lambda_i)$. It can be shown that the geometric multiplicity $\mu_G(\lambda_i)$ is at most the same as the algebraic multiplicity $\mu_A(\lambda_i)$ for all $i = 1, \dots, k$ (the proof is omitted). However, if $\mu_A(\lambda_i) = 1$, then also $\mu_G(\lambda_i) = \mu_A(\lambda_i) = 1$ because each eigenvalue has at least one-dimensional eigenspace.

The eigenvectors related to different eigenvalues are linearly independent.

THEOREM 1.1. *Any eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{C}^n$ corresponding to the distinct eigenvalues $\lambda_1, \dots, \lambda_k \in \mathbb{C}$ of $A \in \mathbb{C}^{n\times n}$ are linearly independent.*

PROOF. The proof is based on induction with respect to the number of considered eigenvalues.

Base case (i.e., any pair of eigenvectors are linearly independent): Without loss of generality, we may only consider the first two eigenvalues $\lambda_1$ and $\lambda_2$; the proof for any other pair would be analogous. Moreover, we may assume that $\lambda_1 \neq 0$ because one of the two eigenvalues must be nonzero, and we can rename the eigenvalues if necessary.

We argue by contradiction: Assume that the corresponding eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly dependent, that is, there exists $0 \neq \alpha \in \mathbb{C}$ such that

$$\boldsymbol{v}_1 = \alpha \boldsymbol{v}_2.$$

Multiplying this equation with $A$ and using the definition of eigenvectors gives

$$\lambda_1 \boldsymbol{v}_1 = \alpha \lambda_2 \boldsymbol{v}_2, \qquad \text{i.e.,} \qquad \boldsymbol{v}_1 = \frac{\lambda_2}{\lambda_1} \alpha \boldsymbol{v}_2.$$

Subtracting the above identities leads to

$$\Big(1 - \frac{\lambda_2}{\lambda_1}\Big) \alpha \boldsymbol{v}_2 = 0.$$

As $\alpha \neq 0$ by assumption and $\boldsymbol{v}_2 \neq 0$ by definition, it must hold that $\lambda_2/\lambda_1 = 1$, which is a contradiction because $\lambda_1$ and $\lambda_2$ are distinct. Hence, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly independent.

Induction assumption: Assume that any $j$ vectors in the set $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ are linearly independent $(1 \le j \le k-1)$.

Induction step: Consider any $j+1$ vectors in $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$. After renumbering, we may assume that they are $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{j+1}$.

Let $\boldsymbol{\alpha} \in \mathbb{C}^{j+1}$ be such that

$$\sum_{i=1}^{j+1} \alpha_i \boldsymbol{v}_i = 0. \tag{43}$$

Multiplying (43) by $A$ gives

$$\sum_{i=1}^{j+1} \alpha_i \lambda_i \boldsymbol{v}_i = 0.$$

Without loss of generality, we may assume that $\lambda_1 \neq 0$; at most one of the distinct eigenvalues is zero and we may rename the eigenvalues and eigenvectors if need be. Hence,

$$\alpha_1 \boldsymbol{v}_1 + \sum_{i=2}^{j+1} \alpha_i \frac{\lambda_i}{\lambda_1} \boldsymbol{v}_i = 0.$$

Subtracting this from (43), we get

$$\sum_{i=2}^{j+1} \alpha_i \Big(1 - \frac{\lambda_i}{\lambda_1}\Big) \boldsymbol{v}_i = 0.$$

Since the $j$ eigenvectors $\boldsymbol{v}_2, \dots, \boldsymbol{v}_{j+1}$ are linearly independent by the induction assumption and $\lambda_i/\lambda_1 \neq 1$ since the considered eigenvalues are distinct, it must hold that $\alpha_i = 0$ for all $i = 2, \dots, j+1$. Using this information, say, in (43), it follows that also $\alpha_1 = 0$, i.e., $\boldsymbol{\alpha} = 0$, and thus $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{j+1}$ are linearly independent. □

If the geometric and the algebraic multiplicity for each eigenvalue of $A \in \mathbb{R}^{n\times n}$ are the same, Theorem 1.1 implies that there exists $n$ linearly independent eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ for $A$: If $\mu_G(\lambda) = \mu_A(\lambda)$ for all eigenvalues, one can find $\mu_A(\lambda)$ linearly independent eigenvectors from each eigenspace $E_\lambda$ of $A$ because $\mu_G(\lambda) = \mu_A(\lambda)$ is its dimension. Since Theorem 1.1 guarantees that eigenvectors corresponding to distinct eigenvalues are automatically linearly independent and the algebraic multiplicities sum to $n$ (cf. (42)), one can indeed altogether find $n$ linearly independent eigenvectors. On the other, if $\mu_G(\lambda) < \mu_A(\lambda)$ for any eigenvalue $\lambda$, there does not exist a *full set* of linearly independent eigenvectors. On the positive side, if $A$ has $n$ distinct eigenvalues, then $\mu_G(\lambda) = \mu_A(\lambda) = 1$ for all eigenvalues and there are $n$ linearly independent eigenvectors.

Assume then that $A \in \mathbb{R}^{n\times n}$ has $n$ linearly independent eigenvectors and stack them as the columns of a matrix:

$$V := \big[\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\big] \in \mathbb{C}^{n\times n}.$$

The matrix product $AV$ gives

$$AV = \big[A\boldsymbol{v}_1, \dots, A\boldsymbol{v}_n\big] = \big[\lambda_1\boldsymbol{v}_1, \dots, \lambda_n\boldsymbol{v}_n\big], \tag{44}$$

where the eigenvalues are repeated according to their algebraic/geometric multiplicity. If $\Lambda \in \mathbb{C}^{n\times n}$ is the diagonal matrix carrying the eigenvalues of $A$ in the corresponding order, i.e.,

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

then (44) can be rewritten as

$$AV = V\Lambda.$$

As $V$ has linearly independent columns, it is invertible (see (2) in Corollary 1.1) and we obtain

$$A = V\Lambda V^{-1}. \tag{45}$$

This is an eigendecomposition for the matrix $A$. If the algebraic and geometric multiplicities are not the same for all eigenvalues, such a decomposition does not exist. If a decomposition of the type (45) exists, $A$ is called *diagonalizable*. Finally, note that a decomposition of the form (45) is never unique: one can change the order of the eigenvalues, the length of the eigenvectors, or even pick different bases for the eigenspaces.

Remark 1.1. *The decomposition (45) has the following intuitive interpretation: The linear map represented by $A \in \mathbb{R}^{n\times n}$ in the standard/Cartesian basis has a diagonal representation $\Lambda \in \mathbb{R}^{n\times n}$ in the eigenbasis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$. To understand this, let $\boldsymbol{\alpha} \in \mathbb{R}^n$ be the coordinates of an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$ in the basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, that is,*

$$\boldsymbol{x} = V\boldsymbol{\alpha} \quad \Longleftrightarrow \quad \boldsymbol{\alpha} = V^{-1}\boldsymbol{x}.$$

*Hence,*

$$A\boldsymbol{x} = V\Lambda V^{-1}\boldsymbol{x} = V(\Lambda\boldsymbol{\alpha}),$$

*which just means that the coordinates of the image vector $A\boldsymbol{x} \in \mathbb{R}^n$ in the eigenbasis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ are $\Lambda\boldsymbol{\alpha} \in \mathbb{R}^n$. In other words, the action of the considered linear map on the coordinates of $\boldsymbol{x}$ in the basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is realized by a diagonal matrix $\Lambda$, i.e., the coordinates change according to the rule $\boldsymbol{\alpha} \mapsto \Lambda\boldsymbol{\alpha}$. This is why a matrix with a representation of the form (45) is called diagonalizable.*

## 2. Hermitian matrices

*Hermitian* matrices are encountered in numerous applications. They have also many nice properties that make them 'easily approachable'. We start our treatment with a couple of definitions that will be put to use immediately afterwards.

Definition 2.1. *A matrix $A \in \mathbb{C}^{n\times n}$ is* Hermitian *if $A^* = A$. In particular, a real matrix is Hermitian if and only if it is* symmetric, *i.e. $A^T = A$.*

Definition 2.2. *A matrix $U \in \mathbb{C}^{n\times n}$ is* unitary *if $U^* = U^{-1}$. In particular, a real matrix is unitary if and only if it is* orthogonal, *i.e. $U^T = U^{-1}$.*

By definition, the diagonal elements of a Hermitian matrix must be real. Observe also that the columns (or the rows) of a unitary matrix form an orthonormal basis for $\mathbb{C}^n$, that is, they are of unit Euclidean length and mutually orthogonal with respect to the Euclidean inner product. Indeed, if $U \in \mathbb{C}^{n\times n}$ is unitary and we write $U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n]$, then

$$U^*U = \begin{bmatrix} \boldsymbol{u}_1^*\boldsymbol{u}_1 & \boldsymbol{u}_1^*\boldsymbol{u}_2 & \cdots & \boldsymbol{u}_1^*\boldsymbol{u}_n \\ \boldsymbol{u}_2^*\boldsymbol{u}_1 & \boldsymbol{u}_2^*\boldsymbol{u}_2 & \cdots & \boldsymbol{u}_2^*\boldsymbol{u}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{u}_n^*\boldsymbol{u}_1 & \boldsymbol{u}_n^*\boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n^*\boldsymbol{u}_n \end{bmatrix} = I, \tag{46}$$

meaning that the off-diagonal inner products vanish and the squared norms on the diagonal equal one.

Our first aim is to show that any Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is unitarily diagonalizable, that is, there exists a (nonunique) decomposition

$$A = Q\Lambda Q^*, \tag{47}$$

where $Q \in \mathbb{C}^{n\times n}$ is unitary and $\Lambda \in \mathbb{R}^{n\times n}$ a diagonal matrix with *real* entries. This property is closely related to the fact that the eigenvalues of a Hermitian matrix are real and eigenvectors corresponding to different eigenvalues are orthogonal.

Lemma 2.1. *The eigenvalues of a Hermitian $A \in \mathbb{C}^{n\times n}$ are real.*

Proof. Let $\lambda$ be an arbitrary eigenvalue of $A$ and let $\boldsymbol{v}$ be a corresponding eigenvector. In particular,

$$\boldsymbol{v}^*A\boldsymbol{v} = \boldsymbol{v}^*(\lambda\boldsymbol{v}) = \lambda\|\boldsymbol{v}\|_2^2.$$

On the other hand, as $A^* = A$,

$$\boldsymbol{v}^*A\boldsymbol{v} = \boldsymbol{v}^*A^*\boldsymbol{v} = (A\boldsymbol{v})^*\boldsymbol{v} = (\lambda\boldsymbol{v})^*\boldsymbol{v} = \overline{\lambda}\|\boldsymbol{v}\|_2^2.$$

Subtracting these identities leads to

$$(\lambda - \overline{\lambda})\|\boldsymbol{v}\|_2^2 = 2\,\mathrm{Im}\lambda\,\|\boldsymbol{v}\|_2^2 = 0.$$

Since $\boldsymbol{v} \neq 0$ by definition, it must hold that $\mathrm{Im}\lambda = 0$, i.e., $\lambda$ is real. □

Lemma 2.2. *Let $\boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{C}^n$ be eigenvectors corresponding to distinct eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$ of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$, respectively. Then $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are orthogonal in the sense that $\boldsymbol{q}_1^*\boldsymbol{q}_2 = 0$.*

PROOF. As the eigenvalues of a Hermitian matrix are real, it holds that

$$\lambda_1 \boldsymbol{q}_1^* \boldsymbol{q}_2 = (\lambda_1 \boldsymbol{q}_1)^* \boldsymbol{q}_2 = (A\boldsymbol{q}_1)^* \boldsymbol{q}_2 = \boldsymbol{q}_1^* A^* \boldsymbol{q}_2 = \boldsymbol{q}_1^* A \boldsymbol{q}_2.$$

On the other hand,

$$\lambda_2 \boldsymbol{q}_1^* \boldsymbol{q}_2 = \boldsymbol{q}_1^* (\lambda_2 \boldsymbol{q}_2) = \boldsymbol{q}_1^* A \boldsymbol{q}_2.$$

Subtracting these equalities gives

$$(\lambda_2 - \lambda_1) \boldsymbol{q}_1^* \boldsymbol{q}_2 = 0,$$

which proves the claim since $\lambda_2 \neq \lambda_1$ by assumption. □

Now we have the necessary tools to establish the unitary decomposition (47).

THEOREM 2.1. *For any Hermitian matrix $A \in \mathbb{C}^{n \times n}$, there exists a unitary $Q \in \mathbb{C}^{n \times n}$ and a diagonal $\Lambda \in \mathbb{R}^{n \times n}$ such that the decomposition* (47) *is valid.*

PROOF. The proof is based on induction with respect to the dimension $n$ of a Hermitian matrix.

Base case ($n = 1$): A Hermitian $1 \times 1$ matrix is a single real number, say, $a \in \mathbb{R}$ that can be decomposed as $a = 1\,a\,1$, i.e., $\Lambda = a$ itself and $Q = 1$.

Induction assumption: The claim is true for $n = k$.

Induction step: Let $A \in \mathbb{C}^{(k+1)\times(k+1)}$ be Hermitian and note that it has at least one eigenvalue $\lambda_1 \in \mathbb{R}$ with a corresponding eigenvector $\boldsymbol{q}_1$ satisfying $\|\boldsymbol{q}_1\|_2 = 1$ (take any eigenvector for $\lambda$ and divide it by its 2-norm). Introduce $k$ auxiliary unit vectors $\boldsymbol{q}_2, \dots, \boldsymbol{q}_{k+1} \in \mathbb{C}^{k+1}$ that are mutually orthogonal as well as orthogonal to $\boldsymbol{q}_1$; such can be found, e.g., via the Gram–Schmidt orthogonalization process that will be discussed in Chapter 3. We denote

$$\tilde{Q} = \big[\boldsymbol{q}_2, \dots, \boldsymbol{q}_{k+1}\big] \in \mathbb{C}^{(k+1)\times k} \qquad \text{and} \qquad Q = \big[\boldsymbol{q}_1,\, \tilde{Q}\big] \in \mathbb{C}^{(k+1)\times(k+1)},$$

and note that $Q$ is unitary (see (46)).

Consider

$$\begin{aligned} Q^* A Q = \big[\boldsymbol{q}_1,\, \tilde{Q}\big]^* A \big[\boldsymbol{q}_1,\, \tilde{Q}\big] &= \begin{bmatrix} \boldsymbol{q}_1^* \\ \tilde{Q}^* \end{bmatrix} \big[A\boldsymbol{q}_1,\, A\tilde{Q}\big] \\ &= \begin{bmatrix} \boldsymbol{q}_1^* A \boldsymbol{q}_1 & \boldsymbol{q}_1^* A \tilde{Q} \\ \tilde{Q}^* A \boldsymbol{q}_1 & \tilde{Q}^* A \tilde{Q} \end{bmatrix} = \begin{bmatrix} \lambda_1 \boldsymbol{q}_1^* \boldsymbol{q}_1 & (\lambda_1 \boldsymbol{q}_1)^* \tilde{Q} \\ \lambda_1 \tilde{Q}^* \boldsymbol{q}_1 & \tilde{Q}^* A \tilde{Q} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \tilde{Q}^* A \tilde{Q} \end{bmatrix}, \end{aligned}$$

where the last step follows from the fact that $\boldsymbol{q}_1$ is of unit length and orthogonal to the columns of $\tilde{Q}$. Since the matrix $\tilde{Q}^* A \tilde{Q} \in \mathbb{C}^{k \times k}$ is obviously Hermitian, by the induction assumption there exists a unitary $Q_1 \in \mathbb{C}^{k \times k}$ and a diagonal matrix $\Lambda_1 \in \mathbb{R}^{k \times k}$ such that

$$\tilde{Q}^* A \tilde{Q} = Q_1^* \Lambda_1 Q_1.$$

In consequence,

$$Q^* A Q = \begin{bmatrix} \lambda_1 & 0 \\ 0 & Q_1^* \Lambda_1 Q_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix}.$$

Due to the unitarity of $Q$, it follows that

$$A = Q \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix} Q^* = \left( Q \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} \right) \begin{bmatrix} \lambda_1 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \left( Q \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} \right)^* \tag{48}$$

This completes the proof: (48) is a unitary decomposition of the form (47) because

$$\left( Q \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} \right)^* \left( Q \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix} Q^* Q \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} = I,$$

where we used the unitarity of both $Q$ and $Q_1$. □

The decomposition (47) can alternatively be given as

$$AQ = Q\Lambda \qquad \Longleftrightarrow \qquad \big[A\boldsymbol{q}_1, \dots, A\boldsymbol{q}_n\big] = \big[\lambda_1\boldsymbol{q}_1, \dots, \lambda_n\boldsymbol{q}_n\big],$$

which demonstrates that the columns of the unitary matrix $Q$ are orthonormal eigenvectors for $A$ and the diagonal elements of $\Lambda$, i.e. $\lambda_1, \dots, \lambda_n$, are the corresponding eigenvalues (repeated according to their algebraic/geometric multiplicity). In particular, the orthogonal eigenvectors $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ form an eigenbasis for the whole of $\mathbb{C}^{n\times n}$ (cf. Lemma 2.1).

As we already know, the eigenvalues of a matrix $A$ are the roots of the corresponding characteristic polynomial. For a Hermitian matrix, the eigenvalues can also be characterized by an optimization problem related to the *Rayleigh quotient*.

Definition 2.3. *For a Hermitian matrix $A \in \mathbb{C}^{n\times n}$, the corresponding Rayleigh quotient is defined as*

$$R(A, \boldsymbol{x}) = \frac{\boldsymbol{x}^* A\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \frac{\boldsymbol{x}^* A\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2} \in \mathbb{R}.$$

*(The Rayleigh quotient is real-valued because the numerator $\boldsymbol{x}^* A\boldsymbol{x} = \boldsymbol{x}^* A^*\boldsymbol{x} = (\boldsymbol{x}^* A\boldsymbol{x})^* = \overline{\boldsymbol{x}^* A\boldsymbol{x}}$ is real as it equals its complex conjugate.)*

Let the (real) eigenvalues of a Hermitian $A \in \mathbb{C}^{n\times n}$ be arranged in increasing order, i.e.,

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n,$$

where the eigenvalues are repeated according to their algebraic (or geometric) multiplicity. It is easy to see that the minimal and maximal values of the Rayleigh quotient are the smallest and largest eigenvalues of $A$, respectively.

Theorem 2.2. *For a Hermitian matrix $A \in \mathbb{C}^{n\times n}$, it holds that*

$$\lambda_1 = \min_{0\neq\boldsymbol{x}\in\mathbb{C}^n} R(A, \boldsymbol{x}) \qquad \textit{and} \qquad \lambda_n = \max_{0\neq\boldsymbol{x}\in\mathbb{C}^n} R(A, \boldsymbol{x}),$$

*where $\lambda_1$ is the smallest eigenvalue of $A$ and $\lambda_n$ the largest.*

Proof. Expand an arbitrary $\boldsymbol{x} \in \mathbb{C}^n$ in an orthonormal eigenbasis of $A$,

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{q}_i, \qquad \boldsymbol{\alpha} \in \mathbb{C}^n.$$

It follows that (homework),

$$\boldsymbol{x}^* A\boldsymbol{x} = \sum_{i=1}^{n} |\alpha_i|^2 \lambda_i \quad \text{ and } \quad \boldsymbol{x}^*\boldsymbol{x} = \sum_{i=1}^{n} |\alpha_i|^2,$$

and the Rayleigh quotient can thus be estimated from below as

$$R(A, \boldsymbol{x}) = \frac{\boldsymbol{x}^* A\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \frac{\sum_{i=1}^{n} |\alpha_i|^2\lambda_i}{\sum_{i=1}^{n} |\alpha_i|^2} \geq \frac{\lambda_1 \sum_{i=1}^{n} |\alpha_i|^2}{\sum_{i=1}^{n} |\alpha_i|^2} = \lambda_1.$$

On the other hand, choosing $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, \dots, n$ directly gives

$$R(A, \boldsymbol{x}) = \frac{\sum_{i=1}^{n} |\alpha_i|^2\lambda_i}{\sum_{i=1}^{n} |\alpha_i|^2} = \lambda_1.$$

This proves the first part of the claim. The second part follows via an analogous argument. □

The other eigenvalues of a Hermitian matrix $A$ can be characterized via

$$\lambda_k = \min_{\dim U = k}\ \max_{0\neq\boldsymbol{x}\in U} \frac{\boldsymbol{x}^* A\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}},$$

where the outer minimization is over subspaces of dimension $k$. This is known as the Courant–Fisher min-max theorem.

## 3. Similarity transformations

If a matrix $A \in \mathbb{C}^{n\times n}$ does not have $n$ linearly independent eigenvectors, it is not diagonalizable. However, one can still look for some other decomposition of the form

$$A = SBS^{-1}, \tag{49}$$

where $B$ is a simpler matrix than the original $A$ in some suitable sense. For example, if $B^n$ is easier to evaluate than $A^n$, then the decomposition (49) is useful when defining matrix valued functions via power series expansion because

$$A^n = SB^nS^{-1}.$$

A decomposition of the form (49) merits a separate definition.

Definition 3.1. *The matrices $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times n}$ are called* similar *if there exists an invertible $S \in \mathbb{C}^{n\times n}$ such that (49) is valid.*

In the spirit of Remark 1.1, two matrices are similar if they represent the same linear mapping in different basis. As an example, if $A$ represents a certain linear map in the Cartesian basis, then $B$ of (49) gives the same linear map in the basis defined by the columns of $S$. In particular, a diagonalizable matrix is represented by a diagonal matrix in an eigenbasis — and the eigenbasis can be chosen to be orthonormal if the matrix in question is Hermitian.

Lemma 3.1. *If $A, B \in \mathbb{C}^{n\times n}$ are similar, then the associated characteristic polynomials are the same, i.e., $p_A = p_B$.*

Proof. Assume that $A = SBS^{-1}$. Using basic properties of the determinant, we deduce

$$\begin{aligned} p_A(\lambda) = \det(A - \lambda I) &= \det\big(SBS^{-1} - \lambda SS^{-1}\big) = \det\big(S(B - \lambda I)S^{-1}\big) \\ &= \det(S)\det(B-\lambda I)\det(S^{-1}) = \det(S)\det(B-\lambda I)\det(S)^{-1} \\ &= \det(B - \lambda I) = p_B(\lambda), \end{aligned}$$

which proves the claim. □

In what follows, we will introduce two decompositions of the form (49) that are valid for any square matrix $A \in \mathbb{C}^{n\times n}$, independently of whether $A$ has a full set of linearly independent eigenvectors or not. We will start with the Schur decomposition and then briefly discuss the Jordan normal form. The latter is a theoretical tool that is rarely numerically computed in practice. We will omit the existence proof for the Jordan normal form to keep the presentation compact.

The Schur decomposition states that any square matrix is unitarily similar to an upper (or lower) triangular matrix. In other words, for each square matrix there exists an orthonormal basis in which the corresponding linear mapping is represented by an upper triangular matrix (cf. Remark 1.1).

Theorem 3.1. *For any $A \in \mathbb{C}^{n\times n}$, there exist a unitary $Q \in \mathbb{C}^{n\times n}$ and an upper triangular $T \in \mathbb{C}^{n\times n}$ such that*

$$A = QTQ^*. \tag{50}$$

Proof. The proof is based on induction with respect to the dimension $n$.

Base case ($n = 1$): A general $1 \times 1$ matrix is a single complex number, say, $a \in \mathbb{C}$ that can be decomposed as $a = 1\,a\,1$, i.e., $T = a$ itself and $Q = 1$.

Induction assumption: The claim is true for $n = k$.

Induction step: We begin as in the proof of Theorem 2.1: Let $A \in \mathbb{C}^{(k+1)\times(k+1)}$ and note that it has at least one eigenvalue $\lambda_1 \in \mathbb{C}$ with a corresponding eigenvector $\boldsymbol{q}_1$ satisfying $\|\boldsymbol{q}_1\|_2 = 1$. Introduce $k$ auxiliary unit vectors $\boldsymbol{q}_2, \dots, \boldsymbol{q}_{k+1} \in \mathbb{C}^{k+1}$ that are mutually orthogonal as well as orthogonal to $\boldsymbol{q}_1$. We denote

$$\tilde{Q} = \big[\boldsymbol{q}_2, \dots, \boldsymbol{q}_{k+1}\big] \in \mathbb{C}^{(k+1)\times k} \qquad \text{and} \qquad Q = \big[\boldsymbol{q}_1, \tilde{Q}\big] \in \mathbb{C}^{(k+1)\times(k+1)},$$

and note that $Q$ is unitary. Exactly as in the proof of Theorem 2.1, we get

$$Q^*AQ = \begin{bmatrix} \boldsymbol{q}_1^*A\boldsymbol{q}_1 & \boldsymbol{q}_1^*A\tilde{Q} \\ \tilde{Q}^*A\boldsymbol{q}_1 & \tilde{Q}^*A\tilde{Q} \end{bmatrix} = \begin{bmatrix} \lambda_1\boldsymbol{q}_1^*\boldsymbol{q}_1 & \boldsymbol{q}_1^*A\tilde{Q} \\ \lambda_1\tilde{Q}^*\boldsymbol{q}_1 & \tilde{Q}^*A\tilde{Q} \end{bmatrix} = \begin{bmatrix} \lambda_1 & \boldsymbol{q}_1^*A\tilde{Q} \\ 0 & \tilde{Q}^*A\tilde{Q} \end{bmatrix},$$

where we used the fact that $\boldsymbol{q}_1$ is a unit vector and orthogonal to the columns of $\tilde{Q}$ by construction (, but we could not get rid of the top right block as $A$ is not assumed to be Hermitian).

By the induction assumption, for $\tilde{Q}^*A\tilde{Q} \in \mathbb{C}^{k\times k}$ there exists an upper triangular $T_1 \in \mathbb{C}^{k\times k}$ and a unitary $Q_1 \in \mathbb{C}^{k\times k}$ such that

$$\tilde{Q}^*A\tilde{Q} = Q_1T_1Q_1^*.$$

Hence,

$$Q^*AQ = \begin{bmatrix} \lambda_1 & \boldsymbol{q}_1^*A\tilde{Q} \\ 0 & Q_1T_1Q_1^* \end{bmatrix} = \begin{bmatrix} \lambda_1 & \boldsymbol{q}_1^*A\tilde{Q}(Q_1Q_1^*) \\ 0 & Q_1T_1Q_1^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix} \begin{bmatrix} \lambda_1 & \boldsymbol{q}_1^*A\tilde{Q}Q_1 \\ 0 & T_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix}$$

and by the unitarity of $Q$,

(51)
$$A = Q\begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix} \begin{bmatrix} \lambda_1 & \boldsymbol{q}_1^*A\tilde{Q}Q_1 \\ 0 & T_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q_1^* \end{bmatrix} Q^* = \left(Q\begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix}\right) \begin{bmatrix} \lambda_1 & \boldsymbol{q}_1^*A\tilde{Q}Q_1 \\ 0 & T_1 \end{bmatrix} \left(Q\begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix}\right)^*.$$

It is easy check that (51) is a decomposition of the required form (50). Indeed, as $T_1 \in \mathbb{C}^{k\times k}$ is upper triangular, so is the midmost matrix in (51). Moreover,

$$\left(Q\begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix}\right)^* \left(Q\begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix}\right) = I,$$

which corresponds exactly to the last formula in the proof of Theorem 2.1. □

Theorem 3.1 states that any square matrix $A$ is similar to an upper triangular $T$. Since $p_A(\lambda) = p_T(\lambda) = \det(T - \lambda I)$ by Lemma 3.1 and the determinant of the upper triangular matrix $T - \lambda I$ is the product of its diagonal elements, the eigenvalues of $A$ are on the diagonal of $T$ repeated according to their algebraic multiplicities (cf. (42)).

To complete this section, we introduce the *Jordan normal form* that indicates any square matrix $A \in \mathbb{C}^{n\times n}$ is similar to an 'almost diagonal' matrix, i.e., to a *Jordan matrix* that is an upper triangular, block diagonal matrix of the form

(52)
$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{n_l}(\lambda_l) \end{bmatrix} \in \mathbb{C}^{n\times n}.$$

Here $J_{n_j}(\lambda_j) \in \mathbb{C}^{n_j\times n_j}$ is a *Jordan block* related to an eigenvalue $\lambda_j$,

$$J_{n_j}(\lambda_j) = \begin{bmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_j & 1 \\ & & & & \lambda_j \end{bmatrix}.$$

Take note that $n_1 + n_2 + \cdots + n_l = n$ and $J$ has nonzero elements only on its main diagonal and on its first superdiagonal.

Each eigenvalue is repeated according to its geometric multiplicity in $J$ of (52). On the other hand, by summing the sizes $n_j$ of the Jordan blocks corresponding to a particular eigenvalue, one gets the algebraic multiplicity of that eigenvalue. In consequence, if the algebraic and the geometric multiplicity are the same for each eigenvalue, all Jordan blocks in (52) are of the size $1 \times 1$ and $J$ is, in fact, a diagonal matrix carrying the eigenvalues repeated according to their algebraic/geometric multiplicities. To be slightly more precise, the total number of nonzero elements, i.e. ones, on the first superdiagonal of $J$ is

$$\sum_{j=1}^{k} \big(\mu_A(\lambda_j) - \mu_G(\lambda_j)\big),$$

where $\lambda_1, \dots, \lambda_k$ are the distinct eigenvalues of $A$.

The following, slightly vague theorem ends this section.

Theorem 3.2. *Any $A \in \mathbb{C}^{n\times n}$ has a decomposition*

$$A = PJP^{-1} \tag{53}$$

*where $J \in \mathbb{C}^{n\times n}$ is as in (52) and the invertible matrix $P \in \mathbb{C}^{n\times n}$ has eigenvectors and so-called* generalized eigenvectors *of $A$ as its columns.*

## 4. Matrix exponential

Similarity transformations can be used to calculate matrix valued functions such as the square root or the exponential of a matrix. In this section, we concentrate on the matrix exponential and its application to qualitative analysis of systems of ordinary differential equations.

For a scalar argument $t \in \mathbb{R}$, the exponential function can be defined, e.g., via its power series expansion:

$$\mathrm{e}^t = \sum_{j=0}^{\infty} \frac{t^j}{j!}. \tag{54}$$

According to basic calculus courses, the series in (54) converges for any $t \in \mathbb{R}$ (and the convergence is, in fact, uniform on any bounded subset of $\mathbb{R}$). In particular, it can be shown that one is allowed to differentiate the series termwise:

$$\frac{d}{dt}\mathrm{e}^t = \sum_{j=0}^{\infty} \frac{d}{dt}\frac{t^j}{j!} = \sum_{j=1}^{\infty} j\frac{t^{(j-1)}}{j!} = \sum_{j=1}^{\infty} \frac{t^{(j-1)}}{(j-1)!} = \sum_{j=0}^{\infty} \frac{t^j}{j} = \mathrm{e}^t.$$

In exactly the same manner, one also obtains that

$$\frac{d}{dt}\mathrm{e}^{at} = a\mathrm{e}^{at}$$

for any constant $a \in \mathbb{C}$. As a consequence, the solution[2] to the initial value problem

$$x'(t) = ax(t) \quad \text{for } t > 0, \qquad x(0) = x_0 \in \mathbb{C}, \tag{55}$$

is obviously $x(t) = x_0\mathrm{e}^{at} = \mathrm{e}^{at}x_0$. An interesting and practically relevant question is how this argumentation can be generalized if the scalar-valued function $x : \mathbb{R}_+ \to \mathbb{C}$ is replaced in (55) by a vector-valued version $\boldsymbol{x} : \mathbb{R}_+ \to \mathbb{C}^n$ and the scalar coefficient $a \in \mathbb{C}$ is replaced by a matrix $A \in \mathbb{C}^{n\times n}$.

To begin with, notice that we can *formally* define $\mathrm{e}^{tA}$ even if $A \in \mathbb{C}^{n\times n}$:

$$\mathrm{e}^{tA} := \sum_{j=0}^{\infty} \frac{(tA)^j}{j!} = \sum_{j=0}^{\infty} \frac{t^j A^j}{j!}, \tag{56}$$

where $A^0$ is defined to be the identity matrix $I \in \mathbb{R}^{n\times n}$. The right-hand side of (56) makes algebraic sense since $A$ is a square matrix. In particular, it seems that $\mathrm{e}^{tA}$ is also a $n \times n$ matrix for any $t \in \mathbb{R}$. However, it is not quite obvious that the series on the right-hand side of (56) converges (in any reasonable sense), and so it is also not obvious if (56) can be taken as the definition of the *matrix exponential* function. To this end, let us define

$$S_N(t) := \sum_{j=0}^{N} \frac{t^j A^j}{j!}$$

[2] In fact, we do not prove on this course that (55) and its vectorial counterpart (58) are uniquely solvable — although they are.

to be the $N$th partial sum in the infinite series on the right-hand side of (56). For example on the course *Euclidean spaces*, it is proven that in any finite-dimensional normed vector space a sequence converges if and only if it is a Cauchy sequence. In our case, this means that $\lim_{N\to\infty} S_N(t)$ exists,[3] i.e., (56) makes sense, if and only if

$$\|S_N(t) - S_M(t)\| \to 0 \qquad \text{as } N, M \to 0 \tag{57}$$

for some matrix norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$.[4]

Let us then prove (57) for an arbitrary *operator norm*; recall that operator norms form a subclass of matrix norms. Without loss of generality we may assume that $N > M$ (the case $M = N$ is trivial and if $M > N$, we can swap the names of the indices). By repetitively applying the triangle inequality and other basic properties of operator norms, we obtain

$$\begin{aligned}\|S_N(t) - S_M(t)\| &= \Big\| \sum_{j=M+1}^{N} \frac{t^j A^j}{j!} \Big\| \leq \sum_{j=M+1}^{N} \Big\| \frac{t^j A^j}{j!} \Big\| \leq \sum_{j=M+1}^{N} \frac{(|t|\|A\|)^j}{j!} \\ &= \sum_{k=0}^{N-(M+1)} \frac{(|t|\|A\|)^{k+(M+1)}}{(k+(M+1))!} \leq \frac{(|t|\|A\|)^{M+1}}{(M+1)!} \sum_{k=0}^{N-(M+1)} \frac{(|t|\|A\|)^k}{k!} \\ &\leq \frac{(|t|\|A\|)^{M+1}}{(M+1)!} \sum_{k=0}^{\infty} \frac{(|t|\|A\|)^k}{k!} = \frac{(|t|\|A\|)^{M+1}}{(M+1)!} \mathrm{e}^{|t|\|A\|},\end{aligned}$$

which converges to zero as $M$ (and $N$) tend to infinity since the factorial in the denominator grows considerably faster than the power in the numerator. Hence, $S_N$ is a Cauchy sequence, and thus it converges for any (fixed) $t \in \mathbb{R}$ and $A \in \mathbb{C}^{n\times n}$. In particular, it makes sense to use (56) as the definition of the matrix exponential function.

Let us then consider the vectorial counterpart of (55), that is, a system of ordinary differential equations of the form

$$\boldsymbol{x}'(t) = A\boldsymbol{x}(t) \quad \text{for } t > 0, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{58}$$

where $A \in \mathbb{C}^{n\times n}$ is the coefficient matrix, $\boldsymbol{x} : \mathbb{R}_+ \to \mathbb{C}^n$ is the solution, and $\boldsymbol{x}_0 \in \mathbb{C}^n$ carries the initial values. Analogously to the case of a single differential equation in (55), the solution to this problems is

$$\boldsymbol{x}(t) = e^{tA}\boldsymbol{x}_0$$

as we will demonstrate next. First of all, the initial condition is satisfied:

$$\mathrm{e}^{0A}\boldsymbol{x}_0 = \boldsymbol{I}\boldsymbol{x}_0 = \boldsymbol{x}_0$$

because all but the first term in the series on the right-hand side of (56) vanish when $t = 0$.[5] As in the scalar-valued case, it can moreover be argued that the series on the right-hand side of (56) can be differentiated term by term, which yields

$$\begin{aligned}\frac{d}{dt}\big(\mathrm{e}^{tA}\boldsymbol{x}_0\big) &= \Big(\sum_{j=0}^{\infty} \frac{d}{dt}\frac{t^j A^j}{j!}\Big)\boldsymbol{x}_0 = \Big(\sum_{j=1}^{\infty} j\frac{t^{(j-1)}A^j}{j!}\Big)\boldsymbol{x}_0 = \Big(\sum_{j=1}^{\infty} A\frac{t^{(j-1)}A^{(j-1)}}{(j-1)!}\Big)\boldsymbol{x}_0 \\ &= \Big(A\sum_{j=0}^{\infty} \frac{t^j A^j}{j}\Big)\boldsymbol{x}_0 = A\big(\mathrm{e}^{tA}\boldsymbol{x}_0\big)\end{aligned}$$

as desired.

[3] Observe that the space of $n \times n$ matrices, i.e. $\mathbb{C}^{n\times n}$, is a vector space (consider scalar multiplication and addition of matrices) and it is also finite-dimensional (any $n^2$ matrices that have only one nonzero element at mutually different locations form a basis for $\mathbb{C}^{n\times n}$).

[4] In fact, it does not matter which matrix norm is used since all norms on a finite-dimensional vector space define the same topology (cf. the third exercise sheet).

[5] It is agreed that $(0\,A)^0 = I$, which can be reasoned by continuity: if $t$ tends to zero from either side, then $(tA)^0 = t^0A^0 = I$ converges to the identity matrix. (This same convention is actually used already in the scalar-valued definition (54): $\mathrm{e}^0 = 0^0 := 1$.)

Similarity transformations can be utilized to write matrix exponentials in more accessible forms; when analyzing this, we drop the 'time variable' to simplify the notation and because the matrix exponential $\mathrm{e}^A$ also has other applications than systems of ordinary differential equations. For a diagonalizable matrix one simply exploits its similarity with a diagonal matrix, and for the other matrices the standard help is the Jordan normal form. Both of these approaches are theoretical tools that are usually not implemented numerically in practice; there are other numerical algorithms for computing the matrix exponential of a given square matrix.

Let us start with a diagonalizable $A \in \mathbb{C}^{n\times n}$ such that

$$A = X\Lambda X^{-1},$$

where $X \in \mathbb{C}^{n\times n}$ is invertible and the diagonal matrix $\Lambda \in \mathbb{C}^{n\times n}$ has the eigenvalues of $A$, i.e. $\lambda_1, \dots, \lambda_n$, on its diagonal (repeated according to their algebraic multiplicity). Clearly,

$$A^j = X\Lambda^j X^{-1} = X \begin{bmatrix} \lambda_1^j & & & \\ & \lambda_2^j & & \\ & & \ddots & \\ & & & \lambda_n^j \end{bmatrix} X^{-1}.$$

It easily follows that

$$\mathrm{e}^A = \sum_{j=0}^{\infty} \frac{A^j}{j!} = X\Big(\sum_{j=0}^{\infty} \frac{\Lambda^j}{j!}\Big)X^{-1} = X\mathrm{e}^{\Lambda}X^{-1}, \tag{59}$$

which further reduces to

$$\mathrm{e}^A = X \begin{bmatrix} \sum_{j=0}^{\infty} \frac{\lambda_1^j}{j!} & & & \\ & \sum_{j=0}^{\infty} \frac{\lambda_2^j}{j!} & & \\ & & \ddots & \\ & & & \sum_{j=0}^{\infty} \frac{\lambda_n^j}{j!} \end{bmatrix} X^{-1} = X \begin{bmatrix} e^{\lambda_1} & & & \\ & e^{\lambda_2} & & \\ & & \ddots & \\ & & & e^{\lambda_n} \end{bmatrix} X^{-1}.$$

This means that calculating the matrix exponential of a diagonalizable $A$ essentially amounts to exponentiating its eigenvalues (as well as computing its eigenvalues and eigenbasis to begin with).

Example 4.1. *The matrix*

$$A := \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}^{-1}$$

*is obviously diagonalizable (as (i) we have already diagonalized it and (ii) it is Hermitian/symmetric). According to the above considerations, the matrix exponential of $A$ can be written as*

$$e^A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} e^{-1} & 0 \\ 0 & e^3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}^{-1}.$$

The exponential of a non-diagonalizable $A \in \mathbb{C}^{n\times n}$ can be transformed into a more tractable form by resorting to the Jordan normal form (53): If $A$, $P$ and $J$ are as in Theorem 3.2, then

$$\mathrm{e}^A = P\mathrm{e}^J P^{-1},$$

which can be deduced in exactly the same way as (59). Furthermore, it can be argued rather straightforwardly that

$$\mathrm{e}^J = \begin{bmatrix} \mathrm{e}^{J_{n_1}(\lambda_1)} & & & \\ & \mathrm{e}^{J_{n_2}(\lambda_2)} & & \\ & & \ddots & \\ & & & \mathrm{e}^{J_{n_l}(\lambda_l)}. \end{bmatrix},$$

where the Jordan blocks $J_{n_j}(\lambda_j)$, $j = 1, \dots, l$, are as in (52). The exponentials of the Jordan blocks, i.e. $\mathrm{e}^{J_{n_j}(\lambda_j)} \in \mathbb{C}^{n_j\times n_j}$, $j = 1, \dots, l$, can still be written down explicitly with the help of the eigenvalues of $A$ and factorials, but we save our strength by not doing this here.

**4.1. Qualitative analysis of differential equations.** The matrix exponential is typically not numerically computed, but it can, e.g., be used to obtain qualitative understanding about the behavior of a solution to a system of differential equations close to so-called *equilibrium points*. Consider the rather general (autonomous) system of ordinary differential equations

$$\boldsymbol{x}'(t) = \boldsymbol{F}(\boldsymbol{x}(t)) \quad \text{for } t > 0, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0 \in \mathbb{R}^n, \tag{60}$$

where $\boldsymbol{x} : \mathbb{R}_+ \to \mathbb{R}^n$ is the solution and $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ is a nice enough function (e.g., twice continuously differentiable). Observe, in particular, that for a real-valued $A$ the system (58) is a special case of (60). An equilibrium point $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ of (60) is such that

$$\boldsymbol{F}(\tilde{\boldsymbol{x}}) = 0.$$

If a solution to (60) ends up at a certain time $\tilde{t} \geq 0$ to an equilibrium point, i.e. $\boldsymbol{x}(\tilde{t}) = \tilde{\boldsymbol{x}}$, it will stay there forever, i.e. $\boldsymbol{x}(t) = \tilde{\boldsymbol{x}}$ for all $t \geq \tilde{t}$, as the right-hand side of (60), and thus also the derivative of the solution in question, vanishes at the equilibrium.

Close to an equilibrium point $\tilde{\boldsymbol{x}}$, the right hand side of (60) can be linearized as

$$\boldsymbol{F}(\boldsymbol{x}) \approx \boldsymbol{F}(\tilde{\boldsymbol{x}}) + D\boldsymbol{F}(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \tilde{\boldsymbol{x}}) = D\boldsymbol{F}(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \tilde{\boldsymbol{x}}),$$

where $D\boldsymbol{F}(\tilde{\boldsymbol{x}}) \in \mathbb{R}^{n\times n}$ is the Jacobian matrix of $\boldsymbol{F}$ evaluated at $\tilde{\boldsymbol{x}}$. By plugging this approximation in (60), one obtains that

$$\boldsymbol{x}'(t) \approx D\boldsymbol{F}(\tilde{\boldsymbol{x}})(\boldsymbol{x}(t) - \tilde{\boldsymbol{x}}) \qquad \Longleftrightarrow \qquad \frac{d}{dt}(\boldsymbol{x}(t) - \tilde{\boldsymbol{x}}) \approx D\boldsymbol{F}(\tilde{\boldsymbol{x}})(\boldsymbol{x}(t) - \tilde{\boldsymbol{x}}),$$

or

$$\boldsymbol{y}'(t) \approx D\boldsymbol{F}(\tilde{\boldsymbol{x}})\boldsymbol{y} \tag{61}$$

after introducing the spatially shifted solution $\boldsymbol{y}(t) := \boldsymbol{x}(t) - \tilde{\boldsymbol{x}}$. The approximation in (61) is good if one only studies the solution of (60) close to the equilibrium point $\tilde{\boldsymbol{x}}$, that is, solutions of (61) close to the origin. In this case, a solution $\boldsymbol{y}(t)$ to (61) is essentially a matrix exponential function defined by the Jacobian matrix $D\boldsymbol{F}(\tilde{x})$; cf. (58) and its solution formula with $A = D\boldsymbol{F}(\tilde{x})$. Thus, the eigenvalues of the Jacobian matrix determine the local behavior of the system (60) close to the equilibrium point.

EXAMPLE 4.2. *As an example, consider a Lotka–Volterra model*

$$\boldsymbol{x}'(t) = \boldsymbol{F}(\boldsymbol{x}(t)) := \begin{bmatrix} \frac{2}{3}x_1(t) - \frac{4}{3}x_1(t)x_2(t) \\ x_1(t)x_2(t) - x_2(t) \end{bmatrix} \quad \text{for } t > 0, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{62}$$

*This model describes the interaction between predator and prey populations (with certain parameter choices). The unknowns $x_1$ and $x_2$ are the number of predators and the number of preys, respectively. A rough understanding about the behavior of the solutions to the Lotka–Volterra system can be obtained by visualizing the vector field $\boldsymbol{F}$ defining the right-hand side of (62): $\boldsymbol{F}(\boldsymbol{z})$ gives a tangent to the trajectory of any solution $\boldsymbol{x}(t)$ of (62) passing through $\boldsymbol{z} \in \mathbb{R}^2$; see Figure 1.*

*Our interest lies with the behavior of the system close to equilibrium points $\tilde{\boldsymbol{x}} \in \mathbb{R}^2$, where $\boldsymbol{F}(\tilde{\boldsymbol{x}}) = 0$. The studied Lotka–Volterra model has two equilibrium points, namely $\tilde{\boldsymbol{x}}_1 = [1, 0.5]^T$ and $\tilde{\boldsymbol{x}}_2 = [0, 0]^T$. Around these points, the behavior of the system can be approximated by the corresponding linearized systems, that is,*

$$\boldsymbol{y}_j'(t) = D\boldsymbol{F}(\tilde{\boldsymbol{x}}_j)\boldsymbol{y}_j(t), \qquad j = 1, 2,$$

*where $\boldsymbol{y}_j(t) = \boldsymbol{x}(t) - \tilde{\boldsymbol{x}}_j$, $j = 1, 2$. For (62),*

$$D\boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} \frac{2}{3} - \frac{4}{3}x_2 & -\frac{4}{3}x_1 \\ x_2 & x_1 - 1 \end{bmatrix},$$

*and so*

$$D\boldsymbol{F}(\tilde{\boldsymbol{x}}_1) = \begin{bmatrix} 0 & -\frac{4}{3} \\ \frac{1}{2} & 0 \end{bmatrix} \qquad \text{and} \qquad D\boldsymbol{F}(\tilde{\boldsymbol{x}}_2) = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & -1 \end{bmatrix}.$$

*Both linearized systems can now be exactly solved and the behavior of their solutions studied using the matrix exponential.*

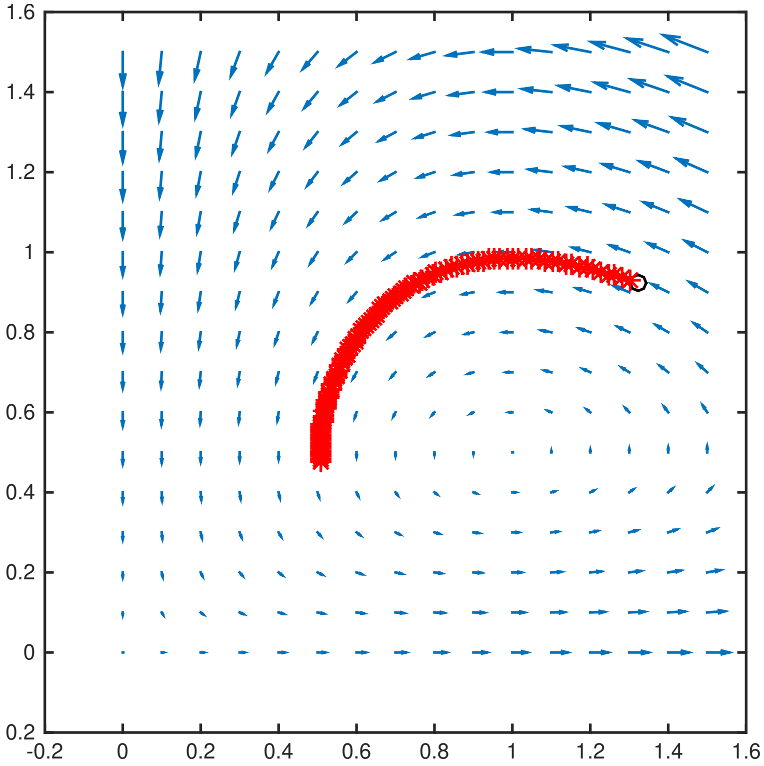

FIGURE 1. The arrows visualize the vector field $\boldsymbol{F} : \mathbb{R}^2 \to \mathbb{R}^2$ for the Lotka–Volterra model of Example 4.2. A trajectory starting from the black circle is visualized by red stars. The system has two equilibrium points, $(1, 0.5)$ and $(0, 0)$.

CHAPTER 3

# Least squares problems

The last theme of these lecture notes are least squares problems that arise, e.g., from the need to fit the parameters of a linear mathematical model to given measurements. To minimize the effect of measurement errors on the estimated parameters, the measurement is typically repeated several times. This leads to an overdetermined problem that includes more observations than model parameters. (In this chapter, we return to the setting where the considered vectors and matrices are real-valued.)

## 1. Motivation, definition and equivalence to normal equation

We start with two examples:

EXAMPLE 1.1. *A simple example of a least squares problem is the determination of a spring constant from several measurements. Assume that we have at our disposal $n$ weights $m_i$, $i = 1, \dots, n$, and can measure the elongation $x_i$ of the spring when the weight $m_i$ is attached to it. In addition, assume that the gravity $g$ is known and we are dealing with an* ideal spring, *meaning that the elongation of the spring under the mass $m_i$ behaves as*

$$x_i = kgm_i, \qquad i = 1, \dots, n,$$

*where $k \in \mathbb{R}$ is the spring constant. By repeating the measurement with all available weights, one obtains a (trivial) linear system for $k$:*

$$\begin{bmatrix} gm_1 \\ gm_2 \\ \vdots \\ gm_n \end{bmatrix} k = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

*Due to measurement errors, this system does not typically have a solution. Instead, one determines (an estimate for) the parameter $k$ by minimizing the 2-norm of the residual: Let*

$$A = \begin{bmatrix} gm_1 \\ gm_2 \\ \vdots \\ gm_n \end{bmatrix} \in \mathbb{R}^{n \times 1} \qquad \text{and} \qquad \boldsymbol{b} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

*and seek $k$ as a solution to*

$$\min_{k \in \mathbb{R}} \|Ak - \boldsymbol{b}\|_2^2.$$

*It follows easily that*

$$\|Ak - \boldsymbol{b}\|_2^2 = (Ak - \boldsymbol{b})^T (Ak - \boldsymbol{b}) = (A^T A)k^2 - 2(A^T b)k + \|\boldsymbol{b}\|_2^2,$$

*which is an equation for an upwards opening parabola in the variable $k$. The spring constant can thus be solved by computing the zero of the derivative with respect to $k$, which gives $k = (A^T A)^{-1} A^T b$.*

EXAMPLE 1.2. *As a second example, let us consider finding an ellipsoid that is the best fit to a given set of points $\boldsymbol{x}^{(i)} \in \mathbb{R}^2$, $i = 1, \dots, n$. There are several ways to define what "best" means and to compute the fit, but for simplicity we have opted to follow an approach that is based on the*

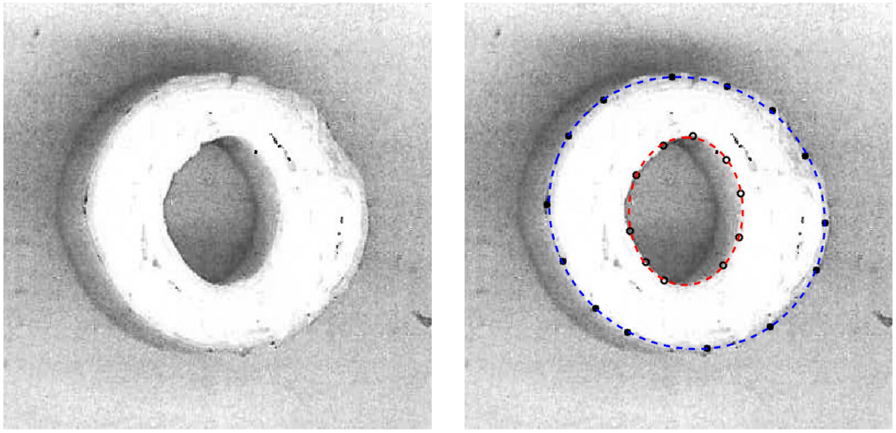

FIGURE 1. A small cylindrical object and two ellipses fitted to its cross section.

*so-called conic section presentation of an ellipse. In this presentation, an ellipse is defined as a set*

$$\big\{(x_1,x_2)\in\mathbb{R}^2 \mid x_1^2+Ax_2^2+Bx_1x_2+Cx_1+Dx_2+E=0, \quad A,B,C,D,E\in\mathbb{R},\ B^2-4A<0\big\}, \tag{63}$$

*which makes fitting an ellipse to a set of points relatively easy, but some work would be needed to find out the semi-axes and center-point of the ellipse.*

*To be more precise, the task is to define the coefficients $A,B,C,D$ and $E$ so that the data points $\boldsymbol{x}^{(i)}\in\mathbb{R}^2$, $i=1,\dots,n$, satisfy the equation defining the set (63) 'as well as possible'. The simplest way to formulate such an optimality condition is to look for $A,B,C,D$ and $E$ that minimize the sum of the squared discrepancies*

$$\sum_{i=1}^{n}\Big(\big(x_1^{(i)}\big)^2+A\big(x_2^{(i)}\big)^2+Bx_1^{(i)}x_2^{(i)}+Cx_1^{(i)}+Dx_2^{(i)}+E\Big)^2. \tag{64}$$

*Let us collect the unknown coefficients $A,B,C,D$ and $E$ into a single vector $\boldsymbol{\alpha}=[A,B,C,D,E]^T\in\mathbb{R}^5$ and set*

$$A=\begin{bmatrix}\big(x_2^{(1)}\big)^2 & x_1^{(1)}x_2^{(1)} & x_1^{(1)} & x_2^{(1)} & 1\\ \big(x_2^{(2)}\big)^2 & x_1^{(2)}x_2^{(2)} & x_1^{(2)} & x_2^{(2)} & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \big(x_2^{(n)}\big)^2 & x_1^{(n)}x_2^{(n)} & x_1^{(n)} & x_2^{(n)} & 1\end{bmatrix} \quad\text{and}\quad \boldsymbol{b}=-\begin{bmatrix}\big(x_1^{(1)}\big)^2\\ \big(x_1^{(2)}\big)^2\\ \vdots\\ \big(x_1^{(n)}\big)^2\end{bmatrix}.$$

*With this notation, the minimization of (64) can be rewritten as follows: find the minimizer $\boldsymbol{\alpha}\in\mathbb{R}^5$ of*

$$\|A\boldsymbol{\alpha}-\boldsymbol{b}\|_2^2,$$

*which is of the general least squares form that we will study in more detail in what follows.*

Motivated by the preceding two examples, we define a *least squares* (LSQ) solution of our original linear system (1) as follows:

DEFINITION 1.1 (LSQ problem). *Let $A\in\mathbb{R}^{m\times n}$ and $\boldsymbol{b}\in\mathbb{R}^m$. A vector $\boldsymbol{x}\in\mathbb{R}^n$ is called a* least squares solution *of the equation (1) if it minimizes the squared discrepancy*

$$\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 \tag{65}$$

*measured in the 2-norm.*

It turns out that minimizing (65) is equivalent to solving the so-called *normal equation*,

$$A^TA\boldsymbol{x}=A^T\boldsymbol{b}. \tag{66}$$

This is not very surprising since (66) characterizes the points where the gradient of the squared sum (65) vanishes (homework). If $N(A) = \{0\}$, then also $N(A^TA) = \{0\}$[1], meaning that $A^TA$ is invertible and (66) has a unique solution. However, it is not quite obvious that (66) always has at least one solution; will prove such an existence result later in this chapter.

Theorem 1.1. *Assume that* (66) *has at least one solution (as it always has!). Then* $\boldsymbol{x} \in \mathbb{R}^n$ *is a minimizer of* (65) *if and only if it solves* (66).

Proof. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a solution to (66) and write an arbitrary $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ in the form $\tilde{\boldsymbol{x}} = \boldsymbol{x}+\boldsymbol{z}$, i.e., set $\boldsymbol{z} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$. Then, it holds that

$$\begin{aligned}
\|A\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 = \|A(\boldsymbol{x}+\boldsymbol{z}) - \boldsymbol{b}\|_2^2 &= \|(A\boldsymbol{x} - \boldsymbol{b}) + A\boldsymbol{z}\|_2^2 = \big((A\boldsymbol{x} - \boldsymbol{b}) + A\boldsymbol{z}\big)^T\big((A\boldsymbol{x} - \boldsymbol{b}) + A\boldsymbol{z}\big) \\
&= \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + (A\boldsymbol{x} - \boldsymbol{b})^T A\boldsymbol{z} + (A\boldsymbol{z})^T(A\boldsymbol{x} - \boldsymbol{b}) + \|A\boldsymbol{z}\|_2^2 \\
&= \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + 2\boldsymbol{z}^T(A^TA\boldsymbol{x} - A^T\boldsymbol{b}) + \|A\boldsymbol{z}\|_2^2 \\
&= \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \|A\boldsymbol{z}\|_2^2 \geq \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2, \qquad (67)
\end{aligned}$$

where the penultimate step follows from $\boldsymbol{x}$ being a solution of the normal equation. In other words, $\boldsymbol{x}$ is a solution of the least squares problem because $\|A\boldsymbol{x}-\boldsymbol{b}\|_2 \leq \|A\tilde{\boldsymbol{x}}-\boldsymbol{b}\|_2$ for an arbitrary $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$.

Let $\tilde{\boldsymbol{x}}$ now also be a least squares solution, which means that equality holds in (67). This implies that $\boldsymbol{z} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$ belongs to $N(A)$. Hence,

$$A^TA\tilde{\boldsymbol{x}} = A^TA(\boldsymbol{x}+\boldsymbol{z}) = A^TA\boldsymbol{x} + A^TA\boldsymbol{z} = A^T\boldsymbol{b},$$

i.e., $\tilde{\boldsymbol{x}}$ is also a solution to the normal equation (66). □

The next goal is to enhance our geometric understanding about the normal equation (66): What is the intuitive reason for it being equivalent to the least squares problem (65)? To this end, we need to introduce the concept of a projection.

## 2. Projection matrices

Projection matrices are related to decompositions of a given vector $\boldsymbol{x} \in \mathbb{R}^n$ into two parts. To this end, let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathbb{R}^n$ and assume that

$$\mathcal{V} \cap \mathcal{W} = \{0\} \qquad \text{and} \qquad \mathcal{V} + \mathcal{W} = \mathbb{R}^n, \qquad (68)$$

where

$$\mathcal{V} + \mathcal{W} = \{\boldsymbol{v} + \boldsymbol{w} \mid \boldsymbol{v} \in \mathcal{V} \text{ and } \boldsymbol{w} \in \mathcal{W}\}.$$

This means that $\mathbb{R}^n$ is the *direct sum* of $\mathcal{V}$ and $\mathcal{W}$, which is expressed as $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{W}$.

Lemma 2.1. *If* $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{W}$, *i.e.* (68) *is valid, then any* $\boldsymbol{x}$ *can be decomposed* uniquely *as*

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{w},$$

*where* $\boldsymbol{v} \in \mathcal{V}$ *and* $\boldsymbol{w} \in \mathcal{W}$.

Proof. Due to (68), it is obvious that any $\boldsymbol{x} \in \mathbb{R}^n$ has the required decomposition, and so we only need to worry about its uniqueness. Let $\boldsymbol{v}, \tilde{\boldsymbol{v}} \in \mathcal{V}$ and $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \mathcal{W}$ be such that

$$\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{x} = \tilde{\boldsymbol{v}} + \tilde{\boldsymbol{w}}.$$

Hence,

$$\mathcal{V} \ni \boldsymbol{v} - \tilde{\boldsymbol{v}} = \tilde{\boldsymbol{w}} - \boldsymbol{w} \in \mathcal{W},$$

and by the first condition in (68), it must hold that

$$\boldsymbol{v} - \tilde{\boldsymbol{v}} = \tilde{\boldsymbol{w}} - \boldsymbol{w} = 0,$$

as the zero vector is the only common element of $\mathcal{V}$ and $\mathcal{W}$. In other words, $\boldsymbol{v} = \tilde{\boldsymbol{v}}$ and $\boldsymbol{w} = \tilde{\boldsymbol{w}}$, and thus the decomposition must be unique. □

[1] Obviously $N(A) \subset N(A^TA)$, but also $N(A^TA) \subset N(A)$: if $\boldsymbol{z} \in N(A^TA)$, then $0 = \boldsymbol{z}^T(A^TA\boldsymbol{z}) = \|A\boldsymbol{z}\|_2^2$, i.e., $\boldsymbol{z} \in N(A)$.

The projections, or projection matrices $P_{\mathcal{V}}, P_{\mathcal{W}} \in \mathbb{R}^{n\times n}$ associated to a decomposition $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{W}$ are defined via

$$P_{\mathcal{V}}\boldsymbol{x} = \boldsymbol{v} \qquad \text{and} \qquad P_{\mathcal{W}}\boldsymbol{x} = \boldsymbol{w},$$

where $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{w} \in \mathcal{W}$ are the unique vectors such that $\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{w}$. It would be straightforward to prove that such matrices exist, i.e., that the functions sending $\boldsymbol{x}$ to its unique components in $\mathcal{V}$ and $\mathcal{W}$ are linear mappings. However, we take a more constructive path and find explicit representations for $P_{\mathcal{V}}$ and $P_{\mathcal{W}}$.

Let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ and $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_l\}$ be basis for $\mathcal{V}$ and $\mathcal{W}$, that is, these sets of vectors are linearly independent and

$$\mathcal{V} = \operatorname{span}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\} \qquad \text{and} \qquad \mathcal{W} = \operatorname{span}\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_l\}.$$

According to (68), any vector in $\mathbb{R}^n$ can be presented as a joint linear combination of $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ and $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_l\}$. Let us prove that the joint set of vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, \boldsymbol{w}_1, \dots, \boldsymbol{w}_l\}$ is also linearly independent: Let $\boldsymbol{\alpha} \in \mathbb{R}^{k+l}$ be such that

$$\sum_{i=1}^{k} \alpha_i \boldsymbol{v}_i + \sum_{j=1}^{l} \alpha_{k+j} \boldsymbol{w}_j = 0 \qquad \Longleftrightarrow \qquad \mathcal{V} \ni \sum_{i=1}^{k} \alpha_i \boldsymbol{v}_i = -\sum_{j=1}^{l} \alpha_{l+k} \boldsymbol{w}_l \in \mathcal{W},$$

where both sides must vanish by virtue of (68). Because $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ and $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_l\}$ are separately linearly independent, this means that the whole vector $\boldsymbol{\alpha} \in \mathbb{R}^{k+l}$ must vanish. Hence, $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k, \boldsymbol{w}_1, \dots, \boldsymbol{w}_l\}$ is a set of linearly independent vectors that span $\mathbb{R}^n$, i.e., they form a basis for $\mathbb{R}^n$. In particular, their total number $k + l$ must equal the dimension of $\mathbb{R}^n$; in the following, we write $l = n - k$.

We stack the two bases as columns of the matrices $W \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{n\times(n-k)}$, respectively:

$$V = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_k] \qquad \text{and} \qquad W = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_{n-k}].$$

With this notation, the two subspaces can be expressed as $\mathcal{V} = R(V)$ and $\mathcal{W} = R(W)$; see (5). The decomposition of $\mathbb{R}^n$ related to $\mathcal{V}$ and $\mathcal{W}$ is thus

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{w} = V\boldsymbol{z}_V + W\boldsymbol{z}_W = [V,\, W]\begin{bmatrix}\boldsymbol{z}_V\\ \boldsymbol{z}_W\end{bmatrix}$$

for some 'coordinate vectors' $\boldsymbol{z}_V \in \mathbb{R}^k$ and $\boldsymbol{z}_W \in \mathbb{R}^{n-k}$. Because the columns of $[W,\, V] \in \mathbb{R}^{n\times n}$ are a (linearly independent) basis for $\mathbb{R}^n$, the matrix $[W,\, V]$ is invertible. Hence,

$$\begin{bmatrix}\boldsymbol{z}_V\\ \boldsymbol{z}_W\end{bmatrix} = [V,\, W]^{-1}\boldsymbol{x}.$$

On the other hand,

$$\boldsymbol{z}_V = [I,\, 0]\begin{bmatrix}\boldsymbol{z}_V\\ \boldsymbol{z}_W\end{bmatrix} \qquad \text{and} \qquad \boldsymbol{z}_W = [0,\, I]\begin{bmatrix}\boldsymbol{z}_V\\ \boldsymbol{z}_W\end{bmatrix},$$

where $I \in \mathbb{R}^{k\times k}$, $0 \in \mathbb{R}^{k\times(n-k)}$ in the first formula and $I \in \mathbb{R}^{(n-k)\times(n-k)}$, $0 \in \mathbb{R}^{(n-k)\times k}$ in the second one. Altogether, we have thus deduced that

$$P_{\mathcal{V}}\boldsymbol{x} = V\boldsymbol{z}_V = V[I,\, 0]\begin{bmatrix}\boldsymbol{z}_V\\ \boldsymbol{z}_W\end{bmatrix} = [V,\, 0][V,\, W]^{-1}\boldsymbol{x}, \tag{69}$$

where the second 0 belongs to $\mathbb{R}^{n\times(n-k)}$. Similarly,

$$P_{\mathcal{W}}\boldsymbol{x} = W\boldsymbol{z}_W = W[0,\, I]\begin{bmatrix}\boldsymbol{z}_V\\ \boldsymbol{z}_W\end{bmatrix} = [0,\, W][V,\, W]^{-1}\boldsymbol{x}, \tag{70}$$

where the latter 0 belongs to $\mathbb{R}^{n\times k}$. In other words, $P_{\mathcal{V}} = [V,\, 0][V,\, W]^{-1}$ and $P_{\mathcal{W}} = [0,\, W][V,\, W]^{-1}$, where the zeros denote zero matrices of appropriate sizes.

By construction, the matrices $P_{\mathcal{V}}$ and $P_{\mathcal{W}}$ must satisfy

$$P_{\mathcal{V}}^2 = P_{\mathcal{V}} \qquad \text{and} \qquad P_{\mathcal{W}}^2 = P_{\mathcal{W}}.$$

Indeed, e.g., $P_{\mathcal{V}}\boldsymbol{x}$ is already an element of $\mathcal{V}$, and so its unique decomposition of the form (68) must be $P_{\mathcal{V}}\boldsymbol{x} = P_{\mathcal{V}}\boldsymbol{x} + 0$, i.e., $P_{\mathcal{V}}^2\boldsymbol{x} = P_{\mathcal{V}}(P_{\mathcal{V}}\boldsymbol{x}) = P_{\mathcal{V}}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. This could also be verified by a direct calculation based on (69). In other words, the projection matrices $P_{\mathcal{V}}$ and $P_{\mathcal{W}}$ are identity maps on their respective ranges. This is a fundamental property that can be taken as the definition of a projection (matrix).

Definition 2.1. *A matrix $P \in \mathbb{R}^{n\times n}$ is a projection if it satisfies*

$$P^2 = P. \tag{71}$$

Each projection matrix is related to a decomposition of the space $\mathbb{R}^n$ into two parts. Let $P \in \mathbb{R}^{n\times n}$ be a projection and consider the decomposition

$$\boldsymbol{x} = P\boldsymbol{x} + (I-P)\boldsymbol{x}, \tag{72}$$

that is, $\mathcal{V} = R(P)$ and $\mathcal{W} = R(I-P)$. Due to the property $P = P^2$ it is easy to see that these two subspaces satisfy the conditions (68): The second one is already proven by (72). On the other hand, if $\boldsymbol{x}$ belongs to both $R(P)$ and $R(I-P)$, it can be given as $\boldsymbol{x} = (I-P)\boldsymbol{z}$ for some $\boldsymbol{z}$, but as $P$ is a projection, it also holds that $P\boldsymbol{x} = \boldsymbol{x}$. Putting these together,

$$\boldsymbol{x} = P\boldsymbol{x} = P(I-P)\boldsymbol{z} = (P-P^2)\boldsymbol{z} = (P-P)\boldsymbol{z} = 0,$$

i.e., the zero vector is the only common element of $R(P)$ and $R(I-P)$. Note that $I-P$ is also a projection:

$$(I-P)(I-P) = I - P - P + P^2 = I - 2P + P = I - P,$$

where we once again utilized the identity $P^2 = P$.

A special projection is related to subspaces $\mathcal{V}$ and $\mathcal{W}$ that are orthogonal.

Definition 2.2. *Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathbb{R}^n$ and $\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be some inner product. The subspaces $\mathcal{V}$ and $\mathcal{W}$ are orthogonal with respect to $\langle\cdot,\cdot\rangle$ if*

$$\langle \boldsymbol{v}, \boldsymbol{w}\rangle = 0$$

*for all $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{w} \in \mathcal{W}$.*

We also need the definition for the orthogonal complement of a subspace $\mathcal{V}$.

Definition 2.3. *Let $\mathcal{V}$ be subspace of $\mathbb{R}^n$ and $\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ some innerpoduct. The orthogonal complement of $\mathcal{V}$ is the subspace*

$$\mathcal{V}^\perp := \{\boldsymbol{w} \in \mathbb{R}^n \mid \langle \boldsymbol{w}, \boldsymbol{v}\rangle = 0 \quad \text{for all } \boldsymbol{v} \in \mathcal{V}\}. \tag{73}$$

*(Proving that $\mathcal{V}^\perp$ is actually a subspace is left as an exercise.)*

It holds that

$$\mathcal{V} \cap \mathcal{V}^\perp = \{0\} \qquad \text{and} \qquad \mathcal{V} + \mathcal{V}^\perp = \mathbb{R}^n, \tag{74}$$

i.e., $\mathcal{V} \oplus \mathcal{V}^\perp = \mathbb{R}^n$. Indeed, if $\boldsymbol{v} \in \mathcal{V}$ also belongs to $\mathcal{V}^\perp$, then by definition

$$\langle \boldsymbol{v}, \boldsymbol{v}\rangle = 0, \qquad \text{i.e.,} \qquad \boldsymbol{v} = 0.$$

On the other hand, if $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ is an orthonormal basis for $\mathcal{V}$,[2] i.e., the vectors are orthogonal (cf. Definition 2.3) and of unit length in the norm defined by the considered inner product, then any $\boldsymbol{x} \in \mathbb{R}^n$ can be written as

$$\boldsymbol{x} = \sum_{i=1}^{k} \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle \boldsymbol{v}_i + \Big(\boldsymbol{x} - \sum_{i=1}^{k} \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle \boldsymbol{v}_i\Big) =: \boldsymbol{v}_{\boldsymbol{x}} + \boldsymbol{w}_{\boldsymbol{x}}, \tag{75}$$

which gives $\boldsymbol{x}$ as a sum of a vector in $\mathcal{V}$ and another in $\mathcal{V}^\perp$: Clearly, $\boldsymbol{v}_{\boldsymbol{x}} \in \mathcal{V}$ and proving that $\boldsymbol{w}_{\boldsymbol{x}} \in \mathcal{V}^\perp$ is left as a homework (write an arbitrary element of $\mathcal{V}$ in the orthonormal basis, take its inner product with $\boldsymbol{w}_{\boldsymbol{x}}$ and employ the orthonormality of the basis vectors). In fact, $\boldsymbol{v}_{\boldsymbol{x}}$ is the *orthogonal* projection of $\boldsymbol{x}$ onto $\mathcal{V}$ and $\boldsymbol{w}_{\boldsymbol{x}}$ the *orthogonal* projection of $\boldsymbol{x}$ onto $\mathcal{V}^\perp$, as will be defined in the following.

For the rest of this chapter, we only consider orthogonal complements with respect to the Euclidean inner product. A projection matrix related to an orthogonal decomposition $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{V}^\perp$ is called an *orthogonal projection* matrix. Let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ and $\{\boldsymbol{v}_1^\perp, \dots, \boldsymbol{v}_{n-k}^\perp\}$ be (not

[2] Such a basis can be constructed by the Gram–Schmidt ortogonalization process as we will learn later.

necessarily orthonormal) bases for $\mathcal{V}$ and $\mathcal{V}^\perp$, respectively, and define the matrices $V \in \mathbb{R}^{n\times k}$ and $V^\perp \in \mathbb{R}^{n\times(n-k)}$ in the standard manner, i.e.,

$$V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \qquad \text{and} \qquad V^\perp = [\boldsymbol{v}_1^\perp, \ldots, \boldsymbol{v}_{n-k}^\perp].$$

Note once again that $\mathcal{V} = R(V) = \operatorname{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ and $\mathcal{V}^\perp = R(V^\perp) = \operatorname{span}(\boldsymbol{v}_1^\perp, \ldots, \boldsymbol{v}_{n-k}^\perp)$. By definition (73), the columns of $V$ are orthogonal to those of $V^\perp$ (in the Euclidean inner product), meaning that

$$V^T V^\perp = 0. \tag{76}$$

On the other hand, the decomposition of $\mathbb{R}^n$ related to $\mathcal{V}$ and $\mathcal{V}^\perp$ is

$$\boldsymbol{x} = V\boldsymbol{z}_V + V^\perp \boldsymbol{z}_{V^\perp} \tag{77}$$

for some yet-to-be-defined $\boldsymbol{z}_V \in \mathbb{R}^k$ and $\boldsymbol{z}_{V^\perp} \in \mathbb{R}^{n-k}$. Multiplying from the left by $V^T$ and using the orthogonlity property (76), we get

$$V^T\boldsymbol{x} = V^T V \boldsymbol{z}_V.$$

Because $N(V^TV) = N(V)$ is trivial as the columns of $V \in \mathbb{R}^{n\times k}$ are linearly independent, $V^TV \in \mathbb{R}^{k\times k}$ is invertible and $\boldsymbol{z}_V$ can be solved as

$$\boldsymbol{z}_V = (V^TV)^{-1}V^T\boldsymbol{x}.$$

The projection matrix related to the decomposition $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{V}^\perp$ is thus

$$P_{\mathcal{V}} = V(V^TV)^{-1}V^T \in \mathbb{R}^{n\times n}, \tag{78}$$

see (77). In particular, the orthogonal projection does not require one to form a basis for $\mathcal{V}^\perp$: the auxiliary basis $\{\boldsymbol{v}_1^\perp, \ldots, \boldsymbol{v}_{n-k}^\perp\}$ of $V^\perp$ does not anymore appear in (78). Moreover, if $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ are orthonormal, then $V$ is orthogonal/unitary, i.e. $V^TV = I$, and so

$$P_{\mathcal{V}}\boldsymbol{x} = VI^{-1}V^T\boldsymbol{x} = V\begin{bmatrix}\boldsymbol{v}_1^T\boldsymbol{x}\\ \vdots \\ \boldsymbol{v}_k^T\boldsymbol{x}\end{bmatrix} = \sum_{i=1}^k (\boldsymbol{v}_i^T\boldsymbol{x})\boldsymbol{v}_i, \tag{79}$$

which corresponds to the first term in (75).

An othonormal projection satisfies some useful identities. First of all, since inversion and transposition commute, we have

$$P_{\mathcal{V}}^T = \big(V(V^TV)^{-1}V^T\big)^T = (V^T)^T\big((V^TV)^T\big)^{-1}V^T = V(V^TV)^{-1}V^T = P_{\mathcal{V}},$$

and hence also,

$$P_{\mathcal{V}}^T(I - P_{\mathcal{V}}) = P_{\mathcal{V}}(I - P_{\mathcal{V}}) = P_{\mathcal{V}} - P_{\mathcal{V}}^2 = P_{\mathcal{V}} - P_{\mathcal{V}} = 0. \tag{80}$$

After transposition, this yields

$$(I - P_{\mathcal{V}})^T P_{\mathcal{V}} = 0. \tag{81}$$

In fact, $I - P_{\mathcal{V}}$ is the orthogonal projection onto $\mathcal{V}^\perp$ (homework).

EXAMPLE 2.1. *Let $\boldsymbol{a}, \boldsymbol{x} \in \mathbb{R}^2$. Consider a decomposition of the vector $\boldsymbol{x}$ into two components, one parallel and the other orthogonal to $\boldsymbol{a}$. Such a decomposition can be constructed as in (75):*

$$\boldsymbol{x} = \frac{\boldsymbol{a}^T\boldsymbol{x}}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a} + \left(\boldsymbol{x} - \frac{\boldsymbol{a}^T\boldsymbol{x}}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}\right) = \frac{1}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{x} + \left(I - \frac{1}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}\boldsymbol{a}^T\right)\boldsymbol{x},$$

*where the division by $\|\boldsymbol{a}\|_2^2$ corresponds to normalization of $\boldsymbol{a}$, i.e., $\boldsymbol{a}/\|\boldsymbol{a}\|_2$ is a unit vector. The two components of this decomposition are orthogonal (see Figure 2) and the related projection matrices are*

$$P = \frac{1}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}\boldsymbol{a}^T. \tag{82}$$

*and $I - P$.*

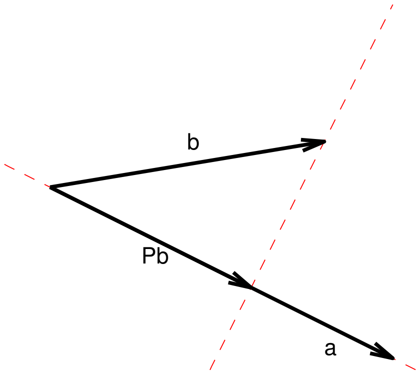

FIGURE 2. Orthogonal projection of $\boldsymbol{a}$ onto $\boldsymbol{b}$.

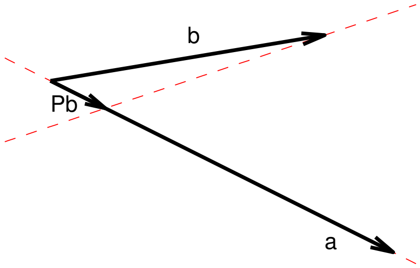

FIGURE 3. One possible oblique projection of $\boldsymbol{a}$ onto $\boldsymbol{b}$.

EXAMPLE 2.2. *Let $\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{x} \in \mathbb{R}^2$ be such that $[\boldsymbol{a}, \boldsymbol{c}] \in \mathbb{R}^{2\times 2}$ is invertible, i.e., $\boldsymbol{a}$ and $\boldsymbol{c}$ are linearly independent. Consider the decomposition of $\boldsymbol{x}$ in the basis defined by $\boldsymbol{a}$ and $\boldsymbol{c}$:*

$$\boldsymbol{x} = \alpha \boldsymbol{a} + \beta \boldsymbol{c}. \tag{83}$$

*The coefficients $\alpha$ and $\beta$ can be solved from the linear system*

$$[\boldsymbol{a}, \boldsymbol{b}] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \boldsymbol{x},$$

*and the parameter $\alpha$ can then be extracted from the solution via*

$$\alpha = [1, 0][\boldsymbol{a}, \boldsymbol{b}]^{-1}\boldsymbol{x},$$

*where $[1, 0] \in \mathbb{R}^{1\times 2}$. Thus, one of the two projection matrices related to this decomposition is*

$$P = [\boldsymbol{a}, 0][\boldsymbol{a}, \boldsymbol{b}]^{-1},$$

*where $[\boldsymbol{a}, 0] \in \mathbb{R}^{2\times 2}$ (cf. (69)).*

## 3. Geometric interpretation of a least squares solution

With the help of projection matrices, we can now understand the geometry behind the least squares problem (65) and the related normal equation (66). Let $P \in \mathbb{R}^m$ be an orthogonal projection from $\mathbb{R}^m$ to $R(A)$. As $P$ keeps elements in $R(A)$ fixed, the least squares functional can be rewritten as

$$\begin{aligned} \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 &= \|PA\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \|P(A\boldsymbol{x} - \boldsymbol{b}) - (I - P)\boldsymbol{b}\|_2^2 \\ &= \|PA\boldsymbol{x} - P\boldsymbol{b}\|^2 - \big(P(A\boldsymbol{x} - \boldsymbol{b})\big)^T (I - P)\boldsymbol{b} - \big((I - P)\boldsymbol{b}\big)^T \big(P(A\boldsymbol{x} - \boldsymbol{b})\big) + \|(I - P)\boldsymbol{b}\|_2^2 \\ &= \|A\boldsymbol{x} - P\boldsymbol{b}\| + \|(I - P)\boldsymbol{b}\|_2^2, \qquad (84) \end{aligned}$$

where the last step follows from (80) or/and (81). Because the second term on the right-hand side of (84) does not depend on $\boldsymbol{x}$, to minimize the least squares functional, $\boldsymbol{x}$ should be chosen such that

$$A\boldsymbol{x} = P\boldsymbol{b}, \tag{85}$$

which has a (not-necessarily-unique) solution because $P\boldsymbol{b} \in R(A)$ by definition. In other words, a minimizer of (65) is such $\boldsymbol{x}$ that $A\boldsymbol{x}$ equals the orthogonal projection of $\boldsymbol{b}$ onto $R(A)$.

If $N(A) = \{0\}$, i.e, $A\boldsymbol{\alpha} = 0$ if and only if $\boldsymbol{\alpha} = 0$, the columns of $A$ are a (linearly independent) basis for $R(A)$. Thus, the definition of an orthogonal projection matrix (78) gives $P = A(A^TA)^{-1}A^T$ and (85) turns into

$$A\boldsymbol{x} = A(A^TA)^{-1}A^T\boldsymbol{b}. \tag{86}$$

In particular, the unique solution of the normal equation (66), i.e.,

$$\boldsymbol{x} = (A^TA)^{-1}A^T\boldsymbol{b}$$

clearly satisfies (86) — as it should because it is the unique minimizer of (65) if $N(A) = N(A^TA) = \{0\}$ according to Theorem 1.1.

## 4. QR decomposition

The normal equation approach to solving the least squares problem has two drawbacks: First of all, if $A$ does not have a trivial nullspace, the normal equation is not uniquely solvable. Secondly, even if $N(A) = \{0\}$, the conditioning of the normal equation is much worse than that of the original problem (1): For example if $A \in \mathbb{R}^{n\times n}$, the condition numbers corresponding to the 2-norm satisfy

$$\kappa_2(A^TA) = \kappa_2(A)^2,$$

which follows from Lemma 2.5 of Chapter 1 and the fourth exercise sheet after noticing that $A^TA$ is symmetric. According to the discussion in Section 1, if the condition number is squared, solving a linear system becomes much more sensitive to, e.g., floating point errors. In some extreme cases, resorting to the normal equation may even lead to loss of information, as demonstrated by the following example.

Example 4.1. *Let*

$$A = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix} \quad \textit{so that} \quad A^T = \begin{bmatrix} 1 & \epsilon & 0 \\ 1 & 0 & \epsilon \end{bmatrix}.$$

*Hence,*

$$A^TA = \begin{bmatrix} 1+\epsilon^2 & 1 \\ 1 & 1+\epsilon^2 \end{bmatrix}.$$

*In floating point arithmetics $fl(1+\epsilon^2) = 1$ when $\epsilon$ is sufficiently small; note that for $\epsilon \ll 1$, the square $\epsilon^2$ is much smaller that $\epsilon$ itself. When this happens,*

$$fl(A^TA) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

*i.e., data is lost and, in this case, the normal equation is even no longer uniquely solvable. In Matlab, this happens if $\epsilon < 10^{-9}$.*

To avoid the dramatic decrease in stability due to the normal equation approach and the related floating point errors, the least squares problem is typically solved using either the QR or the *singular value* (SVD) decomposition for the matrix $A$. Both methods can be understood as ways to compute an orthonormal basis for the subspace $R(A)$.

The simplest way to compute the QR decomposition is via the *Gram–Schmidt orthogonalization process* that is based on the following Lemma. In what follows, "orthogonality" refers to orthogonality in the sense of the Euclidean inner product.

Lemma 4.1. *Let $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_k\} \subset \mathbb{R}^m$, $k < m$, be a set of orthonormal vectors, i.e., $\|\boldsymbol{q}_i\|_2 = 1$, $i = 1, \dots, k$, and*

$$\boldsymbol{q}_i^T\boldsymbol{q}_j = 0 \qquad \textit{for } i \neq j.$$

*In addition, assume that $\boldsymbol{a} \in \mathbb{R}^m$ does not belong to* $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$ *and define*

$$\boldsymbol{q}_{k+1} = \frac{\tilde{\boldsymbol{q}}_{k+1}}{\|\tilde{\boldsymbol{q}}_{k+1}\|_2}, \qquad \textit{where} \quad \tilde{\boldsymbol{q}}_{k+1} = \boldsymbol{a} - \sum_{i=1}^{k}(\boldsymbol{a}^T\boldsymbol{q}_i)\boldsymbol{q}_i. \tag{87}$$

*Then* $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1}\} \subset \mathbb{R}^n$ *is a set of orthonormal vectors and*

$$\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1}) = \text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k, \boldsymbol{a}). \tag{88}$$

Proof. To begin with note that $\tilde{\boldsymbol{q}}_{k+1}$ defined in (87) is a nonzero vector because $\boldsymbol{a} \notin \text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$, i.e., $\boldsymbol{a}$ cannot be given as a linear combination of $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k$.

To prove the orthonormality of the set $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1}\}$ it is enough to prove that $\boldsymbol{q}_{k+1}$ is of unit Euclidean length and orthogonal to $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k$. From the first equation in (87), it is obvious that $\|\boldsymbol{q}_{k+1}\|_2 = 1$. Moreover, since $\boldsymbol{q}_{k+1}$ and $\tilde{\boldsymbol{q}}_{k+1}$ are parallel, it is actually enough to show that $\tilde{\boldsymbol{q}}_{k+1}$ is orthogonal to $\boldsymbol{q}_1, \dots, \boldsymbol{q}_k$: for any $j = 1, \dots, k$, we have

$$\tilde{\boldsymbol{q}}_{k+1}^T \boldsymbol{q}_j = \Big(\boldsymbol{a} - \sum_{i=1}^{k} (\boldsymbol{a}^T \boldsymbol{q}_i) \boldsymbol{q}_i\Big)^T \boldsymbol{q}_j = \boldsymbol{a}^T \boldsymbol{q}_j - \sum_{i=1}^{k} (\boldsymbol{a}^T \boldsymbol{q}_i)(\boldsymbol{q}_i^T \boldsymbol{q}_j) = \boldsymbol{a}^T \boldsymbol{q}_j - \boldsymbol{a}^T \boldsymbol{q}_j = 0$$

due to the orthonormality of $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_k\}$.

Although (88) follows straightforwardly from the definition of linear span, let us anyway carefully prove it for the sake of completeness. Assume first that $\boldsymbol{x} \in \text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1})$, i.e.,

$$\boldsymbol{x} = \sum_{i=1}^{k+1} \alpha_i \boldsymbol{q}_i = \sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i + \alpha_{k+1} \boldsymbol{q}_{k+1}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^{k+1}$. Note that (87) can be rewritten in the form

$$\boldsymbol{q}_{k+1} = \frac{1}{\|\tilde{\boldsymbol{q}}_{k+1}\|_2} \Big(\boldsymbol{a} - \sum_{i=1}^{k} (\boldsymbol{a}^T \boldsymbol{q}_i) \boldsymbol{q}_i\Big).$$

Hence,

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i + \frac{\alpha_{k+1}}{\|\tilde{\boldsymbol{q}}_{k+1}\|_2} \Big(\boldsymbol{a} - \sum_{i=1}^{k} (\boldsymbol{a}^T \boldsymbol{q}_i) \boldsymbol{q}_i\Big) = \sum_{i=1}^{k} \Big(\alpha_i - \frac{\alpha_{k+1} \boldsymbol{a}^T \boldsymbol{q}_i}{\|\tilde{\boldsymbol{q}}_{k+1}\|_2}\Big) \boldsymbol{q}_i + \frac{\alpha_{k+1}}{\|\tilde{\boldsymbol{q}}_{k+1}\|_2} \boldsymbol{a},$$

which is obviously in $\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k, \boldsymbol{a})$, meaning that $\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1}) \subset \text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k, \boldsymbol{a})$.

On the other hand, if $\boldsymbol{x} \in \text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k, \boldsymbol{a})$, then for some $\boldsymbol{\alpha} \in \mathbb{R}^{k+1}$,

$$\begin{aligned}\boldsymbol{x} &= \sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i + \alpha_{k+1} \boldsymbol{a} = \sum_{i=1}^{k} \alpha_i \boldsymbol{q}_i + \alpha_{k+1} \Big(\|\tilde{\boldsymbol{q}}_{k+1}\|_2 \, \boldsymbol{q}_{k+1} + \sum_{i=1}^{k} (\boldsymbol{a}^T \boldsymbol{q}_i) \boldsymbol{q}_i\Big) \\ &= \sum_{i=1}^{k} (\alpha_i + \alpha_{k+1} \boldsymbol{a}^T \boldsymbol{q}_i) \boldsymbol{q}_i + \alpha_{k+1} \|\tilde{\boldsymbol{q}}_{k+1}\|_2 \, \boldsymbol{q}_{k+1},\end{aligned}$$

which clearly belongs to $\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1})$. Hence, also $\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k, \boldsymbol{a}) \subset \text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{k+1})$, which completes the proof. □

The intuitive idea of (87) is that one first subtracts from $\boldsymbol{a}$ its projections onto the one-dimensional subspaces defined by $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$, leaving only the component of $\boldsymbol{a}$ orthogonal to $\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$, and then this component is normalized. In fact, one can write the second equation of (87) in the form

$$\tilde{\boldsymbol{q}}_{k+1} = (I - P_k)\boldsymbol{a},$$

where $P_k \in \mathbb{R}^{m \times m}$ is the orthogonal projection matrix onto the subspace $\text{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$; see (79).

Using Lemma 4.1, it is straightforward to compute an orthonormal basis for the subspace

$$R(A) = \text{span}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n) \subset \mathbb{R}^m,$$

assuming the columns $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ of the matrix $A \in \mathbb{R}^{m \times n}$ are linearly independent, i.e., assuming $N(A) = \{0\}$. Indeed, such basis $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_n\}$ can be recursively obtained via

$$\boldsymbol{q}_j = \frac{\tilde{\boldsymbol{q}}_j}{\|\tilde{\boldsymbol{q}}_j\|_2}, \qquad \text{where} \quad \tilde{\boldsymbol{q}}_j = \boldsymbol{a}_j - \sum_{i=1}^{j-1} (\boldsymbol{a}_j^T \boldsymbol{q}_i) \boldsymbol{q}_i, \qquad \text{for } j = 1, \dots, n.$$

In other words, one first defines $\boldsymbol{q}_1$ by simply normalizing $\boldsymbol{a}_1$, then one computes a unit vector $\boldsymbol{q}_2$ that is orthogonal to $\boldsymbol{q}_1$ and satisfies $\operatorname{span}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \operatorname{span}(\boldsymbol{q}_1, \boldsymbol{a}_2) = \operatorname{span}(\boldsymbol{a}_1, \boldsymbol{a}_2)$, then one continues by computing a unit vector $\boldsymbol{q}_2$ that is orthogonal to both $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ and satisfies

$$\operatorname{span}(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3) = \operatorname{span}(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{a}_3) = \operatorname{span}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3),$$

and so on until $\boldsymbol{q}_n$ is computed and it holds that $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_n) = \operatorname{span}(\boldsymbol{a}_1, \dots, \boldsymbol{a}_n) = R(A)$.

Take note that one can get the original columns of $A$ back via

$$\boldsymbol{a}_j = \|\tilde{\boldsymbol{q}}_j\|_2 \, \boldsymbol{q}_j + \sum_{i=1}^{j-1} (\boldsymbol{a}_j^T \boldsymbol{q}_i) \boldsymbol{q}_i, \qquad j = 1, \dots, n, \tag{89}$$

which demonstrates that, for any $j = 1, \dots, n$, the $j$th column $\boldsymbol{a}_j$ of $A$ can be given as a linear combination of $\boldsymbol{q}_1, \dots, \boldsymbol{q}_j$, i.e., of (only) the first $j$ orthonormal basis vectors of $R(A)$ produced by the Gram–Schmidt process. Defining in the standard manner $Q = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_n] \in \mathbb{R}^{m \times n}$ and collecting the coefficients in the linear combinations of (89) as columns of an upper triangular matrix $R \in \mathbb{R}^{n \times n}$, the equations (89) can be written neatly in a matrix form

$$A = QR.$$

To be more precise, $R$ can be given elementwise as

$$R_{i,j} = \begin{cases} \boldsymbol{a}_j^T \boldsymbol{q}_i & \text{if } i < j, \\ \|\tilde{\boldsymbol{q}}_i\|_2 & \text{if } i = j, \\ 0 & \text{if } i > j. \end{cases}$$

Note also that $Q^T Q = I \in \mathbb{R}^{n \times n}$ because the columns of $Q$ are orthonormal.

Once a QR decomposition has been computed for $A$,[3] the least squares problem (65) can be formulated *equivalently* as the equation

$$R\boldsymbol{x} = Q^T \boldsymbol{b}.$$

There are (at least) two ways to deduce this. First of all, since $N(A^T A) = N(A) = 0$ by assumption, the unique solution of the least squares problem (65) is the unique solution of the normal equation (66). On the other hand, since

$$A^T = R^T Q^T \qquad \text{and} \qquad A^T A = R^T Q^T Q R = R^T I R = R^T R,$$

the normal equation can be written as

$$R^T R \boldsymbol{x} = R^T Q^T \boldsymbol{b} \qquad \Longleftrightarrow \qquad R\boldsymbol{x} = Q^T \boldsymbol{b},$$

where the last step follows by multiplying from the left by the inverse of $R^T$.[4]

The second way is based on using the reformulation (85) of the least squares problem and writing the orthogonal projection matrix $P \in \mathbb{R}^{m \times m}$ onto $R(A)$ with the help of the orthonormal basis $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_n\}$, that is,

$$P = Q\big(Q^T Q\big)^{-1} Q^T = Q I^{-1} Q^T = Q Q^T$$

according to (78). Hence, (85) reduces to

$$A\boldsymbol{x} = QR\boldsymbol{x} = QQ^T \boldsymbol{b} \qquad \Longleftrightarrow \qquad R\boldsymbol{x} = Q^T \boldsymbol{b},$$

where "$\Rightarrow$" follows by multiplying from left with $Q^T$ and "$\Leftarrow$" by multiplying from left with $Q$.

There are two ways to numerically implement the Gram–Schmidt process, or the computation of a QR decomposition, namely classical and modified. In floating point arithmetics these two procedures have very different numerical stability properties. The more stable *modified Gram–Schmidt* procedure can be implemented in Matlab as follows:

[3] Observe that when introducing the QR decomposition, we assumed that the columns of $A \in \mathbb{R}^{m \times n}$ are linearly independent, i.e., $N(A) = \{0\}$, and so everything that follows is conditional on this assumption. In particular, this excludes the case $n > m$.

[4] Observe that both $R$ and $R^T$ are invertible: As triangular square matrices, their eigenvalues are their diagonal elements $\|\tilde{\boldsymbol{q}}_i\|_2 > 0$, $i = 1, \dots, n$. Hence, zero is not an eigenvalue for either of these square matrices, and thus they are both invertible.

```
function [Q,R] = my_gsmith(A)

Q = [];
for i=1:size(A,2)
    q = A(:,i);

    for k=1:size(Q,2)
        R(k,i) = q'*Q(:,k);
        q = q - R(k,i)*Q(:,k);
    end
    R(i,i) = norm(q);
    Q(:,i) = q/R(i,i);
end
```

And the *classical Gram-Schmidt* procedure as follows:

```
function [Q,R] = my_c_gsmith(A)

Q = [];
for i=1:size(A,2)
    q = A(:,i);

    for k=1:size(Q,2)
        R(k,i) = q'*Q(:,k);
    end

    for k=1:size(Q,2)
        q = q - R(k,i)*Q(:,k);
    end
    R(i,i) = norm(q);
    Q(:,i) = q/R(i,i);
end
```

In exact arithmetics, both these algorithms return $Q \in \mathbb{R}^{m\times n}$ such that $Q^TQ = I$ and $R(Q) = R(A)$ as well as an upper triangular $R \in \mathbb{R}^{n\times n}$ such that $A = QR$. However, the two implementations of the Gram–Schmidt process have very different numerical stability properties. The exact definition of numerical stability is in this case rather complicated. One can, e.g., measure the orthogonality of $Q$, i.e. $\|I - Q^TQ\|$, or the accuracy of the decomposition, i.e. $\|A - QR\|$, in some suitable matrix norm. The loss of orthogonality is more prone to stability issues than the accuracy of the decomposition. For the modified Gram–Schmidt, one can prove numerical stability in both of these measures, whereas the classical Gram–Schmidt is not numerically stable.

## 5. Singular value decomposition

The *singular value decomposition* (SVD) is a general matrix factorization that can be formed for *any* matrix. For a symmetric (or Hermitian) and positive definite matrix, the SVD coincides with the unitary diagonalization, or the eigenvalue decomposition; more generally, the eigenvalue and singular value decompositions are almost equivalent for any symmetric matrix, meaning that they can be obtained from one another by simply changing signs of certain matrix elements.

The intuitive idea behind the singular value decomposition is the following: Any matrix $A \in \mathbb{R}^{m\times n}$ can be represented as a diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ between a certain *orthonormal* basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ for $\mathbb{R}^n$ and another *orthonormal* basis $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_m\}$ for $\mathbb{R}^m$.[5] In other words, $A\boldsymbol{x}$ can be evaluated by first computing the coordinates of $\boldsymbol{x}$ in the orthonormal basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, then scaling these coordinates by certain non-negative scalars, and finally interpreting the scaled coordinates as the coordinates of $A\boldsymbol{x}$ in the basis $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_m\}$. If $m > n$, the coordinates/projections of $A\boldsymbol{x}$ corresponding to $\boldsymbol{u}_{n+1}, \dots, \boldsymbol{u}_m$ are automatically set to zero; if

[5] Here and in what follows, "orthonormal" is to be understood in the sense of the Euclidean inner product.

$n > m$, the coordinates/projections of $\boldsymbol{x}$ corresponding to $\boldsymbol{v}_{m+1}, \dots, \boldsymbol{v}_n$ do not affect $A\boldsymbol{x}$ in any way.

Forming the SVD corresponds to finding the aforementioned orthonormal basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ and $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_m\}$, i.e., the right-hand and left-hand *singular vectors*, as well as the corresponding 'scaling scalars'

$$\sigma_1 \geq \sigma_2 \geq \sigma_{\min\{m,n\}} \geq 0$$

that are called the *singular values*, are arranged in decreasing order, and are the diagonal elements of $\Sigma \in \mathbb{R}^{m \times n}$. We denote by $p \leq \min\{m, n\}$ the largest index for which the corresponding singular value is positive, i.e., $\sigma_p > 0$ but $\sigma_{p+1} = 0$ (or $p = \min\{m, n\}$). Note that, according to the above explanation, all information in $A$ is actually encoded in $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_p\}$, $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_p\}$ and $\sigma_1, \dots, \sigma_p$ because the components of $\boldsymbol{x}$ in the directions $\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n$ are either scaled by zero singular values or altogether ignored when evaluating $A\boldsymbol{x}$. It will turn out that $p = \operatorname{rank}(A) = \dim(R(A))$.

Let us formulate the above 'construction' in a matrix form. A SVD for $A \in \mathbb{R}^{m \times n}$ is

$$A = U\Sigma V^T, \tag{90}$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal (or unitary) matrices, i.e., their columns are the aforementioned orthonormal basis for $\mathbb{R}^{m \times m}$ and $\mathbb{R}^{n \times n}$, respectively, and $\Sigma \in \mathbb{R}^{m \times n}$ carries the singular values in the decreasing order on its diagonal. To be more precise, when $n > m$, $\Sigma$ takes the form

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_m & \end{bmatrix},$$

and in the case $m > n$, $\Sigma$ has the form

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \\ & & & \end{bmatrix},$$

that is, there are extra zeros either at the right edge ($n > m$) or at the bottom ($m > n$) of $\Sigma$. The SVD is not unique: the matrices $U$ and $V$ can be chosen in different ways.

Let us verify that the decomposition (90) is concordant with the verbal explanation of the SVD presented in the beginning of this section: For any $\boldsymbol{x} \in \mathbb{R}^n$ we have

$$\begin{aligned} A\boldsymbol{x} = U\Sigma V^T\boldsymbol{x} = U\Sigma \begin{bmatrix} \boldsymbol{v}_1^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{v}_n^T\boldsymbol{x} \end{bmatrix} = U \begin{bmatrix} \sigma_1 \boldsymbol{v}_1^T\boldsymbol{x} \\ \vdots \\ \sigma_{\min\{m,n\}} \boldsymbol{v}_{\min\{m,n\}}^T\boldsymbol{x} \\ 0 \end{bmatrix} &= \sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{v}_i^T\boldsymbol{x})\boldsymbol{u}_i \\ &= \sum_{i=i}^{p} \sigma_i \boldsymbol{u}_i(\boldsymbol{v}_i^T\boldsymbol{x}), \end{aligned} \tag{91}$$

where in the third to last step $0 \in \mathbb{R}^{m-n}$ if $m > n$ and otherwise it should be ignored. As $\boldsymbol{v}_1^T\boldsymbol{x}, \dots, \boldsymbol{v}_n^T\boldsymbol{x}$ are the coordinates/projections of $\boldsymbol{x} \in \mathbb{R}^n$ in the orthonormal basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, the decomposition (90) is exactly of the desired form. In particular, the right-hand side of (90) demonstrates that $A$ is completely determined by the first $p$ singular values and singular vectors.

We still need to prove that any matrix really has a SVD. We start with a lemma that is related to the eigenvalues and eigenvectors of $A^TA$, i.e., entities that will turn out to be closely related to the SVD of $A$ itself.

LEMMA 5.1. *Let $A \in \mathbb{R}^{m \times n}$. Let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \subset \mathbb{R}^n$ be an orthonormal set of eigenvectors for $A^TA \in \mathbb{R}^{n \times n}$ arranged so that the corresponding eigenvalues, $\lambda_1, \dots, \lambda_n$, form a non-increasing* *sequence. Let $p$ denote the number of those $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ that correspond to positive eigenvalues. Then, $A\boldsymbol{v}_i = 0$ for $i = p+1, \dots, n$, and*

$$\boldsymbol{u}_i := \frac{1}{\sqrt{\lambda_i}} A\boldsymbol{v}_i, \qquad i = 1, \dots, p, \tag{92}$$

*form an orthonormal set in $\mathbb{R}^m$. In particular, $0 \le p \le \min\{m, n\}$.*

Proof. To begin with, note that there really exists an orthonormal eigenbasis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \subset \mathbb{R}^n$ for $A^TA \in \mathbb{R}^{n\times n}$ because it is symmetric (or Hermitian); see Theorem 2.1 of Chapter 2 and the succeeding discussion. Moreover, all eigenvalues of $A^TA$ are non-negative since $A^TA$ is positive semi-definite:

$$0 \le \|A\boldsymbol{v}_i\|_2^2 = \boldsymbol{v}_i^T A^T A \boldsymbol{v}_i = \lambda_i \|\boldsymbol{v}_i\|_2^2 = \lambda_i$$

for any $i = 1, \dots, n$. This formula also proves that $A\boldsymbol{v}_i = 0$ if and only if $\lambda_i = 0$, i.e., if and only if $i = p+1, \dots, n$; in particular, none of the vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_p$ defined by (92) is a zero vector.

Let us then take a closer look at the (left-hand singular) vectors defined by (92). It holds that

$$\boldsymbol{u}_i^T \boldsymbol{u}_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} \boldsymbol{v}_i^T A^T A \boldsymbol{v}_j = \frac{\lambda_j}{\sqrt{\lambda_i \lambda_j}} \boldsymbol{v}_i^T \boldsymbol{v}_j$$

which vanishes when $i \ne j$ due to the orthonormality of $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_p\}$. On the other hand,

$$\|\boldsymbol{u}_i\|_2^2 = \boldsymbol{u}_i^T \boldsymbol{u}_i = \frac{\lambda_i}{\sqrt{\lambda_i \lambda_i}} \boldsymbol{v}_i^T \boldsymbol{v}_i = \|\boldsymbol{v}_i\|_2^2 = 1,$$

which completes the proof of the orthonormality of $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_p\}$.

Finally, because $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_p\}$ is an orthonormal set in $\mathbb{R}^n$ and $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_p\}$ is an orthonormal set in $\mathbb{R}^n$, it must hold that $p \le n$ and $p \le m$, i.e., $p \le \min\{m, n\}$. □

As a side note, observe that all positive eigenvalues of $A^TA \in \mathbb{R}^{n\times n}$ are also eigenvalues of $AA^T \in \mathbb{R}^{m\times m}$ with $\boldsymbol{u}_1, \dots, \boldsymbol{u}_p$ being the corresponding orthonormal eigenvectors:

$$AA^T \boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} A(A^T A \boldsymbol{v}_i) = \lambda_i \Big( \frac{1}{\sqrt{\lambda_i}} A\boldsymbol{v}_i \Big) = \lambda_i \boldsymbol{u}_i$$

for $i = 1, \dots, p$.

Theorem 5.1. *Let $A \in \mathbb{R}^{m\times n}$. Then there exists an integer $0 \le p \le \min\{m, n\}$, positive coefficients $\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_p > 0$ and sets of orthonormal vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_p\} \subset \mathbb{R}^n$ and $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_p\} \subset \mathbb{R}^m$ such that*

$$A = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T.$$

Proof. Let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\} \subset \mathbb{R}^n$ be an orthonormal eigenbasis for $A^TA \in \mathbb{R}^n$ organized in the same way as in Lemma 5.1. Let $\boldsymbol{x} \in \mathbb{R}^n$ be arbitrary and write it in the eigenbasis of $A^TA \in \mathbb{R}^n$ as

$$\boldsymbol{x} = \sum_{i=1}^{n} (\boldsymbol{v}_i^T \boldsymbol{x}) \boldsymbol{v}_i.$$

Because $A\boldsymbol{v}_i = 0$ for $i = p+1, \dots, n$ by Lemma 5.1, we have

$$A\boldsymbol{x} = \sum_{i=1}^{n} (\boldsymbol{v}_i^T \boldsymbol{x}) A\boldsymbol{v}_i = \sum_{i=1}^{p} (\boldsymbol{v}_i^T \boldsymbol{x}) A\boldsymbol{v}_i = \sum_{i=1}^{p} (\boldsymbol{v}_i^T \boldsymbol{x}) \sqrt{\lambda_i} \boldsymbol{u}_i = \sum_{i=1}^{p} \sqrt{\lambda_i} \boldsymbol{u}_i \boldsymbol{v}_i^T \boldsymbol{x}$$

where the orthonormal vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_p$ are defined by (92). After defining $\sigma_i = \sqrt{\lambda_i}$, $i = 1, \dots, p$, the proof is complete. □

Let $0 \le p \le \min\{m, n\}$, $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_p\}$ and $\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_p > 0$ be as defined in Lemma 5.1 and Theorem 5.1. Based on Theorem 5.1, the matrix form of SVD, i.e. (90), is obtained by defining $\sigma_{p+1}, \dots, \sigma_{\min\{m,n\}} = 0$ and introducing the missing orthonormal columns of $U$, i.e., $\boldsymbol{u}_{p+1}, \dots, \boldsymbol{u}_m$, via, e.g., the Gram–Schmidt orthogonalization process. It is easy to check that with these choices (91) is valid for any $\boldsymbol{x} \in \mathbb{R}^n$ by virtue of Theorem 5.1, and thus we have proved the existence of (90) for any $A \in \mathbb{R}^{m\times n}$.

Although the 'extra singular vectors' $\boldsymbol{u}_{p+1}, \dots, \boldsymbol{u}_m$ and $\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n$ are not actually needed in constructing $A$ (cf. Theorem 5.1), they can be used to determine $R(A)$, $N(A)$ and their orthogonal complements.

Theorem 5.2. *Let $A \in \mathbb{R}^{m\times n}$, consider its SVD given by* (90) *and denote (still) by $p$ the number of positive singular values of $A$. Then,*

(1) $R(A) = \operatorname{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_p)$,
(2) $N(A) = \operatorname{span}(\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n)$,
(3) $R(A)^\perp = \operatorname{span}(\boldsymbol{u}_{p+1}, \dots, \boldsymbol{u}_m)$,
(4) $N(A)^\perp = \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)$.

*In particular,* $\operatorname{rank}(A) = \dim(R(A)) = p$.

Proof. According to Theorem 5.1,

$$R(A) = \Big\{\boldsymbol{y} \in \mathbb{R}^m \;\Big|\; \boldsymbol{y} = \sum_{i=1}^p \sigma_i(\boldsymbol{v}_i^T\boldsymbol{x})\boldsymbol{u}_i \;\text{ for some } \boldsymbol{x} \in \mathbb{R}^n\Big\}$$

which clearly is a subset of $\operatorname{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_p)$. On the other hand, an arbitrary element in $\operatorname{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_p)$, say,

$$\boldsymbol{y} = \sum_{i=1}^p \alpha_i \boldsymbol{u}_i, \qquad \boldsymbol{\alpha} \in \mathbb{R}^p,$$

can be written as

$$A\Big(\sum_{j=1}^p \frac{\alpha_j}{\sigma_j}\boldsymbol{v}_j\Big) = \sum_{i=1}^p \sigma_i \boldsymbol{u}_i \Big(\sum_{j=1}^p \frac{\alpha_j}{\sigma_j}\boldsymbol{v}_i^T\boldsymbol{v}_j\Big) = \sum_{i=1}^p \sigma_i \frac{\alpha_i}{\sigma_i}\boldsymbol{u}_i = \boldsymbol{y}$$

due to the orthonormality of $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_p\}$. Hence also $\operatorname{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_p) \subset R(A)$, and altogether $R(A) = \operatorname{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_p)$. This also proves that $\operatorname{rank}(A) = p$. Moreover, since the $m - p$ vectors $\boldsymbol{u}_{p+1}, \dots, \boldsymbol{u}_m$ are mutually orthonormal as well as orthogonal to $R(A)$, they must form a basis for the $m - p$ dimensional orthogonal complement $R(A)^\perp$.

Let then $\boldsymbol{x} \in N(A)$ be arbitrary and write it in the orthonormal basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ of $\mathbb{R}^n$ as

$$\boldsymbol{x} = \sum_{j=1}^n \alpha_j \boldsymbol{v}_j, \qquad \boldsymbol{\alpha} \in \mathbb{R}^n.$$

By Theorem 5.1,

$$0 = A\boldsymbol{x} = A\Big(\sum_{j=1}^n \alpha_j \boldsymbol{v}_j\Big) = \sum_{i=1}^p \sigma_i \boldsymbol{u}_i \Big(\sum_{j=1}^n \alpha_j \boldsymbol{v}_i^T \boldsymbol{v}_j\Big) = \sum_{i=1}^p \sigma_i \alpha_i \boldsymbol{u}_i$$

due to the orthonormality of $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$. Since $\sigma_1, \dots, \sigma_p$ are positive and $\boldsymbol{u}_1, \dots, \boldsymbol{u}_p$ are linearly independent, it must hold that $\alpha_1 = \cdots = \alpha_p = 0$, meaning that

$$\boldsymbol{x} = \sum_{j=p+1}^n \alpha_j \boldsymbol{v}_j \in \operatorname{span}(\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n).$$

In other words, $N(A) \subset \operatorname{span}(\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n)$. On the other hand, via an analogous calculation it follows easily that any $\boldsymbol{x} \in \operatorname{span}(\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n)$ belongs to $N(A)$, and so $N(A) = \operatorname{span}(\boldsymbol{v}_{p+1}, \dots, \boldsymbol{v}_n)$. Finally, the remaining orthonormal vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_p$ must form a basis for the $n - (n - p) = p$ dimensional orthogonal complement $N(A)^\perp$.

□

We complete these lecture notes by demonstrating the SVD can be used for solving least squares problems even if $N(A) = 0$.

THEOREM 5.3. *Let $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Consider the SVD of $A$ given by (90) and denote (still) by $p$ the number of positive singular values of $A$. Then, the vector*

$$\boldsymbol{x}^\dagger := \sum_{j=1}^{p} \frac{1}{\sigma_j}(\boldsymbol{u}_j^T\boldsymbol{b})\boldsymbol{v}_j \tag{93}$$

*is a solution to the normal equation (66).*

PROOF. The proof is simply a straightforward computation based on the representations

$$A = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \qquad \text{and} \qquad A^T = \Big(\sum_{j=1}^{p} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T\Big)^T = \sum_{j=1}^{p} \sigma_j \big(\boldsymbol{u}_j \boldsymbol{v}_j^T\big)^T = \sum_{j=1}^{p} \sigma_j \boldsymbol{v}_j \boldsymbol{u}_j^T$$

provided by Theorem 5.1. Indeed,

$$\begin{aligned} A\boldsymbol{x}^\dagger &= \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \Big(\sum_{j=1}^{p} \frac{1}{\sigma_j}(\boldsymbol{u}_j^T\boldsymbol{b})\boldsymbol{v}_j\Big) = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \Big(\sum_{j=1}^{p} \frac{1}{\sigma_j}(\boldsymbol{u}_j^T\boldsymbol{b})(\boldsymbol{v}_i^T\boldsymbol{v}_j)\Big) \\ &= \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i}(\boldsymbol{u}_i^T\boldsymbol{b})\boldsymbol{u}_i = \sum_{i=1}^{p} (\boldsymbol{u}_i^T\boldsymbol{b})\boldsymbol{u}_i, \end{aligned}$$

where we once again used the orthonormality of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\}$. Analogously,

$$A^T(A\boldsymbol{x}^\dagger) = \sum_{j=1}^{p} \sigma_j \boldsymbol{v}_j \boldsymbol{u}_j^T \Big(\sum_{i=1}^{p} (\boldsymbol{u}_i^T\boldsymbol{b})\boldsymbol{u}_i\Big) = \sum_{j=1}^{p} \sigma_j \boldsymbol{v}_j \Big(\sum_{i=1}^{p} (\boldsymbol{u}_i^T\boldsymbol{b})(\boldsymbol{u}_j^T\boldsymbol{u}_i)\Big) = \sum_{j=1}^{p} \sigma_j (\boldsymbol{u}_j^T\boldsymbol{b})\boldsymbol{v}_j,$$

which equals

$$A^T\boldsymbol{b} = \sum_{j=1}^{p} \sigma_j \boldsymbol{v}_j \boldsymbol{u}_j^T \boldsymbol{b} = \sum_{j=1}^{p} \sigma_j (\boldsymbol{u}_j^T\boldsymbol{b})\boldsymbol{v}_j,$$

and so the proof is complete. □

COROLLARY 5.1. *Let $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Consider the SVD of $A$ given by (90) and denote (still) by $p$ the number of positive singular values of $A$. All solutions of the least squares problem (65) are of the form*

$$\boldsymbol{x} = \boldsymbol{x}^\dagger + \boldsymbol{z} \tag{94}$$

*for some $\boldsymbol{z} \in N(A)$.*

PROOF. Since we have (finally!) found a solution $\boldsymbol{x}^\dagger$ to the normal equation (66) independently of whether $N(A) = \{0\}$, by virtue of Theorem 1.1 we know that $\boldsymbol{x}$ is a solution of the least squares problem (65) if and only if it solves the normal equation (66). In consequence, it is enough to prove that all solutions of the normal equation (66) are of the form (94).

Let $\boldsymbol{x}$ be of the form (94). Obviously,

$$A^TA\boldsymbol{x} = A^TA\boldsymbol{x}^\dagger + A^TA\boldsymbol{z} = A^T\boldsymbol{b} + 0 = A^T\boldsymbol{b},$$

and so $\boldsymbol{x}$ solves the normal equation (66). On the other hand, if $\boldsymbol{x}$ is an arbitrary solution of the normal equation, we may define $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}^\dagger$, i.e., $\boldsymbol{x} = \boldsymbol{x}^\dagger + \boldsymbol{z}$. It follows that

$$A^TA\boldsymbol{z} = A^TA\boldsymbol{x} - A^TA\boldsymbol{x}^\dagger = A^T\boldsymbol{b} - A^T\boldsymbol{b} = 0.$$

Hence, $\boldsymbol{z} \in N(A^TA) = N(A)$ and so $\boldsymbol{x}$ is of the form (94). This completes the proof. □

According to Theorems 5.2 and 5.3, the special least squares solution $\boldsymbol{x}^\dagger$ belongs to $\mathrm{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) = N(A)^\perp$. Hence, any (other) solution (94) to the least squares problem (65) satisfies

$$\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}^\dagger + \boldsymbol{z}\|_2^2 = \|\boldsymbol{x}^\dagger\|_2^2 + 2\boldsymbol{z}^T\boldsymbol{x}^\dagger + \|\boldsymbol{z}\|_2^2 = \|\boldsymbol{x}^\dagger\|_2^2 + \|\boldsymbol{z}\|_2^2 \geq \|\boldsymbol{x}^\dagger\|_2^2,$$

and so $\boldsymbol{x}^\dagger$ is the solution to (65) with the smallest Euclidean norm, i.e. the *unique* so-called *minimum norm solution* to (65).

Finally, by comparing (93) to (91), it is rather obvious that one has the representation

$$\boldsymbol{x}^\dagger = A^\dagger \boldsymbol{b}, \qquad \text{where} \quad A^\dagger = V\Sigma^\dagger U^T \in \mathbb{R}^{n\times m}$$

and $\Sigma^\dagger \in \mathbb{R}^{n\times m}$ is of the form

$$\Sigma = \begin{bmatrix} \sigma_1^{-1} & & & & & & & \\ & \sigma_2^{-1} & & & & & & \\ & & \ddots & & & & & \\ & & & \sigma_p^{-1} & & & & \\ & & & & 0 & & & \\ & & & & & \ddots & & \\ & & & & & & 0 & \end{bmatrix}$$

if $m > n$ and of the form

$$\Sigma = \begin{bmatrix} \sigma_1^{-1} & & & & & & \\ & \sigma_2^{-1} & & & & & \\ & & \ddots & & & & \\ & & & \sigma_p^{-1} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \\ & & & & & & \end{bmatrix}$$

if $n > m$. To put it short, $A^\dagger$ is defined based on the SVD of $A$ by reversing the roles of $U$ and $V$, inverting all positive singular values on the diagonal of $\Sigma$, and making the dimensions of $\Sigma^\dagger$ compatible via introduction of extra zeros. The 'almost-an-inverse-matrix' $A^\dagger \in \mathbb{R}^{n\times m}$ is called the *Moore–Penrose pseudoinverse* of $A \in \mathbb{R}^{m\times n}$.

# THE END