

Part I

APPLICATIONS OF LINEAR ALGEBRA

INTRODUCTION

This course “Applications of Linear Algebra” is based on the lectures given by the author to postgraduate students at Tallinn Technical University. Our aim was to acquaint the students with the linear algebra packages LINPACK, EISPACK and LAPACK, and with the theoretical fundamentals of the parts of the packages MATLAB, MAPLE, MATHCAD and MATHEMATICA related to linear algebra. We have tried to explain the linear algebra methods which form the basis for the computing methods used in the packages. We would like to stress that the aim of the course is not to work out concrete computing algorithms but to learn about the basic ideas related to these algorithms. It will be assumed that the reader is acquainted with the basic ideas of algebra.

The author would like to thank Assoc. Prof. Ellen Redi (Tallinn Pedagogical University) whose help in the improvement of the presented material both in its contents and its form has been enormous. Many of the examples and problems were prepared by students Kristiina Krüspan, Kadri Mikk, Reena Prints (Tallinn Pedagogical University), Andrei Filonov, Dmitri Tseluiko (Tartu University) Juhan-Peep Ernits and Heiki Hiisjärv (Tallinn Technical University) within the framework of the TEMPUS-project during their stay at Tampere University of Technology in June, 1997.

The numbers of their examples and problems are marked by an asterisk “*”.

The material is based on the monographs of G.H.Golub and C.F.Van Loan (1996), and G.Strang (1988).

I hope that the course will help the reader interested in applications of linear algebra more to use the linear algebra packages more effectively.

Author.

1 FUNDAMENTATIONS OF LINEAR ALGEBRA

1.1 Vectors

1.1.1 Vector Spaces

One of the fundamental concepts of linear algebra is that of vector space. At the same time it is one of the more often used concepts of algebraic structure in modern mathematics. For example, many function sets studied in mathematical analysis are with respect to their algebraic properties vector spaces. In analysis the notion “linear space” is used instead of the notion “vector space”.

Definition 1.1.1. A set $\mathbf{X}$ is called a *vector space over the number field $\mathbf{K}$* , if to every pair $(\mathbf{x}, \mathbf{y})$ of elements of $\mathbf{X}$ there corresponds a sum $\mathbf{x} + \mathbf{y} \in \mathbf{X}$, and to every pair $(\alpha, \mathbf{x})$, where $\alpha \in \mathbf{K}$ and $\mathbf{x} \in \mathbf{X}$, there corresponds an element $\alpha\mathbf{x} \in \mathbf{X}$, with the properties 1-8:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ (commutability of addition);
2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$ (associativity of addition);
3. $\exists \mathbf{0} \in \mathbf{X} : \mathbf{0} + \mathbf{x} = \mathbf{x}$ (existence of null element);
4. $\forall \mathbf{x} \in \mathbf{X} \Rightarrow \exists -\mathbf{x} \in \mathbf{X} : \mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ (existence of the inverse element);
5. $1 \cdot \mathbf{x} = \mathbf{x}$ (unitarism);
6. $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$ (associativity with respect to number multiplication);
7. $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ (distributivity with respect to vector addition);
8. $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ (distributivity with respect to number addition).

The properties 1-8 are called the vector space axioms. Axioms 1-4 show that $\mathbf{X}$ is a commutative group or an Abelian group with respect to vector addition. The second correspondence is called multiplication of the vector by a number, and it satisfies axioms 5-8. Elements of a vector space are called vectors. If $\mathbf{K} = \mathbf{R}$, then one speaks of a real vector space, and if $\mathbf{K} = \mathbf{C}$, then of a complex vector space. Instead of the notion “vector space” we shall use the abbreviative “space”.

Example 1.1.1. Let us consider the set of all $n \times 1$ -matrices with real elements:

$$\mathbf{X} = \left\{ \mathbf{x} : \mathbf{x} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \wedge \xi_i \in \mathbf{R} \right\}.$$

The sum of two matrices we define in the usual way by the addition of the corresponding elements. By multiplying the matrix by a real number λ we multiply all elements of the matrix by this number. The simple check will show that conditions 1-8 are satisfied. For example, let us check conditions 3 and 4. We construct

$$\mathbf{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \quad -\mathbf{x} = \begin{bmatrix} -\xi_1 \\ \vdots \\ -\xi_n \end{bmatrix}.$$

As

$$\mathbf{0} + \mathbf{x} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = \begin{bmatrix} 0 + \xi_1 \\ \vdots \\ 0 + \xi_n \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = \mathbf{x},$$

the element $\mathbf{0}$ satisfies condition 3 for arbitrary $\mathbf{x} \in \mathbf{X}$, and thus it is the null element of the space $\mathbf{X}$. For the element $-\mathbf{x}$

$$\mathbf{x} + (-\mathbf{x}) = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} + \begin{bmatrix} -\xi_1 \\ \vdots \\ -\xi_n \end{bmatrix} = \begin{bmatrix} \xi_1 - \xi_1 \\ \vdots \\ \xi_n - \xi_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0},$$

i.e., condition 4 is satisfied. Make sure of the validity of the remaining conditions 1-2 and 5-8.

The vector space in example 1.1.1 is called an *n-dimensional real arithmetical space* or in short $\mathbf{R}^n$. Declaring the vector $\mathbf{x}$ of the space $\mathbf{R}^n$ we often use the transposed matrix

$$\mathbf{x} = \begin{bmatrix} \xi_1 & \dots & \xi_n \end{bmatrix}^T.$$

In this presentation we often use punctuation marks (comma, semicolon) to separate the components of the vector, for example

$$\mathbf{x} = \begin{bmatrix} \xi_1, & \dots, & \xi_n \end{bmatrix}^T.$$

Example 1.1.1.* Let U be a set that consists of all pairs of real numbers $\mathbf{a} = (\alpha_1, \alpha_2)$, $\mathbf{b} = (\beta_1, \beta_2)$, ... We define addition and multiplication by a scalar in U as follows:

$$\mathbf{a} + \mathbf{b} = ((\alpha_1^3 + \beta_1^3)^{1/3}, (\alpha_2^3 + \beta_2^3)^{1/3}),$$

$$\alpha \mathbf{a} = (\alpha \alpha_1, \alpha \alpha_2).$$

Is the set U a vector space?

Proposition 1.1.1. Let $\mathbf{X}$ be a vector space. For arbitrary vectors $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ and number $\lambda \in \mathbf{K}$ the following assertions and equalities are valid:

- the null vector $\mathbf{0}$ of the vector space $\mathbf{X}$ is unique;
- the inverse vector $-\mathbf{x}$ of each $\mathbf{x} \in \mathbf{X}$ is unique;
- the uniqueness of the inverse vector allows to define the operation of subtraction by

$$\mathbf{x} - \mathbf{y} \stackrel{def}{=} \mathbf{x} + (-\mathbf{y});$$

- $\mathbf{x} = \mathbf{y} \Leftrightarrow \mathbf{x} - \mathbf{y} = \mathbf{0}$;
- $0\mathbf{x} = \mathbf{0} \quad \forall \mathbf{x} \in \mathbf{X}$;
- $\lambda \mathbf{0} = \mathbf{0} \quad \forall \lambda \in K$;
- $(-1)\mathbf{x} = -\mathbf{x}$;
- $\lambda \mathbf{x} = \mathbf{0} \Leftrightarrow (\lambda = 0 \vee \mathbf{x} = \mathbf{0}).$

Become convinced of the trueness of these assertions! $\square$

Example 1.1.2. Let us consider the set of all $(m \times n)$ -matrices with complex elements. The sum of these matrices will be defined by the addition of the corresponding elements of the matrices. By multiplying the matrix by a complex number λ one will multiply by this number all the elements of the matrix. We leave the check that all conditions 1-8 are satisfied to the reader. This vector space over the complex number field $\mathbf{C}$ will be denoted $\mathbf{C}^{m \times n}$. If we confine ourselves to real matrices, then we shall get a vector space $\mathbf{R}^{m \times n}$ over the number field $\mathbf{R}$. The space $\mathbf{C}^{m \times 1}$ will be identified with the space $\mathbf{C}^m$ and the space $\mathbf{R}^{m \times 1}$ with the space $\mathbf{R}^m$.

Example 1.1.3. The set $\mathbf{F}[\alpha, \beta]$ of all functions $\mathbf{x} : [\alpha, \beta] \rightarrow \mathbf{R}$ is a vector space (prove!) over the number field $\mathbf{R}$ if

$$(\mathbf{x} + \mathbf{y})(t) \stackrel{\text{def}}{=} \mathbf{x}(t) + \mathbf{y}(t) \quad \forall t \in [\alpha, \beta]$$

and

$$(\lambda \mathbf{x})(t) \stackrel{\text{def}}{=} \lambda \mathbf{x}(t) \quad \forall t \in [\alpha, \beta].$$

1.1.2 Subspaces of the Vector Space

Definition 1.2.1. The set $\mathbf{W}$ of vectors of the vector space $\mathbf{X}$ (over the field $\mathbf{K}$) that is a vector space with respect to vector addition and multiplication by a number defined in the vector space $\mathbf{X}$, is called a *subspace* of the vector space $\mathbf{X}$ and denoted $\mathbf{W} \leq \mathbf{X}$.

Proposition 1.2.1. The set $\mathbf{W}$ of vectors of the vector space $\mathbf{X}$ is a subspace of the vector space $\mathbf{X}$ iff for each two vectors $\mathbf{x}, \mathbf{y} \in \mathbf{W}$ and each number $\lambda \in \mathbf{K}$ vectors $\mathbf{x} + \mathbf{y}$ and $\lambda \mathbf{x}$ belong to the set $\mathbf{W}$.

Proof. Necessity is obvious. To prove sufficiency, we have to show that in our case conditions 1-8 for a vector space are satisfied. Let us check condition 1. Let $\mathbf{x}, \mathbf{y} \in \mathbf{W} \subset \mathbf{X}$. By assumption, $\mathbf{x} + \mathbf{y} \in \mathbf{W} \subset \mathbf{X}$. As $\mathbf{X}$ is a vector space, then for $\mathbf{X}$ axiom 1 is satisfied, and then $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$. Therefore, for $\mathbf{W}$ axiom 1 is satisfied, too. Let us test the validity of condition 4. Let $\mathbf{x} \in \mathbf{W} \subset \mathbf{X}$. By assumption, $(-1)\mathbf{x} \in \mathbf{W} \subset \mathbf{X}$. On the other hand, by proposition 1, in $\mathbf{X}$ the equality $(-1)\mathbf{x} = -\mathbf{x}$ holds. Hence the inverse vector $-\mathbf{x}$ belongs to set $\mathbf{W}$ with the vector $\mathbf{x}$, i.e., condition 4 is satisfied. Prove by yourselves the validity of conditions 2, 3 and 5-8. $\square$

Example 1.2.1. The vector space $\mathbf{C}[\alpha, \beta]$ over $\mathbf{R}$ of all functions continuous on $[\alpha, \beta]$ (example 1.1.3) is a subspace of vector space $\mathbf{F}[\alpha, \beta]$. As the sum of two functions continuous on the interval, and the product of such a function by a number are functions continuous on this interval, by proposition 1.2.1, $\mathbf{C}[\alpha, \beta]$ is a subspace of the vector space $\mathbf{F}[\alpha, \beta]$.

Example 1.2.2. Let $\mathbf{P}_n$ be the set of all polynomials $a_0 t^k + a_1 t^{k-1} + \dots + a_{k-1} t + a_k = \mathbf{x}$ ($k \leq n$) of at most degree n with real coefficients. We define addition of two polynomials and multiplication of a polynomial by a real number in the usual way. As a result, we get the vector space $\mathbf{P}_n$ of polynomials of at most degree n . If we denote by $\mathbf{P}_n[\alpha, \beta]$ the vector space of

polynomials of at most degree n defined on the interval $[\alpha, \beta]$, then $\mathbf{P}_n[\alpha, \beta]$ will be a subspace of the vector space $\mathbf{C}[\alpha, \beta]$.

Example 1.2.3.* Let us show that the set $\mathbf{H} = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} : a, b, c \in \mathbf{R} \right\}$ is a subspace of the matrix vector space $\mathbf{R}^{2 \times 2}$.

The set $\mathbf{H}$ is closed with respect to addition and multiplication by scalar since

$$\begin{bmatrix} a & b \\ 0 & c \end{bmatrix} + \begin{bmatrix} d & e \\ 0 & f \end{bmatrix} = \begin{bmatrix} a+d & b+e \\ 0 & c+f \end{bmatrix}$$

and

$$\alpha \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} = \begin{bmatrix} \alpha a & \alpha b \\ 0 & \alpha c \end{bmatrix}.$$

Thus the set $\mathbf{H}$ is a subspace of the matrix vector space $\mathbf{R}^{2 \times 2}$.

Problem 1.2.1.* Prove that the set of all symmetric matrices form a subspace of the vector space of all square matrices $\mathbf{R}^{n \times n}$.

Proposition 1.2.2. If $\mathbf{S}_1, \dots, \mathbf{S}_k$ are subspaces of the vector space $\mathbf{X}$, then the intersection $\mathbf{S} = \mathbf{S}_1 \cap \mathbf{S}_2 \cap \dots \cap \mathbf{S}_k$ of the subspaces is a subspace of the vector space $\mathbf{X}$.

Prove! $\square$

Proposition 1.2.3. If $\mathbf{S}_1, \dots, \mathbf{S}_k$ are subspaces of the space $\mathbf{X}$ and

$$\mathbf{S} = \{\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_k : \mathbf{x}_i \in \mathbf{S}_i \text{ (} i = 1 : k \text{)}\}$$

is the sum of these subspaces, then $\mathbf{S}$ is a subspace of $\mathbf{X}$.

Definition 1.2.2. If each $\mathbf{x} \in \mathbf{S}$ can be expressed uniquely in the form $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_k$ ($\mathbf{x}_i \in \mathbf{S}_i$), then we say that $\mathbf{S}$ is the *direct sum of subspaces* $\mathbf{S}_i$, and it denoted $\mathbf{S} = \mathbf{S}_1 \oplus \mathbf{S}_2 \oplus \dots \oplus \mathbf{S}_k$.

Definition 1.2.3. Each element of the space $\mathbf{X}$ that can be expressed as $\alpha_1 \mathbf{x}_1 + \dots + \alpha_n \mathbf{x}_n$, where $\alpha_i \in \mathbf{K}$, is called a *linear combination* of the elements $\mathbf{x}_1, \dots, \mathbf{x}_n$ of the vector space $\mathbf{X}$ (over the field $\mathbf{K}$).

Definition 1.2.4. The set of all possible linear combination of the set Z is called the *span of the set* $Z \subset \mathbf{X}$.

Example 1.2.4. Let $\mathbf{X} = \mathbf{R}^3$ and $Z = \{ [1; 1; 0]^T, [1; -1; 0]^T \}$. Then $\text{span } Z = \{ [\alpha; \beta; 0]^T : \alpha, \beta \in \mathbf{R} \}$. Prove!

Proposition 1.2.4. The set $\text{span } Z$ of the set $Z \subset \mathbf{X}$ is the least subspace that contains the set Z .

Proof. First, let us prove that $\text{span } Z$ is a subspace of the space $\mathbf{X}$. It is sufficient, by proposition 1.2.1, to show that $\text{span } Z$ is closed with respect to vector addition and multiplication of the vector by a number:

$$\mathbf{x}, \mathbf{y} \in \text{span } Z \Leftrightarrow \mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{u}_i \wedge \mathbf{y} = \sum_{j=1}^m \beta_j \mathbf{v}_j \wedge \alpha_i, \beta_j \in \mathbf{K} \wedge \mathbf{u}_i, \mathbf{v}_j \in Z \Rightarrow$$

$$\mathbf{x} + \mathbf{y} = \sum_{i=1}^n \alpha_i \mathbf{u}_i + \sum_{j=1}^m \beta_j \mathbf{v}_j \wedge \alpha_i, \beta_j \in \mathbf{K} \wedge \mathbf{u}_i, \mathbf{v}_j \in Z \Leftrightarrow \mathbf{x} + \mathbf{y} \in \text{span } Z ;$$

$$\lambda \in \mathbf{K} \wedge \mathbf{x} \in \text{span } Z \Leftrightarrow \lambda \in \mathbf{K} \wedge \mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{u}_i \wedge \mathbf{u}_i \in Z \wedge \alpha_i, \lambda \in \mathbf{K} \Rightarrow$$

$$\lambda \mathbf{x} = \lambda \sum_{i=1}^n \alpha_i \mathbf{u}_i = \sum_{i=1}^n (\lambda \alpha_i) \mathbf{u}_i = \sum_{i=1}^n \beta_i \mathbf{u}_i \wedge \beta_i \in \mathbf{K} \wedge \mathbf{u}_i \in Z \Leftrightarrow \lambda \mathbf{x} \in \text{span } Z .$$

Thus, $\text{span } Z$ is a subspace of the space $\mathbf{X}$. Let us show that $\text{span } Z$ is the least subspace of the space $\mathbf{X}$ that contains the set Z . Let $\mathbf{Y}$ be some subspace of the space $\mathbf{X}$ for which $Z \subset \mathbf{Y}$. As $Z \subset \mathbf{Y}$ and $\mathbf{Y}$ is a subspace, the arbitrary linear combination of the elements of the set Z belongs to the subspace $\mathbf{Y}$. Therefore, $\text{span } Z$ as the set of all such linear combinations belongs to the space $\mathbf{Y}$. $\square$

Corollary 1.2.1. A subset $\mathbf{W}$ of the vector space $\mathbf{X}$ is a subspace iff it coincides with its span, i.e., $\mathbf{W} \leq \mathbf{X} \Leftrightarrow \mathbf{W} = \text{span } \mathbf{W}$.

Problem 1.2.2.* Does the vector $\mathbf{d} = \begin{bmatrix} 8 & 7 & 4 \end{bmatrix}^T$ belong to the subspace $\text{span}\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, when

$$\mathbf{a} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}^T, \mathbf{b} = \begin{bmatrix} 2 & 3 & 1 \end{bmatrix}^T, \mathbf{c} = \begin{bmatrix} 6 & 9 & 3 \end{bmatrix}^T?$$

1.1.3 Linear Dependence of Vectors. Basis of the Vector Space.

Definition 1.3.1. A set of vectors

$$\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$$

in the vector space $\mathbf{X}$ (over the field $\mathbf{K}$) is said to be *linearly dependent* if

$$\exists \alpha_1, \dots, \alpha_k \in \mathbf{K} : |\alpha_1| + \dots + |\alpha_k| \neq 0 \wedge \alpha_1 \mathbf{x}_1 + \dots + \alpha_k \mathbf{x}_k = \mathbf{0}.$$

Definition 1.3.2. A set of vectors in the space $\mathbf{X}$ (over the field $\mathbf{K}$) is said to be *linearly independent* if it is not linearly dependent.

Example 1.3.1.* Let us check if the set $U = \{1 + x, x + x^2, 1 + x^2\}$ is linearly independent in the vector space $\mathbf{P}_n$ ($n \geq 2$) of all polynomials of at most degree n with real coefficients. Let us consider the equality

$$\alpha(1 + x) + \beta(x + x^2) + \gamma(1 + x^2) = 0.$$

It is well-known in algebra that a polynomial is identically null iff all its coefficients are zeros. Thus we get the system

$$\begin{cases} \alpha + \gamma = 0 \\ \alpha + \beta = 0 \\ \beta + \gamma = 0 \end{cases}.$$

This system has only a trivial solution. The set U is linearly independent.

Problem 1.3.1.* Prove that each set of vectors that contains the null vector is linearly dependent.

Problem 1.3.2.* Prove that if the column-vectors of a determinant are linearly dependent, then the determinant equals 0.

Definition 1.3.3. A subset $V = \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}\}$ of the set $U = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of vectors of the vector space $\mathbf{X}$ is called a *maximal linearly independent subset* if V is linearly independent and it is not a proper subset of any linearly independent subset of the set U .

Proposition 1.3.1. If V is a maximal linearly independent subset of the set U , then $\text{span } U = \text{span } V$.

Proof. As $V \subset U$, $\text{span } V \subset \text{span } U$, by the definition of the span. To prove our assertion, we have to show that $\text{span } V \supset \text{span } U$. Let, by antithesis, exist a vector $\mathbf{x}$ of the subspace $\text{span } U$ that does not belong to the subspace $\text{span } V$. Thus, the vector $\mathbf{x}$ cannot be expressed as a linear combination of vectors of V but can be expressed as a linear combination of vectors of U , when at least one vector $\mathbf{x}_j \in U$ is used, at which $\mathbf{x}_j \notin V$ and $\mathbf{x}_j$ is not expressible as a linear combination of vectors of V . Set $V \cup \{\mathbf{x}_j\} \subset U$ is linearly independent and contains the set V as a proper subset. Hence V is not the maximal linearly independent subset. We have got a contradiction to the assumption. Thus $\text{span } V \supset \text{span } U$, Q.E.D. $\square$

Definition 1.3.4. A set $B = \{\mathbf{x}_i\}_{i \in I}$ of vectors of the vector space $\mathbf{X}$ is called a *basis* of the vector space $\mathbf{X}$ if B is linearly independent and each vector $\mathbf{x}$ of the space $\mathbf{X}$ can be expressed as a linear combination of vectors of the set B , $\mathbf{x} = \sum_{i \in I} \alpha_i \mathbf{x}_i$, where coefficients α_i ($i = 1 : n$) are called *coordinates* of the vector $\mathbf{x}$ relative to the basis B .

Definition 1.3.5. If the number of vectors in the basis B of the vector space $\mathbf{X}$, i.e., the number of elements of the set I , is finite, then this number is called the *dimension of the vector space $\mathbf{X}$* and denoted $\dim \mathbf{X}$, and the space $\mathbf{X}$ is called a *finite-dimensional* or a *finite-dimensional vector space*. If the number of vectors in the basis B of the vector space $\mathbf{X}$ is infinite, then the vector space $\mathbf{X}$ is called *infinite-dimensional* or an *infinite-dimensional vector space*.

Proposition 1.3.2. A subset B of the vectors of the vector space $\mathbf{X}$ is a basis of the space iff it is the maximal linearly independent subset.

Example 1.3.2. Vectors

$$\mathbf{e}_k = \underbrace{[0; 0; \dots; 0]}_{k-1 \text{ zeros}}; \underbrace{1; 0; \dots; 0]}_{n-k \text{ zeros}}]^T \quad (k = 1 : n)$$

form a basis in space $\mathbf{R}^n$. Let us check the validity of the conditions in definition 1.3.4. As

$$\sum_{k=1}^n \alpha_k \mathbf{e}_k = \mathbf{0} \Leftrightarrow [\alpha_1, \dots, \alpha_n]^T = [0; \dots; 0]^T \Leftrightarrow \sum_{k=1}^n |\alpha_k| = 0,$$

the vector system $\{\mathbf{e}_k\}_{k=1:n}$ is linearly independent, and, due to

$$[\alpha_1, \dots, \alpha_n]^T = \sum_{k=1}^n \alpha_k \mathbf{e}_k,$$

an arbitrary vector of the space $\mathbf{R}^n$ can be expressed as a linear combination of vectors $\mathbf{e}_k$.

Problem 1.3.3. Vector system

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

forms a basis in space $\mathbf{R}^{2 \times 2}$.

Example 1.3.3. Vector system $\{1; t; t^2; \dots; t^n\}$ forms a basis in vector space $\mathbf{P}_n$ of polynomials of at most degree n . Truly, the set $\{1; t; t^2; \dots; t^n\}$ is linearly independent since

$$\mathbf{x} = a_0 t^n + a_1 t^{n-1} + \dots + a_{n-1} t + a_n = \mathbf{0} \Rightarrow a_k = 0 \quad (k = 1 : n)$$

and each vector of the space $\mathbf{P}_n$ (i.e., arbitrary polynomial of at most degree n) can be expressed in the form

$$\mathbf{x} = a_0 t^n + a_1 t^{n-1} + \dots + a_{n-1} t + a_n.$$

Definition 1.3.6. Two vector spaces $\mathbf{X}$ and $\mathbf{X}'$ are called *isomorphic*, if there exist a one-to-one correspondence between the spaces $\varphi : \mathbf{X} \rightarrow \mathbf{X}'$, such that

$$\begin{aligned} 1) \forall \mathbf{x}, \mathbf{y} \in \mathbf{X} \quad & \varphi(\mathbf{x} + \mathbf{y}) = \varphi(\mathbf{x}) + \varphi(\mathbf{y}); \\ 2) \forall \mathbf{x} \in \mathbf{X}, \forall \alpha \in \mathbf{K} \quad & \varphi(\alpha \mathbf{x}) = \alpha \varphi(\mathbf{x}). \end{aligned}$$

Proposition 1.3.3. All vector spaces (over the same number field $\mathbf{K}$) of the same dimension are isomorphic.

1.1.4 Scalar Product

Definition 1.4.1. A vector space $\mathbf{X}$ over the field $\mathbf{K}$ is called a *space with scalar product* if to each pair of elements $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ there corresponds a certain number $\langle \mathbf{x}, \mathbf{y} \rangle \in \mathbf{K}$, called the *scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$* , such that following condition (the axioms of scalar product) are satisfied:

$$1. \langle \mathbf{x}, \mathbf{x} \rangle \geq 0; \quad \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Rightarrow \mathbf{x} = \mathbf{0};$$

2. $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$, when $\overline{\langle \mathbf{y}, \mathbf{x} \rangle}$ is the conjugate complex number of $\langle \mathbf{x}, \mathbf{y} \rangle$;

3. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ (additivity with respect to the first factor);

4. $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$ (homogeneity with respect to the first factor).

If $\mathbf{X}$ is a vector space over $\mathbf{R}$, then, by the definition, $\langle \mathbf{x}, \mathbf{y} \rangle \in \mathbf{R}$, and condition 1 acquires the form $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$, i.e., in this case scalar product is commutative.

Example 1.4.1. Let us define in $\mathbf{C}^n$ the scalar product of vectors

$$\mathbf{x} = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix}^T \wedge \mathbf{y} = \begin{bmatrix} \eta_1 & \cdots & \eta_n \end{bmatrix}^T$$

by the formula

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n \xi_k \overline{\eta_k}.$$

Let us check the validity of conditions 1-4: $\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{k=1}^n \xi_k \overline{\xi_k} = \sum_{k=1}^n |\xi_k|^2 \geq 0$,

$$\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{k=1}^n |\xi_k|^2 = 0 \Rightarrow \xi_k = 0 \quad (k = 1 : n) \Leftrightarrow \mathbf{x} = \mathbf{0};$$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n \xi_k \overline{\eta_k} = \overline{\sum_{k=1}^n \overline{\xi_k} \eta_k} = \overline{\sum_{k=1}^n \eta_k \overline{\xi_k}} = \overline{\langle \mathbf{y}, \mathbf{x} \rangle};$$

$$\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \sum_{k=1}^n (\xi_k + \eta_k) \overline{\zeta_k} = \sum_{k=1}^n \xi_k \overline{\zeta_k} + \sum_{k=1}^n \eta_k \overline{\zeta_k} = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle;$$

$$\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n \lambda \xi_k \overline{\eta_k} = \lambda \sum_{k=1}^n \xi_k \overline{\eta_k} = \lambda \langle \mathbf{x}, \mathbf{y} \rangle.$$

Example 1.4.2. Let us consider the vector space $\mathbf{L}_2[\alpha, \beta]$ of all functions integrable (in Lebesgue's sense) on the interval $[\alpha, \beta]$. We define the scalar product for such functions by the formula

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_{\alpha}^{\beta} \mathbf{x}(t) \overline{\mathbf{y}(t)} dt.$$

Verify that all the axioms 1-4 of scalar product are satisfied.

Proposition 1.4.1. Scalar product $\langle \mathbf{x}, \mathbf{y} \rangle$ has the following properties:

1. $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ (additivity with respect to the second factor);

2. $\langle \mathbf{x}, \lambda \mathbf{y} \rangle = \overline{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle$ (conjugate homogeneity with respect to the second factor);

3. $\langle \mathbf{x}, \mathbf{0} \rangle = \langle \mathbf{0}, \mathbf{y} \rangle = 0 \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{X}$;

$$4. \langle \lambda \mathbf{x}, \lambda \mathbf{y} \rangle = |\lambda|^2 \langle \mathbf{x}, \mathbf{y} \rangle.$$

Let us *prove* these assertions:

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle &= \overline{\langle \mathbf{y} + \mathbf{z}, \mathbf{x} \rangle} = \overline{\langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{z}, \mathbf{x} \rangle} = \overline{\langle \mathbf{y}, \mathbf{x} \rangle} + \overline{\langle \mathbf{z}, \mathbf{x} \rangle} = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle; \\ \langle \mathbf{x}, \lambda \mathbf{y} \rangle &= \overline{\langle \lambda \mathbf{y}, \mathbf{x} \rangle} = \overline{\lambda \langle \mathbf{y}, \mathbf{x} \rangle} = \overline{\lambda} \overline{\langle \mathbf{y}, \mathbf{x} \rangle} = \overline{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle; \langle \mathbf{x}, \mathbf{0} \rangle = \langle \mathbf{x}, 0\mathbf{x} \rangle = \overline{0} \langle \mathbf{x}, \mathbf{x} \rangle = 0; \\ \langle \lambda \mathbf{x}, \lambda \mathbf{y} \rangle &= \lambda \overline{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle = |\lambda|^2 \langle \mathbf{x}, \mathbf{y} \rangle. \quad \square \end{aligned}$$

Proposition 1.4.2 (Cauchy-Schwartz inequality). For arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ of the vector space with scalar product $\mathbf{X}$ it holds the inequality

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

Proof. If $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, then, by the definition of the scalar product (condition 1) the inequality holds. Now let us consider the case $\langle \mathbf{x}, \mathbf{y} \rangle \neq 0$. We define an auxiliary function

$$\varphi(\lambda) = \langle \mathbf{x} + \lambda \langle \mathbf{x}, \mathbf{y} \rangle \mathbf{y}, \mathbf{x} + \lambda \langle \mathbf{x}, \mathbf{y} \rangle \mathbf{y} \rangle.$$

As for $\lambda \in \mathbf{R}$

$$\begin{aligned} \varphi(\lambda) &= \langle \mathbf{x}, \mathbf{x} \rangle + \lambda \overline{\langle \mathbf{x}, \mathbf{y} \rangle} \langle \mathbf{x}, \mathbf{y} \rangle + \lambda \langle \mathbf{x}, \mathbf{y} \rangle \langle \mathbf{y}, \mathbf{x} \rangle + \lambda^2 |\langle \mathbf{x}, \mathbf{y} \rangle|^2 \langle \mathbf{y}, \mathbf{y} \rangle = \\ &= \lambda^2 |\langle \mathbf{x}, \mathbf{y} \rangle|^2 \langle \mathbf{y}, \mathbf{y} \rangle + 2\lambda |\langle \mathbf{x}, \mathbf{y} \rangle|^2 + \langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \quad \forall \lambda \in \mathbf{R} \Leftrightarrow \\ &\Leftrightarrow |\langle \mathbf{x}, \mathbf{y} \rangle|^4 - |\langle \mathbf{x}, \mathbf{y} \rangle|^2 \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle \leq 0. \end{aligned}$$

The last inequality is equivalent to the inequality $|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle$, and this — to the Cauchy-Schwartz inequality. The Cauchy-Schwartz inequality makes it possible to define the angle between two vectors by the scalar product.

Definition 1.4.2. The angle between arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ of the vector space with scalar product $\mathbf{X}$ is defined by the formula

$$\cos(\widehat{\mathbf{x}, \mathbf{y}}) = \langle \mathbf{x}, \mathbf{y} \rangle / (\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}).$$

Problem 1.4.1.* Show that for each two complex vectors $\mathbf{x}$ and $\mathbf{y}$ the equality

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

holds.

Problem 1.4.2.* The scalar product in the vector space $\mathbf{P}_n[\alpha, \beta]$ of polynomials of at most degree n with real coefficients on $[\alpha, \beta]$ is defined by the formula

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_{\alpha}^{\beta} \mathbf{x}(t) \mathbf{y}(t) dt.$$

Find the angle between the polynomials $\mathbf{x} = t - 1$ and $\mathbf{y} = t^2 + 1$.

1.1.5 Norm of a Vector

Definition 1.5.1. A vector space $\mathbf{X}$ (over the number field $\mathbf{K}$) is called a *normed space*, if to each vector $\mathbf{x} \in \mathbf{X}$ there corresponds a certain non-negative real number $\|\mathbf{x}\|$, called the *norm of the vector*, such that the following conditions are satisfied:

1. $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ (identity axiom);
2. $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$ (homogeneity axiom);
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

Definition 1.5.2. The *distance* $\rho(\mathbf{x}, \mathbf{y})$ between two vectors in the normed space $\mathbf{X}$ is defined by the formula $\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

Proposition 1.5.1 (Hölder inequality). If $1 < p < \infty$, $1/p + 1/q = 1$,

$$\mathbf{x} = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix}^T \in \mathbf{C}^n \wedge \mathbf{y} = \begin{bmatrix} \eta_1 & \cdots & \eta_n \end{bmatrix}^T \in \mathbf{C}^n,$$

then

$$\sum_{k=1}^n |\xi_k \eta_k| \leq \left(\sum_{k=1}^n |\xi_k|^p \right)^{1/p} \left(\sum_{k=1}^n |\eta_k|^q \right)^{1/q}.$$

Proof. See E.Oja, P.Oja (1991, pp. 11-12).

Proposition 1.5.2 (Minkowski inequality). If $1 \leq p < \infty$,

$$\mathbf{x} = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix}^T \in \mathbf{C}^n \wedge \mathbf{y} = \begin{bmatrix} \eta_1 & \cdots & \eta_n \end{bmatrix}^T \in \mathbf{C}^n,$$

then

$$\left(\sum_{k=1}^n |\xi_k + \eta_k|^p \right)^{1/p} \leq \left(\sum_{k=1}^n |\xi_k|^p \right)^{1/p} + \left(\sum_{k=1}^n |\eta_k|^p \right)^{1/p}.$$

Proof. See E.Oja, P.Oja (1991, pp. 10-11).

Example 1.5.1. One defines in $\mathbf{C}^n$ the p -norm ($1 \leq p \leq \infty$) of the vector $\mathbf{x}$ by the formulas

$$\|\mathbf{x}\|_p = (|\xi_1|^p + \cdots + |\xi_n|^p)^{1/p} \quad (1 \leq p < \infty),$$

$$\|\mathbf{x}\|_{\infty} = \max_{1 \leq k \leq n} |\xi_k|.$$

Let us verify that the p -norm ($1 \leq p < \infty$) satisfies the conditions 1-3 in definition 1.5.1:

$$\|\mathbf{x}\|_p = (|\xi_1|^p + \dots + |\xi_n|^p)^{1/p} = 0 \Leftrightarrow \xi_k = 0 \quad (1 \leq k \leq n) \Leftrightarrow \mathbf{x} = \mathbf{0};$$

$$\|\lambda \mathbf{x}\|_p = (|\lambda \xi_1|^p + \dots + |\lambda \xi_n|^p)^{1/p} = [|\lambda|^p (|\xi_1|^p + \dots + |\xi_n|^p)]^{1/p} =$$

$$= |\lambda| (|\xi_1|^p + \dots + |\xi_n|^p)^{1/p} = |\lambda| \|\mathbf{x}\|_p;$$

using the Minkowski inequality, we get

$$\|\mathbf{x} + \mathbf{y}\|_p = \left(\sum_{k=1}^n |\xi_k + \eta_k|^p \right)^{1/p} \leq \left(\sum_{k=1}^n |\xi_k|^p \right)^{1/p} + \left(\sum_{k=1}^n |\eta_k|^p \right)^{1/p} = \|\mathbf{x}\|_p + \|\mathbf{y}\|_p.$$

Verify conditions 1-3 in case of the norm $\|\cdot\|_{\infty}$!

The most often used p -norms are:

$$\|\mathbf{x}\|_1 = |\xi_1| + \dots + |\xi_n|;$$

$$\|\mathbf{x}\|_2 = (|\xi_1|^2 + \dots + |\xi_n|^2)^{1/2};$$

$$\|\mathbf{x}\|_{\infty} = \max_{1 \leq k \leq n} |\xi_k|.$$

Problem 1.5.1.* Let be given the vectors $\mathbf{u} = \begin{bmatrix} 1 & -1 & 3 \end{bmatrix}^T$ ja $\mathbf{v} = \begin{bmatrix} 0 & 3 & 2 \end{bmatrix}^T$. Find

$$\|\mathbf{u} + \mathbf{v}\|_1, \|\mathbf{u} + \mathbf{v}\|_2, \|\mathbf{u} + \mathbf{v}\|_{\infty}, \|\mathbf{u}\|_1 + \|\mathbf{v}\|_1, \|\mathbf{u}\|_2 + \|\mathbf{v}\|_2, \|\mathbf{u}\|_{\infty} + \|\mathbf{v}\|_{\infty},$$

$$\| -5\mathbf{u} \|_1 + 5 \|\mathbf{v}\|_1, \| -5\mathbf{u} \|_2 + 5 \|\mathbf{v}\|_2, \| -5\mathbf{u} \|_{\infty} + 5 \|\mathbf{v}\|_{\infty}.$$

Proposition 1.5.3. All the p -norms of the space $\mathbf{C}^n$ are equivalent, i.e., if $\|\cdot\|_{\alpha}$ and $\|\cdot\|_{\beta}$ are the p -norms of the space $\mathbf{C}^n$, then there exist positive constants c_1 and c_2 , such that

$$c_1 \|\mathbf{x}\|_{\alpha} \leq \|\mathbf{x}\|_{\beta} \leq c_2 \|\mathbf{x}\|_{\alpha} \quad \forall \mathbf{x} \in \mathbf{C}^n.$$

At the same time

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2,$$

$$\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_2 \leq \sqrt{n} \|\mathbf{x}\|_{\infty},$$

$$\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_1 \leq n \|\mathbf{x}\|_{\infty}.$$

Let us *prove* the last three assertions:

$$\begin{aligned}\|\mathbf{x}\|_2 &= (|\xi_1|^2 + \dots + |\xi_n|^2)^{1/2} \leq \left(\sum_{i=1}^n \sum_{j=1}^n |\xi_i| |\xi_j| \right)^{1/2} = \\ &= \left(\left(\sum_{k=1}^n |\xi_k| \right)^2 \right)^{1/2} = \|\mathbf{x}\|_1 ;\end{aligned}$$

Using the Hölder inequality, we get in case $p = q = 2$ that

$$\begin{aligned}\|\mathbf{x}\|_1 &= |\xi_1| + \dots + |\xi_n| = 1 \cdot |\xi_1| + \dots + 1 \cdot |\xi_n| = |1 \cdot \xi_1| + \dots + |1 \cdot \xi_n| \leq \\ &\leq (1^2 + \dots + 1^2)^{1/2} (\xi_1^2 + \dots + \xi_n^2)^{1/2} = \sqrt{n} \|\mathbf{x}\|_2 ; \\ \|\mathbf{x}\|_\infty &= \max_{1 \leq k \leq n} |\xi_k| = \left(\left(\max_{1 \leq k \leq n} |\xi_k| \right)^2 \right)^{1/2} \leq (\xi_1^2 + \dots + \xi_n^2)^{1/2} = \|\mathbf{x}\|_2 ; \\ \|\mathbf{x}\|_2 &= (\xi_1^2 + \dots + \xi_n^2)^{1/2} \leq \left(\left(\max_{1 \leq k \leq n} |\xi_k| \right)^2 + \dots + \left(\max_{1 \leq k \leq n} |\xi_k| \right)^2 \right)^{1/2} = \\ &= (n \left(\max_{1 \leq k \leq n} |\xi_k| \right)^2)^{1/2} = \sqrt{n} \|\mathbf{x}\|_\infty ; \\ \|\mathbf{x}\|_\infty &= \max_{1 \leq k \leq n} |\xi_k| \leq |\xi_1| + \dots + |\xi_n| \leq n \max_{1 \leq k \leq n} |\xi_k| = n \|\mathbf{x}\|_\infty . \quad \square\end{aligned}$$

Proposition 1.5.4. A space with scalar product $\mathbf{X}$ is a normed space with the norm

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

Proof. Let us verify the validity of conditions 1-3:

$$\begin{aligned}\|\mathbf{x}\| &= 0 \Leftrightarrow \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = 0 \Leftrightarrow \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}; \\ \|\lambda \mathbf{x}\| &= \sqrt{\langle \lambda \mathbf{x}, \lambda \mathbf{x} \rangle} = \sqrt{|\lambda|^2 \langle \mathbf{x}, \mathbf{x} \rangle} = |\lambda| \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |\lambda| \|\mathbf{x}\| ; \\ \|\mathbf{x} + \mathbf{y}\| &= \sqrt{\langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle} = \\ &= \sqrt{\|\mathbf{x}\|^2 + \langle \mathbf{x}, \mathbf{y} \rangle + \overline{\langle \mathbf{x}, \mathbf{y} \rangle} + \|\mathbf{y}\|^2} = \sqrt{\|\mathbf{x}\|^2 + 2\Re \langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2} \leq \\ &\leq \sqrt{\|\mathbf{x}\|^2 + 2|\langle \mathbf{x}, \mathbf{y} \rangle| + \|\mathbf{y}\|^2} \leq \sqrt{\|\mathbf{x}\|^2 + 2\|\mathbf{x}\| \|\mathbf{y}\| + \|\mathbf{y}\|^2} \leq \\ &\leq \sqrt{(\|\mathbf{x}\| + \|\mathbf{y}\|)^2} = \|\mathbf{x}\| + \|\mathbf{y}\| .\end{aligned}$$

$\Re(\mathbf{x}, \mathbf{y})$ is the notation for the real part of the complex number $\langle \mathbf{x}, \mathbf{y} \rangle$.

Proposition 1.5.5. In the normed space with scalar product the *parallelogram rule*:

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2).$$

holds.

Proof. By the immediate check, we get

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle + \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle = \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle - \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \\ &= 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2). \end{aligned}$$

Definition 1.5.3. It is said that the sequence $\{\mathbf{x}^{(k)}\}$ of the elements of the space $\mathbf{C}^n$ converges with respect to the p -norm to the element $\mathbf{x} \in \mathbf{C}^n$ if

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_p = 0.$$

In this case we shall write $\mathbf{x}^{(k)} \rightarrow \mathbf{x}$.

Remark 1.5.1. Since all the p -norms of the space $\mathbf{C}^n$ are equivalent, this implies that the convergence of the sequence $\{\mathbf{x}^{(k)}\}$ with respect to the α -norm will yield its convergence with respect to the β -norm.

Problem 1.5.2. Show that if $\mathbf{x} \in \mathbf{C}^n$, then $\lim_{p \rightarrow \infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty$.

Problem 1.5.3. Show that if $\mathbf{x} \in \mathbf{C}^n$, then

$$\|\mathbf{x}\|_p \leq c(\|\Re \mathbf{x}\|_p + \|\Im \mathbf{x}\|_p),$$

where $\mathbf{x} = \Re \mathbf{x} + i\Im \mathbf{x}$ and $\Re \mathbf{x}, \Im \mathbf{x} \in \mathbf{R}^n$. Find such a constant c_n , that

$$c_n(\|\Re \mathbf{x}\|_2 + \|\Im \mathbf{x}\|_2) \leq \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbf{C}^n.$$

Definition 1.5.4. A vector $\hat{\mathbf{x}} \in \mathbf{R}^n$ is called an *approximation* to the vector $\mathbf{x} \in \mathbf{R}^n$ if it differs little from $\mathbf{x}$ in some sense.

Definition 1.5.5. In case of the fixed norm $\|\cdot\|$, the quantity

$$\varepsilon_{abs} = \|\hat{\mathbf{x}} - \mathbf{x}\|$$

is called the *absolute error* of the approximation $\hat{\mathbf{x}}$ to the vector $\mathbf{x}$, and the quantity

$$\varepsilon_{rel} = \|\hat{\mathbf{x}} - \mathbf{x}\| / \|\mathbf{x}\|$$

is called the *relative error* of the approximation ($\mathbf{x} \neq \mathbf{0}$).

In case of the ∞ -norm the relative error can be considered as an index of the correct significant digits. Namely, if $\|\hat{\mathbf{x}} - \mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty \approx 10^{-k}$, then the greatest component of the vector $\hat{\mathbf{x}}$ has k correct significant digits.

Example 1.5.2. Let $\mathbf{x} = [2.543; 0.06356]^T$ and $\hat{\mathbf{x}} = [2.541; 0.06937]^T$. Find ε_{abs} and ε_{rel} , and then the number of the correct significant digits of the greatest component of the approximation $\hat{\mathbf{x}}$ by ε_{rel} . We get $\hat{\mathbf{x}} - \mathbf{x} = [-0.002; 0.00581]^T$, $\varepsilon_{abs} = \|\hat{\mathbf{x}} - \mathbf{x}\|_\infty = 0.00581$ and $\|\mathbf{x}\|_\infty = 2.543$ and $\varepsilon_{rel} \approx 0.0023 \approx 10^{-3} \Rightarrow k = 3$. Thus the greatest component $\hat{\xi}_1$ of $\hat{\mathbf{x}}$ has three correct significant digits. At the same time, the component $\hat{\xi}_2$ has only one correct significant digit.

1.1.6 Orthogonal Vectors

Definition 1.6.1. The vectors $\mathbf{x}$ and $\mathbf{y}$ of the vector space with scalar product $\mathbf{X}$ are called *orthogonal* if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. We write $\mathbf{x} \perp \mathbf{y}$ to indicate the orthogonality of vectors $\mathbf{x}$ and $\mathbf{y}$. A vector $\mathbf{x}$ of the vector space $\mathbf{X}$ is called *orthogonal to the set* $Y \subset \mathbf{X}$ if $\mathbf{x} \perp \mathbf{y} \quad \forall \mathbf{y} \in Y$.

Problem 1.6.1.* Find all vectors that are orthogonal both to the vector $\mathbf{a} = [4 \ 0 \ 6 \ -2 \ 0]^T$ and $\mathbf{b} = [2 \ 1 \ -1 \ 1 \ 1]^T$.

Definition 1.6.2. The sets Y and Z of the vector space $\mathbf{X}$ are called *orthogonal* if $\mathbf{y} \perp \mathbf{z} \quad \forall \mathbf{y} \in Y \text{ and } \forall \mathbf{z} \in Z$.

Definition 1.6.3. A sequence $\{\mathbf{x}^{(k)}\}$ of vectors of the vector space with scalar product $\mathbf{X}$ is called a *Cauchy sequence* if for any $\epsilon > 0$ there is a natural number n_0 such that for all $m \in N$ and $n > n_0$

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(n+m)}\| = \sqrt{\langle \mathbf{x}^{(n)} - \mathbf{x}^{(n+m)}, \mathbf{x}^{(n)} - \mathbf{x}^{(n+m)} \rangle} < \epsilon.$$

Definition 1.6.4. A vector space with scalar product $\mathbf{X}$ is called *complete* if every Cauchy sequence is convergent to a point of the space $\mathbf{X}$.

Definition 1.6.5. A vector space with complex scalar product is called a *Hilbert space* $\mathbf{H}$ if it turns out to be complete with respect to the convergence by the norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Proposition 1.6.1. The space $\mathbf{C}^n$ with the scalar product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n \xi_k \bar{\eta}_k$ is a Hilbert space.

Proposition 1.6.2. The space $\mathbf{L}_2[\alpha, \beta]$ of all square-integrable functions on the interval $[\alpha, \beta]$ with the scalar product $\langle \mathbf{x}, \mathbf{y} \rangle = \int_{\alpha}^{\beta} x(t) \overline{y(t)} dt$ is a Hilbert space.

Proposition 1.6.3. Orthogonality of vectors in the vector space with scalar product $\mathbf{X}$ has the following properties (1-4):

1. $\mathbf{x} \perp \mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{0}$;
2. $\mathbf{x} \perp \mathbf{y} \Leftrightarrow \mathbf{y} \perp \mathbf{x}$;
3. $\mathbf{x} \perp \{\mathbf{y}_1, \dots, \mathbf{y}_k\} \Rightarrow \mathbf{x} \perp (\mathbf{y}_1 + \dots + \mathbf{y}_k)$;
4. $\mathbf{x} \perp \mathbf{y} \Rightarrow \mathbf{x} \perp \lambda \mathbf{y} \quad \forall \lambda \in \mathbf{K}$;

orthogonality of vectors in a Hilbert space has an additional property:

5. $\mathbf{x} \perp \mathbf{y}_n \quad (n = 1; 2; 3; \dots) \wedge \mathbf{y}_n \rightarrow \mathbf{y} \Rightarrow \mathbf{x} \perp \mathbf{y}$.

Let us *prove* these assertions:

$$\mathbf{x} \perp \mathbf{x} \Leftrightarrow \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0};$$

$$\mathbf{x} \perp \mathbf{y} \Leftrightarrow \langle \mathbf{x}, \mathbf{y} \rangle = 0 \Leftrightarrow \overline{\langle \mathbf{y}, \mathbf{x} \rangle} = 0 \Leftrightarrow \langle \mathbf{y}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{y} \perp \mathbf{x};$$

$$\mathbf{x} \perp \{\mathbf{y}_1, \dots, \mathbf{y}_k\} \Leftrightarrow \mathbf{x} \perp \mathbf{y}_1 \wedge \dots \wedge \mathbf{x} \perp \mathbf{y}_k \Leftrightarrow \langle \mathbf{x}, \mathbf{y}_1 \rangle = 0 \wedge \dots \wedge \langle \mathbf{x}, \mathbf{y}_k \rangle = 0$$

$\Rightarrow$

$$\Rightarrow \langle \mathbf{x}, \mathbf{y}_1 \rangle + \dots + \langle \mathbf{x}, \mathbf{y}_k \rangle = 0 \Leftrightarrow \langle \mathbf{x}, \mathbf{y}_1 + \dots + \mathbf{y}_k \rangle = 0 \Leftrightarrow \mathbf{x} \perp (\mathbf{y}_1 + \dots + \mathbf{y}_k);$$

$$\mathbf{x} \perp \mathbf{y} \Leftrightarrow \langle \mathbf{x}, \mathbf{y} \rangle = 0 \Leftrightarrow \overline{\lambda \langle \mathbf{x}, \mathbf{y} \rangle} = 0 \quad \forall \lambda \in \mathbf{K} \Leftrightarrow$$

$$\Leftrightarrow \langle \mathbf{x}, \lambda \mathbf{y} \rangle = 0 \quad \forall \lambda \in \mathbf{K} \Leftrightarrow \mathbf{x} \perp \lambda \mathbf{y};$$

$$\mathbf{x} \perp \mathbf{y}_n \quad \forall n \in \mathbf{N} \wedge \mathbf{y}_n \rightarrow \mathbf{y} \Leftrightarrow \langle \mathbf{x}, \mathbf{y}_n \rangle = 0 \wedge \|\mathbf{y}_n - \mathbf{y}\| \rightarrow 0 \Rightarrow$$

$$\Rightarrow \langle \mathbf{x}, \mathbf{y}_n \rangle = 0 \wedge |\langle \mathbf{x}, \mathbf{y}_n \rangle - \langle \mathbf{x}, \mathbf{y} \rangle| = |\langle \mathbf{x}, \mathbf{y}_n - \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}_n - \mathbf{y}\| \rightarrow 0 \Rightarrow$$

$$\Rightarrow \langle \mathbf{x}, \mathbf{y} \rangle = 0 \Leftrightarrow \mathbf{x} \perp \mathbf{y}.$$

Definition 1.6.6. The *orthogonal complement* of the set $Y \subset \mathbf{X}$ is the set $Y^\perp$ of all vectors of the space $\mathbf{X}$ that are orthogonal to the set Y , i.e.,

$$Y^\perp = \{\mathbf{x} : (\mathbf{x} \in \mathbf{X}) \wedge (\mathbf{x} \perp \mathbf{y} \quad \forall \mathbf{y} \in Y)\}.$$

Problem 1.6.2.* Let $U = \text{span} \left\{ \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T, \begin{bmatrix} 0 & 2 & 1 \end{bmatrix}^T \right\} \subset \mathbf{R}^3$. Find the orthogonal complement of the set U .

Proposition 1.6.4. If $\mathbf{X}$ is a vector space with scalar product, $\mathbf{x} \in \mathbf{X}$, $Y \subset \mathbf{X}$ and $\mathbf{x} \perp Y$, then $\mathbf{x} \perp \text{span } Y$. If, in addition, $\mathbf{X}$ is complete, i.e., is a Hilbert space, then $\mathbf{x} \perp \overline{\text{span } Y}$.

Proof. By assertions 3 and 4 of proposition 1.6.3, $\mathbf{x} \perp \text{span } Y$. If $\mathbf{y} \in \overline{\text{span } Y}$, i.e., $\exists \mathbf{y}_n \in \text{span } Y$ such that $\mathbf{y}_n \rightarrow \mathbf{y}$, then, due to the orthogonality $\mathbf{x} \perp \mathbf{y}_n$ and assertion 5 of proposition 1.6.3, we get $\mathbf{x} \perp \mathbf{y}$, i.e., $\mathbf{x} \perp \overline{\text{span } Y}$.

Proposition 1.6.5. The orthogonal complement $Y^\perp$ of the set $Y \subset \mathbf{X}$ is a subspace of the space $\mathbf{X}$. The orthogonal complement $Y^\perp$ of the set $Y \subset \mathbf{H}$ is a closed subspace of the Hilbert space $\mathbf{H}$, i.e., $Y^\perp$ is a subspace of the space $\mathbf{H}$ that contains all its boundary points.

Proof. Due to the proposition 1.2.1, it is sufficient for the proof of the first assertion of proposition 1.6.5 to show that $Y^\perp$ is closed with respect to vector addition and scalar multiplication. It will follow from assertion 5 of the same proposition, it holds the second assertion of proposition 1.6.5 too.

Proposition 1.6.6. If Y is a closed subspace of the Hilbert space $\mathbf{H}$, then each $\mathbf{x} \in \mathbf{H}$ can be expressed uniquely as the sum $\mathbf{x} = \mathbf{y} + \mathbf{z}$, $\mathbf{y} \in Y$, $\mathbf{z} \in Y^\perp$.

Corollary 1.6.1. If Y is a closed subspace of the Hilbert space, then the space $\mathbf{H}$ can be presented as the direct sum $\mathbf{H} = Y \oplus Y^\perp$ of the closed subspaces Y and $Y^\perp$, and $(Y^\perp)^\perp = Y$.

Definition 1.6.7. The *distance* of the vector $\mathbf{x}$ of the Hilbert space $\mathbf{H}$ from the *subspace* $Y \subset \mathbf{H}$ is defined by the formula

$$\rho(\mathbf{x}, Y) = \inf_{\mathbf{y} \in Y} \|\mathbf{x} - \mathbf{y}\|.$$

Proposition 1.6.7. If Y is a closed subspace of the Hilbert space $\mathbf{H}$ and $\mathbf{x} \in \mathbf{H}$, then there exists a uniquely defined $\mathbf{y} \in Y$ such that $\|\mathbf{x} - \mathbf{y}\| = \rho(\mathbf{x}, Y)$.

Definition 1.6.8. The vector $\mathbf{y}$ in proposition 1.6.7 is called the *orthogonal projection* of $\mathbf{x}$ onto the subspace Y .

Definition 1.6.9. A *vector system* $S = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is called *orthogonal* if $(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i\|^2 \delta_{ij}$, where δ_{ij} is the Kronecker delta. The vector system $S = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is called *orthonormal* if $(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}$.

Example 1.6.1. The vector system $\{\mathbf{e}_k\}$ ($k = 1 : n$), where $\mathbf{e}_k = [0; 0; \dots; 0; 1; 0; \dots; 0]^T$, is orthonormal in $\mathbf{C}^n$.
 $k-1 \text{ zeros}$ $n-k \text{ zeros}$

Example 1.6.2. The vector system

$$\{1/\sqrt{2\pi}, (\cos t)/\sqrt{\pi}, (\sin t)/\sqrt{\pi}, (\cos 2t)/\sqrt{\pi}, (\sin 2t)/\sqrt{\pi}, \dots\}$$

is orthonormal in $L_2[-\pi, \pi]$.

Example 1.6.3. The vector system $\{\exp(i2\pi kt)\}_{k \in \mathbf{Z}}$ is orthonormal in $L_2[0; 1]$. Truly,

$$\begin{aligned} (\mathbf{x}_k, \mathbf{x}_j) &= \int_0^1 \exp(i2\pi kt) \overline{\exp(i2\pi jt)} dt = \int_0^1 \exp(i2\pi(k-j)t) dt = \\ &= \begin{cases} (\exp(i2\pi(k-j)) - 1)/(i2\pi(k-j)) = 0, & \text{kui } k \neq j; \\ 1, & \text{as } k = j. \end{cases} \end{aligned}$$

Proposition 1.6.8. (Gram-Schmidt orthogonalization theorem). If $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is a linearly independent vector system in the vector space with scalar product $\mathbf{H}$, then there exists an orthonormal system $\{\varepsilon_1, \dots, \varepsilon_k\}$ such that $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} = \text{span}\{\varepsilon_1, \dots, \varepsilon_k\}$.

Let us *prove* this assertion by complete induction. In the case $k = 1$, we define $\varepsilon_1 = \mathbf{x}_1 / \|\mathbf{x}_1\|$, and, obviously, $\text{span}\{\mathbf{x}_1\} = \text{span}\{\varepsilon_1\}$. So we have shown the existence of the induction base. We have to show the admissibility of the induction step. Let us assume that the proposition holds for $k = i - 1$, i.e., there exists an orthonormal system $\{\varepsilon_1, \dots, \varepsilon_{i-1}\}$ such that $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_{i-1}\} = \text{span}\{\varepsilon_1, \dots, \varepsilon_{i-1}\}$. Now we consider the vector

$$\mathbf{y}_i = \lambda_1 \varepsilon_1 + \dots + \lambda_{i-1} \varepsilon_{i-1} + \mathbf{x}_i, \quad \lambda_j \in \mathbf{K}.$$

Let us choose the coefficients λ_ν ($\nu = 1: i-1$) so that $\mathbf{y}_i \perp \varepsilon_\nu$ ($\nu = 1: i-1$), i.e., $(\mathbf{y}_i, \varepsilon_\nu) = 0$. We get $i - 1$ conditions:

$$\lambda_\nu (\varepsilon_\nu, \varepsilon_\nu) + (\mathbf{x}_i, \varepsilon_\nu) = 0, \text{ ehk } \lambda_\nu = -(\mathbf{x}_i, \varepsilon_\nu) \quad (\nu = 1: i-1).$$

Thus,

$$\mathbf{y}_i = \mathbf{x}_i - (\mathbf{x}_i, \varepsilon_1) \varepsilon_1 - \dots - (\mathbf{x}_i, \varepsilon_{i-1}) \varepsilon_{i-1}.$$

Now we chose $\varepsilon_i = \mathbf{y}_i / \|\mathbf{y}_i\|$. Since

$$\varepsilon_\nu \in \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_{i-1}\} \quad (\nu = 1: i-1),$$

we get, by the construction of vectors $\mathbf{y}_i$ and ε_i , $\varepsilon_i \in \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_i\}$. Hence

$$\text{span}\{\varepsilon_1, \dots, \varepsilon_i\} \subset \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_i\}.$$

From the representation of the vector $\mathbf{y}_i$ we see that $\mathbf{x}_i$ is a linear combination of vectors $\varepsilon_1, \dots, \varepsilon_i$.

Thus,

$$\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_i\} \subset \text{span}\{\varepsilon_1, \dots, \varepsilon_i\}.$$

Finally,

$$\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_i\} = \text{span}\{\varepsilon_1, \dots, \varepsilon_i\}.$$

Example 1.6.4. Given a vector system $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ in $\mathbf{R}^4$, where

$$\mathbf{x}_1 = [1; 0; 1; 0]^T, \mathbf{x}_2 = [1; 1; 1; 0]^T, \mathbf{x}_3 = [0; 1; 0; 1]^T.$$

Find such an orthogonal system $\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$, for which

$$\text{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} = \text{span}\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}.$$

To apply the orthogonalization process of proposition 1.6.8, we check first the system $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ for the linearly independence (one can omit this process, too, because the situation will be clear in the course of the orthogonalization:

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \xrightarrow{\text{II-I}} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \xrightarrow{\text{III-II}} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow$$

the system $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is linearly independent. Now we find

$$\varepsilon_1 = \mathbf{x}_1 / \|\mathbf{x}_1\| = [1/\sqrt{2}; 0; 1/\sqrt{2}; 0]^T.$$

For $\mathbf{y}_2$ we get:

$$\mathbf{y}_2 = \mathbf{x}_2 - (\mathbf{x}_2, \varepsilon_1)\varepsilon_1 = [1; 1; 1; 0]^T - \sqrt{2}[1/\sqrt{2}; 0; 1/\sqrt{2}; 0]^T = [0; 1; 0; 0]^T.$$

As $\|\mathbf{y}_2\| = 1$, $\varepsilon_2 = \mathbf{y}_2 / \|\mathbf{y}_2\| = [0; 1; 0; 0]^T$. The vector $\mathbf{y}_3$ can be expressed in the form:

$$\begin{aligned} \mathbf{y}_3 &= \mathbf{x}_3 - (\mathbf{x}_3, \varepsilon_1)\varepsilon_1 - (\mathbf{x}_3, \varepsilon_2)\varepsilon_2 = \\ &= [0; 1; 0; 1]^T - 0 \cdot [1/\sqrt{2}; 0; 1/\sqrt{2}; 0]^T - 1 \cdot [0; 1; 0; 0]^T = [0; 0; 0; 1]^T. \end{aligned}$$

Thus,

$$\varepsilon_3 = \mathbf{y}_3 / \|\mathbf{y}_3\| = [0; 0; 0; 1]^T.$$

Example 1.6.5. Given a linearly independent vector system $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ in $\mathbf{L}_2[-1; 1]$, where $\mathbf{x}_1 = 1$, $\mathbf{x}_2 = t$ and $\mathbf{x}_3 = t^2$. Find an orthogonal system $\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$, such that

$$\text{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} = \text{span}\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}.$$

Check that the system $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is linearly independent. The first vector is

$$\varepsilon_1 = \mathbf{x}_1 / \|\mathbf{x}_1\| = 1/\sqrt{2}.$$

The vector $\mathbf{y}_2$ can be expressed in the form:

$$\mathbf{y}_2 = \mathbf{x}_2 - (\mathbf{x}_2, \varepsilon_1)\varepsilon_1 = t - \left(\int_{-1}^1 t \cdot \left(\frac{1}{\sqrt{2}}\right)dt\right)t = t - 0 \cdot t = t.$$

Thus,

$$\varepsilon_2 = \mathbf{y}_2 / \|\mathbf{y}_2\| = t / \sqrt{\int_{-1}^1 t \cdot \bar{t} dt} = t / \sqrt{\frac{2}{3}} = \sqrt{\frac{3}{2}}t.$$

The vector $\mathbf{y}_3$ can be expressed in the form:

$$\begin{aligned} \mathbf{y}_3 &= \mathbf{x}_3 - (\mathbf{x}_3, \varepsilon_1)\varepsilon_1 - (\mathbf{x}_3, \varepsilon_2)\varepsilon_2 = \\ &= t^2 - \left(\int_{-1}^1 t^2 \cdot \left(\frac{1}{\sqrt{2}}\right)dt\right)\frac{1}{\sqrt{2}} - \left(\int_{-1}^1 t^2 \left(\sqrt{\frac{3}{2}}t\right)dt\right)\sqrt{\frac{3}{2}}t = \\ &= t^2 - \frac{1}{2} \cdot \frac{2}{3} - 0 = t^2 - \frac{1}{3}. \end{aligned}$$

Therefore,

$$\begin{aligned} \varepsilon_3 &= \mathbf{y}_3 / \|\mathbf{y}_3\| = (t^2 - \frac{1}{3}) / \sqrt{\int_{-1}^1 (t^2 - \frac{1}{3})(t^2 - \frac{1}{3})dt} = \\ &= (t^2 - \frac{1}{3}) / \sqrt{\frac{2}{5} - \frac{4}{9} + \frac{2}{9}} = \sqrt{\frac{45}{8}}(t^2 - \frac{1}{3}) = \frac{3}{2}\sqrt{\frac{5}{2}}(t^2 - \frac{1}{3}). \end{aligned}$$

The functions $\varepsilon_1, \varepsilon_2$ and ε_3 are the normed Legendre polynomials on $[-1; 1]$.

Problem 1.6.3. Show that a vector system $\{x_1, \dots, x_n\}$ with pairwise orthogonal elements is linearly independent.

1.2 Matrices

1.2.1 Notation for a Matrix and Operations with Matrices

The vector space of all $m \times n$ -matrices with real elements will be denoted by $\mathbf{R}^{m \times n}$ and

$$A \in \mathbf{R}^{m \times n} \Leftrightarrow A = (a_{ik}) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ik} \in \mathbf{R}.$$

The element of the matrix A that stands in the i -th row and k -th column will be denoted by a_{ik} or $A(i, k)$ or $[A]_{ik}$. The main operations with matrices are following:

- transposition of matrices ($\mathbf{R}^{m \times n} \rightarrow \mathbf{R}^{n \times m}$)

$$B = A^T \Leftrightarrow b_{ik} = a_{ki},$$

- addition of matrices ($\mathbf{R}^{m \times n} \times \mathbf{R}^{m \times n} \rightarrow \mathbf{R}^{m \times n}$)

$$C = A + B \Leftrightarrow c_{ik} = a_{ik} + b_{ik},$$

- multiplication of matrices by a number ($\mathbf{R} \times \mathbf{R}^{m \times n} \rightarrow \mathbf{R}^{m \times n}$)

$$B = \lambda A \Leftrightarrow b_{ik} = \lambda a_{ik}$$

- multiplication of matrices ($\mathbf{R}^{m \times p} \times \mathbf{R}^{p \times n} \rightarrow \mathbf{R}^{m \times n}$)

$$C = AB \Leftrightarrow c_{ik} = \sum_{j=1}^p a_{ij}b_{jk}.$$

Problem 2.1.1.* Let

$$A = \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix}, \quad B = \begin{bmatrix} k & n \\ l & p \\ m & r \end{bmatrix}.$$

Find the matrix AB .

Problem 2.1.2.* Let

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \ddots & 0 & 0 \\ 0 & 1 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix} \in \mathbf{R}^{n \times n}.$$

Find the matrix A^{n-1} .

Problem 2.1.3.* Let

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

Prove that

$$A^n = 2^{n-1}A \quad (n \in N).$$

Example 2.1.1.* Let us show that multiplication of matrices is not commutative. Let

$$A = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} -2 & 5 \\ 1 & 2 \end{bmatrix}.$$

We find the products:

$$AB = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} -2 & 5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 13 \\ -4 & 19 \end{bmatrix},$$

$$BA = \begin{bmatrix} -2 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 13 & 2 \\ 7 & 8 \end{bmatrix}.$$

As $AB \neq BA$ does not hold for the example, multiplication of matrices is not commutative in general.

Proposition 2.1.1. If $A \in \mathbf{R}^{m \times p}$ and $B \in \mathbf{R}^{p \times n}$, then

$$(AB)^T = B^T A^T.$$

Proof. If $C = (AB)^T$, then

$$c_{ik} = [(AB)^T]_{ik} = [AB]_{ki} = \sum_{j=1}^p a_{kj} b_{ji}.$$

If $D = B^T A^T$, we also have

$$\begin{aligned} d_{ik} &= [B^T A^T]_{ik} = \sum_{j=1}^p [B^T]_{ij} [A^T]_{jk} = \sum_{j=1}^p [B]_{ji} [A]_{kj} = \\ &= \sum_{j=1}^p a_{kj} b_{ji} = c_{ik}. \quad \square \end{aligned}$$

Definition 2.1.1. A matrix $A \in \mathbf{R}^{n \times n}$ is called *symmetric* if $A^T = A$ and *skew-symmetric* if $A^T = -A$.

Problem 2.1.4.* Is matrix A symmetric or skew-symmetric if

$$a) A = \begin{bmatrix} -1 & 3 & 2 \\ 3 & 1 & 3 \\ 2 & 3 & -1 \end{bmatrix}, \quad b) A = \begin{bmatrix} 0 & 2 & -4 \\ -2 & 1 & -7 \\ 4 & 7 & 2 \end{bmatrix}, \quad c) A = \begin{bmatrix} 2 & -3 & 5 \\ 3 & 1 & 2 \\ -5 & 1 & 4 \end{bmatrix}.$$

Proposition 2.1.2. Each matrix $A \in \mathbf{R}^{n \times n}$ can be expressed as a sum of a symmetric matrix and a skew-symmetric matrix.

Proof. Each matrix $A \in \mathbf{R}^{n \times n}$ can be expressed as $A = B + C$, where $B = (A + A^T)/2$ and $C = (A - A^T)/2$. As

$$B^T = ((A + A^T)/2)^T = (A^T + A)/2 = B$$

and

$$C^T = ((A - A^T)/2)^T = C = (A^T - A)/2 = -C,$$

the proposition holds. $\square$

Problem 2.1.5.* Represent the matrix

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

as a sum of a symmetric and a skew-symmetric matrix.

Definition 2.1.2. If A is a $m \times n$ -matrix with complex elements, i.e., $A \in \mathbf{C}^{m \times n}$, then the *transposed skew-symmetric matrix* A^H will be defined by the equality

$$B = A^H \Leftrightarrow b_{ik} = \overline{a_{ki}}.$$

Definition 2.1.3. A matrix $A \in \mathbf{C}^{n \times n}$ is called an *Hermitian matrix* if

$$A^H = A.$$

Problem 2.1.6.* Is matrix A an Hermitian matrix if

$$a) A = \begin{bmatrix} i & -2+i & -5+3i \\ 2+i & 5i & -2+i \\ 5+3i & 2+i & -8i \end{bmatrix}, \quad b) A = \begin{bmatrix} 5 & 2+3i & 1+i \\ 2-3i & -3 & -2i \\ 1-i & 2i & 0 \end{bmatrix}.$$

Problem 2.1.7.* Let $A \in \mathbf{C}^{m \times n}$. Show that matrices AA^H and $A^H A$ are Hermitian matrices.

The matrix $A \in \mathbf{C}^{m \times n}$ can be expressed both by the column-vectors $\mathbf{c}_k$ ($k = 1 : n$) of the matrix A and by the row-vectors $\mathbf{r}_i^T$ ($i = 1 : m$) of the transpose of matrix A (“pasting” the matrices of the column-vectors or of the transposed row-vectors)

$$A = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_n \end{bmatrix} \equiv \begin{bmatrix} \mathbf{c}_1, & \cdots, & \mathbf{c}_n \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1^T \\ \vdots \\ \mathbf{r}_m^T \end{bmatrix},$$

where $\mathbf{c}_k \in \mathbf{C}^m$ and $\mathbf{r}_i \in \mathbf{C}^n$ and

$$\mathbf{r}_i = \begin{bmatrix} a_{i1} \\ \vdots \\ a_{in} \end{bmatrix}, \quad \mathbf{c}_k = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{mk} \end{bmatrix}.$$

Example 2.1.2. Let us demonstrate these notions on a matrix $A \in \mathbf{R}^{3 \times 2}$:

$$\begin{aligned} A &= \begin{bmatrix} 2 & 3 \\ 4 & 1 \\ 3 & 2 \end{bmatrix} \Rightarrow \mathbf{c}_1 = \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} \wedge \mathbf{c}_2 = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} \wedge \\ \mathbf{r}_1 &= \begin{bmatrix} 2 \\ 3 \end{bmatrix} \wedge \mathbf{r}_2 = \begin{bmatrix} 4 \\ 1 \end{bmatrix} \wedge \mathbf{r}_3 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \wedge \\ \mathbf{r}_1^T &= \begin{bmatrix} 2 & 3 \end{bmatrix} \wedge \mathbf{r}_2^T = \begin{bmatrix} 4 & 1 \end{bmatrix} \wedge \mathbf{r}_3^T = \begin{bmatrix} 3 & 2 \end{bmatrix} \wedge \\ A &= \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1, & \mathbf{c}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \mathbf{r}_3^T \end{bmatrix}. \end{aligned}$$

If $A \in \mathbf{R}^{m \times n}$, then $A(i, :)$ denotes the i -th row of the matrix A , i.e.,

$$A(i, :) = \begin{bmatrix} a_{i1} & \cdots & a_{in} \end{bmatrix},$$

and $A(:, k)$ denotes the k -th column of the matrix A , i.e.,

$$A(:, k) = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{mk} \end{bmatrix}.$$

If $1 \leq p \leq q < n \wedge 1 \leq r \leq m$, then

$$A(r, p : q) = \begin{bmatrix} a_{rp} & \cdots & a_{rq} \end{bmatrix} \in \mathbf{R}^{1 \times (q-p+1)}$$

and if $1 \leq p \leq n \wedge 1 \leq r \leq s \leq m$, then

$$A(r : s, p) = \begin{bmatrix} a_{rp} \\ \vdots \\ a_{sp} \end{bmatrix} \in \mathbf{R}^{s-r+1}.$$

If $A \in \mathbf{R}^{m \times n}$ and $\mathbf{i} = (i_1, \dots, i_p)$ and $\mathbf{k} = (k_1, \dots, k_q)$, where

$$i_1, \dots, i_p \in \{1; 2; \dots; m\} \wedge k_1, \dots, k_q \in \{1; 2; \dots; n\},$$

then the corresponding *submatrix* is

$$A(\mathbf{i}, \mathbf{k}) = \begin{bmatrix} A(i_1, k_1) & \cdots & A(i_1, k_q) \\ \vdots & & \vdots \\ A(i_p, k_1) & \cdots & A(i_p, k_q) \end{bmatrix}.$$

Example 2.1.3. If

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

and $\mathbf{i} = (2; 4)$ and $\mathbf{k} = (1; 3; 5)$, then

$$A(\mathbf{i}, \mathbf{k}) = \begin{bmatrix} 2 & 4 & 3 \\ 4 & 6 & 9 \end{bmatrix}.$$

1.2.2 Band Matrices and Block Matrices

Definition 2.2.1. A matrix whose elements different from zero are only on the main and some adjacent diagonals is called a *band matrix*.

Definition 2.2.2. It is said that the matrix $A \in \mathbf{R}^{m \times n}$ is a *band matrix* with the lower bandwidth p if

$$(i > k + p) \Rightarrow a_{ik} = 0$$

and with the upper bandwidth q if

$$(k > i + q) \Rightarrow a_{ik} = 0,$$

and with the bandwidth $p + q + 1$.

Example 2.2.1. The matrix

$$A = \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & 0 & 0 & 0 \\ 0 & \times & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times & \times \end{bmatrix}$$

is a band matrix because all the elements different from zero are on the main and two lower and one upper diagonals. The lower bandwidth of the matrix A is 2 because $a_{ik} = 0$ as $i > k + 2$, and the upper bandwidth is 1 because $a_{ik} = 0$ as $k > i + 1$. The bandwidth of the matrix is $2 + 1 + 1 = 4$. The elements of the matrix that are necessarily not zeros are denoted by crosses.

Some of the most important types of band matrices are presented in table 2.2.1. If $D \in \mathbf{R}^{m \times n}$ is a diagonal matrix, $q = \min\{m, n\}$ and $d_i = d_{ii}$, then the notation $D = \text{diag}(d_1, \dots, d_q)$. will be used.

Table 2.2.1.

The matrix's type	Lower bandwidth	Upper bandwidth
diagonal matrix	0	0
upper triangular matrix	0	n-1
lower triangular matrix	m-1	0
tridiagonal matrix	1	1
upper tridiagonal matrix	0	1
lower tridiagonal matrix	1	0
upper Hessenberg matrix	1	n-1
lower Hessenberg matrix	m-1	1

Problem 2.2.1.* Find the type, lower bandwidth, upper bandwidth and bandwidth of the matrix A if

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 & 0 \\ 4 & 2 & 1 & 1 & 0 \\ 0 & 2 & 3 & 4 & 1 \\ 0 & 0 & 5 & 4 & 6 \\ 0 & 0 & 0 & 6 & 5 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 1 \\ 2 & 2 & 1 & 0 & \ddots & 0 & 0 \\ 1 & 2 & 3 & 1 & \ddots & 0 & 0 \\ 0 & 1 & 2 & 4 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & \ddots & \cdots & n-1 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 2 & n \end{bmatrix}.$$

Definition 2.2.3. A matrix $A = (A_{\alpha\beta}) \in \mathbf{R}^{m \times n}$ is called a $q \times r$ -block matrix if

$$A = \begin{bmatrix} A_{1;1} & \cdots & A_{1;r} \\ \vdots & & \vdots \\ A_{q;1} & \cdots & A_{q;r} \end{bmatrix} \begin{matrix} m_1 \\ \vdots \\ m_q \end{matrix}, \quad \begin{matrix} n_1 & \cdots & n_r \end{matrix}$$

where $m_1 + \cdots + m_q = m$ and $n_1 + \cdots + n_r = n$ and $A_{\alpha\beta}$ is a $m_\alpha \times n_\beta$ -matrix.

Example 2.2.2. The matrix

$$A = \begin{bmatrix} a & a & a & b & b \\ a & a & a & b & b \\ a & a & a & b & b \\ c & c & c & d & d \end{bmatrix}$$

is a 2×2 -block matrix, where $m_1 = 3$, $m_2 = 1$, $n_1 = 3$ and $n_2 = 2$ and

$$A_{1;1} = \begin{bmatrix} a & a & a \\ a & a & a \\ a & a & a \end{bmatrix}, \quad A_{1;2} = \begin{bmatrix} b & b \\ b & b \\ b & b \end{bmatrix}, \quad A_{2;1} = \begin{bmatrix} c & c & c \end{bmatrix}, \quad A_{2;2} = \begin{bmatrix} d & d \end{bmatrix}.$$

Let

$$B = \begin{bmatrix} B_{1;1} & \dots & B_{1;r} \\ \vdots & & \vdots \\ B_{q;1} & \dots & B_{q;r} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_q \end{matrix},$$

$n_1 \qquad n_r$

and $C = A + B$. Then

$$C = \begin{bmatrix} C_{1;1} & \dots & C_{1;r} \\ \vdots & & \vdots \\ C_{q;1} & \dots & C_{q;r} \end{bmatrix} = \begin{bmatrix} A_{1;1} + B_{1;1} & \dots & A_{1;r} + B_{1;r} \\ \vdots & & \vdots \\ A_{q;1} + B_{q;1} & \dots & A_{q;r} + B_{q;r} \end{bmatrix}.$$

Proposition 2.2.1. If $A \in \mathbf{R}^{m \times p}$, $B \in \mathbf{R}^{p \times n}$ and $C = AB$ are block matrices:

$$A = \begin{bmatrix} A_{1;1} & \dots & A_{1;r} \\ \vdots & & \vdots \\ A_{\alpha;1} & \dots & A_{\alpha;r} \\ \vdots & & \vdots \\ A_{q;1} & \dots & A_{q;r} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_\alpha \\ \\ m_q \end{matrix},$$

$p_1 \qquad p_r$

$$B = \begin{bmatrix} B_{1;1} & \dots & B_{1;\beta} & \dots & B_{1;s} \\ \vdots & & \vdots & & \vdots \\ B_{r;1} & \dots & B_{r;\beta} & \dots & B_{r;s} \end{bmatrix} \begin{matrix} p_1 \\ \\ p_r \end{matrix},$$

$n_1 \qquad n_\beta \qquad n_s$

$$C = \begin{bmatrix} C_{1;1} & \dots & C_{1;\beta} & \dots & C_{1;s} \\ \vdots & & \vdots & & \vdots \\ C_{\alpha;1} & \dots & C_{\alpha;\beta} & \dots & C_{\alpha;s} \\ \vdots & & \vdots & & \vdots \\ C_{q;1} & \dots & C_{q;\beta} & \dots & C_{q;s} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_\alpha \\ \\ m_q \end{matrix},$$

$n_1 \qquad n_\beta \qquad n_s$

where $1 \leq \alpha \leq q$, $1 \leq \beta \leq s$, $m_1 + \dots + m_q = m$, $p_1 + \dots + p_s = p$, $n_1 + \dots + n_r = n$, then

$$C_{\alpha;\beta} = \sum_{\gamma=1}^r A_{\alpha;\gamma} B_{\gamma;\beta} \quad (\alpha = 1 : q \wedge \beta = 1 : s).$$

Proof. Let

$$\lambda = m_1 + \dots + m_{\alpha-1}, \mu = n_1 + \dots + n_{\beta-1}, 1 \leq \gamma \leq r,$$

$$\nu = p_1 + \dots + p_{\gamma-1}, m_0 = n_0 = p_0 = 0.$$

As $[C_{\alpha;\beta}]_{i;k}$ is an element of the block $C_{\alpha;\beta}$ of the matrix C standing in the i -th row and k -th column of this block, and $[A_{\alpha;\gamma}]_{i;j}$ is an element of the block $A_{\alpha;\gamma}$ of the matrix A standing in the i -th row and j -th column of this block, and $[B_{\gamma;\beta}]_{j;k}$ is an element of the block $B_{\gamma;\beta}$ of the matrix B standing in the j -th row and k -th column, then

$$[C_{\alpha;\beta}]_{i;k} = c_{\lambda+i;\mu+k}, [A_{\alpha;\gamma}]_{i;j} = a_{\lambda+i;\nu+j}, [B_{\gamma;\beta}]_{j;k} = b_{\nu+j;\mu+k}.$$

Therefore,

$$\begin{aligned} [C_{\alpha;\beta}]_{i;k} &= c_{\lambda+i;\mu+k} = \sum_{j=1}^p a_{\lambda+i;j} b_{j;\mu+k} = \\ &= \sum_{j=1}^{p_1} a_{\lambda+i;j} b_{j;\mu+k} + \sum_{j=p_1+1}^{p_1+p_2} a_{\lambda+i;j} b_{j;\mu+k} + \dots + \sum_{j=p_1+p_2+\dots+p_{r-1}+1}^p a_{\lambda+i;j} b_{j;\mu+k} = \\ &= \sum_{j=1}^{p_1} [A_{\alpha;1}]_{i;j} [B_{1;\beta}]_{j;k} + \sum_{j=1}^{p_2} [A_{\alpha;2}]_{i;j} [B_{2;\beta}]_{j;k} + \dots + \sum_{j=1}^{p_r} [A_{\alpha;r}]_{i;j} [B_{r;\beta}]_{j;k} = \\ &= [A_{\alpha;1} B_{1;\beta}]_{i;k} + [A_{\alpha;2} B_{2;\beta}]_{i;k} + \dots + [A_{\alpha;r} B_{r;\beta}]_{i;k} = \left[\sum_{j=1}^r A_{\alpha;j} B_{j;\beta} \right]_{i;k}. \end{aligned}$$

Therefore, all the corresponding elements of the matrices $C_{\alpha;\beta}$ and $\sum_{\gamma=1}^s A_{\alpha;\gamma} B_{\gamma;\beta}$ are equal, and our proposition holds. $\square$

Corollary 2.2.1. If $A \in \mathbf{R}^{m \times p}$, $B \in \mathbf{R}^{p \times n}$, $A = \begin{bmatrix} A_1 \\ \vdots \\ A_q \end{bmatrix} \begin{matrix} m_1 \\ \vdots \\ m_q \end{matrix}$,

$$B = \begin{bmatrix} B_1 & \dots & B_r \end{bmatrix},$$

$$n_1 \quad n_r$$

and $m_1 + \dots + m_q = m$ and $n_1 + \dots + n_r = n$, then

$$AB = C = \begin{bmatrix} C_{1;1} & \dots & C_{1;r} \\ \vdots & & \vdots \\ C_{q;1} & \dots & C_{q;r} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_q \end{matrix},$$

$$n_1 \quad n_r$$

where $C_{\alpha\beta} = A_{\alpha}B_{\beta}$ ($\alpha = 1 : q \wedge \beta = 1 : r$).

Corollary 2.2.2. If $A \in \mathbf{R}^{m \times p}$, $B \in \mathbf{R}^{p \times n}$,

$$A = \begin{bmatrix} A_1 & \dots & A_s \end{bmatrix},$$

$$p_1 \quad p_s$$

$$B = \begin{bmatrix} B_1 \\ \vdots \\ B_s \end{bmatrix} \begin{matrix} p_1 \\ \\ p_s \end{matrix}$$

and $p_1 + \dots + p_s = p$, then $AB = C = \sum_{k=1}^p A_k B_k$.

Example 2.2.3. It holds

$$\begin{bmatrix} A_{1;1} & A_{1;2} \\ A_{2;1} & A_{2;2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_{1;1}x_1 + A_{1;2}x_2 \\ A_{2;1}x_1 + A_{2;2}x_2 \end{bmatrix}.$$

Example 2.2.4. It holds

$$\begin{bmatrix} a & a & a & b \\ a & a & a & b \\ a & a & a & b \\ c & c & c & d \\ c & c & c & d \end{bmatrix} \begin{bmatrix} e & f & f \\ e & f & f \\ e & f & f \\ g & h & h \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{bmatrix},$$

where $A = (a)$ is a 3×3 -matrix, $B = (b)$ is a 3×1 -matrix, $C = (c)$ is a 2×3 -matrix, $D = (d)$ is a 2×1 -matrix, $E = (e)$ is a 3×1 -matrix, $F = (f)$ is a 3×2 -matrix, $G = (g)$ is a 1×1 -matrix and $H = (h)$ is a 1×2 -matrix.

Example 2.2.5.* Let us find the product AB of block matrices A and B , when A and B are 3×3 -matrices

$$A = \begin{bmatrix} 1 & 2 & \vdots & 2 \\ 3 & 4 & \vdots & 0 \\ \cdots & \cdots & \vdots & \cdots \\ 0 & 0 & \vdots & -1 \end{bmatrix}, \quad B = \begin{bmatrix} -3 & 1 & 0 & \vdots & 1 \\ 2 & 3 & -1 & \vdots & 1 \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ 0 & 0 & 0 & \vdots & 1 \end{bmatrix}.$$

We denote

$$A = \begin{bmatrix} C & D \\ E & F \end{bmatrix}, \quad B = \begin{bmatrix} G & H \\ K & L \end{bmatrix},$$

where

$$C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad D = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} -1 \end{bmatrix}$$

and

$$G = \begin{bmatrix} -3 & 1 & 0 \\ 2 & 3 & -1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad K = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 1 \end{bmatrix}.$$

We note that the dimensions of the matrices are in accordance with the conditions of multiplication of block matrices. If we denote

$$AB = \begin{bmatrix} R & S \\ T & U \end{bmatrix},$$

then

$$R = CG + DK = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -3 & 1 & 0 \\ 2 & 3 & -1 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 7 & -2 \\ -1 & 15 & -4 \end{bmatrix},$$

$$S = CH + DL = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix},$$

$$T = EG + FK = \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} -3 & 1 & 0 \\ 2 & 3 & -1 \end{bmatrix} + \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

and

$$U = EH + FL = \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} -1 \end{bmatrix}.$$

Thus

$$AB = \begin{bmatrix} 1 & 7 & -2 & 5 \\ -1 & 15 & -4 & 7 \\ 0 & 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 7 & -2 & 5 \\ -1 & 15 & -4 & 7 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

Problem 2.2.2.* Find the product AB of 4×5 -matrix A and 5×4 -matrix B in block form, when

$$A = \begin{bmatrix} 1 & 2 & 3 & \vdots & 0 & 0 \\ 0 & -1 & 4 & \vdots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \vdots & 4 & 1 \\ 0 & 0 & 0 & \vdots & 7 & 5 \end{bmatrix}, B = \begin{bmatrix} 1 & -4 & \vdots & 0 & 0 \\ 2 & 3 & \vdots & 0 & 0 \\ 5 & -1 & \vdots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \vdots & 1 & -1 \\ 0 & 0 & \vdots & 4 & -3 \end{bmatrix}.$$

1.2.3 Determinants

Let us consider an $n \times n$ -matrix, the so-called *matrix of order n*

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Definition 2.3.1. Arbitrary ordering $1, 2, \dots, n$ of indices $i_1, i_2, \dots, i_n$ is called a *permutation*.

Definition 2.3.2. The ordering of two indices in the permutation $i_1 i_2 \dots i_n$ is called *natural* if the smaller index stands before the greater one; in the opposite case, the greater index standing before the smaller one, it is said that the two indices form an *inversion*.

Definition 2.3.3. A *determinant* is a law (mapping, function) that associates with each square matrix A a number, so-called determinant of the

matrix

$$\det(A) \equiv \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} = \sum (-1)^\sigma a_{1,i_1} a_{2,i_2} a_{3,i_3} \dots a_{n,i_n},$$

where the summation goes over all the permutations $i_1 i_2 i_3 \dots i_n$ of indices $1, 2, 3, \dots, n$ and σ is the number of inversions in the permutation $i_1 i_2 i_3 \dots i_n$ of the row indices. We will use expressions: determinant of order n and its rows and columns.

Example 2.3.1. Let us consider the third order determinant

$$\begin{aligned} \det(A) &= \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \\ &= (-1)^0 a_{11} a_{22} a_{33} + (-1)^1 a_{11} a_{23} a_{32} + (-1)^1 a_{12} a_{21} a_{33} + \\ &\quad + (-1)^2 a_{12} a_{23} a_{31} + (-1)^2 a_{13} a_{21} a_{32} + (-1)^3 a_{13} a_{22} a_{31}. \end{aligned}$$

Let us examine the last summand $(-1)^3 a_{13} a_{22} a_{31}$. In the permutation 3 2 1 of the column indices the index 3 forms with the index 2 and the index 1 an inversion. The index 2 does the same with the index 1. So the number of inversions σ in the permutation of the column indices is equal to 3.

Problem 2.3.1.* Which sign has the product

$$a_{1,n} a_{2,n-1} a_{3,n-2} \dots a_{n-1,2} a_{n,1}$$

of elements of a determinant expression.

Properties of determinant

- The determinants of a matrix and its transpose are equal, $\det(A^T) = \det(A)$.
- Multiplying all the elements of the row (column) of the determinant by the same number the determinant will be multiplied by the same number.
- Interchanging two rows (columns) of the determinant the determinant will change its sign.

- If two rows (columns) of the determinant are identical, then the determinant is equal to 0.
- If each element in the row (column) of the determinant is a sum of two summands, then the determinant expands into the sum of two determinants, where in the considered row (column) in the first of them there will be the first summands and in the second of them there will be the second summands, and all the remaining rows (columns) will be identical to those of the given matrix:

$$\begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{k1} + b_{k1} & \dots & a_{kn} + b_{kn} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{k1} & \dots & a_{kn} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ b_{k1} & \dots & b_{kn} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix}.$$

- The determinant will not change if an arbitrary row (column) multiplied by an arbitrary number is added to a row (column).
- The fundamental formulas of the determinant theory (or *theorem of expansion by cofactors*) are valid:

$$a_{i1}A_{k1} + a_{i2}A_{k2} + \dots + a_{in}A_{kn} = \det(A) \cdot \delta_{ik},$$

$$a_{1i}A_{1k} + a_{2i}A_{2k} + \dots + a_{ni}A_{nk} = \det(A) \cdot \delta_{ik},$$

where

$$\delta_{ik} = \begin{cases} 1, & \text{as } i = k \\ 0, & \text{as } i \neq k \end{cases}$$

is the Kronecker symbol and A_{ik} is the product of the number $(-1)^{i+k}$ and the determinant of the $(n-1) \times (n-1)$ -matrix obtained from the given matrix by deleting the i -th row and k -th column.

Example 2.3.2. Let us evaluate the determinant of order n , using the expansion by cofactors by the first column and then by the first row.

$$D_n = \begin{vmatrix} -2 & 1 & 0 & \dots & 0 & 0 \\ 1 & -2 & 1 & \dots & 0 & 0 \\ 0 & 1 & -2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -2 & 1 \\ 0 & 0 & 0 & \dots & 1 & -2 \end{vmatrix} =$$

$$\begin{aligned}
&= (-2)(-1)^{1+1}D_{n-1} + 1 \cdot (-1)^{2+1} \begin{vmatrix} 1 & 0 & \dots & 0 & 0 \\ 1 & -2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -2 & 1 \\ 0 & 0 & \dots & 1 & -2 \end{vmatrix} = \\
&= -2D_{n-1} - D_{n-2}
\end{aligned}$$

or

$$D_n + 2D_{n-1} + D_{n-2} = 0. \quad (1)$$

Equation (1) is a linear homogeneous difference equation with constant coefficients which has the solution of type λ^n . Let us try to find them:

$$\lambda^n + 2\lambda^{n-1} + \lambda^{n-2} = 0 \Leftrightarrow \lambda^{n-2}(\lambda^2 + 2\lambda + 1) = 0.$$

We are interested in a non-trivial solution. So we have get a quadratic equation

$$\lambda^2 + 2\lambda + 1 = 0,$$

to find the solution of the difference equation (1). It has the solutions $\lambda_{1,2} = -1$, and so one of the solutions of equation (1) is $D_n = (-1)^n$. As the number -1 is a double solution of the quadratic equation, $D_n = (-1)^n n$ will be a solution of the equation (1), too. Thus, we have got two linearly independent particular solutions of the linear homogeneous difference equation with constant coefficients. The general solution of the equation can be expressed in form

$$D_n = C_1(-1)^n + C_2(-1)^n n.$$

From the conditions $D_1 = -2$ and $D_2 = 3$ we can find the coefficients C_1 and C_2 :

$$\begin{cases} C_1(-1)^1 + C_2(-1)^1 \cdot 1 = -2 \\ C_1(-1)^2 + C_2(-1)^2 \cdot 2 = 3 \end{cases} \Rightarrow \begin{cases} C_1 = 1 \\ C_2 = 1 \end{cases}$$

So the given problem has the solution

$$D_n = (-1)^n(n + 1).$$

Problem 2.3.2.* Compute the determinant of order n

$$\begin{vmatrix} 7 & 5 & 0 & \cdots & 0 & 0 \\ 2 & 7 & 5 & \ddots & 0 & 0 \\ 0 & 2 & 7 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & \ddots & 7 & 5 \\ 0 & 0 & 0 & \ddots & 2 & 7 \end{vmatrix}.$$

Example 2.3.3. Evaluate the Vandermonde determinant

$$V_n(x_1, x_2, \dots, x_n) = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_1^{n-2} & x_2^{n-2} & x_3^{n-2} & \cdots & x_n^{n-2} \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \cdots & x_n^{n-1} \end{vmatrix}.$$

We subtract x_1 times the penultimate row from the last row, then x_1 times the $(n-2)$ -th row from the penultimate row, then x_1 times $(n-3)$ -th row from the $(n-2)$ -th row etc., in the end x_1 times the second row from the first one. As a result, we get

$$= \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & x_2 - x_1 & x_3 - x_1 & \cdots & x_n - x_1 \\ 0 & x_2^2 - x_1 x_2 & x_3^2 - x_1 x_3 & \cdots & x_n^2 - x_1 x_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & x_2^{n-2} - x_1 x_2^{n-3} & x_3^{n-2} - x_1 x_3^{n-3} & \cdots & x_n^{n-2} - x_1 x_n^{n-3} \\ 0 & x_2^{n-1} - x_1 x_2^{n-2} & x_3^{n-1} - x_1 x_3^{n-2} & \cdots & x_n^{n-1} - x_1 x_n^{n-2} \end{vmatrix}.$$

Using the expression by the first column and factoring out the common factors in the elements, we get

$$= \begin{vmatrix} x_2 - x_1 & x_3 - x_1 & \cdots & x_n - x_1 \\ x_2(x_2 - x_1) & x_3(x_3 - x_1) & \cdots & x_n(x_n - x_1) \\ \cdots & \cdots & \cdots & \cdots \\ x_2^{n-3}(x_2 - x_1) & x_3^{n-3}(x_3 - x_1) & \cdots & x_n^{n-3}(x_n - x_1) \\ x_2^{n-2}(x_2 - x_1) & x_3^{n-2}(x_3 - x_1) & \cdots & x_n^{n-2}(x_n - x_1) \end{vmatrix}.$$

Factoring out from the first columns the common factor $x_2 - x_1$, from the second column $x_3 - x_1$, $\dots$, from the $(n-1)$ -th column $x_n - x_1$, we get

$$= (x_2 - x_1)(x_3 - x_1) \cdots (x_n - x_1) \begin{vmatrix} 1 & 1 & \dots & 1 \\ x_2 & x_3 & \dots & x_n \\ x_2^2 & x_3^2 & \dots & x_n^2 \\ \dots & \dots & \dots & \dots \\ x_2^{n-2} & x_3^{n-2} & \dots & x_n^{n-2} \end{vmatrix}.$$

Using the same operations cycle, results in

$$V_n(x_1, x_2, \dots, x_n) = \prod_{n \geq k > i \geq 1} (x_k - x_i).$$

Proposition 2.3.1 (The Laplace expansion theorem). The so-called Laplace formula

$$\det(A) = \sum M_k A_{n-k}$$

holds, where the summation on the right goes over all determinants (minors) M_k of order k that can be formed of rows $i_1, i_2, \dots, i_k$ and columns $j_1, j_2, \dots, j_k$, and A_{n-k} is the product of the number $(-1)^{i_1+i_2+\dots+i_k+j_1+j_2+\dots+j_k}$ and the determinant of the matrix remaining from the matrix A by deleting the rows $i_1, i_2, \dots, i_k$ and the columns $j_1, j_2, \dots, j_k$ used in forming the minor M_k .

Proof. See Kangro (1962, pp. 37-39). $\square$

Example 2.3.4. Using the Laplace expansion by the first two rows, transform the determinant

$$\begin{vmatrix} a & b & c & 0 \\ d & e & f & 0 \\ 0 & a & b & c \\ 0 & d & e & f \end{vmatrix}.$$

As only three minors are not equal to zero, we get the expansion

$$\begin{vmatrix} a & b & c & 0 \\ d & e & f & 0 \\ 0 & a & b & c \\ 0 & d & e & f \end{vmatrix} = (-1)^{1+2+1+2} \begin{vmatrix} a & b \\ d & e \end{vmatrix} \cdot \begin{vmatrix} b & c \\ e & f \end{vmatrix} +$$

$$+(-1)^{1+2+1+3} \begin{vmatrix} a & c \\ d & f \end{vmatrix} \cdot \begin{vmatrix} a & c \\ d & f \end{vmatrix} + (-1)^{1+2+2+3} \begin{vmatrix} b & c \\ e & f \end{vmatrix} \cdot \begin{vmatrix} 0 & c \\ 0 & f \end{vmatrix}.$$

Problem 2.3.3.* Compute by the use of the Laplace formula the determinant

$$\begin{vmatrix} 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 1 & 5 & 3 \\ 0 & 0 & 0 & 2 & 3 \\ -1 & 1 & 3 & 1 & 2 \\ 2 & 2 & 0 & 0 & 3 \end{vmatrix}.$$

By the Laplace expansion theorem, it holds for each matrix $C = \begin{bmatrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{n1} & \dots & c_{nn} \end{bmatrix}$ the equality

$$\begin{vmatrix} a_{11} & \dots & a_{1n} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} & 0 & \dots & 0 \\ c_{11} & \dots & c_{1n} & b_{11} & \dots & b_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ c_{n1} & \dots & c_{nn} & b_{n1} & \dots & b_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} \cdot \begin{vmatrix} b_{11} & \dots & b_{1n} \\ \dots & \dots & \dots \\ b_{n1} & \dots & b_{nn} \end{vmatrix} \quad (2)$$

Choosing $C = \begin{bmatrix} -1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & -1 \end{bmatrix}$, we transform the determinant

$$\begin{vmatrix} a_{11} & \dots & a_{1n} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} & 0 & \dots & 0 \\ -1 & \dots & 0 & b_{11} & \dots & b_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -1 & b_{n1} & \dots & b_{nn} \end{vmatrix}$$

so that all the elements b_{ij} become zeros. To make $b_{11}, b_{21}, \dots, b_{n1}$ into zeros we have to add to the $(n+1)$ -th column b_{11} times the elements of the first column, b_{21} times the elements of the second column etc, and, in the end, b_{n1} times the elements of the n -th column. Next we make into zeros the elements $b_{12}, b_{22}, \dots, b_{n2}$. For this we add to the $(n+2)$ -th column b_{12} times the first column, b_{22} times the second column etc, and, in the end, b_{n2} times the n -th

column etc. The last step will nullify the elements $b_{1n}, b_{2n}, \dots, b_{nn}$. For this we add to the $2n$ -th column b_{1n} times the first column, b_{2n} times the second column etc, and, in the end, b_{nn} times the n -th column. The result will be

$$\begin{aligned}
& \begin{vmatrix} a_{11} & \dots & a_{1n} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} & 0 & \dots & 0 \\ -1 & \dots & 0 & b_{11} & \dots & b_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -1 & b_{n1} & \dots & b_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & \dots & a_{1n} & d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} & d_{n1} & \dots & d_{nn} \\ -1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -1 & 0 & \dots & 0 \end{vmatrix} = \\
& = (-1)^{n+1+n+2+\dots+2n+1+2+\dots+n} \begin{vmatrix} -1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & -1 \end{vmatrix} \cdot \begin{vmatrix} d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots \\ d_{n1} & \dots & d_{nn} \end{vmatrix} = \\
& = (-1)^{\frac{(1+2n)2n}{2}+n} \begin{vmatrix} d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots \\ d_{n1} & \dots & d_{nn} \end{vmatrix} = \begin{vmatrix} d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots \\ d_{n1} & \dots & d_{nn} \end{vmatrix},
\end{aligned}$$

where

$$d_{ij} = \sum_{k=1}^n a_{ik} b_{kj}. \quad (3)$$

Taking into account (2) and the fact that, by (3), $D = A \cdot B$, we reach the assertion.

Proposition 2.3.1 (the theorem about the determinant of the product of matrices). For arbitrary matrices A and B of order n it holds

$$\det(AB) = (\det A)(\det B).$$

1.2.4 Four Subspaces of a Matrix

Let us consider an $m \times n$ -matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

with real elements. The matrix A can be expressed both by the column-vectors $\mathbf{c}_k$ ($k = 1 : n$) of A or by the row-vectors $\mathbf{r}_i^T$ ($i = 1 : m$) by the transpose of A

$$A = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_n \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1, & \cdots, & \mathbf{c}_n \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1^T \\ \vdots \\ \mathbf{r}_m^T \end{bmatrix},$$

where $\mathbf{r}_i \in \mathbf{R}^n$ and $\mathbf{c}_k \in \mathbf{R}^m$, and $\mathbf{r}_i = \begin{bmatrix} a_{i1} \\ \vdots \\ a_{in} \end{bmatrix}$, $\mathbf{c}_k = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{mk} \end{bmatrix}$.

Definition 2.4.1. The subspace $\text{span}\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ of the set $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ of column-vectors of the matrix A is called the *subspace of column-vectors* of the matrix A , and denoted by $\mathcal{R}(A)$ or $\text{ran}(A)$.

Definition 2.4.2. The subspace $\text{span}\{\mathbf{r}_1, \dots, \mathbf{r}_m\}$ of the set $\{\mathbf{r}_1, \dots, \mathbf{r}_m\}$ of the row-vectors of the matrix A is called the *subspace of row-vectors* of the matrix A , and denoted by $\mathcal{R}(A^T)$ or $\text{ran}(A^T)$.

Definition 2.4.3. The *rank of the matrix* A is the greatest natural number k , for which there exist a minor of order k different from zero. We denote the rank of A by $\text{rank}(A)$.

Let $\text{rank}(A) = r$. Due to the theorem about the rank of the matrix, we get

Proposition 2.4.1. The rank of the matrix is equal to the dimension of the subspace of its row-vectors or column-vectors, i.e.,

$$\text{rank}(A) = \dim \mathcal{R}(A^T) = \dim \mathcal{R}(A) = r.$$

Definition 2.4.4. The (*right*) *null space* of the matrix A is the set of all solutions

$$\mathbf{x} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix}^T$$

of the system of equations

$$A\mathbf{x} = 0. \tag{4}$$

It is a subspace, denoted $\mathcal{N}(A)$ or *null*(A).

Proposition 2.4.2. For every matrix $A \in \mathbf{R}^{m \times n}$ with the rank r ,

$$\dim \mathcal{N}(A) = n - r \wedge \mathcal{N}(A) \perp \mathcal{R}(A^T) \wedge \mathcal{N}(A) \oplus \mathcal{R}(A^T) = \mathbf{R}^n.$$

Proof. The matrix of the system has the rank r , and the number of variables in (4) equals n . Therefore, the number of degrees of freedom of the system is $n - r$. The number of degrees of freedom gives the dimension of the null space. Thus, $\dim \mathcal{N}(A) = n - r$. We can rewrite the system (4) in form

$$\begin{bmatrix} \mathbf{r}_1^T \mathbf{x} \\ \vdots \\ \mathbf{r}_m^T \mathbf{x} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Therefore, $\mathbf{r}_k^T \mathbf{x} = 0 \Leftrightarrow \mathbf{r}_k \perp \mathbf{x}$ ($k = 1 : m$), i.e., the row-vectors of A are orthogonal to any vector of the null space $\mathcal{N}(A)$ of the matrix A . Hence $\mathcal{N}(A) \perp \mathcal{R}(A^T)$. As, in addition, $\dim \mathcal{N}(A) = n - r$ and $\dim \mathcal{R}(A^T) = r$, $\dim \mathcal{N}(A) + \dim \mathcal{R}(A^T) = n$ and the space $\mathbf{R}^n$ can be expressed by the direct sum

$$\mathbf{R}^n = \mathcal{N}(A) \oplus \mathcal{R}(A^T). \quad \square$$

Definition 2.4.5. The (left) null space of the matrix A is the set of all solutions

$$\mathbf{y} = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} = \begin{bmatrix} \eta_1 & \dots & \eta_m \end{bmatrix}^T$$

of the system of equations

$$A^T \mathbf{y} = 0 \tag{5}$$

This subspace is denoted by $\mathcal{N}(A^T)$ or *null*(A^T).

Proposition 2.4.3. For every matrix $A \in \mathbf{R}^{m \times n}$ with the rank r ,

$$\dim \mathcal{N}(A^T) = m - r \wedge \mathcal{N}(A^T) \perp \mathcal{R}(A) \wedge \mathcal{N}(A^T) \oplus \mathcal{R}(A) = \mathbf{R}^m.$$

Proof. The matrix of the system A^T has the rank r , and, the number of variables in (5) equals m . Therefore, the number of degrees of freedom of the system is $m - r$ and

$$\dim \mathcal{N}(A^T) = m - r.$$

The system (5) can be expressed in form

$$\begin{bmatrix} \mathbf{c}_1^T \mathbf{y} \\ \vdots \\ \mathbf{c}_m^T \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

So $\mathbf{c}_k^T \mathbf{y} = 0 \Leftrightarrow \mathbf{c}_k \perp \mathbf{y}$ ($k = 1 : m$) and $\mathcal{N}(A^T) \perp \mathcal{R}(A)$. As $\dim \mathcal{N}(A^T) = m - r$ and $\dim \mathcal{R}(A) = r$, $\dim \mathcal{N}(A^T) + \dim \mathcal{R}(A) = m$, and the space R^m can be expressed by the direct sum

$$\mathbf{R}^m = \mathcal{N}(A^T) \oplus \mathcal{R}(A). \quad \square$$

Example 2.4.1. Let us find the dimensions and bases of the subspaces $\mathcal{R}(A)$, $\mathcal{N}(A)$, $\mathcal{R}(A^T)$ and $\mathcal{N}(A^T)$ for the matrix

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

We will illustrate the assertion of propositions 2.4.2 and 2.4.3 in case of this example.

We start with the examination of the space $\mathcal{R}(A)$. Substituting from the second column of A two times the first column, we get

$$\begin{bmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{bmatrix} \sim;$$

then subtracting from the third column the new second one, from the fourth column the first one and from the fifth column the first one and the new second one, we get

$$\sim \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The symbol " $\sim$ " between the matrices marks that $\mathcal{R}(A)$ is not changed. The last matrix has only two columns different from the null vector $\Rightarrow \dim \mathcal{R}(A) = 2$. The basis in the space $\mathcal{R}(A)$ will be

$$S_{\mathcal{R}(A)} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}.$$

To describe the space $\mathcal{N}(A^T)$, we solve system (5):

$$\begin{bmatrix} 1 & 0 & 1 & \vdots & 0 \\ 2 & 1 & 2 & \vdots & 0 \\ 0 & 1 & 0 & \vdots & 0 \\ 1 & 0 & 1 & \vdots & 0 \\ 1 & 1 & 1 & \vdots & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 1 & \vdots & 0 \\ 0 & 1 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix},$$

i.e.,

$$\begin{cases} \eta_1 + 0\eta_2 + \eta_3 &= 0 \\ \eta_2 &= 0 \end{cases} \Rightarrow \eta_2 = 0 \wedge \eta_3 = p \wedge \eta_1 = -p \Rightarrow$$

$$\mathbf{y} = \begin{bmatrix} -p \\ 0 \\ p \end{bmatrix} = p \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \dim \mathcal{N}(A^T) = 1 \wedge S_{\mathcal{N}(A^T)} = \left\{ \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Let us check by scalar product that $S_{\mathcal{R}(A)} \perp S_{\mathcal{N}(A^T)}$:

$$\begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = 1 \cdot (-1) + 0 \cdot 0 + 1 \cdot 1 = 0;$$

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

The union $S_{\mathcal{R}(A)} \cup S_{\mathcal{N}(A^T)}$ contains three linearly independent vectors of $\mathbf{R}^3$. These vectors form a basis in $\mathbf{R}^3$. Thus, $\mathbf{R}^3 = \mathcal{N}(A^T) \oplus \mathcal{R}(A)$. To describe the space $\mathcal{R}(A^T)$ let us find its dimension and basis:

$$\begin{bmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow$$

$$\dim \mathcal{R}(A^T) = 2 \wedge S_{\mathcal{R}(A^T)} = \left\{ \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

To describe the space $\mathcal{N}(A)$, we solve system (4):

$$\begin{aligned} \begin{bmatrix} 1 & 2 & 0 & 1 & 1 & \vdots & 0 \\ 0 & 1 & 1 & 0 & 1 & \vdots & 0 \\ 1 & 2 & 0 & 1 & 1 & \vdots & 0 \end{bmatrix} &\sim \begin{bmatrix} 1 & 2 & 0 & 1 & 1 & \vdots & 0 \\ 0 & 1 & 1 & 0 & 1 & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 \end{bmatrix} \Rightarrow \\ \begin{cases} \xi_1 + 2\xi_2 + 0\xi_3 + \xi_4 + \xi_5 = 0 \\ \xi_2 + \xi_3 + 0\xi_4 + \xi_5 = 0 \end{cases} &\Rightarrow \begin{cases} \xi_3 = p, \xi_4 = q, \xi_5 = t \\ \xi_2 = -p - t, \xi_1 = 2p - q + t \end{cases} \Rightarrow \\ \mathbf{x} = \begin{bmatrix} 2p - q + t \\ -p - t \\ p \\ q \\ t \end{bmatrix} &= p \begin{bmatrix} 2 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + q \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \\ S_{\mathcal{N}(A)} = \left\{ \begin{bmatrix} 2 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\} &\Rightarrow \dim S_{\mathcal{N}(A)} = 3. \end{aligned}$$

The vectors of the basis $S_{\mathcal{N}(A)}$ are orthogonal to the vectors of the basis $S_{\mathcal{R}(A^T)}$. Thus, $\mathcal{R}(A^T) \perp \mathcal{N}(A)$ and the union $S_{\mathcal{R}(A^T)} \cup S_{\mathcal{N}(A)}$ forms a basis in $\mathbf{R}^5$. Therefore

$$\mathcal{N}(A^T) \oplus \mathcal{R}(A) = \mathbf{R}^5.$$

Problem 2.4.1. Let $A \in \mathbf{R}^{n \times n}$. Show that $\mathcal{N}(A^T A) = \mathcal{N}(A)$.

Problem 2.4.2. Show that

$$\begin{aligned} \mathcal{N}(AB) &\supseteq \mathcal{N}(B) \wedge \mathcal{N}((AB)^T) \supseteq \mathcal{N}(A^T) \wedge \\ \mathcal{R}(AB) &\subseteq \mathcal{R}(A) \wedge \mathcal{R}((AB)^T) \subseteq \mathcal{R}(B^T). \end{aligned}$$

Problem 2.4.3.* Find the dimensions and bases of the subspaces $\mathcal{R}(A)$, $\mathcal{N}(A)$, $\mathcal{R}(A^T)$ and $\mathcal{N}(A^T)$ of the matrix A . Demonstrate the assertion of proposition 2.4.2 and 2.4.3 on the matrix A , where

$$a) A = \begin{bmatrix} -2 & -2 & -10 & 1 \\ 3 & 2 & 12 & -1 \\ -1 & -1 & -5 & 1 \end{bmatrix}, \quad b) A = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & 3 & 5 \\ -2 & 2 & 2 & 2 \end{bmatrix},$$

$$c) A = \begin{bmatrix} 8 & 16 & 2 & 6 \\ 2 & 4 & -1 & 3 \\ 9 & 18 & 2 & 7 \\ 3 & 6 & 0 & 3 \end{bmatrix}, \quad d) A = \begin{bmatrix} 1 & 2 & -1 & 2 & -2 \\ -1 & -2 & 2 & -3 & 3 \\ -1 & -2 & 0 & -1 & 1 \\ -2 & -4 & 0 & -2 & 2 \end{bmatrix}.$$

Problem 2.4.4.* Find the dimensions and bases of the subspaces $\mathcal{R}(AB)$, $\mathcal{N}(AB)$, $\mathcal{R}((AB)^T)$ and $\mathcal{N}((AB)^T)$ of the product AB , where

$$A = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & 3 & 5 \\ -2 & 2 & 2 & 2 \end{bmatrix} \wedge B = \begin{bmatrix} 8 & 16 & 2 & 6 \\ 2 & 4 & -1 & 3 \\ 9 & 18 & 2 & 7 \\ 3 & 6 & 0 & 3 \end{bmatrix}.$$

Compare the results obtained with the results of Problem 2.4.3* in case *b*) and *c*).

1.2.5 Eigenvalues and Eigenvectors of a Matrix

Definition 2.5.1. If

$$A\mathbf{x} = \lambda\mathbf{x}, \tag{6}$$

where $A \in \mathbf{C}^{n \times n}$, $\mathbf{x} \in \mathbf{C}^n$ and λ is a number, then the number λ is called an *eigenvalue* of the matrix A and the vector $\mathbf{x}$ a (*right*) *eigenvector* of the matrix A corresponding to the eigenvalue λ .

Definition 2.5.2. The vector $\mathbf{x}$ is called a (*left*) *eigenvector* of the matrix A if $\mathbf{x}^H A = \lambda \mathbf{x}^H$, where $\mathbf{x}^H$ is the transposed skew-matrix.

Proposition 2.5.1. If $\mathbf{x}$ is a left eigenvector of the matrix A corresponding to the eigenvalue λ , then this $\mathbf{x}$ is a right eigenvector corresponding to the eigenvalue $\bar{\lambda}$.

Proof. We get a chain of assertions:

$$\mathbf{x}^H A = \lambda \mathbf{x}^H \Leftrightarrow (\mathbf{x}^H A)^H = (\lambda \mathbf{x}^H)^H \Leftrightarrow A^H \mathbf{x} = \bar{\lambda} \mathbf{x}. \quad \square$$

It is obvious that if $\mathbf{x}$ is a eigenvector corresponding to the eigenvalue λ , then $c\mathbf{x}$, $c \in \mathbf{C}$ is an eigenvector, too. The equation (6) can be expressed in form

$$(A - \lambda I)\mathbf{x} = 0, \tag{7}$$

where I is the identity matrix of order n . As the null vector is an eigenvector for every square matrix A in eigenvalues problem (6), in following we will confine ourselves to the non-trivial eigenvectors. The equation (7) presents a system of homogeneous linear algebraic equations that has a non-trivial solution iff the matrix $A - \lambda I$ of the system is singular, i.e.,

$$\det(A - \lambda I) = 0. \quad (8)$$

The equation (8) is called the *characteristic equation* of the matrix A , and the polynomial

$$p(\lambda) = \det(A - \lambda I)$$

is called the *characteristic polynomial* of the matrix A . The equation (8) is an algebraic equation of order n with respect to λ , and it can be written down in form:

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \quad (9)$$

According to the fundamental theorem of algebra, the matrix $A \in \mathbf{C}^{n \times n}$ has exactly n eigenvalues, taking into account their multiplicity.

Definition 2.5.3. The set of all eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ of the matrix $A \in \mathbf{C}^{n \times n}$ is called the *spectrum* of the matrix A and denoted by $\lambda(A)$.

Example 2.5.1. Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

We compose the characteristic equation (9) corresponding to the given matrix:

$$\begin{vmatrix} 1 - \lambda & 1 & 1 \\ 1 & 1 - \lambda & 1 \\ 1 & 1 & 1 - \lambda \end{vmatrix} = 0.$$

Calculating the determinant, we get the cubic equation

$$(1 - \lambda)^3 - 3(1 - \lambda) + 2 = 0,$$

with the solutions $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = 3$. Let us find the eigenvectors corresponding to the eigenvalues $\lambda_1 = \lambda_2 = 0$. We replace in system (7) the variable λ by 0 and solve the equation:

$$\begin{bmatrix} 1-0 & 1 & 1 & \vdots & 0 \\ 1 & 1-0 & 1 & \vdots & 0 \\ 1 & 1 & 1-0 & \vdots & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 1 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix}.$$

There is only one independent equation remained:

$$\xi_1 + \xi_2 + \xi_3 = 0.$$

The number of degrees of freedom of the system is 2, and the general solution of the system is

$$\mathbf{x} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} -q - p \\ q \\ p \end{bmatrix} = p \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + q \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix},$$

where p and q are arbitrary real numbers. Thus, the vectors $\mathbf{x}$ that correspond to the eigenvalues $\lambda_1 = \lambda_2 = 0$ form a two-dimensional subspace in the space $\mathbf{R}^3$, and vectors $\mathbf{x}_1 = [-1 \ 0 \ 1]^T$ and $\mathbf{x}_2 = [-1 \ 1 \ 0]^T$ can be chosen for its basis. To find the eigenvector corresponding to the eigenvalue $\lambda_3 = 3$ we have to replace in the system of equations (7) the variable λ by 3. As a result, we get the system of equations:

$$\begin{aligned} & \begin{bmatrix} 1-3 & 1 & 1 & \vdots & 0 \\ 1 & 1-3 & 1 & \vdots & 0 \\ 1 & 1 & 1-3 & \vdots & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & -2 & \vdots & 0 \\ 1 & -2 & 1 & \vdots & 0 \\ -2 & 1 & 1 & \vdots & 0 \end{bmatrix} \sim \\ & \sim \begin{bmatrix} 1 & 1 & -2 & \vdots & 0 \\ 0 & -3 & 3 & \vdots & 0 \\ 0 & 3 & -3 & \vdots & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & -2 & \vdots & 0 \\ 0 & 1 & -1 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1 & \vdots & 0 \\ 0 & 1 & -1 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix}. \end{aligned}$$

The number of degrees of freedom of this system is 1, and the eigenvectors of the matrix A corresponding to the eigenvalue $\lambda_3 = 3$ can be expressed in form

$$\mathbf{x} = \begin{bmatrix} r \\ r \\ r \end{bmatrix} = r \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

They form a one-dimensional subspace in $\mathbf{R}^3$ with the basis vector $\mathbf{x}_3 = [1 \ 1 \ 1]^T$.

Problem 2.5.1.* Find the eigenvalues and eigenvectors of the matrix A , where

$$a) A = \begin{bmatrix} 2 & 3 \\ -1 & 6 \end{bmatrix}, \quad b) A = \begin{bmatrix} 5 & -2 \\ 7 & 4 \end{bmatrix}, \quad c) A = \begin{bmatrix} 1 & -1 \\ 2 & 4 \end{bmatrix}.$$

Problem 2.5.2.* Find the eigenvalues and eigenvectors of the matrix A , when

$$a) A = \begin{bmatrix} 3 & -2 & 2 \\ 0 & 1 & 0 \\ -1 & 1 & 0 \end{bmatrix}, \quad b) A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Proposition 2.5.2. If $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of the matrix A , then

$$\det(A) = \lambda_1 \lambda_2 \cdots \lambda_n.$$

Proof. The left side of the characteristic equation (8) with the zeros $\lambda_1, \dots, \lambda_n$ can be expressed in form

$$\det(A - \lambda I) = (-1)^n (\lambda - \lambda_1) \cdots (\lambda - \lambda_n). \quad (10)$$

If we take in this equation $\lambda = 0$, we get the assertion of the proposition. $\square$

Corollary 2.5.1. Not a single one of the eigenvalues of a regular matrix A is equal to 0.

Proposition 2.5.3. If $\mathbf{x}$ is an eigenvector of the regular matrix A corresponding to the eigenvalue λ , then the same vector $\mathbf{x}$ is as eigenvector of the inverse matrix A^{-1} corresponding to the eigenvalue $1/\lambda$.

To *prove* the assertion we multiply the both sides of the equality (6) on the left by the matrix A^{-1} . We get $A^{-1}A\mathbf{x} = A^{-1}\lambda\mathbf{x}$ or $A^{-1}\mathbf{x} = (1/\lambda)\mathbf{x}$. $\square$

Proposition 2.5.4. If $\mathbf{x}$ is an eigenvector of the matrix A corresponding to the eigenvalue λ , then the same vector $\mathbf{x}$ is an eigenvector of the matrix A^2 corresponding to the eigenvalue λ^2 .

Proof. This assertion follows from the chain:

$$A^2\mathbf{x} = A(A\mathbf{x}) = A(\lambda\mathbf{x}) = \lambda(A\mathbf{x}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x}. \quad \square$$

Problem 2.5.3.* Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of the matrix $A \in \mathbf{C}^{n \times n}$. Prove that $\lambda_1^k, \dots, \lambda_n^k$ are the eigenvalues of the matrix A^k ($k \in \mathbf{N}$).

Problem 2.5.4.* Prove that if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the matrix $A \in \mathbf{C}^{n \times n}$, then $\lambda_1 \pm \mu, \dots, \lambda_n \pm \mu$ are the eigenvalues of the matrix $A \pm \mu I$.

Proposition 2.5.5. The trace of the matrix A , i.e., the sum of the elements on the main diagonal, is equal to the sum of all eigenvalues of the matrix A .

To *prove* the assertion we will use the equality (10). In the expansion of the left side by the powers of the variable λ the coefficient by the power λ^{n-1} is $(-1)^{n-1}(a_{11} + a_{22} + \dots + a_{nn})$ and at the right side it is $(-1)^{n+1}(\lambda_1 + \lambda_2 + \dots + \lambda_n)$. $\square$

Example 2.5.2.* Suppose we know three eigenvalues $\lambda_1 = 4$, $\lambda_2 = 1$ and $\lambda_3 = 6$ of the matrix

$$A = \begin{bmatrix} 4 & 2 & 0 & 4 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & 3 & 3 \\ 0 & 4 & 0 & 7 \end{bmatrix}$$

Let us find the forth eigenvalue of the matrix A and its determinant. Since the trace of the matrix A equals the sum of all eigenvalues,

$$4 + 2 + 3 + 7 = 4 + 1 + 6 + \lambda_4 \Rightarrow \lambda_4 = 5.$$

Computing the determinant, we get

$$\det(A) = \lambda_1 \lambda_2 \lambda_3 \lambda_4 = 4 \cdot 1 \cdot 6 \cdot 5 = 120.$$

Problem 2.5.5.* Suppose we know three eigenvalues $\lambda_1 = 7$, $\lambda_2 = -7$ and $\lambda_3 = 21$ of the matrix

$$A = \begin{bmatrix} 67 & 266 & -30 & 64 \\ -24 & -91 & 12 & -20 \\ -6 & -42 & 10 & -12 \\ 42 & 126 & -21 & 21 \end{bmatrix}.$$

Find the forth eigenvalue of the matrix A and its determinant.

Proposition 2.5.6. The eigenvalues of both an upper triangular or a lower triangular matrix are the elements of the main diagonal.

Proof. Let us consider the case of an upper triangular matrix A . We form the characteristic equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Expanding the determinant we get from here

$$(a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda) = 0. \quad \square$$

Problem 2.5.6.* Find eigenvalues and eigenvectors of the matrix A , where

$$a) A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \quad b) A = \begin{bmatrix} 1 & 2 & 4 & -3 \\ 0 & 1 & 7 & 7 \\ 0 & 0 & 3 & 8 \\ 0 & 0 & 0 & -2 \end{bmatrix}.$$

Proposition 2.5.7. The eigenvectors of the matrix A corresponding to different eigenvalues are linearly independent.

Proof. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ be the eigenvectors of the matrix A corresponding to the different eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ ($k = 2 : n$). We will show that the system of these eigenvectors is linearly independent. Avoiding complexity we shall go through the proof in case $k = 2$. Let us suppose that the antithesis is valid, i.e., the vector system $\{\mathbf{x}_1, \mathbf{x}_2\}$ is linearly independent:

$$\exists(\alpha_1, \alpha_2) : \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 = 0 \quad \wedge \quad |\alpha_1| + |\alpha_2| \neq 0. \quad (11)$$

Multiplying the equality in (11) on the left by matrix A , we get

$$\alpha_1 A\mathbf{x}_1 + \alpha_2 A\mathbf{x}_2 = 0 \quad (12)$$

or

$$\alpha_1 \lambda_1 \mathbf{x}_1 + \alpha_2 \lambda_2 \mathbf{x}_2 = 0. \quad (13)$$

Multiplying the equality in (11) by λ_1 , and subtracting the result from (13), we get

$$\alpha_2(\lambda_2 - \lambda_1)\mathbf{x}_2 = 0.$$

On the left in this equality only the first factor α_2 can equal 0. Analogously, multiplying in (11) by λ_2 , we get the equality $\alpha_1 = 0$. So $|\alpha_1| + |\alpha_2| =$

0, and this is in contradiction with the assumption (11). Therefore, the system of eigenvectors $\{\mathbf{x}_1, \mathbf{x}_2\}$ is linearly independent. $\square$

Let us suppose that the system of eigenvectors $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of the matrix A is linearly independent. Let us form the $n \times n$ -matrix S , choosing the vector $\mathbf{x}_1$, as the first column-vector, the vector $\mathbf{x}_2$ as the second column-vector, $\dots$, the vector $\mathbf{x}_n$ as the n -th column-vector, i.e.,

$$S = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}. \quad (14)$$

Let us denote

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}. \quad (15)$$

For the above example 2.5.1, we get

$$S = \begin{bmatrix} -1 & -1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \wedge \quad \Lambda = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Proposition 2.5.8. If the matrix A has n linearly independent eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ corresponding to the eigenvalues $\lambda_1, \dots, \lambda_n$, then the matrix A can be expressed in form

$$A = SAS^{-1}, \quad (16)$$

where the matrices S and Λ are defined by (14) and (15).

For the *proof* it will suffice to show that

$$AS = S\Lambda. \quad (17)$$

Let us start from the left side of (17):

$$\begin{aligned} AS &= A \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} = \\ &= \begin{bmatrix} A\mathbf{x}_1 & \cdots & A\mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \lambda_1\mathbf{x}_1 & \cdots & \lambda_n\mathbf{x}_n \end{bmatrix}. \end{aligned}$$

From the right side of (17) we get:

$$S\Lambda = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} =$$

$$= \begin{bmatrix} \lambda_1 \mathbf{x}_1 & \cdots & \lambda_n \mathbf{x}_n \end{bmatrix}.$$

Therefore, equality (17) holds, and consequently equality (16), and also the equality

$$\Lambda = S^{-1}AS. \quad \square \quad (18)$$

Example 2.5.3.* Find a 3×3 -matrix A whose eigenvalues and corresponding eigenvectors are:

$$\lambda_1 = 3 \Rightarrow \mathbf{x}_1 = \begin{bmatrix} -3 & 2 & 1 \end{bmatrix}^T,$$

$$\lambda_2 = -2 \Rightarrow \mathbf{x}_2 = \begin{bmatrix} -2 & 1 & 0 \end{bmatrix}^T,$$

$$\lambda_3 = 1 \Rightarrow \mathbf{x}_3 = \begin{bmatrix} -6 & 3 & 1 \end{bmatrix}^T.$$

As the wanted matrix A can be represented in form $A = S\Lambda S^{-1}$, where

$$S = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix} \wedge \Lambda = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

then

$$A = \begin{bmatrix} -3 & -2 & -6 \\ 2 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -3 & -2 & -6 \\ 2 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix}^{-1} =$$

:

$$= \begin{bmatrix} -9 & 4 & -6 \\ 6 & -2 & 3 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 1 & 3 & -3 \\ -1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 6 & -18 \\ 1 & 0 & 9 \\ 2 & 4 & 1 \end{bmatrix}.$$

Problem 2.5.7.* Find a 2×2 -matrix A whose eigenvalues and corresponding eigenvectors are:

$$\lambda_1 = 1 \Rightarrow \mathbf{x}_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \wedge \lambda_2 = 2 \Rightarrow \mathbf{x}_2 = \begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

Problem 2.5.8.* Find a 3×3 -matrix A whose eigenvalues and corresponding eigenvectors are:

$$\lambda_1 = 3 \Rightarrow \mathbf{x}_1 = \begin{bmatrix} -1 & -1 & 1 \end{bmatrix}^T,$$

$$\lambda_2 = -3 \Rightarrow \mathbf{x}_2 = \begin{bmatrix} 2 & 1 & 0 \end{bmatrix}^T,$$

$$\lambda_3 = 5 \Rightarrow \mathbf{x}_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T.$$

Example 2.5.4.* Find matrices A^{100} and A^{155} , where

$$A = \begin{bmatrix} 41 & -30 \\ 56 & -41 \end{bmatrix}.$$

Since

$$\begin{vmatrix} 41 - \lambda & -30 \\ 56 & -41 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda^2 - 1 = 0$$

$$\lambda_1 = 1 \Rightarrow \mathbf{x}_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \wedge \lambda_2 = -1 \Rightarrow \mathbf{x}_2 = \begin{bmatrix} 5 \\ 7 \end{bmatrix},$$

and

$$A = S\Lambda S^{-1} \wedge \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \wedge S = \begin{bmatrix} 3 & 5 \\ 4 & 7 \end{bmatrix} \wedge S^{-1} = \begin{bmatrix} 7 & -5 \\ -4 & 3 \end{bmatrix},$$

then

$$\begin{aligned} A^{100} &= (S\Lambda S^{-1})(S\Lambda S^{-1}) \cdots (S\Lambda S^{-1}) = S\Lambda^{100}S^{-1} = \\ &= \begin{bmatrix} 3 & 5 \\ 4 & 7 \end{bmatrix} \begin{bmatrix} 1^{100} & 0 \\ 0 & (-1)^{100} \end{bmatrix} \begin{bmatrix} 7 & -5 \\ -4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I \end{aligned}$$

and

$$A^{155} = \begin{bmatrix} 3 & 5 \\ 4 & 7 \end{bmatrix} \begin{bmatrix} 1^{155} & 0 \\ 0 & (-1)^{155} \end{bmatrix} \begin{bmatrix} 7 & -5 \\ -4 & 3 \end{bmatrix} = \begin{bmatrix} 41 & -30 \\ 56 & -41 \end{bmatrix} = A.$$

Problem 2.5.9.* Find matrices A^{100} and A^{155} , where

$$a) A = \begin{bmatrix} -5 & 2 \\ -21 & 8 \end{bmatrix}, \quad b) A = \begin{bmatrix} -20 & 42 \\ -9 & 19 \end{bmatrix}.$$

Proposition 2.5.9. If all the eigenvalues of the matrices A and B are single and the matrices A and B are commutative, then they have common eigenvectors.

Proof. Let $\mathbf{x}$ be an eigenvector of the matrix A corresponding to the eigenvalue λ , i.e., it holds (6). Let us multiply both sides (6) on the left by

the matrix B . Due to the commutability of the matrices A and B , we get the chain:

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow B(A\mathbf{x}) = B(\lambda\mathbf{x}) \Leftrightarrow (BA)\mathbf{x} = \lambda(B\mathbf{x}) \Leftrightarrow A(B\mathbf{x}) = \lambda(B\mathbf{x}).$$

Thus, if $\mathbf{x}$ is an eigenvector of the matrix A corresponding to the eigenvalue λ , then $B\mathbf{x}$ is also an eigenvector of the matrix A corresponding to the single eigenvalue λ . If the eigenspace corresponding to the eigenvalue λ is a one-dimensional subspace in $\mathbf{R}^n$, then the vectors $\mathbf{x}$ and $B\mathbf{x}$ are collinear, i.e.,

$$\exists \mu : B\mathbf{x} = \mu\mathbf{x}.$$

Thus, the eigenvector $\mathbf{x}$ of the matrix A corresponding to the eigenvalue λ is also the eigenvector of the matrix B corresponding to the eigenvalue μ . Analogously one can show that each eigenvector of the matrix B is an eigenvector of the matrix A $\square$

Proposition 2.5.10. If the matrices $A, B \in \mathbf{C}^{n \times n}$ have n common linearly independent eigenvectors, then these matrices are commutative.

Proof. Due to proposition 2.5.8, these matrices can be expressed in form

$$A = S\Lambda S^{-1}, \quad B = S\Upsilon S^{-1}, \quad (19)$$

where S is the matrix formed of the eigenvectors as column-vectors, and Λ is a diagonal matrix with eigenvalues of the matrix A on the main diagonal, and Υ is a diagonal matrix with the eigenvalues of the matrix B on the main diagonal. Let us find the products AB and BA , using the representation in (19):

$$AB = S\Lambda S^{-1}S\Upsilon S^{-1} = S\Lambda\Upsilon S^{-1}$$

and

$$BA = S\Upsilon S^{-1}S\Lambda S^{-1} = S\Upsilon\Lambda S^{-1}.$$

As the diagonal matrices Λ and Υ are commutative, $AB = BA$, q.e.d. $\square$

1.2.6 Schur's Decomposition

The eigenvector $\mathbf{x}$ of the matrix $A \in \mathbf{C}^{n \times n}$ determines in the space $\mathbf{C}^n$ a one-dimensional subspace that is invariant with respect to the multiplication by the matrix A on the left.

Definition 2.6.1. The subspace $\mathbf{S} \subseteq \mathbf{C}^n$ is called invariant with respect to the multiplication by the matrix A on the left if $\mathbf{x} \in \mathbf{S} \Rightarrow A\mathbf{x} \in \mathbf{S}$.

Proposition 2.6.1. If $A \in \mathbf{C}^{n \times n}$, $B \in \mathbf{C}^{k \times k}$, $X \in \mathbf{C}^{n \times k}$ and $AX = XB$, then the space $\mathcal{R}(X)$ of the matrix X is invariant with respect to the multiplication by the matrix A on the left and the space $\mathcal{R}(X^T)$ is invariant with respect to the multiplication by the matrix B on the right. In addition, the following connections

$$\dim \mathcal{R}(X) = k \Rightarrow \lambda(B) \subseteq \lambda(A)$$

and

$$\dim \mathcal{R}(X) = k = n \Rightarrow \lambda(B) = \lambda(A).$$

holds.

Proof. If $X = [\mathbf{c}_1 \dots \mathbf{c}_k]$, then

$$AX = A[\mathbf{c}_1 \dots \mathbf{c}_k] = [A\mathbf{c}_1 \dots A\mathbf{c}_k]$$

and

$$\begin{aligned} XB &= [\mathbf{c}_1 \dots \mathbf{c}_k] \begin{bmatrix} b_{1;1} & \dots & b_{1;k} \\ \vdots & & \vdots \\ b_{k;1} & \dots & b_{k;k} \end{bmatrix} = \\ &= \left[b_{1;1}\mathbf{c}_1 + \dots + b_{k;1}\mathbf{c}_k \quad \dots \quad b_{1;k}\mathbf{c}_1 + \dots + b_{k;k}\mathbf{c}_k \right] \end{aligned}$$

and

$$A\mathbf{c}_i = b_{1;i}\mathbf{c}_1 + \dots + b_{k;i}\mathbf{c}_k \quad (i = 1 : k) \Rightarrow A\mathcal{R}(X) \subseteq \mathcal{R}(X).$$

Therefore, the space $\mathcal{R}(X)$ of the column-vectors of the matrix X is invariant with respect to the multiplication by the matrix A on the left. Analogously, one can prove that the space $\mathcal{R}(X^T)$ of the row-vectors is invariant with respect to the multiplication by the matrix B on the right. If $B\mathbf{y} = \lambda\mathbf{y}$, then

$$A(X\mathbf{y}) = (XB)\mathbf{y} = X\lambda\mathbf{y} = \lambda(X\mathbf{y}),$$

i.e., λ is an eigenvalue of the matrix A if λ is a eigenvalue of the matrix B . Naturally,

$$\mathbf{y} \neq \mathbf{0} \xRightarrow{\dim \mathcal{R}(X)=k} X\mathbf{y} \neq \mathbf{0}.$$

therefore, if the column-vectors of the matrix X are linearly independent, then $\lambda(B) \subseteq \lambda(A)$. If $A\mathbf{z} = \lambda\mathbf{z}$ and X is a regular square matrix ($\dim \mathcal{R}(X) = k = n$), then it follows from the equality $AX = XB$ that $A = XBX^{-1}$ and

$$XBX^{-1}\mathbf{z} = \lambda\mathbf{z} \Leftrightarrow B(X^{-1}\mathbf{z}) = \lambda(X^{-1}\mathbf{z}),$$

i.e., every eigenvalue of the matrix A is an eigenvalue of the matrix B , $\lambda(A) \subseteq \lambda(B)$, and thus $\lambda(B) = \lambda(A)$. $\square$

Definition 2.6.2. Matrices $A, B \in \mathbf{C}^{n \times n}$ are said to be *similar* if there exists a regular matrix $X \in \mathbf{C}^{n \times n}$ such that $A = XBX^{-1}$.

Due to proposition 2.6.1 (the last assertion), the spectrum of two similar matrices are equal. We can get this result also directly:

$$\begin{aligned} \det(A - \lambda I) &= \det(XBX^{-1} - X\lambda IX^{-1}) = \\ &= \det(X(B - \lambda I)X^{-1}) = \det(X) \det(B - \lambda I) \det(X^{-1}). \end{aligned}$$

Problem 2.6.1.* Are the matrices A and B similar if

$$a) A = \begin{bmatrix} 1 & i & 0 \\ i & 2 & -1 \\ 0 & i & 1 \end{bmatrix} \wedge B = \begin{bmatrix} 1+i & 7 & 2 \\ 0 & 1 & 9 \\ 0 & 0 & 2-i \end{bmatrix},$$

$$b) A = \begin{bmatrix} 25+25i & 25 & 100 \\ 25 & 100 & 25+25i \\ 25+25i & 25 & 100 \end{bmatrix} \wedge B = \begin{bmatrix} 100 & 35+20i & -5+15i \\ 95+15i & 76+16i & -43+12i \\ 40-20i & -43+12i & 49+9i \end{bmatrix}?$$

Proposition 2.6.2. If $T \in \mathbf{C}^{n \times n}$ and

$$T = \begin{bmatrix} T_{1;1} & T_{1;2} \\ 0 & T_{2;2} \end{bmatrix} \begin{matrix} p \\ q \end{matrix}.$$

then $\lambda(T) = \lambda(T_{1;1}) \cup \lambda(T_{2;2})$.

Proof. If $T\mathbf{x} = \lambda\mathbf{x}$, i.e., $\lambda \in \lambda(T)$, $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$, $\mathbf{x}_1 \in \mathbf{C}^p$ and $\mathbf{x}_2 \in \mathbf{C}^q$, then

$$\begin{bmatrix} T_{1;1} & T_{1;2} \\ 0 & T_{2;2} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \Rightarrow \begin{cases} T_{1;1}\mathbf{x}_1 + T_{1;2}\mathbf{x}_2 = \lambda\mathbf{x}_1 \\ T_{2;2}\mathbf{x}_2 = \lambda\mathbf{x}_2 \end{cases}.$$

If $\mathbf{x}_2 \neq \mathbf{0}$, then

$$T_{2,2}\mathbf{x}_2 = \lambda\mathbf{x}_2 \Rightarrow \lambda \in \lambda(T_{2,2}).$$

If $\mathbf{x}_2 = \mathbf{0}$, then

$$T_{1,1}\mathbf{x}_1 = \lambda\mathbf{x}_1 \Rightarrow \lambda \in \lambda(T_{1,1}).$$

Thus,

$$\lambda(T) \subset \lambda(T_{1,1}) \cup \lambda(T_{2,2}).$$

Since the potencies of the sets $\lambda(T_{1,1}) \cup \lambda(T_{2,2})$ and $\lambda(T)$ are equal, then the proposition holds. $\square$

Example 2.6.1. Using proposition 2.6.2, let us find the spectrum of the matrix

$$A = \begin{bmatrix} 1 & 1 & 5 & 6 \\ -1 & 1 & 7 & 3 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & -4 & 3 \end{bmatrix}.$$

First, we find the eigenvalues of the matrices $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ and $\begin{bmatrix} 2 & 1 \\ -4 & 3 \end{bmatrix}$:

$$\begin{vmatrix} 1-\lambda & 1 \\ -1 & 1-\lambda \end{vmatrix} = 0 \Rightarrow \begin{cases} \lambda_1 = 1+i \\ \lambda_2 = 1-i \end{cases},$$

$$\begin{vmatrix} 2-\lambda & 1 \\ -4 & 3-\lambda \end{vmatrix} = 0 \Rightarrow \begin{cases} \lambda_3 = (5+i\sqrt{15})/2 \\ \lambda_4 = (5-i\sqrt{15})/2 \end{cases}.$$

Thus, the spectrum of the matrix is $\lambda(A) = \{1+i; 1-i; (5+i\sqrt{15})/2; (5-i\sqrt{15})/2\}$.

Problem 2.6.2.* Find by the use of proposition 2.6.2 the spectrum of the matrix A if

$$a) A = \begin{bmatrix} 2 & -3 & 17 & 36 \\ 4 & 6 & 11 & -13 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 3 & 8 \end{bmatrix}, \quad b) A = \begin{bmatrix} 2 & 17 & -2 \\ 0 & -2 & -1 \\ 0 & 5 & 2 \end{bmatrix}.$$

Definition 2.6.3. A matrix $Q \in \mathbf{C}^{n \times n}$ is called a *unitary matrix* if $Q^H Q = Q Q^H = I$.

Problem 2.6.3.* Is the matrix Q a unitary matrix if

$$\begin{aligned} a) \quad Q &= \begin{bmatrix} -\frac{1}{2} & \frac{1}{2}\sqrt{3} \\ \frac{1}{2}\sqrt{3} & \frac{1}{2} \end{bmatrix}, \quad b) \quad Q = \begin{bmatrix} \cos x & i \sin x \\ i \sin x & \cos x \end{bmatrix}, \\ c) \quad Q &= \begin{bmatrix} \frac{2}{5}\sqrt{5} & \frac{1}{5}i\sqrt{5} \\ -\frac{1}{5}\sqrt{5} & \frac{2}{5}i\sqrt{5} \end{bmatrix}. \end{aligned}$$

Proposition 2.6.3 (the QR factorisation theorem) . If $A \in \mathbf{C}^{m \times n}$, then the matrix A can be expressed in form $A = QR$, where matrix $Q \in \mathbf{C}^{m \times m}$ is unitary matrix and matrix $R \in \mathbf{C}^{m \times n}$ is an upper triangular matrix.

Proposition 2.6.4. If $A \in \mathbf{C}^{n \times n}$, $B \in \mathbf{C}^{p \times p}$, $X \in \mathbf{C}^{n \times p}$,

$$AX = XB \tag{20}$$

and $\text{rank}(X) = p$, then there exists a unitary matrix $Q \in \mathbf{C}^{n \times n}$ such that

$$Q^H A Q = T = \begin{bmatrix} T_{1;1} & T_{1;2} \\ 0 & T_{2;2} \end{bmatrix} \quad \begin{matrix} p \\ n-p \end{matrix},$$

where $\lambda(T_{1;1}) = \lambda(A) \cap \lambda(B)$.

Proof. Let us consider for the matrix X its QR factorization $X = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$, where $Q \in \mathbf{C}^{n \times n}$ and $R_1 \in \mathbf{C}^{p \times p}$. Substracting the factorization of the matrix X into equality (20), we get

$$AQ \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} B \Leftrightarrow Q^H A Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} B.$$

The spectrums of the matrices $Q^H A Q$ and A coincide, i.e. $\lambda(Q^H A Q) = \lambda(A)$. Representing the matrix A in form

$$Q^H A Q = \begin{bmatrix} T_{1;1} & T_{1;2} \\ T_{2;1} & T_{2;2} \end{bmatrix} \quad \begin{matrix} p \\ n-p \end{matrix},$$

we find

$$\begin{bmatrix} T_{1;1} & T_{1;2} \\ T_{2;1} & T_{2;2} \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} R_1 B \\ 0 \end{bmatrix} \Rightarrow \begin{cases} T_{1;1} R_1 = R_1 B \\ T_{2;1} R_1 = 0 \end{cases} \xrightarrow[\det R_1 \neq 0]{\text{prop. 2.6.1}} \begin{cases} \lambda(T_{1;1}) = \lambda(B) \\ T_{2;1} = 0 \end{cases}.$$

Therefore, the proposition holds. $\square$

Remark 2.6.1. Proposition 2.6.4 makes it possible, if we know an invariant subspace of the given matrix, to transform it by unitary similarity transformations into a triangular block form.

Proposition 2.6.5 (Schur's decomposition). If $A \in \mathbf{C}^{n \times n}$, then there exists a unitary matrix $Q \in \mathbf{C}^{n \times n}$ such that

$$Q^H A Q = T = D + N, \quad (21)$$

where $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ and $N \in \mathbf{C}^{n \times n}$ is a strictly upper triangular matrix, i.e., an upper triangular matrix with zeros on the main diagonal. The matrix Q can be formed so that the eigenvalues of the matrix A are in the given order on the main diagonal of the matrix D .

To *prove* this assertion we will use the method of complete induction. As the assertion holds for $n = 1$, the base for the induction exists. We are going now to show the admissibility of the induction steps. We suppose that the assertion holds for the matrices whose order is less or equal to $k - 1$. Let us show that the assertion will be valid for k , too. If $A\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{x} \neq \mathbf{0}$, then, by lemma 2.6.4, choosing $X = \mathbf{x}$, $B = \lambda$, there exists a unitary matrix U such that

$$U^H A U = T = \begin{bmatrix} \lambda & w^H \\ 0 & C \end{bmatrix} \begin{matrix} 1 \\ k-1 \end{matrix},$$

Since $C \in \mathbf{C}^{(k-1) \times (k-1)}$, the assertion is valid for this matrix, i.e., there exists a unitary matrix $\hat{U}$ such that $\hat{U}^H C \hat{U}$ is an upper triangular matrix. If $Q = U \text{diag}(1; \hat{U})$, then

$$\begin{aligned} Q^H A Q &= \begin{bmatrix} 1 & 0 \\ 0 & \hat{U}^H \end{bmatrix} U^H A U \begin{bmatrix} 1 & 0 \\ 0 & \hat{U} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \hat{U}^H \end{bmatrix} \begin{bmatrix} \lambda & w^H \\ 0 & C \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \hat{U} \end{bmatrix} = \\ &= \begin{bmatrix} \lambda & w^H \\ 0 & \hat{U}^H C \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \hat{U} \end{bmatrix} = \begin{bmatrix} \lambda & w^H \hat{U} \\ 0 & \hat{U}^H C \hat{U} \end{bmatrix}, \end{aligned}$$

and so the matrix $Q^H A Q$ is an upper triangular matrix. $\square$

Example 2.6.2. Let

$$A = \begin{bmatrix} 3 & 8 \\ -2 & 3 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} 2i/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & -2i/\sqrt{5} \end{bmatrix}.$$

Let us show that Q is unitary matrix. For this we, first, find the product $Q^H A Q$. The checking of the matrix Q for unitarity gives:

$$Q^H Q = \begin{bmatrix} -2i/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2i/\sqrt{5} \end{bmatrix} \begin{bmatrix} 2i/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & -2i/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$Q Q^H = \begin{bmatrix} 2i/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & -2i/\sqrt{5} \end{bmatrix} \begin{bmatrix} -2i/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2i/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

$$Q Q^H = \begin{bmatrix} 2i/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & -2i/\sqrt{5} \end{bmatrix} \begin{bmatrix} -2i/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2i/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Let us find the product

$$\begin{aligned} Q^H A Q &= \begin{bmatrix} -2i/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2i/\sqrt{5} \end{bmatrix} \begin{bmatrix} 3 & 8 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} 2i/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & -2i/\sqrt{5} \end{bmatrix} = \\ &= \begin{bmatrix} 3 + 4i & -6 \\ 0 & 3 - 4i \end{bmatrix}. \end{aligned}$$

Consequently, we have obtained the Schur decomposition of the matrix A .

Now (21) can be represented in the form $AQ = QT$. Replacing $Q = [\mathbf{q}_1 \cdots \mathbf{q}_n]$, where the vectors $\mathbf{q}_i$ are called *Schur vectors*, into the last equality, we get

$$A[\mathbf{q}_1 \cdots \mathbf{q}_n] = [\mathbf{q}_1 \cdots \mathbf{q}_n]T$$

or

$$\begin{aligned} &[A\mathbf{q}_1 \cdots A\mathbf{q}_n] = \\ &= \begin{bmatrix} \lambda_1 \mathbf{q}_1 & \lambda_2 \mathbf{q}_2 + n_{1,2} \mathbf{q}_1 & \cdots & \lambda_n \mathbf{q}_n + n_{1,n} \mathbf{q}_1 + n_{2,n} \mathbf{q}_2 + \cdots + n_{n-1,n} \mathbf{q}_{n-1} \end{bmatrix} \end{aligned}$$

or

$$A\mathbf{q}_i = \lambda_i \mathbf{q}_i + n_{1,i} \mathbf{q}_1 + \cdots + n_{i-1,i} \mathbf{q}_{i-1} = \lambda_i \mathbf{q}_i + \sum_{k=1}^{i-1} n_{ki} \mathbf{q}_k \quad (i = 1: n).$$

From this equality it turns out that all subspaces $S_k = \text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ ($k = 1: n$) are invariant with respect to multiplication by the matrix A on the left, and Schur vector $\mathbf{q}_i$ is an eigenvector of the matrix A if and only if in the i -th column of the matrix N there are only zeros.

Definition 2.6.4. If $A \in \mathbf{C}^{n \times n}$ and $A^H A = A A^H$, then A is called a *normal matrix*.

Exercise 2.6.4.* Check the normality of A if

$$a) A = \begin{bmatrix} 1 & -1 & -1 \\ 1 & i & 1 \\ -1 & -1 & 1 \end{bmatrix}, \quad b) A = \begin{bmatrix} i & -1 & i \\ 1 & i & 1 \\ i & -1 & i \end{bmatrix}, \quad c) A = \begin{bmatrix} i & i & i \\ -i & i & -i \\ i & i & i \end{bmatrix}.$$

Proposition 2.6.6. A matrix $A \in \mathbf{C}^{n \times n}$ is normal matrix iff there exists a unitary matrix $Q \in \mathbf{C}^{n \times n}$, satisfying the condition $Q^H A Q = D = \text{diag}(\lambda_1, \dots, \lambda_n)$.

Proof. If the matrix A is unitarily similar to the diagonal matrix D , then

$$Q^H A Q = D \Leftrightarrow A = Q D Q^H \Rightarrow A^H A = Q D^H Q^H Q D Q^H = Q D^H D Q^H \wedge$$

$$A A^H = Q D Q^H Q D^H Q^H = Q D D^H Q^H$$

and since diagonal matrices are commutative, then $A^H A = A A^H$ and the matrix A is normal. Vice versa, if the matrix A is normal and the Schur decomposition of this matrix is $Q^H A Q = T$, then T is also normal because

$$T^H T = Q^H A^H Q Q^H A Q = Q^H A^H A Q$$

and

$$T T^H = Q^H A Q Q^H A^H Q = Q^H A A^H Q.$$

Since a triangular matrix is normal only if it is a diagonal matrix, then it has been proved that a unitary matrix is similar to a diagonal matrix. $\square$

Proposition 2.6.7 (block-diagonal-decomposition). Let

$$Q^H A Q = T = \begin{bmatrix} T_{1;1} & T_{1;2} & \cdots & T_{1;q} \\ 0 & T_{2;2} & \cdots & T_{2;q} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & T_{q;q} \end{bmatrix}$$

be the Schur decomposition of the matrix $A \in \mathbf{C}^{n \times n}$, where the blocks $T_{i;i}$ are square matrices. If $\lambda(T_{i;i}) \cap \lambda(T_{j;j}) = \emptyset$ ($i \neq j$), then there exists a regular matrix $Y \in \mathbf{C}^{n \times n}$ such that

$$(QY)^{-1}A(QY) = \text{diag}(T_{1;1}, \dots, T_{q;q}).$$

Corollary 2.6.1. If $A \in \mathbf{C}^{n \times n}$, then there exists a regular matrix X such that

$$X^{-1}AX = \text{diag}(\lambda_1 I + N_1, \dots, \lambda_q I + N_q) \quad N_i \in \mathbf{C}^{n_i \times n_i},$$

where $\lambda_1, \dots, \lambda_q$, $n_1 + \dots + n_q = n$ and each N_i is a strictly upper triangular matrix.

Proposition 2.6.8 (Jordan decomposition). If $A \in \mathbf{C}^{n \times n}$, then there exists a regular matrix $X \in \mathbf{C}^{n \times n}$ such that $X^{-1}AX = J = \text{diag}(J_1, \dots, J_t)$, where $m_1 + \dots + m_t = n$,

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & \lambda_i \end{bmatrix}$$

is an $m_i \times m_i$ Jordan block, and the matrix J is called the *Jordan normal form* of the matrix A .

Proof. See Lancaster (1982, p. 143).

Example 2.6.3. Using "Maple", we find the Jordan decompositions $A = XJX^{-1}$ of two matrices:

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

and

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

1.2.7 Norms and Condition Numbers of a Matrix

Definition 2.7.1. A mapping $f : \mathbf{R}^{m \times n} \rightarrow \mathbf{R}$ is called the *norming of a matrix* and the obtained value the *matrix norm* if the following three conditions are satisfied:

$$\begin{aligned} f(A) &\geq 0 & A &\in \mathbf{R}^{m \times n}, & (f(A) = 0 \Leftrightarrow A = 0) \\ f(A + B) &\leq f(A) + f(B) & A, B &\in \mathbf{R}^{m \times n}, \\ f(\alpha A) &= |\alpha| f(A) & \alpha &\in \mathbf{R}, A \in \mathbf{R}^{m \times n}. \end{aligned}$$

The matrix norm will be denoted $f(A) = \|A\|$.

The most frequently used norms in linear algebra are the *Frobenius norm*

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} \quad (22)$$

and the *p-norms* ($p \geq 1$),

$$\|A\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}. \quad (23)$$

From (23) it follows that

$$\|A\|_p \stackrel{\mathbf{x} \neq \mathbf{0}}{\geq} \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p$$

or

$$\|A\mathbf{x}\|_p \stackrel{\mathbf{x} \neq \mathbf{0}}{\leq} \|A\|_p \|\mathbf{x}\|_p \quad (24)$$

Let us verify that *p*-norm satisfies the conditions of the matrix norm. We find that

$$\|A\mathbf{x}\|_p \geq 0 \wedge \|\mathbf{x}\|_p > 0 \Rightarrow \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p \geq 0 \Rightarrow \|A\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p \geq 0,$$

$$\|A\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p = 0 \Leftrightarrow \|A\mathbf{x}\|_p = 0 \quad \forall \mathbf{x} \in \mathbf{R}^n \Leftrightarrow A = 0,$$

further

$$\|A + B\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \|(A + B)\mathbf{x}\|_p / \|\mathbf{x}\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \|A\mathbf{x} + B\mathbf{x}\|_p / \|\mathbf{x}\|_p \leq$$

$$\begin{aligned} \leq \sup_{\mathbf{x} \neq \mathbf{0}} (\|A\mathbf{x}\|_p + \|B\mathbf{x}\|_p) / \|\mathbf{x}\|_p &\leq \sup_{\mathbf{x} \neq \mathbf{0}} \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p + \sup_{\mathbf{x} \neq \mathbf{0}} \|B\mathbf{x}\|_p / \|\mathbf{x}\|_p = \\ &= \|A\|_p + \|B\|_p \end{aligned}$$

and

$$\begin{aligned} \|\alpha A\|_p &= \sup_{\mathbf{x} \neq \mathbf{0}} \|(\alpha A)\mathbf{x}\|_p / \|\mathbf{x}\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} |\alpha| \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p = \\ &= |\alpha| \sup_{\mathbf{x} \neq \mathbf{0}} \|A\mathbf{x}\|_p / \|\mathbf{x}\|_p = |\alpha| \|A\|_p . \end{aligned}$$

Exercise 2.7.1. Verify that the Frobenius norm satisfies the conditions of the matrix norm.

Exercise 2.7.2.* Compute the Frobenius norm $\|A\|_F$ if

$$a) A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 4 \\ 2 & 1 & 3 \end{bmatrix}, \quad b) A = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 3 & 0 & 5 & 4 \\ 1 & 1 & 1 & 2 \\ 1 & 3 & 2 & 2 \end{bmatrix}, \quad c) A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 0 & 1 & 0 \\ 3 & 4 & 3 & 4 & 3 \\ 5 & 5 & 5 & 5 & 5 \end{bmatrix}.$$

Definition 2.7.2. For the fixed matrix norm the value

$$k(A) = \|A\| \|A^{-1}\|$$

is called the *condition number* corresponding to the regular square matrix $A \in \mathbf{R}^{n \times n}$.

The condition number corresponding to the Frobenius norm will be denoted $k_F(A)$ and the condition number corresponding to the p -norm will be denoted $k_p(A)$. For a singular square matrix $A \in \mathbf{R}^{n \times n}$ we will define $k(A) = +\infty$.

Exercise 2.7.2. Show that if $A \in \mathbf{R}^{n \times n}$, then

$$\begin{aligned} \frac{1}{n} k_2(A) &\leq k_1(A) \leq n k_2(A), \\ \frac{1}{n} k_\infty(A) &\leq k_2(A) \leq n k_\infty(A), \\ \frac{1}{n^2} k_1(A) &\leq k_\infty(A) \leq n^2 k_1(A). \end{aligned}$$

Proposition 2.7.1. Rule (23) for the calculation of the norm $\|A\|_p$ can be transformed to the form

$$\|A\|_p = \sup_{\|x\|_p=1} \|Ax\|_p. \quad (25)$$

Proof. Using the third property of the norm and the homogeneity of multiplication of a vector by a matrix, we have

$$\frac{\|Ax\|_p}{\|x\|_p} = \left\| \frac{1}{\|x\|_p} Ax \right\|_p = \left\| A \frac{x}{\|x\|_p} \right\|_p,$$

where $\left\| \frac{x}{\|x\|_p} \right\|_p = 1$. $\square$

Proposition 2.7.2. If $A \in \mathbf{R}^{m \times n}$, $B \in \mathbf{R}^{n \times q}$ and $p \geq 1$, then $\|AB\|_p \leq \|A\|_p \|B\|_p$.

Proof. Using (24) and (25), we find that

$$\begin{aligned} \|AB\|_p &= \sup_{\|x\|_p=1} \|(AB)x\|_p = \sup_{\|x\|_p=1} \|A(Bx)\|_p \leq \sup_{\|x\|_p=1} \|A\|_p \|Bx\|_p = \\ &= \|A\|_p \sup_{\|x\|_p=1} \|Bx\|_p = \|A\|_p \|B\|_p. \quad \square \end{aligned}$$

Remark 2.7.1. Since $k_p(A) = \|A\|_p \|A^{-1}\|_p \geq \|AA^{-1}\|_p = \|I\|_p = 1$, then always $k_p(A) \geq 1$.

Remark 2.7.2. For each $A \in \mathbf{R}^{m \times n}$ and $\mathbf{x} \in \mathbf{R}^n$ and for arbitrary vector norm $\|\cdot\|_\alpha$ on $\mathbf{R}^n$ and $\|\cdot\|_\beta$ on $\mathbf{R}^m$ the relation

$$\|A\mathbf{x}\|_\beta \leq \|A\|_{\alpha,\beta} \|\mathbf{x}\|_\alpha,$$

holds, where $\|A\|_{\alpha,\beta}$ is a matrix norm defined by

$$\|A\|_{\alpha,\beta} = \sup_{\mathbf{x} \neq \mathbf{0}} \|A\mathbf{x}\|_\beta / \|\mathbf{x}\|_\alpha.$$

Since the set $\{\mathbf{x} \in \mathbf{R}^n : \|\mathbf{x}\|_\alpha = 1\}$ is compact and $\|\cdot\|_\beta$ is continuous, it follows that

$$\|A\|_{\alpha,\beta} = \max_{\|\mathbf{x}\|_\alpha=1} \|A\mathbf{x}\|_\beta = \|A\mathbf{x}^*\|_\beta$$

for some $\mathbf{x}^* \in \mathbf{R}^n$ with $\|\mathbf{x}^*\|_\alpha = 1$.

Definiton 2.7.3. If $k(A)$ is relatively small, then the matrix A is called a *well-conditioned matrix*, but if $k(A)$ is great, then an *ill-conditioned matrix*.

Definition 2.7.4. A norm $\|A\|$ of the square matrix A is said to be *consistent* with the vector norm $\|\mathbf{x}\|$ if

$$\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$$

and it is submultiplicative, i.e.,

$$\|AB\| \leq \|A\| \|B\|.$$

Definition 2.7.5. The norm $\|A\|$ of the square matrix, consistent with the vector norm $\|\mathbf{x}\|$ is said to be *subordinate to the vector norm $\|\mathbf{x}\|$* if for any matrix A there exists a vector $\mathbf{x} = \mathbf{x}(A) \neq \mathbf{0}$ such that $\|A\mathbf{x}\| = \|A\| \|\mathbf{x}\|$.

Proposition 2.7.3. For arbitrary vector norm $\|\mathbf{x}\|$ there exists at least one matrix norm $\|A\|$ subordinate (and thus at least one consistent) to this vector norm, and this is

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Remark 2.7.3. Not all matrix norms satisfy the submultiplicative property $\|AB\| \leq \|A\| \|B\|$. For example, if we define $\|A\|_{\Delta} = \max_{i,j} |a_{ij}|$, then

for the matrices $A = B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ we have $\|A\|_{\Delta} = \|B\|_{\Delta} = 1$ and

$$\|AB\|_{\Delta} = \left\| \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \right\|_{\Delta} = 2 > \|A\|_{\Delta} \|B\|_{\Delta}.$$

Proposition 2.7.4. If $A \in \mathbf{R}^{m \times n}$, then the following relations between the matrix norms hold:

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|, \quad (26)$$

$$\|A\|_{\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|, \quad (27)$$

$$\begin{aligned}
\|A\|_2 &\leq \|A\|_F \leq \sqrt{n} \|A\|_2, \\
\|A\|_\Delta &\leq \|A\|_2 \leq \sqrt{mn} \|A\|_\Delta, \\
\frac{1}{\sqrt{n}} \|A\|_\infty &\leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty, \\
\frac{1}{\sqrt{n}} \|A\|_1 &\leq \|A\|_2 \leq \sqrt{n} \|A\|_1.
\end{aligned}$$

If $A \in \mathbf{C}^{m \times n}$, $1 \leq i_1 \leq i_2 \leq m$ and $1 \leq j_1 \leq j_2 \leq n$, then

$$\|A(i_1 : i_2, j_1 : j_2)\|_p \leq \|A\|_p.$$

Prove the relation (27). We have

$$\begin{aligned}
\|Ax\|_\infty &= \max_i \left| \sum_{j=1}^n a_{ij} \xi_j \right| \leq \max_i \sum_{j=1}^n |a_{ij}| |\xi_j| \leq \\
&\leq \max_i \sum_{j=1}^n |a_{ij}| \|x\|_\infty = \|x\|_\infty \sum_{j=1}^n |a_{kj}|,
\end{aligned}$$

where we suppose that maximum has been gained if the index i obtains the value k . We have the estimation

$$\|A\|_\infty \leq \max_i \sum_{j=1}^n |a_{ij}|.$$

Let

$$z = \begin{pmatrix} \varsigma_1 & \dots & \varsigma_n \end{pmatrix}^T$$

and

$$\varsigma_j = \begin{cases} 1, & \text{if } a_{kj} \geq 0; \\ -1, & \text{if } a_{kj} < 0. \end{cases}$$

Since $\|z\|_\infty = 1$, then

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_p \geq \|Az\|_\infty = \max_i \left| \sum_{j=1}^n a_{ij} \varsigma_j \right| \geq \sum_{j=1}^n a_{kj} \varsigma_k = \sum_{j=1}^n |a_{kj}|,$$

and thus

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|. \quad \square$$

Example 2.7.1. Let us calculate the norms $\|A\|_1$ and $\|A\|_\infty$ for the matrix A if

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

From (26) and (27) we find that

$$\left\| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right\|_1 =$$

$$= \max(|a_{11}| + |a_{21}| + |a_{31}|, |a_{12}| + |a_{22}| + |a_{32}|, |a_{13}| + |a_{23}| + |a_{33}|)$$

and

$$\left\| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right\|_\infty =$$

$$= \max(|a_{11}| + |a_{12}| + |a_{13}|, |a_{21}| + |a_{22}| + |a_{23}|, |a_{31}| + |a_{32}| + |a_{33}|).$$

Example 2.7.2. Let us calculate the inverse matrix A^{-1} of the matrix A , the norms $\|A\|_1$, $\|A^{-1}\|_1$, $\|A\|_2$, $\|A^{-1}\|_2$, $\|A\|_\infty$, $\|A^{-1}\|_\infty$ and the condition numbers of matrix A $k_1(A)$, $k_2(A)$, $k_\infty(A)$ if

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

It follows that

$$A^{-1} = \begin{bmatrix} -\frac{3}{4} & -\frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{2} & -1 & -\frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{2} & -\frac{3}{4} \end{bmatrix},$$

$$\|A\|_1 = \|A\|_\infty = 4, \|A\|_2 = 2 + \sqrt{2}, \|A^{-1}\|_1 = \|A^{-1}\|_\infty = 2$$

and

$$\|A^{-1}\|_2 = 1 + \frac{1}{2}\sqrt{2}, \quad k_1(A) = k_\infty(A) = 2, \quad k_2(A) = \frac{1}{2}(2 + \sqrt{2})^2.$$

If formulae (26) and (27) enable us to calculate easily 1-norm and ∞ -norm, respectively, then the calculation of the 2-norm is more complicated. The matrix 2-norm is called also the *matrix spectral norm*.

Proposition 2.7.5. If $A \in \mathbf{R}^{m \times n}$, then

$$\|A\|_2 = \sqrt{\max_{\mu \in \lambda(A^T A)} \mu},$$

i.e., $\|A\|_2$ is the square root of the largest eigenvalue of $A^T A$.

Proof. To calculate $\|A\|_2$, we find first $\|A\|_2^2$. Thus,

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2 \Rightarrow \|A\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2^2 = \max_{\mathbf{x}^T \mathbf{x}=1} \mathbf{x}^T A^T A \mathbf{x}.$$

Let $A^T A = B \in \mathbf{R}^{n \times n}$. The matrix B is a symmetric matrix because

$$B^T = (A^T A)^T = A^T A$$

and

$$\begin{aligned} \mathbf{x}^T A^T A \mathbf{x} &= \mathbf{x}^T B \mathbf{x} = \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix} \begin{bmatrix} b_{1;1} & \cdots & b_{1;n} \\ \vdots & \cdots & \vdots \\ b_{n;1} & \cdots & b_{n;n} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = \\ &= \begin{bmatrix} \xi_1 & \cdots & \xi_n \end{bmatrix} \begin{bmatrix} b_{1;1}\xi_1 + \cdots + b_{1;n}\xi_n \\ \cdots & \cdots & \cdots \\ b_{n;1}\xi_1 + \cdots + b_{n;n}\xi_n \end{bmatrix} = \\ &= \left[\xi_1 \sum_{j=1}^n b_{1;j}\xi_j + \cdots + \xi_n \sum_{j=1}^n b_{n;j}\xi_j \right], \\ &\quad \mathbf{x}^T \mathbf{x} = \sum_{j=1}^n \xi_j^2, \end{aligned}$$

then $\mathbf{x}^T A^T A \mathbf{x}$ is a function of n variables $\xi_1, \dots, \xi_n$ and

$$\begin{aligned} \frac{\partial (\mathbf{x}^T A^T A \mathbf{x})}{\partial \xi_i} &= \sum_{j=1}^n b_{i;j}\xi_j + \sum_{j=1}^n b_{j;i}\xi_j \stackrel{b_{i;j}=b_{j;i}}{=} 2 \sum_{j=1}^n b_{i;j}\xi_j = 2 [A^T A \mathbf{x}]_i, \\ \frac{\partial (\mathbf{x}^T \mathbf{x})}{\partial \xi_i} &= 2\xi_i = 2 [\mathbf{x}]_i. \end{aligned}$$

The problem of finding $\max_{\mathbf{x}^T \mathbf{x}=1} \mathbf{x}^T A^T A \mathbf{x}$ is a problem of finding the relative extremum. To solve our problem we form an auxiliary function

$$\Phi(\xi_1; \dots, \xi_n; \rho) = \mathbf{x}^T A^T A \mathbf{x} + \mu (1 - \mathbf{x}^T \mathbf{x}).$$

To find the stationary points of Φ , we form the system of equations:

$$\frac{\partial \Phi}{\partial \xi_i} = 0 \quad (i = 1 : n) \quad \wedge \quad \frac{\partial \Phi}{\partial \rho} = 0,$$

i.e.,

$$\begin{cases} 2 [A^T A \mathbf{x}]_i - 2\mu [\mathbf{x}]_i = 0 & (i = 1 : n) \\ 1 - \mathbf{x}^T \mathbf{x} = 0 \end{cases}$$

or

$$\begin{cases} A^T A \mathbf{x} = \mu \mathbf{x}, \\ \|\mathbf{x}\|_2 = 1. \end{cases}$$

Thus, any stationary point for relative extremum is the normed vector $\mathbf{x}$ corresponding to an eigenvalue of $A^T A$. Let us express from the relation $A^T A \mathbf{x} = \mu \mathbf{x}$ the eigenvalue μ . We obtain that $\mu = \mathbf{x}^T A^T A \mathbf{x}$, where $\|\mathbf{x}\|_2 = 1$. Comparing this result with the original formula for finding $\|A\|_2^2$, we notice that $\|A\|_2^2 = \max_{\mu \in \lambda(A^T A)} \mu$. Thus,

$$\|A\|_2 = \sqrt{\max_{\mu \in \lambda(A^T A)} \mu},$$

i.e., $\|A\|_2$ is the square root of the largest eigenvalue of $A^T A$.

Corollary 2.7.1. If matrix $A \in \mathbf{R}^{m \times n}$ is symmetric, then

$$\|A\|_2 = \max_{\lambda \in \lambda(A)} \lambda.$$

Example 2.7.3. Let us calculate the inverse matrix A^{-1} of the matrix A , the norms $\|A\|_1$, $\|A^{-1}\|_1$, $\|A\|_2$, $\|A^{-1}\|_2$, $\|A\|_\infty$, $\|A^{-1}\|_\infty$ and the condition numbers of the matrix A $k_1(A)$, $k_2(A)$, $k_\infty(A)$ if

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1.00000001 \end{bmatrix}.$$

We obtain that

$$A^{-1} = \begin{bmatrix} 1.0 \times 10^8 & -1.0 \times 10^8 \\ -1.0 \times 10^8 & 1.0 \times 10^8 \end{bmatrix}, \quad \|A\|_1 = \|A\|_2 \approx \|A\|_\infty \approx 2,$$

$$\|A^{-1}\|_1 = \|A^{-1}\|_\infty \approx \|A^{-1}\|_2 \approx 2 \times 10^8, \quad k_1(A) = k_\infty(A) \approx 4 \times 10^8.$$

Example 2.7.4. Let us see how the almost singularity (the value of the determinant is close to zero) and ill condition of the matrix are related. For the matrix

$$A_n = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ 0 & 0 & 1 & \cdots & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbf{R}^{n \times n}$$

$\det(A_n) = 1$ but $k_\infty(A_n) = n2^{n-1}$. In contrast, for the diagonal matrix

$$D_n = \text{diag}(\varepsilon, \dots, \varepsilon) \in \mathbf{R}^{n \times n}$$

$k_p(D_n) = 1$ but $\det(D_n) = \varepsilon^n$ for an arbitrarily small ε .

Exercise 2.7.4.* Find the inverse A^{-1} of the matrix A , the norms $\|A\|_1, \|A^{-1}\|_1, \|A\|_2, \|A^{-1}\|_2, \|A\|_\infty, \|A^{-1}\|_\infty$ and the condition numbers $k_1(A), k_2(A), k_\infty(A)$ if

$$a) A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad b) A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad c) A = \begin{bmatrix} -2 & -1 & 2 & -1 \\ 1 & 2 & 1 & -2 \\ 2 & -1 & 2 & 1 \\ 0 & 2 & 0 & 1 \end{bmatrix}.$$

1.2.8 Cayley-Hamilton Theorem

Proposition 2.8.1 (Cayley-Hamilton theorem). If $A \in C^{n \times n}$ and

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

then $p(A) = 0$, i.e., the matrix A satisfies its characteristic equation.

Proof. According to proposition 2.6.7, there exists a regular matrix $X \in C^{n \times n}$ such that

$$X^{-1}AX = J \equiv \text{diag}(J_1, \dots, J_t),$$

where

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_i \end{bmatrix}$$

is an upper bidiagonal $m_i \times m_i$ -matrix (Jordan block) that has on its main diagonal the eigenvalue λ_i of the matrix A (at least m_i -multiple eigenvalue of the matrix A since to this eigenvalue may correspond some more Jordan blocks) and $m_1 + \dots + m_t = n$. Since $J_i - \lambda_i I = (\delta_{k; j-1})$, then $(\delta_{k; j-1})^{m_i} = 0$ and $(J_i - \lambda_i I)^{m_i} = 0$. If $p(\lambda)$ is the characteristic polynomial of the matrix A and the zeros of this polynomial are $\lambda_1, \dots, \lambda_t$, then

$$p(\lambda) = (-1)^n (\lambda - \lambda_1)^{m_1} \dots (\lambda - \lambda_t)^{m_t}$$

and

$$p(J) = (-1)^n (J - \lambda_1 I)^{m_1} \dots (J - \lambda_t I)^{m_t}.$$

We show that $p(J) = 0$. Let the matrix J have the block form:

$$J = \begin{bmatrix} J_1 & 0 & 0 & \cdots & 0 \\ 0 & J_2 & 0 & \cdots & 0 \\ 0 & 0 & J_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & J_t \end{bmatrix} \begin{matrix} m_1 \\ m_2 \\ m_3 \\ \vdots \\ m_t \end{matrix}.$$

We obtain that

$$\begin{aligned} p(J) &= (-1)^n (J - \lambda_1 I)^{m_1} \dots (J - \lambda_t I)^{m_t} = \\ &= (-1)^n \begin{bmatrix} J_1 - \lambda_1 I & 0 & 0 & \cdots & 0 \\ 0 & J_2 - \lambda_1 I & 0 & \cdots & 0 \\ 0 & 0 & J_3 - \lambda_1 I & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & J_t - \lambda_1 I \end{bmatrix}^{m_1} \dots \\ &\quad \dots \begin{bmatrix} J_1 - \lambda_t I & 0 & 0 & \cdots & 0 \\ 0 & J_2 - \lambda_t I & 0 & \cdots & 0 \\ 0 & 0 & J_3 - \lambda_t I & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & J_t - \lambda_t I \end{bmatrix}^{m_t} = \\ &= (-1)^n \begin{bmatrix} (J_1 - \lambda_1 I)^{m_1} & 0 & 0 & \cdots & 0 \\ 0 & (J_2 - \lambda_1 I)^{m_1} & 0 & \cdots & 0 \\ 0 & 0 & (J_3 - \lambda_1 I)^{m_1} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (J_t - \lambda_1 I)^{m_1} \end{bmatrix} \dots \end{aligned}$$

$$\begin{aligned}
& \cdots \begin{bmatrix} (J_1 - \lambda_t I)^{m_t} & 0 & 0 & \cdots & 0 \\ 0 & (J_2 - \lambda_t I)^{m_t} & 0 & \cdots & 0 \\ 0 & 0 & (J_3 - \lambda_t I)^{m_t} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (J_t - \lambda_t I)^{m_t} \end{bmatrix} = \\
& = (-1)^n \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & (J_2 - \lambda_1 I)^{m_1} & 0 & \cdots & 0 \\ 0 & 0 & (J_3 - \lambda_1 I)^{m_1} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (J_t - \lambda_1 I)^{m_1} \end{bmatrix} \cdots \\
& \cdots \begin{bmatrix} (J_1 - \lambda_t I)^{m_t} & 0 & 0 & \cdots & 0 \\ 0 & (J_2 - \lambda_t I)^{m_t} & 0 & \cdots & 0 \\ 0 & 0 & (J_3 - \lambda_t I)^{m_t} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = \\
& = (-1)^n \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = 0.
\end{aligned}$$

From the relation $X^{-1}AX = J$ it follows that $A = XJX^{-1}$. We complete the proof with

$$\begin{aligned}
& p(A) = p(XJX^{-1}) = \\
& = (-1)^n (XJX^{-1} - X\lambda_1 I X^{-1})(XJX^{-1} - X\lambda_2 I X^{-1}) \cdots (XJX^{-1} - X\lambda_n I X^{-1}) = \\
& = (-1)^n X(J - \lambda_1 I)X^{-1}X(J - \lambda_2 I)X^{-1} \cdots X(J - \lambda_n I)X^{-1} = Xp(J)X^{-1} = 0. \quad \square
\end{aligned}$$

Example 2.8.1. Verify the assertion of the Cayley-Hamilton theorem for the matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

We construct the characteristic polynomial

$$p(\lambda) = \det(A - \lambda I) = \begin{vmatrix} a - \lambda & b \\ c & d - \lambda \end{vmatrix} = \lambda^2 - (a + d)\lambda + ad - cb$$

and find

$$\begin{aligned} p(A) &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}^2 - (a + d) \begin{bmatrix} a & b \\ c & d \end{bmatrix} + (ad - bc) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \\ &= \begin{bmatrix} a^2 + bc - a^2 - ad - bc & ab + bd - ab - bd \\ ac + cd - ac - cd & bc + d^2 - ad - d^2 + ad - bc \end{bmatrix} = 0. \end{aligned}$$

Exercise 2.8.1.* Let

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

Compute A^2 , and using the Cayley-Hamilton theorem, find the matrix

$$A^7 - 3A^6 + A^4 + 3A^3 - 2A^2 + 3I.$$

Definition 2.8.1. A polynomial $q(\lambda)$ is called a *nullifying polynomial* of the matrix $A \in C^{n \times n}$ if $q(A) = 0$.

The characteristic polynomial of the matrix $A \in C^{n \times n}$ is a nullifying polynomial of this matrix (by the Cayley-Hamilton theorem).

Definition 2.8.2. The nullifying polynomial of the matrix $A \in C^{n \times n}$ of the lowest degree is called the *minimal polynomial* of the matrix A .

Exercise 2.8.1. Verify that the characteristic polynomial of matrix $A \in C^{n \times n}$ is divisible by the minimal polynomial of the matrix A without remainder.

Proposition 2.8.2. Let $p(\lambda)$ and $\psi(\lambda)$ be the characteristic polynomial and the minimal polynomial of the matrix A , respectively. Let the greatest common divisor of the matrix $(I\lambda - A)^\vee$, that is, the matrix of algebraic complements of the elements of the matrix $(I\lambda - A)$, be $d(\lambda)$. Then,

$$p(\lambda) = d(\lambda)\psi(\lambda).$$

Proof. See Lancaster (1982, p. 123-124).

Example 2.8.2. Find the characteristic polynomial and the minimal polynomial of the matrix $D = \text{diag}(a, a, b, b)$. First we find that

$$\begin{aligned} (I\lambda - D)^\vee &= \text{diag}(\lambda - a, \lambda - a, \lambda - b, \lambda - b)^\vee = \\ &= \begin{bmatrix} \lambda - a & 0 & 0 & 0 \\ 0 & \lambda - a & 0 & 0 \\ 0 & 0 & \lambda - b & 0 \\ 0 & 0 & 0 & \lambda - b \end{bmatrix}^\vee = \\ &= \begin{bmatrix} (\lambda - a)(\lambda - b)^2 & 0 & 0 & 0 \\ 0 & (\lambda - a)(\lambda - b)^2 & 0 & 0 \\ 0 & 0 & (\lambda - a)^2(\lambda - b) & 0 \\ 0 & 0 & 0 & (\lambda - a)^2(\lambda - b) \end{bmatrix} \end{aligned}$$

and the greatest common divisor of the elements of the matrix $(I\lambda - D)^\vee$ is $d(\lambda) = (\lambda - a)(\lambda - b)$. By the assertion of proposition 2.8.2 $\psi(\lambda) = (\lambda - a)(\lambda - b)$. Let us check

$$\begin{aligned} \psi(D) &= (D - aI)(D - bI) = \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & b - a & 0 \\ 0 & 0 & 0 & b - a \end{bmatrix} \begin{bmatrix} a - b & 0 & 0 & 0 \\ 0 & a - b & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = 0. \end{aligned}$$

Indeed, $\psi(\lambda)$ is the nullifying polynomial. It is easy to verify that no polynomial of first degree can nullify the matrix A . Thus, $\psi(\lambda)$ is the minimal polynomial of the matrix A .

Example 2.8.3. Find the characteristic and the minimal polynomials of the matrices

$$A = \begin{bmatrix} 6 & 2 & -2 \\ -2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 6 & 2 & 2 \\ -2 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

First we find the characteristic polynomials:

$$p_A(\lambda) = \begin{vmatrix} 6 - \lambda & 2 & -2 \\ -2 & 2 - \lambda & 2 \\ 2 & 2 & 2 - \lambda \end{vmatrix} = \lambda^3 - 10\lambda^2 + 32\lambda - 32$$

and

$$p_B(\lambda) = \begin{vmatrix} 6 - \lambda & 2 & 2 \\ -2 & 2 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{vmatrix} = \lambda^3 - 10\lambda^2 + 32\lambda - 32.$$

Next we find the minimal polynomials:

$$\begin{aligned} (I\lambda - A)^\vee &= \\ &= \begin{bmatrix} \lambda - 6 & -2 & 2 \\ 2 & \lambda - 2 & -2 \\ -2 & -2 & \lambda - 2 \end{bmatrix}^\vee = \begin{bmatrix} \lambda^2 - 4\lambda & -2\lambda + 8 & 2\lambda - 8 \\ 2\lambda - 8 & \lambda^2 - 8\lambda + 16 & 2\lambda - 8 \\ -2\lambda + 8 & 2\lambda - 8 & \lambda^2 - 8\lambda + 16 \end{bmatrix} = \\ &= \begin{bmatrix} \lambda(\lambda - 4) & -2(\lambda - 4) & 2(\lambda - 4) \\ 2(\lambda - 4) & (\lambda - 4)^2 & 2(\lambda - 4) \\ -2(\lambda - 4) & 2(\lambda - 4) & (\lambda - 4)^2 \end{bmatrix} \Rightarrow d_A(\lambda) = \lambda - 4 \Rightarrow \\ &\Rightarrow \psi_A(\lambda) = \frac{\lambda^3 - 10\lambda^2 + 32\lambda - 32}{\lambda - 4} = \lambda^2 - 6\lambda + 8 \end{aligned}$$

and

$$\begin{aligned} (I\lambda - B)^\vee &= \\ &= \begin{bmatrix} \lambda - 6 & -2 & -2 \\ 2 & \lambda - 2 & 0 \\ 0 & 0 & \lambda - 2 \end{bmatrix}^\vee = \begin{bmatrix} (\lambda - 2)^2 & -2(\lambda - 2) & 0 \\ 2(\lambda - 2) & (\lambda - 2)(\lambda - 6) & 0 \\ 2(\lambda - 2) & -4 & (\lambda - 4)^2 \end{bmatrix} \Rightarrow \\ &\Rightarrow d_B(\lambda) = 1 \Rightarrow \psi_B(\lambda) = \frac{\lambda^3 - 10\lambda^2 + 32\lambda - 32}{1} = \lambda^3 - 10\lambda^2 + 32\lambda - 32. \end{aligned}$$

1.2.9 Functions of Matrices

Let us consider a matrix $A \in \mathbf{C}^{n \times n}$ and a function of a complex variable

$$f(z), \quad f : \mathbf{C} \rightarrow \mathbf{C}.$$

There are a lot of possibilities to define a function of matrix $f(A)$ starting from the function of a complex variable $f(z)$. The simplest of these possibilities seems to be the substituting the variable "z" by the variable "A". For example,

$$f(z) = z^2 + 3z - 7 \rightarrow f(A) = A^2 + 3A - 7I$$

and

$$f(z) = \frac{4 + 5z}{3 - 8z} \rightarrow f(A) = (4I + 5A)(3I - 8A)^{-1},$$

as well

$$\begin{aligned} e^z &= \sum_{k=0}^{\infty} \frac{z^k}{k!} \rightarrow e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}, \\ \cos z &= \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k}}{(2k)!} \rightarrow \cos A = \sum_{k=0}^{\infty} (-1)^k \frac{A^{2k}}{(2k)!}, \\ \sin z &= \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k+1}}{(2k+1)!} \rightarrow \sin A = \sum_{k=0}^{\infty} (-1)^k \frac{A^{2k+1}}{(2k+1)!}, \\ \ln(I + z) &= \sum_{k=1}^{\infty} (-1)^{k+1} \frac{z^k}{k} \rightarrow \ln(I + A) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{A^k}{k}. \end{aligned}$$

It turns out that this approach is not very practical for solving problems.

Definition 2.9.1. If $A \in \mathbf{C}^{n \times n}$, $f(z)$ is analytic in the open domain $\mathfrak{D}$, Γ is a closed simple line (does not cut itself) in $\mathfrak{D}$ and the spectrum $\lambda(A)$ of the matrix A is included in domain $\mathfrak{D}_\Gamma$ enfolded by Γ , then

$$f(A) \stackrel{def}{=} \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - A)^{-1} dz, \quad (28)$$

where the integral is applied to the matrix element by element.

Remark 2.9.1. Formula (28) is an analogue to the Cauchy integral formula proved for functions of a complex variable.

Example 2.9.1. Let $f(z) = z$ and $A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$. Check how to calculate by rule (28). Since $\lambda_1 = a$ and $\lambda_2 = c$ are the eigenvalues of A , then let us choose the line $\Gamma : |z| = r$, where $r > \max(|a|, |c|)$. The function $f(z) = z$ is analytic in domain $\mathfrak{D}_\Gamma$. First we find

$$\begin{aligned} (zI - A)^{-1} &= \begin{bmatrix} z - a & -b \\ 0 & z - c \end{bmatrix}^{-1} = \\ &= \begin{bmatrix} 1/(z - a) & (b/(a - c))(1/(z - a) - 1/(z - c)) \\ 0 & 1/(z - c) \end{bmatrix}, \end{aligned}$$

and then

$$\begin{aligned}
f\left(\begin{bmatrix} a & b \\ 0 & c \end{bmatrix}\right) &= \frac{1}{2\pi i} \oint_{|z|=r} \begin{bmatrix} z/(z-a) & (b/(a-c))(z/(z-a) - z/(z-c)) \\ 0 & 1/(z-c) \end{bmatrix} dz = \\
&= \begin{bmatrix} \frac{1}{2\pi i} \oint_{|z|=r} z/(z-a) dz & \frac{1}{2\pi i} \oint_{|z|=r} (b/(a-c))(z/(z-a) - z/(z-c)) dz \\ 0 & \frac{1}{2\pi i} \oint_{|z|=r} 1/(z-c) dz \end{bmatrix} = \\
&= \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}.
\end{aligned}$$

Proposition 2.9.1. If $f(A) = (f_{k;j})$, then

$$f_{k;j} = \frac{1}{2\pi i} \oint_{\Gamma} f(z) \mathbf{e}_k^T (zI - A)^{-1} \mathbf{e}_j dz,$$

where $\mathbf{e}_k$ is a vector in space $\mathbf{C}^n$ whose k -th component is one and the remaining ones are zeros.

Proof. Let $B = (b_{k;j}) = (zI - A)^{-1}$. Then,

$$\mathbf{e}_k^T (zI - A)^{-1} \mathbf{e}_j = \mathbf{e}_k^T B \mathbf{e}_j = \begin{bmatrix} b_{k;1} & \cdots & b_{k;n} \end{bmatrix} \mathbf{e}_j = b_{k;j}.$$

Since the matrix is integrated element by element, we obtain that

$$\frac{1}{2\pi i} \oint_{\Gamma} f(z) \mathbf{e}_k^T (zI - A)^{-1} \mathbf{e}_j dz = \frac{1}{2\pi i} \oint_{\Gamma} f(z) b_{k;j} dz = f_{k;j}. \quad \square$$

Proposition 2.9.2. If $A \in \mathbf{C}^{n \times n}$, there $\exists f(A)$, i.e., the conditions of definition 2.9.1 are satisfied, and

$$A = X B X^{-1} = X \operatorname{diag}(B_1, \dots, B_p) X^{-1}, \quad B_k \in \mathbf{C}^{n_k \times n_k}, \quad (29)$$

then

$$f(A) = X f(B) X^{-1} = X \operatorname{diag}(f(B_1), \dots, f(B_p)) X^{-1}. \quad (30)$$

Proof. From (28), (29) and $XX^{-1} = I$ we find that

$$\begin{aligned}
f(A) &= \frac{1}{2\pi i} \oint_{\Gamma} f(z) (zI - A)^{-1} dz = \frac{1}{2\pi i} \oint_{\Gamma} f(z) (X z I X^{-1} - X B X^{-1})^{-1} dz = \\
&= \frac{1}{2\pi i} \oint_{\Gamma} X f(z) (zI - B)^{-1} X^{-1} dz = X \frac{1}{2\pi i} \oint_{\Gamma} f(z) (zI - B)^{-1} dz X^{-1} = X f(B) X^{-1}.
\end{aligned}$$

Since

$$(Iz - B)^{-1} = \text{diag}((Iz - B_1)^{-1}, \dots, (Iz - B_p)^{-1})$$

and

$$\begin{aligned} f(B) &= \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - B)^{-1} dz = \\ &= \text{diag}\left(\frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - B_1)^{-1} dz; \dots; \text{diag}\left(\frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - B_p)^{-1} dz\right) = \right. \\ &= \text{diag}(f(B_1); \dots; f(B_p)), \end{aligned}$$

then

$$f(A) = X f(B) X^{-1} = X \text{diag}(f(B_1), \dots, f(B_p)) X^{-1},$$

which was to be proved. $\square$

Proposition 2.9.3. If $A \in \mathbf{C}^{n \times n}$ and $X^{-1}AX = \text{diag}(J_1, \dots, J_p)$ is the Jordan normal form of the matrix A , where

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_i \end{bmatrix}$$

is an $m_i \times m_i$ Jordan block, $m_1 + \dots + m_p = n$, and $f(z)$ is analytic on an open set that includes the spectrum $\lambda(A)$ of the matrix A , then

$$f(A) = X \text{diag}(f(J_1), \dots, f(J_p)) X^{-1}, \quad (31)$$

where

$$f(J_i) = \begin{bmatrix} f(\lambda_i) & f'(\lambda_i) & \cdots & \cdots & f^{(m_i-1)}(\lambda_i)/(m_i-1)! \\ 0 & f(\lambda_i) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & f'(\lambda_i) \\ 0 & \cdots & \cdots & \cdots & f(\lambda_i) \end{bmatrix}. \quad (32)$$

Proof. Using Proposition 2.9.2 it is sufficient to consider only the value $F(G)$, where $G = \lambda I + E$ is a $q \times q$ Jordan block and $E = (\delta_{i,j-1})$. Let the matrix $zI - G$ be regular. Since

$$E^k = (\delta_{i,j-k}) \Rightarrow (k \geq q \Rightarrow E^k = 0),$$

then

$$(zI - G)^{-1} = \sum_{k=0}^{q-1} \frac{E^k}{(z - \lambda)^{k+1}}$$

and

$$\begin{aligned} f(G) &= \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - G)^{-1} dz = \frac{1}{2\pi i} \oint_{\Gamma} f(z) \sum_{k=0}^{q-1} \frac{E^k}{(z - \lambda)^{k+1}} dz = \\ &= \sum_{k=0}^{q-1} E^k \frac{1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{(z - \lambda)^{k+1}} dz = \sum_{k=0}^{q-1} \frac{f^{(k)}(\lambda)}{k!} E^k. \end{aligned}$$

Now, taking into account the condition $E^k = (\delta_{i,j-k})$, the assertion holds. $\square$

Example 2.9.2. Find $\cos A$ if

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

Since $\lambda_1 = \dots = \lambda_5 = 0$ and the function $\cos z$ is analytic in the neighbourhood of 0, then for the calculation of $\cos A$ we can apply the algorithm given in Proposition 2.9.3. We use for the calculation of Jordan decomposition of the matrix A "Maple":

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

Therefore, $J = \text{diag}(J_1, J_2)$, where

$$J_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad \wedge \quad J_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Using (3), we find the matrices $\cos J_1$ and $\cos J_2$:

$$\cos J_1 = \begin{bmatrix} \cos 0 & (-\sin 0)/1! & (-\cos 0)/2! \\ 0 & \cos 0 & (-\sin 0)/1! \\ 0 & 0 & \cos 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$\cos J_2 = \begin{bmatrix} \cos 0 & (-\sin 0)/1! \\ 0 & \cos 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

After that, using (2), we find the matrix wanted:

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

Corollary 2.9.1. If $A \in \mathbf{C}^{n \times n}$, $A = X \operatorname{diag}(\lambda_1, \dots, \lambda_n) X^{-1}$ and there $\exists f(A)$, then

$$f(A) = X \operatorname{diag}(f(\lambda_1), \dots, f(\lambda_n)) X^{-1}.$$

Proof. This is a special case of Proposition 2.9.3. All the Jordan blocks are 1×1 .

Example 2.9.3. If $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the matrix $A \in \mathbf{C}^{n \times n}$ and $\mathbf{x}_1, \dots, \mathbf{x}_n$ are the corresponding linearly independent eigenvectors, i.e., $\mathbf{x}_1, \dots, \mathbf{x}_n$ generate a basis in $\mathbf{C}^n$, then $X = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}$, and from the analyticity of $\exp z$, $\cos z$, $\sin z$ in the whole finite part of complex plane it follows that

$$\begin{aligned} \exp A &= X \operatorname{diag}(\exp \lambda_1, \dots, \exp \lambda_n) X^{-1} = X (\exp \Lambda) X^{-1}, \\ \cos A &= X \operatorname{diag}(\cos \lambda_1, \dots, \cos \lambda_n) X^{-1} = X (\cos \Lambda) X^{-1}, \\ \sin A &= X \operatorname{diag}(\sin \lambda_1, \dots, \sin \lambda_n) X^{-1} = X (\sin \Lambda) X^{-1}, \end{aligned}$$

where $\Lambda, \exp \Lambda, \cos \Lambda, \sin \Lambda, \exp A, \cos A, \sin A \in \mathbf{C}^{n \times n}$ and

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}, \quad \exp \Lambda = \begin{bmatrix} \exp \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \exp \lambda_n \end{bmatrix},$$

$$\cos \Lambda = \begin{bmatrix} \cos \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \cos \lambda_n \end{bmatrix}, \quad \sin \Lambda = \begin{bmatrix} \sin \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sin \lambda_n \end{bmatrix}.$$

Next we consider the problem arising in the approximation of the function $f(A)$ by the function $g(A)$. This kind of problem arises, for example, if we replace $f(A)$ with its Taylor polynomial of degree q .

Proposition 2.9.4. Let $A \in \mathbf{C}^{n \times n}$, $X^{-1}AX = \text{diag}(J_1, \dots, J_p)$, where

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_i \end{bmatrix}$$

is an $m_i \times m_i$ Jordan block and $m_1 + \dots + m_p = n$. If the functions $f(z)$ and $g(z)$ are analytic on an open set containing the spectrum $\lambda(A)$ of the matrix A , then

$$\|f(A) - g(A)\|_2 \leq k_2(X) \max_{1 \leq i \leq p \wedge 0 \leq r \leq m_i - 1} m_i \frac{|f^{(r)}(\lambda_i) - g^{(r)}(\lambda_i)|}{r!}. \quad (33)$$

Proof. Choosing $h(z) = f(z) - g(z)$ we have

$$\begin{aligned} \|f(A) - g(A)\|_2 &= \|X \text{diag}(h(J_1), \dots, h(J_p)) X^{-1}\|_2 \leq \\ &\leq \|X\|_2 \|\text{diag}(h(J_1), \dots, h(J_p))\|_2 \|X^{-1}\|_2 \leq k_2(X) \max_{1 \leq i \leq p} \|h(J_i)\|_2. \end{aligned}$$

Using Proposition 2.9.3 and inequality $\|B\|_2 \leq n \cdot \max_{i,j} |b_{i,j}|$ we find that

$$\|h(J_i)\|_2 \leq m_i \max_{0 \leq r \leq m_i - 1} \frac{|h^{(r)}(\lambda_i)|}{r!}$$

and thus, the assertion holds. $\square$

Example 2.9.4. Let

$$A = \begin{bmatrix} 1/10 & 1/10 & 0 \\ 0 & 1/10 & 1/10 \\ 0 & 0 & 1/10 \end{bmatrix}.$$

We estimate the difference $\sin A - A$.

Since $\lambda_1 = \lambda_2 = \lambda_3 = .1$ and the functions $f(z) = \sin z$ and $g(z) = z$ are analytic in the neighbourhood of .1, then we can apply the estimation (33) obtained in Proposition 2.9.4. First, we use "Maple" for finding the Jordan decomposition of the matrix A :

$$A = XJX^{-1} = \begin{bmatrix} \frac{1}{100} & 0 & 1 \\ 0 & \frac{1}{10} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{10} & 1 & 0 \\ 0 & \frac{1}{10} & 1 \\ 0 & 0 & \frac{1}{10} \end{bmatrix} \begin{bmatrix} 100 & 0 & -100 \\ 0 & 10 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence, there is only one Jordan block in the Jordan decomposition of the matrix A , i.e.

$$J = J_1 = \text{diag}(J_1).$$

Second, we use "Maple" for finding the condition number of X :

$$k_2(X) \approx 200.01.$$

Since

$$f(z) - g(z) = \sin z - z \Rightarrow |f(.1) - g(.1)|/0! = |\sin .1 - .1| \approx 1.6658 \times 10^{-4},$$

$$f'(z) - g'(z) = \cos z - 1 \Rightarrow |f'(1) - g'(1)|/1! = |\cos .1 - 1| \approx 4.9958 \times 10^{-3}$$

and

$$f''(z) - g''(z) = -\sin z \Rightarrow |f''(1) - g''(1)|/2! = |-\sin .1|/2 \approx 4.9917 \times 10^{-2},$$

then, by estimation (30), we have that

$$\|\sin A - A\|_2 \leq 200.01 \cdot 3 \cdot 4.9917 \times 10^{-2} \approx 29.952.$$

It is known that matrix X in the Jordan decomposition of A is not uniquely determined. We try to choose the matrix X so that the condition number $k_2(X)$ should be minimal. Applying the Filipov algorithm to find the Jordan decomposition of A (see Proposition 2.5.2.1), we obtain that

$$A = XJX^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 100 \end{bmatrix} \begin{bmatrix} \frac{1}{10} & 1 & 0 \\ 0 & \frac{1}{10} & 1 \\ 0 & 0 & \frac{1}{10} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{10} & 0 \\ 0 & 0 & \frac{1}{100} \end{bmatrix},$$

where

$$k_2(X) = 100.$$

It turns out that

$$A = XJX^{-1} = \begin{bmatrix} \frac{1}{10} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10 \end{bmatrix} \begin{bmatrix} \frac{1}{10} & 1 & 0 \\ 0 & \frac{1}{10} & 1 \\ 0 & 0 & \frac{1}{10} \end{bmatrix} \begin{bmatrix} 10 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{10} \end{bmatrix}$$

is also a Jordan decomposition of A , where

$$k_2(X) = 10.$$

Therefore, the best estimation we can have by Proposition 2.9.4 is

$$\|\sin A - A\|_2 \leq 10 \cdot 3 \cdot 4.9917 \times 10^{-2} \approx 1.4975.$$

Otherwise, in this example for calculating $\sin A$ we can apply the algorithm given in Proposition 2.9.3. Using the formula (32), we have that

$$\begin{aligned} \sin J &= \begin{bmatrix} \sin .1 & (\cos .1)/1! & (-\sin .1)/2! \\ 0 & \sin .1 & (\cos .1)/1! \\ 0 & 0 & \sin .1 \end{bmatrix} = \\ &= \begin{bmatrix} .099833 & .995 & -.049917 \\ 0 & .099833 & .995 \\ 0 & 0 & .099833 \end{bmatrix}. \end{aligned}$$

By formula (31) we calculate the value of the function in question:

$$\begin{aligned} \sin A &= \begin{bmatrix} \frac{1}{100} & 0 & 1 \\ 0 & \frac{1}{10} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} .099833 & .995 & -.049917 \\ 0 & .099833 & .995 \\ 0 & 0 & .099833 \end{bmatrix} \begin{bmatrix} 100 & 0 & -100 \\ 0 & 10 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \\ &= \begin{bmatrix} .099833 & .0995 & -.00049917 \\ 0 & .099833 & .0995 \\ 0 & 0 & .099833 \end{bmatrix}. \end{aligned}$$

Hence,

$$\sin A - A = \begin{bmatrix} .099833 - .1 & .0995 - .1 & -.00049917 \\ 0 & .099833 - .1 & .0995 - .1 \\ 0 & 0 & .099833 - .1 \end{bmatrix} =$$

$$= \begin{bmatrix} -1.67 \times 10^{-4} & -.0005 & -4.9917 \times 10^{-4} \\ 0 & -1.67 \times 10^{-4} & -.0005 \\ 0 & 0 & -1.67 \times 10^{-4} \end{bmatrix}$$

and

$$\|\sin A - A\|_2 \approx 8.8098 \times 10^{-4}.$$

As a result of this example, we can assert that estimation (33) proved in Proposition 2.9.5 is quite rough.

Proposition 2.9.5. If the Maclaurin expansion of the function $f(z)$

$$f(z) = \sum_{k=0}^{\infty} c_k z^k$$

is convergent in the circle containing the spectrum $\lambda(A)$ of the matrix $A \in \mathbf{C}^{n \times n}$, then

$$f(A) = \sum_{k=0}^{\infty} c_k A^k.$$

Prove this assertion with an additional assumption that the matrix A has a basis consisting of its eigenvectors. In this case, by Corollary 2.9.1,

$$\begin{aligned} f(A) &= X \operatorname{diag}(f(\lambda_1), \dots, f(\lambda_n)) X^{-1} = \\ &= X \operatorname{diag}\left(\sum_{k=0}^{\infty} c_k \lambda_1^k, \dots, \sum_{k=0}^{\infty} c_k \lambda_n^k\right) X^{-1} = \\ &= X \left(\sum_{k=0}^{\infty} c_k D^k\right) X^{-1} = \sum_{k=0}^{\infty} c_k (X D X^{-1})^k = \sum_{k=0}^{\infty} c_k A^k. \quad \square \end{aligned}$$

Proposition 2.9.6. If the Maclaurin series of the function $f(z)$

$$f(z) = \sum_{k=0}^{\infty} c_k z^k$$

is convergent in the circle containing the spectrum $\lambda(A)$ of the matrix $A \in \mathbf{C}^{n \times n}$, then

$$\|f(A) - \sum_{k=0}^q c_k A^k\|_2 \leq \frac{n}{(q+1)!} \max_{0 \leq s \leq 1} \|A^{q+1} f^{(q+1)}(As)\|_2.$$

Proof. Let us define the matrix $E(s)$ by

$$f(As) = \sum_{k=0}^q c_k(As)^k + E(s) \quad (0 \leq s \leq 1). \quad (34)$$

If $f_{i,j}(s) = [f(As)]_{i,j}$, then $f_{i,j}(s)$ is analytic, and therefore,

$$f_{i,j}(s) = \sum_{k=0}^q \frac{f_{i,j}(0)}{k!} s^k + \frac{f_{i,j}^{(q+1)}(\varepsilon_{i,j})}{(q+1)!} s^{q+1}, \quad (35)$$

where $0 \leq \varepsilon_{i,j} \leq s \leq 1$. By comparing the powers of the variable s in (34) and (35), we conclude that $[E(s)]_{i,j}$ has the form

$$\varepsilon_{i,j}(s) = \frac{f_{i,j}^{(q+1)}(\varepsilon_{i,j})}{(q+1)!} s^{q+1}.$$

If $f_{i,j}^{(q+1)} = [A^{q+1} f^{(q+1)}(As)]_{i,j}$, then

$$|\varepsilon_{i,j}(s)| \leq \max_{0 \leq s \leq 1} \frac{|f_{i,j}^{(q+1)}(\varepsilon_{i,j})|}{(q+1)!} \leq n \max_{0 \leq s \leq 1} \|A^{q+1} f^{(q+1)}(As)\|_2. \quad \square$$

Exercise 2.9.1. Prove that for an arbitrary matrix $A \in \mathbf{C}^{n \times n}$

$$I + \cos(2A) = 2 \cos^2 A$$

and

$$\sin(2A) = 2 \sin A \cdot \cos A.$$

Exercise 2.9.2. Apply Proposition 2.9.6 to the estimation of the errors in the approximate equalities

$$\sin A \approx \sum_{k=0}^q (-1)^k \frac{A^{2k+1}}{(2k+1)!}$$

and

$$\cos A \approx \sum_{k=0}^q (-1)^k \frac{A^{2k}}{(2k)!}.$$

Proposition 2.9.7 (Sylvester theorem). If all eigenvalues λ_k of the matrix $A \in \mathbf{C}^{n \times n}$ are different, then

$$f(A) = \sum_{k=1}^n f(\lambda_k) \frac{\prod_{i \neq k} (A - \lambda_i I)}{\prod_{i \neq k} (\lambda_k - \lambda_i)} \quad (36)$$

or

$$f(A) = \frac{1}{\Delta} \sum_{k=1}^n \Delta_k A^{k-1}, \quad (37)$$

where Δ_k ($k = 1; 2; \dots; n$) is the determinant obtained from the Vandermonde determinant

$$\Delta = \begin{vmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \dots & \dots & \dots & \dots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \dots & \lambda_n^{n-1} \end{vmatrix}$$

by replacing the k -th row vector

$$(\lambda_1^{k-1} \lambda_2^{k-1} \dots \lambda_n^{k-1})$$

by the vector

$$(f(\lambda_1) \ f(\lambda_2) \ \dots \ f(\lambda_n)).$$

Example 2.9.4. Calculate $\exp A$ if $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$.

First we find the eigenvalues of A :

$$\begin{vmatrix} 1-\lambda & 1 \\ -1 & 1-\lambda \end{vmatrix} = 0 \Rightarrow (1-\lambda)^2 + 1 = 0 \Rightarrow \begin{cases} \lambda_1 = 1+i \\ \lambda_2 = 1-i \end{cases}$$

Then we use formula (36):

$$\begin{aligned} \exp \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} &= \frac{A - (1-i)I}{1+i-1+i} \exp(1+i) + \frac{A - (1+i)I}{1-i-1-i} \exp(1-i) = \\ &= \frac{\begin{bmatrix} i & 1 \\ -1 & i \end{bmatrix}}{2i} \exp(1+i) + \frac{\begin{bmatrix} -i & 1 \\ -1 & -i \end{bmatrix}}{-2i} = \\ &= e \cdot \begin{bmatrix} (\exp i + \exp(-i))/2 & (\exp i - \exp(-i))/2i \\ -(\exp i - \exp(-i))/2i & (\exp i + \exp(-i))/2 \end{bmatrix} = e \cdot \begin{bmatrix} \cos 1 & \sin 1 \\ -\sin 1 & \cos 1 \end{bmatrix} \end{aligned}$$

Now applying (37), we have

$$\begin{aligned} \exp \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} &= (1/\det \begin{bmatrix} 1 & 1 \\ 1+i & 1-i \end{bmatrix}) [I \det \begin{bmatrix} \exp(1+i) & \exp(1-i) \\ 1+i & 1-i \end{bmatrix} + \\ &+ \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \det \begin{bmatrix} 1 & 1 \\ \exp(1+i) & \exp(1-i) \end{bmatrix}] \end{aligned}$$

$$\begin{aligned}
&= \frac{e}{-2i} \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} [(1-i) \exp i - (1+i) \exp(-i)] + \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} [\exp(-i) - \exp i] \right\} = \\
&= \frac{e}{-2i} \begin{bmatrix} -i \exp i - i \exp(-i) & \exp(-i) - \exp i \\ -\exp(-i) + \exp i & -i \exp i - i \exp(-i) \end{bmatrix} = e \cdot \begin{bmatrix} \cos 1 & \sin 1 \\ -\sin 1 & \cos 1 \end{bmatrix}.
\end{aligned}$$

We solve this problem once more using the formula $\exp A = S \exp \Lambda S^{-1}$, where S is the matrix formed of the eigenvalues of A . Find the eigenvalues of A :

$$\lambda_1 = 1 + i \Rightarrow \begin{bmatrix} -i & 1 & \vdots & 0 \\ -1 & -i & \vdots & 0 \end{bmatrix} \Rightarrow x_1 = \begin{bmatrix} 1 \\ i \end{bmatrix}$$

and

$$\lambda_2 = 1 - i \Rightarrow \begin{bmatrix} i & 1 & \vdots & 0 \\ -1 & i & \vdots & 0 \end{bmatrix} \Rightarrow x_2 = \begin{bmatrix} 1 \\ -i \end{bmatrix},$$

and the matrix

$$S = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}.$$

Hence

$$\begin{aligned}
\exp A &= S \exp \Lambda S^{-1} = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} \begin{bmatrix} e^{1+i} & 0 \\ 0 & e^{1-i} \end{bmatrix} \begin{bmatrix} 1/2 & -i/2 \\ 1/2 & i/2 \end{bmatrix} = \\
&= \begin{bmatrix} e^{1+i} & e^{1-i} \\ ie^{1+i} & -ie^{1-i} \end{bmatrix} \begin{bmatrix} 1/2 & -i/2 \\ 1/2 & i/2 \end{bmatrix} = \\
&= e \begin{bmatrix} (\exp i + \exp(-i))/2 & (\exp i - \exp(-i))/2i \\ -(\exp i - \exp(-i))/2i & (\exp i + \exp(-i))/2 \end{bmatrix} = \\
&= e \begin{bmatrix} \cos i & \sin i \\ -\sin i & \cos i \end{bmatrix}.
\end{aligned}$$

2 Computation Methods in Linear Algebra

2.1 LU-Factorization

2.1.1 Solution of Triangular Systems

Let us consider the solution of a 2×2 –lower triangular system

$$\begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (l_{11}l_{22} \neq 0)$$

by forward substitution. From the first equation we obtain $\xi_1 = b_1/l_{11}$, and then from the second one $\xi_2 = (b_2 - l_{21}\xi_1)/l_{22}$.

Proposition 1.1.1 (*forward substitution*). If $L = (l_{ij}) \in \mathbf{R}^{n \times n}$ is a lower triangular matrix, $\prod_{i=1}^n l_{ii} \neq 0$ and $L\mathbf{x} = \mathbf{b}$, then the solution is

$$\xi_i = (b_i - \sum_{k=1}^{i-1} l_{ik}\xi_k)/l_{ii} \quad (i=1:n).$$

Solve the 2×2 –upper triangular system

$$\begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (u_{11}u_{22} \neq 0)$$

by back substitution. From the second equation we obtain $\xi_2 = b_2/u_{22}$, and then from the first one $\xi_1 = (b_1 - u_{12}\xi_2)/u_{11}$.

Proposition 1.1.2 (*back substitution*). If $U = (u_{ij}) \in \mathbf{R}^{n \times n}$ is an upper triangular matrix, $\prod_{i=1}^n u_{ii} \neq 0$ and $U\mathbf{x} = \mathbf{b}$, then the solution is

$$\xi_i = (b_i - \sum_{k=i+1}^n u_{ik}\xi_k)/u_{ii} \quad (i=1:n).$$

In the case of forward substitution as well as in the case of back substitution the solution of the system with a regular $n \times n$ –triangular matrix requires $1 + 3 + \dots + (2n - 1) = n^2$ operations.

Proposition 1.1.3 (*forward substitution: row version*). If $L \in \mathbf{R}^{n \times n}$ is lower triangular, $\prod_{i=1}^n l_{ii} \neq 0$, $L\mathbf{x} = \mathbf{b}$, and ξ_1 has been found, then after substitution of ξ_1 into the equations from the second to the n -th, we obtain a new $(n - 1) \times (n - 1)$ –lower triangular system

$$L(2:n; 2:n)\mathbf{x}(2:n) = \mathbf{b}(2:n) - \mathbf{x}(1)L(2:n; 1).$$

Proposition 1.1.4 (*back substitution:column version*). If $U \in \mathbf{R}^{n \times n}$ is upper triangular, $\prod_{i=1}^n u_{ii} \neq 0$, $U\mathbf{x} = \mathbf{b}$, and ξ_n has been found, then after

the substitution of ξ_n into the equations from the first to the $(n-1)$ -th, we obtain a new $(n-1) \times (n-1)$ -upper triangular system

$$U(1 : (n-1); 1 : (n-1)) \mathbf{x}(1 : (n-1)) = \mathbf{b}(1 : (n-1)) - \mathbf{x}(n)U(1 : (n-1); n).$$

Now we consider the simultaneous solution of several systems with a common system matrix. Let us consider the system $LX = B$, where $L \in \mathbf{R}^{n \times n}$ is a regular lower triangular matrix, $B \in \mathbf{R}^{n \times q}$ and the unknown is $X \in \mathbf{R}^{n \times q}$. We represent this system in block form

$$\begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \cdots & L_{NN} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{bmatrix}, \quad (1)$$

where the diagonal blocks are square. From the equation $L_{11}X_1 = B_1$ we can find X_1 . By using for system (1) the row version given in Proposition 1.1.3, we obtain

$$\begin{bmatrix} L_{22} & 0 & \cdots & 0 \\ L_{32} & L_{33} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{N2} & L_{N3} & \cdots & L_{NN} \end{bmatrix} \begin{bmatrix} X_2 \\ X_3 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} B_2 - L_{21}X_1 \\ B_3 - L_{31}X_1 \\ \vdots \\ B_N - L_{N1}X_1 \end{bmatrix}.$$

Continuing in this way we obtain the solution of system (1).

Proposition 1.1.5. Triangular matrices have the following properties:

- the inverse of an upper (lower) triangular matrix is upper (lower) triangular;
- the product of two upper (lower) triangular matrices is upper (lower) triangular;
- the inverse of an unit upper (lower) triangular matrix is unit upper (lower) triangular;
- the product of two unit upper (lower) triangular matrices is upper (lower) triangular.
- **Exercise 1.1.1.*** Prove Proposition 1.1.5.

2.1.2 Gauss Transformation and LU-Factorization

Under certain conditions the system matrix A of the equation $A\mathbf{x} = \mathbf{b}$ can be expressed in the form of a product of a *unit lower triangular* matrix L with units on the main diagonal and an upper triangular matrix U , and as the result, one has to solve two systems with triangular matrices.

Proposition 1.2.1 (*LU-method*). If $A \in \mathbf{R}^{n \times n}$, $A = LU$, where L is unit lower triangular, U is regular upper triangular and $A\mathbf{x} = \mathbf{b}$, then $LU\mathbf{x} = \mathbf{b}$, and for the solution of the system one has first to solve the system $L\mathbf{y} = \mathbf{b}$ and then the system $U\mathbf{x} = \mathbf{y}$.

Example 1.2.1. Solve the system

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

using *LU-method*. Since

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix},$$

then, by Proposition 1.2.1, we have to solve first the system

$$\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}.$$

The solution of this system is $\eta_1 = 1$ and $\eta_2 = 5 - 3 \cdot 1 = 2$. Second, solving the system

$$\begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

we find that $\xi_2 = -1$ and $\xi_1 = 1 - 2 \cdot (-1) = 3$. Thus, $\mathbf{x} = \begin{bmatrix} 3 & -1 \end{bmatrix}^T$.

The Gaussian elimination method considered in the main course of linear algebra for solution of systems of linear equations is also applicable also to the *LU-factorization*. Let $\mathbf{x} \in \mathbf{R}^m$, where $\xi_k \neq 0$. If

$$\tau_i = \xi_i / \xi_k \quad (i = (k+1) : m) \quad \mathbf{t}^{(k)} = \begin{bmatrix} 0 & \cdots & 0 & \tau_{k+1} & \cdots & \tau_m \\ k \text{ zeros} \end{bmatrix}$$

and

$$M_k = I - \mathbf{t}^{(k)} \mathbf{e}_k^T, \quad (2)$$

then

$$M_k \mathbf{x} = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & & 1 & 0 & & 0 \\ 0 & & -\tau_{k+1} & 1 & & 0 \\ \vdots & \vdots & & \vdots & \ddots & \\ 0 & \cdots & -\tau_m & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_k \\ \xi_{k+1} \\ \vdots \\ \xi_m \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Definition 1.2.1. A matrix M_k of the form (2) is called a *Gauss matrix*, the components $\mathbf{t}((k+1) : n)$ are called *Gauss multipliers*, and the vector $\mathbf{t}^{(k)}$ is called the *Gauss vector*. The transformation defined with the Gauss matrix M_k is called the *Gauss transformation*.

Definition 1.2.2. The value

$$d_k = \begin{cases} a_{11}, & \text{if } k = 1, \\ \det(A(1 : k, 1 : k)) / \det(A(1 : k-1, 1 : k-1)), & \text{if } k = 2 : p, \end{cases}$$

is called the k -th *pivot* of the matrix $A \in \mathbf{R}^{m \times n}$, where $p = \min(m, n)$ and $\det(A(1 : i, 1 : i)) \neq 0$ ($i = 1 : p-1$).

If $A \in \mathbf{R}^{n \times n}$, then for the nonzero pivots of A the Gauss matrices $M_1, \dots, M_{n-1}$ can be found such that $M_{n-1}M_{n-2} \cdots M_2M_1A = U$ is upper triangular.

Example 1.2.2. Let us consider the finding of the Gauss matrices M_1 and M_2 and the upper triangular matrix U for

$$A = \begin{bmatrix} 2 & 2 & -1 \\ 4 & 5 & 2 \\ -2 & 1 & 2 \end{bmatrix}$$

By relation (2), we obtain that

$$\begin{aligned} M_1 &= I - \mathbf{t}^{(1)} \mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 4/2 \\ (-2)/2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Thus,

$$M_1 A = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & -1 \\ 4 & 5 & 2 \\ -2 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 2 & -1 \\ 0 & 1 & 4 \\ 0 & 3 & 1 \end{bmatrix}$$

and

$$\begin{aligned} M_2 &= I - \mathbf{t}^{(2)} \mathbf{e}_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{bmatrix}. \end{aligned}$$

Therefore,

$$U = M_2 M_1 A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & -1 \\ 0 & 1 & 4 \\ 0 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & -1 \\ 0 & 1 & 4 \\ 0 & 0 & -11 \end{bmatrix}.$$

Note that matrix $A^{(k-1)} = M_{k-1} \cdots M_1 A$ is upper triangular in columns 1 to $k-1$, and for the calculation of the elements of the Gauss matrix M_k we use the matrix vector $A^{(k-1)}(k : m, k)$. The calculation of M_k is possible if $a_{kk}^{(k-1)} \neq 0$. Moreover, $M_k^{-1} = I + \mathbf{t}^{(k)} \mathbf{e}_k^T$. If to choose

$$L = M_1^{-1} \cdots M_{n-1}^{-1},$$

then

$$A = LU.$$

We stress that in our treatment the lower triangular matrix L is a unit lower triangular matrix.

Proposition 1.2.2. If $\det(A(1:k, 1:k)) \neq 0$ for $(k = 1:n-1)$, then $A \in R^{n \times n}$ has an LU factorization. If the LU factorization exists and A is regular, then the LU factorization is unique and $\det(A) = u_{11} \cdots u_{nn}$.

Proof. Suppose $k-1$ steps have been taken and the matrix $A^{(k-1)} = M_{k-1} \cdots M_1 A$ has been found. The element $a_{kk}^{(k-1)}$ is the k -th pivot of A and $\det(A(1:k, 1:k)) = a_{11}^{(k-1)} \cdots a_{kk}^{(k-1)}$. Hence, if $A(1:k, 1:k)$ is regular, then $a_{kk}^{(k-1)} \neq 0$, and A has an LU factorization. Let us suppose that the regular matrix A has two LU factorizations $A = L_1 U_1$ and $A = L_2 U_2$. We have

$L_1U_1 = L_2U_2$ or $L_2^{-1}L_1 = U_2U_1^{-1}$. Since $L_2^{-1}L_1$ is unit lower triangular and $U_2U_1^{-1}$ is upper triangular, then $L_2^{-1}L_1 = I$, $U_2U_1^{-1} = I$ and $L_2 = L_1$ and also $U_2 = U_1$. $\square$

Example 1.2.3.* Find the LU factorization of the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 8 & 7 \end{bmatrix}.$$

Find the Gauss matrix M_1 for A :

$$\begin{aligned} M_1 &= I - \mathbf{t}^{(1)}\mathbf{e}_1^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 8/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 4 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -4 & 1 \end{bmatrix}. \end{aligned}$$

Thus,

$$M_1A = \begin{bmatrix} 1 & 0 \\ -4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 8 & 7 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}, \quad M_1^{-1} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix}$$

and

$$L = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}, \quad A = LU = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}.$$

Example 1.2.4.* Find the LU factorization of

$$A = \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 6 & 9 & 8 \end{bmatrix}.$$

Find the Gauss matrix M_1 for A :

$$\begin{aligned} M_1 &= I - \mathbf{t}^{(1)}\mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 6/2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Thus,

$$M_1 A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 6 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 0 & 0 & -1 \end{bmatrix}, \quad M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}$$

and since $M_1 A$ is upper triangular, then $M_2 = I$ and

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 0 & 0 & -1 \end{bmatrix},$$

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 0 & 0 & -1 \end{bmatrix}.$$

Example 1.2.5.* By using the LU factorization, solve the system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 6 & 9 & 8 \end{bmatrix} \quad \wedge \quad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 5 \end{bmatrix}.$$

In example 1.2.4 we found the LU factorization for A :

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 0 & 0 & -1 \end{bmatrix}.$$

By solving the system $L\mathbf{y} = \mathbf{b}$, i.e.,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 5 \end{bmatrix},$$

we obtain

$$\mathbf{y} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix}.$$

By solving the system $U\mathbf{x} = \mathbf{y}$, i.e.,

$$\begin{bmatrix} 2 & 3 & 3 \\ 0 & 5 & 7 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix},$$

we obtain that

$$\mathbf{x} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

Exercise 1.2.1.* Find the LU factorization of A if

$$a) A = \begin{bmatrix} 3 & 1 \\ 6 & 7 \end{bmatrix}, \quad b) A = \begin{bmatrix} 1 & 0 \\ 8 & 1 \end{bmatrix}, \quad c) A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Exercise 1.2.2.* By using the LU factorization solve the system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \quad \wedge \quad \mathbf{b} = \begin{bmatrix} 2 \\ -3 \\ 4 \end{bmatrix}.$$

If the principal minors of a rectangular matrix $A \in \mathbf{R}^{m \times n}$ are nonzero, i.e.,

$$\det(A(1:k, 1:k)) \neq 0 \quad (k = 1 : \min(m, n)),$$

then A has an LU factorization.

Example 1.2.6. The following equalities hold:

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

$$\begin{bmatrix} 2 & 8 & 4 \\ 1 & 6 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 8 & 4 \\ 0 & 2 & 3 \end{bmatrix}.$$

As is known from the main course of algebra, the direct application of the Gaussian elimination, therefore also the direct realization of the LU factorization fails, if at least one of the principal minors is singular. It turns out that for a regular matrix it is possible after an appropriate interchange of matrix rows to find the LU factorization. Permutation matrices are used for interchanging the matrix rows (columns).

Definition 1.2.3. A *permutation matrix* $P \in \mathbf{R}^{n \times n}$ is the identity I with its rows reordered.

Example 1.2.7. Consider the effect of multiplying a 4×4 matrix A by a concrete permutation matrix P .

$$PA = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{11} & a_{12} & a_{13} & a_{14} \end{bmatrix}$$

Multiplying by the permutation matrix P on the left, we obtain a new matrix, where the rows of initial the matrix are reordered exactly in the same way as the rows of the identity I are reordered for getting P . Multiplying on the right,

$$AP = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} a_{14} & a_{13} & a_{11} & a_{12} \\ a_{24} & a_{23} & a_{21} & a_{22} \\ a_{34} & a_{33} & a_{31} & a_{32} \\ a_{44} & a_{43} & a_{41} & a_{42} \end{bmatrix},$$

we obtain a new matrix, where the columns of the initial matrix are reordered in the same way as the columns of the identity I are reordered for getting P . The following holds

Proposition 1.2.3. If $A \in \mathbf{R}^{n \times n}$ and $\det(A) \neq 0$, then there exists a permutation matrix $P \in \mathbf{R}^{n \times n}$ such that all the principal minors of PA are nonzero, and consequently, there exists the LU factorization

$$PA = LU.$$

Example 1.2.8.* Let

$$A = \begin{bmatrix} 0 & -2 & 2 \\ 1 & 2 & -1 \\ 3 & 5 & -8 \end{bmatrix}.$$

Find for a certain permutation matrix $P \in \mathbf{R}^{3 \times 3}$ the LU factorization of PA . Interchange the first and second rows of A , i.e., choose

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and find for the matrix

$$PA = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -2 & 2 \\ 1 & 2 & -1 \\ 3 & 5 & -8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 3 & 5 & -8 \end{bmatrix}$$

the Gauss matrix

$$M_1 = I - \mathbf{t}^{(1)} \mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 3/1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}.$$

Thus,

$$M_1 PA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 3 & 5 & -8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 0 & -1 & -5 \end{bmatrix}, \quad M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}$$

and

$$M_2 = I - \mathbf{t}^{(2)} \mathbf{e}_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ (-1)/(-2) \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}$$

and

$$M_2 M_1 PA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 0 & -1 & -5 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 0 & 0 & -6 \end{bmatrix}, \quad M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}.$$

Consequently,

$$L = M_1^{-1} M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & \frac{1}{2} & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 0 & 0 & -6 \end{bmatrix}$$

and

$$PA = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 2 \\ 0 & 0 & -6 \end{bmatrix}.$$

Exercise 1.2.3.* Find for a certain permutation matrix P the LU factorization of PA if

$$a) A = \begin{bmatrix} 0 & 2 \\ 3 & 5 \end{bmatrix}, \quad b) A = \begin{bmatrix} 0 & 7 \\ 5 & 3 \end{bmatrix}, \quad c) A = \begin{bmatrix} 0 & 5 & 7 \\ 2 & 3 & 3 \\ 6 & 9 & 8 \end{bmatrix}.$$

2.2 QR Factorization

2.2.1 Householder Reflection

Definition 2.1.1. If $\mathbf{v} \in \mathbf{R}^n$ and $\mathbf{v} \neq \mathbf{0}$, then the matrix of the form

$$H = I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}} \quad (1)$$

is called the *Householder matrix* or *Householder reflection* and the vector $\mathbf{v}$ is called the *Householder vector*.

Proposition 2.1.1. The Householder matrix H is symmetric and orthogonal. The Householder reflection reflects every vector $\mathbf{x} \in \mathbf{R}^n$ in the hyperplane $\text{span}\{\mathbf{v}\}^\perp$.

Proof.

$$H^T = I^T - 2 \frac{(\mathbf{v}\mathbf{v}^T)^T}{\mathbf{v}^T\mathbf{v}} = I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}} = H$$

and

$$HH^T = H^TH = (I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}})^2 = I - 4 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}} + 4 \frac{\mathbf{v}(\mathbf{v}^T\mathbf{v})\mathbf{v}^T}{(\mathbf{v}^T\mathbf{v})(\mathbf{v}^T\mathbf{v})} = I.$$

To prove the third part of the assertion, we choose on the hyperplane $\text{span}\{\mathbf{v}\}^\perp$ an orthogonal basis $\{\mathbf{a}_1, \dots, \mathbf{a}_{n-1}\}$. Hence, $\mathbf{v} \perp \mathbf{a}_i$ ($i = 1 : n-1$) and $\mathbf{v}^T\mathbf{a}_i = 0$ ($i = 1 : n-1$). If

$$\mathbf{x} = \beta\mathbf{v} + \beta_1\mathbf{a}_1 + \dots + \beta_{n-1}\mathbf{a}_{n-1},$$

then

$$\begin{aligned} H\mathbf{x} &= H(\beta\mathbf{v}) + H(\beta_1\mathbf{a}_1) + \dots + H(\beta_{n-1}\mathbf{a}_{n-1}) = \\ &= \beta(I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}})\mathbf{v} + \beta_1(I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}})\mathbf{a}_1 + \dots + \beta_{n-1}(I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}})\mathbf{a}_{n-1} = \\ &= \beta(\mathbf{v} - 2 \frac{\mathbf{v}(\mathbf{v}^T\mathbf{v})}{\mathbf{v}^T\mathbf{v}}) + \beta_1(\mathbf{a}_1 - 2 \frac{\mathbf{v}(\mathbf{v}^T\mathbf{a}_1)}{\mathbf{v}^T\mathbf{v}}) + \dots + \beta_{n-1}(\mathbf{a}_{n-1} - 2 \frac{\mathbf{v}(\mathbf{v}^T\mathbf{a}_{n-1})}{\mathbf{v}^T\mathbf{v}}) = \\ &= -\beta\mathbf{v} + \beta_1\mathbf{a}_1 + \dots + \beta_{n-1}\mathbf{a}_{n-1}, \end{aligned}$$

i.e., the vectors $\mathbf{x}$ and $H\mathbf{x}$ have the same orthogonal projection onto the hyperplane $\text{span}\{\mathbf{v}\}^\perp$

$$\beta_1\mathbf{a}_1 + \dots + \beta_{n-1}\mathbf{a}_{n-1},$$

but projections onto the vector $\mathbf{v}$ have opposite directions. Thus $H\mathbf{x}$ is the reflection of $\mathbf{x}$ in the hyperplane $\text{span}\{\mathbf{v}\}^\perp$. It is significant to note that the Householder matrix H depends only on the direction of Householder vector $\mathbf{v}$ and does not depend on the sign of the direction and length of $\mathbf{v}$.

Proposition 2.1.2. If $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{v} = \mathbf{x} \pm \|\mathbf{x}\|_2 \mathbf{e}_1$, then vector $H\mathbf{x}$, where H is the Householder matrix denoted by (1), has the same direction as $\mathbf{e}_1$, i.e., the Householder reflection H applied to the vector $\mathbf{x}$ annihilates all but the first component of the vector $\mathbf{x}$.

Proof. Our aim is to determine for a nonzero vector $\mathbf{x}$ the Householder vector $\mathbf{v}$ so that $H\mathbf{x} \in \text{span}\{\mathbf{e}_1\}$. Since

$$H\mathbf{x} = (I - 2\frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}})\mathbf{x} = \mathbf{x} - 2\frac{\mathbf{v}(\mathbf{v}^T\mathbf{x})}{\mathbf{v}^T\mathbf{v}} = \mathbf{x} - 2\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{v}$$

and $H\mathbf{x} \in \text{span}\{\mathbf{e}_1\}$, then $\mathbf{v} \in \text{span}\{\mathbf{x}, \mathbf{e}_1\}$. By choosing $\mathbf{v} = \mathbf{x} + \alpha\mathbf{e}_1$, we obtain that

$$\begin{aligned}\mathbf{v}^T\mathbf{x} &= \mathbf{x}^T\mathbf{x} + \alpha\mathbf{e}_1^T\mathbf{x} = \mathbf{x}^T\mathbf{x} + \alpha\xi_1, \\ \mathbf{v}^T\mathbf{v} &= (\mathbf{x}^T + \alpha\mathbf{e}_1^T)(\mathbf{x} + \alpha\mathbf{e}_1) = \mathbf{x}^T\mathbf{x} + 2\alpha\xi_1 + \alpha^2\end{aligned}$$

and

$$\begin{aligned}H\mathbf{x} &= \mathbf{x} - 2\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{v} = \mathbf{x} - 2\frac{\mathbf{x}^T\mathbf{x} + \alpha\xi_1}{\mathbf{x}^T\mathbf{x} + 2\alpha\xi_1 + \alpha^2}(\mathbf{x} + \alpha\mathbf{e}_1) = \\ &= (1 - 2\frac{\mathbf{x}^T\mathbf{x} + \alpha\xi_1}{\mathbf{x}^T\mathbf{x} + 2\alpha\xi_1 + \alpha^2})\mathbf{x} - 2\alpha\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{e}_1.\end{aligned}$$

Choose α so that in the latter representation of $H\mathbf{x}$ the coefficient of $\mathbf{x}$ is zero, i.e.,

$$\begin{aligned}1 - 2\frac{\mathbf{x}^T\mathbf{x} + \alpha\xi_1}{\mathbf{x}^T\mathbf{x} + 2\alpha\xi_1 + \alpha^2} &= 0 \Leftrightarrow \\ \Leftrightarrow \mathbf{x}^T\mathbf{x} + 2\alpha\xi_1 + \alpha^2 - 2\mathbf{x}^T\mathbf{x} - 2\alpha\xi_1 &= 0 \Leftrightarrow \\ \Leftrightarrow \mathbf{x}^T\mathbf{x} = \alpha^2 &\Leftrightarrow \|\mathbf{x}\|_2 = \pm\alpha.\end{aligned}$$

For this choice $\alpha = \pm \|\mathbf{x}\|_2$ we have $\mathbf{v} = \mathbf{x} \pm \|\mathbf{x}\|_2 \mathbf{e}_1$ and

$$H\mathbf{x} = -2\alpha\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{e}_1 = -2\alpha\frac{\mathbf{x}^T\mathbf{x} \pm \alpha\xi_1}{\mathbf{x}^T\mathbf{x} \pm 2\alpha\xi_1 + \mathbf{x}^T\mathbf{x}}\mathbf{e}_1 = -\alpha\mathbf{e}_1 = \mp \|\mathbf{x}\|_2 \mathbf{e}_1.$$

Example 2.1.1 Let $\mathbf{x} = [2 \ 6 \ -3]^T$. Find the Householder vector $\mathbf{v}$ and hence to it the Householder transformation that annihilates the two last coordinates of the vector $\mathbf{x}$. By Proposition 2.1.1 we compute $\mathbf{v} = \mathbf{x} \pm \|\mathbf{x}\|_2 \mathbf{e}_1 =$

$[2 \ 6 \ -3]^T \pm 7\mathbf{e}_1$. Choose the sign plus for coefficient of $\mathbf{e}_1$ and we obtain $\mathbf{v} = [9 \ 6 \ -3]^T$. Find the Householder matrix H that depends only on direction of $\mathbf{v}$,

$$\begin{aligned} H &= I - \frac{2}{\mathbf{v}^T \mathbf{v}} \mathbf{v} \mathbf{v}^T = I - \frac{2}{14} \begin{bmatrix} 3 \\ 2 \\ -1 \end{bmatrix} \begin{bmatrix} 3 & 2 & -1 \end{bmatrix} = \\ &= I - \frac{1}{7} \begin{bmatrix} 9 & 6 & -3 \\ 6 & 4 & -2 \\ -3 & -2 & 1 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} -2 & -6 & 3 \\ -6 & 3 & 2 \\ 3 & 2 & 6 \end{bmatrix}. \end{aligned}$$

Check,

$$H\mathbf{x} = \frac{1}{7} \begin{bmatrix} -2 & -6 & 3 \\ -6 & 3 & 2 \\ 3 & 2 & 6 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ -3 \end{bmatrix} = \begin{bmatrix} -7 \\ 0 \\ 0 \end{bmatrix}.$$

Exercise 2..1.1.* Find the Householder matrix H such that $H\mathbf{x} \in \text{span}\{\mathbf{e}_1\}$, where $\mathbf{x} = [-3 \ 1 \ -5 \ 1]^T$.

Let $Q_i \in \mathbf{R}^{n \times n}$ ($i=1:r$) be the Householder matrices. Consider the product of these matrices

$$Q = Q_1 \cdots Q_r,$$

where

$$Q_j = I - \beta_j \mathbf{v}^{(j)} \mathbf{v}^{(j)T}$$

and each $\mathbf{v}^{(j)}$ has the form

$$\mathbf{v}^{(j)} = \begin{bmatrix} 0 & \cdots & 0 & 1 & \nu_{j+1}^{(j)} & \cdots & \nu_n^{(j)} \end{bmatrix}^T.$$

$j-1$ zeros

The matrix Q can be written in the form

$$Q = I + WY^T, \tag{2}$$

where W and Y are $n \times r$ -matrices. The answer to the question how to find representation (2) is given by the following proposition.

Proposition 2.1.3. Suppose $Q = I + WY^T \in \mathbf{R}^{n \times n}$ is an orthogonal matrix with $W, Y \in \mathbf{R}^{n \times j}$. If $H = I - \beta \mathbf{v} \mathbf{v}^T$, where $\mathbf{v} \in \mathbf{R}^n$ and $\mathbf{z} = -\beta Q \mathbf{v}$, then

$$Q_+ = QH = I + W_+ Y_+^T,$$

where $W_+ = [W \ z]$ and $Y_+ = [Y \ \mathbf{v}]$, and consequently, $W_+, Y_+ \in \mathbf{R}^{n \times (j+1)}$.

Proof. Since

$$\begin{aligned} QH &= (I + WY^T)(I - \beta\mathbf{v}\mathbf{v}^T) = I + WY^T - \beta(I + WY^T)\mathbf{v}\mathbf{v}^T = \\ &= I + WY^T - \beta Q\mathbf{v}\mathbf{v}^T = I + WY^T - \mathbf{z}\mathbf{v}^T \end{aligned}$$

and

$$I + [W \ z] \begin{bmatrix} Y^T \\ \mathbf{v}^T \end{bmatrix} = I + [WY^T + \mathbf{z}\mathbf{v}^T] = I + WY^T + \mathbf{z}\mathbf{v}^T,$$

then $QH = I + W_+Y_+^T$, and the assertion of the proposition holds.

2.2.2 Givens Rotations

The Householder reflection is effective for intruducing zeros if there are "many" components to be annihilated. If it is necessary to nullify only one component or sometimes a couple of components, then usually the Givens method is used. The *Givens rotation* is realized by an $n \times n$ -matrix

$$G(i, k, \theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} i \\ \\ k \\ \\ i \\ k \end{matrix}$$

where $c = \cos \theta$ and $s = \sin \theta$. This matrix $G(i, k, \theta)$ is evidently orthogonal. If $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} = G(i, k, \theta)^T \mathbf{x}$, then

$$\eta_j = \begin{cases} c\xi_i - s\xi_k, & j = i, \\ s\xi_i + c\xi_k, & j = k, \\ \xi_j, & j \neq i, k. \end{cases}$$

By setting

$$c = \frac{\xi_i}{\sqrt{\xi_i^2 + \xi_k^2}} \wedge s = \frac{-\xi_k}{\sqrt{\xi_i^2 + \xi_k^2}} ,$$

we get $\eta_k = 0$.

Example 2.2.1. Consider the annihilation of the last component of the vector $\mathbf{x} = [2 \ 6 \ -3]^T$ given in example 2.1.1 with Givens rotations. Find the values of c and s :

$$c = \frac{\xi_2}{\sqrt{\xi_2^2 + \xi_3^2}} = \frac{6}{\sqrt{45}} = \frac{2\sqrt{5}}{5} ,$$

$$s = \frac{-\xi_3}{\sqrt{\xi_2^2 + \xi_3^2}} = \frac{3}{\sqrt{45}} = \frac{\sqrt{5}}{5} .$$

Check

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ 0 & \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{15\sqrt{5}}{5} \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 3\sqrt{5} \\ 0 \end{bmatrix} .$$

2.2.3 Householder QR Factorization

Apply the Householder reflection to the matrix $A \in \mathbf{R}^{m \times n}$ ($m \geq n$) to obtain the QR factorization.

Example 2.3.1. Suppose $A \in \mathbf{R}^{5 \times 4}$ and assume that the Householder matrices H_1 and H_2 have been computed so that

$$H_2 H_1 A = \begin{bmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \otimes & \times \\ 0 & 0 & \otimes & \times \\ 0 & 0 & \otimes & \times \end{bmatrix} .$$

Concentrating on the highlighted vector $\begin{bmatrix} \otimes \\ \otimes \\ \otimes \end{bmatrix}$, we determine a Householder

matrix $\widetilde{H}_3$ such that

$$\widetilde{H}_3 \begin{bmatrix} \otimes \\ \otimes \\ \otimes \end{bmatrix} = \begin{bmatrix} \times \\ 0 \\ 0 \end{bmatrix}$$

Choosing $H_3 = \text{diag}(I_2, \widetilde{H}_3)$, we get

$$H_3 H_2 H_1 A = \begin{bmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \otimes \\ 0 & 0 & 0 & \otimes \end{bmatrix}.$$

Next consider the highlighted vector $\begin{bmatrix} \otimes \\ \otimes \end{bmatrix}$ and determine $\widetilde{H}_4$ such that

$$\widetilde{H}_4 \begin{bmatrix} \otimes \\ \otimes \end{bmatrix} = \begin{bmatrix} \times \\ 0 \end{bmatrix}.$$

Choosing $H_4 = \text{diag}(I_3, \widetilde{H}_4)$, we have

$$H_4 H_3 H_2 H_1 A = \begin{bmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 \end{bmatrix} = R.$$

By setting $Q = H_1 H_2 H_3 H_4$, we obtain $QR = H_1 H_2 H_3 H_4 H_4 H_3 H_2 H_1 A = A$.

Proposition 2.3.1. If $A \in \mathbf{R}^{m \times n}$ ($m \geq n$), then there exist Householder matrices H_i such that

$$Q = \begin{cases} H_1 \cdots H_n, & \text{kui } m > n, \\ H_1 \cdots H_{n-1}, & \text{kui } m = n \end{cases},$$

$$R = \begin{cases} H_n \cdots H_1 A, & \text{kui } m > n, \\ H_1 \cdots H_{n-1} A, & \text{kui } m = n \end{cases}$$

and

$$A = QR,$$

where $Q \in \mathbf{R}^{m \times m}$ is orthogonal and $R \in \mathbf{R}^{m \times n}$ is upper triangular.

Example 2.3.2. Find the Householder QR factorization for

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & -1 & -1 \end{bmatrix}.$$

In example 2.1.1 the Householder matrix for the transformation of the first column vector $[2 \ 6 \ -3]^T$ of A has been found:

$$H_1 = \frac{1}{7} \begin{bmatrix} -2 & -6 & 3 \\ -6 & 3 & 2 \\ 3 & 2 & 6 \end{bmatrix}.$$

Then

$$H_1 A = \frac{1}{7} \begin{bmatrix} -2 & -6 & 3 \\ -6 & 3 & 2 \\ 3 & 2 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & -1 & -1 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} -49 & -15 & -5 \\ 0 & 4 & -8 \\ 0 & -2 & -3 \end{bmatrix}.$$

To find $\widetilde{H}_2$, we compute the Householder vector

$$\mathbf{v} = \begin{bmatrix} 4 \\ -2 \end{bmatrix} - \sqrt{20} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 - 2\sqrt{5} \\ -2 \end{bmatrix}.$$

Hence

$$\widetilde{H}_2 = I - 2 \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} = \cdots = \frac{\sqrt{5}}{5} \begin{bmatrix} 2 & -1 \\ -1 & -2 \end{bmatrix}$$

and

$$H_2 = \text{diag}(I_1, \widetilde{H}_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ 0 & -\frac{\sqrt{5}}{5} & -\frac{2\sqrt{5}}{5} \end{bmatrix}$$

and also

$$\begin{aligned} R = H_2 H_1 A &= \frac{1}{7} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ 0 & -\frac{\sqrt{5}}{5} & -\frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} -49 & -15 & -5 \\ 0 & 4 & -8 \\ 0 & -2 & -3 \end{bmatrix} = \\ &= \begin{bmatrix} -7 & -\frac{15}{7} & -\frac{5}{7} \\ 0 & \frac{2\sqrt{5}}{7} & -\frac{13\sqrt{5}}{35} \\ 0 & 0 & \frac{2\sqrt{5}}{5} \end{bmatrix}. \end{aligned}$$

Find also the orthogonal matrix

$$\begin{aligned} Q = H_1 H_2 &= \frac{1}{7} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ 0 & -\frac{\sqrt{5}}{5} & -\frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} -2 & -6 & 3 \\ -6 & 3 & 2 \\ 3 & 2 & 6 \end{bmatrix} = \\ &= \frac{\sqrt{5}}{35} \begin{bmatrix} -2\sqrt{5} & -15 & 0 \\ -6\sqrt{5} & 4 & -7 \\ 3\sqrt{5} & -2 & -14 \end{bmatrix} \end{aligned}$$

and check the result

$$QR = \frac{\sqrt{5}}{35} \begin{bmatrix} -2\sqrt{5} & -15 & 0 \\ -6\sqrt{5} & 4 & -7 \\ 3\sqrt{5} & -2 & -14 \end{bmatrix} \begin{bmatrix} -7 & -\frac{15}{7} & -\frac{5}{7} \\ 0 & \frac{2\sqrt{5}}{7} & -\frac{13\sqrt{5}}{35} \\ 0 & 0 & \frac{2\sqrt{5}}{5} \end{bmatrix} = A.$$

Example 2.3.3*. Find the Householder QR factorization of

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

The vector that has to be transformed is $\mathbf{x} = \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}^T$, where $\|\mathbf{x}\|_2 = \sqrt{1^2 + 2^2 + 2^2} = 3$. Construct the vector

$$\mathbf{v} = \mathbf{x} \pm \|\mathbf{x}\|_2 \mathbf{e}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \pm 3\mathbf{e}_1.$$

Choose a minus sign for the coefficient of $\mathbf{e}_1$ and take into account that H depends only on the direction of $\mathbf{v}$:

$$\mathbf{v} = \begin{bmatrix} -2 & 2 & 2 \end{bmatrix}^T \sim \begin{bmatrix} -1 & 1 & 1 \end{bmatrix}^T.$$

Find the Householder matrix

$$H_1 = I - \frac{2}{\mathbf{v}^T \mathbf{v}} \mathbf{v} \mathbf{v}^T = I - \frac{2}{(-1)(-1) + 1 \cdot 1 + 1 \cdot 1} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 1 \end{bmatrix} =$$

$$= I - \frac{2}{3} \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{bmatrix}.$$

Verify that H_1 annihilates all the elements of the first column of A but the first one.

$$H_1 A = \frac{1}{3} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 9 & 9 \\ 0 & 3 \\ 0 & -3 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}.$$

Further we transform the vector $\mathbf{x} = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$, where $\|\mathbf{x}\|_2 = \sqrt{2}$. Find the Householder vector according to $\mathbf{x}$

$$\mathbf{v} = \mathbf{x} \pm \|\mathbf{x}\|_2 \mathbf{e}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \pm \sqrt{2} \mathbf{e}_1.$$

Choose a minus sign for coefficient of $\mathbf{e}_1$:

$$\mathbf{v} = \begin{bmatrix} -1 - \sqrt{2} \\ 1 \end{bmatrix}.$$

Using this vector we obtain the Householder matrix

$$\begin{aligned} \widetilde{H}_2 &= I - \frac{2}{\mathbf{v}^T \mathbf{v}} \mathbf{v} \mathbf{v}^T = I - \frac{2}{(-1 - \sqrt{2})^2 + 1} \begin{bmatrix} -1 - \sqrt{2} \\ 1 \end{bmatrix} \begin{bmatrix} -1 - \sqrt{2} & 1 \end{bmatrix} = \\ &= I - \frac{2}{2(2 - \sqrt{2})} \begin{bmatrix} 3 - 2\sqrt{2} & \sqrt{2} - 1 \\ \sqrt{2} - 1 & 1 \end{bmatrix} = \frac{1}{2 - \sqrt{2}} \begin{bmatrix} -1 + \sqrt{2} & -\sqrt{2} + 1 \\ -\sqrt{2} + 1 & -\sqrt{2} + 1 \end{bmatrix} = \\ &= \frac{\sqrt{2} - 1}{2 - \sqrt{2}} \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix} \end{aligned}$$

and find that

$$H_2 = \text{diag}(I_1, \widetilde{H}_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2}/2 & -\sqrt{2}/2 \\ 0 & -\sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}.$$

Thus,

$$R = H_2 H_1 A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & -\frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix}$$

and

$$Q = H_1 H_2 = \frac{1}{3} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & -\frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{2}{3}\sqrt{2} \\ \frac{2}{3} & \frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{2} \\ \frac{2}{3} & -\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{2} \end{bmatrix}.$$

Let us check the result:

$$QR = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{2}{3}\sqrt{2} \\ \frac{2}{3} & \frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{2} \\ \frac{2}{3} & -\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} = A.$$

Exercise 2.3.1.* Find the QR factorization of A if

$$a) A = \begin{bmatrix} 0 & 0 \\ 1 & 3 \\ 0 & 2 \end{bmatrix}; \quad b) A = \begin{bmatrix} 5 & 9 \\ 12 & 7 \end{bmatrix}; \quad c) A = \begin{bmatrix} 3 & 3 & 0 \\ 3 & 5 & 0 \\ 0 & 0 & 6 \end{bmatrix};$$

$$d) A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 3 & 1 & 0 & 0 \\ 5 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

2.2.4 Givens QR Factorization

Next we consider how to use the Givens rotations to compute the QR factorization of a given matrix.

Example 2.4.1. Consider for $A \in \mathbf{R}^{4 \times 3}$ the idea of the *Givens QR factorization*:

$$A = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{G_1^T(3,4)} \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{G_2^T(2,3)} \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{G_3^T(1,2)}$$

$$\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{G_4^T(3,4)} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{G_5^T(2,3)} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{G_6^T(3,4)}$$

$$\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix} = R$$

The orthogonal matrix has the form:

$$Q = G_1(3, 4)G_2(2, 3)G_3(1, 2)G_4(3, 4)G_5(2, 3)G_6(3, 4).$$

Example 2.4.2.* Find the Givens QR factorization of

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & 1 & -1 \end{bmatrix}.$$

Let us annihilate the element $A(3, 1)$ of A . For this we construct the Givens matrix $G_1(2, 3)$. Find the values c and s :

$$c = \frac{6}{\sqrt{6^2 + (-3)^2}} = \frac{6}{\sqrt{45}} = \frac{2\sqrt{5}}{5}, \quad s = \frac{3}{\sqrt{6^2 + (-3)^2}} = \frac{3}{\sqrt{45}} = \frac{\sqrt{5}}{5}.$$

Thus, we have

$$G_1(2, 3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & -\frac{1}{5}\sqrt{5} & \frac{2}{5}\sqrt{5} \end{bmatrix}$$

and

$$\begin{aligned} A^{(1)} = G_1^T(2, 3)A &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2}{5}\sqrt{5} & -\frac{1}{5}\sqrt{5} \\ 0 & \frac{1}{5}\sqrt{5} & \frac{2}{5}\sqrt{5} \end{bmatrix} \begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & 1 & -1 \end{bmatrix} = \\ &= \begin{bmatrix} 2 & 0 & 1 \\ 3\sqrt{5} & \frac{3}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & \frac{4}{5}\sqrt{5} & -\frac{2}{5}\sqrt{5} \end{bmatrix}. \end{aligned}$$

For the annihilation of the element $A^{(1)}(2, 1)$ of $A^{(1)}$ we construct the Givens matrix $G_2(1, 2)$. Find the values c and s :

$$c = \frac{2}{\sqrt{2^2 + (3\sqrt{5})^2}} = \frac{2}{7}, \quad s = \frac{-3\sqrt{5}}{\sqrt{2^2 + (3\sqrt{5})^2}} = -\frac{3\sqrt{5}}{7}.$$

Thus,

$$G_2(1, 2) = \begin{bmatrix} \frac{2}{7} & -\frac{3}{7}\sqrt{5} & 0 \\ \frac{3}{7}\sqrt{5} & \frac{2}{7} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$\begin{aligned} A^{(2)} &= G_2^T(1, 2)A^{(1)} = \begin{bmatrix} \frac{2}{7} & \frac{3}{7}\sqrt{5} & 0 \\ -\frac{3}{7}\sqrt{5} & \frac{2}{7} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 1 \\ 3\sqrt{5} & \frac{3}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & \frac{4}{5}\sqrt{5} & -\frac{2}{5}\sqrt{5} \end{bmatrix} = \\ &= \begin{bmatrix} 7 & \frac{9}{7} & \frac{5}{7} \\ 0 & \frac{6}{35}\sqrt{5} & -\frac{13}{35}\sqrt{5} \\ 0 & \frac{4}{5}\sqrt{5} & -\frac{2}{5}\sqrt{5} \end{bmatrix}. \end{aligned}$$

To annihilate the element $A^{(2)}(3, 2)$ of $A^{(2)}$ we construct the Givens matrix $G_3(2, 3)$. Find the values of c and s :

$$c = \frac{\frac{6}{35}\sqrt{5}}{\sqrt{\left(\frac{6}{35}\sqrt{5}\right)^2 + \left(\frac{4}{5}\sqrt{5}\right)^2}} = \frac{\frac{6}{35}\sqrt{5}}{\frac{2}{7}\sqrt{41}} = \frac{3}{205}\sqrt{205}$$

and

$$s = \frac{-\frac{4}{5}\sqrt{5}}{\sqrt{\left(\frac{6}{35}\sqrt{5}\right)^2 + \left(\frac{4}{5}\sqrt{5}\right)^2}} = -\frac{14}{205}\sqrt{205}.$$

Thus,

$$G_3(2, 3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{205}\sqrt{205} & -\frac{14}{205}\sqrt{205} \\ 0 & \frac{14}{205}\sqrt{205} & \frac{3}{205}\sqrt{205} \end{bmatrix}$$

and

$$R = G_3^T(2, 3)A^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{205}\sqrt{205} & \frac{14}{205}\sqrt{205} \\ 0 & -\frac{14}{205}\sqrt{205} & \frac{3}{205}\sqrt{205} \end{bmatrix} \begin{bmatrix} 7 & \frac{9}{7} & \frac{5}{7} \\ 0 & \frac{6}{35}\sqrt{5} & -\frac{13}{35}\sqrt{5} \\ 0 & \frac{4}{5}\sqrt{5} & -\frac{2}{5}\sqrt{5} \end{bmatrix} =$$

:

$$= \begin{bmatrix} 7 & \frac{9}{7} & \frac{5}{7} \\ 0 & \frac{2}{7}\sqrt{41} & -\frac{47}{287}\sqrt{41} \\ 0 & 0 & \frac{4}{41}\sqrt{41} \end{bmatrix}$$

and

$$\begin{aligned}
Q &= G_1(2, 3)G_2(1, 2)G_3(2, 3) = \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{2}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & -\frac{1}{5}\sqrt{5} & \frac{2}{5}\sqrt{5} \end{bmatrix} \begin{bmatrix} \frac{2}{7} & -\frac{3\sqrt{5}}{7} & 0 \\ \frac{3\sqrt{5}}{7} & \frac{2}{7} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{205}\sqrt{205} & -\frac{14}{205}\sqrt{205} \\ 0 & \frac{14}{205}\sqrt{205} & \frac{3}{205}\sqrt{205} \end{bmatrix} = \\
&= \begin{bmatrix} \frac{2}{7} & -\frac{9}{287}\sqrt{41} & \frac{6}{41}\sqrt{41} \\ \frac{6}{7} & \frac{22}{287}\sqrt{41} & -\frac{1}{41}\sqrt{41} \\ -\frac{3}{7} & \frac{38}{287}\sqrt{41} & \frac{2}{41}\sqrt{41} \end{bmatrix}.
\end{aligned}$$

Let us check:

$$\begin{aligned}
QR &= \begin{bmatrix} \frac{2}{7} & -\frac{9}{287}\sqrt{41} & \frac{6}{41}\sqrt{41} \\ \frac{6}{7} & \frac{22}{287}\sqrt{41} & -\frac{1}{41}\sqrt{41} \\ -\frac{3}{7} & \frac{38}{287}\sqrt{41} & \frac{2}{41}\sqrt{41} \end{bmatrix} \begin{bmatrix} 7 & \frac{9}{7} & \frac{5}{7} \\ 0 & \frac{2}{7}\sqrt{41} & -\frac{47}{287}\sqrt{41} \\ 0 & 0 & \frac{4}{41}\sqrt{41} \end{bmatrix} = \\
&= \begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & 1 & -1 \end{bmatrix} = A.
\end{aligned}$$

:

Exercise 2.4.1. Find the Givens QR factorization of the matrix A in example 2.3.2.

Exercise 2.4.2.* Find the Givens QR factorization of A if

$$a) A = \begin{bmatrix} -12 & 1 \\ 4 & 0 \\ 3 & 3 \end{bmatrix}, \quad b) A = \begin{bmatrix} -6 & 2 \\ 8 & 4 \end{bmatrix}, \quad c) A = \begin{bmatrix} 12 & -3 & 1 \\ -3 & 1 & 2 \\ 4 & -\frac{4}{3} & -1 \end{bmatrix}.$$

2.2.5 Main Theorem of QR Factorization

Proposition 2.5.1. If $A = [\mathbf{a}_1 \cdots \mathbf{a}_n] \in \mathbf{R}^{m \times n}$ ($m \geq n$) with linearly independent column vectors $\mathbf{a}_i$ ($i = 1:n$) can be factored into $A = QR$, where $Q = [\mathbf{q}_1 \cdots \mathbf{q}_m] \in \mathbf{R}^{m \times m}$ and $R \in \mathbf{R}^{m \times n}$, then

$$\text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_k\} = \text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_k\} \quad (k = 1:n). \quad (3)$$

In particular, if

$$Q_1 = Q(1 : m, 1 : n), \quad Q_2 = Q(1 : m, n + 1 : m), \quad R_1 = R(1 : n, 1 : n),$$

then

$$\mathcal{R}(A) = \mathcal{R}(Q_1) \tag{4}$$

$$\mathcal{R}(A)^\perp = \mathcal{R}(Q_2) \tag{5}$$

and

$$A = Q_1 R_1, \tag{6}$$

Proof. If $A = QR$, then

$$a_{ik} = \sum_{j=1}^m q_{ij} r_{jk} \stackrel{r_{jk}=0}{\underset{j>k}{=}} \sum_{j=1}^k q_{ij} r_{jk} \quad (i = 1 : m, \quad k = 1 : n)$$

or

$$\mathbf{a}_k = \sum_{j=1}^k r_{jk} \mathbf{q}_j \quad (k = 1 : n).$$

Thus, $\mathbf{a}_k \in \text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ and $\text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_k\} \subset \text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$. Since $\text{rank}(A) = n$, then $\text{rank}(\text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_k\}) = k$, and relation (3) holds. Relation (3) for $k = n$ yields relation (4), and this yields (5). From

$$a_{ik} = \sum_{j=1}^m q_{ij} r_{jk} = \sum_{j=1}^n q_{ij} r_{jk}$$

results assertion (6). $\square$

2.3 Singular Value Decomposition

2.3.1 Existence of Singular Value Decomposition

Proposition 3.1.1. If $V_1 \in \mathbf{R}^{n \times r}$ ($r < n$) has orthonormal columns, then there exists $V_2 \in \mathbf{R}^{n \times (n-r)}$ such that $V = [V_1 \ V_2]$ is orthogonal, where the orthogonal complement $\mathcal{R}(V_1)^\perp$ of the span of column vectors of the matrix V_1 is equal to the span $\mathcal{R}(V_2)$ of the column vectors of the matrix V_2 , i.e., $\mathcal{R}(V_1)^\perp = \mathcal{R}(V_2)$.

Proof is based on the Gram-Schmidt orthogonalization. $\square$

Proposition 3.1.2. If $\mathbf{x} \in \mathbf{R}^n$ and $Q \in \mathbf{R}^{m \times n}$ has orthonormal columns, then $\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2$.

Proof. If $Q \in \mathbf{R}^{m \times n}$ has orthonormal columns, then $Q^T Q = I_n$ and

$$\|Q\mathbf{x}\|_2^2 = (Q\mathbf{x})^T Q\mathbf{x} = \mathbf{x}^T Q^T Q\mathbf{x} = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|_2^2. \quad \square$$

Proposition 3.1.3. Let $A \in \mathbf{R}^{m \times n}$. If $Q \in \mathbf{R}^{m \times m}$ and $Z \in \mathbf{R}^{n \times n}$ are orthogonal, then

$$\|QAZ\|_F = \|A\|_F$$

and

$$\|QAZ\|_2 = \|A\|_2. \quad (1)$$

Prove relation (1):

$$\begin{aligned} \|QAZ\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|QAZ\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|QA(Z\mathbf{x})\|_2 = \\ &= \max_{\|\mathbf{z}\|_2=1} \|QA\mathbf{z}\|_2 = \max_{\|\mathbf{z}\|_2=1} \|Q(A\mathbf{z})\|_2 = \max_{\|\mathbf{z}\|_2=1} \|A\mathbf{z}\|_2 = \|A\|_2. \quad \square \end{aligned}$$

Proposition 3.1.4 (*existence theorem of the singular value decomposition*). If $A \in \mathbf{R}^{m \times n}$, then there exist orthogonal matrices

$$U = [\mathbf{u}_1 \cdots \mathbf{u}_m] \in \mathbf{R}^{m \times m}$$

and

$$V = [\mathbf{v}_1 \cdots \mathbf{v}_n] \in \mathbf{R}^{n \times n},$$

such that

$$U^T A V = \Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbf{R}^{m \times n} \quad (p = \min\{m, n\}) \quad (2)$$

with

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0.$$

Proof. By the definition of the matrix 2-norm there exist vectors $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} \in \mathbf{R}^m$ such that $A\mathbf{x} = \sigma\mathbf{y}$, where $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$ and $\sigma = \|A\|_2$. By Proposition 3.1.1, there exist matrices $V_2 \in \mathbf{R}^{n \times (n-1)}$ and $U_2 \in \mathbf{R}^{m \times (m-1)}$

such that $V = [\mathbf{x} \ V_2]$ and $U = [\mathbf{y} \ U_2]$ are orthogonal. Using this notation, we obtain

$$\begin{aligned} U^T AV &= \begin{bmatrix} \mathbf{y}^T \\ U_2^T \end{bmatrix} A \begin{bmatrix} \mathbf{x} & V_2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}^T \\ U_2^T \end{bmatrix} \begin{bmatrix} A\mathbf{x} & AV_2 \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{y}^T \\ U_2^T \end{bmatrix} \begin{bmatrix} \sigma\mathbf{y} & AV_2 \end{bmatrix} = \begin{bmatrix} \sigma\mathbf{y}^T\mathbf{y} & \mathbf{y}^T AV_2 \\ \sigma U_2^T\mathbf{y} & U_2^T AV_2 \end{bmatrix} = \\ &= \begin{bmatrix} \sigma & \mathbf{w}^T \\ \mathbf{0} & B \end{bmatrix} = A_1 \end{aligned}$$

with $\mathbf{w} = V_2^T A^T \mathbf{y}$ and $B = U_2^T AV_2$. Since

$$A_1 \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \sigma & \mathbf{w}^T \\ \mathbf{0} & B \end{bmatrix} \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \sigma^2 + \mathbf{w}^T \mathbf{w} \\ B\mathbf{w} \end{bmatrix},$$

then

$$\left\| A_1 \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} \right\|_2^2 \geq (\sigma^2 + \mathbf{w}^T \mathbf{w})^2.$$

On the other hand,

$$\left\| A_1 \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} \right\|_2^2 \leq \|A_1\|_2^2 \left\| \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} \right\|_2^2 = \|A_1\|_2^2 (\sigma^2 + \mathbf{w}^T \mathbf{w}),$$

and therefore,

$$\|A_1\|_2^2 \geq \sigma^2 + \mathbf{w}^T \mathbf{w} = \|A\|_2^2 + \mathbf{w}^T \mathbf{w}.$$

By Proposition 3.1.3, we find that $\|A_1\|_2^2 = \|A\|_2^2$. Consequently, $\mathbf{w}^T \mathbf{w} = 0$ and $\mathbf{w} = \mathbf{0}$. We obtain

$$U^T AV = \begin{bmatrix} \sigma & \mathbf{0}^T \\ \mathbf{0} & B \end{bmatrix}$$

or

$$A = U \begin{bmatrix} \sigma & \mathbf{0}^T \\ \mathbf{0} & B \end{bmatrix} V^T$$

and

$$A^T A = V \begin{bmatrix} \sigma & \mathbf{0}^T \\ \mathbf{0} & B^T \end{bmatrix} U^T U \begin{bmatrix} \sigma & \mathbf{0}^T \\ \mathbf{0} & B \end{bmatrix} V^T = V \begin{bmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & B^T B \end{bmatrix} V^T.$$

Thus, the matrices $A^T A$ and $\begin{bmatrix} \sigma^2 & \mathbf{0}^T \\ \mathbf{0} & B^T B \end{bmatrix}$ are similar, and they have the same eigenvalues. Consequently,

$$\lambda(A^T A) = \{\sigma^2\} \cup \lambda(B^T B),$$

where σ^2 as $\|A\|_2^2$ is the greatest eigenvalue of $A^T A$. Note that since $A^T A$ is symmetric then all eigenvalues of $A^T A$ are non-negative. The Reasoning used for the matrix A will be used in the next step for the matrix B etc. So, on the main diagonal of Σ there are the square roots of the eigenvalues of $A^T A$, more exactly, the first $p = \min\{m, n\}$ of them in descending order. $\square$

Definition 3.1.1. The relation in form (2) is called the *singular value decomposition* of the matrix $A \in \mathbf{R}^{m \times n}$. The elements σ_i ($i = 1 : \min\{m, n\}$) on the main diagonal of Σ are called the *singular values* of the matrix A .

2.3.2 Properties of Singular Value Decomposition

Relation (2) yields the relations

$$AV = U\Sigma \tag{3}$$

and

$$A^T U = V\Sigma^T. \tag{4}$$

Proposition 3.2.1. If $A \in \mathbf{R}^{m \times n}$, $A = U\Sigma V^T$, $U = [\mathbf{u}_1 \ \dots \ \mathbf{u}_m] \in \mathbf{R}^{m \times m}$ and $V = [\mathbf{v}_1 \ \dots \ \mathbf{v}_n] \in \mathbf{R}^{n \times n}$, then for each $i = 1 : \min\{m, n\}$ the following holds

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i, \tag{5}$$

$$A^T \mathbf{u}_i = \sigma_i \mathbf{v}_i, \tag{6}$$

$$\|A\|_F = \sigma_1^2 + \dots + \sigma_p^2 \quad (p = \min\{m, n\}),$$

$$\|A\|_2 = \sigma_1$$

and

$$\min_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_n \quad (m \geq n).$$

Proof. Suppose $n > m$. Consider relation (3) which can be written in the form

$$A[\mathbf{v}_1 \cdots \mathbf{v}_n] = [\mathbf{u}_1 \cdots \mathbf{u}_m] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_m & 0 \end{bmatrix}$$

or

$$[A\mathbf{v}_1 \cdots A\mathbf{v}_n] = \begin{bmatrix} \sigma_1 \mathbf{u}_1 & \cdots & \sigma_m \mathbf{u}_m & 0 \end{bmatrix}.$$

The latter is (5) for the elements in the m first columns of the matrix. Consider relation (4) that can be written in the form

$$A^T[\mathbf{u}_1 \cdots \mathbf{u}_m] = [\mathbf{v}_1 \cdots \mathbf{v}_n] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

or

$$[A^T \mathbf{u}_1 \cdots A^T \mathbf{u}_m] = \begin{bmatrix} \sigma_1 \mathbf{v}_1 & \cdots & \sigma_m \mathbf{v}_m \end{bmatrix},$$

which represents relation (6) by elements. We note that "0" denotes also certain blocks consisting of zeros. $\square$

Proposition 3.2.2. If the singular values in the singular value decomposition (2) of $A \in \mathbf{R}^{m \times n}$ satisfy the inequalities

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0,$$

then

1. $\text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\} = \mathcal{R}(A);$
2. $\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_r\} = \mathcal{R}(A^T);$
3. $\text{span}\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\} = \mathcal{N}(A^T);$
4. $\text{span}\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\} = \mathcal{N}(A);$
5. $\text{rank}(A) = r;$

6. the singular values of A are equal to the semi-axes of the hyperellipsoid $E = \{A\mathbf{x} : \|\mathbf{x}\| = 1\}$;
7. $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.

Prove the first of these properties. Consider the relation $A = U\Sigma V^T$. Since

$$[\Sigma V^T]_{jk} = \sum_{s=1}^n \sigma_{js} v_{sk}^T = \begin{cases} \sigma_j v_{kj}, & \text{if } j = 1 : r, \\ 0, & \text{if } j = r+1 : m, \end{cases}$$

then

$$a_{ik} = [U\Sigma V^T]_{ik} = \sum_{j=1}^m u_{ij} [\Sigma V^T]_{jk} = \sum_{j=1}^r u_{ij} \sigma_j v_{kj}$$

or

$$\mathbf{a}_k = \sum_{j=1}^r \sigma_j v_{kj} \mathbf{u}_j.$$

Thus,

$$\mathbf{a}_k \in \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\} \ (k = 1 : n) \Rightarrow \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\} = \mathcal{R}(A). \quad \square$$

Proposition 3.2.3. If $A \in \mathbf{R}^{m \times n}$ and $A = U\Sigma V^T$ is a singular value decomposition of the matrix A , then the column-vectors of $U \in \mathbf{R}^{m \times m}$ are the normed eigenvectors of AA^T and the column-vectors of $V \in \mathbf{R}^{n \times n}$ are the normed eigenvectors of $A^T A$. Singular values of the matrix A can be found as square roots of the eigenvalues of $A^T A$ or AA^T .

Proof. Proceeding from the singular value decomposition of the matrix A we will find expressions of AA^T and $A^T A$:

$$AA^T = U\Sigma V^T V \Sigma^T U^T = U(\Sigma \Sigma^T)U^T \quad (7)$$

and

$$A^T A = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T. \quad (8)$$

Since the matrices $\Sigma \Sigma^T$ and $\Sigma^T \Sigma$ are diagonal matrices, the orthogonal matrices U and V in the expressions (7) and (8) must be formed by the eigenvectors of the matrices AA^T and $A^T A$ respectively. $\square$

2.3.3 Algorithm of Singular Value Decomposition

Algorithm 3.3.1. To find the singular value decomposition of the matrix $A \in \mathbf{R}^{m \times n}$ one has to:

I Find the eigenvalues of the matrix $A^T A$ and arrange them in descending order.

II Find the number of nonzero eigenvalues of the matrix $A^T A$.

III Find the orthogonal eigenvectors of the matrix $A^T A$ corresponding to the eigenvalues, and arrange them in the same order to form the column-vectors of the matrix $V \in \mathbf{R}^{n \times n}$.

IV Form a diagonal matrix $\Sigma \in \mathbf{R}^{m \times n}$ placing on the leading diagonal the square roots $\sigma_i = \sqrt{\lambda_i}$ of $p = \min\{m, n\}$ first eigenvalues of the matrix $A^T A$ obtained in I in descending order.

V Find the first column-vectors of the matrix $U \in \mathbf{R}^{m \times m}$:

$$\mathbf{u}_i = \sigma_i^{-1} A \mathbf{v}_i \quad (i = 1 : r). \quad (9)$$

VI Add to the matrix U the rest of $m - r$ vectors using the Gram-Schmidt orthogonalization process. $\square$

Example 3.3.1. Let us find the singular value decomposition of the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \in R^{3 \times 2}.$$

I Find the eigenvalues of the matrix $A^T A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$:

$$\lambda_1 = 3, \quad \lambda_2 = 1.$$

II Find the number of nonzero eigenvalues of the matrix $A^T A$: $r = 2$.

III Find the orthonormal eigenvectors of the matrix $A^T A$ corresponding to the eigenvalues λ_1 and λ_2 :

$$\mathbf{v}_1 = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix} \text{ and } \mathbf{v}_2 = \begin{bmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{bmatrix} \text{ forming a matrix}$$

$$V = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \in R^{2 \times 2}.$$

IV Find the singular value matrix $\Sigma \in R^{3 \times 2}$:

$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{1} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

on the leading diagonal of which are the square roots of the eigenvalues of the matrix $A^T A$ (in descending order) and the rest of the entries of the matrix Σ are zeros.

V Find the first two column-vectors of the matrix $U \in R^{3 \times 3}$ using the formula (9)

$$\mathbf{u}_1 = \sigma_1^{-1} A \mathbf{v}_1 = \frac{\sqrt{3}}{3} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} \sqrt{6}/3 \\ \sqrt{6}/6 \\ \sqrt{6}/6 \end{bmatrix}$$

and

$$\mathbf{u}_2 = \sigma_2^{-1} A \mathbf{v}_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} 0 \\ -\sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}.$$

VI To find the vector $\mathbf{u}_3$ we shall first find, applying the Gram-Schmidt process, a vector $\hat{\mathbf{u}}_3$ perpendicular to $\mathbf{u}_1$ and $\mathbf{u}_2$:

$$\hat{\mathbf{u}}_3 = \mathbf{e}_1 - (\mathbf{u}_1^T \mathbf{e}_1) \mathbf{u}_1 - (\mathbf{u}_2^T \mathbf{e}_1) \mathbf{u}_2 = \begin{bmatrix} 1/3 & -1/3 & -1/3 \end{bmatrix}^T.$$

Norming the vector $\hat{\mathbf{u}}_3$, we get

$$\mathbf{u}_3 = \begin{bmatrix} \sqrt{3}/3 \\ -\sqrt{3}/3 \\ -\sqrt{3}/3 \end{bmatrix}.$$

Hence

$$U = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{bmatrix} = \begin{bmatrix} \sqrt{6}/3 & 0 & \sqrt{3}/3 \\ \sqrt{6}/6 & \sqrt{2}/2 & -\sqrt{3}/3 \\ \sqrt{6}/6 & -\sqrt{2}/2 & -\sqrt{3}/3 \end{bmatrix}$$

and the singular value decomposition of the matrix A is

$$A = \begin{bmatrix} \sqrt{6}/3 & 0 & \sqrt{3}/3 \\ \sqrt{6}/6 & -\sqrt{2}/2 & -\sqrt{3}/3 \\ \sqrt{6}/6 & \sqrt{2}/2 & -\sqrt{3}/3 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}.$$

Example 3.3.2. Let us find the singular value decomposition of the matrix $A = \begin{bmatrix} 2 & 1 & -2 \end{bmatrix}$.

I Find the eigenvalues of the matrix $A^T A$:

$$\det(A^T A - \lambda I) = 0 \Leftrightarrow \begin{vmatrix} 4 - \lambda & 2 & -4 \\ 2 & 1 - \lambda & -2 \\ -4 & -2 & 4 - \lambda \end{vmatrix} = 0 \Rightarrow \begin{cases} \lambda_1 = 9, \\ \lambda_2 = 0, \\ \lambda_3 = 0. \end{cases}$$

II Find the number of the nonzero eigenvalues of the matrix $A^T A$: $r = 1$.

III Find the eigenvector of the matrix $A^T A$:

$$\lambda_1 = 9 \Rightarrow \mathbf{v}_1 = \begin{bmatrix} -2/3 & -1/3 & 2/3 \end{bmatrix}^T,$$

$$\lambda_{2,3} = 0 \Rightarrow \begin{cases} \mathbf{v}_2 = \begin{bmatrix} -\sqrt{5}/5 & 2\sqrt{5}/5 & 0 \end{bmatrix}^T, \\ \mathbf{v}_3 = \begin{bmatrix} 4\sqrt{5}/15 & 2\sqrt{5}/15 & 5\sqrt{5}/15 \end{bmatrix}^T. \end{cases}$$

Since the eigenvalue 0 is multiple, the Gram-Schmidt orthogonalization process is used to find the vector $\mathbf{v}_3$. We compile the orthonormal matrix V :

$$V = \begin{bmatrix} -2/3 & -\sqrt{5}/5 & 4\sqrt{5}/15 \\ -1/3 & 2\sqrt{5}/5 & 2\sqrt{5}/15 \\ 2/3 & 0 & 5\sqrt{5}/15 \end{bmatrix}.$$

IV Form the singular value matrix:

$$\Sigma = \begin{bmatrix} 3 & 0 & 0 \end{bmatrix}.$$

V Calculate the unique column-vector of the matrix U applying the formula (9):

$$\mathbf{u}_1 = \frac{1}{3} A \mathbf{v}_1 = \frac{1}{3} \begin{bmatrix} 2 & 1 & -2 \end{bmatrix} \begin{bmatrix} -2/3 & -1/3 & 2/3 \end{bmatrix}^T = \begin{bmatrix} -1 \end{bmatrix}.$$

Thus the singular value decomposition of the matrix A is

$$A = U \Sigma V^T = \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} -2/3 & -1/3 & 2/3 \\ \sqrt{5}/5 & 2\sqrt{5}/5 & 0 \\ 4\sqrt{5}/15 & 2\sqrt{5}/15 & 5\sqrt{5}/15 \end{bmatrix}.$$

Example 3.3.3.* Let us find the singular value decomposition of the matrix

$$A = \begin{bmatrix} 2 & 2 & 2 & 2 \\ \frac{17}{10} & \frac{1}{10} & -\frac{17}{10} & -\frac{1}{10} \\ \frac{3}{5} & \frac{9}{5} & -\frac{3}{5} & -\frac{9}{5} \end{bmatrix}.$$

The given 3×4 matrix A has three nonzero singular values. Therefore it is enough to find nonzero singular values of the matrix A using the 3×3 matrix AA^T (not the 4×4 matrix $A^T A$). Since

$$AA^T = \begin{bmatrix} 2 & 2 & 2 & 2 \\ \frac{17}{10} & \frac{1}{10} & -\frac{17}{10} & -\frac{1}{10} \\ \frac{3}{5} & \frac{9}{5} & -\frac{3}{5} & -\frac{9}{5} \end{bmatrix} \begin{bmatrix} 2 & \frac{17}{10} & \frac{3}{5} \\ 2 & \frac{1}{10} & -\frac{3}{5} \\ 2 & -\frac{17}{10} & -\frac{3}{5} \\ 2 & -\frac{1}{10} & -\frac{9}{5} \end{bmatrix} = \begin{bmatrix} 16 & 0 & 0 \\ 0 & \frac{29}{5} & \frac{12}{5} \\ 0 & \frac{12}{5} & \frac{36}{5} \end{bmatrix},$$

then the characteristic equation of AA^T is

$$\begin{vmatrix} 16 - \lambda & 0 & 0 \\ 0 & \frac{29}{5} - \lambda & \frac{12}{5} \\ 0 & \frac{12}{5} & \frac{36}{5} - \lambda \end{vmatrix} = 0$$

or

$$(16 - \lambda)(36 - 13\lambda + \lambda^2) = 0,$$

and the solutions of this equation are $\lambda_1 = 16$, $\lambda_2 = 9$ and $\lambda_3 = 4$. Since $\lambda_i = \sigma_i^2$ and the matrix Σ is a 3×4 matrix, then on the leading diagonal of the matrix Σ there are the singular values of the matrix A in descending order, and all other elements of the matrix Σ are zeros:

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix}.$$

The matrix U has for column-vectors the orthonormed eigenvectors of the matrix AA^T :

$$\lambda_1 = 16 \Rightarrow \mathbf{u}_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T;$$

$$\lambda_2 = 9 \Rightarrow \mathbf{u}_2 = \begin{bmatrix} 0 & \frac{3}{5} & \frac{4}{5} \end{bmatrix}^T;$$

$$\lambda_3 = 4 \Rightarrow \mathbf{u}_3 = \begin{bmatrix} 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix}^T.$$

Collecting the vectors $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$, we obtain the matrix

$$U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{5} & -\frac{4}{5} \\ 0 & \frac{4}{5} & \frac{3}{5} \end{bmatrix}.$$

According to the relation (6), we shall find the first three column-vectors of the matrix V (the matrix Σ has three nonzero entries on its leading diagonal) using the formula

$$\mathbf{v}_i = \frac{1}{\sigma_i} A^T \mathbf{u}_i.$$

Hence

$$\mathbf{v}_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}.$$

To calculate the vector $\mathbf{v}_4$, we find first, using the Gram-Schmidt orthogonalization process, the vector $\hat{\mathbf{v}}_4$ perpendicular to the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$:

$$\begin{aligned} \hat{\mathbf{v}}_4 &= \mathbf{e}_1 - (\mathbf{v}_1^T \mathbf{e}_1) \mathbf{v}_1 - (\mathbf{v}_2^T \mathbf{e}_1) \mathbf{v}_2 - (\mathbf{v}_3^T \mathbf{e}_1) \mathbf{v}_3 = \\ &= \mathbf{e}_1 - \frac{1}{2} \mathbf{v}_1 - \frac{1}{2} \mathbf{v}_2 + \frac{1}{2} \mathbf{v}_3 = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \end{bmatrix}^T. \end{aligned}$$

Since $\|\hat{\mathbf{v}}_4\|_2 = \frac{1}{2}$, then

$$\mathbf{v}_4 = 2\hat{\mathbf{v}}_4 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$$

and

$$V = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

Let us check the result:

$$U\Sigma V^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{5} & -\frac{4}{5} \\ 0 & \frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} =$$

$$= \begin{bmatrix} 2 & 2 & 2 & 2 \\ \frac{17}{10} & \frac{1}{10} & -\frac{17}{10} & -\frac{1}{10} \\ \frac{3}{5} & \frac{9}{5} & -\frac{3}{5} & -\frac{9}{5} \end{bmatrix} = A$$

and

$$U^T AV = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{5} & \frac{4}{5} \\ 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} 2 & 2 & 2 & 2 \\ \frac{17}{10} & \frac{1}{10} & -\frac{17}{10} & -\frac{1}{10} \\ \frac{3}{5} & \frac{9}{5} & -\frac{3}{5} & -\frac{9}{5} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} =$$

:

$$= \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix} = \Sigma.$$

Problem 3.3.1.* Applying the singular value decomposition of the matrix A obtained in example 3.3.3, find the bases of the subspace of the column-vectors $\mathcal{R}(A)$, the right null space $\mathcal{N}(A)$, the subspace of the row-vectors $\mathcal{R}(A^T)$, and the left null space $\mathcal{N}(A^T)$ of the matrix A .

Problem 3.3.2.* Find the singular value decomposition and the QR factorization of the matrix $A = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$.

Problem 3.3.3.* Find the singular value decomposition of the matrix $A = \begin{bmatrix} -\frac{5}{2} + 3\sqrt{3} & \frac{5}{2}\sqrt{3} + 3 \end{bmatrix}$.

2.4 Pseudoinverse Matrix

2.4.1 Least-Squares Method

Let us consider the solution of a system of linear equations

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

by the least-squares method in the case where the condition of the Kronecker-Capelli theorem is not satisfied, i.e., the system has no solution in an ordinary sense.

Example 4.1.1. Let the system be

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix},$$

where $\mathbf{b} = [\beta_1 \ \beta_2 \ \beta_3]^T \notin \mathcal{R}(A)$ and $\text{rank}(A) = 2$. Let $\mathbf{p}$ be the orthogonal projection of the vector $\mathbf{b}$ onto the space $\mathcal{R}(A)$. Since the vector $\mathbf{p} \in \mathcal{R}(A)$ and $\text{rank}(A) = 2$, the system $A\bar{\mathbf{x}} = \mathbf{p}$ has a unique solution. Taking into consideration that $\mathbf{R}^3 = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$, we get $\mathbf{b} - \mathbf{p} \in \mathcal{N}(A^T) \Leftrightarrow A^T(\mathbf{b} - \mathbf{p}) = \mathbf{0}$ and $A^T(\mathbf{b} - A\bar{\mathbf{x}}) = \mathbf{0}$ or

$$A^T A \bar{\mathbf{x}} = A^T \mathbf{b}. \quad (2)$$

The matrix $A^T A$ of the system (2) is regular since $\text{rank}(A) = 2$. Therefore the system (2) is uniquely solvable on the given conditions and

$$\bar{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}. \quad (3)$$

By minimizing the square of the norm of discrepancy $A\mathbf{x} - \mathbf{b}$

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) = (\mathbf{x}^T A^T - \mathbf{b}^T)^T (A\mathbf{x} - \mathbf{b})$$

($\text{grad} \|A\mathbf{x} - \mathbf{b}\|_2^2 = 0$), we obtain the same system (2), and hence the same solution $\bar{\mathbf{x}}$ determined by the formula (3), the *least-square solution* of the equation (1).

The line of reasoning given in example 4.1.1 can be realized also in a more general case.

Definition 4.1.1. If $A \in \mathbf{R}^{m \times n}$, then system (2) is called the *system of normal equations* of system (1).

Proposition 4.1.1. If $A \in \mathbf{R}^{m \times n}$, $\mathbf{b} \notin \mathcal{R}(A)$ and suppose $\text{rank}(A) = n$, then the system of normal equations (2) of system (1) is uniquely solvable and the least-squares solution $\bar{\mathbf{x}}$ of the system (1) is given by (3).

Example 4.1.2.* Let us solve by the least-squares method the system of equations

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

We form the system of normal equations $A^T A \bar{\mathbf{x}} = A^T \mathbf{b}$:

$$\begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} \bar{\mathbf{x}} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 9 & 9 \\ 9 & 11 \end{bmatrix} \bar{\mathbf{x}} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}.$$

Thus,

$$\begin{cases} 9\bar{\xi}_1 + 9\bar{\xi}_2 = 5 \\ 9\bar{\xi}_1 + 11\bar{\xi}_2 = 5 \end{cases} \Rightarrow \begin{cases} \bar{\xi}_1 = \frac{5}{9} \\ \bar{\xi}_2 = 0 \end{cases}.$$

If $A \in \mathbf{R}^{m \times n}$, $\mathbf{b} \notin \mathcal{R}(A)$ and $\text{rank}(A) < n$, then the system of normal equations (2) has an infinite number of solutions, which can be all expressed as

$$\bar{\mathbf{x}} = \bar{\mathbf{x}}_r + \bar{\mathbf{x}}_n,$$

where $\bar{\mathbf{x}}_r \in \mathcal{R}(A^T)$ and $\bar{\mathbf{x}}_n \in \mathcal{N}(A)$. From among the solutions $\bar{\mathbf{x}}$ we will find the one having the least norm, the so-called optimum solution $\mathbf{x}^+$. From the orthogonality of the vectors $\bar{\mathbf{x}}_r$ and $\bar{\mathbf{x}}_n$ it follows that

$$\|\bar{\mathbf{x}}\|_2^2 = \|\bar{\mathbf{x}}_r\|_2^2 + \|\bar{\mathbf{x}}_n\|_2^2.$$

Since from $\bar{\mathbf{x}}_n \in \mathcal{N}(A)$ it follows $A\bar{\mathbf{x}}_n = 0$, then

$$A\bar{\mathbf{x}} = \mathbf{p} \Leftrightarrow A(\bar{\mathbf{x}}_r + \bar{\mathbf{x}}_n) = \mathbf{p} \Leftrightarrow A\bar{\mathbf{x}}_r + A\bar{\mathbf{x}}_n = \mathbf{p} \Rightarrow A\bar{\mathbf{x}}_r = \mathbf{p}$$

and $\bar{\mathbf{x}}_r \in \mathcal{R}(A^T)$ is the optimum solution $\mathbf{x}^+$ of the equation $A\bar{\mathbf{x}} = \mathbf{p}$. Thus, $\mathbf{x}^+ = \bar{\mathbf{x}}$. $\square$

2.4.2 Pseudoinverse Matrix and Optimum Solution

Next we will consider the algorithm for finding the optimum solution.

Example 4.2.1. Let $\mathbf{b} = [\beta_1 \ \beta_2 \ \beta_3]^T$ and

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where $\sigma_1 \neq 0$ and $\sigma_2 \neq 0$. We will find the optimum solution of the system

$$\Sigma \mathbf{x} = \mathbf{b}.$$

The orthogonal projection of the vector $\mathbf{b}$ on the space $\mathcal{R}(\Sigma)$ is $\mathbf{p} = \begin{bmatrix} \beta_1 & \beta_2 & 0 \end{bmatrix}^T$, and $\mathbf{b} - \mathbf{p} = \begin{bmatrix} 0 & 0 & \beta_3 \end{bmatrix}^T$. To find the solution $\bar{\mathbf{x}}$, one must solve the system

$$\Sigma \bar{\mathbf{x}} = \mathbf{p},$$

i.e.,

$$\begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{\xi}_1 \\ \bar{\xi}_2 \\ \bar{\xi}_3 \\ \bar{\xi}_4 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ 0 \end{bmatrix}$$

or

$$\begin{cases} \sigma_1 \bar{\xi}_1 + 0 \bar{\xi}_2 + 0 \bar{\xi}_3 + 0 \bar{\xi}_4 = \beta_1 \\ 0 \bar{\xi}_1 + \sigma_2 \bar{\xi}_2 + 0 \bar{\xi}_3 + 0 \bar{\xi}_4 = \beta_2 \\ 0 \bar{\xi}_1 + 0 \bar{\xi}_2 + 0 \bar{\xi}_3 + 0 \bar{\xi}_4 = 0 \end{cases} \Rightarrow \begin{cases} \bar{\xi}_1 = \beta_1 / \sigma_1 \\ \bar{\xi}_2 = \beta_2 / \sigma_2 \\ \bar{\xi}_3 = \gamma \\ \bar{\xi}_4 = \delta \end{cases},$$

where $\gamma, \delta \in \mathbf{R}$ are arbitrary. Taking $\gamma = \delta = 0$, we obtain the solution with the least 2-norm

$$\mathbf{x}^+ = \begin{bmatrix} \beta_1 / \sigma_1 & \beta_2 / \sigma_2 & 0 & 0 \end{bmatrix}^T.$$

We state that $\mathbf{x}^+$ can be expressed also by

$$\mathbf{x}^+ = \begin{bmatrix} \beta_1 / \sigma_1 \\ \beta_2 / \sigma_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 / \sigma_1 & 0 & 0 \\ 0 & 1 / \sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

The optimum solution $\mathbf{x}^+$ of the given example can be obtained from the vector $\mathbf{b}$ by multiplying it on the left by the matrix

$$\Sigma^+ = \begin{bmatrix} 1 / \sigma_1 & 0 & 0 \\ 0 & 1 / \sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The matrix Σ^+ is obtained from the matrix Σ by transposing and afterwards replacing the nonzero entries by their reciprocals. Hence $\mathbf{x}^+ = A^+ \mathbf{b}$.

Let us generalize the result obtained in example 4.2.1.

Proposition 4.2.1. If

$$\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbf{R}^{m \times n} \quad (p = \min\{m, n\}) \quad (1)$$

and

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_p, \quad (2)$$

then the optimum solution $\mathbf{x}^+$ of the system

$$\Sigma \mathbf{x} = \mathbf{b}$$

is given by

$$\mathbf{x}^+ = \Sigma^+ \mathbf{b},$$

where

$$\Sigma^+ = \text{diag}(1/\sigma_1, \dots, 1/\sigma_r, 0, \dots, 0) \in \mathbf{R}^{n \times m}. \quad (3)$$

Definition 4.2.1. Let

$$A = U \Sigma V^T$$

be the singular value decomposition of the matrix $A \in \mathbf{R}^{m \times n}$. The *pseudoinverse matrix* of the matrix A is a matrix

$$A^+ = V \Sigma^+ U^T,$$

where Σ and Σ^+ are given by relations (1-3).

Problem 4.2.1. Let $A \in \mathbf{R}^{n \times n}$ and $\det(A) \neq 0$. Show that $A^+ = A^{-1}$.

Problem 4.2.2. Let us find the pseudoinverse matrix of the matrix $A = \begin{bmatrix} 2 & 1 & -2 \end{bmatrix}$ given in example 3.3.2. We found the singular value decomposition of the matrix A in this example

$$A = U \Sigma V^T = \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} -2/3 & -1/3 & 2/3 \\ \sqrt{5}/5 & 2\sqrt{5}/5 & 0 \\ 4\sqrt{5}/15 & 2\sqrt{5}/15 & 5\sqrt{5}/15 \end{bmatrix}.$$

Using definition 4.2.1,

$$A^+ = V \Sigma^+ U^T,$$

i.e.,

$$A^+ = \begin{bmatrix} -2/3 & \sqrt{5}/5 & 4\sqrt{5}/15 \\ -1/3 & 2\sqrt{5}/5 & 2\sqrt{5}/15 \\ 2/3 & 0 & 5\sqrt{5}/15 \end{bmatrix} \begin{bmatrix} 1/3 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} = \begin{bmatrix} 2/9 \\ -1/9 \\ -2/9 \end{bmatrix}.$$

Proposition 4.2.2. If $A \in \mathbf{R}^{m \times n}$, then the optimum solution $\mathbf{x}^+$ of the system $A\mathbf{x} = \mathbf{p}$ (in the sense of least-squares) is given by

$$\mathbf{x}^+ = A^+ \mathbf{b}.$$

Proof. When a vector is multiplied by the orthogonal matrix U^T , its 2-norm is conserved. Therefore,

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|U\Sigma V^T \mathbf{x} - \mathbf{b}\|_2 = \|\Sigma V^T \mathbf{x} - U^T \mathbf{b}\|_2.$$

Let substitute $\mathbf{y} = V^T \mathbf{x}$. Hence

$$\min_{\mathbf{x} \in \mathbf{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{y} \in \mathbf{R}^n} \|\Sigma \mathbf{y} - U^T \mathbf{b}\|_2.$$

Proposition 4.2.1 implies that the minimizing vector for the expression $\|\Sigma \mathbf{y} - U^T \mathbf{b}\|_2$ is the vector

$$\mathbf{y}^+ = \Sigma^+ U^T \mathbf{b}$$

and the vector

$$\mathbf{x}^+ = V \mathbf{y}^+ = V \Sigma^+ U^T \mathbf{b} = A^+ \mathbf{b}$$

minimizes the expression $\|A\mathbf{x} - \mathbf{b}\|_2$. $\square$

Example 4.2.3. Let us find the optimum solution of the system

$$2\xi_1 + \xi_2 - 2\xi_3 = 9.$$

In example 4.2.2, we found the pseudoinverse matrix

$$A^+ = \begin{bmatrix} 2/9 \\ -1/9 \\ -2/9 \end{bmatrix}$$

of the matrix of the system $A = \begin{bmatrix} 2 & 1 & -2 \end{bmatrix}$.

In virtue of proposition 4.2.2, we get the optimum solution

$$\mathbf{x}^+ = A^+ \mathbf{b} = \begin{bmatrix} 2/9 \\ -1/9 \\ -2/9 \end{bmatrix} \begin{bmatrix} 9 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}.$$

Example 4.2.4. Let us find the optimum solution of the system

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

In example 3.3.1, we found the singular value decomposition of the system matrix A

$$A = \begin{bmatrix} \sqrt{6}/3 & 0 & \sqrt{3}/3 \\ \sqrt{6}/6 & \sqrt{2}/2 & -\sqrt{3}/3 \\ \sqrt{6}/6 & -\sqrt{2}/2 & -\sqrt{3}/3 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}.$$

Using definition 4.2.1, we will find the pseudoinverse matrix

$$\begin{aligned} A^+ &= \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{6}/3 & \sqrt{6}/6 & \sqrt{6}/6 \\ 0 & \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{3}/3 & -\sqrt{3}/3 & -\sqrt{3}/3 \end{bmatrix} = \\ &= \begin{bmatrix} 1/3 & 2/3 & -1/3 \\ 1/3 & -1/3 & 2/3 \end{bmatrix}. \end{aligned}$$

The optimum solution of the system will be

$$\mathbf{x}^+ = A^+ \mathbf{b} = \begin{bmatrix} 1/3 & 2/3 & -1/3 \\ 1/3 & -1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 5/3 \end{bmatrix}.$$

Problem 4.2.2.* Find the pseudoinverse of the matrix $A = [0]$ and explain the result. *Answer:* $A^+ = [0]$.

Problem 4.2.3.* Find the pseudoinverse of the matrix A

$$a) A = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad b) A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix}.$$

Problem 4.2.4.* What is the pseudoinverse matrix of the matrix A with orthogonal columns? *Answer:* $A^+ = A^T$.

Problem 4.2.5.* Find the optimum solution of the system

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Proposition 4.2.3 (*Conditions of Moore-Penrose.*) If $A \in \mathbf{R}^{m \times n}$, then the conditions

$$AXA = A, \quad XAX = X, \quad (AX)^T = AX, \quad (XA)^T = XA$$

are satisfied only by one matrix $X \in \mathbf{R}^{n \times m}$, and this is A^+ .

Problem 4.2.6.* A matrix A is called a *projectionmatrix* if

$$A^2 = A \wedge A^T = A.$$

Check the Moore-Penrose conditions for the projectionmatrix. Does $A^+ = A$?

2.5 Jordan Form of a Matrix

2.5.1 Matrix Diagonalization

In proposition 1.2.6.8 on the Jordan decomposition it is stated that if $A \in \mathbf{C}^{n \times n}$, then there exists such a regular $X \in \mathbf{C}^{n \times n}$ that

$$X^{-1}AX = J = \text{diag}(J_1, \dots, J_t), \quad (1)$$

where $m_1 + \dots + m_t = n$ and

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_i \end{bmatrix} \in \mathbf{C}^{m_i \times m_i}$$

is a *Jordan block* or *Jordan box*, and the matrix J is called a *Jordan canonical form* or *Jordan normal form* of the matrix A . The number of Jordan blocks in decomposition (1) equals the number of the linearly independent eigenvectors of the matrix A . Namely, to each linearly independent eigenvector corresponds one block. Hence if the matrix A has a basis of eigenvectors, then all the Jordan blocks are 1×1 blocks, and the Jordan normal form coincides with the diagonal form of the matrix given in proposition 1.2.5.8 $S^{-1}AS = \Lambda$, where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ and the matrix S has for columns the linearly independent eigenvectors of the matrix A corresponding to these eigenvalues.

Example 5.1.1.* Let us find the Jordan form of the matrix

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

We shall find the eigenvalues of the matrix A :

$$\det(A - \lambda I) = 0 \Leftrightarrow (\lambda^2 - 1)(\lambda - 2) \Rightarrow \begin{cases} \lambda_1 = 1, \\ \lambda_2 = -1, \\ \lambda_3 = 2. \end{cases}$$

Now we shall find the eigenvectors corresponding to these eigenvalues:

$$\begin{aligned} \lambda_1 = 1 &\rightarrow \begin{bmatrix} 3-1 & -1 & 2 & \vdots & 0 \\ 1 & -1-1 & 1 & \vdots & 0 \\ -1 & 1 & 0-1 & \vdots & 0 \end{bmatrix} \xrightarrow{I \leftrightarrow II} \\ &\sim \begin{bmatrix} 1 & -2 & 1 & \vdots & 0 \\ 2 & -1 & 2 & \vdots & 0 \\ -1 & 1 & -1 & \vdots & 0 \end{bmatrix} \xrightarrow[\text{III}+I]{\text{II}-2\cdot I} \begin{bmatrix} 1 & -2 & 1 & \vdots & 0 \\ 0 & 3 & 0 & \vdots & 0 \\ 0 & -1 & 0 & \vdots & 0 \end{bmatrix} \Rightarrow \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}; \\ \lambda_2 = -1 &\rightarrow \mathbf{x}_2 = \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix}; \quad \lambda_3 = 2 \rightarrow \mathbf{x}_3 = \begin{bmatrix} 5 \\ 1 \\ -2 \end{bmatrix}. \end{aligned}$$

We compile the matrix of eigenvectors of the matrix A

$$S = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3] = \begin{bmatrix} 1 & -1 & 5 \\ 0 & -2 & 1 \\ -1 & 1 & -2 \end{bmatrix}$$

and find the inverse matrix

$$S^{-1} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & -\frac{3}{2} \\ \frac{1}{6} & -\frac{1}{2} & \frac{1}{6} \\ \frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix}.$$

As the result, we obtain

$$\begin{aligned} S^{-1}AS &= \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & -\frac{3}{2} \\ \frac{1}{6} & -\frac{1}{2} & \frac{1}{6} \\ \frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & -1 & 2 \\ 1 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 5 \\ 0 & -2 & 1 \\ -1 & 1 & -2 \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \Lambda. \end{aligned}$$

Proposition 5.1.1. Any Hermitian (symmetric) matrix $A \in \mathbf{C}^{n \times n}$ ($A \in \mathbf{R}^{n \times n}$) can be diagonalized using a unitary matrix $U \in \mathbf{C}^{n \times n}$ (an orthogonal matrix $Q \in \mathbf{R}^{n \times n}$), i.e., there exists such $U \in \mathbf{C}^{n \times n}$ ($Q \in \mathbf{R}^{n \times n}$), that

$$U^H A U = \Lambda \quad (Q^T A Q = \Lambda). \quad (2)$$

Proof. The Schur factorisation (proposition 1.2.6.5) implies that the Hermitian matrix $A \in \mathbf{C}^{n \times n}$ can be given in the form

$$U^H A U = T, \quad (3)$$

where $U \in \mathbf{C}^{n \times n}$ is a unitary matrix and $T \in \mathbf{C}^{n \times n}$ is an upper triangular matrix. Finding the transpose conjugate matrices of both sides of (3), we get

$$U^H A^H U = T^H.$$

In virtue of the Hermitian matrix definition $A^H = A$, we find that

$$U^H A U = T^H. \quad (4)$$

From (3) and (4) it follows that $T = D$. The diagonal elements of the diagonal matrix D similar to the matrix A are the eigenvalues of the matrix A . The assertion about the symmetric matrix $A \in \mathbf{R}^{n \times n}$ is a special case of the complex version. $\square$

Problem 5.1.1.* Let

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

Find such an orthogonal matrix $Q \in \mathbf{R}^{4 \times 4}$, that $Q^T A Q = \Lambda$, where Λ is a diagonal matrix.

Problem 5.1.2.* Let

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

Find such a unitary matrix $U \in \mathbf{C}^{n \times n}$, that $U^H A U = \Lambda$, where Λ is a diagonal matrix.

Not every square matrix can be put in form (2). Proposition 1.2.6.6 implies that only a normal matrix A ($A^H A = A A^H$) can be expressed in form (2). In the general case of the diagonalization of a matrix one must confine himself to the Jordan normal form (1).

2.5.2 Analysis of Jordan Form of a Matrix

It is not sufficient to find the eigenvalues of the matrix to obtain the Jordan form of this matrix.

Example 5.2.1. Let us find the Jordan matrices J of the matrices

$$T = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, A = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

It is easy to find out that the spectra of T, A, B and I are the same, $\lambda(T) = \lambda(A) = \lambda(B) = \lambda(I) = \{1; 1\}$. Let us find the eigenvectors corresponding to $\lambda = 1$:

$$T \rightarrow \begin{bmatrix} 0 & 2 & : 0 \\ 0 & 0 & : 0 \end{bmatrix} \rightarrow \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix};$$

$$\begin{aligned}
A &\rightarrow \begin{bmatrix} 1 & -1 & \vdots & 0 \\ 1 & -1 & \vdots & 0 \end{bmatrix} \rightarrow \mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \\
B &\rightarrow \begin{bmatrix} 0 & 0 & \vdots & 0 \\ 1 & 0 & \vdots & 0 \end{bmatrix} \rightarrow \mathbf{x}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \\
I &\rightarrow \begin{bmatrix} 0 & 0 & \vdots & 0 \\ 0 & 0 & \vdots & 0 \end{bmatrix} \rightarrow \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \wedge \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.
\end{aligned}$$

We see that the matrices T , A and B have only one independent eigenvector and only one Jordan block corresponding to the eigenvalue $\lambda = 1$, and thus the matrices T , A and B have the same Jordan matrix

$$J = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The matrix I has two linearly independent eigenvectors and, consequently, two Jordan blocks, and the corresponding Jordan matrix coincides with the matrix I .

Problem 5.2.1. Verify that the matrix

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbf{R}^{n \times n}$$

corresponds the one-block Jordan matrix

$$\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \ddots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbf{R}^{n \times n}.$$

Example 5.2.2. Let us consider the Jordan matrix

$$J = \begin{bmatrix} 3 & 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} J_1 & & \\ & J_2 & \\ & & J_3 \end{bmatrix}. \quad (5)$$

Let us find the eigenvectors corresponding to the eigenvalue $\lambda = 3$ of multiplicity 2:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & -3 & 1 & 0 & \vdots & 0 \\ 0 & 0 & 0 & -3 & 0 & \vdots & 0 \\ 0 & 0 & 0 & 0 & -3 & \vdots & 0 \end{bmatrix} \Rightarrow \mathbf{x} = \begin{bmatrix} p \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Therefore, one linearly independent eigenvector $\mathbf{e}_1$ and one Jordan block corresponds to the eigenvalue $\lambda = 3$:

$$\begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}.$$

Let us find the eigenvectors corresponding to the eigenvalue $\lambda = 0$ of multiplicity 3:

$$\begin{bmatrix} 3 & 1 & 0 & 0 & 0 & \vdots & 0 \\ 0 & 3 & 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 \end{bmatrix} \Rightarrow \mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ q \\ 0 \\ r \end{bmatrix}.$$

Hence two linearly independent eigenvectors $\mathbf{e}_3$ and $\mathbf{e}_5$ and two Jordan blocks correspond to the eigenvalue $\lambda = 0$

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

and

$$[0].$$

The question arises, what conditions must the 5×5 matrix A satisfy to have for the corresponding Jordan matrix the J given by (5)? How do we find the regular matrix X such that

$$X^{-1}AX = J? \tag{6}$$

The first condition is $\lambda(A) = \lambda(J)$, but it is not sufficient. The eigenvalues of the matrix A must be also considered. We express the relation (6) in the form $AX = XJ$ or

$$A \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_5 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_5 \end{bmatrix} \begin{bmatrix} 3 & 1 & & & \\ & 3 & & & \\ & & 0 & 1 & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}.$$

Having multiplied the matrices, we get the formulas

$$A\mathbf{x}_1 = 3\mathbf{x}_1, \quad A\mathbf{x}_2 = 3\mathbf{x}_2 + \mathbf{x}_1 \quad (7)$$

and

$$A\mathbf{x}_3 = 0\mathbf{x}_3, \quad A\mathbf{x}_4 = 0\mathbf{x}_4 + \mathbf{x}_3, \quad A\mathbf{x}_5 = 0\mathbf{x}_5. \quad (8)$$

From the formulas (7) and (8) it follows that similarly to the matrix J the matrix A must have three eigenvectors $\mathbf{x}_1$, $\mathbf{x}_3$ and $\mathbf{x}_5$. In addition, the matrix A must have two *generalized* eigenvectors or two *first order flag vectors* $\mathbf{x}_2$ and $\mathbf{x}_4$. It is said that the vector $\mathbf{x}_2$ belongs to the chain that begins with the vector $\mathbf{x}_1$ and is defined by the formula (7). This chain determines the Jordan block J_1 . The two first formula of (8) define the second chain consisting of the vectors $\mathbf{x}_3$ and $\mathbf{x}_4$, and this chain, in its turn, defines the Jordan block J_2 . The last of the formulas (8) defines the third chain consisting of the vector $\mathbf{x}_5$, and this chain, in its turn, defines the Jordan block J_3 .

Proposition 5.2.1. The determination of the Jordan form of the matrix $A \in \mathbf{C}^{n \times n}$ reduces to the finding of chains. Every chain starts on the eigenvector of the matrix A and for every value of the index $i = 1 : n$

$$A\mathbf{x}_i = \lambda_i\mathbf{x}_i \vee A\mathbf{x}_i = \lambda_i\mathbf{x}_i + \mathbf{x}_{i-1}. \quad (9)$$

The vectors $\mathbf{x}_i$ are the column vectors of the matrix X , and every chain determines one Jordan block.

2.5.3 Algorithm of Filipov

If $n = 1$, then the Jordan block coincides with the given matrix and formula (9) is true. Let us suppose that the Jordan form of the matrix A is found by applying the Jordan block construction formula (9) if the order of the matrix A is smaller than n . we will use mathematical induction.

I step. Assume that A is singular, $\dim \mathcal{R}(A) = r < n$. Considering the corresponding $r \times r$ matrix we find that in this case the construction based on formulas (9) is realizable. Namely, in the space $\mathcal{R}(A)$ there are r independent vectors $\mathbf{w}_i$ such that the following relations hold

$$A\mathbf{w}_i = \lambda_i \mathbf{w}_i \vee A\mathbf{w}_i = \lambda_i \mathbf{w}_i + \mathbf{w}_{i-1}. \quad (10)$$

II step. Let us suppose that $\dim \mathcal{R}(A) \cap \mathcal{N}(A) = p$. Every vector of the null space $\mathcal{N}(A)$ is an eigenvector of the matrix A corresponding to the eigenvalue of the matrix A $\lambda = 0$. Therefore, there must be p chains on the I step which begin with the eigenvectors corresponding to the eigenvalue 0. We are interested in the last vector of each such chain. Since the vectors $\mathbf{w}_i$ belonging to the subspace $\mathcal{R}(A) \cap \mathcal{N}(A)$ must also belong to the space $\mathcal{R}(A)$, then they have to be the linear combinations of the column vectors of the matrix A

$$\mathbf{w}_i = A\mathbf{y}_i$$

with some $\mathbf{y}_i$. Therefore, the vector $\mathbf{y}_i$ follows the vector $\mathbf{w}_i$ in the chain corresponding to the eigenvalue $\lambda = 0$.

III step. Since $\dim \mathcal{N}(A) = n - p$, there must be $n - r - p$ more linearly independent vectors $\mathbf{z}_i$ of the space $\mathcal{N}(A)$ in the orthogonal complement of the subspace $\mathcal{R}(A) \cap \mathcal{N}(A)$.

Proposition 5.3.1. The algorithm of Filipov defines r vectors $\mathbf{w}_i$, p vectors $\mathbf{y}_i$ and $n - r - p$ vectors $\mathbf{z}_i$, which determine the Jordan chains. These vectors are linearly independent, they can be chosen for the column-vectors of the matrix X , and $J = X^{-1}AX$.

Proof. See Strang (1988, p. 457). $\square$

Example 5.3.1. Let us find the Jordan normal form of the matrix

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

using the algorithm of Filipov.

I step. From the form of the matrix $\lambda(A) = \{0; 0; 0\}$ and $\mathcal{R}(A) = \text{span}\{\mathbf{e}_1\}$. Hence $r = 1$ and there is a vector $\mathbf{w}_1 = \mathbf{e}_1$ from this subspace $\mathcal{R}(A)$ satisfying the condition (10).

II step. Let us find the basis of the null space $\mathcal{N}(A)$ of the matrix A :

$$\begin{bmatrix} 0 & 1 & 2 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix} \Rightarrow \mathbf{n}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \wedge \mathbf{n}_2 = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}.$$

The vector $\mathbf{n}_1$ belongs to the subspace $\mathcal{R}(A) \cap \mathcal{N}(A)$ and $p = \dim \mathcal{R}(A) \cap \mathcal{N}(A) = 1$. We solve the system

$$\begin{bmatrix} 0 & 1 & 2 & \vdots & 1 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix} \Rightarrow \mathbf{y}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

III step. We take for the vector $\mathbf{z}_1$ the vector $\mathbf{n}_2$ and form the matrix X :

$$X = \begin{bmatrix} \mathbf{w}_1 & \mathbf{y}_1 & \mathbf{z}_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{bmatrix}.$$

Now we find the inverse matrix

$$\begin{bmatrix} 1 & 0 & 0 & \vdots & 1 & 0 & 0 \\ 0 & 1 & 2 & \vdots & 0 & 1 & 0 \\ 0 & 0 & -1 & \vdots & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & \vdots & 1 & 0 & 0 \\ 0 & 1 & 0 & \vdots & 0 & 1 & 2 \\ 0 & 0 & 1 & \vdots & 0 & 0 & -1 \end{bmatrix} \Rightarrow$$

$$X^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{bmatrix}$$

and the Jordan matrix

$$X^{-1}AX = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The software package “Maple” gives for the Jordan decomposition:

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

Since the matrix X in the Jordan decomposition of the matrix A is not uniquely defined, then for many problems it is of interest to choose the matrix X so that the conditional number $k(X)$ is the least. Such a problem arose in example 1.2.9.4.

Problem 5.3.1. Find the Jordan decomposition of the matrix

$$A = \begin{bmatrix} 3 & 1 & 0 & 0 \\ -4 & -1 & 0 & 0 \\ 7 & 1 & 2 & 1 \\ -17 & -6 & -1 & 0 \end{bmatrix}.$$

Problem 5.3.2.* Find the Jordan decomposition of the matrix

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

Problem 5.3.3.* Find the Jordan decomposition of the matrix

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

Problem 5.3.4. Let the Jordan decomposition of the matrix $A \in \mathbf{R}^{n \times n}$ be $A = MJM^{-1}$. Show that $A^2 = A \Rightarrow J^2 = J$.

2.6 Strict Methods of Solving Linear Algebraic Systems of Equations

2.6.1 LDM^T Decomposition and LDL^T Decomposition of a Matrix

Next we will consider the special cases of LU factorizations of square matrices.

Proposition 6.1.1. If all the principal minors of the matrix $A \in \mathbf{R}^{n \times n}$ are different from zero, then there exist lower triangular matrices L and M with the unit leading diagonal and a diagonal matrix $D = \text{diag}(d_1, \dots, d_n)$ that

$$A = LDM^T, \quad (1)$$

and the decomposition (1) is unique.

Proof. Since all the principal minors of the matrix $A \in \mathbf{R}^{n \times n}$ are nonzero, then the proposition 1.2.2 implies that there exists a unique LU factorization of the matrix A

$$A = LU. \quad (2)$$

Let $D = \text{diag}(d_1, \dots, d_n)$, where $d_i = u_{ii}$ ($i = 1: n$). From the regularity of the matrix A it follows that the matrix D is regular. Therefore, $\exists D^{-1}$ and $M^T = D^{-1}U$ is an upper triangular matrix. Hence

$$A = LU = LD(D^{-1}U) = LDM^T.$$

The uniqueness of the decomposition (1) follows from the uniqueness of the factorization (2). $\square$

Definition 6.1.1. The decomposition (1) is called the LDM^T decomposition of the regular matrix $A \in \mathbf{R}^{n \times n}$.

Example 6.1.1.* Let us find the LDM^T decomposition of the matrix

$$A = \begin{bmatrix} 1 & 2 & 0 & 1 \\ -1 & -1 & -3 & 0 \\ -1 & -3 & 2 & 2 \\ 2 & 4 & 0 & 1 \end{bmatrix}.$$

We state that if the principal minors of the matrix A are different from zero, then, by transforming the matrix A to the triangular form by the Gauss transformation, we find simultaneously both the matrix L and the matrix U . Namely, the entry l_{ij} ($i > j$) of the lower triangular matrix L equals the factor by which the j -th row must be multiplied when it is subtracted from

the i -th row to delete the entry in the i -th row. We find

$$\begin{aligned} & \begin{bmatrix} 1 & 2 & 0 & 1 \\ -1 & -1 & -3 & 0 \\ -1 & -3 & 2 & 2 \\ 2 & 4 & 0 & 1 \end{bmatrix} \xrightarrow{\begin{matrix} l_{21} = -1 \\ l_{31} = -1 \\ l_{41} = 2 \end{matrix}} \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & -3 & 1 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & 0 & -1 \end{bmatrix} \xrightarrow{\begin{matrix} l_{32} = -1 \\ l_{42} = 0 \end{matrix}} \\ & \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & -1 & 4 \\ 0 & 0 & 0 & -1 \end{bmatrix} \xrightarrow{l_{43} = 0} \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & -1 & 4 \\ 0 & 0 & 0 & -1 \end{bmatrix} = U \end{aligned}$$

and

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

Let us check:

$$LU = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & -1 & 4 \\ 0 & 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 & 1 \\ -1 & -1 & -3 & 0 \\ -1 & -3 & 2 & 2 \\ 2 & 4 & 0 & 1 \end{bmatrix} = A.$$

Proposition 6.1.2. If the regular matrix $A \in \mathbf{R}^{n \times n}$ is symmetric and the LDM^T decomposition of it has the form (1), then $L = M$, i.e.,

$$A = LDL^T. \quad (3)$$

Proof. From decomposition (1) it follows that

$$AM^{-T} = LD.$$

Multiplying both sides of the last equality on the left by matrix M^{-1} , we get

$$M^{-1}AM^{-T} = M^{-1}LD. \quad (4)$$

The matrix $M^{-1}AM^{-T}$ is symmetric since

$$(M^{-1}AM^{-T})^T = M^{-1}A^TM^{-T} = M^{-1}AM^{-T}.$$

The matrix $M^{-1}AM^{-T}$ is a lower triangular matrix since both M^{-1} and $AM^{-T} = LD$ are lower triangular matrices. In virtue of relation (4), the matrix $M^{-1}LD$ is also symmetric and lower triangular. Therefore, the matrix $M^{-1}LD$ is diagonal. Since the matrix D is regular, then also the matrix $M^{-1}L$ is diagonal. In addition, the matrix $M^{-1}L$ is a lower triangular matrix with the unit diagonal. Hence $M^{-1}L = I$ or $L = M$. $\square$

Problem 6.1.1.* Find the LU factorization, LDM^T decomposition and LDL^T decomposition of the matrix

$$A = \begin{bmatrix} 1 & -1 & 2 & 0 \\ -1 & 2 & -3 & 1 \\ 2 & -3 & 1 & 3 \\ 0 & 1 & 3 & -4 \end{bmatrix}.$$

2.6.2 Positive Definite Systems

Definition 6.2.1. The matrix $A \in \mathbf{R}^{n \times n}$ is a *positive definite matrix* if

$$\mathbf{x}^T A \mathbf{x} > 0$$

for all nonzero vector $\mathbf{x} \in \mathbf{R}^n$.

Example 6.2.1. The matrix

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

is a positive definite one since for $\forall \mathbf{x} = \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix}^T \in \mathbf{R}^2$

$$\begin{aligned} \mathbf{x}^T A \mathbf{x} &= \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix} \begin{bmatrix} 2\xi_1 + \xi_2 \\ \xi_1 + \xi_2 \end{bmatrix} = \\ &= 2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2 = \xi_1^2 + (\xi_1 + \xi_2)^2 > 0. \end{aligned}$$

Problem 6.2.1.* Show that the matrix

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

is positive definite.

Proposition 6.2.1. If $A \in \mathbf{R}^{n \times n}$ is a positive definite matrix and the column vectors of the matrix $X \in \mathbf{R}^{n \times k}$ are linearly independent, then the matrix

$$B = X^T A X \in \mathbf{R}^{k \times k}$$

is also positive defined.

Proof. If for the vector $\mathbf{z} \in \mathbf{R}^k$ the relation

$$0 \geq \mathbf{z}^T B \mathbf{z}$$

holds, then

$$0 \geq \mathbf{z}^T B \mathbf{z} = \mathbf{z}^T X^T A X \mathbf{z} = (X \mathbf{z})^T A (X \mathbf{z})$$

and from the positive definiteness of the matrix A it follows that $X \mathbf{z} = \mathbf{0}$. Since the column vectors of the matrix X are linearly independent, then from $X \mathbf{z} = \mathbf{0}$ it follows that $\mathbf{z} = \mathbf{0}$. Hence from conditions $\mathbf{z} \in \mathbf{R}^k$ and $\mathbf{z} \neq \mathbf{0}$ it follows that $\mathbf{z}^T B \mathbf{z} > 0$, i.e., the matrix B is positive definite. $\square$

Corollary 6.2.1. If the matrix $A \in \mathbf{R}^{n \times n}$ is positive definite, then all the submatrices of the matrix A obtained by deleting the rows and columns of the matrix A with the same numbers are positive definite and all the elements on the leading diagonal of the matrix are positive.

Proof. If $\mathbf{v} \in \mathbf{R}^k$ ($k \leq n$) is a vector with natural number coordinates satisfying the condition

$$1 \leq \nu_1 < \dots < \nu_k \leq n,$$

then

$$X = I_n(:, \mathbf{v}) \in \mathbf{R}^{n \times k}$$

is a matrix obtained from the unit matrix I_n by taking the column-vectors with indices $\nu_1, \dots, \nu_k$. Hence the column-vectors of the matrix X are linearly independent, and proposition 6.2.1 implies that the matrix $X^T A X$ is positive definite. The matrix $X^T A X$ is a submatrix of the matrix A obtained from the rows and columns with numbers $\nu_1, \dots, \nu_k$ of the matrix A . Therefore, all the submatrices of the matrix A obtained by deleting the rows and columns of the matrix A with the same numbers are positive definite. Taking $k = 1$, we get the second part of the statement. $\square$

Corollary 6.2.2. If $A \in \mathbf{R}^{n \times n}$ is positive definite, then the matrix A has a decomposition $A = LDM^T$ and all the leading diagonal elements of the matrix D are positive.

Proof. On the ground of corollary 6.2.1, all the submatrices $A(1 : k, 1 : k)$ ($1 \leq k \leq n$) of the matrix A are positive definite, and, therefore, regular matrices, and proposition 6.1.1 implies the existence of the LDM^T decomposition. Taking $X = L^{-T}$ in proposition 6.2.1, we find that the matrix

$$B = DM^T L^{-T} = L^{-1}AL^{-T}$$

is positive definite. Since the matrix $M^T L^{-T}$ is an upper triangular matrix with the unit diagonal, the matrices B and D have the same leading diagonal and the elements on it must be positive, provided that B is positive definite. $\square$

Proposition 6.2.2 (*Cholesky factorization*). If the matrix $A \in \mathbf{R}^{n \times n}$ is symmetric and positive definite, then there exists exactly one lower triangular matrix G with the positive leading diagonal such that

$$A = GG^T. \quad (5)$$

Proof. In virtue of proposition 6.1.2, there exist and are uniquely defined the lower triangular matrix L with the unit diagonal and the diagonal matrix $D = \text{diag}(d_1, \dots, d_n)$ such that the decomposition (3) holds, i.e., $A = LDL^T$. The corollary 6.2.2 provides that elements d_k of the matrix D are positive. Therefore, the matrix

$$G = L\sqrt{D} = L \cdot \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n}) \in \mathbf{R}^{n \times n}$$

is a lower triangular matrix with the positive leading diagonal, and equality (5) holds. The uniqueness of the factorization follows from the uniqueness of the decomposition (3). $\square$

The factorization (5) is known as the *Cholesky factorization*. The matrix G is called the *Cholesky triangular matrix* of the matrix A . To solve the system of equations

$$Ax = b$$

having the symmetric and positive definite matrix A , one has to find the Cholesky triangular matrix of the matrix A . Secondly, one has to solve the system with the triangular matrix

$$Gy = b.$$

Thirdly, one has to solve the system

$$G^T \mathbf{x} = \mathbf{y}.$$

The Cholesky factorization can be found step by step.

Proposition 6.2.3. If the matrix $A \in \mathbf{R}^{n \times n}$ is symmetric and positive definite, then, denoting

$$A = \begin{bmatrix} \alpha & \mathbf{v}^T \\ \mathbf{v} & B \end{bmatrix},$$

the matrix A can be expressed by

$$A = \begin{bmatrix} \alpha & \mathbf{v}^T \\ \mathbf{v} & B \end{bmatrix} = \begin{bmatrix} \beta & \mathbf{0}^T \\ \mathbf{v}/\beta & I_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix} \begin{bmatrix} \beta & \mathbf{v}^T/\beta \\ \mathbf{0} & I_{n-1} \end{bmatrix}, \quad (6)$$

where $\beta = \sqrt{\alpha}$. The matrix $B - \mathbf{v}\mathbf{v}^T/\alpha$ is positive definite. If

$$B - \mathbf{v}\mathbf{v}^T/\alpha = G_1 G_1^T,$$

then $A = GG^T$, where

$$G = \begin{bmatrix} \beta & \mathbf{0}^T \\ \mathbf{v}/\beta & G_1 \end{bmatrix}.$$

Proof. Let us check the accuracy of the decomposition (6):

$$\begin{aligned} & \begin{bmatrix} \beta & \mathbf{0}^T \\ \mathbf{v}/\beta & I_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix} \begin{bmatrix} \beta & \mathbf{v}^T/\beta \\ \mathbf{0} & I_{n-1} \end{bmatrix} = \\ &= \begin{bmatrix} \beta & \mathbf{0}^T \\ \mathbf{v}/\beta & B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix} \begin{bmatrix} \beta & \mathbf{v}^T/\beta \\ \mathbf{0} & I_{n-1} \end{bmatrix} = \begin{bmatrix} \beta^2 & \mathbf{v}^T \\ \mathbf{v} & \mathbf{v}\mathbf{v}^T/\beta^2 + B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix} = \\ &= \begin{bmatrix} \alpha & \mathbf{v}^T \\ \mathbf{v} & B \end{bmatrix} = A. \end{aligned}$$

If

$$X = \begin{bmatrix} 1 & -\mathbf{v}^T/\alpha \\ \mathbf{0} & I_{n-1} \end{bmatrix},$$

then

$$X^T A X = \begin{bmatrix} 1 & \mathbf{0}^T \\ -\mathbf{v}/\alpha & I_{n-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{v}^T \\ \mathbf{v} & B \end{bmatrix} \begin{bmatrix} 1 & -\mathbf{v}^T/\alpha \\ \mathbf{0} & I_{n-1} \end{bmatrix} =$$

$$= \begin{bmatrix} \alpha & \mathbf{v}^T \\ \mathbf{0} & B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix} \begin{bmatrix} 1 & -\mathbf{v}^T/\alpha \\ \mathbf{0} & I_{n-1} \end{bmatrix} = \begin{bmatrix} \alpha & \mathbf{0}^T \\ \mathbf{0} & B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix}.$$

Since the matrix A is positive definite and the column-vector system of the matrix X is linearly independent, proposition 6.2.1 implies the positive definiteness of the matrix

$$\begin{bmatrix} \alpha & \mathbf{0}^T \\ \mathbf{0} & B - \mathbf{v}\mathbf{v}^T/\alpha \end{bmatrix}$$

and from corollary 6.2.1 it follows that the matrix $B - \mathbf{v}\mathbf{v}^T/\alpha$ is likewise positive definite. So we can, analogously to the partition of the matrix A into blocks, decompose the matrix $B - \mathbf{v}\mathbf{v}^T/\alpha$ into blocks, etc.

Example 6.2.2. Let us find the LU , LDM^T , LDL^T and Cholesky factorizations of the matrix

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 8 & 4 \\ 0 & 4 & 13 \end{bmatrix}.$$

The principal minors of the matrix A are nonzero. We find

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 8 & 4 \\ 0 & 4 & 13 \end{bmatrix} \xrightarrow[l_{31}=0]{l_{21}=2} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 4 & 4 \\ 0 & 4 & 13 \end{bmatrix} \xrightarrow[l_{32}=1]{} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 4 & 4 \\ 0 & 0 & 9 \end{bmatrix} = U,$$

and

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

and also

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 4 & 4 \\ 0 & 0 & 9 \end{bmatrix}.$$

Knowing the LU factorization of the matrix A , we will find the LDM^T decomposition, LDL^T decomposition and Cholesky factorization of it:

$$A = LDM^T = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = LDL^T$$

and

$$A = GG^T = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 0 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}.$$

Let us find the Cholesky factorization of the matrix A also step by step using the algorithm given in proposition 6.2.3. Since at the first step

$$\alpha_1 = 1, \beta_1 = \sqrt{\alpha_1} = 1, \mathbf{v}_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, B_1 = \begin{bmatrix} 8 & 4 \\ 4 & 13 \end{bmatrix},$$

then

$$B_1 - \mathbf{v}_1 \mathbf{v}_1^T / \alpha_1 = \begin{bmatrix} 8 & 4 \\ 4 & 13 \end{bmatrix} - \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \end{bmatrix} / 1 = \begin{bmatrix} 4 & 4 \\ 4 & 13 \end{bmatrix}.$$

On the next step,

$$\alpha_2 = 4, \beta_2 = \sqrt{4} = 2, \mathbf{v}_2 = \begin{bmatrix} 4 \end{bmatrix}, B_2 = \begin{bmatrix} 13 \end{bmatrix}$$

and

$$B_2 - \mathbf{v}_2 \mathbf{v}_2^T / \alpha_2 = \begin{bmatrix} 13 \end{bmatrix} - \begin{bmatrix} 4 \end{bmatrix} \begin{bmatrix} 4 \end{bmatrix}^T / 4 = \begin{bmatrix} 9 \end{bmatrix} = \begin{bmatrix} 3 \end{bmatrix} \begin{bmatrix} 3 \end{bmatrix}^T.$$

Therefore,

$$G_2 = \begin{bmatrix} 3 \end{bmatrix}$$

and

$$G_1 = \begin{bmatrix} \beta_2 & 0 \\ \mathbf{v}_2 / \beta_2 & G_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 2 & 3 \end{bmatrix}$$

and

$$G = \begin{bmatrix} \beta_1 & 0 \\ \mathbf{v}_1 / \beta_1 & G_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 0 & 2 & 3 \end{bmatrix}.$$

Problem 6.2.2.* Find the Cholesky factorization of the positive definite matrix

$$A = \begin{bmatrix} 1 & 2 & -1 & 2 \\ 2 & 8 & 6 & 0 \\ -1 & 6 & 21 & -2 \\ 2 & 0 & -2 & 25 \end{bmatrix}.$$

Problem 6.2.3.* Solve the system of equations $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 10 & -10 \\ 1 & -10 & 14 \end{bmatrix} \quad \wedge \quad \mathbf{b} = \begin{bmatrix} 2 \\ -2 \\ 6 \end{bmatrix},$$

when the Cholesky factorization of the matrix A is given

$$A = GG^T \quad \wedge \quad G = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & -3 & 2 \end{bmatrix}.$$

2.6.3 Positive Semidefinite Matrices

Definition 6.3.1. A matrix $A \in \mathbf{R}^{n \times n}$ is called a *positive semidefinite matrix* if

$$\forall \mathbf{x} \in \mathbf{R}^n \Rightarrow \mathbf{x}^T A \mathbf{x} \geq 0.$$

Example 6.3.1. The matrix

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

is positive semidefinite since for $\forall \mathbf{x} = \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix}^T \in \mathbf{R}^2$

$$\begin{aligned} \mathbf{x}^T A \mathbf{x} &= \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \\ &= \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix} \begin{bmatrix} \xi_1 + \xi_2 \\ \xi_1 + \xi_2 \end{bmatrix} = (\xi_1 + \xi_2)^2 \geq 0, \end{aligned}$$

and in the case $\xi_1 = -\xi_2 \wedge \xi_1 \neq 0$ we see that $\mathbf{x}^T A \mathbf{x} = 0$, but $\mathbf{x} \neq \mathbf{0}$, i.e., the matrix A is a positive semidefinite matrix, but it is not positive definite.

Problem 6.3.1.* Show that the matrix

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

is positive semidefinite.

Proposition 6.3.1. If $A \in \mathbf{R}^{n \times n}$ is a symmetric positive semidefinite matrix, then

$$|a_{ij}| \leq \sqrt{a_{ii}a_{jj}}, \quad (7)$$

$$a_{ii} = 0 \Rightarrow A(i, :) = A(:, i) = 0, \quad (8)$$

$$|a_{ij}| \leq (a_{ii} + a_{jj})/2 \quad (9)$$

and

$$\max_{i,j} |a_{ij}| = \max_i a_{ii}. \quad (10)$$

Proof. Let $\alpha \in \mathbf{R}$ and $\mathbf{x} = \mathbf{e}_i + \alpha \mathbf{e}_j \in \mathbf{R}^n$. Since the matrix A is positive semidefinite and symmetric, then

$$\begin{aligned} 0 \leq \mathbf{x}^T A \mathbf{x} &= \begin{bmatrix} \mathbf{0}^T & 1 & \mathbf{0}^T & \alpha & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{nn} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ 1 \\ \mathbf{0} \\ \alpha \\ \mathbf{0} \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{0}^T & 1 & \mathbf{0}^T & \alpha & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} a_{1i} + \alpha a_{1j} \\ \vdots \\ a_{ni} + \alpha a_{nj} \end{bmatrix} = a_{ii} + \alpha a_{ij} + \alpha a_{ji} + \alpha^2 a_{jj}, \end{aligned}$$

and

$$a_{ii} + 2\alpha a_{ij} + \alpha^2 a_{jj} \geq 0. \quad (11)$$

Condition (11) is satisfied exactly when

$$a_{ij}^2 - a_{ii}a_{jj} \leq 0,$$

from which, in its turn, it follows (7), and from it (8). Fixing in inequality (11) $\alpha = \pm 1$ and taking into consideration the symmetry of the matrix A , we get

$$a_{ii} + a_{jj} \geq -2a_{ij},$$

$$a_{ii} + a_{jj} \geq 2a_{ij},$$

and assertions (9) and (10). $\square$

Problem 6.3.2. Show that the algorithm of the Cholesky factorization $A = GG^T$ is applicable (with small changes) also to the symmetric positive semidefinite matrix A .

2.6.4 Polar Decomposition of a Matrix and Method of Square Roots

Proposition 6.4.1 (*on the reduced singular value decomposition of a matrix*). If the matrix $A \in \mathbf{R}^{m \times n}$ ($m \geq n$) has the singular value decomposition $A = U\Sigma V^T$, where $U \in \mathbf{R}^{m \times m}$ and $V \in \mathbf{R}^{n \times n}$ are orthogonal matrices and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n) \in \mathbf{R}^{m \times n}$, then the *reduced singular value decomposition* of this matrix A is

$$A = U_1 \Sigma_1 V^T,$$

where $U_1 = U(:, 1:n)$ and $\Sigma_1 = \Sigma(1:n, :)$.

Proof. If one uses the representation of the matrices U and V by the column-vectors

$$U = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix}$$

and

$$V = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix},$$

then

$$\begin{aligned} A = U\Sigma V^T &= \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \\ \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 \mathbf{v}_1^T \\ \vdots \\ \sigma_n \mathbf{v}_n^T \\ \mathbf{0} \end{bmatrix} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \dots + \sigma_n \mathbf{u}_n \mathbf{v}_n^T + 0 \end{aligned}$$

and

$$\begin{aligned} A = U_1 \Sigma_1 V^T &= \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} = \\ &= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \dots + \sigma_n \mathbf{u}_n \mathbf{v}_n^T. \end{aligned}$$

Example 6.4.1. Find the reduced singular value decomposition of the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \in \mathbf{R}^{3 \times 2}.$$

The singular value decomposition $A = U\Sigma V^T$ of the matrix A was found in example 3.3.1. It is

$$A = \begin{bmatrix} \sqrt{6}/3 & 0 & \sqrt{3}/3 \\ \sqrt{6}/6 & -\sqrt{2}/2 & -\sqrt{3}/3 \\ \sqrt{6}/6 & \sqrt{2}/2 & -\sqrt{3}/3 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}.$$

According to proposition 6.4.1, the reduced singular value decomposition has the form

$$A = U_1 \Sigma_1 V^T,$$

where $U_1 = U(:, 1:n)$ and $\Sigma_1 = \Sigma(1:n, :)$, i.e.,

$$A = \begin{bmatrix} \sqrt{6}/3 & 0 \\ \sqrt{6}/6 & -\sqrt{2}/2 \\ \sqrt{6}/6 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}.$$

Proposition 6.4.2. If the matrix $A \in \mathbf{R}^{m \times n}$ has the reduced singular value decomposition

$$A = U_1 \Sigma_1 V^T,$$

then the matrix A can be written in the form

$$A = ZP, \tag{12}$$

where $Z = U_1 V^T \in \mathbf{R}^{m \times n}$ is a matrix with orthogonal columns and $P = V \Sigma_1 V^T$ is a symmetric positive semidefinite matrix.

Proof. Since $A = U_1 \Sigma_1 V^T$, then

$$A = U_1 (V^T V) \Sigma_1 V^T = (U_1 V^T) (V \Sigma_1 V^T) = ZP.$$

Let us check the correctness of the assertion of the proposition. Firstly, Z is a matrix with orthonormal columns since

$$Z^T Z = (U_1 V^T)^T (U_1 V^T) = V (U_1^T U_1) V^T = V V^T = I.$$

Secondly, $P = V \Sigma_1 V^T$ is a positive semidefinite matrix since

$$\mathbf{x}^T P \mathbf{x} = \mathbf{x}^T V \Sigma_1 V^T \mathbf{x} = (V^T \mathbf{x})^T \Sigma_1 (V^T \mathbf{x}) = \sum_{i=1}^n \sigma_i \eta_i^2 \geq 0 \quad (\forall \mathbf{x} \in \mathbf{R}^n),$$

where $\eta_i = \sum_{k=1}^n v_{ki} \xi_k$. $\square$

Definition 6.4.1. The factorization of the matrix $A \in \mathbf{R}^{m \times n}$ in the form (12) is called the *polar decomposition*.

Example 6.4.2. Find the polar decomposition of the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \in R^{3 \times 2}.$$

In example 6.4.1 the reduced singular value decomposition $A = U_1 \Sigma_1 V^T$ of the matrix A was found. Let us find the factors Z and P occurring in the polar decomposition of the matrix A :

$$\begin{aligned} Z = U_1 V^T &= \begin{bmatrix} \sqrt{6}/3 & 0 \\ \sqrt{6}/6 & -\sqrt{2}/2 \\ \sqrt{6}/6 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} = \\ &= \begin{bmatrix} \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{3} - \frac{1}{2} & \frac{1}{6}\sqrt{3} + \frac{1}{2} \\ \frac{1}{6}\sqrt{3} + \frac{1}{2} & \frac{1}{6}\sqrt{3} - \frac{1}{2} \end{bmatrix}, \\ P = V \Sigma_1 V^T &= \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2}\sqrt{3} + \frac{1}{2} & \frac{1}{2}\sqrt{3} - \frac{1}{2} \\ \frac{1}{2}\sqrt{3} - \frac{1}{2} & \frac{1}{2}\sqrt{3} + \frac{1}{2} \end{bmatrix}. \end{aligned}$$

Hence the polar decomposition of the matrix A is

$$A = ZP = \begin{bmatrix} \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{3} - \frac{1}{2} & \frac{1}{6}\sqrt{3} + \frac{1}{2} \\ \frac{1}{6}\sqrt{3} + \frac{1}{2} & \frac{1}{6}\sqrt{3} - \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2}\sqrt{3} + \frac{1}{2} & \frac{1}{2}\sqrt{3} - \frac{1}{2} \\ \frac{1}{2}\sqrt{3} - \frac{1}{2} & \frac{1}{2}\sqrt{3} + \frac{1}{2} \end{bmatrix}.$$

Problem 6.4.1.* Find the polar decomposition of the matrix

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

Definition 6.4.2. Let $A \in \mathbf{R}^{m \times n}$. If the matrix $X \in \mathbf{R}^{m \times n}$ satisfies the equation $X^2 = A$, then the matrix X is the *square root of the matrix A* .

Proposition 6.4.3. If

$$A = GG^T$$

is the Cholesky factorization of the symmetric positive semidefinite matrix $A \in \mathbf{R}^{n \times n}$ and

$$G = U\Sigma V^T$$

is the singular value decomposition of the matrix G and

$$X = U\Sigma U^T,$$

then

$$X^2 = A,$$

i.e., the matrix X is the square root of the matrix A , where X is a symmetric positive semidefinite matrix. Only one such X exists.

Proof. We find

$$\begin{aligned} A = GG^T &= (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^T V \Sigma U^T = U\Sigma^2 U^T = \\ &= U\Sigma(U^T U)\Sigma U^T = (U\Sigma U^T)(U\Sigma U^T) = X^2. \end{aligned}$$

Show that the matrix X is a uniquely defined symmetric positive semidefinite matrix! $\square$

Example 6.4.3. Let us find the square root of the matrix

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

The matrix A is symmetric and positive semidefinite (see example 6.3.1), and

$$A = GG^T = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

Since

$$G = U\Sigma V^T = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

then

$$X = U\Sigma U^T =$$

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

Problem 6.4.2.* Find the square root of the matrix

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

2.6.5 Systems with Band Matrices

In many applications the matrix A of the system of equations $A\mathbf{x} = \mathbf{b}$ is a band matrix, i.e., the unknown quantity ξ_i appears with a nonzero coefficient only in the i -th equation and some “neighbouring” equations.

Proposition 6.5.1. Let $A = LU$ be the LU factorization of the band matrix $A \in \mathbf{R}^{n \times n}$. If the upper band width of the matrix A is q and the lower band width is p , then the matrix U has the upper band width q and the matrix L has the lower band width p .

Proof. We will prove it by induction. In the case $n = 1$, this assertion is valid. Let us show the admissibility of the step of induction. Let the proposition be correct for an $(n - 1) \times (n - 1)$ matrix A . Let the matrix A be given in the form

$$A = \begin{bmatrix} \alpha & \mathbf{w}^T \\ \mathbf{v} & B \end{bmatrix}.$$

The following equality is valid:

$$A = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{v}/\alpha & I_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & B - \mathbf{v}\mathbf{w}^T/\alpha \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{w}^T \\ 0 & I_{n-1} \end{bmatrix}.$$

Since in the vectors $\mathbf{v}$ and $\mathbf{w}$ at the most only the first p and q coordinates are different from zero, then the matrix $B - \mathbf{v}\mathbf{w}^T/\alpha$ has the upper band width p and the lower band width q . The matrix $B - \mathbf{v}\mathbf{w}^T/\alpha$ is a $(n - 1) \times (n - 1)$

matrix, and, hence $B - \mathbf{v}\mathbf{w}^T/\alpha = L_1U_1$, where U_1 has the upper band width q and L_1 has the lower band width p . The matrices

$$L = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{v}/\alpha & L_1 \end{bmatrix}$$

and

$$U = \begin{bmatrix} \alpha & \mathbf{w}^T \\ \mathbf{0} & U_1 \end{bmatrix}$$

have the band width p and q , respectively, and $A = LU$. $\square$

Problem 6.5.1. Find for the LU factorization of the matrix given in example 6.2.1 the upper and lower band width for the matrices A , L and U .

2.6.6 Block Systems

Let us consider the system in the form

$$\begin{bmatrix} D_1 & F_1 & & \cdots & 0 \\ E_1 & D_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & D_{n-1} & F_{n-1} \\ 0 & \cdots & & E_{n-1} & D_n \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-1} \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_{n-1} \\ \mathbf{b}_n \end{bmatrix}, \quad (13)$$

where $D_i, E_i, F_i \in \mathbf{R}^{q \times q}$ and $\mathbf{x}_i, \mathbf{b}_i \in \mathbf{R}^q$. If we represent the matrix A in the form

$$A = \begin{bmatrix} I & & & \cdots & 0 \\ L_1 & I & & & \vdots \\ & \ddots & \ddots & & \\ \vdots & & \ddots & I & \\ 0 & \cdots & & L_{n-1} & I \end{bmatrix} \begin{bmatrix} U_1 & F_1 & & \cdots & 0 \\ & U_2 & \ddots & & \vdots \\ & & \ddots & \ddots & \\ \vdots & & & U_{n-1} & F_{n-1} \\ 0 & \cdots & & & U_n \end{bmatrix},$$

then

$$A = \begin{bmatrix} U_1 & F_1 & & \cdots & & 0 \\ L_1 U_1 & L_1 F_1 + U_2 & F_2 & & & \vdots \\ & L_2 U_2 & L_2 F_2 + U_3 & F_3 & & \\ & & L_3 U_3 & \ddots & \ddots & \\ \vdots & & & \ddots & \ddots & F_{n-1} \\ 0 & \cdots & & & L_{n-1} U_{n-1} & L_{n-1} F_{n-1} + U_n \end{bmatrix}.$$

Let us find step by step the blocks L_i and U_i :

$$\begin{aligned} U_1 &= D_1 \rightarrow \text{solve } L_1 U_1 = E_1 \rightarrow \\ &\rightarrow U_2 = D_2 - L_1 F_1 \rightarrow \text{solve } L_2 U_2 = E_2 \rightarrow \cdots \rightarrow \\ &\rightarrow U_{n-1} = D_{n-1} - L_{n-2} F_{n-2} \rightarrow \text{solve } L_{n-1} U_{n-1} = E_{n-1} \rightarrow \\ &\rightarrow U_n = D_n - L_{n-1} F_{n-1}. \end{aligned}$$

To solve system (13), one must first solve the system

$$\begin{bmatrix} I & & \cdots & 0 \\ L_1 & I & & \vdots \\ & \ddots & \ddots & \\ \vdots & & \ddots & I \\ 0 & \cdots & L_{n-1} & I \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{n-1} \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_{n-1} \\ \mathbf{b}_n \end{bmatrix}.$$

We find that

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{b}_1 \rightarrow L_1 \mathbf{y}_1 + \mathbf{y}_2 = \mathbf{b}_2 \rightarrow \mathbf{y}_2 = \mathbf{b}_2 - L_1 \mathbf{y}_1 \rightarrow \cdots \rightarrow \\ &\rightarrow L_{i-1} \mathbf{y}_{i-1} + \mathbf{y}_i = \mathbf{b}_i \rightarrow \mathbf{y}_i = \mathbf{b}_i - L_{i-1} \mathbf{y}_{i-1} \rightarrow \cdots \rightarrow \\ &\rightarrow L_{n-1} \mathbf{y}_{n-1} + \mathbf{y}_n = \mathbf{b}_n \rightarrow \mathbf{y}_n = \mathbf{b}_n - L_{n-1} \mathbf{y}_{n-1}. \end{aligned}$$

Secondly, we have to solve the system

$$\begin{bmatrix} U_1 & F_1 & & \cdots & 0 \\ & U_2 & \ddots & & \vdots \\ & & \ddots & \ddots & \\ \vdots & & & U_{n-1} & F_{n-1} \\ 0 & \cdots & & & U_n \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-1} \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{n-1} \\ \mathbf{y}_n \end{bmatrix}.$$

Example 6.6.1.* Let us solve the system of equations $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & -1 & 2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 2 & 2 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & -1 & 1 & 2 & -1 \end{bmatrix} \wedge \mathbf{b} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -4 \\ 3 \\ 0 \\ 1 \end{bmatrix}.$$

This is a block system given by relation (13) since

$$A = \begin{bmatrix} D_1 & F_1 & 0 \\ E_1 & D_2 & F_2 \\ 0 & E_2 & D_3 \end{bmatrix},$$

where $D_i, E_i, F_i \in \mathbf{R}^{2 \times 2}$ and

$$\begin{aligned} D_1 &= \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} \wedge F_1 = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \wedge E_1 = \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix} \wedge D_2 = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix} \\ &\wedge F_2 = \begin{bmatrix} -1 & 1 \\ 2 & 1 \end{bmatrix} \wedge E_2 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \wedge D_3 = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}. \end{aligned}$$

We will express the matrix A in the form

$$\begin{aligned} A = LU &= \begin{bmatrix} I_2 & 0 & 0 \\ L_1 & I_2 & 0 \\ 0 & L_2 & I_2 \end{bmatrix} \begin{bmatrix} U_1 & F_1 & 0 \\ 0 & U_2 & F_2 \\ 0 & 0 & U_3 \end{bmatrix} = \\ &= \begin{bmatrix} U_1 & F_1 & 0 \\ L_1 U_1 & L_1 F_1 + U_2 & F_2 \\ L_2 U_2 & L_2 F_2 + U_3 \end{bmatrix}. \end{aligned}$$

Now we find

$$\begin{aligned} U_1 &= D_1 = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}, \\ L_1 U_1 &= E_1 \Rightarrow L_1 = \begin{bmatrix} -2 & 2 \\ 1 & 0 \end{bmatrix}, \\ L_1 F_1 + U_2 &= D_2 \Rightarrow U_2 = \begin{bmatrix} 4 & 3 \\ -3 & 0 \end{bmatrix}, \end{aligned}$$

$$L_2 U_2 = E_2 \Rightarrow L_2 = \begin{bmatrix} 1/3 & 1/9 \\ 1/3 & 7/9 \end{bmatrix},$$

$$L_2 F_2 + U_3 = D_3 \Rightarrow U_3 = \begin{bmatrix} 10/9 & 5/9 \\ 7/9 & -19/9 \end{bmatrix}$$

and

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -2 & 2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1/3 & 1/9 & 1 & 0 \\ 0 & 0 & 1/3 & 7/9 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 4 & 3 & -1 & 1 \\ 0 & 0 & -3 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 10/9 & 5/9 \\ 0 & 0 & 0 & 0 & 7/9 & -19/9 \end{bmatrix}.$$

To find a solution of the system $A\mathbf{x} = \mathbf{b}$, we shall solve two systems $L\mathbf{y} = \mathbf{b}$ and $U\mathbf{x} = \mathbf{y}$. The system $L\mathbf{y} = \mathbf{b}$ can be expressed in the form

$$\begin{bmatrix} I_2 & 0 & 0 \\ L_1 & I_2 & 0 \\ 0 & L_2 & I_2 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix},$$

where

$$\begin{aligned} \mathbf{b}_1 &= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \wedge \mathbf{b}_2 = \begin{bmatrix} \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} -4 \\ 3 \end{bmatrix} \wedge \\ &\wedge \mathbf{b}_3 = \begin{bmatrix} \beta_5 \\ \beta_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{b}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \wedge \mathbf{y}_2 = \mathbf{b}_2 - L_1 \mathbf{y}_1 = \begin{bmatrix} -4 \\ 2 \end{bmatrix} \wedge \\ &\wedge \mathbf{y}_3 = \mathbf{b}_3 - L_2 \mathbf{y}_2. \end{aligned}$$

Solving the system $U\mathbf{x} = \mathbf{y}$, which can be given in the form

$$\begin{bmatrix} U_1 & F_1 & 0 \\ 0 & U_2 & F_2 \\ 0 & 0 & U_3 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix},$$

we obtain

$$\mathbf{x}_3 = U_3^{-1} \mathbf{y}_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \wedge \mathbf{x}_2 = U_2^{-1} (\mathbf{y}_2 - F_2 \mathbf{x}_3) = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \wedge$$

$$\wedge \mathbf{x}_1 = U_1^{-1}(\mathbf{y}_1 - F_1 \mathbf{x}_2) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Thus,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 \end{bmatrix}^T.$$

2.6.7 Solution of the Systems of Equations by QR Method

Let us consider the system

$$A\mathbf{x} = \mathbf{b}, \quad (14)$$

where $A = QR$ is a QR factorization of the regular matrix $A \in \mathbf{R}^{n \times n}$, while $Q \in \mathbf{R}^{n \times n}$ is an orthogonal matrix and $R \in \mathbf{R}^{n \times n}$ is an upper triangular matrix. Substituting in (14) the matrix A by its QR factorization, we get

$$QR\mathbf{x} = \mathbf{b}. \quad (15)$$

Multiplying the both sides of equality (15) on the left by the matrix Q^T , we find

$$R\mathbf{x} = Q^T \mathbf{b}. \quad (16)$$

System (16) has an upper triangular matrix R . From the regularity of the matrix A it follows the regularity of the matrix R . Hence system (16) is uniquely solvable. For this the substitution given in proposition 1.1.2 will be used backwards.

Example 6.7.1. Let us solve the system

$$\begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & -1 & -1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad (17)$$

using the QR method.

In example 2.3.2 the QR factorization of the matrix of the system

$$\begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & -1 & -1 \end{bmatrix} = \frac{\sqrt{5}}{35} \begin{bmatrix} -2\sqrt{5} & -15 & 0 \\ -6\sqrt{5} & 4 & -7 \\ 3\sqrt{5} & -2 & -14 \end{bmatrix} \begin{bmatrix} -7 & -\frac{15}{7} & -\frac{5}{7} \\ 0 & \frac{2\sqrt{5}}{7} & -\frac{13\sqrt{5}}{35} \\ 0 & 0 & \frac{2\sqrt{5}}{5} \end{bmatrix}$$

was found. We respect system (17) in form (16):

$$\begin{bmatrix} -7 & -\frac{15}{7} & -\frac{5}{7} \\ 0 & \frac{2}{7}\sqrt{5} & -\frac{13}{35}\sqrt{5} \\ 0 & 0 & \frac{2}{5}\sqrt{5} \end{bmatrix} \mathbf{x} = \frac{\sqrt{5}}{35} \begin{bmatrix} -2\sqrt{5} & -6\sqrt{5} & 3\sqrt{5} \\ -15 & 4 & -2 \\ 0 & -7 & -14 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

i.e.,

$$\begin{bmatrix} -7 & -\frac{15}{7} & -\frac{5}{7} \\ 0 & \frac{2}{7}\sqrt{5} & -\frac{13}{35}\sqrt{5} \\ 0 & 0 & \frac{2}{5}\sqrt{5} \end{bmatrix} \mathbf{x} = \begin{bmatrix} -\frac{1}{7} \\ -\frac{17}{35}\sqrt{5} \\ -\frac{2}{5}\sqrt{5} \end{bmatrix}.$$

We solve the obtained system with the upper triangular matrix using the backwards substitution. The result is $\mathbf{x} = \begin{bmatrix} 1 & -3 & -1 \end{bmatrix}^T$.

Let us consider the solving of system (14), where $A = QR$ is a QR factorization of the regular matrix $A \in \mathbf{R}^{m \times n}$ ($m \geq n$), where $Q \in \mathbf{R}^{m \times m}$ is an orthogonal matrix and $R \in \mathbf{R}^{n \times n}$ is an upper triangular matrix, by the least-squares method. Let

$$Q^T A = R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

and

$$Q^T \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix},$$

where $R_1 \in \mathbf{R}^{n \times n}$, $\mathbf{c} \in \mathbf{R}^n$ and $\mathbf{d} \in \mathbf{R}^{m-n}$. We find

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|_2^2 &= \|Q^T A\mathbf{x} - Q^T \mathbf{b}\|_2^2 = \left\| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \right\|_2^2 = \\ &= \left\| \begin{bmatrix} R_1 \mathbf{x} - \mathbf{c} \\ \mathbf{0} - \mathbf{d} \end{bmatrix} \right\|_2^2 = \|R_1 \mathbf{x} - \mathbf{c}\|_2^2 + \|\mathbf{d}\|_2^2. \end{aligned}$$

Since the quantity $\|\mathbf{d}\|_2^2$ is a constant, we can minimize only the quantity

$$\|R_1 \mathbf{x} - \mathbf{c}\|_2^2,$$

and the minimal value of it is 0. Really, from the condition $\dim A = n$ it follows that the matrix R_1 is regular. Hence the system

$$R_1 \mathbf{x}_{LS} = \mathbf{c},$$

where the symbol $\mathbf{x}_{LS}$ denotes the least-squares solution of system (14), is uniquely solvable.

Example 6.7.2. Let us find the least squares solution of the system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

using the QR method.

Using the software package “Maple”, we obtain the QR factorization of the matrix of the system

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

From this factorization it appears that

$$R_1 = \begin{bmatrix} -\sqrt{2} & 0 & 0 \\ 0 & -\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

To get the vector $\mathbf{c}$, we find

$$Q^T \mathbf{b} = \begin{bmatrix} -\frac{1}{2}\sqrt{2} & 0 & 0 & -\frac{1}{2}\sqrt{2} & 0 \\ 0 & -\frac{1}{2}\sqrt{2} & 0 & 0 & \frac{1}{2}\sqrt{2} \\ 0 & 0 & -1 & 0 & 0 \\ -\frac{1}{2}\sqrt{2} & 0 & 0 & \frac{1}{2}\sqrt{2} & 0 \\ 0 & \frac{1}{2}\sqrt{2} & 0 & 0 & \frac{1}{2}\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \\ -1 \\ -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{bmatrix}.$$

Hence

$$\mathbf{c} = \begin{bmatrix} -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} & -1 \end{bmatrix}^T.$$

We get for the concrete form of the system $R_1 \mathbf{x}_{LS} = \mathbf{c}$

$$\begin{bmatrix} -\sqrt{2} & 0 & 0 \\ 0 & -\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}_{LS} = \begin{bmatrix} -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \\ -1 \end{bmatrix},$$

from which it follows that

$$\mathbf{x}_{LS} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -1 \end{bmatrix}^T.$$

Example 6.7.3.* Let us find the least-squares solution of the system

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

In example 2.3.3 it was found the QR factorization of the matrix of the system:

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2}{3}\sqrt{2} \\ \frac{2}{3} & \frac{1}{2}\sqrt{2} & -\frac{1}{6}\sqrt{2} \\ \frac{2}{3} & -\frac{1}{2}\sqrt{2} & -\frac{1}{6}\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} = QR.$$

Omitting the last row of zeros in the matrix R , we get

$$R_1 = \begin{bmatrix} 3 & 3 \\ 0 & \sqrt{2} \end{bmatrix}.$$

Now we find

$$Q^T \mathbf{b} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3}\sqrt{2} \\ 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{2}{3}\sqrt{2} & -\frac{1}{6}\sqrt{2} & -\frac{1}{6}\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ 0 \\ \frac{1}{3}\sqrt{2} \end{bmatrix}$$

Taking the first two components of this vector (the matrix R_1 has two rows), we obtain

$$\mathbf{c} = \begin{bmatrix} \frac{5}{3} \\ 0 \end{bmatrix}.$$

We get the least-squares solution of the initial system from $R_1 \mathbf{x}_{LS} = \mathbf{c}$, i.e.,

$$\begin{bmatrix} 3 & 3 \\ 0 & \sqrt{2} \end{bmatrix} \mathbf{x}_{LS} = \begin{bmatrix} \frac{5}{3} \\ 0 \end{bmatrix} \Rightarrow \mathbf{x}_{LS} = \begin{bmatrix} \frac{5}{9} \\ 0 \end{bmatrix}.$$

Problem 6.7.1.* Solve the system of equations

$$\begin{bmatrix} 12 & -3 & 1 \\ -3 & 1 & 2 \\ 4 & -\frac{4}{3} & -1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}$$

knowing the QR factorization of the system matrix

$$\begin{bmatrix} 12 & -3 & 1 \\ -3 & 1 & 2 \\ 4 & -\frac{4}{3} & -1 \end{bmatrix} = \begin{bmatrix} \frac{12}{13} & -\frac{5}{13} & 0 \\ -\frac{3}{13} & -\frac{36}{65} & -\frac{52}{65} \\ \frac{4}{13} & \frac{48}{65} & -\frac{39}{65} \end{bmatrix} \begin{bmatrix} 13 & -\frac{133}{39} & \frac{2}{13} \\ 0 & -\frac{5}{13} & -\frac{29}{13} \\ 0 & 0 & -1 \end{bmatrix}.$$

Problem 6.7.2.* Find the least-squares solution of the system

$$\begin{bmatrix} 0 & 2 \\ 1 & 3 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_1 \end{bmatrix} = \begin{bmatrix} -3 \\ 4 \\ 2 \end{bmatrix}.$$

2.7 Iterative Solution of Systems of Equations

2.7.1 Powers of a Matrix and Inverse Matrix

Proposition 7.1.1. If $F \in \mathbf{R}^{n \times n}$ and $\|F\|_p < 1$, then $I - F$ is a regular matrix and

$$(I - F)^{-1} = \sum_{k=0}^{\infty} F^k,$$

while

$$\|(I - F)^{-1}\|_p \leq \frac{1}{1 - \|F\|_p}.$$

Proof. If we suppose the contrary to the assertion that the matrix $I - F$ is singular, then there exists a nonzero vector $\mathbf{x} \in \mathbf{R}^n$ that $(I - F)\mathbf{x} = \mathbf{0}$, i.e., $\mathbf{x} = F\mathbf{x}$ and $\|\mathbf{x}\|_p = \|F\mathbf{x}\|_p$, $\|F\|_p \geq 1$. Hence the matrix $I - F$ is a regular matrix. To find the matrix $(I - F)^{-1}$, we consider the identity

$$\left(\sum_{k=0}^n F^k \right) (I - F) = I - F^{n+1}.$$

Since

$$\|F^k\|_p \leq \|F\|_p^k \wedge \|F\|_p < 1 \Rightarrow \lim_{k \rightarrow \infty} F^k = 0,$$

then

$$\left(\lim_{n \rightarrow \infty} \sum_{k=0}^n F^k \right) (I - F) = I,$$

that implies

$$(I - F)^{-1} = \lim_{n \rightarrow \infty} \sum_{k=0}^n F^k = \sum_{k=0}^{\infty} F^k$$

and

$$\|(I - F)^{-1}\|_p \leq \sum_{k=0}^n \|F\|_p^k \leq \frac{1}{1 - \|F\|_p},$$

which was to be proved. $\square$

Proposition 7.1.2. Let $Q^H A Q = T = D + N$ be the Schur factorization of the matrix $A \in \mathbf{C}^{n \times n}$, while D is a diagonal matrix and N is a strictly upper triangular matrix (on the leading diagonal there are zeros). Let λ and μ be respectively the greatest and the least modulus eigenvalues of the matrix A . If $\theta \geq 0$, then for all $k \geq 0$

$$\|A^k\|_2 \leq (1 + \theta)^{n-1} \left(|\lambda| + \frac{\|N\|_F}{1 + \theta} \right)^k.$$

If A is a regular matrix and the number θ is such that

$$(1 + \theta)|\mu| > \|N\|_F,$$

then for all $k \geq 0$

$$\|A^{-k}\|_2 \leq (1 + \theta)^{n-1} \left(\frac{1}{|\mu| - \|F\|_F / (1 + \theta)} \right)^k.$$

Proof. See Golub, Loan (1996, pp 336-337). $\square$

The formula

$$B^{-1} = A^{-1} - B^{-1}(B - A)A^{-1}$$

which is easily checked shows how the inverse matrix changes when the matrix A is substituted by the matrix B . The modification of this formula is a *formula of Sherman-Morrison-Woodbury* given in the following proposition.

Proposition 7.1.3. If $A \in \mathbf{R}^{n \times n}$ and $U, V \in \mathbf{R}^{n \times k}$, while matrices A and $I + V^T A^{-1} U$ are regular,

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1}.$$

Proof. See Golub, Loan (1996, pp. 50). $\square$

2.7.2 Jacobi's and Gauss-Seidel Method

Let $A \in \mathbf{C}^{n \times n}$ and $a_{ii} \neq 0$ ($i = 1 : n$). We will consider the solution of the system of equations

$$A\mathbf{x} = \mathbf{b} \quad (1)$$

by an iterative method.

Definition 7.2.1. The *approximation* or the *approximate value* of the solution $\mathbf{x}$ of system (1) is a vector that in certain sense differs little from the vector $\mathbf{x}$. Let us represent system (1) in the form

$$\xi_i = (\beta_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} \xi_j) / a_{ii} \quad (i = 1 : n).$$

Jacobi's iterative process is defined by the algorithm

$$\xi_i^{(k+1)} = (\beta_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} \xi_j^{(k)}) / a_{ii} \quad (i = 1 : n). \quad (2)$$

The Gauss-Seidel iterative process is defined by the algorithm

$$\xi_i^{(k+1)} = (\beta_i - \sum_{j=1}^{i-1} a_{ij} \xi_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} \xi_j^{(k)}) / a_{ii} \quad (i = 1 : n). \quad (3)$$

In case of both Jacobi's and the Gauss-Seidel iterative processes the transition from the approximation $\mathbf{x}^{(k)} = \{\xi_i^{(k)}\}$ of the solution of system (1) to the next approximation $\mathbf{x}^{(k+1)} = \{\xi_i^{(k+1)}\}$ can be described using the matrices

$$L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \ddots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

and $D = \text{diag}(a_{11}, \dots, a_{nn})$, while $A = L + D + U$. For example, Jacobi's algorithm can be represented as

$$M_J \mathbf{x}^{(k+1)} = N_J \mathbf{x}^{(k)} + b, \quad (4)$$

where $M_J = D$ and $N_J = -(L + U)$. The Gauss-Seidel algorithm (3) can be represented as

$$M_G \mathbf{x}^{(k+1)} = N_G \mathbf{x}^{(k)} + b, \quad (5)$$

where $M_G = D + L$ and $N_G = -U$.

Example 7.2.1.* Let us solve the system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 1 & -1 \\ -1 & 3 & 0 \\ 1 & 0 & -2 \end{bmatrix} \wedge \mathbf{b} = \begin{bmatrix} 0 \\ 2 \\ -3 \end{bmatrix}$$

by Jacobi's method.

Let us represent the matrix A in the form

$$A = L + D + U = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -2 \end{bmatrix} + \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We will form the matrices M_j and N_j :

$$M_j = D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -2 \end{bmatrix}, \quad N_j = -(L + U) = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

The algorithm of Jacobi's iterative process can be given in the form

$$\mathbf{x}^{(k+1)} = M_j^{-1} N_j \mathbf{x}^{(k)} + M_j^{-1} \mathbf{b}.$$

Since

$$M_j^{-1} N_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & -1/2 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}$$

and

$$M_j^{-1} \mathbf{b} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & -1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ -3 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix},$$

then

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \mathbf{x}^{(k)} + \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix}.$$

If we take for the initial approximation $\mathbf{x}^{(0)} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$, then we shall have

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ .66667 \\ 1.5 \end{bmatrix},$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ .66667 \\ 1.5 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} .83333 \\ .66667 \\ 1.5 \end{bmatrix},$$

:

$$\mathbf{x}^{(3)} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} .83333 \\ .66667 \\ 1.5 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} .83333 \\ .94444 \\ 1.9167 \end{bmatrix},$$

$$\mathbf{x}^{(4)} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} .83333 \\ .94444 \\ 1.9167 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} .97226 \\ .94444 \\ 1.9167 \end{bmatrix},$$

$$\mathbf{x}^{(5)} = \begin{bmatrix} 0 & -1 & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} .97226 \\ .94444 \\ 1.9167 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} .97226 \\ .99075 \\ 1.9861 \end{bmatrix}$$

and so on (the exact solution of this equation is $\mathbf{x} = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix}^T$).

Problem 7.2.1.* Solve the system given in example 7.2.1 by the Gauss-Seidel method.

2.7.3 Decomposition of the System Matrix and Convergence of the Iterative process

Both Jacobi's and the Gauss-Seidel algorithms are of the type

$$M\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + b, \quad (6)$$

where $A = M - N$. Assume that the *decomposition of the matrix* A is given. Applying the iterative algorithm, it is important that the linear system (6) with the system matrix M is easy to solve. By Jacobi's method M is a diagonal matrix and by the Gauss-Seidel method it is a lower triangular matrix. It appears that the convergence of the iterative process given by (6) depends on the spectral radius of the matrix $M^{-1}N$.

Definition 7.3.1. The quantity

$$\rho(G) = \max\{|\lambda| : \lambda \in \lambda(G)\}$$

is the *spectral radius of the matrix* $G \in \mathbf{C}^{n \times n}$.

Proposition 7.3.1. Let $A = M - N$ be the decomposition of the regular matrix $A \in \mathbf{R}^{n \times n}$ and let $\mathbf{b} \in \mathbf{R}^n$. If the matrix M is regular and

$$\rho(M^{-1}N) < 1, \quad (7)$$

then the sequence of the approximation $\{\mathbf{x}^{(k)}\}$ defined by algorithm (6) converges to the solution $\mathbf{x} = A^{-1}\mathbf{b}$ of system (1) by arbitrary initial approximation $\mathbf{x}^{(0)}$.

Proof. Let us get

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}. \quad (8)$$

Since the exact solution satisfies the equality

$$M\mathbf{x} = N\mathbf{x} + \mathbf{b}, \quad (9)$$

then from equalities (6) and (9) we get

$$M(\mathbf{x}^{(k+1)} - \mathbf{x}) = N(\mathbf{x}^{(k)} - \mathbf{x}).$$

Taking into account (8), we find for arbitrary non-negative integer k the relation

$$M\mathbf{e}^{(k+1)} = N\mathbf{e}^{(k)}$$

or

$$\mathbf{e}^{(k+1)} = M^{-1}N\mathbf{e}^{(k)} = (M^{-1}N)^{k+1}\mathbf{e}^{(0)}.$$

By virtue of proposition 7.1.2, it follows from inequality (7) that

$$\lim_{k \rightarrow \infty} (M^{-1}N)^k = 0.$$

Thus,

$$\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}. \quad \square$$

Example 7.3.1.* Let the matrix of the system $A\mathbf{x} = \mathbf{b}$ be

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

We will prove that the sequence of the approximations $\{\mathbf{x}^{(k)}\}$ defined by Jacobi's algorithm converges to the solution of the system for any initial approximation $\mathbf{x}^{(0)}$.

Since

$$A = L + D + U = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$M_j = D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad N_j = -(L + U) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

and

$$M_j^{-1}N_j = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

then

$$\lambda(M_j^{-1}N_j) = \left\{0, \frac{1}{2}i\sqrt{2}, -\frac{1}{2}i\sqrt{2}\right\}$$

and

$$\rho(M_j^{-1}N_j) = \max\{|\lambda| : \lambda \in \lambda(M_j^{-1}N_j)\} = \frac{1}{2}\sqrt{2} < 1.$$

We note that the matrix A is regular. Hence, by virtue of proposition 7.3.1, the sequence of the approximations $\{\mathbf{x}^{(k)}\}$ defined by Jacobi's algorithm converges to the solution of the system $\mathbf{x} = A^{-1}\mathbf{b}$ for any initial approximation $\mathbf{x}^{(0)}$.

Problem 7.3.1.* Solve the system

$$\begin{bmatrix} -2 & 1 & -1 \\ -1 & 4 & -2 \\ -1 & -2 & -2 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 3 \\ 2 \\ -4 \end{bmatrix}$$

both by Jacobi's and the Gauss-Seidel method. Prove that the sequences of the approximations defined by these algorithms converge to the solution of this system for any initial approximation $\mathbf{x}^{(0)}$.

Problem 7.3.2.* Solve the system

$$\begin{bmatrix} 2 & 1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix}$$

both by Jacobi's and the Gauss-Seidel method. Prove that the sequences of the approximations defined by these algorithms converge to the solution of this system for arbitrary initial approximation $\mathbf{x}^{(0)}$.

Definition 7.3.2. The matrix $A \in \mathbf{C}^{n \times n}$ is a *matrix with the strictly dominant diagonal* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \quad (i = 1 : n).$$

Remark 7.3.1. If the matrix $A \in \mathbf{R}^{n \times n}$ is a matrix with the strictly dominant diagonal, then the spectral radius $\rho(M_J^{-1}N_J)$ of the matrix $M_J^{-1}N_J$ satisfies the condition

$$\rho(M_J^{-1}N_J) < 1,$$

i.e., the iteration given by formula (4) converges.

Proof. See Golub, Loan (1996, pp. 120, 512). $\square$

Proposition 7.3.2. If $A \in \mathbf{R}^{n \times n}$ is a symmetric positive definite matrix, then the Gauss-Seidel iterative process converges for arbitrary $\mathbf{x}^{(0)}$.

Proof. Let us denote $A = L + D + L^T$, where L is a strictly lower triangular matrix (zeros on the leading diagonal) and D is a diagonal matrix. Since the matrix A is positive definite, then, by virtue of corollary 6.2.1, the matrix D is also positive definite. Hence $\sqrt{D}$ exists. The matrices A and $L + D$ are regular. Therefore, by virtue of proposition 7.3.1, to prove the convergence of the Gauss-Seidel iterative process, it is sufficient to show that the spectral radius $\rho(G)$ of the matrix $G = -(L + D)^{-1}U$ satisfies the condition $\rho(G) < 1$. Let $G_1 = D^{1/2}GD^{-1/2}$. Since the similar matrices G and G_1 have the same spectrum, then it is sufficient to check the condition $\rho(G_1) < 1$. Let $L_1 = D^{-1/2}LD^{-1/2}$. We find that

$$\begin{aligned} G_1 &= D^{1/2}GD^{-1/2} = -D^{1/2}(L + D)^{-1}L^TD^{-1/2} = \\ &= -D^{1/2}(D^{1/2}L_1D^{1/2} + D^{1/2}ID^{1/2})^{-1}D^{1/2}L_1^TD^{1/2}D^{-1/2} = \\ &= -D^{1/2}[D^{1/2}(L_1 + I)D^{1/2}]^{-1}D^{1/2}L_1^TD^{1/2}D^{-1/2} = \\ &= -D^{1/2}D^{-1/2}(L_1 + I)^{-1}D^{-1/2}D^{1/2}L_1^TD^{1/2}D^{-1/2} = -(L_1 + I)^{-1}L_1^T. \end{aligned}$$

Hence it is sufficient to prove that $\rho(G_1) < 1$. If $G_1\mathbf{x} = \lambda\mathbf{x}$, while $\mathbf{x}^H\mathbf{x} = 1$, then

$$-(I + L_1)^{-1}L_1^T\mathbf{x} = \lambda\mathbf{x},$$

$$-L_1^T \mathbf{x} = \lambda(I + L_1)\mathbf{x}$$

and

$$-\mathbf{x}^H L_1^T \mathbf{x} = \lambda \mathbf{x}^H \mathbf{x} + \lambda \mathbf{x}^H L_1 \mathbf{x}$$

and

$$-\mathbf{x}^H L_1^T \mathbf{x} = \lambda + \lambda \mathbf{x}^H L_1 \mathbf{x}$$

If we set $\mathbf{x}^H L_1 \mathbf{x} = \alpha + i\beta$, then $\mathbf{x}^H L_1^T \mathbf{x} = \alpha - i\beta$ and

$$-\alpha + i\beta\lambda = \lambda(1 + \alpha + i\beta),$$

and also

$$\lambda = \frac{-\alpha + i\beta}{1 + \alpha + i\beta}.$$

Thus,

$$|\lambda|^2 = \frac{\alpha^2 + \beta^2}{\alpha^2 + 2\alpha + 1 + \beta^2}. \quad (10)$$

Since, the matrix A is positive definite, then, by virtue of proposition 6.2.1, the matrix $D^{-1/2}AD^{-1/2}$ is also positive definite and

$$\begin{aligned} D^{-1/2}AD^{-1/2} &= D^{-1/2}(D + L + L^T)D^{-1/2} = \\ &= D^{-1/2}(D^{1/2}ID^{1/2} + D^{1/2}L_1D^{1/2} + D^{1/2}L_1^TD^{1/2})D^{-1/2} = \\ &= I + L_1 + L_1^T. \end{aligned}$$

Hence

$$\begin{aligned} 0 &< \mathbf{x}^H(I + L_1 + L_1^T)\mathbf{x} = 1 + \mathbf{x}^H L_1 \mathbf{x} + \mathbf{x}^H L_1^T \mathbf{x} = 1 + 2\alpha \\ &= 1 + \alpha + i\beta + \alpha - i\beta = 2\alpha + 1, \end{aligned}$$

and condition (10) implies $|\lambda| < 1$, i.e., $\rho(G_1) < 1$. $\square$

2.7.4 Acceleration of the Convergence of an Iterative Process

If the quantity $\rho(M_G^{-1}N_G)$ is smaller than one but close to one, then the Gauss-Seidel method converges, but very slowly. A problem arises how do we accelerate the convergence of the sequence of approximations $\{\mathbf{x}^{(k)}\}$? One of the processes of acceleration of the convergence is the so-called *method of relaxation*. The relaxation method is based on the algorithm

$$\xi_i^{(k+1)} = \omega(\beta_i - \sum_{j=1}^{i-1} a_{ij}\xi_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}\xi_j^{(k)})/a_{ii} + (1-\omega)\xi_i^{(k)} \quad (i = 1 : n),$$

which can be written in the matrix form

$$M_\omega \mathbf{x}^{(k+1)} = N_\omega \mathbf{x}^{(k)} + \omega \mathbf{b},$$

where $M_\omega = D + \omega L$ and $N_\omega = (1-\omega)D - \omega U$. The problem lies in finding the parameter ω so that $\rho(M_\omega^{-1}N_\omega)$ were the least. By certain additional conditions, this problem can be solved.

The second way of acceleration of the convergence is the so-called *Chebyshev method of convergence acceleration*. Let us suppose that we have found using algorithm (6), the approximations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}$ of the solution $\mathbf{x}$ of system (1). Let

$$\mathbf{y}^{(k)} = \sum_{j=0}^k \nu_j(k) \mathbf{x}^{(j)}. \quad (11)$$

The problem lies in finding the coefficients $\nu_j(k)$ in formula (11) so that the error vector $\mathbf{y}^{(k)} - \mathbf{x}$ of the approximation $\mathbf{y}^{(k)}$ is smaller than the error vector $\mathbf{x}^{(k)} - \mathbf{x}$. In case of

$$\mathbf{x}^{(j)} = \mathbf{x} \quad (j = 1 : k),$$

it is natural to demand that $\mathbf{y}^{(k)} = \mathbf{x}$. This is just a case when

$$\sum_{j=0}^k \nu_j(k) = 1. \quad (12)$$

How do we choose the factors $\nu_j(k)$ so that they would satisfy (12), and the error vector were the shortest? Since

$$\mathbf{x}^{(k)} - \mathbf{x} = (M^{-1}N)^k \mathbf{e}^{(0)},$$

then

$$\begin{aligned}\mathbf{y}^{(k)} - \mathbf{x} &= \sum_{j=0}^k \nu_j(k)(\mathbf{x}^{(j)} - \mathbf{x}) = \sum_{j=0}^k \nu_j(k)(M^{-1}N)^j \mathbf{e}^{(0)} = \\ &= \sum_{j=0}^k \nu_j(k)G^j \mathbf{e}^{(0)} = p_k(G)\mathbf{e}^{(0)},\end{aligned}$$

where $G = M^{-1}N$ and

$$p_k(z) = \sum_{j=0}^k \nu_j(k)z^j.$$

From condition (12), it follows that $p_k(1) = 1$. In addition,

$$\|\mathbf{y}^{(k)} - \mathbf{x}\|_2 \leq \|p_k(G)\|_2 \|\mathbf{e}^{(0)}\|_2. \quad (13)$$

We confine ourselves further to the case of a symmetric matrix G . Let the eigenvalues λ_i of the symmetric matrix G satisfy the chain of inequalities

$$-1 < \alpha \leq \lambda_n \leq \dots \leq \lambda_1 \leq \beta < 1.$$

If λ is the eigenvalue corresponding to the eigenvector $\mathbf{x}$, then

$$p_k(G)\mathbf{x} = \sum_{j=0}^k \nu_j(k)G^j \mathbf{x} = \sum_{j=0}^k \nu_j(k)\lambda^j \mathbf{x},$$

i.e., the vector $\mathbf{x}$ is also the eigenvector of the matrix $p_k(G)$ corresponding to the eigenvalue

$$\sum_{j=0}^k \nu_j(k)\lambda^j = p_k(\lambda).$$

In the case of the symmetric matrix G , the matrix $p_k(G)$ is also symmetric. Hence

$$\|p_k(G)\|_2 = \max_{\lambda_i \in \lambda(G)} \left| \sum_{j=0}^k \nu_j(k)\lambda_i^j \right| \leq \max_{\lambda \in [\alpha, \beta]} |p_k(\lambda)|.$$

To decrease the quantity $\|p_k(G)\|_2$ one must find a polynomial $p_k(z)$ that has small values on the segment $[\alpha, \beta]$ and satisfies the condition $p_k(1) = 1$. The *Chebyshev* polynomials have these properties. The *Chebyshev* polynomials are defined on the segment $[-1; 1]$ by the recurrence relation

$$c_j(z) = 2zc_{j-1}(z) - c_{j-2}(z) \quad (j = 2 :),$$

while $c_0(z) = 1$ and $c_1(z) = z$. These polynomials satisfy the inequality

$$|c_j(z)| \leq 1 \quad z \in [-1; 1]$$

and $c_j(1) = 1$, and the values $|c_j(z)|$ grow quickly outside the segment $[-1; 1]$. The polynomial

$$p_k(z) = \frac{c_k(-1 + 2(z - \alpha)/(\beta - \alpha))}{c_k(\mu)}, \quad (14)$$

where

$$\mu = -1 + 2\frac{1 - \alpha}{\beta - \alpha} = 1 + 2\frac{1 - \beta}{\beta - \alpha} > 1,$$

satisfies the conditions $p_k(1) = 1$ and

$$|p_k(z)| \leq 1 \quad z \in [\alpha; \beta].$$

Taking into account relations (13) and (14), we find

$$\|\mathbf{y}^{(k)} - \mathbf{x}\|_2 \leq \frac{\|\mathbf{x}^{(k)} - \mathbf{x}\|_2}{|c_k(\mu)|}.$$

Therefore, the greater is μ , the greater is $|c_k(\mu)|$, and the greater will be the *Chebyshev's acceleration*.

2.8 Numeric Stability

In the following part we will consider how the deviations of the entries of the regular matrix $A \in \mathbf{R}^{n \times n}$ and vector $\mathbf{b} \in \mathbf{R}^n$ will cause the deviations of the solution $\mathbf{x}$ of the linear system

$$A\mathbf{x} = \mathbf{b}. \quad (1)$$

2.8.1 Singular Value Decomposition and Numeric Stability

Definition 8.1.1. The ε -rank of the matrix A is defined by the formula

$$\text{rank}(A, \varepsilon) = \min_{\|A-B\| \leq \varepsilon} \text{rank}(B).$$

Example 8.1.1.* Let us find the ε -rank of the matrix

$$A = \begin{bmatrix} 0.1 & 0.01 \\ -0.1 & 0.01 \end{bmatrix}$$

if $\varepsilon = 0.14$.

Evidently,

$$0 \leq \text{rank}(A, \varepsilon) \leq 2.$$

If the equality

$$\text{rank}(A, \varepsilon) = 0,$$

holds, then

$$\exists B : \text{rank}(B) = 0 \wedge \|A - B\|_2 \leq \varepsilon.$$

Since

$$\text{rank}(B) = 0 \Leftrightarrow B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

then

$$\begin{aligned} \|A - B\|_2 &= \left\| \begin{bmatrix} 0.1 & 0.01 \\ -0.1 & 0.01 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\|_2 = \\ &= \left\| \begin{bmatrix} 0.1 & 0.01 \\ -0.1 & 0.01 \end{bmatrix} \right\|_2 = \max \left\{ \sqrt{\lambda_1}, \sqrt{\lambda_2} \right\}, \end{aligned}$$

where λ_1 and λ_2 are the eigenvalues of the matrix

$$\begin{bmatrix} 0.1 & -0.1 \\ 0.01 & 0.01 \end{bmatrix} \begin{bmatrix} 0.1 & 0.01 \\ -0.1 & 0.01 \end{bmatrix} = \begin{bmatrix} 0.02 & 0 \\ 0 & 0.0002 \end{bmatrix}.$$

Therefore, $\lambda_1 = 0.02$ and $\lambda_2 = 0.0002$ and

$$\|A - B\|_2 = \sqrt{0.02} \approx 0.14142 > 0.14 = \varepsilon.$$

This is a contradiction, and it means that $\text{rank}(A, \varepsilon) > 0$. Let us show that $\text{rank}(A, \varepsilon) = 1$. Now, for the matrix

$$B = \begin{bmatrix} 0.1 & 0.01 \\ 0 & 0 \end{bmatrix}$$

$\text{rank}(B) = 1$ and

$$\begin{aligned} \|A - B\|_2 &= \left\| \begin{bmatrix} 0.1 & 0.01 \\ -0.1 & 0.01 \end{bmatrix} - \begin{bmatrix} 0.1 & 0.01 \\ 0 & 0 \end{bmatrix} \right\|_2 = \\ &= \left\| \begin{bmatrix} 0 & 0 \\ -0.1 & 0.01 \end{bmatrix} \right\|_2 = \max \left\{ \sqrt{\lambda_1}, \sqrt{\lambda_2} \right\}, \end{aligned}$$

where λ_1 and λ_2 are the eigenvalues of the matrix

$$\begin{bmatrix} 0 & -0.1 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ -0.1 & 0.01 \end{bmatrix} = \begin{bmatrix} .01 & -.001 \\ -.001 & .0001 \end{bmatrix}.$$

Thus, $\lambda_1 = 0.01$ and $\lambda_2 = 0.0001$, and

$$\|A - B\|_2 = \sqrt{0.01} \approx 0.1 < 0.14 = \varepsilon.$$

But this means that $\text{rank}(A, \varepsilon) = 1$.

Proposition 8.1.1. Let $A = U\Sigma V^T$ be the singular value decomposition of a matrix $A \in \mathbf{R}^{m \times n}$. If $k < r = \text{rank}(A)$ and

$$A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

then

$$\min_{\text{rank}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

Proof. Since

$$\begin{aligned} U^T A_k V &= U^T \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T V = \\ &= \begin{bmatrix} \mathbf{u}_1^T & \cdots & \mathbf{u}_m^T \end{bmatrix}^T \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{u}_1^T & \cdots & \mathbf{u}_m^T \end{bmatrix}^T \begin{bmatrix} \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_1 & \cdots & \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_n \end{bmatrix} = \end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_m^T \end{bmatrix} \begin{bmatrix} \sigma_1 \mathbf{u}_1 & \cdots & \sigma_k \mathbf{u}_k & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} = \\
&= \begin{bmatrix} \sigma_1 \mathbf{u}_1^T \mathbf{u}_1 & \cdots & \sigma_k \mathbf{u}_1^T \mathbf{u}_k & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 \mathbf{u}_m^T \mathbf{u}_1 & \cdots & \sigma_1 \mathbf{u}_m^T \mathbf{u}_k & 0 & \cdots & 0 \end{bmatrix} = \\
&= \text{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0) \in \mathbf{R}^{m \times n}
\end{aligned}$$

and

$$U^T(A - A_k)V = \text{diag}(0, \dots, 0, \sigma_{k+1}, \dots, \sigma_p),$$

while $p = \min\{m, n\}$. Since the Euclid norm of the matrix $A - A_k$ equals the greatest entry of the matrix $U^T(A - A_k)V$, then

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

Let $B \in \mathbf{R}^{m \times n}$ be a matrix for which $\text{rank}(B) = k$. We can find the orthonormal vectors $\mathbf{x}_1, \dots, \mathbf{x}_{n-k}$, such that the null space of the matrix B is a linear span of the vectors $\mathbf{x}_1, \dots, \mathbf{x}_{n-k}$, i.e.,

$$\mathcal{N}(B) = \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_{n-k}\}.$$

Since in the space $\mathbf{R}^n$ $n + 1$ vectors are linearly dependent, then

$$\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_{n-k}\} \cap \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_{k+1}\} \neq \{\mathbf{0}\}.$$

If $\mathbf{z}$ is a unit vector (by the Euclidean norm) from this intersection, then $B\mathbf{z} = \mathbf{0}$ and

$$\begin{aligned}
A\mathbf{z} &= \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^T \sum_{i=1}^{k+1} (\mathbf{v}_i^T \mathbf{z}) \mathbf{v}_i = \sum_{j=1}^r \sigma_j \mathbf{u}_j \sum_{i=1}^{k+1} (\mathbf{v}_i^T \mathbf{z}) (\mathbf{v}_j^T \mathbf{v}_i) = \\
&= \sum_{j=1}^r \sigma_j \mathbf{u}_j \sum_{i=1}^{k+1} (\mathbf{v}_i^T \mathbf{z}) \delta_{ij} = \sum_{i=1}^{k+1} \sigma_i (\mathbf{v}_i^T \mathbf{z}) \mathbf{u}_i.
\end{aligned}$$

Hence

$$\|A - B\|_2^2 \geq \|(A - B)\mathbf{z}\|_2^2 = \|A\mathbf{z}\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (\mathbf{v}_i^T \mathbf{z})^2 \geq$$

$$\geq \sum_{i=1}^{k+1} \sigma_{k+1}^2 (\mathbf{v}_i^T \mathbf{z})^2 = \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\mathbf{v}_i^T \mathbf{z})^2 = \sigma_{k+1}^2. \quad \square$$

Corollary 8.1.1. If the matrix $A \in \mathbf{R}^{n \times n}$ is regular, then the least singular value σ_n of the matrix A determines the distance of the matrix A from the nearest singular matrix.

Corollary 8.1.2. If $r_\varepsilon = \text{rank}(A, \varepsilon)$, then

$$\sigma_1 \geq \dots \geq \sigma_{r_\varepsilon} > \varepsilon \geq \sigma_{r_\varepsilon+1} \geq \dots \geq \sigma_p \quad (p = \min\{m, n\}).$$

Problem 8.1.1.* Use corollary 8.1.2 to solve the problem given in example 8.1.1.

Proposition 8.1.2. If

$$A = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^T = U \Sigma V^T$$

is the singular value decomposition of the regular matrix $A \in \mathbf{R}^{n \times n}$, then the solution $\mathbf{x}$ of system (1) can be expressed in the form

$$\mathbf{x} = A^{-1} \mathbf{b} = (U \Sigma V^T)^{-1} \mathbf{b} = \sum_{i=1}^n \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i. \quad (2)$$

Proof. Let us check the correctness of the assertion of the proposition:

$$\begin{aligned} \mathbf{x} = A^{-1} \mathbf{b} &= (U \Sigma V^T)^{-1} \mathbf{b} = V \Sigma^+ U^T \mathbf{b} = \\ &= \sum_{i=1}^n \mathbf{v}_i \mathbf{u}_i^T \mathbf{b} / \sigma_i = \sum_{i=1}^n \mathbf{v}_i (\mathbf{u}_i^T \mathbf{b}) / \sigma_i = \sum_{i=1}^n \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i. \end{aligned}$$

Corollary 8.1.3. From the representation of the solution in form (2) it appears that small deviations of the entries of the matrix A could cause great deviations of the solution $\mathbf{x}$ if σ_n is small.

Example 8.1.2.* For which system

$$\begin{bmatrix} 54/125 & -169/250 \\ 53/125 & -54/125 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

or

$$\begin{bmatrix} 41/16250 & -99/13000 \\ 46/15101 & -17/3250 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

small deviations of the entries of the system matrix can cause greater deviations of the solution $\mathbf{x}$? Applying the package “Maple”, we find the singular value decomposition of both system matrices:

$$\begin{bmatrix} 54/125 & -169/250 \\ 53/125 & -54/125 \end{bmatrix} = \begin{bmatrix} -.8 & -.6 \\ -.6 & .8 \end{bmatrix} \begin{bmatrix} 1.0 & 0 \\ 0 & .1 \end{bmatrix} \begin{bmatrix} -.6 & .8 \\ .8 & .6 \end{bmatrix}$$

and

$$\begin{bmatrix} 41/16250 & -99/13000 \\ 46/15101 & -17/3250 \end{bmatrix} = \begin{bmatrix} -.8 & -.6 \\ -.6 & .8 \end{bmatrix} \begin{bmatrix} .01 & 0 \\ 0 & .001 \end{bmatrix} \begin{bmatrix} -.38462 & .92308 \\ .92308 & .38462 \end{bmatrix}.$$

We see that the least singular value σ_2 of the first system matrix is hundred times smaller than the least singular value of the second system matrix. Therefore, in virtue of corollary 8.1.3, we can state that the first system is more stable than the second one, i.e., small deviations of the entries of the second system matrix can cause greater deviations of the solution $\mathbf{x}$ than the deviation of the same order of the entries of the first system matrix.

Problem 8.1.2.* Which of the two systems

$$\begin{bmatrix} 5400/169 & -3940/169 \\ 14650/169 & -5400/169 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 7 \end{bmatrix}$$

or

$$\begin{bmatrix} 141/845 & -841/850 \\ 291/676 & -141/845 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 7 \end{bmatrix}$$

is more stable?

2.8.2 Taylor Development

We can obtain the exact evaluation of the sensibility of system (1) using the system depending on a parameter

$$(A + \delta F)\mathbf{x}(\delta) = \mathbf{b} + \delta \mathbf{f}, \quad (3)$$

where $F \in \mathbf{R}^{n \times n}$, $\mathbf{f} \in \mathbf{R}^n$ and $\mathbf{x}(0) = \mathbf{x}$. If A is a regular matrix, then $\mathbf{x}(\delta)$ is a differentiable function of the parameter δ in some neighbourhood of the value 0. Differentiating the both sides of equality (3) with respect to the parameter δ , we get

$$F\mathbf{x}(\delta) + (A + \delta F)\frac{d\mathbf{x}}{d\delta}(\delta) = \mathbf{f}$$

and

$$\frac{d\mathbf{x}}{d\delta}(\delta) = (A + \delta F)^{-1}(\mathbf{f} - F\mathbf{x}(\delta)) \quad (4)$$

It follows from relation (4) that

$$\frac{d\mathbf{x}}{d\delta}(0) = A^{-1}(\mathbf{f} - F\mathbf{x}).$$

Let us write the first order Taylor formula for the function $\mathbf{x}(\delta)$

$$\mathbf{x}(\delta) = \mathbf{x} + \delta \frac{d\mathbf{x}}{d\delta}(0) + O(\delta^2) = \mathbf{x} + \delta A^{-1}(\mathbf{f} - F\mathbf{x}) + O(\delta^2).$$

As the result, we obtain for arbitrary vector norm and for the matrix norm corresponding to it

$$\begin{aligned} \frac{\|\mathbf{x}(\delta) - \mathbf{x}\|}{\|\mathbf{x}\|} &= \frac{\|(\delta \frac{d\mathbf{x}}{d\delta}(0) + O(\delta^2))\|}{\|\mathbf{x}\|} = \frac{\|\delta A^{-1}(\mathbf{f} - F\mathbf{x}) + O(\delta^2)\|}{\|\mathbf{x}\|} \leq \\ &\leq |\delta| \frac{\|A^{-1}\| (\|\mathbf{f}\| + \|F\| \cdot \|\mathbf{x}\|)}{\|\mathbf{x}\|} + O(\delta^2) \leq |\delta| \|A^{-1}\| \left(\frac{\|\mathbf{f}\|}{\|\mathbf{x}\|} + \|F\| \right) + O(\delta^2). \end{aligned}$$

Taking into account that from relation (1) it follows $\|\mathbf{b}\| \leq \|A\| \|\mathbf{x}\|$, we obtain the estimate

$$\frac{\|\mathbf{x}(\delta) - \mathbf{x}\|}{\|\mathbf{x}\|} \leq |\delta| \|A^{-1}\| \|A\| \left(\frac{\|\mathbf{f}\|}{\|\mathbf{b}\|} + \frac{\|F\|}{\|A\|} \right) + O(\delta^2),$$

or

$$\frac{\|\mathbf{x}(\delta) - \mathbf{x}\|}{\|\mathbf{x}\|} \leq k(A)(\varepsilon_{rel}(A) + \varepsilon_{rel}(\mathbf{b})) + O(\delta^2),$$

where $\varepsilon_{rel}(A) = |\delta| \frac{\|F\|}{\|A\|}$ and $\varepsilon_{rel}(\mathbf{b}) = |\delta| \frac{\|\mathbf{f}\|}{\|\mathbf{b}\|}$ are the relative errors of the matrix A and the vector $\mathbf{b}$, respectively.

Proposition 8.2.1. If $A \in \mathbf{R}^{n \times n}$ is a regular matrix, then the relative error $\varepsilon_{rel}(\mathbf{x})$ of the solution of the linear system (1) corresponding to the relative error $\varepsilon_{rel}(A)$ of the matrix A and the relative error $\varepsilon_{rel}(\mathbf{b})$, of the vector $\mathbf{b}$ is given by

$$\varepsilon_{rel}(\mathbf{x}) \approx k(A)(\varepsilon_{rel}(A) + \varepsilon_{rel}(\mathbf{b})).$$

Corollary 8.2.1. In case of the Euclidean norm the estimate the

$$\varepsilon_{rel}(\mathbf{x}) \approx \frac{\sigma_1}{\sigma_n}(\varepsilon_{rel}(A) + \varepsilon_{rel}(\mathbf{b})).$$

holds.

Proof. The relation $\|A\|_2 = \sigma_1$ holds. As the matrix A is regular, it follows from its singular value decomposition $A = U\Sigma V^T$ that $A^{-1} = V\Sigma^+ U^T$, where $\Sigma^+ = \text{diag}(1/\sigma_1, \dots, 1/\sigma_n)$. Since $\max_{1 \leq i \leq n} 1/\sigma_i = 1/\sigma_n$, then $\|A^{-1}\|_2 = 1/\sigma_n$ and $k_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sigma_1/\sigma_n$. $\square$

Example 8.2.1.* Let us estimate the relative error $\varepsilon_{rel}(\mathbf{x})$ of the solution $\mathbf{x}$ of the system $A\mathbf{x} = \mathbf{b}$ in the case of the Euclidean norm if

$$A = \begin{bmatrix} 80 & 18 & 24 \\ 60 & -24 & -32 \\ 0 & 4/5 & -3/5 \end{bmatrix},$$

$$\varepsilon_{rel}(A) = 0.09 \text{ and } \varepsilon_{rel}(\mathbf{b}) = 0.01.$$

Let us find the singular decomposition of the matrix A

$$\begin{aligned} A &= \begin{bmatrix} 80 & 18 & 24 \\ 60 & -24 & -32 \\ 0 & 4/5 & -3/5 \end{bmatrix} = \\ &= \begin{bmatrix} -.8 & .6 & 0 \\ -.6 & -.8 & 0 \\ 0 & 0 & 1.0 \end{bmatrix} \begin{bmatrix} 100.0 & 0 & 0 \\ 0 & 50.0 & 0 \\ 0 & 0 & 1.0 \end{bmatrix} \begin{bmatrix} -1.0 & 0 & 0 \\ 0 & .6 & .8 \\ 0 & .8 & -.6 \end{bmatrix}. \end{aligned}$$

In virtue of corollary 8.2.1, we can state

$$\varepsilon_{rel}(\mathbf{x}) \approx \frac{\sigma_1}{\sigma_3}(\varepsilon_{rel}(A) + \varepsilon_{rel}(\mathbf{b})) = \frac{100}{1}(0.09 + 0.01) = 10.$$

Problem 8.2.1.* Estimate the relative error $\varepsilon_{rel}(\mathbf{x})$ of the solution $\mathbf{x}$ of the system $A\mathbf{x} = \mathbf{b}$ in case of the Euclidean norm if

$$A = \begin{bmatrix} 65 & -144/13 & 60/13 \\ 156 & 60/13 & -25/13 \\ 0 & 5/13 & 12/13 \end{bmatrix},$$

$$\varepsilon_{rel}(A) = 0.008 \text{ and } \varepsilon_{rel}(\mathbf{b}) = 0.002.$$

Remark 8.2.1. Kahan (1966) proved that

$$\frac{1}{k_p(A)} = \min_{A+\Delta A \text{ singular}} \frac{\|\Delta A\|_p}{\|A\|_p}$$

and Rice (1966) proved that

$$k(A) = \lim_{\varepsilon \rightarrow 0} \sup_{\|\Delta A\| \leq \varepsilon \|A\|} \frac{\|(A + \Delta A)^{-1} - A^{-1}\|}{\varepsilon} \cdot \frac{1}{\|A^{-1}\|}.$$

2.8.3 Strict Estimations

The assertion of proposition 8.2.1 is of a local kind. Namely, this proposition is proved for relatively small deviations. Next we will give the evaluation of the deviation of the solution in general case. First we will prove one necessary auxiliary result.

Proposition 8.3.1. If the matrix $A \in \mathbf{R}^{n \times n}$ is a regular matrix and

$$r \equiv \|A^{-1}E\|_p < 1,$$

then $A + E$ is regular and

$$\|(A + E)^{-1} - A^{-1}\|_p \leq \frac{\|E\|_p \|A^{-1}\|_p^2}{1 - r}.$$

Proof. A regular matrix can be represented in the form

$$A + E = A(I - F),$$

where $F = -A^{-1}E$. Since $\|F\|_p = r < 1$, then, in virtue of proposition 7.1.1, the matrix $I - F$ is regular and

$$\|(I - F)^{-1}\|_p < \frac{1}{1 - r}.$$

Hence

$$(A + E)^{-1} = (I - F)^{-1}A^{-1}$$

and

$$\|(A + E)^{-1}\|_p \leq \frac{\|A^{-1}\|_p}{1 - r}.$$

From the equality

$$B^{-1} = A^{-1} - B^{-1}(B - A)A^{-1}$$

it follows that

$$(A + E)^{-1} - A^{-1} = -A^{-1}E(A + E)^{-1}$$

and

$$\|(A + E)^{-1} - A^{-1}\|_p \leq \|A^{-1}\|_p \|E\|_p \|(A + E)^{-1}\|_p \leq \frac{\|A^{-1}\|_p^2 \|E\|_p}{1 - r}. \quad \square$$

Proposition 8.3.2. Let

$$\begin{aligned} Ax &= \mathbf{b}, & A &\in \mathbf{R}^{n \times n}, \mathbf{0} \neq \mathbf{b} \in \mathbf{R}^n, \\ (A + \Delta A)\mathbf{y} &= \mathbf{b} + \Delta \mathbf{b}, & \Delta A &\in \mathbf{R}^{n \times n}, \Delta \mathbf{b} \in \mathbf{R}^n, \end{aligned}$$

while $\|\Delta A\| \leq \delta \|A\|$ and $\|\Delta \mathbf{b}\| \leq \delta \|\mathbf{b}\|$. If $\delta \cdot k(A) = r < 1$, then $A + \Delta A$ is a regular matrix, and

$$\frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} \leq \frac{1 + r}{1 - r}$$

and

$$\frac{\|\mathbf{y} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{2\delta}{1 - r}k(A).$$

Proof. Since $\|A^{-1}\Delta A\| \leq \delta \|A^{-1}\| \|A\| = r < 1$, then, in virtue of proposition 8.3.1, $A + \Delta A$ is a regular matrix. Applying proposition 7.1.1 and the equality

$$(I + A^{-1}\Delta A)\mathbf{y} = \mathbf{x} + A^{-1}\Delta \mathbf{b},$$

we find

$$\|\mathbf{y}\| \leq \|(I + A^{-1}\Delta A)^{-1}\| (\|\mathbf{x}\| + \delta \|A^{-1}\| \|\mathbf{b}\|) \leq$$

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

In addition, we find that $\|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$. Hence

$$\|\mathbf{y}\| \leq \frac{1}{1-r} (\|\mathbf{x}\| + r \|\mathbf{x}\|),$$

and that means that the first part of the assertion is true. Since the relation

$$\mathbf{y} - \mathbf{x} = A^{-1}\Delta\mathbf{b} - A^{-1}\Delta\mathbf{y},$$

holds, then

$$\|\mathbf{y} - \mathbf{x}\| \leq \delta \|A^{-1}\| \|\mathbf{b}\| + \delta \|A^{-1}\| \|A\| \|\mathbf{y}\|$$

and therefore,

$$\begin{aligned} \frac{\|\mathbf{y} - \mathbf{x}\|}{\|\mathbf{x}\|} &\leq \delta k(A) \frac{\|\mathbf{b}\|}{\|A\| \|\mathbf{x}\|} + \delta k(A) \frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} \leq \\ &\leq \delta k(A) \left(1 + \frac{1+r}{1-r} \right) = \frac{2\delta}{1-r} k(A). \quad \square \end{aligned}$$

Example 8.3.1.* Let

$$A = \begin{bmatrix} 52.8 & 60.4 \\ 29.6 & 52.8 \end{bmatrix} \wedge \|\Delta A\|_2 \leq 8 \wedge \mathbf{b} = \begin{bmatrix} 40 \\ 30 \end{bmatrix} \wedge \|\Delta\mathbf{b}\|_2 \leq 4.$$

We will estimate the relative error of the solution of the system $A\mathbf{x} = \mathbf{b}$ using the Euclidean norm. Since

$$A^T A = \begin{bmatrix} 52.8 & 29.6 \\ 60.4 & 52.8 \end{bmatrix} \begin{bmatrix} 52.8 & 60.4 \\ 29.6 & 52.8 \end{bmatrix} = \begin{bmatrix} 3664 & 4752 \\ 4752 & 6436 \end{bmatrix}$$

and the eigenvalues $A^T A$ are $\lambda_1 = 10000$ and $\lambda_2 = 100$, then

$$\|A\|_2 = \max \left\{ \sqrt{\lambda_1}, \sqrt{\lambda_2} \right\} = 100.$$

Let us find the Euclidean norm of the vector $\mathbf{b}$:

$$\|\mathbf{b}\|_2 = \sqrt{40^2 + 30^2} = 50.$$

If one takes $\delta = 0.08$, then the conditions $\|\Delta A\| \leq \delta \|A\|$ and $\|\Delta \mathbf{b}\| \leq \delta \|\mathbf{b}\|$ are satisfied. From the singular value decomposition of the matrix A

$$\begin{bmatrix} 52.8 & 60.4 \\ 29.6 & 52.8 \end{bmatrix} = \begin{bmatrix} -.8 & -.6 \\ -.6 & .8 \end{bmatrix} \begin{bmatrix} 100.0 & 0 \\ 0 & 10.0 \end{bmatrix} \begin{bmatrix} -.6 & -.8 \\ -.8 & .6 \end{bmatrix}$$

we find that

$$k_2(A) = \frac{\sigma_1}{\sigma_2} = \frac{100}{10} = 10.$$

Hence

$$r = \delta \cdot k_2(A) = 0.08 \cdot 10 = 0.8 < 1,$$

and, in virtue of proposition 8.3.2, we get

$$\frac{\|\Delta \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{2\delta}{1-r} k_2(A) = \frac{2 \cdot 0.8}{1-0.8} \cdot 10 = 8.$$

Problem 8.3.1.* Let

$$A = \begin{bmatrix} 65 & -12 \\ -156 & -5 \end{bmatrix} \wedge \|\Delta A\|_2 \leq 169/15 \wedge \mathbf{b} = \begin{bmatrix} 120 \\ 50 \end{bmatrix} \wedge \|\Delta \mathbf{b}\|_2 \leq 13/15.$$

Find the relative error of the solution of the system $A\mathbf{x} = \mathbf{b}$. Use the Euclidean norm.

References

- G.H.Golub and C.F.Van Loan (1996). Matrix Computations, John Hopkins University Press, London.
- G.Kangro (1962). Higher Algebra, Estonian State Publishers, Tallinn (in Estonian).
- P.Lankaster (1982). Theory of Matrices, Nauka, Moscow (in Russian).
- E.Oja and P.Oja (1991). Functional Analysis, Tartu University Press, Tartu (in Estonian).
- G.Strang (1988). Linear Algebra and its Applications, Harcourt Brace Jovanovich College Publishers, Orlando, Florida.