# Linear Algebra Done Wrong

Sergei Treil

Department of Mathematics, Brown University

# Preface

The title of the book sounds a bit mysterious. Why should anyone read this book if it presents the subject in a wrong way? What is particularly done "wrong" in the book?

Before answering these questions, let me first describe the target audience of this text. This book appeared as lecture notes for the course "Honors Linear Algebra". It supposed to be a *first* linear algebra course for mathematically advanced students. It is intended for a student who, while not yet very familiar with abstract reasoning, is willing to study more rigorous mathematics than what is presented in a "cookbook style" calculus type course. Besides being a first course in linear algebra it is also supposed to be a first course introducing a student to *rigorous* proof, formal definitions—in short, to the style of modern theoretical (abstract) mathematics. The target audience explains the very specific blend of elementary ideas and concrete examples, which are usually presented in introductory linear algebra texts with more abstract definitions and constructions typical for advanced books.

Another specific of the book is that it is not written by or for an algebraist. So, I tried to emphasize the topics that are important for analysis, geometry, probability, etc., and did not include some traditional topics. For example, I am only considering vector spaces over the fields of real or complex numbers. Linear spaces over other fields are not considered at all, since I feel time required to introduce and explain abstract fields would be better spent on some more classical topics, which will be required in other disciplines. And later, when the students study general fields in an abstract algebra course they will understand that many of the constructions studied in this book will also work for general fields.

Also, I treat only finite-dimensional spaces in this book and a basis always means a finite basis. The reason is that it is impossible to say something non-trivial about infinite-dimensional spaces without introducing convergence, norms, completeness etc., i.e. the basics of functional analysis. And this is definitely a subject for a separate course (text). So, I do not consider infinite Hamel bases here: they are not needed in most applications to analysis and geometry, and I feel they belong in an abstract algebra course.

**Notes for the instructor.** There are several details that distinguish this text from standard advanced linear algebra textbooks. First concerns the definitions of bases, linearly independent, and generating sets. In the book I first define a basis as a system with the property that any vector admits a unique representation as a linear combination. And then linear independence and generating system properties appear naturally as halves of the basis property, one being uniqueness and the other being existence of the representation.

The reason for this approach is that I feel the concept of a basis is a much more important notion than linear independence: in most applications we really do not care about linear independence, we need a system to be a basis. For example, when solving a homogeneous system, we are not just looking for linearly independent solutions, but for the correct number of linearly independent solutions, i.e. for a basis in the solution space.

And it is easy to explain to students, why bases are important: they allow us to introduce coordinates, and work with $\mathbb{R}^n$ (or $\mathbb{C}^n$) instead of working with an abstract vector space. Furthermore, we need coordinates to perform computations using computers, and computers are well adapted to working with matrices. Also, I really do not know a simple motivation for the notion of linear independence.

Another detail is that I introduce linear transformations before teaching how to solve linear systems. A disadvantage is that we did not prove until Chapter 2 that only a square matrix can be invertible as well as some other important facts. However, having already defined linear transformation allows more systematic presentation of row reduction. Also, I spend a lot of time (two sections) motivating matrix multiplication. I hope that I explained well why such a strange looking rule of multiplication is, in fact, a very natural one, and we really do not have any choice here.

Many important facts about bases, linear transformations, etc., like the fact that any two bases in a vector space have the same number of vectors, are proved in Chapter 2 by counting pivots in the row reduction. While most of these facts have "coordinate free" proofs, formally not involving Gaussian

elimination, a careful analysis of the proofs reveals that the Gaussian elimination and counting of the pivots do not disappear, they are just hidden in most of the proofs. So, instead of presenting very elegant (but not easy for a beginner to understand) "coordinate-free" proofs, which are typically presented in advanced linear algebra books, we use "row reduction" proofs, more common for the "calculus type" texts. The advantage here is that it is easy to see the common idea behind all the proofs, and such proofs are easier to understand and to remember for a reader who is not very mathematically sophisticated.

I also present in Section 8 of Chapter 2 a simple and easy to remember formalism for the change of basis formula.

Chapter 3 deals with determinants. I spent a lot of time presenting a motivation for the determinant, and only much later give formal definitions. Determinants are introduced as a way to compute volumes. It is shown that if we allow signed volumes, to make the determinant linear in each column (and at that point students should be well aware that the linearity helps a lot, and that allowing negative volumes is a very small price to pay for it), and assume some very natural properties, then we do not have any choice and arrive to the classical definition of the determinant. I would like to emphasize that initially I do not postulate antisymmetry of the determinant; I deduce it from other very natural properties of volume.

Note, that while formally in Chapters 1–3 I was dealing mainly with real spaces, everything there holds for complex spaces, and moreover, even for the spaces over arbitrary fields.

Chapter 4 is an introduction to spectral theory, and that is where the complex space $\mathbb{C}^n$ naturally appears. It was formally defined in the beginning of the book, and the definition of a complex vector space was also given there, but before Chapter 4 the main object was the real space $\mathbb{R}^n$. Now the appearance of complex eigenvalues shows that for spectral theory the most natural space is the complex space $\mathbb{C}^n$, even if we are initially dealing with real matrices (operators in real spaces). The main accent here is on the diagonalization, and the notion of a basis of eigesnspaces is also introduced.

Chapter 5 dealing with inner product spaces comes after spectral theory, because I wanted to do both the complex and the real cases simultaneously, and spectral theory provides a strong motivation for complex spaces. Other then the motivation, Chapters 4 and 5 do not depend on each other, and an instructor may do Chapter 5 first.

Although I present the Jordan canonical form in Chapter 9, I usually do not have time to cover it during a one-semester course. I prefer to spend more time on topics discussed in Chapters 6 and 7 such as diagonalization

of normal and self-adjoint operators, polar and singular values decomposition, the structure of orthogonal matrices and orientation, and the theory of quadratic forms.

I feel that these topics are more important for applications, then the Jordan canonical form, despite the definite beauty of the latter. However, I added Chapter 9 so the instructor may skip some of the topics in Chapters 6 and 7 and present the Jordan Decomposition Theorem instead.

I also included (new for 2009) Chapter 8, dealing with dual spaces and tensors. I feel that the material there, especially sections about tensors, is a bit too advanced for a first year linear algebra course, but some topics (for example, change of coordinates in the dual space) can be easily included in the syllabus. And it can be used as an introduction to tensors in a more advanced course. Note, that the results presented in this chapter are true for an arbitrary field.

I had tried to present the material in the book rather informally, preferring intuitive geometric reasoning to formal algebraic manipulations, so to a purist the book may seem not sufficiently rigorous. Throughout the book I usually (when it does not lead to the confusion) identify a linear transformation and its matrix. This allows for a simpler notation, and I feel that overemphasizing the difference between a transformation and its matrix may confuse an inexperienced student. Only when the difference is crucial, for example when analyzing how the matrix of a transformation changes under the change of the basis, I use a special notation to distinguish between a transformation and its matrix.

# Contents

*Chapter 1*

# Basic Notions

## 1. Vector spaces

A vector space $V$ is a collection of objects, called vectors (denoted in this book by lowercase bold letters, like $\mathbf{v}$), along with two operations, addition of vectors and multiplication by a number (scalar) [1] , such that the following 8 properties (the so-called *axioms* of a vector space) hold:

The first 4 properties deal with the addition:

1. Commutativity: $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$ for all $\mathbf{v}, \mathbf{w} \in V$;
2. Associativity: $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$;
3. Zero vector: there exists a special vector, denoted by $\mathbf{0}$ such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for all $\mathbf{v} \in V$;
4. Additive inverse: For every vector $\mathbf{v} \in V$ there exists a vector $\mathbf{w} \in V$ such that $\mathbf{v} + \mathbf{w} = \mathbf{0}$. Such additive inverse is usually denoted as $-\mathbf{v}$;

   *The next two properties concern multiplication:*

5. Multiplicative identity: $1\mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in V$;

A question arises, "How one can memorize the above properties?" And the answer is that one does not need to, see below!

[1] We need some visual distinction between vectors and other objects, so in this book we use bold lowercase letters for vectors and regular lowercase letters for numbers (scalars). In some (more advanced) books Latin letters are reserved for vectors, while Greek letters are used for scalars; in even more advanced texts any letter can be used for anything and the reader must understand from the context what each symbol means. I think it is helpful, especially for a beginner to have some visual distinction between different objects, so a bold lowercase letters will always denote a vector. And on a blackboard an arrow (like in $\vec{v}$) is used to identify a vector.

6. Multiplicative associativity: $(\alpha\beta)\mathbf{v} = \alpha(\beta\mathbf{v})$ for all $\mathbf{v} \in V$ and all scalars $\alpha$, $\beta$;

   *And finally, two distributive properties, which connect multiplication and addition:*

7. $\alpha(\mathbf{u}+\mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$ for all $\mathbf{u}, \mathbf{v} \in V$ and all scalars $\alpha$;
8. $(\alpha+\beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$ for all $\mathbf{v} \in V$ and all scalars $\alpha$, $\beta$.

**Remark.** The above properties seem hard to memorize, but it is not necessary. They are simply the familiar rules of algebraic manipulations with numbers, that you know from high school. The only new twist here is that you have to understand what operations you can apply to what objects. You can add vectors, and you can multiply a vector by a number (scalar). Of course, you can do with number all possible manipulations that you have learned before. But, you cannot multiply two vectors, or add a number to a vector.

**Remark.** It is easy to prove that zero vector $\mathbf{0}$ is unique, and that given $\mathbf{v} \in V$ its additive inverse $-\mathbf{v}$ is also unique.

It is also not hard to show using properties 5, 6 and 8 that $\mathbf{0} = 0\mathbf{v}$ for any $\mathbf{v} \in V$, and that $-\mathbf{v} = (-1)\mathbf{v}$. Note, that to do this one still needs to use other properties of a vector space in the proofs, in particular properties 3 and 4.

If the scalars are the usual real numbers, we call the space $V$ a *real* vector space. If the scalars are the complex numbers, i.e. if we can multiply vectors by complex numbers, we call the space $V$ a *complex* vector space.

Note, that any complex vector space is a real vector space as well (if we can multiply by complex numbers, we can multiply by real numbers), but not the other way around.

If you do not know what a field is, do not worry, since in this book we consider only the case of real and complex spaces.

It is also possible to consider a situation when the scalars are elements of an arbitrary field $\mathbb{F}$. In this case we say that $V$ is a vector space over the field $\mathbb{F}$. Although many of the constructions in the book (in particular, everything in Chapters 1–3) work for general fields, in this text we are considering only real and complex vector spaces.

If we do not specify the set of scalars, or use a letter $\mathbb{F}$ for it, then the results are true for both real and complex spaces. If we need to distinguish real and complex cases, we will explicitly say which case we are considering.

Note, that in the definition of a vector space over an arbitrary field, we *require* the set of scalars to be a field, so we can always divide (without a remainder) by a non-zero scalar. Thus, it is possible to consider vector space over rationals, but not over the integers.

### 1.1. Examples.

**Example.** The space $\mathbb{R}^n$ consists of all columns of size $n$,

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

whose entries are real numbers. Addition and multiplication are defined entrywise, i.e.

$$\alpha \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \alpha v_1 \\ \alpha v_2 \\ \vdots \\ \alpha v_n \end{pmatrix}, \qquad \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_n \end{pmatrix}$$

**Example.** The space $\mathbb{C}^n$ also consists of columns of size $n$, only the entries now are complex numbers. Addition and multiplication are defined exactly as in the case of $\mathbb{R}^n$, the only difference is that we can now multiply vectors by *complex* numbers, i.e. $\mathbb{C}^n$ is a *complex* vector space.

Many results in this text are true for both $\mathbb{R}^n$ and $\mathbb{C}^n$. In such cases we will use notation $\mathbb{F}^n$.

**Example.** The space $M_{m\times n}$ (also denoted as $M_{m,n}$) of $m \times n$ matrices: the addition and multiplication by scalars are defined entrywise. If we allow only real entries (and so only multiplication only by reals), then we have a real vector space; if we allow complex entries and multiplication by complex numbers, we then have a complex vector space.

Formally, we have to distinguish between between real and complex cases, i.e. write something like $M^{\mathbb{R}}_{m,n}$ or $M^{\mathbb{C}}_{m,n}$. However, in most situations there is no difference between real and complex case, and there is no need to specify which case we are considering. If there is a difference we say explicitly which case we are considering.

**Remark.** As we mentioned above, the axioms of a vector space are just the familiar rules of algebraic manipulations with (real or complex) numbers, so if we put scalars (numbers) for the vectors, all axioms will be satisfied. Thus, the set $\mathbb{R}$ of real numbers is a real vector space, and the set $\mathbb{C}$ of complex numbers is a complex vector space.

More importantly, since in the above examples all vector operations (addition and multiplication by a scalar) are performed entrywise, for these examples the axioms of a vector space are automatically satisfied because they are satisfied for scalars (can you see why?). So, we do not have to

check the axioms, we get the fact that the above examples are indeed vector spaces for free!

The same can be applied to the next example, the coefficients of the polynomials play the role of entries there.

**Example.** The space $\mathbb{P}_n$ of polynomials of degree at most $n$, consists of all polynomials $p$ of form

$$p(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n,$$

where $t$ is the independent variable. Note, that some, or even all, coefficients $a_k$ can be 0.

In the case of real coefficients $a_k$ we have a real vector space, complex coefficient give us a complex vector space. Again, we will specify whether we treating real or complex case only when it is essential; otherwise everything applies to both cases.

**Question:** What are zero vectors in each of the above examples?

**1.2. Matrix notation.** An $m \times n$ matrix is a rectangular array with $m$ rows and $n$ columns. Elements of the array are called *entries* of the matrix.

It is often convenient to denote matrix entries by indexed letters: the first index denotes the number of the row, where the entry is, and the second one is the number of the column. For example

$$A = (a_{j,k})_{j=1,\ k=1}^{m,\ n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \ldots & a_{m,n} \end{pmatrix} \tag{1.1}$$

is a general way to write an $m \times n$ matrix.

Very often for a matrix $A$ the entry in row number $j$ and column number $k$ is denoted by $A_{j,k}$ or $(A)_{j,k}$, and sometimes as in example (1.1) above the same letter but in lowercase is used for the matrix entries.

Given a matrix $A$, its *transpose* (or transposed matrix) $A^T$, is defined by transforming the rows of $A$ into the columns. For example

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

So, the columns of $A^T$ are the rows of $A$ and vice versa, the rows of $A^T$ are the columns of $A$.

The formal definition is as follows: $(A^T)_{j,k} = (A)_{k,j}$ meaning that the entry of $A^T$ in the row number $j$ and column number $k$ equals the entry of $A$ in the row number $k$ and row number $j$.

The transpose of a matrix has a very nice interpretation in terms of linear transformations, namely it gives the so-called *adjoint* transformation. We will study this in detail later, but for now transposition will be just a useful formal operation.

One of the first uses of the transpose is that we can write a column vector $\mathbf{x} \in \mathbb{F}^n$ (recall that $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$) as $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$. If we put the column vertically, it will use significantly more space.

**Exercises.**

**1.1.** Let $\mathbf{x} = (1, 2, 3)^T$, $\mathbf{y} = (y_1, y_2, y_3)^T$, $\mathbf{z} = (4, 2, 1)^T$. Compute $2\mathbf{x}$, $3\mathbf{y}$, $\mathbf{x} + 2\mathbf{y} - 3\mathbf{z}$.

**1.2.** Which of the following sets (with natural addition and multiplication by a scalar) are vector spaces. Justify your answer.

a) The set of all continuous functions on the interval $[0, 1]$;

b) The set of all non-negative functions on the interval $[0, 1]$;

c) The set of all polynomials of degree *exactly* $n$;

d) The set of all symmetric $n \times n$ matrices, i.e. the set of matrices $A = \{a_{j,k}\}_{j,k=1}^n$ such that $A^T = A$.

**1.3.** True or false:

a) Every vector space contains a zero vector;

b) A vector space can have more than one zero vector;

c) An $m \times n$ matrix has $m$ rows and $n$ columns;

d) If $f$ and $g$ are polynomials of degree $n$, then $f + g$ is also a polynomial of degree $n$;

e) If $f$ and $g$ are polynomials of degree at most $n$, then $f + g$ is also a polynomial of degree at most $n$

**1.4.** Prove that a zero vector $\mathbf{0}$ of a vector space $V$ is unique.

**1.5.** What matrix is the zero vector of the space $M_{2\times 3}$?

**1.6.** Prove that the additive inverse, defined in Axiom 4 of a vector space is unique.

**1.7.** Prove that $0\mathbf{v} = \mathbf{0}$ for any vector $\mathbf{v} \in V$.

**1.8.** Prove that for any vector $\mathbf{v}$ its additive inverse $-\mathbf{v}$ is given by $(-1)\mathbf{v}$.

## 2. Linear combinations, bases.

Let $V$ be a vector space, and let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p \in V$ be a collection of vectors. A *linear combination* of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ is a sum of form

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_p\mathbf{v}_p = \sum_{k=1}^{p} \alpha_k\mathbf{v}_k.$$

**Definition 2.1.** A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_n \in V$ is called a basis (for the vector space $V$) if any vector $\mathbf{v} \in V$ admits a *unique* representation as a linear combination

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_n\mathbf{v}_n = \sum_{k=1}^{n} \alpha_k\mathbf{v}_k.$$

The coefficients $\alpha_1, \alpha_2, \dots, \alpha_n$ are called *coordinates* of the vector $\mathbf{v}$ (in the basis, or with respect to the basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$).

Another way to say that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis is to say that for any possible choice of the right side $\mathbf{v}$, the equation $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_m\mathbf{v}_n = \mathbf{v}$ (with unknowns $x_k$) has a unique solution.

Before discussing any properties of bases[2], let us give a few examples, showing that such objects exist, and that it makes sense to study them.

**Example 2.2.** In the first example the space $V$ is $\mathbb{F}^n$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Consider vectors

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

(the vector $\mathbf{e}_k$ has all entries 0 except the entry number $k$, which is 1). The system of vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is a basis in $\mathbb{F}^n$. Indeed, any vector

$$\mathbf{v} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{F}^n$$

can be represented as the linear combination

$$\mathbf{v} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots x_n\mathbf{e}_n = \sum_{k=1}^{n} x_k\mathbf{e}_k$$

[2] the plural for the "basis" is *bases*, the same as the plural for "base"

and this representation is unique. The system $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n \in \mathbb{F}^n$ is called *the standard basis* in $\mathbb{F}^n$

**Example 2.3.** In this example the space is the space $\mathbb{P}_n$ of the polynomials of degree at most $n$. Consider vectors (polynomials) $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n \in \mathbb{P}_n$ defined by

$$\mathbf{e}_0 := 1, \quad \mathbf{e}_1 := t, \quad \mathbf{e}_2 := t^2, \quad \mathbf{e}_3 := t^3, \quad \ldots, \quad \mathbf{e}_n := t^n.$$

Clearly, any polynomial $p$, $p(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n$ admits a unique representation

$$p = a_0\mathbf{e}_0 + a_1\mathbf{e}_1 + \ldots + a_n\mathbf{e}_n.$$

So the system $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n \in \mathbb{P}_n$ is a basis in $\mathbb{P}_n$. We will call it the standard basis in $\mathbb{P}_n$.

**Remark 2.4.** If a vector space $V$ has a basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, then any vector $\mathbf{v}$ is uniquely defined by its coefficients in the decomposition $\mathbf{v} = \sum_{k=1}^n \alpha_k \mathbf{v}_k$. So, if we stack the coefficients $\alpha_k$ in a column, we can operate with them as if they were column vectors, i.e. as with elements of $\mathbb{F}^n$ (again here $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$, but everything also works for an abstract field $\mathbb{F}$).

This is a very important remark, that will be used throughout the book. It allows us to translate any statement about the standard column space $\mathbb{F}^n$ to a vector space $V$ with a basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$

Namely, if $\mathbf{v} = \sum_{k=1}^n \alpha_k \mathbf{v}_k$ and $\mathbf{w} = \sum_{k=1}^n \beta_k \mathbf{v}_k$, then

$$\mathbf{v} + \mathbf{w} = \sum_{k=1}^n \alpha_k \mathbf{v}_k + \sum_{k=1}^n \beta_k \mathbf{v}_k = \sum_{k=1}^n (\alpha_k + \beta_k)\mathbf{v}_k,$$

i.e. to get the column of coordinates of the sum one just need to add the columns of coordinates of the summands. Similarly, to get the coordinates of $\alpha\mathbf{v}$ we need simply to multiply the column of coordinates of $\mathbf{v}$ by $\alpha$.

**2.1. Generating and linearly independent systems.** The definition of a basis says that any vector admits a unique representation as a linear combination. This statement is in fact two statements, namely that the representation exists and that it is unique. Let us analyze these two statements separately.

If we only consider the existence we get the following notion

**Definition 2.5.** A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p \in V$ is called a generating system (also a *spanning system*, or a *complete system*) in $V$ if any vector $\mathbf{v} \in V$ admits representation as a linear combination

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_p\mathbf{v}_p = \sum_{k=1}^p \alpha_k \mathbf{v}_k.$$

The only difference from the definition of a basis is that we do not assume that the representation above is unique.

The words *generating*, *spanning* and *complete* here are synonyms. I personally prefer the term complete, because of my operator theory background. Generating and spanning are more often used in linear algebra textbooks.

Clearly, any basis is a generating (complete) system. Also, if we have a basis, say $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, and we add to it several vectors, say $\mathbf{v}_{n+1}, \ldots, \mathbf{v}_p$, then the new system will be a generating (complete) system. Indeed, we can represent any vector as a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, and just ignore the new ones (by putting corresponding coefficients $\alpha_k = 0$).

Now, let us turn our attention to the uniqueness. We do not want to worry about existence, so let us consider the zero vector $\mathbf{0}$, which always admits a representation as a linear combination.

**Definition.** A linear combination $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_p\mathbf{v}_p$ is called *trivial* if $\alpha_k = 0\ \forall k$.

A trivial linear combination is always (for all choices of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$) equal to $\mathbf{0}$, and that is probably the reason for the name.

**Definition.** A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p \in V$ is called *linearly independent* if only the trivial linear combination ($\sum_{k=1}^{p} \alpha_k\mathbf{v}_k$ with $\alpha_k = \mathbf{0}\ \forall k$) of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ equals $\mathbf{0}$.

In other words, the system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ is linearly independent iff the equation $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_p\mathbf{v}_p = \mathbf{0}$ (with unknowns $x_k$) has only trivial solution $x_1 = x_2 = \ldots = x_p = 0$.

If a system is not linearly independent, it is called *linearly dependent*. By negating the definition of linear independence, we get the following

**Definition.** A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ is called *linearly dependent* if $\mathbf{0}$ can be represented as a nontrivial linear combination, $\mathbf{0} = \sum_{k=1}^{p} \alpha_k\mathbf{v}_k$. Non-trivial here means that at least one of the coefficient $\alpha_k$ is non-zero. This can be (and usually is) written as $\sum_{k=1}^{p} |\alpha_k| \neq 0$.

So, restating the definition we can say, that a system is linearly dependent if and only if there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_p$, $\sum_{k=1}^{p} |\alpha_k| \neq 0$ such that

$$\sum_{k=1}^{p} \alpha_k\mathbf{v}_k = \mathbf{0}.$$

An alternative definition (in terms of equations) is that a system $\mathbf{v}_1$, $\mathbf{v}_2, \ldots, \mathbf{v}_p$ is linearly dependent iff the equation

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_p\mathbf{v}_p = \mathbf{0}$$

(with unknowns $x_k$) has a non-trivial solution. Non-trivial, once again means that at least one of $x_k$ is different from 0, and it can be written as $\sum_{k=1}^{p} |x_k| \neq 0$.

The following proposition gives an alternative description of linearly dependent systems.

**Proposition 2.6.** *A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p \in V$ is linearly dependent if and only if one of the vectors $\mathbf{v}_k$ can be represented as a linear combination of the other vectors,*

$$\mathbf{v}_k = \sum_{\substack{j=1 \\ j\neq k}}^{p} \beta_j \mathbf{v}_j. \tag{2.1}$$

**Proof.** Suppose the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ is linearly dependent. Then there exist scalars $\alpha_k$, $\sum_{k=1}^{p} |\alpha_k| \neq 0$ such that

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \ldots + \alpha_p \mathbf{v}_p = \mathbf{0}.$$

Let $k$ be the index such that $\alpha_k \neq 0$. Then, moving all terms except $\alpha_k \mathbf{v}_k$ to the right side we get

$$\alpha_k \mathbf{v}_k = -\sum_{\substack{j=1 \\ j\neq k}}^{p} \alpha_j \mathbf{v}_j.$$

Dividing both sides by $\alpha_k$ we get (2.1) with $\beta_j = -\alpha_j/\alpha_k$.

On the other hand, if (2.1) holds, $\mathbf{0}$ can be represented as a non-trivial linear combination

$$\mathbf{v}_k - \sum_{\substack{j=1 \\ j\neq k}}^{p} \beta_j \mathbf{v}_j = \mathbf{0}.$$

□

Obviously, any basis is a linearly independent system. Indeed, if a system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis, $\mathbf{0}$ admits a unique representation

$$\mathbf{0} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \ldots + \alpha_n \mathbf{v}_n = \sum_{k=1}^{n} \alpha_k \mathbf{v}_k.$$

Since the trivial linear combination always gives $\mathbf{0}$, the trivial linear combination must be the *only one* giving $\mathbf{0}$.

So, as we already discussed, if a system is a basis it is a complete (generating) and linearly independent system. The following proposition shows that the converse implication is also true.

**Proposition 2.7.** *A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in V$ is a basis if and only if it is linearly independent and complete (generating).*

In many textbooks a basis is defined as a complete and linearly independent system. By Proposition 2.7 this definition is equivalent to ours.

**Proof.** We already know that a basis is always linearly independent and complete, so in one direction the proposition is already proved.

Let us prove the other direction. Suppose a system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is linearly independent and complete. Take an arbitrary vector $\mathbf{v} \in V$. Since the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is linearly complete (generating), $\mathbf{v}$ can be represented as

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \ldots + \alpha_n \mathbf{v}_n = \sum_{k=1}^{n} \alpha_k \mathbf{v}_k.$$

We only need to show that this representation is unique.

Suppose $\mathbf{v}$ admits another representation

$$\mathbf{v} = \sum_{k=1}^{n} \widetilde{\alpha}_k \mathbf{v}_k.$$

Then

$$\sum_{k=1}^{n} (\alpha_k - \widetilde{\alpha}_k)\mathbf{v}_k = \sum_{k=1}^{n} \alpha_k \mathbf{v}_k - \sum_{k=1}^{n} \widetilde{\alpha}_k \mathbf{v}_k = \mathbf{v} - \mathbf{v} = \mathbf{0}.$$

Since the system is linearly independent, $\alpha_k - \widetilde{\alpha}_k = 0 \ \forall k$, and thus the representation $\mathbf{v} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \ldots + \alpha_n \mathbf{v}_n$ is unique. □

**Remark.** In many textbooks a basis is defined as a complete and linearly independent system (by Proposition 2.7 this definition is equivalent to ours). Although this definition is more common than one presented in this text, I prefer the latter. It emphasizes the main property of a basis, namely that any vector admits a unique representation as a linear combination.

**Proposition 2.8.** *Any (finite) generating system contains a basis.*

**Proof.** Suppose $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p \in V$ is a generating (complete) set. If it is linearly independent, it is a basis, and we are done.

Suppose it is not linearly independent, i.e. it is linearly dependent. Then there exists a vector $\mathbf{v}_k$ which can be represented as a linear combination of the vectors $\mathbf{v}_j$, $j \neq k$.

Since $\mathbf{v}_k$ can be represented as a linear combination of vectors $\mathbf{v}_j$, $j \neq k$, any linear combination of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ can be represented as a linear combination of the same vectors without $\mathbf{v}_k$ (i.e. the vectors $\mathbf{v}_j$, $1 \leq j \leq p$, $j \neq k$). So, if we delete the vector $\mathbf{v}_k$, the new system will still be a complete one.

If the new system is linearly independent, we are done. If not, we repeat the procedure.

Repeating this procedure finitely many times we arrive to a linearly independent and complete system, because otherwise we delete all vectors and end up with an empty set.

So, any finite complete (generating) set contains a complete linearly independent subset, i.e. a basis. □

**Exercises.**

**2.1.** Find a basis in the space of $3 \times 2$ matrices $M_{3\times 2}$.

**2.2.** True or false:

a) Any set containing a zero vector is linearly dependent

b) A basis must contain $\mathbf{0}$;

c) subsets of linearly dependent sets are linearly dependent;

d) subsets of linearly independent sets are linearly independent;

e) If $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_n\mathbf{v}_n = \mathbf{0}$ then all scalars $\alpha_k$ are zero;

**2.3.** Recall, that a matrix is called *symmetric* if $A^T = A$. Write down a basis in the space of *symmetric* $2 \times 2$ matrices (there are many possible answers). How many elements are in the basis?

**2.4.** Write down a basis for the space of

a) $3 \times 3$ symmetric matrices;

b) $n \times n$ symmetric matrices;

c) $n \times n$ *antisymmetric* ($A^T = -A$) matrices;

**2.5.** Let a system of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ be linearly independent but not generating. Show that it is possible to find a vector $\mathbf{v}_{r+1}$ such that the system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}$ is linearly independent. **Hint:** Take for $\mathbf{v}_{r+1}$ any vector that cannot be represented as a linear combination $\sum_{k=1}^{r} \alpha_k\mathbf{v}_k$ and show that the system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}$ is linearly independent.

**2.6.** Is it possible that vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are linearly dependent, but the vectors $\mathbf{w}_1 = \mathbf{v}_1 + \mathbf{v}_2$, $\mathbf{w}_2 = \mathbf{v}_2 + \mathbf{v}_3$ and $\mathbf{w}_3 = \mathbf{v}_3 + \mathbf{v}_1$ are linearly *independent*?

## 3. Linear Transformations. Matrix–vector multiplication

The words "transformation", "transform", "mapping", "map", "operator", "function" all denote the same object.

A *transformation* $T$ from a set $X$ to a set $Y$ is a rule that for each argument (input) $x \in X$ assigns a value (output) $y = T(x) \in Y$.

The set $X$ is called the *domain* of $T$, and the set $Y$ is called the *target space* or *codomain* of $T$.

We write $T : X \to Y$ to say that $T$ is a transformation with the domain $X$ and the target space $Y$.

**Definition.** Let $V$, $W$ be vector spaces (over the same field $\mathbb{F}$). A transformation $T : V \to W$ is called linear if

1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ $\forall \mathbf{u}, \mathbf{v} \in V$;
2. $T(\alpha \mathbf{v}) = \alpha T(\mathbf{v})$ for all $\mathbf{v} \in V$ and for all scalars $\alpha \in \mathbb{F}$.

Properties 1 and 2 together are equivalent to the following one:

$$T(\alpha \mathbf{u} + \beta \mathbf{v}) = \alpha T(\mathbf{u}) + \beta T(\mathbf{v}) \qquad \text{for all } \mathbf{u}, \mathbf{v} \in V \quad \text{and for all scalars } \alpha, \beta.$$

**3.1. Examples.** You dealt with linear transformation before, may be without even suspecting it, as the examples below show.

**Example.** Differentiation: Let $V = \mathbb{P}_n$ (the set of polynomials of degree at most $n$), $W = \mathbb{P}_{n-1}$, and let $T : \mathbb{P}_n \to \mathbb{P}_{n-1}$ be the differentiation operator,

$$T(p) := p' \qquad \forall p \in \mathbb{P}_n.$$

Since $(f + g)' = f' + g'$ and $(\alpha f)' = \alpha f'$, this is a linear transformation.

**Example.** Rotation: in this example $V = W = \mathbb{R}^2$ (the usual coordinate plane), and a transformation $T_\gamma : \mathbb{R}^2 \to \mathbb{R}^2$ takes a vector in $\mathbb{R}^2$ and rotates it counterclockwise by $\gamma$ radians. Since $T_\gamma$ rotates the plane as a whole, it rotates as a whole the parallelogram used to define a sum of two vectors (parallelogram law). Therefore the property 1 of linear transformation holds. It is also easy to see that the property 2 is also true.

**Example.** Reflection: in this example again $V = W = \mathbb{R}^2$, and the transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ is the reflection in the first coordinate axis, see the fig. It can also be shown geometrically, that this transformation is linear, but we will use another way to show that.

Namely, it is easy to write a formula for $T$,

$$T\Big(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\Big) = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}$$

and from this formula it is easy to check that the transformation is linear.

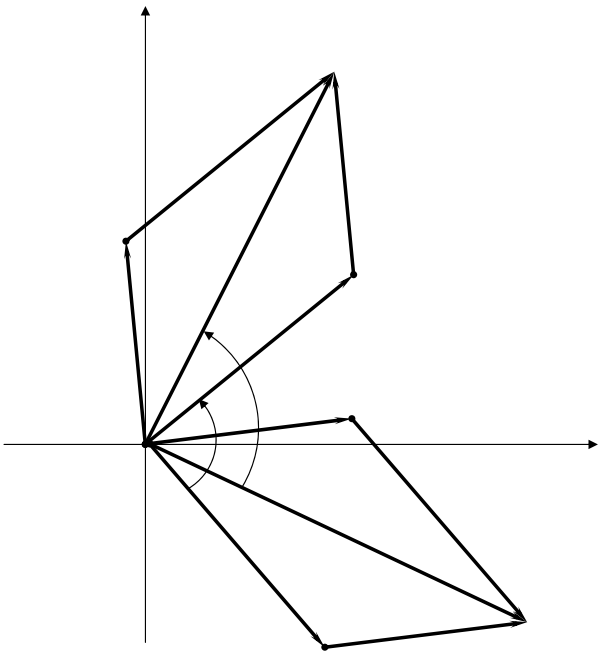

**Figure 1.** Rotation

**Example.** Let us investigate linear transformations $T : \mathbb{R} \to \mathbb{R}$. Any such transformation is given by the formula

$$T(x) = ax \qquad \text{where } a = T(1).$$

Indeed,

$$T(x) = T(x \times 1) = xT(1) = xa = ax.$$

So, any linear transformation of $\mathbb{R}$ is just a multiplication by a constant.

**3.2. Linear transformations $\mathbb{F}^n \to \mathbb{F}^m$. Matrix–column multiplication.** It turns out that a linear transformation $T : \mathbb{F}^n \to \mathbb{F}^m$ also can be represented as a multiplication, not by a scalar, but by a matrix.

Let us see how. Let $T : \mathbb{F}^n \to \mathbb{F}^m$ be a linear transformation. What information do we need to compute $T(\mathbf{x})$ for all vectors $\mathbf{x} \in \mathbb{F}^n$? My claim is that it is sufficient to know how $T$ acts on the standard basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ of $\mathbb{F}^n$. Namely, it is sufficient to know $n$ vectors in $\mathbb{F}^m$ (i.e. the vectors of size $m$),

$$\mathbf{a}_1 = T(\mathbf{e}_1), \quad \mathbf{a}_2 := T(\mathbf{e}_2), \quad \dots\ , \quad \mathbf{a}_n := T(\mathbf{e}_n).$$

Indeed, let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Then $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_n\mathbf{e}_n = \sum_{k=1}^{n} x_k\mathbf{e}_k$ and

$$T(\mathbf{x}) = T(\sum_{k=1}^{n} x_k\mathbf{e}_k) = \sum_{k=1}^{n} T(x_k\mathbf{e}_k) = \sum_{k=1}^{n} x_k T(\mathbf{e}_k) = \sum_{k=1}^{n} x_k\mathbf{a}_k.$$

So, if we join the vectors (columns) $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ together in a matrix $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ ($\mathbf{a}_k$ being the $k$th column of $A$, $k = 1, 2, \ldots, n$), this matrix contains all the information about $T$.

Let us show how one should define the product of a matrix and a vector (column) to represent the transformation $T$ as a product, $T(\mathbf{x}) = A\mathbf{x}$. Let

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \ldots & a_{m,n} \end{pmatrix}.$$

Recall, that the column number $k$ of $A$ is the vector $\mathbf{a}_k$, i.e.

$$\mathbf{a}_k = \begin{pmatrix} a_{1,k} \\ a_{2,k} \\ \vdots \\ a_{m,k} \end{pmatrix}.$$

Then if we want $A\mathbf{x} = T(\mathbf{x})$ we get

$$A\mathbf{x} = \sum_{k=1}^{n} x_k\mathbf{a}_k = x_1 \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{m,1} \end{pmatrix} + x_2 \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ \vdots \\ a_{m,2} \end{pmatrix} + \ldots + x_n \begin{pmatrix} a_{1,n} \\ a_{2,n} \\ \vdots \\ a_{m,n} \end{pmatrix}.$$

So, the matrix–vector multiplication should be performed by the following *column by coordinate* rule:

| multiply each column of the matrix by the corresponding coordinate of the vector. |
|---|

**Example.**

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 3 \end{pmatrix} + 2 \begin{pmatrix} 2 \\ 2 \end{pmatrix} + 3 \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 14 \\ 10 \end{pmatrix}.$$

The "column by coordinate" rule is very well adapted for parallel computing. It will be also very important in different theoretical constructions later.

However, when doing computations manually, it is more convenient to compute the result one entry at a time. This can be expressed as the following *row by column* rule:

> To get the entry number $k$ of the result, one need to multiply row number $k$ of the matrix by the vector, that is, if $A\mathbf{x} = \mathbf{y}$, then $y_k = \sum_{j=1}^{n} a_{k,j}x_j$, $k = 1, 2, \dots m$;

here $x_j$ and $y_k$ are coordinates of the vectors $\mathbf{x}$ and $\mathbf{y}$ respectively, and $a_{j,k}$ are the entries of the matrix $A$.

**Example.**

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1\cdot 1 + 2\cdot 2 + 3\cdot 3 \\ 4\cdot 1 + 5\cdot 2 + 6\cdot 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 32 \end{pmatrix}$$

**3.3. Linear transformations and generating sets.** As we discussed above, linear transformation $T$ (acting from $\mathbb{F}^n$ to $\mathbb{F}^m$) is completely defined by its values on the standard basis in $\mathbb{F}^n$.

The fact that we consider the standard basis is not essential, one can consider any basis, even any generating (spanning) set. Namely,

> A linear transformation $T : V \to W$ is completely defined by its values on a generating set (in particular by its values on a basis).

So, if $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a generating set (in particular, if it is a basis) in $V$, and $T$ and $T_1$ are linear transformations $T, T_1 : V \to W$ such that

$$T\mathbf{v}_k = T_1\mathbf{v}_k, \qquad k = 1, 2, \dots, n$$

then $T = T_1$.

The proof of this statement is trivial and left as an exercise.

**3.4. Conclusions.**

- To get the matrix of a linear transformation $T : \mathbb{F}^n \to \mathbb{F}^m$ one needs to join the vectors $\mathbf{a}_k = T\mathbf{e}_k$ (where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is the standard basis in $\mathbb{F}^n$) into a matrix: $k$th column of the matrix is $\mathbf{a}_k$, $k = 1, 2, \dots, n$.
- If the matrix $A$ of the linear transformation $T$ is known, then $T(\mathbf{x})$ can be found by the matrix–vector multiplication, $T(\mathbf{x}) = A\mathbf{x}$. To perform matrix–vector multiplication one can use either "column by coordinate" or "row by column" rule.

The latter seems more appropriate for manual computations. The former is well adapted for parallel computers, and will be used in different theoretical constructions.

For a linear transformation $T : \mathbb{F}^n \to \mathbb{F}^m$, its matrix is usually denoted as $[T]$. However, very often people do not distinguish between a linear transformation and its matrix, and use the same symbol for both. When it does not lead to confusion, we will also use the same symbol for a transformation and its matrix.

The notation $T\mathbf{v}$ is often used instead of $T(\mathbf{v})$.

Since a linear transformation is essentially a multiplication, the notation $T\mathbf{v}$ is often used instead of $T(\mathbf{v})$. We will also use this notation. Note that the usual order of algebraic operations apply, i.e. $T\mathbf{v} + \mathbf{u}$ means $T(\mathbf{v}) + \mathbf{u}$, not $T(\mathbf{v} + \mathbf{u})$.

In the matrix vector multiplication using the "row by column" rule be sure that you have the same number of entries in the row and in the column. The entries in the row and in the column should end simultaneously: if not, the multiplication is not defined.

**Remark.** In the matrix–vector multiplication $A\mathbf{x}$ the number of columns of the matrix $A$ matrix must coincide with the size of the vector $\mathbf{x}$, i.e. a vector in $\mathbb{F}^n$ can only be multiplied by an $m \times n$ matrix.

It makes sense, since an $m \times n$ matrix defines a linear transformation $\mathbb{F}^n \to \mathbb{F}^m$, so vector $\mathbf{x}$ must belong to $\mathbb{F}^n$.

The easiest way to remember this is to remember that if performing multiplication you run out of some elements faster, then the multiplication is not defined. For example, if using the "row by column" rule you run out of row entries, but still have some unused entries in the vector, the multiplication is not defined. It is also not defined if you run out of vector's entries, but still have unused entries in the row.

**Remark.** One does not have to restrict himself to the case of $\mathbb{F}^n$ with standard basis: everything described in this section works for transformation between arbitrary vector spaces as long as there is a basis in the domain and in the target space. Of course, if one changes a basis, the matrix of the linear transformation will be different. This will be discussed later in Section 8.

**Exercises.**

**3.1.** Multiply:

a) $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$;

b) $\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix}$;

c) $\begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$;

d) $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$.

**3.2.** Let a linear transformation in $\mathbb{R}^2$ be the reflection in the line $x_1 = x_2$. Find its matrix.

**3.3.** For each linear transformation below find it matrix

a) $T : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $T(x, y)^T = (x + 2y, 2x - 5y, 7y)^T$;

b) $T : \mathbb{R}^4 \to \mathbb{R}^3$ defined by $T(x_1, x_2, x_3, x_4)^T = (x_1+x_2+x_3+x_4, x_2-x_4, x_1+3x_2+6x_4)^T$;

c) $T : \mathbb{P}_n \to \mathbb{P}_n$, $Tf(t) = f'(t)$ (find the matrix with respect to the standard basis $1, t, t^2, \dots, t^n$);

d) $T : \mathbb{P}_n \to \mathbb{P}_n$, $Tf(t) = 2f(t) + 3f'(t) - 4f''(t)$ (again with respect to the standard basis $1, t, t^2, \dots, t^n$).

**3.4.** Find $3 \times 3$ matrices representing the transformations of $\mathbb{R}^3$ which:

a) project every vector onto $x$-$y$ plane;

b) reflect every vector through $x$-$y$ plane;

c) rotate the $x$-$y$ plane through $30°$, leaving $z$-axis alone.

**3.5.** Let $A$ be a linear transformation. If $\mathbf{z}$ is the center of the straight interval $[\mathbf{x}, \mathbf{y}]$, show that $A\mathbf{z}$ is the center of the interval $[A\mathbf{x}, A\mathbf{y}]$. **Hint:** What does it mean that $\mathbf{z}$ is the center of the interval $[\mathbf{x}, \mathbf{y}]$?

**3.6.** The set $\mathbb{C}$ of complex numbers can be canonically identified with the space $\mathbb{R}^2$ by treating each $z = x + iy \in \mathbb{C}$ as a column $(x, y)^T \in \mathbb{R}^2$.

a) Treating $\mathbb{C}$ as a complex vector space, show that the multiplication by $\alpha = a + ib \in \mathbb{C}$ is a linear transformation in $\mathbb{C}$. What is its matrix?

b) Treating $\mathbb{C}$ as the real vector space $\mathbb{R}^2$ show that the multiplication by $\alpha = a + ib$ defines a linear transformation there. What is its matrix?

c) Define $T(x+iy) = 2x-y+i(x-3y)$. Show that this transformation is not a linear transformation in the complex vectors space $\mathbb{C}$, but if we treat $\mathbb{C}$ as the real vector space $\mathbb{R}^2$ then it is a linear transformation there (i.e. that $T$ is a *real linear* but not a *complex linear* transformation).
Find the matrix of the real liner transformation $T$.

**3.7.** Show that any linear transformation in $\mathbb{C}$ (treated as a complex vector space) is a multiplication by $\alpha \in \mathbb{C}$.

## 4. Linear transformations as a vector space

What operations can we perform with linear transformations? We can always multiply a linear transformation for a scalar, i.e. if we have a linear

transformation $T : V \to W$ and a scalar $\alpha$ we can define a new transformation $\alpha T$ by

$$(\alpha T)\mathbf{v} = \alpha(T\mathbf{v}) \qquad \forall \mathbf{v} \in V.$$

It is easy to check that $\alpha T$ is also a linear transformation:

$$\begin{aligned}(\alpha T)(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2) &= \alpha(T(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2)) && \text{by the definition of } \alpha T \\ &= \alpha(\alpha_1 T\mathbf{v}_1 + \alpha_2 T\mathbf{v}_2) && \text{by the linearity of } T \\ &= \alpha_1\alpha T\mathbf{v}_1 + \alpha_2\alpha T\mathbf{v}_2 = \alpha_1(\alpha T)\mathbf{v}_1 + \alpha_2(\alpha T)\mathbf{v}_2\end{aligned}$$

If $T_1$ and $T_2$ are linear transformations with the same domain and target space ($T_1 : V \to W$ and $T_2 : V \to W$, or in short $T_1, T_2 : V \to W$), then we can add these transformations, i.e. define a new transformation $T = (T_1 + T_2) : V \to W$ by

$$(T_1 + T_2)\mathbf{v} = T_1\mathbf{v} + T_2\mathbf{v} \qquad \forall \mathbf{v} \in V.$$

It is easy to check that the transformation $T_1 + T_2$ is a linear one, one just needs to repeat the above reasoning for the linearity of $\alpha T$.

So, if we fix vector spaces $V$ and $W$ and consider the collection of all linear transformations from $V$ to $W$ (let us denote it by $\mathcal{L}(V, W)$), we can define 2 operations on $\mathcal{L}(V, W)$: multiplication by a scalar and addition. It can be easily shown that these operations satisfy the axioms of a vector space, defined in Section 1.

This should come as no surprise for the reader, since axioms of a vector space essentially mean that operation on vectors follow standard rules of algebra. And the operations on linear transformations are defined as to satisfy these rules!

As an illustration, let us write down a formal proof of the first distributive law (axiom 7) of a vector space. We want to show that $\alpha(T_1 + T_2) = \alpha T_1 + \alpha T_2$. For any $\mathbf{v} \in V$

$$\begin{aligned}\alpha(T_1 + T_2)\mathbf{v} &= \alpha((T_1 + T_2)\mathbf{v}) && \text{by the definition of multiplication} \\ &= \alpha(T_1\mathbf{v} + T_2\mathbf{v}) && \text{by the definition of the sum} \\ &= \alpha T_1\mathbf{v} + \alpha T_2\mathbf{v} && \text{by Axiom 7 for } W \\ &= (\alpha T_1 + \alpha T_2)\mathbf{v} && \text{by the definition of the sum}\end{aligned}$$

So indeed $\alpha(T_1 + T_2) = \alpha T_1 + \alpha T_2$.

**Remark.** Linear operations (addition and multiplication by a scalar) on linear transformations $T : \mathbb{F}^n \to \mathbb{F}^m$ correspond to the respective operations on their matrices. Since we know that the set of $m \times n$ matrices is a vector space, this immediately implies that $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ is a vector space.

We presented the abstract proof above, first of all because it work for general spaces, for example, for spaces without a basis, where we cannot

work with coordinates. Secondly, the reasonings similar to the abstract one presented here, are used in many places, so the reader will benefit from understanding it.

And as the reader gains some mathematical sophistication, he/she will see that this abstract reasoning is indeed a very simple one, that can be performed almost automatically.

## 5. Composition of linear transformations and matrix multiplication.

**5.1. Definition of the matrix multiplication.** Knowing matrix–vector multiplication, one can easily guess what is the natural way to define the product $AB$ of two matrices: Let us multiply by $A$ each column of $B$ (matrix-vector multiplication) and join the resulting column-vectors into a matrix. Formally,

| if $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r$ are the columns of $B$, then $A\mathbf{b}_1, A\mathbf{b}_2, \dots, A\mathbf{b}_r$ are the columns of the matrix $AB$. |
|---|

Recalling the *row by column* rule for the matrix–vector multiplication we get the following *row by column rule for the matrices*

| the entry $(AB)_{j,k}$ (the entry in the row $j$ and column $k$) of the product $AB$ is defined by<br>$(AB)_{j,k} = (\text{row } \#j \text{ of } A) \cdot (\text{column } \#k \text{ of } B)$ |
|---|

Formally it can be rewritten as

$$(AB)_{j,k} = \sum_l a_{j,l} b_{l,k},$$

if $a_{j,k}$ and $b_{j,k}$ are entries of the matrices $A$ and $B$ respectively.

I intentionally did not speak about sizes of the matrices $A$ and $B$, but if we recall the row by column rule for the matrix–vector multiplication, we can see that in order for the multiplication to be defined, the size of a row of $A$ should be equal to the size of a column of $B$.

In other words the product $AB$ is defined if and only if $A$ is an $m \times n$ and $B$ is $n \times r$ matrix.

**5.2. Motivation: composition of linear transformations.** One can ask yourself here: Why are we using such a complicated rule of multiplication? Why don't we just multiply matrices entrywise?

And the answer is, that the multiplication, as it is defined above, arises naturally from the composition of linear transformations.

Suppose we have two linear transformations, $T_1 : \mathbb{F}^n \to \mathbb{F}^m$ and $T_2 : \mathbb{F}^r \to \mathbb{F}^n$. Define the composition $T = T_1 \circ T_2$ of the transformations $T_1$, $T_2$ as

$$T(\mathbf{x}) = T_1(T_2(\mathbf{x})) \qquad \forall x \in \mathbb{F}^r.$$

We will usually identify a linear transformation and its matrix, but in the next few paragraphs we will distinguish them

Note that $T_2(\mathbf{x}) \in \mathbb{F}^n$. Since $T_1 : \mathbb{F}^n \to \mathbb{F}^m$, the expression $T_1(T_2(\mathbf{x}))$ is well defined and the result belongs to $\mathbb{F}^m$. So, $T : \mathbb{F}^r \to \mathbb{F}^m$.

It is easy to show that $T$ is a linear transformation (exercise), so it is defined by an $m \times r$ matrix. How one can find this matrix, knowing the matrices of $T_1$ and $T_2$?

Let $A$ be the matrix of $T_1$ and $B$ be the matrix of $T_2$. As we discussed in the previous section, the columns of $T$ are vectors $T(\mathbf{e}_1), T(\mathbf{e}_2), \dots, T(\mathbf{e}_r)$, where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_r$ is the standard basis in $\mathbb{F}^r$. For $k = 1, 2, \dots, r$ we have

$$T(\mathbf{e}_k) = T_1(T_2(\mathbf{e}_k)) = T_1(B\mathbf{e}_k) = T_1(\mathbf{b}_k) = A\mathbf{b}_k$$

(operators $T_2$ and $T_1$ are simply the multiplication by $B$ and $A$ respectively).

So, the columns of the matrix of $T$ are $A\mathbf{b}_1, A\mathbf{b}_2, \dots, A\mathbf{b}_r$, and that is exactly how the matrix $AB$ was defined!

Let us return to identifying again a linear transformation with its matrix. Since the matrix multiplication agrees with the composition, we can (and will) write $T_1T_2$ instead of $T_1 \circ T_2$ and $T_1T_2\mathbf{x}$ instead of $T_1(T_2(\mathbf{x}))$.

**Note:** order of transformations!

Note that in the composition $T_1T_2$ the transformation $T_2$ is applied first! The way to remember this is to see that in $T_1T_2\mathbf{x}$ the transformation $T_2$ meets $\mathbf{x}$ fist.

**Remark.** There is another way of checking the dimensions of matrices in a product, different form the row by column rule: for a composition $T_1T_2$ to be defined it is necessary that $T_2\mathbf{x}$ belongs to the domain of $T_1$. If $T_2$ acts from some space, say $\mathbb{F}^r$ to $\mathbb{F}^n$, then $T_1$ must act from $\mathbb{F}^n$ to some space, say $\mathbb{F}^m$. So, in order for $T_1T_2$ to be defined the matrices of $T_1$ and $T_2$ should be of sizes $m \times n$ and $n \times r$ respectively—the same condition as obtained from the *row by column* rule.

**Example.** Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be the reflection in the line $x_1 = 3x_2$. It is a linear transformation, so let us find its matrix. To find the matrix, we need to compute $T\mathbf{e}_1$ and $T\mathbf{e}_2$. However, the direct computation of $T\mathbf{e}_1$ and $T\mathbf{e}_2$ involves significantly more trigonometry than a sane person is willing to remember.

An easier way to find the matrix of $T$ is to represent it as a composition of simple linear transformation. Namely, let $\gamma$ be the angle between the $x_1$ axis and the line $x_1 = 3x_2$, and let $T_0$ be the reflection in the $x_1$-axis. Then to get the reflection $T$ we can first rotate the plane by the angle $-\gamma$, moving the line $x_1 = 3x_2$ to the $x_1$-axis, then reflect everything in the $x_1$

axis, and then rotate the plane by $\gamma$, taking everything back. Formally it can be written as

$$T = R_\gamma T_0 R_{-\gamma}$$

(note the order of terms!), where $R_\gamma$ is the rotation by $\gamma$. The matrix of $T_0$ is easy to compute,

$$T_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

the rotation matrices are known

$$R_\gamma = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma, \end{pmatrix},$$
$$R_{-\gamma} = \begin{pmatrix} \cos(-\gamma) & -\sin(-\gamma) \\ \sin(-\gamma) & \cos(-\gamma), \end{pmatrix} = \begin{pmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma, \end{pmatrix}$$

To compute $\sin\gamma$ and $\cos\gamma$ take a vector in the line $x_1 = 3x_2$, say a vector $(3,1)^T$. Then

$$\cos\gamma = \frac{\text{first coordinate}}{\text{length}} = \frac{3}{\sqrt{3^2+1^2}} = \frac{3}{\sqrt{10}}$$

and similarly

$$\sin\gamma = \frac{\text{second coordinate}}{\text{length}} = \frac{1}{\sqrt{3^2+1^2}} = \frac{1}{\sqrt{10}}$$

Gathering everything together we get

$$T = R_\gamma T_0 R_{-\gamma} = \frac{1}{\sqrt{10}}\begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\frac{1}{\sqrt{10}}\begin{pmatrix} 3 & 1 \\ -1 & 3 \end{pmatrix}$$
$$= \frac{1}{10}\begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ -1 & 3 \end{pmatrix}$$

It remains only to perform matrix multiplication here to get the final result. □

**5.3. Properties of matrix multiplication.** Matrix multiplication enjoys a lot of properties, familiar to us from high school algebra:

1. Associativity: $A(BC) = (AB)C$, provided that either left or right side is well defined; we therefore can (and will) simply write $ABC$ in this case.
2. Distributivity: $A(B+C) = AB + AC$, $(A+B)C = AC + BC$, provided either left or right side of each equation is well defined.
3. One can take scalar multiplies out: $A(\alpha B) = (\alpha A)B = \alpha(AB) = \alpha AB$.

These properties are easy to prove. One should prove the corresponding properties for linear transformations, and they almost trivially follow from the definitions. The properties of linear transformations then imply the properties for the matrix multiplication.

The new twist here is that the commutativity fails:

> matrix multiplication is non-commutative, i.e. generally for matrices $AB \neq BA$.

One can see easily it would be unreasonable to expect the commutativity of matrix multiplication. Indeed, let $A$ and $B$ be matrices of sizes $m \times n$ and $n \times r$ respectively. Then the product $AB$ is well defined, but if $m \neq r$, $BA$ is not defined.

Even when both products are well defined, for example, when $A$ and $B$ are $n \times n$ (square) matrices, the multiplication is still non-commutative. If we just pick the matrices $A$ and $B$ at random, the chances are that $AB \neq BA$: we have to be very lucky to get $AB = BA$.

**5.4. Transposed matrices and multiplication.** Given a matrix $A$, its *transpose* (or transposed matrix) $A^T$ is defined by transforming the rows of $A$ into the columns. For example

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

So, the columns of $A^T$ are the rows of $A$ and vice versa, the rows of $A^T$ are the columns of $A$.

The formal definition is as follows: $(A^T)_{j,k} = (A)_{k,j}$ meaning that the entry of $A^T$ in the row number $j$ and column number $k$ equals the entry of $A$ in the row number $k$ and row number $j$.

The transpose of a matrix has a very nice interpretation in terms of linear transformations, namely it gives the so-called *adjoint* transformation. We will study this in detail later, but for now transposition will be just a useful formal operation.

One of the first uses of the transpose is that we can write a column vector $\mathbf{x} \in \mathbb{F}^n$ as $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$. If we put the column vertically, it will use significantly more space.

A simple analysis of the row by columns rule shows that

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

i.e. when you take the transpose of the product, you change the order of the terms.

**5.5. Trace and matrix multiplication.** For a square $(n \times n)$ matrix $A = (a_{j,k})$ its trace (denoted by $\operatorname{trace} A$) is the sum of the diagonal entries

$$\operatorname{trace} A = \sum_{k=1}^{n} a_{k,k}.$$

**Theorem 5.1.** *Let $A$ and $B$ be matrices of size $m \times n$ and $n \times m$ respectively (so the both products $AB$ and $BA$ are well defined). Then*

$$\operatorname{trace}(AB) = \operatorname{trace}(BA)$$

We leave the proof of this theorem as an exercise, see Problem 5.6 below. There are essentially two ways of proving this theorem. One is to compute the diagonal entries of $AB$ and of $BA$ and compare their sums. This method requires some proficiency in manipulating sums in $\sum$ notation.

If you are not comfortable with algebraic manipulations, there is another way. We can consider two linear transformations, $T$ and $T_1$, acting from $M_{n\times m}$ to $\mathbb{F} = \mathbb{F}^1$ defined by

$$T(X) = \operatorname{trace}(AX), \qquad T_1(X) = \operatorname{trace}(XA)$$

To prove the theorem it is sufficient to show that $T = T_1$; the equality for $X = B$ gives the theorem.

Since a linear transformation is completely defined by its values on a generating system, we need just to check the equality on some simple matrices, for example on matrices $X_{j,k}$, which has all entries 0 except the entry 1 in the intersection of $j$th column and $k$th row.

**Exercises.**

**5.1.** Let

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & -2 \end{pmatrix}, \; C = \begin{pmatrix} 1 & -2 & 3 \\ -2 & 1 & -1 \end{pmatrix}, \; D = \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix}$$

a) Mark all the products that are defined, and give the dimensions of the result: $AB$, $BA$, $ABC$, $ABD$, $BC$, $BC^T$, $B^TC$, $DC$, $D^TC^T$.

b) Compute $AB$, $A(3B + C)$, $B^TA$, $A(BD)$, $(AB)D$.

**5.2.** Let $T_\gamma$ be the matrix of rotation by $\gamma$ in $\mathbb{R}^2$. Check by matrix multiplication that $T_\gamma T_{-\gamma} = T_{-\gamma} T_\gamma = I$

**5.3.** Multiply two rotation matrices $T_\alpha$ and $T_\beta$ (it is a rare case when the multiplication is commutative, i.e. $T_\alpha T_\beta = T_\beta T_\alpha$, so the order is not essential). Deduce formulas for $\sin(\alpha + \beta)$ and $\cos(\alpha + \beta)$ from here.

**5.4.** Find the matrix of the orthogonal projection in $\mathbb{R}^2$ onto the line $x_1 = -2x_2$. **Hint:** What is the matrix of the projection onto the coordinate axis $x_1$?

**5.5.** Find linear transformations $A, B : \mathbb{R}^2 \to \mathbb{R}^2$ such that $AB = \mathbf{0}$ but $BA \neq \mathbf{0}$.

**5.6.** Prove Theorem 5.1, i.e. prove that $\operatorname{trace}(AB) = \operatorname{trace}(BA)$.

**5.7.** Construct a non-zero matrix $A$ such that $A^2 = \mathbf{0}$.

**5.8.** Find the matrix of the reflection through the line $y = -2x/3$. Perform all the multiplications.

## 6. Invertible transformations and matrices. Isomorphisms

**6.1. Identity transformation and identity matrix.** Among all linear transformations, there is a special one, the identity transformation (operator) $I$, $I\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x}$.

To be precise, there are infinitely many identity transformations: for any vector space $V$, there is the identity transformation $I = I_V : V \to V$, $I_V\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in V$. However, when it is does not lead to the confusion we will use the same symbol $I$ for all identity operators (transformations). We will use the notation $I_V$ only we want to emphasize in what space the transformation is acting.

Often, the symbol $E$ is used in Linear Algebra textbooks for the identity matrix. I prefer $I$, since it is used in operator theory.

Clearly, if $I : \mathbb{F}^n \to \mathbb{F}^n$ is the identity transformation in $\mathbb{F}^n$, its matrix is the $n \times n$ matrix

$$I = I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

(1 on the main diagonal and 0 everywhere else). When we want to emphasize the size of the matrix, we use the notation $I_n$; otherwise we just use $I$.

Clearly, for an arbitrary linear transformation $A$, the equalities

$$AI = A, \qquad IA = A$$

hold (whenever the product is defined).

### 6.2. Invertible transformations.

**Definition.** Let $A : V \to W$ be a linear transformation. We say that the transformation $A$ is *left invertible* if there exist a linear transformation $B : W \to V$ such that

$$BA = I \qquad (I = I_V \text{ here}).$$

The transformation $A$ is called *right invertible* if there exists a linear transformation $C : W \to V$ such that

$$AC = I \qquad (\text{here } I = I_W).$$

The transformations $B$ and $C$ are called *left* and *right inverses* of $A$. Note, that we did not assume the uniqueness of $B$ or $C$ here, and generally left and right inverses are not unique.

**Definition.** A linear transformation $A : V \to W$ is called *invertible* if it is both right and left invertible.

**Theorem 6.1.** *If a linear transformation $A : V \to W$ is invertible, then its left and right inverses $B$ and $C$ are unique and coincide.*

**Corollary.** *A transformation $A : V \to W$ is invertible if and only if there exists a unique linear transformation (denoted $A^{-1}$), $A^{-1} : W \to V$ such that*

$$A^{-1}A = I_V, \qquad AA^{-1} = I_W.$$

Very often this property is used as the definition of an invertible transformation

The transformation $A^{-1}$ is called the *inverse* of $A$.

**Proof of Theorem 6.1.** Let $BA = I$ and $AC = I$. Then

$$BAC = B(AC) = BI = B.$$

On the other hand

$$BAC = (BA)C = IC = C,$$

and therefore $B = C$.

Suppose for some transformation $B_1$ we have $B_1A = I$. Repeating the above reasoning with $B_1$ instead of $B$ we get $B_1 = C$. Therefore the left inverse $B$ is unique. The uniqueness of $C$ is proved similarly. □

**Definition.** A matrix is called *invertible* (resp. *left invertible, right invertible*) if the corresponding linear transformation is invertible (resp. left invertible, right invertible).

Theorem 6.1 asserts that a matrix $A$ is invertible if there exists a unique matrix $A^{-1}$ such that $A^{-1}A = I$, $AA^{-1} = I$. The matrix $A^{-1}$ is called (surprise) the *inverse* of $A$.

*Examples.*

1. The identity transformation (matrix) is invertible, $I^{-1} = I$;
2. The rotation $R_\gamma$

$$R_\gamma = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix}$$

is invertible, and the inverse is given by $(R_\gamma)^{-1} = R_{-\gamma}$. This equality is clear from the geometric description of $R_\gamma$, and it also can be checked by the matrix multiplication;

3. The column $(1,1)^T$ is left invertible but not right invertible. One of the possible left inverses in the row $(1/2, 1/2)$.

   To show that this matrix is not right invertible, we just notice that there are more than one left inverse. **Exercise:** describe all left inverses of this matrix.

4. The row $(1,1)$ is right invertible, but not left invertible. The column $(1/2, 1/2)^T$ is a possible right inverse.

An invertible matrix must be square (to be proved later)

**Remark 6.2.** An invertible matrix *must* be square ($n \times n$). Moreover, if a square matrix $A$ has either left or right inverse, it is invertible. So, it is sufficient to check only one of the identities $AA^{-1} = I$, $A^{-1}A = I$.

This fact will be proved later. Until we prove this fact, we will not use it. I presented it here only to stop students from trying wrong directions.

6.2.1. *Properties of the inverse transformation.*

Inverse of a product: $(AB)^{-1} = B^{-1}A^{-1}$. Note the change of order

**Theorem 6.3** (Inverse of the product)**.** *If linear transformations $A$ and $B$ are invertible (and such that the product $AB$ is defined), then the product $AB$ is invertible and*

$$(AB)^{-1} = B^{-1}A^{-1}$$

*(note the change of the order!)*

**Proof.** Direct computation shows:

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$$

and similarly

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}IB = B^{-1}B = I$$

□

**Remark 6.4.** The invertibility of the product $AB$ does not imply the invertibility of the factors $A$ and $B$ (can you think of an example?). However, if one of the factors (either $A$ or $B$) and the product $AB$ are invertible, then the second factor is also invertible.

We leave the proof of this fact as an exercise.

**Theorem 6.5** (Inverse of $A^T$)**.** *If a matrix $A$ is invertible, then $A^T$ is also invertible and*

$$(A^T)^{-1} = (A^{-1})^T$$

**Proof.** Using $(AB)^T = B^T A^T$ we get

$$(A^{-1})^T A^T = (AA^{-1})^T = I^T = I,$$

and similarly

$$A^T(A^{-1})^T = (A^{-1}A)^T = I^T = I.$$

□

And finally, if $A$ is invertible, then $A^{-1}$ is also invertible, $(A^{-1})^{-1} = A$.

So, let us summarize the main properties of the inverse:

1. If $A$ is invertible, then $A^{-1}$ is also invertible, $(A^{-1})^{-1} = A$;
2. If $A$ and $B$ are invertible and the product $AB$ is defined, then $AB$ is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.
3. If $A$ is invertible, then $A^T$ is also invertible and $(A^T)^{-1} = (A^{-1})^T$.

**6.3. Isomorphism. Isomorphic spaces.** An invertible linear transformation $A : V \to W$ is called an *isomorphism.* We did not introduce anything new here, it is just another name for the object we already studied.

Two vector spaces $V$ and $W$ are called *isomorphic* (denoted $V \cong W$) if there is an isomorphism $A : V \to W$.

Isomorphic spaces can be considered as different representation of the same space, meaning that all properties and constructions involving vector space operations are preserved under isomorphism.

The theorem below illustrates this statement.

**Theorem 6.6.** *Let $A : V \to W$ be an isomorphism, and let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be a basis in $V$. Then the system $A\mathbf{v}_1, A\mathbf{v}_2, \dots, A\mathbf{v}_n$ is a basis in $W$.*

We leave the proof of the theorem as an exercise.

**Remark.** In the above theorem one can replace "basis" by "linearly independent", or "generating", or "linearly dependent"—all these properties are preserved under isomorphisms.

**Remark.** If $A$ is an isomorphism, then so is $A^{-1}$. Therefore in the above theorem we can state that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis if and only if $A\mathbf{v}_1, A\mathbf{v}_2$, $\dots, A\mathbf{v}_n$ is a basis.

The converse to the Theorem 6.6 is also true

**Theorem 6.7.** *Let $A : V \to W$ be a linear map,and let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ be bases in $V$ and $W$ respectively. If $A\mathbf{v}_k = \mathbf{w}_k$, $k = 1, 2, \dots, n$, then $A$ is an isomorphism.*

**Proof.** Define the inverse transformation $A^{-1}$ by $A^{-1}\mathbf{w}_k = \mathbf{v}_k$, $k = 1$, $2, \dots, n$ (as we know, a linear transformation is defined by its values on a basis). □

*Examples.*

1. Let $A : \mathbb{F}^{n+1} \to \mathbb{P}_n^{\mathbb{F}}$ ($\mathbb{P}_n^{\mathbb{F}}$ is the set of polynomials $\sum_{k=0}^{n} a_k t^k$, $\alpha_k \in \mathbb{F}$ of degree at most $n$) is defined by

$$A\mathbf{e}_1 = 1,\ A\mathbf{e}_2 = t, \dots, A\mathbf{e}_n = t^{n-1},\ A\mathbf{e}_{n+1} = t^n$$

By Theorem 6.7 $A$ is an isomorphism, so $\mathbb{P}_n^{\mathbb{F}} \cong \mathbb{F}^{n+1}$.

2. Let $V$ be a vector space (over $\mathbb{F}$) with a basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$. Define transformation $A : \mathbb{F}^n \to V$ by
$$A\mathbf{e}_k = \mathbf{v}_k, \qquad k = 1, 2, \dots, n,$$
where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is the standard basis in $\mathbb{F}^n$. Again by Theorem 6.7 $A$ is an isomorphism, so $V \cong \mathbb{F}^n$.

Any real vector space with a basis is isomorphic to $\mathbb{R}^n$ (for some $n$). Similarly, any complex vector space with a basis is isomorphic to $\mathbb{C}^n$.

3. The space $M_{2\times 3}^{\mathbb{F}}$ of $2 \times 3$ matrices with entries in $\mathbb{F}$ is isomorphic to $\mathbb{R}^6$;

4. More generally, $M_{m\times n}^{\mathbb{F}} \cong \mathbb{F}^{m\cdot n}$

### 6.4. Invertibility and equations.

Doesn't this remind you of a basis?

**Theorem 6.8.** *Let $A : X \to Y$ be a linear transformation. Then $A$ is invertible if and only if for any right side $\mathbf{b} \in Y$ the equation*
$$A\mathbf{x} = \mathbf{b}$$
*has a unique solution $\mathbf{x} \in X$.*

**Proof.** Suppose $A$ is invertible. Then $\mathbf{x} = A^{-1}\mathbf{b}$ solves the equation $A\mathbf{x} = \mathbf{b}$. To show that the solution is unique, suppose that for some other vector $\mathbf{x}_1 \in X$
$$A\mathbf{x}_1 = \mathbf{b}$$
Multiplying this identity by $A^{-1}$ from the left we get
$$A^{-1}A\mathbf{x}_1 = A^{-1}\mathbf{b},$$
and therefore $\mathbf{x}_1 = A^{-1}\mathbf{b} = \mathbf{x}$. Note that both identities, $AA^{-1} = I$ and $A^{-1}A = I$ were used here.

Let us now suppose that the equation $A\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x}$ for any $\mathbf{b} \in Y$. Let us use symbol $\mathbf{y}$ instead of $\mathbf{b}$. We know that given $\mathbf{y} \in Y$ the equation
$$A\mathbf{x} = \mathbf{y}$$
has a unique solution $\mathbf{x} \in X$. Let us call this solution $B(\mathbf{y})$.

Note that $B(\mathbf{y})$ is defined for all $\mathbf{y} \in Y$, so we defined a transformation $B : Y \to X$.

Let us check that $B$ is a linear transformation. We need to show that $B(\alpha\mathbf{y}_1+\beta\mathbf{y}_2) = \alpha B(\mathbf{y}_1)+\beta B(\mathbf{y}_2)$. Let $\mathbf{x}_k := B(\mathbf{y}_k)$, $k = 1, 2$, i.e. $A\mathbf{x}_k = \mathbf{y}_k$, $k = 1, 2$. Then
$$A(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2) = \alpha A\mathbf{x}_1 + \beta A\mathbf{x}_2 = \alpha\mathbf{y}_1 + \beta\mathbf{y}_2,$$
which means
$$B(\alpha\mathbf{y}_1 + \beta\mathbf{y}_2) = \alpha B(\mathbf{y}_1) + \beta B(\mathbf{y}_2).$$

And finally, let us show that $B$ is indeed the inverse of $A$. Take $\mathbf{x} \in X$ and let $\mathbf{y} = A\mathbf{x}$, so by the definition of $B$ we have $\mathbf{x} = B\mathbf{y}$. Then for all $\mathbf{x} \in X$

$$BA\mathbf{x} = B\mathbf{y} = \mathbf{x},$$

so $BA = I$. Similarly, for arbitrary $\mathbf{y} \in Y$ let $\mathbf{x} = B\mathbf{y}$, so $\mathbf{y} = A\mathbf{x}$. Then for all $\mathbf{y} \in Y$

$$AB\mathbf{y} = A\mathbf{x} = \mathbf{y}$$

so $AB = I$. □

Recalling the definition of a basis we get the following corollary of Theorems 6.6 and 6.7.

**Corollary 6.9.** *An $m \times n$ matrix is invertible if and only if its columns form a basis in $\mathbb{F}^m$.*

**Exercises.**

**6.1.** Prove, that if $A : V \to W$ is an isomorphism (i.e. an invertible linear transformation) and $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis in $V$, then $A\mathbf{v}_1, A\mathbf{v}_2, \dots, A\mathbf{v}_n$ is a basis in $W$.

**6.2.** Find all right inverses to the $1 \times 2$ matrix (row) $A = (1, 1)$. Conclude from here that the row $A$ is not left invertible.

**6.3.** Find all left inverses of the column $(1, 2, 3)^T$

**6.4.** Is the column $(1, 2, 3)^T$ right invertible? Justify

**6.5.** Find two matrices $A$ and $B$ that $AB$ is invertible, but $A$ and $B$ are not. **Hint:** square matrices $A$ and $B$ would not work. **Remark:** It is easy to construct such $A$ and $B$ in the case when $AB$ is a $1 \times 1$ matrix (a scalar). But can you get $2 \times 2$ matrix $AB$? $3 \times 3$? $n \times n$?

**6.6.** Suppose the product $AB$ is invertible. Show that $A$ is right invertible and $B$ is left invertible. **Hint:** you can just write formulas for right and left inverses.

**6.7.** Let $A$ and $AB$ be invertible (assuming that the product $AB$ is well defined). Prove that $B$ is invertible.

**6.8.** Let $A$ be $n \times n$ matrix. Prove that if $A^2 = \mathbf{0}$ then $A$ is not invertible

**6.9.** Suppose $AB = \mathbf{0}$ for some non-zero matrix $B$. Can $A$ be invertible? Justify.

**6.10.** Write matrices of the linear transformations $T_1$ and $T_2$ in $\mathbb{F}^5$, defined as follows: $T_1$ interchanges the coordinates $x_2$ and $x_4$ of the vector $\mathbf{x}$, and $T_2$ just adds to the coordinate $x_2$ $a$ times the coordinate $x_4$, and does not change other coordinates, i.e.

$$T_1 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_4 \\ x_3 \\ x_2 \\ x_5 \end{pmatrix}, \qquad T_2 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 + ax_4 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix};$$

here $a$ is some fixed number.

Show that $T_1$ and $T_2$ are invertible transformations, and write the matrices of the inverses. **Hint:** it may be simpler, if you first describe the inverse transformation, and then find its matrix, rather than trying to guess (or compute) the inverses of the matrices $T_1$, $T_2$.

**6.11.** Find the matrix of the rotation in $\mathbb{R}^3$ through the angle $\alpha$ around the vector $(1, 2, 3)^T$. We assume that the rotation is counterclockwise if we sit at the tip of the vector and looking at the origin.

You can present the answer as a product of several matrices: you don't have to perform the multiplication.

**6.12.** Give examples of matrices (say $2 \times 2$) such that:

a) $A + B$ is not invertible although both $A$ and $B$ are invertible;

b) $A + B$ is invertible although both $A$ and $B$ are not invertible;

c) All of $A$, $B$ and $A + B$ are invertible

**6.13.** Let $A$ be an invertible symmetric ($A^T = A$) matrix. Is the inverse of $A$ symmetric? Justify.

## 7. Subspaces.

A *subspace* of a vector space $V$ is a non-empty subset $V_0 \subset V$ of $V$ which is closed under the vector addition and multiplication by scalars, i.e.

1. If $\mathbf{v} \in V_0$ then $\alpha\mathbf{v} \in V_0$ for all scalars $\alpha$;
2. For any $\mathbf{u}, \mathbf{v} \in V_0$ the sum $\mathbf{u} + \mathbf{v} \in V_0$;

Again, the conditions 1 and 2 can be replaced by the following one:

$$\alpha\mathbf{u} + \beta\mathbf{v} \in V_0 \qquad \text{for all } \mathbf{u}, \mathbf{v} \in V_0, \text{ and for all scalars } \alpha, \beta.$$

Note, that a subspace $V_0 \subset V$ with the operations (vector addition and multiplication by scalars) inherited from $V$, is a vector space. Indeed, because all operations are inherited from the vector space $V$ they must satisfy all eight axioms of the vector space. The only thing that could possibly go wrong, is that the result of some operation does not belong to $V_0$. But the definition of a subspace prohibits this!

Now let us consider some examples:

1. *Trivial* subspaces of a space $V$, namely $V$ itself and $\{\mathbf{0}\}$ (the subspace consisting only of zero vector). Note, that the empty set $\varnothing$ is not a vector space, since it does not contain a zero vector, so it is not a subspace.

With each linear transformation $A : V \to W$ we can associate the following two subspaces:

2. The *null space*, or *kernel* of $A$, which is denoted as $\operatorname{Null} A$ or $\operatorname{Ker} A$ and consists of all vectors $\mathbf{v} \in V$ such that $A\mathbf{v} = \mathbf{0}$
3. The range $\operatorname{Ran} A$ is defined as the set of all vectors $\mathbf{w} \in W$ which can be represented as $\mathbf{w} = A\mathbf{v}$ for some $\mathbf{v} \in V$.

If $A$ is a matrix, i.e. $A : \mathbb{F}^m \to \mathbb{F}^n$, then recalling *column by coordinate* rule of the matrix–vector multiplication, we can see that any vector $\mathbf{w} \in \operatorname{Ran} A$ can be represented as a linear combination of columns of the matrix $A$. That explains why the term column space (and notation $\operatorname{Col} A$) is often used for the range of the matrix. So, for a matrix $A$, the notation $\operatorname{Col} A$ is often used instead of $\operatorname{Ran} A$.

And now the last example.

4. Given a system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r \in V$ its *linear span* (sometimes called simply *span*) $\mathcal{L}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$ is the collection of all vectors $\mathbf{v} \in V$ that can be represented as a linear combination $\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_r\mathbf{v}_r$ of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$. The notation $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$ is also used instead of $\mathcal{L}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$

It is easy to check that in all of these examples we indeed have subspaces. We leave this an an exercise for the reader. Some of the statements will be proved later in the text.

**Exercises.**

**7.1.** Let $X$ and $Y$ be subspaces of a vector space $V$. Prove that $X \cap Y$ is a subspace of $V$.

**7.2.** Let $V$ be a vector space. For $X, Y \subset V$ the sum $X + Y$ is the collection of all vectors $\mathbf{v}$ which can be represented as $\mathbf{v} = \mathbf{x} + \mathbf{y}$, $\mathbf{x} \in X$, $\mathbf{y} \in Y$. Show that if $X$ and $Y$ are *subspaces* of $V$, then $X + Y$ is also a subspace.

**7.3.** Let $X$ be a subspace of a vector space $V$, and let $\mathbf{v} \in V$, $\mathbf{v} \notin X$. Prove that if $\mathbf{x} \in X$ then $\mathbf{x} + \mathbf{v} \notin X$.

**7.4.** Let $X$ and $Y$ be subspaces of a vector space $V$. Using the previous exercise, show that $X \cup Y$ is a subspace if and only if $X \subset Y$ or $Y \subset X$.

**7.5.** What is the smallest subspace of the space of $4 \times 4$ matrices which contains all upper triangular matrices ($a_{j,k} = 0$ for all $j > k$), and all symmetric matrices ($A = A^T$)? What is the largest subspace contained in both of those subspaces?

## 8. Application to computer graphics.

In this section we give some ideas of how linear algebra is used in computer graphics. We will not go into the details, but just explain some ideas. In particular we explain why manipulation with 3 dimensional images are reduced to multiplications of $4 \times 4$ matrices.

**8.1. 2-dimensional manipulation.** The $x$-$y$ plane (more precisely, a rectangle there) is a good model of a computer monitor. Any object on a monitor is represented as a collection of *pixels*, each pixel is assigned a specific color. Position of each pixel is determined by the column and row, which play role of $x$ and $y$ coordinates on the plane. So a rectangle on a plane with $x$-$y$ coordinates is a good model for a computer screen: and a graphical object is just a collection of points.

**Remark.** There are two types of graphical objects: bitmap objects, where every pixel of an object is described, and vector object, where we describe only *critical points*, and graphic engine connects them to reconstruct the object. A (digital) photo is a good example of a bitmap object: every pixel of it is described. Bitmap object can contain a lot of points, so manipulations with bitmaps require a lot of computing power. Anybody who has edited digital photos in a bitmap manipulation program, like Adobe Photoshop, knows that one needs quite a powerful computer, and even with modern and powerful computers manipulations can take some time.

That is the reason that most of the objects, appearing on a computer screen are vector ones: the computer only needs to memorize critical points. For example, to describe a polygon, one needs only to give the coordinates of its vertices, and which vertex is connected with which. Of course, not all objects on a computer screen can be represented as polygons, some, like letters, have curved smooth boundaries. But there are standard methods allowing one to draw smooth curves through a collection of points, for example Bezier splines, used in PostScript and Adobe PDF (and in many other formats).

Anyhow, this is the subject of a completely different book, and we will not discuss it here. For us a graphical object will be a collection of points (either wireframe model, or bitmap) and we would like to show how one can perform some manipulations with such objects.

The simplest transformation is a translation (shift), where each point (vector) $\mathbf{v}$ is translated by $\mathbf{a}$, i.e. the vector $\mathbf{v}$ is replaced by $\mathbf{v}+\mathbf{a}$ (notation $\mathbf{v} \mapsto \mathbf{v}+\mathbf{a}$ is used for this). Vector addition is very well adapted to computers, so the translation is easy to implement.

Note, that the translation is not a linear transformation (if $\mathbf{a} \neq \mathbf{0}$): while it preserves the straight lines, it does not preserve $\mathbf{0}$.

All other transformation used in computer graphics are linear. The first one that comes to mind is rotation. The rotation by $\gamma$ around the origin $\mathbf{0}$ is given by the multiplication by the rotation matrix $R_\gamma$ we discussed above,

$$R_\gamma = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix}.$$

If we want to rotate around a point $\mathbf{a}$, we first need to translate the picture by $-\mathbf{a}$, moving the point $\mathbf{a}$ to $\mathbf{0}$, then rotate around $\mathbf{0}$ (multiply by $R_\gamma$) and then translate everything back by $\mathbf{a}$.

Another very useful transformation is scaling, given by a matrix

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix},$$

$a, b \geq 0$. If $a = b$ it is *uniform scaling* which enlarges (reduces) an object, preserving its shape. If $a \neq b$ then $x$ and $y$ coordinates scale differently; the object becomes "taller" or "wider".

Another often used transformation is *reflection*: for example the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

defines the reflection through $x$-axis.

We will show later in the book, that any linear transformation in $\mathbb{R}^2$ can be represented either as a composition of scaling rotations and reflections. However it is sometimes convenient to consider some different transformations, like the shear transformation, given by the matrix

$$\begin{pmatrix} 1 & \tan\varphi \\ 0 & 1 \end{pmatrix}.$$

This transformation makes all objects slanted, the horizontal lines remain horizontal, but vertical lines go to the slanted lines at the angle $\varphi$ to the horizontal ones.

**8.2. 3-dimensional graphics.** Three-dimensional graphics is more complicated. First we need to be able to manipulate 3-dimensional objects, and then we need to represent it on 2-dimensional plane (monitor).

The manipulations with 3-dimensional objects is pretty straightforward, we have the same basic transformations: translation, reflection through a plane, scaling, rotation. Matrices of these transformations are very similar to the matrices of their $2 \times 2$ counterparts. For example the matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}, \qquad \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

represent respectively reflection through $x$-$y$ plane, scaling, and rotation around $z$-axis.

Note, that the above rotation is essentially 2-dimensional transformation, it does not change $z$ coordinate. Similarly, one can write matrices for the other 2 *elementary rotations* around $x$ and around $y$ axes. It will be

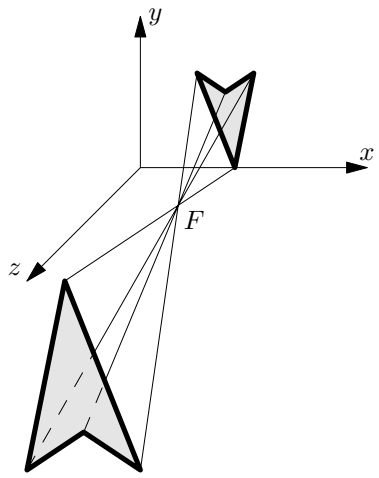


**Figure 2.** Perspective projection onto $x$-$y$ plane: $F$ is the center (focal point) of the projection

shown later that a rotation around an arbitrary axis can be represented as a composition of elementary rotations.

So, we know how to manipulate 3-dimensional objects. Let us now discuss how to represent such objects on a 2-dimensional plane. The simplest way is to project it to a plane, say to the $x$-$y$ plane. To perform such projection one just needs to replace $z$ coordinate by 0, the matrix of this *projection* is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Such method is often used in technical illustrations. Rotating an object and projecting it is equivalent to looking at it from different points. However, this method does not give a very realistic picture, because it does not take into account the perspective, the fact that the objects that are further away look smaller.

To get a more realistic picture one needs to use the so-called *perspective projection*. To define a perspective projection one needs to pick a point (the center of projection or the focal point) and a plane to project onto. Then each point in $\mathbb{R}^3$ is projected into a point on the plane such that the point, its image and the *center of the projection* lie on the same line, see Fig. 2.

This is exactly how a camera works, and it is a reasonable first approximation of how our eyes work.

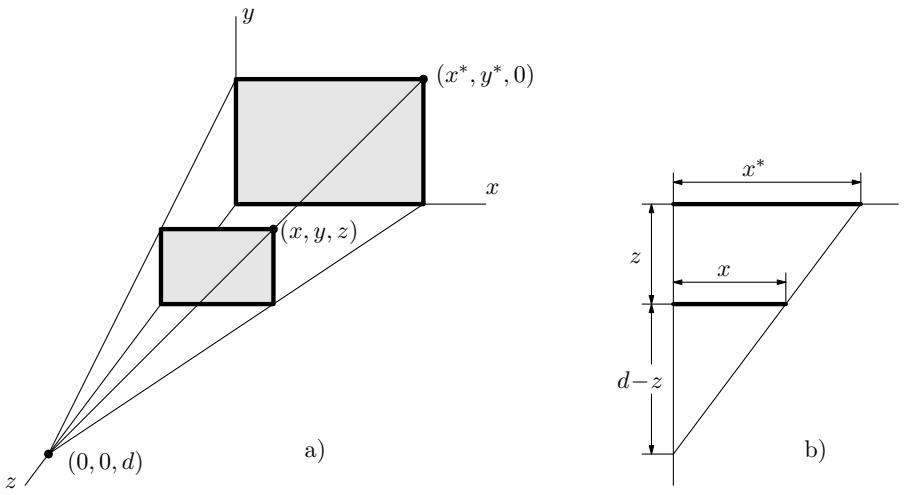


**Figure 3.** Finding coordinates $x^*$, $y^*$ of the perspective projection of the point $(x, y, z)^T$.

Let us get a formula for the projection. Assume that the focal point is $(0, 0, d)^T$ and that we are projecting onto $x$-$y$ plane, see Fig. 3 a). Consider a point $\mathbf{v} = (x, y, z)^T$, and let $\mathbf{v}^* = (x^*, y^*, 0)^T$ be its projection. Analyzing similar triangles see Fig. 3 b), we get that

$$\frac{x^*}{d} = \frac{x}{d-z},$$

so

$$x^* = \frac{xd}{d-z} = \frac{x}{1-z/d},$$

and similarly

$$y^* = \frac{y}{1-z/d}.$$

Note, that this formula also works if $z > d$ and if $z < 0$: you can draw the corresponding similar triangles to check it.

Thus the perspective projection maps a point $(x, y, z)^T$ to the point $\left(\frac{x}{1-z/d}, \frac{y}{1-z/d}, 0\right)^T$.

This transformation is definitely not linear (because of $z$ in the denominator). However it is still possible to represent it as a linear transformation. To do this let us introduce the so-called *homogeneous coordinates*.

In the homogeneous coordinates, every point in $\mathbb{R}^3$ is represented by 4 coordinates, the last, 4th coordinate playing role of the scaling coefficient. Thus, to get usual 3-dimensional coordinates of the vector $\mathbf{v} = (x, y, z)^T$ from its homogeneous coordinates $(x_1, x_2, x_3, x_4)^T$ one needs to divide all

entries by the last coordinate $x_4$ and take the first 3 coordinates [3] (if $x_4 = 0$ this recipe does not work, so we assume that the case $x_4 = 0$ corresponds to the point at infinity).

Thus in homogeneous coordinates the vector $\mathbf{v}^*$ can be represented as $(x, y, 0, 1 - z/d)^T$, so in homogeneous coordinates the perspective projection is a linear transformation:

$$\begin{pmatrix} x \\ y \\ 0 \\ 1 - z/d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1/d & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}.$$

Note that in the homogeneous coordinates the translation is also a linear transformation:

$$\begin{pmatrix} x + a_1 \\ y + a_2 \\ z + a_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & a_1 \\ 0 & 1 & 0 & a_2 \\ 0 & 0 & 1 & a_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}.$$

But what happen if the center of projection is not a point $(0, 0, d)^T$ but some arbitrary point $(d_1, d_2, d_3)^T$. Then we first need to apply the translation by $-(d_1, d_2, 0)^T$ to move the center to $(0, 0, d_3)^T$ while preserving the $x$-$y$ plane, apply the projection, and then move everything back translating it by $(d_1, d_2, 0)^T$. Similarly, if the plane we project to is not $x$-$y$ plane, we move it to the $x$-$y$ plane by using rotations and translations, and so on.

All these operations are just multiplications by $4 \times 4$ matrices. That explains why modern graphic cards have $4 \times 4$ matrix operations embedded in the processor.

Of course, here we only touched the mathematics behind 3-dimensional graphics, there is much more. For example, how to determine which parts of the object are visible and which are hidden, how to make realistic lighting, shades, etc.

**Exercises.**

**8.1.** What vector in $\mathbb{R}^3$ has homogeneous coordinates $(10, 20, 30, 5)$?

[3] If we multiply homogeneous coordinates of a point in $\mathbb{R}^2$ by a non-zero scalar, we do not change the point. In other words, in homogeneous coordinates a point in $\mathbb{R}^3$ is represented by a line through 0 in $\mathbb{R}^4$.

**8.2.** Show that a rotation through $\gamma$ can be represented as a composition of two shear-and-scale transformations

$$T_1 = \begin{pmatrix} 1 & 0 \\ \sin\gamma & \cos\gamma \end{pmatrix}, \qquad T_2 = \begin{pmatrix} \sec\gamma & -\tan\gamma \\ 0 & 1 \end{pmatrix}.$$

In what order the transformations should be taken?

**8.3.** Multiplication of a 2-vector by an arbitrary $2 \times 2$ matrix usually requires 4 multiplications.

Suppose a $2 \times 1000$ matrix $D$ contains coordinates of 1000 points in $\mathbb{R}^2$. How many multiplications are required to transform these points using 2 arbitrary $2 \times 2$ matrices $A$ and $B$. Compare 2 possibilities, $A(BD)$ and $(AB)D$.

**8.4.** Write $4 \times 4$ matrix performing perspective projection to $x$-$y$ plane with center $(d_1, d_2, d_3)^T$.

**8.5.** A transformation $T$ in $\mathbb{R}^3$ is a rotation about the line $y = x+3$ in the $x$-$y$ plane through an angle $\gamma$. Write a $4 \times 4$ matrix corresponding to this transformation.

You can leave the result as a product of matrices.

*Chapter 2*

# Systems of linear equations

### 1. Different faces of linear systems.

There exist several points of view on what a system of linear equations, or in short a *linear system* is. The first, naïve one is, that it is simply a collection of $m$ linear equations with $n$ unknowns $x_1, x_2, \ldots, x_n$,

$$\begin{cases} a_{1,1}x_1 & + & a_{1,2}x_2 & + & \ldots & + & a_{1,n}x_n & = & b_1 \\ a_{2,1}x_1 & + & a_{2,2}x_2 & + & \ldots & + & a_{2,n}x_n & = & b_2 \\ \quad\ldots & & & & & & & & \\ a_{m,1}x_1 & + & a_{m,2}x_2 & + & \ldots & + & a_{m,n}x_n & = & b_m \end{cases} .$$

To solve the system is to find *all* $n$-tuples of numbers $x_1, x_2, \ldots, x_n$ which satisfy all $m$ equations simultaneously.

If we denote $\mathbf{x} := (x_1, x_2, \ldots, x_n)^T \in \mathbb{F}^n$, $\mathbf{b} = (b_1, b_2, \ldots, b_m)^T \in \mathbb{F}^m$, and

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \ldots & a_{m,n} \end{pmatrix},$$

then the above linear system can be written in the *matrix form* (as a *matrix-vector equation*)

$$A\mathbf{x} = \mathbf{b}.$$

To solve the above equation is to find all vectors $\mathbf{x} \in \mathbb{F}^n$ satisfying $A\mathbf{x} = \mathbf{b}$.

And finally, recalling the "column by coordinate" rule of the matrix-vector multiplication, we can write the system as a *vector equation*

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots + x_n\mathbf{a}_n = \mathbf{b},$$

where $\mathbf{a}_k$ is the $k$th column of the matrix $A$, $\mathbf{a}_k = (a_{1,k}, a_{2,k}, \ldots, a_{m,k})^T$, $k = 1, 2, \ldots, n$.

Note, these three examples are essentially just different representations of the same mathematical object.

Before explaining how to solve a linear system, let us notice that it does not matter what we call the unknowns, $x_k$, $y_k$ or something else. So, all the information necessary to solve the system is contained in the matrix $A$, which is called *the coefficient matrix* of the system and in the vector (right side) $\mathbf{b}$. Hence, all the information we need is contained in the following matrix

$$\left(\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \ldots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \ldots & a_{m,n} & b_m \end{array}\right)$$

which is obtained by attaching the column $\mathbf{b}$ to the matrix $A$. This matrix is called the *augmented matrix* of the system. We will usually put the vertical line separating $A$ and $\mathbf{b}$ to distinguish between the augmented matrix and the coefficient matrix.

## 2. Solution of a linear system. Echelon and reduced echelon forms

Linear system are solved by the *Gauss–Jordan elimination* (which is sometimes called *row reduction*). By performing operations on rows of the augmented matrix of the system (i.e. on the equations), we reduce it to a simple form, the so-called *echelon form*. When the system is in the *echelon form*, one can easily write the solution.

**2.1. Row operations.** There are three types of row operations we use:

1. Row exchange: interchange two rows of the matrix;
2. Scaling: multiply a row by a non-zero scalar $a$;
3. Row replacement: replace a row # $k$ by its sum with a constant multiple of a row # $j$; all other rows remain intact;

It is clear that the operations 1 and 2 do not change the set of solutions of the system; they essentially do not change the system.

As for the operation 3, one can easily see that it does not lose solutions. Namely, let a "new" system be obtained from an "old" one by a row operation of type 3. Then any solution of the "old" system is a solution of the "new" one.

To see that we do not gain anything extra, i.e. that any solution of the "new" system is also a solution of the "old" one, we just notice that row operation of type 3 are *reversible*, i.e. the "old' system also can be obtained from the "new" one by applying a row operation of type 3 (can you say which one?)

2.1.1. *Row operations and multiplication by elementary matrices.* There is another, more "advanced" explanation why the above row operations are legal. Namely, every row operation is equivalent to the multiplication of the matrix from the left by one of the special elementary matrices.

Namely, the multiplication by the matrix

$$
\begin{array}{c}
\\
\\
\\
\\
j \\
\\
\\
\\
k \\
\\
\\
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc} & & & & j & & & & & k & \end{array} \\
\left(\begin{array}{ccccccccccc}
1 & & & & \vdots & & & & \vdots & & \\
& \ddots & & & \vdots & & & & \vdots & 0 & \\
& & & 1 & \vdots & & & & \vdots & & \\
\dots & \dots & \dots & 0 & \dots & \dots & \dots & 1 & & & \\
& & & \vdots & 1 & & & \vdots & & & \\
& & & \vdots & & \ddots & & \vdots & & & \\
& & & \vdots & & & 1 & \vdots & & & \\
\dots & \dots & \dots & 1 & \dots & \dots & \dots & 0 & & & \\
& & & & & & & & 1 & & \\
& & 0 & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{array}\right)
\end{array}
$$

just interchanges the rows number $j$ and number $k$. Multiplication by the matrix

$$
\begin{array}{c}
\\
\\
\\
k \\
\\
\\
\end{array}
\begin{array}{c}
\left(\begin{array}{ccccccc}
1 & & & \vdots & & & \\
& \ddots & & \vdots & & 0 & \\
& & 1 & 0 & & & \\
\dots & \dots & 0 & a & 0 & & \\
& & & 0 & 1 & & \\
& 0 & & & & \ddots & 0 \\
& & & & & 0 & 1
\end{array}\right)
\end{array}
$$

multiplies the row number $k$ by $a$. Finally, multiplication by the matrix

$$\begin{array}{c} \\ \\ j \\ \\ k \\ \\ \end{array} \begin{pmatrix} 1 & & \vdots & & \vdots & & \\ & \ddots & \vdots & & \vdots & 0 & \\ \dots & \dots & 1 & \dots & 0 & & \\ & & \vdots & \ddots & \vdots & & \\ \dots & \dots & a & \dots & 1 & & \\ & 0 & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

adds to the row $\#k$ row $\#j$ multiplied by $a$, and leaves all other rows intact.

A way to describe (or to remember) these elementary matrices: they are obtained from $I$ by applying the corresponding row operation to it

To see, that the multiplication by these matrices works as advertised, one can just see how the multiplications act on vectors (columns).

Note that all these matrices are invertible (compare with reversibility of row operations). The inverse of the first matrix is the matrix itself. To get the inverse of the second one, one just replaces $a$ by $1/a$. And finally, the inverse of the third matrix is obtained by replacing $a$ by $-a$. To see that the inverses are indeed obtained this way, one again can simply check how they act on columns.

So, performing a row operation on the augmented matrix of the system $A\mathbf{x} = \mathbf{b}$ is equivalent to the multiplication of the system (from the left) by a special invertible matrix $E$. Left multiplying the equality $A\mathbf{x} = \mathbf{b}$ by $E$ we get that any solution of the equation

$$A\mathbf{x} = \mathbf{b}$$

is also a solution of

$$EA\mathbf{x} = E\mathbf{b}.$$

Multiplying this equation (from the left) by $E^{-1}$ we get that any of its solutions is a solution of the equation

$$E^{-1}EA\mathbf{x} = E^{-1}E\mathbf{b},$$

which is the original equation $A\mathbf{x} = \mathbf{b}$. So, a row operation does not change the solution set of a system.

**2.2. Row reduction.** The main step of row reduction consists of three sub-steps:

1. Find the leftmost non-zero column of the matrix;
2. Make sure, by applying row operations of type 1 (row exchange), if necessary, that the first (the upper) entry of this column is non-zero. This entry will be called the *pivot entry* or simply the *pivot*;

3. "Kill" (i.e. make them 0) all non-zero entries below the pivot by adding (subtracting) an appropriate multiple of the first row from the rows number $2, 3, \dots, m$.

We apply the main step to a matrix, then we leave the first row alone and apply the main step to rows $2, \dots, m$, then to rows $3, \dots, m$, etc.

The point to remember is that after we subtract a multiple of a row from all rows below it (step 3), we leave it alone and do not change it in any way, not even interchange it with another row.

After applying the main step finitely many times (at most $m$), we get what is called the *echelon form* of the matrix.

2.2.1. *An example of row reduction.* Let us consider the following linear system:

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 1 \\ 3x_1 + 2x_2 + x_3 = 7 \\ 2x_1 + x_2 + 2x_3 = 1 \end{cases}$$

The augmented matrix of the system is

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 3 & 2 & 1 & 7 \\ 2 & 1 & 2 & 1 \end{array}\right)$$

Subtracting 3·Row#1 from the second row, and subtracting 2·Row#1 from the third one we get:

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 3 & 2 & 1 & 7 \\ 2 & 1 & 2 & 1 \end{array}\right) \begin{array}{l} \\ -3R_1 \\ -2R_1 \end{array} \quad \sim \quad \left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & -4 & -8 & 4 \\ 0 & -3 & -4 & -1 \end{array}\right)$$

Multiplying the second row by $-1/4$ we get

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 1 & 2 & -1 \\ 0 & -3 & -4 & -1 \end{array}\right)$$

Adding 3·Row#2 to the third row we obtain

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 1 & 2 & -1 \\ 0 & -3 & -4 & -1 \end{array}\right) -3R_2 \quad \sim \quad \left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 2 & -4 \end{array}\right)$$

Now we can use the so called *back substitution* to solve the system. Namely, from the last row (equation) we get $x_3 = -2$. Then from the second equation we get

$$x_2 = -1 - 2x_3 = -1 - 2(-2) = 3,$$

and finally, from the first row (equation)

$$x_1 = 1 - 2x_2 - 3x_3 = 1 - 6 + 6 = 1.$$

So, the solution is

$$\begin{cases} x_1 = 1 \\ x_2 = 3, \\ x_3 = -2, \end{cases}$$

or in vector form

$$\mathbf{x} = \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix}$$

or $\mathbf{x} = (1, 3, -2)^T$. We can check the solution by multiplying $A\mathbf{x}$, where $A$ is the coefficient matrix.

Instead of using back substitution, we can do row reduction from bottom to top, killing all the entries above the main diagonal of the coefficient matrix: we start by multiplying the last row by $1/2$, and the rest is pretty self-explanatory:

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 1 & -2 \end{array}\right) \begin{matrix} \\ \\ -2R_3 \end{matrix} \begin{matrix} -3R_3 \\ \\ \end{matrix} \sim \left(\begin{array}{ccc|c} 1 & 2 & 0 & 7 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -2 \end{array}\right) -2R_2$$

$$\sim \left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -2 \end{array}\right)$$

and we just read the solution $\mathbf{x} = (1, 3, -2)^T$ off the augmented matrix.

We leave it as an exercise to the reader to formulate the algorithm for the backward phase of the row reduction.

**2.3. Echelon form.** A matrix is in *echelon form* if it satisfies the following two conditions:

1. All zero rows (i.e. the rows with all entries equal 0), if any, are below all non-zero entries.

For a non-zero row, let us call the leftmost non-zero entry the *leading entry*. Then the second property of the echelon form can be formulated as follows:

2. For any non-zero row its leading entry is strictly to the right of the leading entry in the previous row.

Pivots: leading (leftmost non-zero entries) in a row.

The leading entry in each row in echelon form is also called *pivot entry*, or simply *pivot*, because these entries are exactly the *pivots* we used in the row reduction.

A particular case of the echelon form is the so-called *triangular* form. We got this form in our example above. In this form the coefficient matrix is square ($n \times n$), all its entries on the main diagonal are non-zero, and all the

entries below the main diagonal are zero. The right side, i.e. the rightmost column of the augmented matrix can be arbitrary.

After the backward phase of the row reduction, we get what the so-called *reduced echelon form* of the matrix: coefficient matrix equal $I$, as in the above example, is a particular case of the reduced echelon form.

The general definition is as follows: we say that a matrix is in the *reduced echelon form*, if it is in the echelon form and

3. All pivot entries are equal 1;
4. All entries above the pivots are 0. Note, that all entries below the pivots are also 0 because of the echelon form.

To get reduced echelon form from echelon form, we work from the bottom to the top and from the right to the left, using row replacement to kill all entries above the pivots.

An example of the reduced echelon form is the system with the coefficient matrix equal $I$. In this case, one just reads the solution from the reduced echelon form. In general case, one can also easily read the solution from the reduced echelon form. For example, let the reduced echelon form of the system (augmented matrix) be

$$\left(\begin{array}{ccccc|c} \boxed{1} & 2 & 0 & 0 & 0 & 1 \\ 0 & 0 & \boxed{1} & 5 & 0 & 2 \\ 0 & 0 & 0 & 0 & \boxed{1} & 3 \end{array}\right);$$

here we boxed the pivots. The idea is to move the variables, corresponding to the columns without pivot (the so-called *free variables*) to the right side. Then we can just write the solution.

$$\begin{cases} x_1 = 1 - 2x_2 \\ x_2 \text{ is free} \\ x_3 = 2 - 5x_4 \\ x_4 \text{ is free} \\ x_5 = 3 \end{cases}$$

or in the vector form

$$\mathbf{x} = \begin{pmatrix} 1 - 2x_2 \\ x_2 \\ 2 - 5x_4 \\ x_4 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 3 \end{pmatrix} + x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 0 \\ -5 \\ 1 \\ 0 \end{pmatrix}, \quad x_2, x_4 \in \mathbb{F}.$$

One can also find the solution from the echelon form by using back substitution: the idea is to work from bottom to top, moving all free variables to the right side.

**Exercises.**

**2.1.** Write the systems of equations below in matrix form and as vector equations:

a) $\begin{cases} x_1 + 2x_2 - x_3 = -1 \\ 2x_1 + 2x_2 + x_3 = 1 \\ 3x_1 + 5x_2 - 2x_3 = -1 \end{cases}$

b) $\begin{cases} x_1 - 2x_2 - x_3 = 1 \\ 2x_1 - 3x_2 + x_3 = 6 \\ 3x_1 - 5x_2 = 7 \\ x_1 + 5x_3 = 9 \end{cases}$

c) $\begin{cases} x_1 + 2x_2 + 2x_4 = 6 \\ 3x_1 + 5x_2 - x_3 + 6x_4 = 17 \\ 2x_1 + 4x_2 + x_3 + 2x_4 = 12 \\ 2x_1 - 7x_3 + 11x_4 = 7 \end{cases}$

d) $\begin{cases} x_1 - 4x_2 - x_3 + x_4 = 3 \\ 2x_1 - 8x_2 + x_3 - 4x_4 = 9 \\ -x_1 + 4x_2 - 2x_3 + 5x_4 = -6 \end{cases}$

e) $\begin{cases} x_1 + 2x_2 - x_3 + 3x_4 = 2 \\ 2x_1 + 4x_2 - x_3 + 6x_4 = 5 \\ x_2 + 2x_4 = 3 \end{cases}$

f) $\begin{cases} 2x_1 - 2x_2 - x_3 + 6x_4 - 2x_5 = 1 \\ x_1 - x_2 + x_3 + 2x_4 - x_5 = 2 \\ 4x_1 - 4x_2 + 5x_3 + 7x_4 - x_5 = 6 \end{cases}$

g) $\begin{cases} 3x_1 - x_2 + x_3 - x_4 + 2x_5 = 5 \\ x_1 - x_2 - x_3 - 2x_4 - x_5 = 2 \\ 5x_1 - 2x_2 + x_3 - 3x_4 + 3x_5 = 10 \\ 2x_1 - x_2 - 2x_4 + x_5 = 5 \end{cases}$

Solve the systems. Write the answers in the vector form.

**2.2.** Find *all* solutions of the vector equation

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_3\mathbf{v}_3 = \mathbf{0},$$

where $\mathbf{v}_1 = (1,1,0)^T$, $\mathbf{v}_2 = (0,1,1)^T$ and $\mathbf{v}_3 = (1,0,1)^T$. What conclusion can you make about linear independence (dependence) of the system of vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$?

## 3. Analyzing the pivots.

All questions about existence of a solution and it uniqueness can be answered by analyzing pivots in the echelon (reduced echelon) form of the augmented matrix of the system. First of all, let us investigate the question of when is the equation $A\mathbf{x} = \mathbf{b}$ *inconsistent*, i.e. when it does not have a solution. The answer follows immediately, if one just thinks about it:

a system is inconsistent (does not have a solution) if and only if there is a pivot in the last column of an echelon form of the *augmented* matrix, i.e. iff an echelon form of the augmented matrix has a row $\left(\begin{array}{ccccc|c} 0 & 0 & \dots & 0 & b \end{array}\right)$, $b \neq 0$ in it.

Indeed, such a row correspond to the equation $0x_1 + 0x_2 + \ldots + 0x_n = b \neq 0$ that does not have a solution. If we don't have such a row, we just make the reduced echelon form and then read the solution off it.

Now, three more statements. Note, they all deal with the *coefficient matrix*, and not with the augmented matrix of the system.

1. A solution (if it exists) is unique iff there are no free variables, that is if and only if the echelon form of the coefficient matrix has a pivot in every column;
2. Equation $A\mathbf{x} = \mathbf{b}$ is consistent for all right sides $\mathbf{b}$ if and only if the echelon form of the coefficient matrix has a pivot in every row.
3. Equation $A\mathbf{x} = \mathbf{b}$ has a *unique solution* for any right side $\mathbf{b}$ if and only if echelon form of the coefficient matrix $A$ has a pivot in every column and every row.

The first statement is trivial, because free variables are responsible for all non-uniqueness. I should only emphasize that this statement *does not say anything* about the existence.

The second statement is a tiny bit more complicated. If we have a pivot in every row of the coefficient matrix, we cannot have the pivot in the last column of the *augmented* matrix, so the system is always consistent, no matter what the right side $\mathbf{b}$ is.

Let us show that if we have a zero row in the echelon form of the coefficient matrix $A$, then we can pick a right side $\mathbf{b}$ such that the system $A\mathbf{x} = \mathbf{b}$ is not consistent. Let $A_e$ echelon form of the coefficient matrix $A$. Then

$$A_e = EA,$$

where $E$ is the product of elementary matrices, corresponding to the row operations, $E = E_N \dots E_2 E_1$. If $A_e$ has a zero row, then the last row is also zero. Therefore, if we put $\mathbf{b}_e = (0, \dots, 0, 1)^T$ (all entries are 0, except the last one), then the equation

$$A_e\mathbf{x} = \mathbf{b}_e$$

does not have a solution. Multiplying this equation by $E^{-1}$ from the left, and recalling that $E^{-1}A_e = A$, we get that the equation

$$A\mathbf{x} = E^{-1}\mathbf{b}_e$$

does not have a solution.

Finally, statement 3 immediately follows from statements 1 and 2. □

From the above analysis of pivots we get several very important corollaries. The main observation we use is

| In echelon form, any row and any column have no more than 1 pivot in it (it can have 0 pivots) |
|---|

**3.1. Corollaries about linear independence and bases. Dimension.** Questions as to when a system of vectors in $\mathbb{F}^n$ is a basis, a linearly independent or a spanning system, can be easily answered by the row reduction.

**Proposition 3.1.** *Let us have a system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in \mathbb{F}^n$, and let $A = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$ be an $n \times m$ matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$. Then*

1. *The system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ is linearly independent iff echelon form of $A$ has a pivot in every column;*
2. *The system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ is complete in $\mathbb{F}^n$ (spanning, generating) iff echelon form of $A$ has a pivot in every row;*
3. *The system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ is a basis in $\mathbb{F}^n$ iff echelon form of $A$ has a pivot in every column and in every row.*

**Proof.** The system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in \mathbb{F}^n$ is linearly independent if and only if the equation

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_m\mathbf{v}_m = \mathbf{0}$$

has the unique (trivial) solution $x_1 = x_2 = \ldots = x_m = 0$, or equivalently, the equation $A\mathbf{x} = \mathbf{0}$ has unique solution $\mathbf{x} = \mathbf{0}$. By statement 1 above, it happens if and only if there is a pivot in every column of the matrix.

Similarly, the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in \mathbb{F}^n$ is complete in $\mathbb{F}^n$ if and only if the equation

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_m\mathbf{v}_m = \mathbf{b}$$

has a solution for any right side $\mathbf{b} \in \mathbb{F}^n$. By statement 2 above, it happens if and only if there is a pivot in every row in echelon form of the matrix.

And finally, the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in \mathbb{F}^n$ is a basis in $\mathbb{F}^n$ if and only if the equation

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_m\mathbf{v}_m = \mathbf{b}$$

has unique solution for any right side $\mathbf{b} \in \mathbb{F}^n$. By statement 3 this happens if and only if there is a pivot in every column and in every row of echelon form of $A$. □

**Proposition 3.2.** *Any linearly independent system of vectors in $\mathbb{F}^n$ cannot have more than $n$ vectors in it.*

**Proof.** Let a system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in \mathbb{F}^n$ be linearly independent, and let $A = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$ be the $n \times m$ matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$. By Proposition 3.1 echelon form of $A$ must have a pivot in every column, which is impossible if $m > n$ (number of pivots cannot be more than number of rows). □

**Proposition 3.3.** *Any two bases in a vector space $V$ have the same number of vectors in them.*

**Proof.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$ be two different bases in $V$. Without loss of generality we can assume that $n \le m$. Consider an isomorphism $A : \mathbb{F}^n \to V$ defined by

$$A\mathbf{e}_k = \mathbf{v}_k, \qquad k = 1, 2, \dots n,$$

where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is the standard basis in $\mathbb{R}^n$.

Since $A^{-1}$ is also an isomorphism, the system

$$A^{-1}\mathbf{w}_1, A^{-1}\mathbf{w}_2, \dots, A^{-1}\mathbf{w}_m$$

is a basis (see Theorem 6.6 in Chapter 1). So it is linearly independent, and by Proposition 3.2, $m \le n$. Together with the assumption $n \le m$ this implies that $m = n$. □

The statement below is a particular case of the above proposition.

**Proposition 3.4.** *Any basis in $\mathbb{F}^n$ must have exactly $n$ vectors in it.*

**Proof.** This fact follows immediately from the previous proposition, but there is also a direct proof. Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ be a basis in $\mathbb{F}^n$ and let $A$ be the $n \times m$ matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$. The fact that the system is a basis, means that the equation

$$A\mathbf{x} = \mathbf{b}$$

has a unique solution for any (all possible) right side $\mathbf{b}$. The existence means that there is a pivot in every row (of a reduced echelon form of the matrix), hence the number of pivots is exactly $n$. The uniqueness mean that there is pivot in every column of the coefficient matrix (its echelon form), so

$$m = \text{number of columns} = \text{number of pivots} = n$$

□

**Proposition 3.5.** *Any spanning (generating) set in $\mathbb{F}^n$ must have at least $n$ vectors.*

**Proof.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ be a complete system in $\mathbb{F}^n$, and let $A$ be $n \times m$ matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$. Statement 2 of Proposition 3.1 implies

that echelon form of $A$ has a pivot in every row. Since number of pivots cannot exceed the number of columns, $n \leq m$. □

### 3.2. Corollaries about invertible matrices.

**Proposition 3.6.** *A matrix $A$ is invertible if and only if its echelon form has pivot in every column and every row.*

**Proof.** As it was discussed in the beginning of the section, the equation $A\mathbf{x} = \mathbf{b}$ has a unique solution for any right side $\mathbf{b}$ if and only if the echelon form of $A$ has pivot in every row and every column. But, we know, see Theorem 6.8 in Chapter 1, that the matrix (linear transformation) $A$ is invertible if and only if the equation $A\mathbf{x} = \mathbf{b}$ has a unique solution for any possible right side $\mathbf{b}$.

There is also an alternative proof. We know that a matrix is invertible if and only if its columns form a basis in (see Corollary 6.9 in Section 6.4, Chapter 1). Proposition 3.4 above states that it happens if and only if there is a pivot in every row and every column. □

The above proposition immediately implies the following

**Corollary 3.7.** *An invertible matrix must be square ($n \times n$).*

**Proposition 3.8.** *If a square ($n \times n$) matrix is left invertible, or if it is right invertible, then it is invertible. In other words, to check the invertibility of a square matrix $A$ it is sufficient to check only one of the conditions $AA^{-1} = I$, $A^{-1}A = I$.*

Note, that this proposition applies only to square matrices!

**Proof.** We know that matrix $A$ is invertible if and only if the equation $A\mathbf{x} = \mathbf{b}$ has a unique solution for any right side $\mathbf{b}$. This happens if and only if the echelon form of the matrix $A$ has pivots in every row and in every column.

If a matrix $A$ is left invertible, the equation $A\mathbf{x} = \mathbf{0}$ has unique solution $\mathbf{x} = \mathbf{0}$. Indeed, if $B$ is a left inverse of $A$ (i.e. $BA = I$), and $\mathbf{x}$ satisfies

$$A\mathbf{x} = \mathbf{0},$$

then multiplying this identity by $B$ from the left we get $\mathbf{x} = \mathbf{0}$, so the solution is unique. Therefore, the echelon form of $A$ has pivots in every column (no free variables). If the matrix $A$ is square ($n \times n$), the echelon form also has pivots in every row ($n$ pivots, and a row can have no more than 1 pivot), so the matrix is invertible.

If a matrix $A$ is right invertible, and $C$ is its right inverse ($AC = I$), then for $\mathbf{x} = C\mathbf{b}$, $\mathbf{b} \in \mathbb{F}^n$

$$A\mathbf{x} = AC\mathbf{b} = I\mathbf{b} = \mathbf{b}.$$

Therefore, for any right side $\mathbf{b}$ the equation $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x} = C\mathbf{b}$. Thus, echelon form of $A$ has a pivot in every row. If $A$ is square, it also has a pivot in every column, so $A$ is invertible. □

**Exercises.**

**3.1.** For what value of $b$ the system

$$\begin{pmatrix} 1 & 2 & 2 \\ 2 & 4 & 6 \\ 1 & 2 & 3 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 4 \\ b \end{pmatrix}$$

has a solution. Find the general solution of the system for this value of $b$.

**3.2.** Determine, if the vectors

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

are linearly independent or not.

Do these four vectors span $\mathbb{R}^4$? (In other words, is it a generating system?) What about $\mathbb{C}^4$?

**3.3.** Determine, which of the following systems of vectors are bases in $\mathbb{R}^3$:

a) $(1, 2, -1)^T$, $(1, 0, 2)^T$, $(2, 1, 1)^T$;

b) $(-1, 3, 2)^T$, $(-3, 1, 3)^T$, $(2, 10, 2)^T$;

c) $(67, 13, -47)^T$, $(\pi, -7.84, 0)^T$, $(3, 0, 0)^T$.

Which of the systems are bases in $\mathbb{C}^3$?

**3.4.** Do the polynomials $x^3 + 2x$, $x^2 + x + 1$, $x^3 + 5$ generate (span) $\mathbb{P}_3$? Justify your answer.

**3.5.** Can 5 vectors in $\mathbb{F}^4$ be linearly independent? Justify your answer.

**3.6.** Prove or disprove: If the columns of a square $(n \times n)$ matrix $A$ are linearly independent, so are the columns of $A^2 = AA$.

**3.7.** Prove or disprove: If the columns of a square $(n \times n)$ matrix $A$ are linearly independent, so are the rows of $A^3 = AAA$.

**3.8.** Show that if the equation $A\mathbf{x} = \mathbf{0}$ has unique solution (i.e. if echelon form of $A$ has pivot in every column), then $A$ is left invertible. **Hint:** elementary matrices may help.
**Note:** It was shown in the text that if $A$ is left invertible, then the equation $A\mathbf{x} = \mathbf{0}$ has unique solution. But here you are asked to prove the converse of this statement, which was not proved.

**Remark:** This can be a very hard problem, for it requires deep understanding of the subject. However, when you understand what to do, the problem becomes almost trivial.

**3.9.** Is the reduced echelon form of a matrix unique? Justify your conclusion.

Namely, suppose that by performing some row operations (not necessarily following any algorithm) we end up with a reduced echelon matrix. Do we always end up with the same matrix, or can we get different ones? Note that we are only allowed to perform row operations, the "column operations"' are forbidden.

**Hint:** What happens if you start with an invertible matrix? Also, are the pivots always in the same columns, or can it be different depending on what row operations you perform? If you can tell what the pivot columns are without reverting to row operations, then the positions of pivot columns do not depend on them.

## 4. Finding $A^{-1}$ by row reduction.

As it was discussed above, an invertible matrix must be square, and its echelon form must have pivots in every row and every column. Therefore reduced echelon form of an invertible matrix is the identity matrix $I$. Therefore,

| Any invertible matrix is row equivalent (i.e. can be reduced by row operations) to the identity matrix. |
|---|

Now let us state a simple algorithm of finding the inverse of an $n \times n$ matrix:

1. Form an *augmented* $n \times 2n$ matrix $(A \,|\, I)$ by writing the $n \times n$ identity matrix right of $A$;
2. Performing row operations on the augmented matrix transform $A$ to the identity matrix $I$;
3. The matrix $I$ that we added will be automatically transformed to $A^{-1}$;
4. If it is impossible to transform $A$ to the identity by row operations, $A$ is not invertible.

There are several possible explanations of the above algorithm. The first, a naïve one, is as follows: we know that (for an invertible $A$) the vector $A^{-1}\mathbf{b}$ is the solution of the equation $A\mathbf{x} = \mathbf{b}$. So to find the column number $k$ of $A^{-1}$ we need to find the solution of $A\mathbf{x} = \mathbf{e}_k$, where $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ is the standard basis in $\mathbb{R}^n$. The above algorithm just solves the equations

$$A\mathbf{x} = \mathbf{e}_k, \qquad k = 1, 2, \ldots, n$$

simultaneously!

Let us also present another, more "advanced" explanation. As we discussed above, every row operation can be realized as a left multiplication

by an elementary matrix. Let $E_1, E_2, \ldots, E_N$ be the elementary matrices corresponding to the row operation we performed, and let $E = E_N \cdots E_2 E_1$ be their product.[1] We know that the row operations transform $A$ to identity, i.e. $EA = I$, so $E = A^{-1}$. But the same row operations transform the augmented matrix $(A \,|\, I)$ to $(EA \,|\, E) = (I \,|\, A^{-1})$. □

This "advanced" explanation using elementary matrices implies an important proposition that will be often used later.

**Theorem 4.1.** *Any invertible matrix can be represented as a product of elementary matrices.*

**Proof.** As we discussed in the previous paragraph, $A^{-1} = E_N \cdots E_2 E_1$, so

$$A = (A^{-1})^{-1} = E_1^{-1} E_2^{-1} \cdots E_N^{-1}.$$

□

**An Example.** Suppose we want to find the inverse of the matrix

$$\begin{pmatrix} 1 & 4 & -2 \\ -2 & -7 & 7 \\ 3 & 11 & -6 \end{pmatrix}.$$

Augmenting the identity matrix to it and performing row reduction we get

$$\left(\begin{array}{ccc|ccc} 1 & 4 & -2 & 1 & 0 & 0 \\ -2 & -7 & 7 & 0 & 1 & 0 \\ 3 & 11 & -6 & 0 & 0 & 1 \end{array}\right)\begin{matrix} \\ +2R_1 \\ -3R_1 \end{matrix} \sim \left(\begin{array}{ccc|ccc} 1 & 4 & -2 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & -1 & 0 & -3 & 0 & 1 \end{array}\right)\begin{matrix} \\ \\ +R_2 \end{matrix} \sim$$

$$\left(\begin{array}{ccc|ccc} 1 & 4 & -2 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & 0 & 3 & -1 & 1 & 1 \end{array}\right)\begin{matrix} \\ \\ \times 3 \end{matrix} \sim \left(\begin{array}{ccc|ccc} 3 & 12 & -6 & 3 & 0 & 0 \\ 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & 0 & 3 & -1 & 1 & 1 \end{array}\right)\begin{matrix} +2R_3 \\ -R_3 \\ \\ \end{matrix} \sim$$

Here in the last row operation we multiplied the first row by 3 to avoid fractions in the backward phase of row reduction. Continuing with the row reduction we get

$$\left(\begin{array}{ccc|ccc} 3 & 12 & 0 & 1 & 2 & 2 \\ 0 & 1 & 0 & 3 & 0 & -1 \\ 0 & 0 & 3 & -1 & 1 & 1 \end{array}\right)\begin{matrix} -12R_2 \\ \\ \\ \end{matrix} \sim \left(\begin{array}{ccc|ccc} 3 & 0 & 0 & -35 & 2 & 14 \\ 0 & 1 & 0 & 3 & 0 & -1 \\ 0 & 0 & 3 & -1 & 1 & 1 \end{array}\right)$$

Dividing the first and the last row by 3 we get the inverse matrix

$$\begin{pmatrix} -35/3 & 2/3 & 14/3 \\ 3 & 0 & -1 \\ -1/3 & 1/3 & 1/3 \end{pmatrix}$$

[1] Although it does not matter here, but please notice, that if the row operation $E_1$ was performed first, $E_1$ must be the rightmost term in the product

**Exercises.**

**4.1.** Find the inverse of the matrices

$$\begin{pmatrix} 1 & 2 & 1 \\ 3 & 7 & 3 \\ 2 & 3 & 4 \end{pmatrix}, \qquad \begin{pmatrix} 1 & -1 & 2 \\ 1 & 1 & -2 \\ 1 & 1 & 4 \end{pmatrix}.$$

Show all steps

## 5. Dimension. Finite-dimensional spaces.

**Definition.** The dimension $\dim V$ of a vector space $V$ is the number of vectors in a basis.

For a vector space consisting only of zero vector $\mathbf{0}$ we put $\dim V = 0$. If $V$ does not have a (finite) basis, we put $\dim V = \infty$.

If $\dim V$ is finite, we call the space $V$ *finite-dimensional*; otherwise we call it *infinite-dimensional*.

Proposition 3.3 asserts that the dimension is well defined, i.e. that it does not depend on the choice of a basis.

Proposition 2.8 from Chapter 1 states that any finite spanning system in a vector space $V$ contains a basis. This immediately implies the following

**Proposition 5.1.** *A vector space $V$ is finite-dimensional if and only if it has a finite spanning system.*

Suppose, that we have a system of vectors in a finite-dimensional vector space, and we want to check if it is a basis (or if it is linearly independent, or if it is complete)? Probably the simplest way is to use an isomorphism $A : V \to \mathbb{R}^n$, $n = \dim E$ to move the problem to $\mathbb{R}^n$, where all such questions can be answered by row reduction (studying pivots).

Note, that if $\dim V = n$, then there always exists an isomorphism $A : V \to \mathbb{R}^n$. Indeed, if $\dim V = n$ then there exists a basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in V$, and one can define an isomorphism $A : V \to \mathbb{R}^n$ by

$$A\mathbf{v}_k = \mathbf{e}_k, \qquad k = 1, 2, \dots, n.$$

As an example, let us give the following two corollaries of the above Propositions 3.2, 3.5:

**Proposition 5.2.** *Any linearly independent system in a finite-dimensional vector space $V$ cannot have more than $\dim V$ vectors in it.*

**Proof.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in V$ be a linearly independent system, and let $A : V \to \mathbb{R}^n$ be an isomorphism. Then $A\mathbf{v}_1, A\mathbf{v}_2, \dots, A\mathbf{v}_m$ is a linearly independent system in $\mathbb{R}^n$, and by Proposition 3.2 $m \le n$. □

**Proposition 5.3.** *Any generating system in a finite-dimensional vector space $V$ must have at least* $\dim V$ *vectors in it.*

**Proof.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m \in V$ be a complete system in $V$, and let $A : V \to \mathbb{R}^n$ be an isomorphism. Then $A\mathbf{v}_1, A\mathbf{v}_2, \dots, A\mathbf{v}_m$ is a complete system in $\mathbb{R}^n$, and by Proposition 3.5 $m \geq n$. □

**5.1. Completing a linearly independent system to a basis.** The following statement will play an important role later.

**Proposition 5.4** (Completion to a basis)**.** *A linearly independent system of vectors in a finite-dimensional space can be completed to a basis, i.e. if $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ are linearly independent vectors in a finite-dimensional vector space $V$ then one can find vectors $\mathbf{v}_{r+1}, \mathbf{v}_{r+2} \dots, \mathbf{v}_n$ such that the system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis in $V$.*

**Proof.** Let $n = \dim V$ and let $r < n$ (if $r = n$ then the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ is already a basis, and the case $r > n$ is impossible). Take any vector not belonging to $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$ and call it $\mathbf{v}_{r+1}$ (one can always do that because the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ is not generating). By Exercise 2.5 from Chapter 1 the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r, \mathbf{v}_{r+1}$ is linearly independent. Repeat the procedure with the new system to get vector $\mathbf{v}_{r+2}$, and so on.

We will stop the process when we get a generating system. Note, that the process cannot continue infinitely, because a linearly independent system of vectors in $V$ cannot have more than $n = \dim V$ vectors. □

**Exercises.**

**5.1.** True or false:

a) Every vector space that is generated by a finite set has a basis;

b) Every vector space has a (finite) basis;

c) A vector space cannot have more than one basis;

d) If a vector space has a finite basis, then the number of vectors in every basis is the same.

e) The dimension of $\mathbb{P}_n$ is $n$;

f) The dimension on $M_{m\times n}$ is $m + n$;

g) If vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ generate (span) the vector space $V$, then every vector in $V$ can be written as a linear combination of vector $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in only one way;

h) Every subspace of a finite-dimensional space is finite-dimensional;

i) If $V$ is a vector space having dimension $n$, then $V$ has exactly one subspace of dimension 0 and exactly one subspace of dimension $n$.

**5.2.** Prove that if $V$ is a vector space having dimension $n$, then a system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in $V$ is linearly independent if and only if it spans $V$.

**5.3.** Prove that a linearly independent system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in a vector space $V$ is a basis if and only if $n = \dim V$.

**5.4.** (An old problem revisited: now this problem should be easy) Is it possible that vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are linearly dependent, but the vectors $\mathbf{w}_1 = \mathbf{v}_1 + \mathbf{v}_2$, $\mathbf{w}_2 = \mathbf{v}_2 + \mathbf{v}_3$ and $\mathbf{w}_3 = \mathbf{v}_3 + \mathbf{v}_1$ are linearly *independent*? **Hint:** What dimension can the subspace $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ have?

**5.5.** Let vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be a basis in $V$. Show that $\mathbf{u} + \mathbf{v} + \mathbf{w}$, $\mathbf{v} + \mathbf{w}$, $\mathbf{w}$ is also a basis in $V$.

**5.6.** Consider in the space $\mathbb{R}^5$ vectors $\mathbf{v}_1 = (2, -1, 1, 5, -3)^T$, $\mathbf{v}_2 = (3, -2, 0, 0, 0)^T$, $\mathbf{v}_3 = (1, 1, 50, -921, 0)^T$.

a) Prove that these vectors are linearly independent.

b) Complete this system of vectors to a basis.

If you do part b) first you can do everything without any computations.

## 6. General solution of a linear system.

In this short section we discuss the structure of the general solution (i.e. of the solution set) of a linear system.

We call a system $A\mathbf{x} = \mathbf{b}$ homogeneous if the right side $\mathbf{b} = \mathbf{0}$, i.e. a homogeneous system is a system of form $A\mathbf{x} = \mathbf{0}$.

With each system

$$A\mathbf{x} = \mathbf{b}$$

we can associate a homogeneous system just by putting $\mathbf{b} = \mathbf{0}$.

**Theorem 6.1** (General solution of a linear equation)**.** *Let a vector $\mathbf{x}_1$ satisfy the equation $A\mathbf{x} = \mathbf{b}$, and let $H$ be the set of all solutions of the associated homogeneous system*

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

*Then the set*

$$\{\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_\mathrm{h} : \mathbf{x}_\mathrm{h} \in H\}$$

*is the set of all solutions of the equation $A\mathbf{x} = \mathbf{b}$.*

In other words, this theorem can be stated as

$$\boxed{\begin{array}{l}\text{General solution}\\ \text{of } A\mathbf{x} = \mathbf{b}\end{array}} = \boxed{\begin{array}{l}\text{A particular solu-}\\ \text{tion of } A\mathbf{x} = \mathbf{b}\end{array}} + \boxed{\begin{array}{l}\text{General solution}\\ \text{of } A\mathbf{x} = \mathbf{0}\end{array}}.$$

**Proof.** Fix a vector $\mathbf{x}_1$ satisfying $A\mathbf{x}_1 = \mathbf{b}$. Let a vector $\mathbf{x}_\mathrm{h}$ satisfy $A\mathbf{x}_h = \mathbf{0}$. Then for $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_\mathrm{h}$ we have

$$A\mathbf{x} = A(\mathbf{x}_1 + \mathbf{x}_\mathrm{h}) = A\mathbf{x}_1 + A\mathbf{x}_\mathrm{h} = \mathbf{b} + \mathbf{0} = \mathbf{b},$$

so any $\mathbf{x}$ of form

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_\mathrm{h}, \qquad \mathbf{x}_\mathrm{h} \in H$$

is a solution of $A\mathbf{x} = \mathbf{b}$.

Now let $\mathbf{x}$ satisfy $A\mathbf{x} = \mathbf{b}$. Then for $\mathbf{x}_\mathrm{h} := \mathbf{x} - \mathbf{x}_1$ we get

$$A\mathbf{x}_\mathrm{h} = A(\mathbf{x} - \mathbf{x}_1) = A\mathbf{x} - A\mathbf{x}_1 = \mathbf{b} - \mathbf{b} = \mathbf{0},$$

so $\mathbf{x}_\mathrm{h} \in H$. Therefore *any* solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$ can be represented as $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_\mathrm{h}$ with some $\mathbf{x}_\mathrm{h} \in H$. □

The power of this theorem is in its generality. It applies to all linear equations, we do not have to assume here that vector spaces are finite-dimensional. You will meet this theorem in differential equations, integral equations, partial differential equations, etc. Besides showing the structure of the solution set, this theorem allows one to separate investigation of uniqueness from the study of existence. Namely, to study uniqueness, we only need to analyze uniqueness of the homogeneous equation $A\mathbf{x} = \mathbf{0}$, which always has a solution.

There is an immediate application in this course: this theorem allows us to check a solution of a system $A\mathbf{x} = \mathbf{b}$. For example, consider a system

$$\begin{pmatrix} 2 & 3 & 1 & 4 & -9 \\ 1 & 1 & 1 & 1 & -3 \\ 1 & 1 & 1 & 2 & -5 \\ 2 & 2 & 2 & 3 & -8 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 17 \\ 6 \\ 8 \\ 14 \end{pmatrix}.$$

Performing row reduction one can find the solution of this system

$$\mathbf{x} = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 2 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} 2 \\ -1 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \qquad x_3, x_5 \in \mathbb{F}. \tag{6.1}$$

The parameters $x_3$, $x_5$ can be denoted here by any other letters, $t$ and $s$, for example; we are keeping notation $x_3$ and $x_5$ here only to remind us that the parameters came from the corresponding free variables.

Now, let us suppose, that we are just given this solution, and we want to check whether or not it is correct. Of course, we can repeat the row operations, but this is too time consuming. Moreover, if the solution was obtained by some non-standard method, it can look differently from what we get from the row reduction. For example, the formula

$$\mathbf{x} = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 2 \\ 0 \end{pmatrix} + s \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 1 \end{pmatrix}, \qquad s, t \in \mathbb{F} \tag{6.2}$$

gives the same set as (6.1) (can you say why?); here we just replaced the last vector in (6.1) by its sum with the second one. So, this formula is different from the solution we got from the row reduction, but it is nevertheless correct.

The simplest way to check that (6.1) and (6.2) give us correct solutions, is to check that the first vector $(3, 1, 0, 2, 0)^T$ satisfies the equation $A\mathbf{x} = \mathbf{b}$, and that the other two (the ones with the parameters $x_3$ and $x_5$ or $s$ and $t$ in front of them) should satisfy the associated homogeneous equation $A\mathbf{x} = \mathbf{0}$.

If this checks out, we will be assured that any vector $\mathbf{x}$ defined by (6.1) or (6.2) is indeed a solution.

Note, that this method of checking the solution does not guarantee that (6.1) (or (6.2)) gives us all the solutions. For example, if we just somehow miss out the term with $x_3$, the above method of checking will still work fine.

So, how can we guarantee, that we did not miss any free variable, and there should not be extra term in (6.1)?

What comes to mind, is to count the pivots again. In this example, if one does row operations, the number of pivots is 3. So indeed, there should be 2 free variables, and it looks like we did not miss anything in (6.1).

To be able to *prove* this, we will need new notions of fundamental subspaces and of rank of a matrix. I should also mention that in this particular example, one does not have to perform all row operations to check that there are only 2 free variables, and that formulas (6.1) and (6.2) both give correct general solutions.

### Exercises.

**6.1.** True or false

a) Any system of linear equations has at least one solution;

b) Any system of linear equations has at most one solution;

c) Any homogeneous system of linear equations has at least one solution;

d) Any system of $n$ linear equations in $n$ unknowns has at least one solution;

e) Any system of $n$ linear equations in $n$ unknowns has at most one solution;

f) If the homogeneous system corresponding to a given system of a linear equations has a solution, then the given system has a solution;

g) If the coefficient matrix of a homogeneous system of $n$ linear equations in $n$ unknowns is invertible, then the system has no non-zero solution;

h) The solution set of any system of $m$ equations in $n$ unknowns is a subspace in $\mathbb{R}^n$;

i) The solution set of any homogeneous system of $m$ equations in $n$ unknowns is a subspace in $\mathbb{R}^n$.

**6.2.** Find a $2 \times 3$ system (2 equations with 3 unknowns) such that its general solution has a form

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + s \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \qquad s \in \mathbb{R}.$$

## 7. Fundamental subspaces of a matrix. Rank.

As we discussed above in Section 7 of Chapter 1, with any linear transformation $A : V \to W$ we can associate two subspaces, namely, its kernel, or null space

$$\operatorname{Ker} A = \operatorname{Null} A := \{\mathbf{v} \in V : A\mathbf{v} = \mathbf{0}\} \subset V,$$

and its range

$$\operatorname{Ran} A = \{\mathbf{w} \in W : \mathbf{w} = A\mathbf{v} \text{ for some } \mathbf{v} \in V\} \subset W.$$

In other words, the kernel $\operatorname{Ker} A$ is the solution set of the homogeneous equation $A\mathbf{x} = \mathbf{0}$, and the range $\operatorname{Ran} A$ is exactly the set of all right sides $\mathbf{b} \in W$ for which the equation $A\mathbf{x} = \mathbf{b}$ has a solution.

If $A$ is an $m \times n$ matrix, i.e. a mapping from $\mathbb{F}^n$ to $\mathbb{F}^m$, then it follows from the "column by coordinate" rule of the matrix multiplication that any vector $\mathbf{w} \in \operatorname{Ran} A$ can be represented as a linear combination of columns of $A$. This explains the name *column space* (notation $\operatorname{Col} A$), which is often used instead of $\operatorname{Ran} A$.

If $A$ is a matrix, then in addition to $\operatorname{Ran} A$ and $\operatorname{Ker} A$ one can also consider the range and kernel for the transposed matrix $A^T$. Often the term *row space* is used for $\operatorname{Ran} A^T$ and the term *left null space* is used for $\operatorname{Ker} A^T$ (but usually no special notation is introduced).

The four subspaces $\operatorname{Ran} A$, $\operatorname{Ker} A$, $\operatorname{Ran} A^T$, $\operatorname{Ker} A^T$ are called the *fundamental subspaces* of the matrix $A$. In this section we will study important relations between the dimensions of the four fundamental subspaces.

We will need the following definition, which is one of the fundamental notions of Linear Algebra

**Definition.** Given a linear transformation (matrix) $A$ its rank, $\operatorname{rank} A$, is the dimension of the range of $A$

$$\operatorname{rank} A := \dim \operatorname{Ran} A.$$

**7.1. Computing fundamental subspaces and rank.** To compute the fundamental subspaces and rank of a matrix, one needs to do echelon reduction. Namely, let $A$ be the matrix, and $A_{\mathrm{e}}$ be its echelon form

1. The pivot columns of the *original* matrix $A$ (i.e. the columns where after row operations we will have pivots in the echelon form) give us a basis (one of many possible) in $\operatorname{Ran} A$.

2. The pivot *rows* of the echelon from $A_{\mathrm{e}}$ give us a basis in the row space. Of course, it is possible just to transpose the matrix, and then do row operations. But if we already have the echelon form of $A$, say by computing $\operatorname{Ran} A$, then we get $\operatorname{Ran} A^T$ for free.
3. To find a basis in the null space $\operatorname{Ker} A$ one needs to solve the homogeneous equation $A\mathbf{x} = \mathbf{0}$: the details will be seen from the example below.

**Example.** Consider a matrix

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

Performing row operations we get the echelon form

$$\begin{pmatrix} \boxed{1} & 1 & 2 & 2 & 1 \\ 0 & 0 & \boxed{-3} & -3 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(the pivots are boxed here). So, the columns 1 and 3 of the *original matrix*, i.e. the columns

$$\begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 1 \\ 3 \\ -1 \end{pmatrix}$$

give us a basis in $\operatorname{Ran} A$. We also get a basis for the row space $\operatorname{Ran} A^T$ for free: the first and second row of the *echelon form* of $A$, i.e. the vectors

$$\begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ -3 \\ -3 \\ -1 \end{pmatrix}$$

(we put the vectors vertically here. The question of whether to put vectors here vertically as columns, or horizontally as rows is is really a matter of convention. Our reason for putting them vertically is that although we call $\operatorname{Ran} A^T$ the *row space* we define it as a column space of $A^T$)

To compute the basis in the null space $\operatorname{Ker} A$ we need to solve the equation $A\mathbf{x} = \mathbf{0}$. Compute the *reduced* echelon form of $A$, which in this example

is

$$\begin{pmatrix} \boxed{1} & 1 & 0 & 0 & 1/3 \\ 0 & 0 & \boxed{1} & 1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note, that when solving the homogeneous equation $A\mathbf{x} = \mathbf{0}$, it is not necessary to write the whole augmented matrix, it is sufficient to work with the coefficient matrix. Indeed, in this case the last column of the augmented matrix is the column of zeroes, which does not change under row operations. So, we can just keep this column in mind, without actually writing it. Keeping this last zero column in mind, we can read the solution off the reduced echelon form above:

$$\begin{cases} x_1 = -x_2 - \frac{1}{3}x_5, \\ x_2 \text{ is free}, \\ x_3 = -x_4 - \frac{1}{3}x_5 \\ x_4 \text{ is free}, \\ x_5 \text{ is free}, \end{cases}$$

or, in the vector form

$$\mathbf{x} = \begin{pmatrix} -x_2 - \frac{1}{3}x_5 \\ x_2 \\ -x_4 - \frac{1}{3}x_5 \\ x_4 \\ x_5 \end{pmatrix} = x_2 \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} -1/3 \\ 0 \\ -1/3 \\ 0 \\ 1 \end{pmatrix} \tag{7.1}$$

The vectors at each free variable, i.e. in our case the vectors

$$\begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -1/3 \\ 0 \\ -1/3 \\ 0 \\ 1 \end{pmatrix}$$

form a basis in $\operatorname{Ker} A$.

Unfortunately, there is no shortcut for finding a basis in $\operatorname{Ker} A^T$, one must solve the equation $A^T\mathbf{x} = \mathbf{0}$. The knowledge of the echelon form of $A$ does not help here.

**7.2. Explanation of the computing bases in the fundamental subspaces.** So, why do the above methods indeed give us bases in the fundamental subspaces?

7.2.1. *The null space* $\operatorname{Ker} A$. The case of the null space $\operatorname{Ker} A$ is probably the simplest one: since we solved the equation $A\mathbf{x} = \mathbf{0}$, i.e. found all the solutions, then any vector in $\operatorname{Ker} A$ is a linear combination of the vectors we obtained. Thus, the vectors we obtained form a spanning system in $\operatorname{Ker} A$. To see that the system is linearly independent, let us multiply each vector by the corresponding free variable and add everything, see (7.1). Then for each free variable $x_k$, the entry number $k$ of the resulting vector is exactly $x_k$, see (7.1) again, so the only way this vector (the linear combination) can be $\mathbf{0}$ is when all free variables are 0.

7.2.2. *The column space* $\operatorname{Ran} A$. Let us now explain why the method for finding a basis in the column space $\operatorname{Ran} A$ works. First of all, notice that the pivot columns of the *reduced echelon* form $A_{\text{re}}$ of $A$ form a basis in $\operatorname{Ran} A_{\text{re}}$ (not in the column space of the original matrix, but of its reduced echelon form!). Since row operations are just left multiplications by invertible matrices, they do not change linear independence. Therefore, the pivot columns of the *original* matrix $A$ are linearly independent.

Let us now show that the pivot columns of $A$ span the column space of $A$. Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ be the pivot columns of $A$, and let $\mathbf{v}$ be an arbitrary column of $A$. We want to show that $\mathbf{v}$ can be represented as a linear combination of the pivot columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$,

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_r\mathbf{v}_r.$$

The reduced echelon form $A_{\text{re}}$ is obtained from $A$ by the left multiplication

$$A_{\text{re}} = EA,$$

where $E$ is a product of elementary matrices, so $E$ is an invertible matrix. The vectors $E\mathbf{v}_1, E\mathbf{v}_2, \dots, E\mathbf{v}_r$ are the pivot columns of $A_{\text{re}}$, and the column $\mathbf{v}$ of $A$ is transformed to the column $E\mathbf{v}$ of $A_{\text{re}}$. Since the pivot columns of $A_{\text{re}}$ form a basis in $\operatorname{Ran} A_{\text{re}}$, vector $E\mathbf{v}$ can be represented as a linear combination

$$E\mathbf{v} = \alpha_1 E\mathbf{v}_1 + \alpha_2 E\mathbf{v}_2 + \ldots + \alpha_r E\mathbf{v}_r.$$

Multiplying this equality by $E^{-1}$ from the left we get the representation

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_r\mathbf{v}_r,$$

so indeed the pivot columns of $A$ span $\operatorname{Ran} A$.

7.2.3. *The row space* $\operatorname{Ran} A^T$. It is easy to see that the pivot rows of the *echelon form* $A_{\text{e}}$ of $A$ are linearly independent. Indeed, let $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ be the transposed (since we agreed always to put vectors vertically) pivot rows of $A_{\text{e}}$. Suppose

$$\alpha_1\mathbf{w}_1 + \alpha_2\mathbf{w}_2 + \ldots + \alpha_r\mathbf{w}_r = \mathbf{0}.$$

Consider the first non-zero entry of $\mathbf{w}_1$. Since for all other vectors $\mathbf{w}_2, \mathbf{w}_3, \dots, \mathbf{w}_r$ the corresponding entries equal 0 (by the definition of echelon form), we can conclude that $\alpha_1 = 0$. So we can just ignore the first term in the sum.

Consider now the first non-zero entry of $\mathbf{w}_2$. The corresponding entries of the vectors $\mathbf{w}_3, \dots, \mathbf{w}_r$ are 0, so $\alpha_2 = 0$. Repeating this procedure, we get that $\alpha_k = 0 \ \forall k = 1, 2, \dots, r$.

To see that vectors $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ span the row space, one can notice that *row operations do not change the row space*. This can be obtained directly from analyzing row operations, but we present here a more formal way to demonstrate this fact.

For a transformation $A$ and a set $X$ let us denote by $A(X)$ the set of all elements $y$ which can represented as $y = A(x)$, $x \in X$,

$$A(X) := \{y = A(x) : x \in X\}.$$

If $A$ is an $m \times n$ matrix, and $A_{\mathrm{e}}$ is its echelon form, $A_{\mathrm{e}}$ is obtained from $A$ be left multiplication

$$A_{\mathrm{e}} = EA,$$

where $E$ is an $m \times m$ invertible matrix (the product of the corresponding elementary matrices). Then

$$\operatorname{Ran} A_{\mathrm{e}}^T = \operatorname{Ran}(A^T E^T) = A^T(\operatorname{Ran} E^T) = A^T(\mathbb{R}^m) = \operatorname{Ran} A^T,$$

so indeed $\operatorname{Ran} A^T = \operatorname{Ran} A_{\mathrm{e}}^T$.

**7.3. The Rank Theorem. Dimensions of fundamental subspaces.** There are many applications in which one needs to find a basis in column space or in the null space of a matrix. For example, as it was shown above, solving a homogeneous equation $A\mathbf{x} = \mathbf{0}$ amounts to finding a basis in the null space $\operatorname{Ker} A$. Finding a basis in the column space means simply extracting a basis from a spanning set, by removing unnecessary vectors (columns).

However, the most important application of the above methods of computing bases of fundamental subspaces is the relations between their dimensions.

**Theorem 7.1** (The Rank Theorem)**.** *For a matrix $A$*

$$\operatorname{rank} A = \operatorname{rank} A^T.$$

This theorem is often stated as follows:

The *column rank* of a matrix coincides with its *row rank.*

The proof of this theorem is trivial, since dimensions of both $\operatorname{Ran} A$ and $\operatorname{Ran} A^T$ are equal to the number of pivots in the echelon form of $A$.

The following theorem is gives us important relations between dimensions of the fundamental spaces. It is often also called the Rank Theorem

**Theorem 7.2.** *Let $A$ be an $m \times n$ matrix, i.e. a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then*

1. $\dim \operatorname{Ker} A + \dim \operatorname{Ran} A = \dim \operatorname{Ker} A + \operatorname{rank} A = n$ *(dimension of the domain of $A$);*
2. $\dim \operatorname{Ker} A^T + \dim \operatorname{Ran} A^T = \dim \operatorname{Ker} A^T + \operatorname{rank} A^T = \dim \operatorname{Ker} A^T + \operatorname{rank} A = m$ *(dimension of the target space of $A$);*

**Proof.** The proof, modulo the above algorithms of finding bases in the fundamental subspaces, is almost trivial. The first statement is simply the fact that the number of free variables ($\dim \operatorname{Ker} A$) plus the number of basic variables (i.e. the number of pivots, i.e. $\operatorname{rank} A$) adds up to the number of columns (i.e. to $n$).

The second statement, if one takes into account that $\operatorname{rank} A = \operatorname{rank} A^T$ is simply the first statement applied to $A^T$. □

As an application of the above theorem, let us recall the example from Section 6. There we considered a system

$$\begin{pmatrix} 2 & 3 & 1 & 4 & -9 \\ 1 & 1 & 1 & 1 & -3 \\ 1 & 1 & 1 & 2 & -5 \\ 2 & 2 & 2 & 3 & -8 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 17 \\ 6 \\ 8 \\ 14 \end{pmatrix},$$

and we claimed that its general solution given by

$$\mathbf{x} = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 2 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} 2 \\ -1 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \qquad x_3, x_5 \in \mathbb{F},$$

or by

$$\mathbf{x} = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 2 \\ 0 \end{pmatrix} + s \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 1 \end{pmatrix}, \qquad s, t \in \mathbb{F}.$$

We checked in Section 6 that a vector $\mathbf{x}$ given by either formula is indeed a solution of the equation. But, how can we guarantee that any of the formulas describe *all* solutions?

First of all, we know that in either formula, the last 2 vectors (the ones multiplied by the parameters) belong to $\operatorname{Ker} A$. It is easy to see that in either case both vectors are linearly independent (two vectors are linearly dependent if and only if one is a multiple of the other).

Now, let us count dimensions: interchanging the first and the second rows and performing first round of row operations

$$\begin{matrix} \\ -2R_1 \\ -R_1 \\ -2R_1 \end{matrix}\begin{pmatrix} 1 & 1 & 1 & 1 & -3 \\ 2 & 3 & 1 & 4 & -9 \\ 1 & 1 & 1 & 2 & -5 \\ 2 & 2 & 2 & 3 & -8 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 & 1 & -3 \\ 0 & 1 & -1 & 2 & -3 \\ 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}$$

we see that there are three pivots already, so $\operatorname{rank} A \geq 3$. (Actually, we already can see that the rank is 3, but it is enough just to have the estimate here). By Theorem 7.2, $\operatorname{rank} A + \dim \operatorname{Ker} A = 5$, hence $\dim \operatorname{Ker} A \leq 2$, and therefore there cannot be more than 2 linearly independent vectors in $\operatorname{Ker} A$. Therefore, last 2 vectors in either formula form a basis in $\operatorname{Ker} A$, so either formula give all solutions of the equation.

An important corollary of the rank theorem, is the following theorem connecting existence and uniqueness for linear equations.

**Theorem 7.3.** *Let $A$ be an $m \times n$ matrix. Then the equation*

$$A\mathbf{x} = \mathbf{b}$$

*has a solution for every $\mathbf{b} \in \mathbb{R}^m$ if and only if the dual equation*

$$A^T\mathbf{x} = \mathbf{0}$$

*has a unique (only the trivial) solution. (Note, that in the second equation we have $A^T$, not $A$).*

**Proof.** The proof follows immediately from Theorem 7.2 by counting the dimensions. We leave the details as an exercise to the reader. □

There is a very nice geometric interpretation of the second rank theorem (Theorem 7.2). Namely, statement 1 of the theorem says, that if a transformation $A : \mathbb{F}^n \to \mathbb{F}^m$ has trivial kernel ($\operatorname{Ker} A = \{\mathbf{0}\}$), then the dimensions of the domain $\mathbb{F}^n$ and of the range $\operatorname{Ran} A$ coincide. If the kernel is non-trivial, then the transformation "kills" $\dim \operatorname{Ker} A$ dimensions, so $\dim \operatorname{Ran} A = n - \dim \operatorname{Ker} A$.

**7.4. Completion of a linearly independent system to a basis.** As Proposition 5.4 from Section 5 above asserts, any linearly independent system can be completed to a basis, i.e. given inearly independent vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ in a finite-dimensional vector space $V$, one can find vectors

$\mathbf{v}_{r+1}, \mathbf{v}_{r+2} \dots, \mathbf{v}_n$ such that the system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis in $V$.

Theoretically, the proof of this proposition give us an algorithm of finding the vectors $\mathbf{v}_{r+1}, \mathbf{v}_{r+2} \dots, \mathbf{v}_n$, but this algorithm does not look too practical.

Ideas of this section give us a more practical way to perform the completion to a basis.

First of all, notice that if an $m \times n$ matrix is in an echelon form, then its non-zero rows (which are clearly linearly independent) can be easily completed to a basis in the whole space $\mathbb{F}^n$; one just needs to add some rows in appropriate places, so the resulting matrix is still in the echelon form and has pivots in every column.

Then, the non-zero rows of the new matrix form a basis, and we can order it any way we want, because property of being basis does not depend on the ordering.

Suppose now that we have linearly independent vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$, $\mathbf{v}_k \in \mathbb{F}^n$. Consider the matrix $A$ with rows $\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_r^T$ and perform row operations to get the echelon form $A_{\mathrm{e}}$. As we discussed above, the rows of $A_{\mathrm{e}}$ can be easily completed to a basis in $\mathbb{R}^n$. And it turns out that the same vectors that complete rows of $A_{\mathrm{e}}$ to a basis complete to a basis the original vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$.

To see that, let vectors $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ complete the rows of $A_{\mathrm{e}}$ to a basis in $\mathbb{F}^n$. Then, if we add to a matrix $A_{\mathrm{e}}$ rows $\mathbf{v}_{r+1}^T, \dots, \mathbf{v}_n^T$, we get an invertible matrix. Let call this matrix $\widetilde{A}_{\mathrm{e}}$, and let $\widetilde{A}$ be the matrix obtained from $A$ by adding rows $\mathbf{v}_{r+1}^T, \dots, \mathbf{v}_n^T$. The matrix $\widetilde{A}_{\mathrm{e}}$ can be obtained from $\widetilde{A}$ by row operations, so

$$\widetilde{A}_{\mathrm{e}} = E\widetilde{A},$$

where $E$ is the product of the corresponding elementary matrices. Then $\widetilde{A} = E^{-1}$ and $\widetilde{A}$ is invertible as a product of invertible matrices.

But that means that the rows of $\widetilde{A}$ form a basis in $\mathbb{F}^n$, which is exactly what we need.

**Remark.** The method of completion to a basis described above may be not the simplest one, but one of its principal advantages is that it works for vector spaces over an arbitrary field.

**Exercises.**

**7.1.** True or false:

a) The rank of a matrix is equal to the number of its non-zero columns;

b) The $m \times n$ zero matrix is the only $m \times n$ matrix having rank 0;

c) Elementary row operations preserve rank;

d) Elementary column operations do not necessarily preserve rank;

e) The rank of a matrix is equal to the maximum number of linearly independent columns in the matrix;

f) The rank of a matrix is equal to the maximum number of linearly independent rows in the matrix;

g) The rank of an $n \times n$ matrix is at most $n$;

h) An $n \times n$ matrix having rank $n$ is invertible.

**7.2.** A $54 \times 37$ matrix has rank 31. What are dimensions of all 4 fundamental subspaces?

**7.3.** Compute rank and find bases of all four fundamental subspaces for the matrices

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 & 1 & 1 \\ 1 & 4 & 0 & 1 & 2 \\ 0 & 2 & -3 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

**7.4.** Prove that if $A : X \to Y$ and $V$ is a subspace of $X$ then $\dim AV \le \operatorname{rank} A$. ($AV$ here means the subspace $V$ transformed by the transformation $A$, i.e. any vector in $AV$ can be represented as $A\mathbf{v}$, $\mathbf{v} \in V$). Deduce from here that $\operatorname{rank}(AB) \le \operatorname{rank} A$.

**Remark:** Here one can use the fact that if $V \subset W$ then $\dim V \le \dim W$. Do you understand why is it true?

**7.5.** Prove that if $A : X \to Y$ and $V$ is a subspace of $X$ then $\dim AV \le \dim V$. Deduce from here that $\operatorname{rank}(AB) \le \operatorname{rank} B$.

**7.6.** Prove that if the product $AB$ of two $n \times n$ matrices is invertible, then both $A$ and $B$ are invertible. Even if you know about determinants, do not use them, we did not cover them yet. **Hint:** use previous 2 problems.

**7.7.** Prove that if $A\mathbf{x} = \mathbf{0}$ has unique solution, then the equation $A^T\mathbf{x} = \mathbf{b}$ has a solution for every right side $\mathbf{b}$.
**Hint:** count pivots

**7.8.** Write a matrix with the required property, or explain why no such matrix exists

a) Column space contains $(1,0,0)^T$, $(0,0,1)^T$, row space contains $(1,1)^T$, $(1,2)^T$;

b) Column space is spanned by $(1,1,1)^T$, nullspace is spanned by $(1,2,3)^T$;

c) Column space is $\mathbb{R}^4$, row space is $\mathbb{R}^3$.

**Hint:** Check first if the dimensions add up.

**7.9.** If $A$ has the same four fundamental subspaces as $B$, does $A = B$?

**7.10.** Complete the rows of a matrix

$$\begin{pmatrix} e^3 & 3 & 4 & 0 & -\pi & 6 & -2 \\ 0 & 0 & 2 & -1 & \pi^e & 1 & 1 \\ 0 & 0 & 0 & 0 & 3 & -3 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

to a basis in $\mathbb{R}^7$.

**7.11.** For a matrix

$$\begin{pmatrix} 1 & 2 & -1 & 2 & 3 \\ 2 & 2 & 1 & 5 & 5 \\ 3 & 6 & -3 & 0 & 24 \\ -1 & -4 & 4 & -7 & 11 \end{pmatrix}$$

find bases in its column and row spaces.

**7.12.** For the matrix in the previous problem, complete the basis in the row space to a basis in $\mathbb{R}^5$

**7.13.** For the matrix

$$A = \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix}$$

compute $\operatorname{Ran} A$ and $\operatorname{Ker} A$. What can you say about relation between these subspaces?

**7.14.** Is it possible that for a real matrix $A$ that $\operatorname{Ran} A = \operatorname{Ker} A^T$? Is it possible for a complex $A$?

**7.15.** Complete the vectors $(1, 2, -1, 2, 3)^T$, $(2, 2, 1, 5, 5)^T$, $(-1, -4, 4, 7, -11)^T$ to a basis in $\mathbb{R}^5$.

## 8. Representation of a linear transformation in arbitrary bases. Change of coordinates formula.

The material we have learned about linear transformations and their matrices can be easily extended to transformations in abstract vector spaces with finite bases. In this section we will distinguish between a linear transformation $T$ and its matrix, the reason being that we consider different bases, so a linear transformation can have different matrix representation.

**8.1. Coordinate vector.** Let $V$ be a vector space with a basis $\mathcal{B} := \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$. Any vector $\mathbf{v} \in V$ admits a unique representation as a linear combination

$$\mathbf{v} = x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \ldots + x_n\mathbf{b}_n = \sum_{k=1}^{n} x_k\mathbf{b}_k.$$

The numbers $x_1, x_2, \ldots, x_n$ are called the *coordinates* of the vector $\mathbf{v}$ in the basis $\mathcal{B}$. It is convenient to join these coordinates into the so-called *coordinate vector* of $\mathbf{v}$ relative to the basis $\mathcal{B}$, which is the column vector

$$[\mathbf{v}]_{\mathcal{B}} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{F}^n.$$

Note that the mapping

$$\mathbf{v} \mapsto [\mathbf{v}]_{\mathcal{B}}$$

is an isomorphism between $V$ and $\mathbb{F}^n$. It transforms the basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ to the standard basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ in $\mathbb{F}^n$.

**8.2. Matrix of a linear transformation.** Let $T : V \to W$ be a linear transformation, and let $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$, $\mathcal{B} := \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_m\}$ be bases in $V$ and $W$ respectively.

A matrix of the transformation $T$ in (or with respect to) the bases $\mathcal{A}$ and $\mathcal{B}$ is an $m \times n$ matrix, denoted by $[T]_{\mathcal{B}\mathcal{A}}$, which relates the coordinate vectors $[T\mathbf{v}]_{\mathcal{B}}$ and $[\mathbf{v}]_{\mathcal{A}}$,

$$[T\mathbf{v}]_{\mathcal{B}} = [T]_{\mathcal{B}\mathcal{A}}[\mathbf{v}]_{\mathcal{A}};$$

notice the balance of symbols $\mathcal{A}$ and $\mathcal{B}$ here: this is the reason we put the first basis $\mathcal{A}$ into the second position.

The matrix $[T]_{\mathcal{B}\mathcal{A}}$ is easy to find: its $k$th column is just the coordinate vector $[T\mathbf{a}_k]_{\mathcal{B}}$ (compare this with finding the matrix of a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$).

As in the case of standard bases, composition of linear transformations is equivalent to multiplication of their matrices: one only has to be a bit more careful about bases. Namely, let $T_1 : X \to Y$ and $T_2 : Y \to Z$ be linear transformation, and let $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ be bases in $X$, $Y$ and $Z$ respectively. Then for the composition $T = T_2 T_1$,

$$T : X \to Z, \qquad T\mathbf{x} := T_2(T_1(\mathbf{x}))$$

we have

$$[T]_{\mathcal{C}\mathcal{A}} = [T_2 T_1]_{\mathcal{C}\mathcal{A}} = [T_2]_{\mathcal{C}\mathcal{B}}[T_1]_{\mathcal{B}\mathcal{A}} \tag{8.1}$$

(notice again the balance of indices here).

The proof here goes exactly as in the case of $\mathbb{F}^n$ spaces with standard bases, so we do not repeat it here. Another possibility is to transfer everything to the spaces $\mathbb{F}^n$ via the coordinate isomorphisms $\mathbf{v} \mapsto [\mathbf{v}]_{\mathcal{B}}$. Then one does not need any proof, everything follows from the results about matrix multiplication.

**8.3. Change of coordinate matrix.** Let us have two bases $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$ and $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ in a vector space $V$. Consider the identity transformation $I = I_V$ and its matrix $[I]_{\mathcal{B}\mathcal{A}}$ in these bases. By the definition

$$[\mathbf{v}]_{\mathcal{B}} = [I]_{\mathcal{B}\mathcal{A}}[\mathbf{v}]_{\mathcal{A}}, \qquad \forall \mathbf{v} \in V,$$

i.e. for any vector $\mathbf{v} \in V$ the matrix $[I]_{\mathcal{BA}}$ transforms its coordinates in the basis $\mathcal{A}$ into coordinates in the basis $\mathcal{B}$. The matrix $[I]_{\mathcal{BA}}$ is often called the *change of coordinates* (from the basis $\mathcal{A}$ to the basis $\mathcal{B}$) matrix.

The matrix $[I]_{\mathcal{BA}}$ is easy to compute: according to the general rule of finding the matrix of a linear transformation, its $k$th column is the coordinate representation $[\mathbf{a}_k]_{\mathcal{B}}$ of $k$th element of the basis $\mathcal{A}$

Note that

$$[I]_{\mathcal{AB}} = ([I]_{\mathcal{BA}})^{-1},$$

(follows immediately from the multiplication of matrices rule (8.1)), so any change of coordinate matrix is always invertible.

8.3.1. *An example: change of coordinates from the standard basis.* Let our space $V$ be $\mathbb{F}^n$, and let us have a basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ there. We also have the standard basis $\mathcal{S} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ there. The change of coordinates matrix $[I]_{\mathcal{SB}}$ is easy to compute:

$$[I]_{\mathcal{SB}} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n] =: B,$$

i.e. it is just the matrix $B$ whose $k$th column is the vector (column) $\mathbf{v}_k$. And in the other direction

$$[I]_{\mathcal{BS}} = ([I]_{\mathcal{SB}})^{-1} = B^{-1}.$$

For example, consider a basis

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 1 \end{pmatrix} \right\}$$

in $\mathbb{F}^2$, and let $\mathcal{S}$ denote the standard basis there. Then

$$[I]_{\mathcal{SB}} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} =: B$$

and

$$[I]_{\mathcal{BS}} = [I]_{\mathcal{SB}}^{-1} = B^{-1} = \frac{1}{3} \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}$$

(we know how to compute inverses, and it is also easy to check that the above matrix is indeed the inverse of $B$)

8.3.2. *An example: going through the standard basis.* In the space of polynomials of degree at most 1 we have bases

$$\mathcal{A} = \{1, 1+x\}, \quad \text{and} \quad \mathcal{B} = \{1+2x, 1-2x\},$$

and we want to find the change of coordinate matrix $[I]_{\mathcal{BA}}$.

Of course, we can always take vectors from the basis $\mathcal{A}$ and try to decompose them in the basis $\mathcal{B}$; it involves solving linear systems, and we know how to do that.

However, I think the following way is simpler. In $\mathbb{P}_1$ we also have the standard basis $\mathcal{S} = \{1, x\}$, and for this basis

$$[I]_{\mathcal{SA}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} =: A, \qquad [I]_{\mathcal{SB}} = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix} =: B,$$

and taking the inverses

$$[I]_{\mathcal{AS}} = A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \qquad [I]_{\mathcal{BS}} = B^{-1} = \frac{1}{4}\begin{pmatrix} 2 & 1 \\ 2 & -1 \end{pmatrix}.$$

Then

$$[I]_{\mathcal{BA}} = [I]_{\mathcal{BS}}[I]_{\mathcal{SA}} = B^{-1}A = \frac{1}{4}\begin{pmatrix} 2 & 1 \\ 2 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

and

$$[I]_{\mathcal{AB}} = [I]_{\mathcal{AS}}[I]_{\mathcal{SB}} = A^{-1}B = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}$$

Notice the balance of indices here.

**8.4. Matrix of a transformation and change of coordinates.** Let $T : V \to W$ be a linear transformation, and let $\mathcal{A}$, $\widetilde{\mathcal{A}}$ be two bases in $V$ and let $\mathcal{B}$, $\widetilde{\mathcal{B}}$ be two bases in $W$. Suppose we know the matrix $[T]_{\mathcal{BA}}$, and we would like to find the matrix representation with respect to new bases $\widetilde{\mathcal{A}}$, $\widetilde{\mathcal{B}}$, i.e. the matrix $[T]_{\widetilde{\mathcal{B}}\widetilde{\mathcal{A}}}$. The rule is very simple:

to get the matrix in the "new" bases one has to surround the matrix in the "old" bases by change of coordinates matrices.

I did not mention here what change of coordinate matrix should go where, because we don't have any choice if we follow the balance of indices rule. Namely, matrix representation of a linear transformation changes according to the formula

$$\boxed{[T]_{\widetilde{\mathcal{B}}\widetilde{\mathcal{A}}} = [I]_{\widetilde{\mathcal{B}}\mathcal{B}}[T]_{\mathcal{BA}}[I]_{\mathcal{A}\widetilde{\mathcal{A}}}}$$

Notice the balance of indices.

The proof can be done just by analyzing what each of the matrices does.

**8.5. Case of one basis: similar matrices.** Let $V$ be a vector space and let $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$ be a basis in $V$. Consider a linear transformation $T : V \to V$ and let $[T]_{\mathcal{AA}}$ be its matrix in this basis (we use the same basis for "inputs" and "outputs")

The case when we use the same basis for "inputs" and "outputs" is very important (because in this case we can multiply a matrix by itself), so let us study this case a bit more carefully. Notice, that very often in this case the shorter notation $[T]_{\mathcal{A}}$ is used instead of $[T]_{\mathcal{AA}}$. However, the two index notation $[T]_{\mathcal{AA}}$ is better adapted to the balance of indices rule, so I recommend using it (or at least always keep it in mind) when doing change of coordinates.

$[T]_{\mathcal{A}}$ is often used instead of $[T]_{\mathcal{AA}}$. It is shorter, but two index notation is better adapted to the balance of indices rule.

Let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ be another basis in $V$. By the change of coordinate rule above

$$[T]_{\mathcal{B}\mathcal{B}} = [I]_{\mathcal{B}\mathcal{A}}[T]_{\mathcal{A}\mathcal{A}}[I]_{\mathcal{A}\mathcal{B}}$$

Recalling that

$$[I]_{\mathcal{B}\mathcal{A}} = [I]^{-1}_{\mathcal{A}\mathcal{B}}$$

and denoting $Q := [I]_{\mathcal{A}\mathcal{B}}$, we can rewrite the above formula as

$$[T]_{\mathcal{B}\mathcal{B}} = Q^{-1}[T]_{\mathcal{A}\mathcal{A}}Q.$$

This gives a motivation for the following definition

**Definition 8.1.** We say that a matrix $A$ is similar to a matrix $B$ if there exists an invertible matrix $Q$ such that $A = Q^{-1}BQ$.

Since an invertible matrix must be square, it follows from counting dimensions, that similar matrices $A$ and $B$ have to be square and of the same size. If $A$ is similar to $B$, i.e. if $A = Q^{-1}BQ$, then

$$B = QAQ^{-1} = (Q^{-1})^{-1}A(Q^{-1})$$

(since $Q^{-1}$ is invertible), therefore $B$ is similar to $A$. So, we can just say that $A$ and $B$ are *similar*.

The above reasoning shows, that it does not matter where to put $Q$ and where $Q^{-1}$: one can use the formula $A = QBQ^{-1}$ in the definition of similarity.

The above discussion shows, that one can treat similar matrices as different matrix representation of the same linear operator (transformation).

**Exercises.**

**8.1.** True or false

a) Every change of coordinate matrix is square;

b) Every change of coordinate matrix is invertible;

c) The matrices $A$ and $B$ are called similar if $B = Q^T AQ$ for some matrix $Q$;

d) The matrices $A$ and $B$ are called similar if $B = Q^{-1}AQ$ for some matrix $Q$;

e) Similar matrices do not need to be square.

**8.2.** Consider the system of vectors

$$(1, 2, 1, 1)^T, \quad (0, 1, 3, 1)^T, \quad (0, 3, 2, 0)^T, \quad (0, 1, 0, 0)^T.$$

a) Prove that it is a basis in $\mathbb{F}^4$. Try to do minimal amount of computations.

b) Find the change of coordinate matrix that changes the coordinates in this basis to the standard coordinates in $\mathbb{F}^4$ (i.e. to the coordinates in the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_4$).

**8.3.** Find the change of coordinates matrix that changes the coordinates in the basis $1, 1+t$ in $\mathbb{P}_1$ to the coordinates in the basis $1-t, 2t$.

**8.4.** Let $T$ be the linear operator in $\mathbb{F}^2$ defined (in the standard coordinates) by

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3x+y \\ x-2y \end{pmatrix}$$

Find the matrix of $T$ in the standard basis and in the basis

$$(1,1)^T, \qquad (1,2)^T.$$

**8.5.** Prove, that if $A$ and $B$ are similar matrices then $\operatorname{trace} A = \operatorname{trace} B$. **Hint:** recall how $\operatorname{trace}(XY)$ and $\operatorname{trace}(YX)$ are related.

**8.6.** Are the matrices

$$\begin{pmatrix} 1 & 3 \\ 2 & 2 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 0 & 2 \\ 4 & 2 \end{pmatrix}$$

similar? Justify.

*Chapter 3*

# Determinants

## 1. Introduction.

The reader probably already met determinants in calculus or algebra, at least the determinants of $2 \times 2$ and $3 \times 3$ matrices. For a $2 \times 2$ matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

the determinant is simply $ad - bc$; the determinant of a $3 \times 3$ matrix can be found by the "Star of David" rule.

In this chapter we would like to introduce determinants for $n \times n$ matrices. I don't want just to give a formal definition. First I want to give some motivation, and then derive some properties the determinant should have. Then if we want to have these properties, we do not have any choice, and arrive to several equivalent definitions of the determinant.

It is more convenient to start not with the determinant of a matrix, but with determinant of a system of vectors. There is no real difference here, since we always can join vectors together (say as columns) to form a matrix.

Let us have $n$ vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in $\mathbb{R}^n$ (notice that the number of vectors coincides with dimension), and we want to find the $n$-dimensional volume of the parallelepiped determined by these vectors.

The parallelepiped determined by the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ can be defined as the collection of all vectors $\mathbf{v} \in \mathbb{R}^n$ that can be represented as

$$\mathbf{v} = t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \ldots + t_n\mathbf{v}_n, \qquad 0 \le t_k \le 1 \quad \forall k = 1, 2, \ldots, n.$$

It can be easily visualized when $n = 2$ (parallelogram) and $n = 3$ (parallelepiped). So, what is the $n$-dimensional volume?

If $n = 2$ it is area; if $n = 3$ it is indeed the volume. In dimension 1 is it just the length.

Finally, let us introduce some notation. For a system of vectors (columns) $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ we will denote its determinant (that we are going to construct) as $D(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$. If we join these vectors in a matrix $A$ (column number $k$ of $A$ is $\mathbf{v}_k$), then we will use the notation $\det A$,

$$\det A = D(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$

Also, for a matrix

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} \end{pmatrix}$$

its determinant is often is denoted by

$$\begin{vmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} \end{vmatrix}.$$

## 2. What properties determinant should have.

We know, that for dimensions 2 and 3 "volume" of a parallelepiped is determined by the *base times height* rule: if we pick one vector, then height is the distance from this vector to the subspace spanned by the remaining vectors, and the base is the $(n-1)$-dimensional volume of the parallelepiped determined by the remaining vectors.

Now let us generalize this idea to higher dimensions. For a moment we do not care about how exactly to determine height and base. We will show, that if we assume that the base and the height satisfy some natural properties, then we do not have any choice, and the volume (determinant) is uniquely defined.

**2.1. Linearity in each argument.** First of all, if we multiply vector $\mathbf{v}_1$ by a positive number $a$, then the height (i.e. the distance to the linear span $\mathcal{L}(\mathbf{v}_2, \ldots, \mathbf{v}_n)$) is multiplied by $a$. If we admit negative heights (and negative volumes), then this property holds for all scalars $a$, and so the determinant $D(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ of the system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ should satisfy

$$D(\alpha\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = \alpha D(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n).$$

Of course, there is nothing special about vector $\mathbf{v}_1$, so for any index $k$

$$D(\mathbf{v}_1, \dots, \underset{k}{\alpha\mathbf{v}_k}, \dots, \mathbf{v}_n) = \alpha D(\mathbf{v}_1, \dots, \underset{k}{\mathbf{v}_k}, \dots, \mathbf{v}_n) \tag{2.1}$$

To get the next property, let us notice that if we add 2 vectors, then the "height" of the result should be equal the sum of the "heights" of summands, i.e. that

$$D(\mathbf{v}_1, \dots, \underbrace{\mathbf{u}_k + \mathbf{v}_k}_{k}, \dots, \mathbf{v}_n) = \tag{2.2}$$

$$D(\mathbf{v}_1, \dots, \underset{k}{\mathbf{u}_k}, \dots, \mathbf{v}_n) + D(\mathbf{v}_1, \dots, \underset{k}{\mathbf{v}_k}, \dots, \mathbf{v}_n)$$

In other words, the above two properties say that the determinant of $n$ vectors is *linear in each argument (vector)*, meaning that if we fix $n-1$ vectors and interpret the remaining vector as a variable (argument), we get a linear function.

**Remark.** We already know that *linearity* is a very nice property, that helps in many situations. So, admitting negative heights (and therefore negative volumes) is a very small price to pay to get linearity, since we can always put on the absolute value afterwards.

In fact, by admitting negative heights, we did not sacrifice anything! To the contrary, we even gained something, because the sign of the determinant contains some information about the system of vectors (orientation).

**2.2. Preservation under "column replacement".** The next property also seems natural. Namely, if we take a vector, say $\mathbf{v}_j$, and add to it a multiple of another vector $\mathbf{v}_k$, the "height" does not change, so

$$D(\mathbf{v}_1, \dots, \underbrace{\mathbf{v}_j + \alpha\mathbf{v}_k}_{j}, \dots, \underset{k}{\mathbf{v}_k}, \dots, \mathbf{v}_n) \tag{2.3}$$

$$= D(\mathbf{v}_1, \dots, \underset{j}{\mathbf{v}_j}, \dots, \underset{k}{\mathbf{v}_k}, \dots, \mathbf{v}_n)$$

In other words, if we apply the *column operation* of the third type, the determinant does not change.

**Remark.** Although it is not essential here, let us notice that the second part of linearity (property (2.2)) is not independent: it can be deduced from properties (2.1) and (2.3).

We leave the proof as an exercise for the reader.

**2.3. Antisymmetry.** The next property the determinant should have, is that if we interchange 2 vectors, the determinant changes sign:

$$D(\mathbf{v}_1, \dots, \underset{j}{\mathbf{v}_k}, \dots, \underset{k}{\mathbf{v}_j}, \dots, \mathbf{v}_n) = -D(\mathbf{v}_1, \dots, \underset{j}{\mathbf{v}_j}, \dots, \underset{k}{\mathbf{v}_k}, \dots, \mathbf{v}_n). \tag{2.4}$$

Functions of several variables that change sign when one interchanges any two arguments are called *antisymmetric*.

At first sight this property does not look natural, but it can be deduced from the previous ones. Namely, applying property (2.3) three times, and then using (2.1) we get

$$\begin{aligned}
&D(\mathbf{v}_1,\dots,\underset{j}{\mathbf{v}_j},\dots,\underset{k}{\mathbf{v}_k},\dots,\mathbf{v}_n) = \\
&\quad = D(\mathbf{v}_1,\dots,\underset{j}{\mathbf{v}_j},\dots,\underbrace{\mathbf{v}_k-\mathbf{v}_j}_{k},\dots,\mathbf{v}_n) \\
&\quad = D(\mathbf{v}_1,\dots,\underbrace{\mathbf{v}_j+(\mathbf{v}_k-\mathbf{v}_j)}_{j},\dots,\underbrace{\mathbf{v}_k-\mathbf{v}_j}_{k},\dots,\mathbf{v}_n) \\
&\quad = D(\mathbf{v}_1,\dots,\underset{j}{\mathbf{v}_k},\dots,\underbrace{\mathbf{v}_k-\mathbf{v}_j}_{k},\dots,\mathbf{v}_n) \\
&\quad = D(\mathbf{v}_1,\dots,\underset{j}{\mathbf{v}_k},\dots,\underbrace{(\mathbf{v}_k-\mathbf{v}_j)-\mathbf{v}_k}_{k},\dots,\mathbf{v}_n) \\
&\quad = D(\mathbf{v}_1,\dots,\underset{j}{\mathbf{v}_k},\dots,\underset{k}{-\mathbf{v}_j},\dots,\mathbf{v}_n) \\
&\quad = -D(\mathbf{v}_1,\dots,\underset{j}{\mathbf{v}_k},\dots,\underset{k}{\mathbf{v}_j},\dots,\mathbf{v}_n).
\end{aligned}$$

**2.4. Normalization.** The last property is the easiest one. For the standard basis $\mathbf{e}_1,\mathbf{e}_2,\dots,\mathbf{e}_n$ in $\mathbb{R}^n$ the corresponding parallelepiped is the $n$-dimensional unit cube, so

$$D(\mathbf{e}_1,\mathbf{e}_2,\dots,\mathbf{e}_n)=1. \tag{2.5}$$

In matrix notation this can be written as

$$\det(I)=1$$

## 3. Constructing the determinant.

The plan of the game is now as follows: using the properties that as we decided in Section 2 the determinant should have, we derive other properties of the determinant, some of them highly non-trivial. We will show how to use these properties to compute the determinant using our old friend—row reduction.

Later, in Section 4, we will show that the determinant, i.e. a function with the desired properties exists and unique. After all we have to be sure that the object we are computing and studying exists.

While our initial geometric motivation for determinant and its properties came from considering vectors in the real vector space $\mathbb{R}^n$, so they relate only to matrices with real entries, all the constructions below use only algebraic operations (addition, multiplication, division) and are applicable to matrices with complex entries, and even with entries in an arbitrary field.

So in what follows we are constructing determinant not just for real matrices, but for complex matrices as well (and also for matrices with entries in an arbitrary field). The nice geometric motivation for the properties works only in the real case, but after we decided on the properties of the determinant (see properties 1–3 below) everything works in the general case.

**3.1. Basic properties.** We will use the following basic properties of the determinant:

1. Determinant is linear in each column, i.e. in vector notation for every index $k$

$$D(\mathbf{v}_1, \dots, \underbrace{\alpha\mathbf{u}_k + \beta\mathbf{v}_k}_{k}, \dots, \mathbf{v}_n) = \alpha D(\mathbf{v}_1, \dots, \underset{k}{\mathbf{u}_k}, \dots, \mathbf{v}_n) + \beta D(\mathbf{v}_1, \dots, \underset{k}{\mathbf{v}_k}, \dots, \mathbf{v}_n)$$

   for all scalars $\alpha$, $\beta$.
2. Determinant is *antisymmetric*, i.e. if one interchanges two columns, the determinant changes sign.
3. Normalization property: $\det I = 1$.

All these properties were discussed above in Section 2. The first property is just the (2.1) and (2.2) combined. The second one is (2.4), and the last one is the normalization property (2.5). Note, that we did not use property (2.3): it can be deduced from the above three. These three properties completely define determinant!

**3.2. Properties of determinant deduced from the basic properties.**

**Proposition 3.1.** *For a square matrix $A$ the following statements hold:*

1. *If $A$ has a zero column, then* $\det A = 0$.
2. *If $A$ has two equal columns, then* $\det A = 0$*;*
3. *If one column of $A$ is a multiple of another, then* $\det A = 0$*;*
4. *If columns of $A$ are linearly dependent, i.e. if the matrix is not invertible, then* $\det A = 0$.

**Proof.** Statement 1 follows immediately from linearity. If we multiply the zero column by zero, we do not change the matrix and its determinant. But by the property 1 above, we should get 0.

The fact that determinant is antisymmetric, implies statement 2. Indeed, if we interchange two equal columns, we change nothing, so the determinant remains the same. On the other hand, interchanging two columns

changes sign of determinant, so

$$\det A = -\det A,$$

which is possible only if $\det A = 0$.

Statement 3 is immediate corollary of statement 2 and linearity.

To prove the last statement, let us first suppose that the first vector $\mathbf{v}_1$ is a linear combination of the other vectors,

$$\mathbf{v}_1 = \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 + \ldots + \alpha_n\mathbf{v}_n = \sum_{k=2}^{n} \alpha_k\mathbf{v}_k.$$

Then by linearity we have (in vector notation)

$$D(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = D\left(\big(\sum_{k=2}^{n} \alpha_k\mathbf{v}_k\big), \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\right) = \sum_{k=2}^{n} \alpha_k D(\mathbf{v}_k, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n)$$

and each determinant in the sum is zero because of two equal columns.

Let us now consider general case, i.e. let us assume that the system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is linearly dependent. Then one of the vectors, say $\mathbf{v}_k$ can be represented as a linear combination of the others. Interchanging this vector with $\mathbf{v}_1$ we arrive to the situation we just treated, so

$$D(\mathbf{v}_1, \ldots, \underset{k}{\mathbf{v}_k}, \ldots, \mathbf{v}_n) = -D(\mathbf{v}_k, \ldots, \underset{k}{\mathbf{v}_1}, \ldots, \mathbf{v}_n) = -0 = 0,$$

so the determinant in this case is also 0. □

The next proposition generalizes property (2.3). As we already have said above, this property can be deduced from the three "basic" properties of the determinant, we are using in this section.

Note, that adding to a column a multiple of itself is prohibited here. We can only add multiples of the other columns.

**Proposition 3.2.** *The determinant does not change if we add to a column a linear combination of the other columns (leaving the other columns intact). In particular, the determinant is preserved under "column replacement" (column operation of third type).*

**Proof.** Fix a vector $\mathbf{v}_k$, and let $\mathbf{u}$ be a linear combination of the other vectors,

$$\mathbf{u} = \sum_{j\neq k} \alpha_j\mathbf{v}_j.$$

Then by linearity

$$D(\mathbf{v}_1, \ldots, \underbrace{\mathbf{v}_k + \mathbf{u}}_{k}, \ldots, \mathbf{v}_n) = D(\mathbf{v}_1, \ldots, \underset{k}{\mathbf{v}_k}, \ldots, \mathbf{v}_n) + D(\mathbf{v}_1, \ldots, \underset{k}{\mathbf{u}}, \ldots, \mathbf{v}_n),$$

and by Proposition 3.1 the last term is zero. □

**3.3. Determinants of diagonal and triangular matrices.** Now we are ready to compute determinant for some important special classes of matrices. The first class is the so-called *diagonal* matrices. Let us recall that a square matrix $A = \{a_{j,k}\}_{j,j=1}^{n}$ is called *diagonal* if all entries *off the main diagonal* are zero, i.e. if $a_{j,k} = 0$ for all $j \neq k$. We will often use the notation $\operatorname{diag}\{a_1, a_2, \dots, a_n\}$ for the diagonal matrix

$$\begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & a_n \end{pmatrix}.$$

Since a diagonal matrix $\operatorname{diag}\{a_1, a_2, \dots, a_n\}$ can be obtained from the identity matrix $I$ by multiplying column number $k$ by $a_k$,

| Determinant of a diagonal matrix equal the product of the diagonal entries, $$\det(\operatorname{diag}\{a_1, a_2, \dots, a_n\}) = a_1 a_2 \dots a_n.$$ |
|---|

The next important class is the class of so-called *triangular* matrices. A square matrix $A = \{a_{j,k}\}_{j,j=1}^{n}$ is called *upper triangular* if all entries *below* the main diagonal are 0, i.e. if $a_{j,k} = 0$ for all $k < j$. A square matrix is called *lower triangular* if all entries *above* the main are 0, i.e if $a_{j,k} = 0$ for all $j < k$. We call a matrix *triangular*, if it is either lower or upper triangular matrix.

It is easy to see that

| Determinant of a *triangular* matrix equals to the product of the diagonal entries, $$\det A = a_{1,1} a_{2,2} \dots a_{n,n}.$$ |
|---|

Indeed, if a triangular matrix has zero on the main diagonal, it is not invertible (this can easily be checked by column operations) and therefore both sides equal zero. If all diagonal entries are non-zero, then using column replacement (column operations of third type) one can transform the matrix into a diagonal one with the same diagonal entries: For upper triangular matrix one should first subtract appropriate multiples of the first column from the columns number $2, 3, \dots, n$, "killing" all entries in the first row, then subtract appropriate multiples of the second column from columns number $3, \dots, n$, and so on.

To treat the case of lower triangular matrices one has to do "column reduction" from the left to the right, i.e. first subtract appropriate multiples of the last column from columns number $n-1, \dots, 2, 1$, and so on.

**3.4. Computing the determinant.** Now we know how to compute determinants, using their properties: one just needs to do column reduction (i.e. row reduction for $A^T$) keeping track of column operations changing the determinant. Fortunately, the most often used operation—row replacement, i.e. operation of third type does not change the determinant. So we only need to keep track of interchanging of columns and of multiplication of column by a scalar.

If an echelon form of $A^T$ does not have pivots in every column (and row), then $A$ is not invertible, so $\det A = 0$. If $A$ is invertible, we arrive at a triangular matrix, and $\det A$ is the product of diagonal entries times the correction from column interchanges and multiplications.

The above algorithm implies that $\det A$ can be zero only if a matrix $A$ is not invertible. Combining this with the last statement of Proposition 3.1 we get

**Proposition 3.3.** $\det A = 0$ *if and only if* $A$ *is not invertible. An equivalent statement:* $\det A \neq 0$ *if and only if* $A$ *is invertible.*

Note, that although we now know how to compute determinants, the determinant is still not defined. One can ask: why don't we define it as the result we get from the above algorithm? The problem is that formally this result is not well defined: that means we did not prove that different sequences of column operations yield the same answer.

**3.5. Determinants of a transpose and of a product. Determinants of elementary matrices.** In this section we prove two important theorems.

**Theorem 3.4** (Determinant of a transpose)**.** *For a square matrix* $A$*,*

$$\det A = \det(A^T).$$

This theorem implies that for all statement about columns we discussed above, the corresponding statements about rows are also true. In particular, determinants behave under *row operations* the same way they behave under *column operations*. So, we can use row operations to compute determinants.

**Theorem 3.5** (Determinant of a product)**.** *For* $n \times n$ *matrices* $A$ *and* $B$

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

*In other words*

*Determinant of a product equals product of determinants.*

To prove both theorems we need the following lemma.

**Lemma 3.6.** *For a square matrix $A$ and an elementary matrix $E$ (of the same size)*

$$\det(AE) = (\det A)(\det E)$$

**Proof.** The proof can be done just by direct checking: determinants of special matrices are easy to compute; right multiplication by an elementary matrix is a column operation, and effect of column operations on the determinant is well known.

This can look like a lucky coincidence, that the determinants of elementary matrices agree with the corresponding column operations, but it is not a coincidence at all.

Namely, for a column operation the corresponding elementary matrix can be obtained from the identity matrix $I$ by this column operation. So, its determinant is 1 (determinant of $I$) times the effect of the column operation.

And that is all! It may be hard to realize at first, but the above paragraph is a *complete and rigorous* proof of the lemma! □

Applying $N$ times Lemma 3.6 we get the following corollary.

**Corollary 3.7.** *For any matrix $A$ and any sequence of elementary matrices $E_1, E_2, \dots, E_N$ (all matrices are $n \times n$)*

$$\det(AE_1E_2 \dots E_N) = (\det A)(\det E_1)(\det E_2) \dots (\det E_N)$$

**Lemma 3.8.** *Any invertible matrix is a product of elementary matrices.*

**Proof.** We know that any invertible matrix is row equivalent to the identity matrix, which is its reduced echelon form. So

$$I = E_N E_{N-1} \dots E_2 E_1 A,$$

and therefore any invertible matrix can be represented as a product of elementary matrices,

$$A = E_1^{-1} E_2^{-1} \dots E_{N-1}^{-1} E_N^{-1} I = E_1^{-1} E_2^{-1} \dots E_{N-1}^{-1} E_N^{-1}$$

(the inverse of an elementary matrix is an elementary matrix). □

**Proof of Theorem 3.4.** First of all, it can be easily checked, that for an elementary matrix $E$ we have $\det E = \det(E^T)$. Notice, that it is sufficient to prove the theorem only for invertible matrices $A$, since if $A$ is not invertible then $A^T$ is also not invertible, and both determinants are zero.

By Lemma 3.8 matrix $A$ can be represented as a product of elementary matrices,

$$A = E_1 E_2 \dots E_N,$$

and by Corollary 3.7 the determinant of $A$ is the product of determinants of the elementary matrices. Since taking the transpose just transposes each elementary matrix and reverses their order, Corollary 3.7 implies that $\det A = \det A^T$. □

**Proof of Theorem 3.5.** Let us first suppose that the matrix $B$ is invertible. Then Lemma 3.8 implies that $B$ can be represented as a product of elementary matrices

$$B = E_1 E_2 \dots E_N,$$

and so by Corollary 3.7

$$\det(AB) = (\det A)[(\det E_1)(\det E_2)\dots(\det E_N)] = (\det A)(\det B).$$

If $B$ is not invertible, then the product $AB$ is also not invertible, and the theorem just says that $0 = 0$.

To check that the product $AB = C$ is not invertible, let us assume that it is invertible. Then multiplying the identity $AB = C$ by $C^{-1}$ from the left, we get $C^{-1}AB = I$, so $C^{-1}A$ is a left inverse of $B$. So $B$ is left invertible, and since it is square, it is invertible. We got a contradiction. □

**3.6. Summary of properties of determinant.** First of all, let us say once more, that *the determinant is defined only for square matrices!* Since we now know that $\det A = \det(A^T)$, the statements that we knew about columns are true for rows too.

1. Determinant is linear in each row (column) when the other rows (columns) are fixed.
2. If one interchanges two rows (columns) of a matrix $A$, the determinant changes sign.
3. For a triangular (in particular, for a diagonal) matrix its determinant is the product of the diagonal entries. In particular, $\det I = 1$.
4. If a matrix $A$ has a zero row (or column), $\det A = 0$.
5. If a matrix $A$ has two equal rows (columns), $\det A = 0$.
6. If one of the rows (columns) of $A$ is a linear combination of the other rows (columns), i.e. if the matrix is not invertible, then $\det A = 0$;
   More generally,
7. $\det A = 0$ if and only if $A$ is not invertible, or equivalently
8. $\det A \neq 0$ if and only if $A$ is invertible.
9. $\det A$ does not change if we add to a row (column) a linear combination of the other rows (columns). In particular, the determinant is preserved under the row (column) replacement, i.e. under the row (column) operation of the third kind.

10. $\det A^T = \det A$.
11. $\det(AB) = (\det A)(\det B)$.
    And finally,
12. If $A$ is an $n \times n$ matrix, then $\det(aA) = a^n \det A$.

The last property follows from the linearity of the determinant, if we recall that to multiply a matrix $A$ by $a$ we have to multiply each row by $a$, and that each multiplication multiplies the determinant by $a$.

**Exercises.**

**3.1.** If $A$ is an $n \times n$ matrix, how are the determinants $\det A$ and $\det(5A)$ related? **Remark:** $\det(5A) = 5 \det A$ only in the trivial case of $1 \times 1$ matrices

**3.2.** How are the determinants $\det A$ and $\det B$ related if

a)
$$A = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}, \qquad B = \begin{pmatrix} 2a_1 & 3a_2 & 5a_3 \\ 2b_1 & 3b_2 & 5b_3 \\ 2c_1 & 3c_2 & 5c_3 \end{pmatrix};$$

b)
$$A = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}, \qquad B = \begin{pmatrix} 3a_1 & 4a_2 + 5a_1 & 5a_3 \\ 3b_1 & 4b_2 + 5b_1 & 5b_3 \\ 3c_1 & 4c_2 + 5c_1 & 5c_3 \end{pmatrix}.$$

**3.3.** Using column or row operations compute the determinants

$$\begin{vmatrix} 0 & 1 & 2 \\ -1 & 0 & -3 \\ 2 & 3 & 0 \end{vmatrix}, \qquad \begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix}, \qquad \begin{vmatrix} 1 & 0 & -2 & 3 \\ -3 & 1 & 1 & 2 \\ 0 & 4 & -1 & 1 \\ 2 & 3 & 0 & 1 \end{vmatrix}, \qquad \begin{vmatrix} 1 & x \\ 1 & y \end{vmatrix}.$$

**3.4.** A square $(n \times n)$ matrix is called skew-symmetric (or antisymmetric) if $A^T = -A$. Prove that if $A$ is skew-symmetric and $n$ is odd, then $\det A = 0$. Is this true for even $n$?

**3.5.** A square matrix is called *nilpotent* if $A^k = \mathbf{0}$ for some positive integer $k$. Show that for a nilpotent matrix $A$ $\det A = 0$.

**3.6.** Prove that if the matrices $A$ and $B$ are similar, than $\det A = \det B$.

**3.7.** A real square matrix $Q$ is called orthogonal if $Q^T Q = I$. Prove that if $Q$ is an orthogonal matrix then $\det Q = \pm 1$.

**3.8.** Show that
$$\begin{vmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{vmatrix} = (z - x)(z - y)(y - x).$$
This is a particular case of the so-called Vandermonde determinant.

**3.9.** Let points $A$, $B$ and $C$ in the plane $\mathbb{R}^2$ have coordinates $(x_1, y_1)$, $(x_2, y_2)$ and $(x_3, y_3)$ respectively. Show that the area of triangle $ABC$ is the absolute value of

$$\frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}.$$

**Hint:** use row operations and geometric interpretation of $2\times 2$ determinants (area).

**3.10.** Let $A$ be a square matrix. Show that block triangular matrices

$$\begin{pmatrix} I & * \\ \mathbf{0} & A \end{pmatrix}, \qquad \begin{pmatrix} A & * \\ \mathbf{0} & I \end{pmatrix}, \qquad \begin{pmatrix} I & \mathbf{0} \\ * & A \end{pmatrix}, \qquad \begin{pmatrix} A & \mathbf{0} \\ * & I \end{pmatrix}$$

all have determinant equal to $\det A$. Here $*$ can be anything.

The following problems illustrate the power of block matrix notation.

**3.11.** Use the previous problem to show that if $A$ and $C$ are square matrices, then

$$\det \begin{pmatrix} A & B \\ \mathbf{0} & C \end{pmatrix} = \det A \det C.$$

**Hint:** $\begin{pmatrix} A & B \\ \mathbf{0} & C \end{pmatrix} = \begin{pmatrix} I & B \\ \mathbf{0} & C \end{pmatrix}\begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix}$.

**3.12.** Let $A$ be $m \times n$ and $B$ be $n \times m$ matrices. Prove that

$$\det \begin{pmatrix} 0 & A \\ -B & I \end{pmatrix} = \det(AB).$$

**Hint:** While it is possible to transform the matrix by row operations to a form where the determinant is easy to compute, the easiest way is to right multiply the matrix by $\begin{pmatrix} I & 0 \\ B & I \end{pmatrix}$.

## 4. Formal definition. Existence and uniqueness of the determinant.

In this section we arrive to the formal definition of the determinant. We show that a function, satisfying the basic properties 1, 2, 3 from Section 3 exists, and moreover, such function is unique, i.e. we do not have any choice in constructing the determinant.

Consider an $n \times n$ matrix $A = \{a_{j,k}\}_{j,k=1}^n$, and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be its columns, i.e.

$$\mathbf{v}_k = \begin{pmatrix} a_{1,k} \\ a_{2,k} \\ \vdots \\ a_{n,k} \end{pmatrix} = a_{1,k}\mathbf{e}_1 + a_{2,k}\mathbf{e}_2 + \ldots + a_{n,k}\mathbf{e}_n = \sum_{j=1}^{n} a_{j,k}\mathbf{e}_j.$$

Using linearity of the determinant we expand it in the first column $\mathbf{v}_1$:

$$D(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = D(\sum_{j=1}^{n} a_{j,1}\mathbf{e}_j, \mathbf{v}_2, \dots, \mathbf{v}_n) = \sum_{j=1}^{n} a_{j,1} D(\mathbf{e}_j, \mathbf{v}_2, \dots, \mathbf{v}_n). \tag{4.1}$$

Then we expand it in the second column, then in the third, and so on. We get

$$D(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \dots \sum_{j_n=1}^{n} a_{j_1,1} a_{j_2,2} \dots a_{j_n,n} D(\mathbf{e}_{j_1}.\mathbf{e}_{j_2}, \dots \mathbf{e}_{j_n}).$$

Notice, that we have to use a different index of summation for each column: we call them $j_1, j_2, \dots, j_n$; the index $j_1$ here is the same as the index $j$ in (4.1).

It is a huge sum, it contains $n^n$ terms. Fortunately, some of the terms are zero. Namely, if any 2 of the indices $j_1, j_2, \dots, j_n$ coincide, the determinant $D(\mathbf{e}_{j_1}.\mathbf{e}_{j_2}, \dots \mathbf{e}_{j_n})$ is zero, because there are two equal columns here.

So, let us rewrite the sum, omitting all zero terms. The most convenient way to do that is using the notion of a *permutation*. Informally, a permutation of an ordered set $\{1, 2, \dots, n\}$ is a rearrangement of its elements. A convenient formal way to represent such a rearrangement is by using a function

$$\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\},$$

where $\sigma(1), \sigma(2), \dots, \sigma(n)$ gives the new order of the set $1, 2, \dots, n$. In other words, the permutation $\sigma$ rearranges the ordered set $1, 2, \dots, n$ into $\sigma(1), \sigma(2), \dots, \sigma(n)$.

Such function $\sigma$ has to be one-to-one (different values for different arguments) and onto (assumes all possible values from the target space). The functions which are one-to-one and onto are called *bijections*, and they give one-to-one correspondence between the domain and the target space.[1]

Although it is not directly relevant here, let us notice, that it is well-known in combinatorics, that the number of different permutations of the set $\{1, 2, \dots, n\}$ is exactly $n!$. The set of all permutations of the set $\{1, 2, \dots, n\}$ will be denoted $\mathrm{Perm}(n)$.

---

[1] There is another canonical way to represent permutation by a bijection $\sigma$, namely in this representation $\sigma(k)$ gives new position of the element number $k$. In this representation $\sigma$ rearranges $\sigma(1), \sigma(2), \dots, \sigma(n)$ into $1, 2, \dots, n$.

While in the first representation it is easy to write the function if you know the rearrangement of the set $1, 2, \dots, n$, the second one is more adapted to the composition of permutations: it coincides with the composition of functions. Namely if we first perform the permutation that correspond to a function $\sigma$ and then one that correspond to $\tau$, the resulting permutation will correspond to $\tau \circ \sigma$.

Using the notion of a permutation, we can rewrite the determinant as

$$D(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = \sum_{\sigma \in \mathrm{Perm}(n)} a_{\sigma(1),1} a_{\sigma(2),2} \dots a_{\sigma(n),n} D(\mathbf{e}_{\sigma(1)}, \mathbf{e}_{\sigma(2)}, \dots, \mathbf{e}_{\sigma(n)}).$$

The matrix with columns $\mathbf{e}_{\sigma(1)}, \mathbf{e}_{\sigma(2)}, \dots, \mathbf{e}_{\sigma(n)}$ can be obtained from the identity matrix by finitely many column interchanges, so the determinant

$$D(\mathbf{e}_{\sigma(1)}, \mathbf{e}_{\sigma(2)}, \dots, \mathbf{e}_{\sigma(n)})$$

is 1 or $-1$ depending on the number of column interchanges.

To formalize that, we (informally) define the *sign* (denoted $\operatorname{sign}\sigma$) of a permutation $\sigma$ to be 1 if an even number of interchanges is necessary to rearrange the $n$-tuple $1, 2, \dots, n$ into $\sigma(1), \sigma(2), \dots, \sigma(n)$, and $\operatorname{sign}(\sigma) = -1$ if the number of interchanges is odd.

It is a well-known fact from the combinatorics, that the sign of permutation is well defined, i.e. that although there are infinitely many ways to get the $n$-tuple $\sigma(1), \sigma(2), \dots, \sigma(n)$ from $1, 2, \dots, n$, the number of interchanges is either always odd or always even.

One of the ways to show that is to introduce an alternative definition. Let $K = K(\sigma)$ be the number of *disorders* of $\sigma$, i.e. the number of pairs $(j, k)$, $j, k \in \{1, 2, \dots, n\}$, $j < k$ such that $\sigma(j) > \sigma(k)$, and see if the number is even or odd. We call the permutation $\sigma$ *odd* if $K$ is odd and *even* if $K$ is even. Then define $\operatorname{sign}\sigma := (-1)^{K(\sigma)}$; note that this way $\operatorname{sign}\sigma$ is well defined.

We want to show that $\operatorname{sign}\sigma = (-1)^{K(\sigma)}$ can indeed be computed by rearranging the $n$-tuple $1, 2, \dots, n$ into $\sigma(1), \sigma(2), \dots, \sigma(n)$ and counting the number of interchanges, as was described above.

If $\sigma(k) = k\ \forall k$, then the number of *disorders* $K(\sigma)$ is 0, so *sign* of such *identity* permutation is 1. Note also, that any elementary transpose, which interchange two neighbors, changes the sign of a permutation, because it changes (increases or decreases) the number of *disorders* exactly by 1. So, to get from a permutation to another one always needs an even number of elementary transposes if the permutations have the same sign, and an odd number if the signs are different.

Finally, any interchange of two entries can be achieved by an odd number of elementary transposes. This implies that *sign* changes under an interchange of two entries. So, to get from $1, 2, \dots, n$ to an even permutation (positive sign) one always need even number of interchanges, and odd number of interchanges is needed to get an odd permutation (negative sign).

So, if we want determinant to satisfy basic properties 1–3 from Section 3, we must define it as

$$\det A = \sum_{\sigma \in \mathrm{Perm}(n)} a_{\sigma(1),1} a_{\sigma(2),2} \dots a_{\sigma(n),n} \operatorname{sign}(\sigma), \tag{4.2}$$

where the sum is taken over all permutations of the set $\{1, 2, \dots, n\}$.

If we define the determinant this way, it is easy to check that it satisfies the basic properties 1–3 from Section 3. Indeed, it is linear in each column, because for each column every term (product) in the sum contains exactly one entry from this column.

Interchanging two columns of $A$ just adds an extra interchange to the permutation, so right side in (4.2) changes sign. Finally, for the identity matrix $I$, the right side of (4.2) is 1 (it has one non-zero term).

**Exercises.**

**4.1.** Suppose the permutation $\sigma$ takes $(1, 2, 3, 4, 5)$ to $(5, 4, 1, 2, 3)$.

a) Find sign of $\sigma$;

b) What does $\sigma^2 := \sigma \circ \sigma$ do to $(1, 2, 3, 4, 5)$?

c) What does the inverse permutation $\sigma^{-1}$ do to $(1, 2, 3, 4, 5)$?

d) What is the sign of $\sigma^{-1}$?

**4.2.** Let $P$ be a *permutation matrix*, i.e. an $n \times n$ matrix consisting of zeroes and ones and such that there is exactly one 1 in every row and every column.

a) Can you describe the corresponding linear transformation? That will explain the name.

b) Show that $P$ is invertible. Can you describe $P^{-1}$?

c) Show that for some $N > 0$

$$P^N := \underbrace{PP \dots P}_{N \text{ times}} = I.$$

Use the fact that there are only finitely many permutations.

**4.3.** Why is there an even number of permutations of $(1, 2, \dots, 9)$ and why are exactly half of them odd permutations? **Hint:** This problem can be hard to solve in terms of permutations, but there is a very simple solution using determinants.

**4.4.** If $\sigma$ is an odd permutation, explain why $\sigma^2$ is even but $\sigma^{-1}$ is odd.

**4.5.** How many multiplications and additions is required to compute the determinant using formal definition (4.2) of the determinant of an $n \times n$ matrix? Do not count the operations needed to compute $\operatorname{sign} \sigma$.

## 5. Cofactor expansion.

For an $n \times n$ matrix $A = \{a_{j,k}\}_{j,k=1}^{n}$ let $A_{j,k}$ denotes the $(n-1) \times (n-1)$ matrix obtained from $A$ by crossing out row number $j$ and column number $k$.

**Theorem 5.1** (Cofactor expansion of determinant)**.** *Let $A$ be an $n \times n$ matrix. For each $j$, $1 \le j \le n$, determinant of $A$ can be expanded in the row number $j$ as*

$$\det A = a_{j,1}(-1)^{j+1} \det A_{j,1} + a_{j,2}(-1)^{j+2} \det A_{j,2} + \ldots + a_{j,n}(-1)^{j+n} \det A_{j,n} = \sum_{k=1}^{n} a_{j,k}(-1)^{j+k} \det A_{j,k}.$$

*Similarly, for each $k$, $1 \le k \le n$, the determinant can be expanded in the column number $k$,*

$$\det A = \sum_{j=1}^{n} a_{j,k}(-1)^{j+k} \det A_{j,k}.$$

**Proof.** Let us first prove the formula for the expansion in row number 1. The formula for expansion in row number $k$ then can be obtained from it by interchanging rows number 1 and $k$. Since $\det A = \det A^T$, column expansion follows automatically.

Let us first consider a special case, when the first row has one non-zero term $a_{1,1}$. Performing column operations on columns $2, 3, \ldots, n$ we transform $A$ to the lower triangular form. The determinant of $A$ then can be computed as

$$\boxed{\begin{array}{l}\text{the product of diagonal}\\ \text{entries of the triangular}\\ \text{matrix}\end{array}} \times \boxed{\begin{array}{l}\text{correcting factor from}\\ \text{the column operations}\end{array}}.$$

But the product of all diagonal entries except the first one (i.e. without $a_{1,1}$) times the correcting factor is exactly $\det A_{1,1}$, so in this particular case $\det A = a_{1,1} \det A_{1,1}$.

Let us now consider the case when all entries in the first row except $a_{1,2}$ are zeroes. This case can be reduced to the previous one by interchanging columns number 1 and 2, and therefore in this case $\det A = (-1)a_{1,2} \det A_{1,2}$.

The case when $a_{1,3}$ is the only non-zero entry in the first row, can be reduced to the previous one by interchanging rows 2 and 3, so in this case $\det A = a_{1,3} \det A_{1,3}$.

Repeating this procedure we get that in the case when $a_{1,k}$ is the only non-zero entry in the first row $\det A = (-1)^{1+k} a_{1,k} \det A_{1,k}$. [2]

In the general case, linearity of the determinant in each row implies that

$$\det A = \det A^{(1)} + \det A^{(2)} + \ldots + \det A^{(n)} = \sum_{k=1}^{n} \det A^{(k)}$$

where the matrix $A^{(k)}$ is obtained from $A$ by replacing all entries in the first row except $a_{1,k}$ by 0. As we just discussed above

$$\det A^{(k)} = (-1)^{1+k} a_{1,k} \det A_{1,k},$$

so

$$\det A = \sum_{k=1}^{n} (-1)^{1+k} a_{1,k} \det A_{1,k}.$$

To get the cofactor expansion in the second row, we can interchange the first and second rows and apply the above formula. The row exchange changes the sign, so we get

$$\det A = -\sum_{k=1}^{n} (-1)^{1+k} a_{2,k} \det A_{2,k} = \sum_{k=1}^{n} (-1)^{2+k} a_{2,k} \det A_{2,k}.$$

Exchanging rows 3 and 2 and expanding in the second row we get formula

$$\det A = \sum_{k=1}^{n} (-1)^{3+k} a_{3,k} \det A_{3,k},$$

and so on.

To expand the determinant $\det A$ in a column one need to apply the row expansion formula for $A^T$. □

**Definition.** The numbers

$$C_{j,k} = (-1)^{j+k} \det A_{j,k}$$

are called *cofactors*.

---

[2] In the case when $a_{1,k}$ is the only non-zero entry in the first row it may be tempting to exchange columns number 1 and number $k$, to reduce the problem to the case $a_{1,1} \neq 0$. However, when we exchange columns 1 and $k$ we change the order of other columns: if we just cross out column number $k$, then column number 1 will be the first of the remaining columns. But, if we exchange columns 1 and $k$, and then cross out column $k$ (which is now the first one), then the column 1 will be now column number $k-1$. To avoid the complications of keeping track of the order of columns, we can, as we did above, exchange columns number $k$ and $k-1$, reducing everything to the situation we treated on the previous step. Such an operation does not change the order for the rest of the columns.

Using this notation, the formula for expansion of the determinant in the row number $j$ can be rewritten as

$$\det A = a_{j,1}C_{j,1} + a_{j,2}C_{j,2} + \ldots + a_{j,n}C_{j,n} = \sum_{k=1}^{n} a_{j,k}C_{j,k}.$$

Similarly, expansion in the column number $k$ can be written as

$$\det A = a_{1,k}C_{1,k} + a_{2,k}C_{2,k} + \ldots + a_{n,k}C_{n,k} = \sum_{j=1}^{n} a_{j,k}C_{j,k}$$

Very often the cofactor expansion formula is used as the definition of determinant.

**Remark.** Very often the cofactor expansion formula is used as the definition of determinant. It is not difficult to show that the quantity given by this formula satisfies the basic properties of the determinant: the normalization property is trivial, the proof of antisymmetry is easy. However, the proof of linearity is a bit tedious (although not too difficult).

**Remark.** Although it looks very nice, the cofactor expansion formula is not suitable for computing determinant of matrices bigger than $3 \times 3$.

As one can count it requires more than $n!$ multiplications (to be precise it requires $\sum_{k=2}^{n} n!/k!$ multiplications), and $n!$ grows very rapidly. For example, cofactor expansion of a $20 \times 20$ matrix require more than $20! \approx 2.4 \cdot 10^{18}$ multiplications. It would take a computer performing a *billion* multiplications per second over 77 years to perform 20! multiplications; performing the multiplications required for the cofactor expansion of the determinant of a $20 \times 20$ matrix will require more than 132 years.[3]

On the other hand, computing the determinant of an $n \times n$ matrix using *row reduction* requires $(n^3 + 2n - 3)/3$ multiplications (and about the same number of additions). It would take a computer performing a *million* operations per second (very slow, by today's standards) a fraction of a second to compute the determinant of a $100 \times 100$ matrix by row reduction.

It can only be practical to apply the cofactor expansion formula in higher dimensions if a row (or a column) has a lot of zero entries.

However, the cofactor expansion formula is of great theoretical importance, as the next section shows.

**5.1. Cofactor formula for the inverse matrix.** The matrix $C = \{C_{j,k}\}_{j,k=1}^{n}$ whose entries are cofactors of a given matrix $A$ is called the *cofactor matrix* of $A$.

**Theorem 5.2.** *Let $A$ be an invertible matrix and let $C$ be its cofactor matrix. Then*

$$A^{-1} = \frac{1}{\det A}\, C^T.$$

[3] The reader can check the numbers sung, for example, WolframAlpha

**Proof.** Let us find the product $AC^T$. The diagonal entry number $j$ is obtained by multiplying $j$th row of $A$ by $j$th column of $A$ (i.e. $j$th row of $C$), so

$$(AC^T)_{j,j} = a_{j,1}C_{j,1} + a_{j,2}C_{j,2} + \ldots + a_{j,n}C_{j,n} = \det A,$$

by the cofactor expansion formula.

To get the off diagonal terms we need to multiply $k$th row of $A$ by $j$th column of $C^T$, $j \neq k$, to get

$$a_{k,1}C_{j,1} + a_{k,2}C_{j,2} + \ldots + a_{k,n}C_{j,n}.$$

It follows from the cofactor expansions formula (expanding in $j$th row) that this is the determinant of the matrix obtained from $A$ by replacing row number $j$ by the row number $k$ (and leaving all other rows as they were). But the rows $j$ and $k$ of this matrix coincide, so the determinant is 0. So, all off-diagonal entries of $AC^T$ are zeroes (and all diagonal ones equal $\det A$), thus

$$AC^T = (\det A)\, I.$$

That means that the matrix $\frac{1}{\det A}\, C^T$ is a right inverse of $A$, and since $A$ is square, it is the inverse. □

Recalling that for an invertible matrix $A$ the equation $A\mathbf{x} = \mathbf{b}$ has a unique solution

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{\det A} C^T\mathbf{b},$$

we get the following corollary of the above theorem.

**Corollary 5.3** (Cramer's rule)**.** *For an invertible matrix $A$ the entry number $k$ of the solution of the equation $A\mathbf{x} = \mathbf{b}$ is given by the formula*

$$x_k = \frac{\det B_k}{\det A},$$

*where the matrix $B_k$ is obtained from $A$ by replacing column number $k$ of $A$ by the vector $\mathbf{b}$.*

### 5.2. Some applications of the cofactor formula for the inverse.

**Example** (Inverting $2 \times 2$ matrices)**.** The cofactor formula really shines when one needs to invert a $2 \times 2$ matrix

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

The cofactors are just entries ($1 \times 1$ matrices), the cofactor matrix is

$$\begin{pmatrix} d & -c \\ -b & a \end{pmatrix},$$

so the inverse matrix $A^{-1}$ is given by the formula

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

While the cofactor formula for the inverse does not look practical for dimensions higher than 3, it has a great theoretical value, as the examples below illustrate.

**Example** (Matrix with integer inverse)**.** Suppose that we want to construct a matrix $A$ with integer entries, such that its inverse also has integer entries (inverting such a matrix would make a nice homework problem: no messing with fractions). If $\det A = 1$ and its entries are integer, the cofactor formula for inverses implies that $A^{-1}$ also have integer entries.

Note, that it is easy to construct an integer matrix $A$ with $\det A = 1$: one should start with a triangular matrix with 1 on the main diagonal, and then apply several row or column replacements (operations of the third type) to make the matrix look generic.

**Example** (Inverse of a polynomial matrix)**.** Another example is to consider a *polynomial matrix* $A(x)$, i.e. a matrix whose entries are not numbers but polynomials $a_{j,k}(x)$ of the variable $x$. If $\det A(x) \equiv 1$, then the inverse matrix $A^{-1}(x)$ is also a polynomial matrix.

If $\det A(x) = p(x) \not\equiv 0$, it follows from the cofactor expansion that $p(x)$ is a polynomial, so $A^{-1}(x)$ has rational entries: moreover, $p(x)$ is a multiple of each denominator.

**Exercises.**

**5.1.** Evaluate the determinants using any method

$$\begin{vmatrix} 0 & 1 & 1 \\ 1 & 2 & -5 \\ 6 & -4 & 3 \end{vmatrix}, \qquad \begin{vmatrix} 1 & -2 & 3 & -12 \\ -5 & 12 & -14 & 19 \\ -9 & 22 & -20 & 31 \\ -4 & 9 & -14 & 15 \end{vmatrix}.$$

**5.2.** Use row (column) expansion to evaluate the determinants. Note, that you don't need to use the first row (column): picking row (column) with many zeroes will simplify your calculations.

$$\begin{vmatrix} 1 & 2 & 0 \\ 1 & 1 & 5 \\ 1 & -3 & 0 \end{vmatrix}, \qquad \begin{vmatrix} 4 & -6 & -4 & 4 \\ 2 & 1 & 0 & 0 \\ 0 & -3 & 1 & 3 \\ -2 & 2 & -3 & -5 \end{vmatrix}.$$

**5.3.** For the $n \times n$ matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & a_0 \\ -1 & 0 & 0 & \dots & 0 & a_1 \\ 0 & -1 & 0 & \dots & 0 & a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & a_{n-2} \\ 0 & 0 & 0 & \dots & -1 & a_{n-1} \end{pmatrix}$$

compute $\det(A + tI)$, where $I$ is $n \times n$ identity matrix. You should get a nice expression involving $a_0, a_1, \dots, a_{n-1}$ and $t$. Row expansion and induction is probably the best way to go.

**5.4.** Using cofactor formula compute inverses of the matrices

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \qquad \begin{pmatrix} 19 & -17 \\ 3 & -2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ 3 & 5 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 0 & 1 & 1 \end{pmatrix}.$$

**5.5.** Let $D_n$ be the determinant of the $n \times n$ tridiagonal matrix

$$\begin{pmatrix} 1 & -1 & & & 0 \\ 1 & 1 & -1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & 1 & -1 \\ 0 & & & 1 & 1 \end{pmatrix}.$$

Using cofactor expansion show that $D_n = D_{n-1} + D_{n-2}$. This yields that the sequence $D_n$ is the Fibonacci sequence $1, 2, 3, 5, 8, 13, 21, \dots$

**5.6.** Vandermonde determinant revisited. Our goal is to prove the formula

$$\begin{vmatrix} 1 & c_0 & c_0^2 & \dots & c_0^n \\ 1 & c_1 & c_1^2 & \dots & c_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & c_n & c_n^2 & \dots & c_n^n \end{vmatrix} = \prod_{0 \le j < k \le n} (c_k - c_j)$$

for the $(n+1) \times (n+1)$ Vandermonde determinant.

We will apply induction. To do this

a) Check that the formula holds for $n = 1$, $n = 2$.

b) Call the variable $c_n$ in the last row $x$, and show that the determinant is a polynomial of degree $n$, $A_0 + A_1 x + A_2 x^2 + \dots + A_n x^n$, with the coefficients $A_k$ depending on $c_0, c_1, \dots, c_{n-1}$.

c) Show that the polynomial has zeroes at $x = c_0, c_1, \dots, c_{n-1}$, so it can be represented as $A_n \cdot (x - c_0)(x - c_1) \dots (x - c_{n-1})$, where $A_n$ as above.

d) Assuming that the formula for the Vandermonde determinant is true for $n - 1$, compute $A_n$ and prove the formula for $n$.

**5.7.** How many multiplication is needed to compute the determinant of an $n \times n$ matrix using the cofactor expansion? Prove the formula.

## 6. Minors and rank.

For a matrix $A$ let us consider its $k \times k$ *submatrix*, obtained by taking $k$ rows and $k$ columns. The determinant of this matrix is called a *minor* of order $k$. Note, that an $m \times n$ matrix has $\binom{m}{k} \cdot \binom{n}{k}$ different $k \times k$ submatrices, and so it has $\binom{m}{k} \cdot \binom{n}{k}$ minors of order $k$.

**Theorem 6.1.** *For a non-zero matrix $A$ its rank equals to the maximal integer $k$ such that there exists a non-zero minor of order $k$.*

**Proof.** Let us first show, that if $k > \operatorname{rank} A$ then all minors of order $k$ are 0. Indeed, since the dimension of the column space $\operatorname{Ran} A$ is $\operatorname{rank} A < k$, any $k$ columns of $A$ are linearly dependent. Therefore, for any $k \times k$ submatrix of $A$ its columns are linearly dependent, and so all minors of order $k$ are 0.

To complete the proof we need to show that there exists a non-zero minor of order $k = \operatorname{rank} A$. There can be many such minors, but probably the easiest way to get such a minor is to take pivot rows and pivot columns (i.e. rows and columns of the original matrix, containing a pivot). This $k \times k$ submatrix has the same pivots as the original matrix, so it is invertible (pivot in every column and every row) and its determinant is non-zero. □

This theorem does not look very useful, because it is much easier to perform row reduction than to compute all minors. However, it is of great theoretical importance, as the following corollary shows.

**Corollary 6.2.** *Let $A = A(x)$ be an $m \times n$ polynomial matrix (i.e. a matrix whose entries are polynomials of $x$). Then* $\operatorname{rank} A(x)$ *is constant everywhere, except maybe finitely many points, where the rank is smaller.*

**Proof.** Let $r$ be the largest integer such that $\operatorname{rank} A(x) = r$ for some $x$. To show that such $r$ exists, we first try $r = \min\{m, n\}$. If there exists $x$ such that $\operatorname{rank} A(x) = r$, we have found $r$. If not, we replace $r$ by $r - 1$ and try again. After finitely many steps we either stop or hit 0. So, $r$ exists.

Let $x_0$ be a point such that $\operatorname{rank} A(x_0) = r$, and let $M$ be a minor of order $k$ such that $M(x_0) \neq 0$. Since $M(x)$ is the determinant of a $k \times k$ polynomial matrix, $M(x)$ is a polynomial. Since $M(x_0) \neq 0$, it is not identically zero, so it can be zero only at finitely many points. So, everywhere except maybe finitely many points $\operatorname{rank} A(x) \geq r$. But by the definition of $r$, $\operatorname{rank} A(x) \leq r$ for all $x$. □

## 7. Review exercises for Chapter 3.

**7.1.** True or false

a) Determinant is only defined for square matrices.

b) If two rows or columns of $A$ are identical, then $\det A = 0$.

c) If $B$ is the matrix obtained from $A$ by interchanging two rows (or columns), then $\det B = \det A$.

d) If $B$ is the matrix obtained from $A$ by multiplying a row (column) of $A$ by a scalar $\alpha$, then $\det B = \det A$.

e) If $B$ is the matrix obtained from $A$ by adding a multiple of a row to some other row, then $\det B = \det A$.

f) The determinant of a triangular matrix is the product of its diagonal entries.

g) $\det(A^T) = -\det(A)$.

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

i) A matrix $A$ is invertible if and only if $\det A \neq 0$.

j) If $A$ is an invertible matrix, then $\det(A^{-1}) = 1/\det(A)$.

**7.2.** Let $A$ be an $n \times n$ matrix. How are $\det(3A)$, $\det(-A)$ and $\det(A^2)$ related to $\det A$.

**7.3.** If the entries of both $A$ and $A^{-1}$ are integers, is it possible that $\det A = 3$? **Hint:** what is $\det(A)\det(A^{-1})$?

**7.4.** Let $\mathbf{v}_1, \mathbf{v}_2$ be vectors in $\mathbb{R}^2$ and let $A$ be the $2 \times 2$ matrix with columns $\mathbf{v}_1, \mathbf{v}_2$. Prove that $|\det A|$ is the area of the parallelogram with two sides given by the vectors $\mathbf{v}_1, \mathbf{v}_2$.

Consider first the case when $\mathbf{v}_1 = (x_1, 0)^T$. To treat general case $\mathbf{v}_1 = (x_1, y_1)^T$ left multiply $A$ by a rotation matrix that transforms vector $\mathbf{v}_1$ into $(\widetilde{x}_1, 0)^T$. **Hint:** what is the determinant of a rotation matrix?

The following problem illustrates relation between the sign of the determinant and the so-called *orientation* of a system of vectors.

**7.5.** Let $\mathbf{v}_1$, $\mathbf{v}_2$ be vectors in $\mathbb{R}^2$. Show that $D(\mathbf{v}_1, \mathbf{v}_2) > 0$ if and only if there exists a rotation $T_\alpha$ such that the vector $T_\alpha \mathbf{v}_1$ is parallel to $\mathbf{e}_1$ (and looking in the same direction), and $T_\alpha \mathbf{v}_2$ is in the upper half-plane $x_2 > 0$ (the same half-plane as $\mathbf{e}_2$).

**Hint:** What is the determinant of a rotation matrix?

*Chapter 4*

# Introduction to spectral theory (eigenvalues and eigenvectors)

Spectral theory is the main tool that helps us to understand the structure of a linear operator. In this chapter we consider only operators acting from a vector space to itself (or, equivalently, $n \times n$ matrices). If we have such a linear transformation $A : V \to V$, we can multiply it by itself, take any power of it, or any polynomial.

The main idea of spectral theory is to split the operator into simple blocks and analyze each block separately.

To explain the main idea, let us consider *difference equations*. Many processes can be described by the equations of the following type

$$\mathbf{x}_{n+1} = A\mathbf{x}_n, \qquad n = 0, 1, 2, \dots,$$

where $A : V \to V$ is a linear transformation, and $\mathbf{x}_n$ is the state of the system at the time $n$. Given the initial state $\mathbf{x}_0$ we would like to know the state $\mathbf{x}_n$ at the time $n$, analyze the long time behavior of $\mathbf{x}_n$, etc. [1]

[1] The difference equations are discrete time analogues of the *differential equation* $\mathbf{x}'(t) = A\mathbf{x}(t)$. To solve the differential equation, one needs to compute $e^{tA} := \sum_{k=0}^{\infty} t^k A^n/k!$, and spectral theory also helps in doing this.

At the first glance the problem looks trivial: the solution $\mathbf{x}_n$ is given by the formula $\mathbf{x}_n = A^n\mathbf{x}_0$. But what if $n$ is huge: thousands, millions? Or what if we want to analyze the behavior of $\mathbf{x}_n$ as $n \to \infty$?

Here the idea of *eigenvalues* and *eigenvectors* comes in. Suppose that $A\mathbf{x}_0 = \lambda\mathbf{x}_0$, where $\lambda$ is some scalar. Then $A^2\mathbf{x}_0 = \lambda^2\mathbf{x}_0$, $A^3\mathbf{x}_0 = \lambda^3\mathbf{x}_0, \dots,$ $A^n\mathbf{x}_0 = \lambda^n\mathbf{x}_0$, so the behavior of the solution is very well understood

In this section we will consider only operators in finite-dimensional spaces. Spectral theory in infinitely many dimensions is significantly more complicated, and most of the results presented here fail in infinite-dimensional setting.

## 1. Main definitions

**1.1. Eigenvalues, eigenvectors, spectrum.** A scalar $\lambda$ is called an *eigenvalue* of an operator $A : V \to V$ if there exists a *non-zero* vector $\mathbf{v} \in V$ such that

$$A\mathbf{v} = \lambda\mathbf{v}.$$

The vector $\mathbf{v}$ is called the *eigenvector* of $A$ (corresponding to the eigenvalue $\lambda$).

If we know that $\lambda$ is an eigenvalue, the eigenvectors are easy to find: one just has to solve the equation $A\mathbf{x} = \lambda\mathbf{x}$, or, equivalently

$$(A - \lambda I)\mathbf{x} = \mathbf{0}.$$

So, finding *all* eigenvectors, corresponding to an eigenvalue $\lambda$ is simply finding the nullspace of $A - \lambda I$. The nullspace $\operatorname{Ker}(A - \lambda I)$, i.e. the set of all eigenvectors and $\mathbf{0}$ vector, is called the *eigenspace*.

The set of all eigenvalues of an operator $A$ is called *spectrum* of $A$, and is usually denoted $\sigma(A)$.

**1.2. Finding eigenvalues: characteristic polynomials.** A scalar $\lambda$ is an eigenvalue if and only if the nullspace $\operatorname{Ker}(A - \lambda I)$ is non-trivial (so the equation $(A - \lambda I)\mathbf{x} = \mathbf{0}$ has a non-trivial solution).

Let $A$ act on $\mathbb{F}^n$ (i.e. $A : \mathbb{F}^n \to \mathbb{F}^n$). Since the matrix of $A$ is square, $A - \lambda I$ has a non-trivial nullspace if and only if it is not invertible. We know that a square matrix is not invertible if and only if its determinant is 0. Therefore

$$\boxed{\lambda \in \sigma(A),\ \text{i.e. } \lambda \text{ is an eigenvalue of } A \quad \Longleftrightarrow \quad \det(A - \lambda I) = 0}$$

If $A$ is an $n \times n$ matrix, the determinant $\det(A - \lambda I)$ is a polynomial of degree $n$ of the variable $\lambda$. This polynomial is called the *characteristic polynomial* of $A$. So, to find all eigenvalues of $A$ one just needs to compute the characteristic polynomial and find all its roots.

This method of finding the spectrum of an operator is not very practical in higher dimensions. Finding roots of a polynomial of high degree can be a very difficult problem, and it is impossible to solve the equation of degree higher than 4 in radicals. So, in higher dimensions different numerical methods of finding eigenvalues and eigenvectors are used.

**1.3. Finding characteristic polynomial and eigenvalues of an abstract operator.** So we know how to find the spectrum of a matrix. But how do we find eigenvalues of an operator acting in an abstract vector space? The recipe is simple:

| Take an arbitrary basis, and compute eigenvalues of the matrix of the operator in this basis. |
|---|

But how do we know that the result does not depend on a choice of the basis?

There can be several possible explanations. One is based on the notion of *similar matrices*. Let us recall that square matrices $A$ and $B$ are called similar if there exist an invertible matrix $S$ such that

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

Note, that determinants of similar matrices coincide. Indeed

$$\det A = \det(SBS^{-1}) = \det S \det B \det S^{-1} = \det B$$

because $\det S^{-1} = 1/\det S$. Note that if $A = SBS^{-1}$ then

$$A - \lambda I = SBS^{-1} - \lambda SIS^{-1} = S(BS^{-1} - \lambda IS^{-1}) = S(B - \lambda I)S^{-1},$$

so the matrices $A - \lambda I$ and $B - \lambda I$ are similar. Therefore

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

i.e.

| characteristic polynomials of similar matrices coincide. |
|---|

If $T : V \to V$ is a linear transformation, and $\mathcal{A}$ and $\mathcal{B}$ are two bases in $V$, then

$$[T]_{\mathcal{A}\mathcal{A}} = [I]_{\mathcal{A}\mathcal{B}}[T]_{\mathcal{B}\mathcal{B}}[I]_{\mathcal{B}\mathcal{A}}$$

and since $[I]_{\mathcal{B}\mathcal{A}} = ([I]_{\mathcal{A}\mathcal{B}})^{-1}$ the matrices $[T]_{\mathcal{A}\mathcal{A}}$ and $[T]_{\mathcal{B}\mathcal{B}}$ are similar.

In other words, matrices of a linear transformation in different bases are similar.

Therefore, we can define the characteristic polynomial of an operator as the characteristic polynomial of its matrix in some basis. As we have discussed above, the result does not depend on the choice of the basis, so characteristic polynomial of an operator is well defined.

**1.4. Complex vs real spaces.** The fundamental theorem of algebra asserts that any polynomial (of degree at least 1) has a complex root. That implies that an operator in a finite-dimensional *complex* vector space has at least one eigenvalue, so its spectrum is non-empty.

On the other hand it is easy to construct a linear transformation in a real vector space without *real* eigenvalues, the rotation $R_\alpha$, $\alpha \neq \pi n$ in $\mathbb{R}^2$ being one of examples. Since it is usually assumed that eigenvalues should belong to the field of scalars (if an operator acts in a vector space over a field $\mathbb{F}$ the eigenvalues should be in $\mathbb{F}$), such operators have empty spectrum.

Thus, the complex case (i.e. operators acting in complex vector spaces) seems to be the most natural setting for the spectral theory. Since $\mathbb{R} \subset \mathbb{C}$, we can always treat a real $n \times n$ matrix as an operator in $\mathbb{C}^n$ to allow complex eigenvalues. Treating real matrices as operators in $\mathbb{C}^n$ is typical in the spectral theory, and we will follow this agreement. Finding eigenvalues of a matrix (unless otherwise specified) will always mean finding all *complex* eigenvalues and not restricting oneself only to real ones.

Note that an operator in an abstract real vector space also can be interpreted as an operator in a complex space. A naïve approach would be to fix a basis (recall that all spaces in this chapter are finite-dimensional), and then work with coordinates in this basis allowing complex coordinates: that will be essentially move from operators in $\mathbb{R}^n$ to operators $\mathbb{C}^n$ described above.

This construction describes what is known as the *complexification* of a real vector space, and the result does not depend on the choice of a basis. A "high brow" abstract construction of the complexification, explaining why the result does not depend on the choice of a basis is described below in Section 8.2 of Chapter 5.

**1.5. Multiplicities of eigenvalues.** Let us remind the reader, that if $p$ is a polynomial, and $\lambda$ is its root (i.e. $p(\lambda) = 0$) then $z - \lambda$ divides $p(z)$, i.e. $p$ can be represented as $p(z) = (z - \lambda)q(z)$, where $q$ is some polynomial. If $q(\lambda) = 0$, then $q$ also can be divided by $z - \lambda$, so $(z - \lambda)^2$ divides $p$ and so on.

The largest positive integer $k$ such that $(z - \lambda)^k$ divides $p(z)$ is called the multiplicity of the root $\lambda$.

If $\lambda$ is an eigenvalue of an operator (matrix) $A$, then it is a root of the characteristic polynomial $p(z) = \det(A - zI)$. The multiplicity of this root is called the *(algebraic) multiplicity* of the eigenvalue $\lambda$.

Any polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$ of degree $n$ has exactly $n$ complex roots, counting multiplicity. The words *counting multiplicities* mean that if a root has multiplicity $d$ we have to list (count) it $d$ times. In other words,

$p$ can be represented as

$$p(z) = a_n(z - \lambda_1)(z - \lambda_2)\dots(z - \lambda_n).$$

where $\lambda_1, \lambda_2, \dots, \lambda_n$ are its complex roots, counting multiplicities.

There is another notion of multiplicity of an eigenvalue: the dimension of the eigenspace $\operatorname{Ker}(A - \lambda I)$ is called *geometric multiplicity* of the eigenvalue $\lambda$.

Geometric multiplicity is not as widely used as algebraic multiplicity. So, when people say simply "multiplicity" they usually mean *algebraic multiplicity*.

Let us mention, that algebraic and geometric multiplicities of an eigenvalue can differ.

**Proposition 1.1.** *Geometric multiplicity of an eigenvalue cannot exceed its algebraic multiplicity.*

**Proof.** See Exercise 1.9 below. □

### 1.6. Trace and determinant.

**Theorem 1.2.** *Let $A$ be $n \times n$ matrix, and let $\lambda_1, \lambda_2, \dots, \lambda_n$ be its (complex) eigenvalues (counting multiplicities). Then*

1. $\operatorname{trace} A = \lambda_1 + \lambda_2 + \dots + \lambda_n$.
2. $\det A = \lambda_1 \lambda_2 \dots \lambda_n$.

**Proof.** See Exercises 1.10, 1.11 below. □

**1.7. Eigenvalues of a triangular matrix.** Computing eigenvalues is equivalent to finding roots of a characteristic polynomial of a matrix (or using some numerical method), which can be quite time consuming. However, there is one particular case, when we can just read eigenvalues off the matrix. Namely

| eigenvalues of a triangular matrix (counting multiplicities) are exactly the diagonal entries $a_{1,1}, a_{2,2}, \dots, a_{n,n}$ |
|---|

By triangular here we mean either upper or lower triangular matrix. Since a diagonal matrix is a particular case of a triangular matrix (it is both upper and lower triangular

| the eigenvalues of a diagonal matrix are its diagonal entries |
|---|

The proof of the statement about triangular matrices is trivial: we need to subtract $\lambda$ from the diagonal entries of $A$, and use the fact that determinant of a triangular matrix is the product of its diagonal entries. We get

the characteristic polynomial

$$\det(A - \lambda I) = (a_{1,1} - \lambda)(a_{2,2} - \lambda) \dots (a_{n,n} - \lambda)$$

and its roots are exactly $a_{1,1}, a_{2,2}, \dots, a_{n,n}$.

**Exercises.**

**1.1.** True or false:

a) Every linear operator in an $n$-dimensional vector space has $n$ distinct eigenvalues;

b) If a matrix has one eigenvector, it has infinitely many eigenvectors;

c) There exists a square real matrix with no real eigenvalues;

d) There exists a square matrix with no (complex) eigenvectors;

e) Similar matrices always have the same eigenvalues;

f) Similar matrices always have the same eigenvectors;

g) The sum of two eigenvectors of a matrix $A$ is always an eigenvector;

h) The sum of two eigenvectors of a matrix $A$ corresponding to the same eigenvalue $\lambda$ is always an eigenvector.

**1.2.** Find characteristic polynomials, eigenvalues and eigenvectors of the following matrices:

$$\begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 1 \\ -1 & 4 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 3 & 3 \\ -3 & -5 & -3 \\ 3 & 3 & 1 \end{pmatrix}.$$

**1.3.** Compute eigenvalues and eigenvectors of the rotation matrix

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Note, that the eigenvalues (and eigenvectors) do not need to be real.

**1.4.** Compute characteristic polynomials and eigenvalues of the following matrices:

$$\begin{pmatrix} 1 & 2 & 5 & 67 \\ 0 & 2 & 3 & 6 \\ 0 & 0 & -2 & 5 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 1 & 0 & 2 \\ 0 & \pi & 43 & 2 \\ 0 & 0 & 16 & 1 \\ 0 & 0 & 0 & 54 \end{pmatrix}, \qquad \begin{pmatrix} 4 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 2 & 4 & e & 0 \\ 3 & 3 & 1 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 4 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 \\ 3 & 3 & 1 & 1 \end{pmatrix}.$$

Do not expand the characteristic polynomials, leave them as products.

**1.5.** Prove that eigenvalues (counting multiplicities) of a triangular matrix coincide with its diagonal entries

**1.6.** An operator $A$ is called *nilpotent* if $A^k = \mathbf{0}$ for some $k$. Prove that if $A$ is nilpotent, then $\sigma(A) = \{0\}$ (i.e. that 0 is the only eigenvalue of $A$).

**1.7.** Show that characteristic polynomial of a block triangular matrix

$$\begin{pmatrix} A & * \\ \mathbf{0} & B \end{pmatrix},$$

where $A$ and $B$ are square matrices, coincides with $\det(A-\lambda I)\det(B-\lambda I)$. (Use Exercise 3.11 from Chapter 3).

**1.8.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be a basis in a vector space $V$. Assume also that the first $k$ vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ of the basis are eigenvectors of an operator $A$, corresponding to an eigenvalue $\lambda$ (i.e. that $A\mathbf{v}_j = \lambda\mathbf{v}_j$, $j = 1, 2, \dots, k$). Show that in this basis the matrix of the operator $A$ has block triangular form

$$\begin{pmatrix} \lambda I_k & * \\ \mathbf{0} & B \end{pmatrix},$$

where $I_k$ is $k \times k$ identity matrix and $B$ is some $(n-k) \times (n-k)$ matrix.

**1.9.** Use the two previous exercises to prove that geometric multiplicity of an eigenvalue cannot exceed its algebraic multiplicity.

**1.10.** Prove that determinant of a matrix $A$ is the product of its eigenvalues (counting multiplicities).

**Hint:** first show that $\det(A-\lambda I) = (\lambda_1-\lambda)(\lambda_2-\lambda)\dots(\lambda_n-\lambda)$, where $\lambda_1, \lambda_2, \dots, \lambda_n$ are eigenvalues (counting multiplicities). Then compare the free terms (terms without $\lambda$) or plug in $\lambda = 0$ to get the conclusion.

**1.11.** Prove that the trace of a matrix equals the sum of eigenvalues in three steps. First, compute the coefficient of $\lambda^{n-1}$ in the right side of the equality

$$\det(A-\lambda I) = (\lambda_1-\lambda)(\lambda_2-\lambda)\dots(\lambda_n-\lambda).$$

Then show that $\det(A-\lambda I)$ can be represented as

$$\det(A-\lambda I) = (a_{1,1}-\lambda)(a_{2,2}-\lambda)\dots(a_{n,n}-\lambda) + q(\lambda)$$

where $q(\lambda)$ is polynomial of degree at most $n-2$. And finally, comparing the coefficients of $\lambda^{n-1}$ get the conclusion.

## 2. Diagonalization.

One of the application of the spectral theory is the diagonalization of operators, which means given an operator to find a basis in which the matrix of the operator is diagonal. Such basis does not always exists, i.e not all operators can be diagonalized (are diagonalizable). Importance of diagonalizable operators comes from the fact that the powers, and more general function of diagonal matrices are easy to compute, so if we diagonalize an operator we can easily compute functions of it.

We will explain how to compute functions of diagonalizable operators in this section. We also give a necessary and sufficient condition for an operator to be diagonalizable, as well as some simple sufficient conditions.

Note also that for operators in $\mathbb{F}^n$ (matrices) the diagonalizability of $A$ means that it can be represented as $A = SDS^{-1}$, where $D$ is a diagonal matrix (and $S$, of course, is invertible); we will explain this shortly.

Unless otherwise specified, all results in this section hold for both complex and real vector spaces (and even for spaces over arbitrary fields).

**2.1. Preliminaries.** Suppose an operator $A$ in a vector space $V$ is such that $V$ has a basis $\mathcal{B} = \mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ of eigenvectors of $A$, with $\lambda_1, \lambda_2, \dots, \lambda_n$ being the corresponding eigenvalues. Then the matrix of $A$ in this basis is the diagonal matrix with $\lambda_1, \lambda_2, \dots, \lambda_n$ on the diagonal

$$[A]_{\mathcal{B}\mathcal{B}} = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}. \tag{2.1}$$

On the other hand, if the matrix of an operator $A$ in a basis $\mathcal{B} = \mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is given by (2.1) then trivially $A\mathbf{b}_k = \lambda_k \mathbf{b}_k$, i.e $\lambda_k$ are eigenvalues and $\mathbf{b}_k$ are corresponding eigenvectors.

Note that the above reasoning hods for both complex and real vector spaces (and even for vector spaces over arbitrary fields)

Applying the above reasoning to operators in $\mathbb{F}^n$ (matrices) we immediately get the following theorem. Note, that while in this book $\mathbb{F}$ is either $\mathbb{C}$ or $\mathbb{R}$, this theorem hods for an arbitrary field $\mathbb{F}$.

**Theorem 2.1.** *A matrix $A$ (with values in $\mathbb{F}$) admits a representation $A = SDS^{-1}$, where $D$ is a diagonal matrix and $S$ is an invertible one (both with entries in $\mathbb{F}$) if and only if there exists a basis in $\mathbb{F}^n$ of eigenvectors of $A$.*

*Moreover, in this case diagonal entries of $D$ are the eigenvalues and the columns of $S$ are the corresponding eigenvectors (column number $k$ corresponds to $k$th diagonal entry of $D$).*

**Proof.** Let $D = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$, and let $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ be the columns of $S$ (note that since $S$ is invertible its columns form a basis in $\mathbb{F}^n$). Then the identity $A = SDS^{-1}$ means that $D = [A]_{\mathcal{B},\mathcal{B}}$.

Indeed, $S = [I]_{\mathcal{S},\mathcal{B}}$ is the change of the coordinates matrix from $\mathcal{B}$ to the standard basis $\mathcal{S}$, so we get from $A = SDS^{-1}$ that $D = S^{-1}AS = [I]_{\mathcal{B},\mathcal{S}} A [I]_{\mathcal{S},\mathcal{B}}]$, which means exactly that $D = [A]_{\mathcal{B},\mathcal{B}}$.

And as we just discussed above, $[A]_{\mathcal{B},\mathcal{B}} = D = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ if and only if $\lambda_k$ are the eigenvalues and $\mathbf{b}_k$ are the corresponding eigenvectors of $A$. □

**Remark.** Note if a matrix admits the representation $A = SDS^{-1}$ with a diagonal matrix $D$, then a simple direct calculation shows that the columns of $S$ are eigenvectors of $A$ and diagonal entries of $D$ are corresponding eigenvalues. This gives an alternative proof of the corresponding statement in Theorem 2.1.

As we discussed above, a diagonalizable operator $A : V \to V$ has exactly $n = \dim V$ eigenvalues (counting multiplicities). Any operator in a complex vector space $V$ has $n$ eigenvalues (counting multiplicities); an operator in a real space, on the other hand, could have no real eigenvalues.

We will, as it is customary in the spectral theory, treat real matrices as operators in the complex space $\mathbb{C}^n$, thus allowing complex eigenvalues and eigenvectors. Unless otherwise specified we will mean by the diagonalization of a matrix its *complex* diagonalization, i.e. a representation $A = SDS^{-1}$ where matrices $S$ and $D$ can have complex entries.

The question when a real matrix admits a real diagonalization ($A = SDS^{-1}$ with real matrices $S$ and $D$) is in fact a very simple one, see Theorem 2.9 below.

**2.2. Some motivations: functions of operators.** Let the matrix of an operator $A$ in a basis $\mathcal{B} = \mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is a diagonal one given by (2.1). Then it is easy to find an $N$th power of the operator $A$. Namely, the matrix of $A^N$ in the basis $\mathcal{B}$ is

$$[A^N]_{\mathcal{B}\mathcal{B}} = \operatorname{diag}\{\lambda_1^N, \lambda_2^N, \dots, \lambda_n^N\} = \begin{pmatrix} \lambda_1^N & & & 0 \\ & \lambda_2^N & & \\ & & \ddots & \\ 0 & & & \lambda_n^N \end{pmatrix}.$$

Moreover, functions of the operator are also very easy to compute: for example the operator (matrix) exponent $e^{tA}$ is defined as $e^{tA} = I + tA + \dfrac{t^2A^2}{2!} + \dfrac{t^3A^3}{3!} + \ldots = \displaystyle\sum_{k=0}^{\infty} \frac{t^kA^k}{k!}$, and its matrix in the basis $\mathcal{B}$ is

$$[e^{tA}]_{\mathcal{B}\mathcal{B}} = \operatorname{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_n t}\} = \begin{pmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n t} \end{pmatrix}.$$

Let now $A$ be an operator in $\mathbb{F}^n$. To find the matrices of the operators $A^N$ and $e^{tA}$ in the standard basis $\mathcal{S}$, we need to recall that the change of coordinate matrix $[I]_{\mathcal{S}\mathcal{B}}$ is the matrix with columns $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$. Let us

call this matrix $S$, then according to the change of coordinates formula we have

$$A = [A]_{\mathcal{SS}} = S \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} S^{-1} = SDS^{-1},$$

where we use $D$ for the diagonal matrix in the middle.

Similarly

$$A^N = SD^N S^{-1} = S \begin{pmatrix} \lambda_1^N & & & 0 \\ & \lambda_2^N & & \\ & & \ddots & \\ 0 & & & \lambda_n^N \end{pmatrix} S^{-1} ,$$

and similarly for $e^{tA}$.

Another way of thinking about powers (or other functions) of diagonalizable operators is to see that if operator $A$ can be represented as $A = SDS^{-1}$, then

$$A^N = \underbrace{(SDS^{-1})(SDS^{-1})\dots(SDS^{-1})}_{N \text{ times}} = SD^N S^{-1}$$

and it is easy to compute the $N$th power of a diagonal matrix.

**2.3. The case of $n$ distinct eigenvalues.** We now present very simple *sufficient* condition for an operator to be diagonalizable, see Corollary 2.3 below.

**Theorem 2.2.** *Let $\lambda_1, \lambda_2, \dots, \lambda_r$ be distinct eigenvalues of $A$, and let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ be the corresponding eigenvectors. Then vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ are linearly independent.*

**Proof.** We will use induction on $r$. The case $r = 1$ is trivial, because by the definition an eigenvector is non-zero, and a system consisting of one non-zero vector is linearly independent.

Suppose that the statement of the theorem is true for $r-1$. Suppose there exists a non-trivial linear combination

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_r\mathbf{v}_r = \sum_{k=1}^{r} c_k\mathbf{v}_k = \mathbf{0}. \tag{2.2}$$

Applying $A - \lambda_r I$ to (2.2) and using the fact that $(A - \lambda_r I)\mathbf{v}_r = \mathbf{0}$ we get

$$\sum_{k=1}^{r-1} c_k(\lambda_k - \lambda_r)\mathbf{v}_k = \mathbf{0}.$$

By the induction hypothesis vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{r-1}$ are linearly independent, so $c_k(\lambda_k - \lambda_r) = 0$ for $k = 1, 2, \dots, r-1$. Since $\lambda_k \neq \lambda_r$ we can conclude that $c_k = 0$ for $k < r$. Then it follows from (2.2) that $c_r = 0$, i.e. we have the trivial linear combination. □

**Corollary 2.3.** *If an operator $A : V \to V$ has exactly $n = \dim V$ distinct eigenvalues, then it is diagonalizable.*

While very simple, this is a very important statement, and it will be used a lot! Please remember it.

**Proof.** For each eigenvalue $\lambda_k$ let $\mathbf{v}_k$ be a corresponding eigenvector (just pick one eigenvector for each eigenvalue). By Theorem 2.2 the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is linearly independent, and since it consists of exactly $n = \dim V$ vectors it is a basis. □

**2.4. Bases of subspaces (AKA direct sums of subspaces).** To describe diagonalizable operators we need to introduce some new definitions.

Let $V_1, V_2, \dots, V_p$ be subspaces of a vector space $V$. We say that the system of subspaces is a basis in $V$ if any vector $\mathbf{v} \in V$ admits a unique representation as a sum

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_p = \sum_{k=1}^{p} \mathbf{v}_k, \qquad \mathbf{v}_k \in V_k. \tag{2.3}$$

We also say, that a system of subspaces $V_1, V_2, \dots, V_p$ is linearly independent if the equation

$$\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_p = \mathbf{0}, \qquad \mathbf{v}_k \in V_k$$

has only trivial solution ($\mathbf{v}_k = \mathbf{0}\ \forall k = 1, 2, \dots, p$).

Another way to phrase that is to say that a system of subspaces $V_1, V_2, \dots, V_p$ is linearly independent if and only if any system of non-zero vectors $\mathbf{v}_k$, where $\mathbf{v}_k \in V_k$, is linearly independent.

We say that the system of subspaces $V_1, V_2, \dots, V_p$ is generating (or complete, or spanning) if any vector $\mathbf{v} \in V$ admits representation as (2.3) (not necessarily unique).

**Remark 2.4.** From the above definition one can immediately see that Theorem 2.2 in fact states that the system of eigenspaces $E_k$ of an operator $A$

$$E_k := \operatorname{Ker}(A - \lambda_k I), \qquad \lambda_k \in \sigma(A),$$

is linearly independent.

**Remark 2.5.** It is easy to see that similarly to the bases of vectors, a system of subspaces $V_1, V_2, \dots, V_p$ is a basis if and only if it is generating and linearly independent. We leave the proof of this fact as an exercise for the reader.

There is a simple example of a basis of subspaces. Let $V$ be a vector space with a basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Split the set of indices $1, 2, \ldots, n$ into $p$ subsets $\Lambda_1, \Lambda_2, \ldots, \Lambda_p$, and define subspaces $V_k := \operatorname{span}\{\mathbf{v}_j : j \in \Lambda_k\}$. Clearly the subspaces $V_k$ form a basis of $V$.

The following theorem shows that in the finite-dimensional case it is essentially the only possible example of a basis of subspaces.

**Theorem 2.6.** *Let $V_1, V_2, \ldots, V_p$ be a basis of subspaces, and let us have in each subspace $V_k$ a basis (of vectors) $\mathcal{B}_k$[2]. Then the union $\cup_k \mathcal{B}_k$ of these bases is a basis in $V$.*

To prove the theorem we need the following lemma

**Lemma 2.7.** *Let $V_1, V_2, \ldots, V_p$ be a linearly independent family of subspaces, and let us have in each subspace $V_k$ a linearly independent system $\mathcal{B}_k$ of vectors [3] Then the union $\mathcal{B} := \cup_k \mathcal{B}_k$ is a linearly independent system.*

**Proof.** The proof of the lemma is almost trivial, if one thinks a bit about it. The main difficulty in writing the proof is a choice of a appropriate notation. Instead of using two indices (one for the number $k$ and the other for the number of a vector in $\mathcal{B}_k$, let us use "flat" notation.

Namely, let $n$ be the number of vectors in $\mathcal{B} := \cup_k \mathcal{B}_k$. Let us order the set $\mathcal{B}$, for example as follows: first list all vectors from $\mathcal{B}_1$, then all vectors in $\mathcal{B}_2$, etc, listing all vectors from $\mathcal{B}_p$ last.

This way, we index all vectors in $\mathcal{B}$ by integers $1, 2, \ldots, n$, and the set of indices $\{1, 2, \ldots, n\}$ splits into the sets $\Lambda_1, \Lambda_2, \ldots, \Lambda_p$ such that the set $\mathcal{B}_k$ consists of vectors $\mathbf{b}_j : j \in \Lambda_k$.

Suppose we have a non-trivial linear combination

$$c_1\mathbf{b}_1 + c_2\mathbf{b}_2 + \ldots + c_n\mathbf{b}_n = \sum_{j=1}^{n} c_j\mathbf{b}_j = \mathbf{0}. \tag{2.4}$$

Denote

$$\mathbf{v}_k = \sum_{j \in \Lambda_k} c_j\mathbf{b}_j.$$

Then (2.4) can be rewritten as

$$\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_p = \mathbf{0}.$$

[2] We do not list the vectors in $\mathcal{B}_k$, one just should keep in mind that each $\mathcal{B}_k$ consists of finitely many vectors in $V_k$

[3] Again, here we do not name each vector in $\mathcal{B}_k$ individually, we just keep in mind that each set $\mathcal{B}_k$ consists of finitely many vectors.

Since $\mathbf{v}_k \in V_k$ and the system of subspaces $V_k$ is linearly independent, $\mathbf{v}_k = \mathbf{0}$ $\forall k$. Than means that for every $k$

$$\sum_{j\in\Lambda_k} c_j \mathbf{b}_j = \mathbf{0},$$

and since the system of vectors $\mathbf{b}_j : j \in \Lambda_k$ (i.e. the system $\mathcal{B}_k$) are linearly independent, we have $c_j = 0$ for all $j \in \Lambda_k$. Since it is true for all $\Lambda_k$, we can conclude that $c_j = 0$ for all $j$. □

**Proof of Theorem 2.6.** To prove the theorem we will use the same notation as in the proof of Lemma 2.7, i.e. the system $\mathcal{B}_k$ consists of vectors $\mathbf{b}_j$, $j \in \Lambda_k$.

Lemma 2.7 asserts that the system of vectors $\mathbf{b}_j$, $j = 1, 2, \dots, n$ is linearly independent, so it only remains to show that the system is complete.

Since the system of subspaces $V_1, V_2, \dots, V_p$ is a basis, any vector $\mathbf{v} \in V$ can be represented as

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_p = \sum_{k=1}^{p} \mathbf{v}_k, \qquad \mathbf{v}_k \in V_k.$$

Since the vectors $\mathbf{b}_j$, $j \in \Lambda_k$ form a basis in $V_k$, the vectors $\mathbf{v}_k$ can be represented as

$$\mathbf{v}_k = \sum_{j\in\Lambda_k} c_j \mathbf{b}_j,$$

and therefore $\mathbf{v} = \sum_{j=1}^{n} c_j \mathbf{b}_j$. □

**2.5. Criterion of diagonalizability.** First of all let us recall a simple necessary condition. Since the eigenvalues (counting multiplicities) of a diagonal matrix $D = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ are exactly $\lambda_1, \lambda_2, \dots, \lambda_n$, we see that if an operator $A : V \to V$ is diagonalizable, it has exactly $n = \dim V$ eigenvalues (counting multiplicities).

Theorem below holds for both real and complex vector spaces (and even for spaces over genera fields).

**Theorem 2.8.** *Let an operator $A : V \to V$ has exactly $n = \dim V$ eigenvalues (counting multiplicities)[4]. Then $A$ is diagonalizable if and only if for each eigenvalue $\lambda$ the dimension of the eigenspace $\operatorname{Ker}(A - \lambda I)$ (i.e. the geometric multiplicity of $\lambda$) coincides with the algebraic multiplicity of $\lambda$.*

**Proof.** First of all let us note, that for a diagonal matrix, the algebraic and geometric multiplicities of eigenvalues coincide, and therefore the same holds for the diagonalizable operators.

[4]Since any operator in a complex vector space has exactly $n$ eigenvalues (counting multiplicities), this assumption is moot in the complex case.

Let us now prove the other implication. Let $\lambda_1, \lambda_2, \dots, \lambda_p$ be eigenvalues of $A$, and let $E_k := \operatorname{Ker}(A - \lambda_k I)$ be the corresponding eigenspaces. According to Remark 2.4, the subspaces $E_k$, $k = 1, 2, \dots, p$ are linearly independent.

Let $\mathcal{B}_k$ be a basis in $E_k$. By Lemma 2.7 the system $\mathcal{B} = \cup_k \mathcal{B}_k$ is a linearly independent system of vectors.

We know that each $\mathcal{B}_k$ consists of $\dim E_k$(= multiplicity of $\lambda_k$) vectors. So the number of vectors in $\mathcal{B}$ equal to the sum of multiplicities of eigenvalues $\lambda_k$. But the sum of multiplicities of the eigenvalues is the number of eigenvalues counting multiplicities, which is exactly $n = \dim V$. So, we have a linearly independent system of $n = \dim V$ eigenvectors, which means it is a basis. □

**2.6. Real factorization.** The theorem below is, in fact, already proven (iT is essentially Theorem 2.8 for real spaces). We state it here to summarize the situation with real diagonalization of real matrices.

**Theorem 2.9.** *A real $n \times n$ matrix $A$ admits a real factorization (i.e. representation $A = SDS^{-1}$ where $S$ and $D$ are real matrices, $D$ is diagonal and $S$ is invertible) if and only if it admits complex factorization and all eigenvalues of $A$ are real.*

**2.7. Some example.**

2.7.1. *Real eigenvalues.* Consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 8 & 1 \end{pmatrix}.$$

Its characteristic polynomial is equal to

$$\begin{vmatrix} 1-\lambda & 2 \\ 8 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 16$$

and its roots (eigenvalues) are $\lambda = 5$ and $\lambda = -3$. For the eigenvalue $\lambda = 5$

$$A - 5I = \begin{pmatrix} 1-5 & 2 \\ 8 & 1-5 \end{pmatrix} = \begin{pmatrix} -4 & 2 \\ 8 & -4 \end{pmatrix}$$

A basis in its nullspace consists of one vector $(1, 2)^T$, so this is the corresponding eigenvector.

Similarly, for $\lambda = -3$

$$A - \lambda I = A + 3I = \begin{pmatrix} 4 & 2 \\ 8 & 4 \end{pmatrix}$$

and the eigenspace $\mathrm{Ker}(A+3I)$ is spanned by the vector $(1,-2)^T$. The matrix $A$ can be diagonalized as

$$A = \begin{pmatrix} 1 & 2 \\ 8 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}^{-1}$$

2.7.2. *Complex eigenvalues.* Consider the matrix

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

Its characteristic polynomial is

$$\begin{vmatrix} 1-\lambda & 2 \\ -2 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 + 2^2$$

and the eigenvalues (roots of the characteristic polynomial are $\lambda = 1 \pm 2i$. For $\lambda = 1+2i$

$$A - \lambda I = \begin{pmatrix} -2i & 2 \\ -2 & -2i \end{pmatrix}$$

This matrix has rank 1, so the eigenspace $\mathrm{Ker}(A-\lambda T)$ is spanned by one vector, for example by $(1,i)^T$.

Since the matrix $A$ is real, we do not need to compute an eigenvector for $\lambda = 1-2i$: we can get it for free by taking the complex conjugate of the above eigenvector, see Exercise 2.2 below. So, for $\lambda = 1-2i$ a corresponding eigenvector is $(1,-i)^T$, and so the matrix $A$ can be diagonalized as

$$A = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} 1+2i & 0 \\ 0 & 1-2i \end{pmatrix} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}^{-1}.$$

2.7.3. *A non-diagonalizable matrix.* Consider the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Its characteristic polynomial is

$$\begin{vmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^2,$$

so $A$ has an eigenvalue 1 of multiplicity 2. But, it is easy to see that $\dim \mathrm{Ker}(A-I) = 1$ (1 pivot, so $2-1=1$ free variable). Therefore, the geometric multiplicity of the eigenvalue 1 is different from its algebraic multiplicity, so $A$ is not diagonalizable.

There is also an explanation which does not use Theorem 2.8. Namely, we got that the eigenspace $\mathrm{Ker}(A-1I)$ is one dimensional (spanned by the vector $(1,0)^T$). If $A$ were diagonalizable, it would have a diagonal form

$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ in some basis,[5] and so the dimension of the eigenspace wold be 2. Therefore $A$ cannot be diagonalized.

**Exercises.**

**2.1.** Let $A$ be $n \times n$ matrix. True or false:

a) $A^T$ has the same eigenvalues as $A$.

b) $A^T$ has the same eigenvectors as $A$.

c) If $A$ is is diagonalizable, then so is $A^T$.

Justify your conclusions.

**2.2.** Let $A$ be a square matrix with real entries, and let $\lambda$ be its complex eigenvalue. Suppose $\mathbf{v} = (v_1, v_2, \dots, v_n)^T$ is a corresponding eigenvector, $A\mathbf{v} = \lambda\mathbf{v}$. Prove that the $\overline{\lambda}$ is an eigenvalue of $A$ and $A\overline{\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$. Here $\overline{\mathbf{v}}$ is the complex conjugate of the vector $\mathbf{v}$, $\overline{\mathbf{v}} := (\overline{v}_1, \overline{v}_2, \dots, \overline{v}_n)^T$.

**2.3.** Let

$$A = \begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix}.$$

Find $A^{2004}$ by diagonalizing $A$.

**2.4.** Construct a matrix $A$ with eigenvalues 1 and 3 and corresponding eigenvectors $(1, 2)^T$ and $(1, 1)^T$. Is such a matrix unique?

**2.5.** Diagonalize the following matrices, if possible:

a) $\begin{pmatrix} 4 & -2 \\ 1 & 1 \end{pmatrix}$.

b) $\begin{pmatrix} -1 & -1 \\ 6 & 4 \end{pmatrix}$.

c) $\begin{pmatrix} -2 & 2 & 6 \\ 5 & 1 & -6 \\ -5 & 2 & 9 \end{pmatrix}$ ($\lambda = 2$ is one of the eigenvalues)

**2.6.** Consider the matrix

$$A = \begin{pmatrix} 2 & 6 & -6 \\ 0 & 5 & -2 \\ 0 & 0 & 4 \end{pmatrix}.$$

a) Find its eigenvalues. Is it possible to find the eigenvalues without computing?

b) Is this matrix diagonalizable? Find out without computing anything.

c) If the matrix is diagonalizable, diagonalize it.

---

[5] Note, that the only linear transformation having matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ in some basis is the identity transformation $I$. Since $A$ is definitely not the identity, we can immediately conclude that $A$ cannot be diagonalized, so counting dimension of the eigenspace is not necessary.

**2.7.** Diagonalize the matrix

$$\begin{pmatrix} 2 & 0 & 6 \\ 0 & 2 & 4 \\ 0 & 0 & 4 \end{pmatrix}.$$

**2.8.** Find all square roots of the matrix

$$A = \begin{pmatrix} 5 & 2 \\ -3 & 0 \end{pmatrix}$$

i.e. find all matrices $B$ such that $B^2 = A$. **Hint:** Finding a square root of a diagonal matrix is easy. You can leave your answer as a product.

**2.9.** Let us recall that the famous Fibonacci sequence:

$$0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots$$

is defined as follows: we put $\varphi_0 = 0$, $\varphi_1 = 1$ and define

$$\varphi_{n+2} = \varphi_{n+1} + \varphi_n.$$

We want to find a formula for $\varphi_n$. To do this

a) Find a $2 \times 2$ matrix $A$ such that

$$\begin{pmatrix} \varphi_{n+2} \\ \varphi_{n+1} \end{pmatrix} = A \begin{pmatrix} \varphi_{n+1} \\ \varphi_n \end{pmatrix}$$

**Hint:** Combine the trivial equation $\varphi_{n+1} = \varphi_{n+1}$ with the Fibonacci relation $\varphi_{n+2} = \varphi_{n+1} + \varphi_n$.

b) Diagonalize $A$ and find a formula for $A^n$.

c) Noticing that

$$\begin{pmatrix} \varphi_{n+1} \\ \varphi_n \end{pmatrix} = A^n \begin{pmatrix} \varphi_1 \\ \varphi_0 \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

find a formula for $\varphi_n$. (You will need to compute an inverse and perform multiplication here).

d) Show that the vector $(\varphi_{n+1}/\varphi_n, 1)^T$ converges to an eigenvector of $A$. What do you think, is it a coincidence?

**2.10.** Let $A$ be a $5 \times 5$ matrix with 3 eigenvalues (not counting multiplicities). Suppose we know that one eigenspace is three-dimensional. Can you say if $A$ is diagonalizable?

**2.11.** Give an example of a $3 \times 3$ matrix which cannot be diagonalized. After you constructed the matrix, can you make it "generic", so no special structure of the matrix could be seen?

**2.12.** Let a non-zero matrix $A$ satisfy $A^5 = \mathbf{0}$. Prove that $A$ cannot be diagonalized. More generally, any non-zero nilpotent matrix, i.e. a non-zero matrix satisfying $A^N = \mathbf{0}$ for some $N$ cannot be diagonalized.

**2.13.** Eigenvalues of a transposition:

a) Consider the transformation $T$ in the space $M_{2\times 2}$ of $2\times 2$ matrices, $T(A) = A^T$. Find all its eigenvalues and eigenvectors. Is it possible to diagonalize this transformation? **Hint:** While it is possible to write a matrix of this linear transformation in some basis, compute characteristic polynomial, and so on, it is easier to find eigenvalues and eigenvectors directly from the definition.

b) Can you do the same problem but in the space of $n \times n$ matrices?

**2.14.** Prove that two subspaces $V_1$ and $V_2$ are linearly independent if and only if $V_1 \cap V_2 = \{\mathbf{0}\}$.

*Chapter 5*

# Inner product spaces

Theory of inner product spaces is developed only for real and complex spaces, so $\mathbb{F}$ in this Chapter is always $\mathbb{R}$ or $\mathbb{C}$; the results usually do not generalize to spaces over arbitrary fields.

Most of the results and calculations in this chapter hold (and the results have the same statements) in both real and complex cases. In rare situations when there is a difference between real and complex case, we state explicitly which case is considered: otherwise everything holds for both cases.

Finally, when the results and calculations hold for both complex and real cases, we use formulas for the complex case; in the real case they give correct, although sometimes a bit more complicated, formulas.

## 1. Inner product in $\mathbb{R}^n$ and $\mathbb{C}^n$. Inner product spaces.

**1.1. Inner product and norm in $\mathbb{R}^n$.** In dimensions 2 and 3, we defined the length of a vector $\mathbf{x}$ (i.e. the distance from its endpoint to the origin) by the Pythagorean rule, for example in $\mathbb{R}^3$ the length of the vector is defined as

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + x_3^2}.$$

It is natural to generalize this formula for all $n$, to define the *norm* of the vector $\mathbf{x} \in \mathbb{R}^n$ as

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

The word *norm* is used as a fancy replacement for the word length.

The *dot product* in $\mathbb{R}^3$ was defined as $\mathbf{x} \cdot \mathbf{y} = x_1y_2 + x_2y_2 + x_3y_3$, where $\mathbf{x} = (x_1, x_2, x_3)^T$ and $\mathbf{y} = (y_1, y_2, y_3)^T$.

While the notation $\mathbf{x} \cdot \mathbf{y}$ and term "dot product" is often used for the inner product, for reasons which will be clear later, we prefer the notation $(\mathbf{x}, \mathbf{y})$

Similarly, in $\mathbb{R}^n$ one can define the *inner product* $(\mathbf{x}, \mathbf{y})$ of two vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ by

$$(\mathbf{x}, \mathbf{y}) := x_1y_1 + x_2y_2 + \dots + x_ny_n = \mathbf{y}^T\mathbf{x},$$

so $\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})}$.

Note, that $\mathbf{y}^T\mathbf{x} = \mathbf{x}^T\mathbf{y}$, and we use the notation $\mathbf{y}^T\mathbf{x}$ only to be consistent.

**1.2. Inner product and norm in $\mathbb{C}^n$.** Let us now define norm and inner product for $\mathbb{C}^n$. As we have seen before, the complex space $\mathbb{C}^n$ is the most natural space from the point of view of spectral theory: even if one starts from a matrix with real coefficients (or operator on a real vectors space), the eigenvalues can be complex, and one needs to work in a complex space.

For a complex number $z = x + iy$, we have $|z|^2 = x^2 + y^2 = z\overline{z}$. If $\mathbf{z} \in \mathbb{C}^n$ is given by

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} x_1 + iy_1 \\ x_2 + iy_2 \\ \vdots \\ x_n + iy_n \end{pmatrix},$$

it is natural to define its norm $\|\mathbf{z}\|$ by

$$\|\mathbf{z}\|^2 = \sum_{k=1}^{n} (x_k^2 + y_k^2) = \sum_{k=1}^{n} |z_k|^2.$$

Let us try to define an inner product on $\mathbb{C}^n$ such that $\|\mathbf{z}\|^2 = (\mathbf{z}, \mathbf{z})$. One of the choices is to define $(\mathbf{z}, \mathbf{w})$ by

$$(\mathbf{z}, \mathbf{w}) = z_1\overline{w}_1 + z_2\overline{w}_2 + \dots + z_n\overline{w}_n = \sum_{k=1}^{n} z_k\overline{w}_k,$$

and that will be our definition of the *standard* inner product in $\mathbb{C}^n$.

To simplify the notation, let us introduce a new notion. For a matrix $A$ let us define its *Hermitian adjoint*, or simply *adjoint* $A^*$ by $A^* = \overline{A}^T$, meaning that we take the transpose of the matrix, and then take the complex conjugate of each entry. Note, that for a real matrix $A$, $A^* = A^T$.

Using the notion of $A^*$, one can write the standard inner product in $\mathbb{C}^n$ as

$$(\mathbf{z}, \mathbf{w}) = \mathbf{w}^*\mathbf{z}.$$

**Remark.** It is easy to see that one can define a different inner product in $\mathbb{C}^n$ such that $\|\mathbf{z}\|^2 = (\mathbf{z}, \mathbf{z})$, namely the inner product given by

$$(\mathbf{z}, \mathbf{w})_1 = \overline{z}_1 w_1 + \overline{z}_2 w_2 + \ldots + \overline{z}_n w_n = \mathbf{z}^* \mathbf{w}.$$

We did not specify what properties we want the inner product to satisfy, but $\mathbf{z}^*\mathbf{w}$ and $\mathbf{w}^*\mathbf{z}$ are the only reasonable choices giving $\|\mathbf{z}\|^2 = (\mathbf{z}, \mathbf{z})$.

Note, that the above two choices of the inner product are essentially equivalent: the only difference between them is notational, because $(\mathbf{z}, \mathbf{w})_1 = (\mathbf{w}, \mathbf{z})$.

While the second choice of the inner product looks more natural, the first one, $(\mathbf{z}, \mathbf{w}) = \mathbf{w}^*\mathbf{z}$ is more widely used, so we will use it as well.

**1.3. Inner product spaces.** The inner product we defined for $\mathbb{R}^n$ and $\mathbb{C}^n$ satisfies the following properties:

1. (Conjugate) symmetry: $(\mathbf{x}, \mathbf{y}) = \overline{(\mathbf{y}, \mathbf{x})}$; note, that for a real space, this property is just symmetry, $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$;
2. Linearity: $(\alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z}) = \alpha(\mathbf{x}, \mathbf{z}) + \beta(\mathbf{y}, \mathbf{z})$ for all vector $\mathbf{x}, \mathbf{y}, \mathbf{z}$ and all scalars $\alpha, \beta$;
3. Non-negativity: $(\mathbf{x}, \mathbf{x}) \geq 0$ $\forall \mathbf{x}$;
4. Non-degeneracy: $(\mathbf{x}, \mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Let $V$ be a (complex or real) vector space. An *inner product* on $V$ is a function, that assign to each pair of vectors $\mathbf{x}$, $\mathbf{y}$ a scalar, denoted by $(\mathbf{x}, \mathbf{y})$ such that the above properties 1–4 are satisfied.

Note that for a real space $V$ we assume that $(\mathbf{x}, \mathbf{y})$ is always real, and for a complex space the inner product $(\mathbf{x}, \mathbf{y})$ can be complex.

A space $V$ together with an inner product on it is called *an inner product space*. Given an inner product space, one defines the norm on it by

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})}.$$

1.3.1. *Examples.*

**Example 1.1.** Let $V$ be $\mathbb{R}^n$ or $\mathbb{C}^n$. We already have an inner product $(\mathbf{x}, \mathbf{y}) = \mathbf{y}^*\mathbf{x} = \sum_{k=1}^n x_k \overline{y}_k$ defined above.

This inner product is called the *standard* inner product in $\mathbb{R}^n$ or $\mathbb{C}^n$

We will use symbol $\mathbb{F}$ to denote both $\mathbb{C}$ and $\mathbb{R}$. When we have some statement about the space $\mathbb{F}^n$, it means the statement is true for both $\mathbb{R}^n$ and $\mathbb{C}^n$.

**Example 1.2.** Let $V$ be the space $\mathbb{P}_n$ of polynomials of degree at most $n$. Define the inner product by

$$(f, g) = \int_{-1}^{1} f(t)\overline{g(t)}dt.$$

It is easy to check, that the above properties 1–4 are satisfied.

This definition works both for complex and real cases. In the real case we only allow polynomials with real coefficients, and we do not need the complex conjugate here.

Let us recall, that for a square matrix $A$, its *trace* is defined as the sum of the diagonal entries,

$$\operatorname{trace} A := \sum_{k=1}^{n} a_{k,k}.$$

**Example 1.3.** For the space $M_{m\times n}$ of $m \times n$ matrices let us define the so-called Frobenius inner product by

$$(A, B) = \operatorname{trace}(B^* A).$$

Again, it is easy to check that the properties 1–4 are satisfied, i.e. that we indeed defined an inner product.

Note, that

$$\operatorname{trace}(B^* A) = \sum_{j,k} A_{j,k}\overline{B}_{j,k},$$

so this inner product coincides with the standard inner product in $\mathbb{C}^{mn}$.

**1.4. Properties of inner product.** The statements we get in this section are true for any abstract inner product space, not only for $\mathbb{F}^n$. To prove them we use only properties 1–4 of the inner product.

First of all let us notice, that properties 1 and 2 imply that

$2'$. $(\mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z}) = \overline{\alpha}(\mathbf{x}, \mathbf{y}) + \overline{\beta}(\mathbf{x}, \mathbf{z})$.

Indeed,

$$(\mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z}) = \overline{(\alpha\mathbf{y} + \beta\mathbf{z}, \mathbf{x})} = \overline{\alpha(\mathbf{y}, \mathbf{x}) + \beta(\mathbf{z}, \mathbf{x})} = \\ = \overline{\alpha}\overline{(\mathbf{y}, \mathbf{x})} + \overline{\beta}\,\overline{(\mathbf{z}, \mathbf{x})} = \overline{\alpha}(\mathbf{x}, \mathbf{y}) + \overline{\beta}(\mathbf{x}, \mathbf{z})$$

Note also that property 2 implies that for all vectors $\mathbf{x}$

$$(\mathbf{0}, \mathbf{x}) = (\mathbf{x}, \mathbf{0}) = 0.$$

**Lemma 1.4.** *Let $\mathbf{x}$ be a vector in an inner product space $V$. Then $\mathbf{x} = \mathbf{0}$ if and only if*

$$(\mathbf{x}, \mathbf{y}) = 0 \qquad \forall \mathbf{y} \in V. \tag{1.1}$$

**Proof.** Since $(\mathbf{0}, \mathbf{y}) = 0$ we only need to show that (1.1) implies $\mathbf{x} = \mathbf{0}$. Putting $\mathbf{y} = \mathbf{x}$ in (1.1) we get $(\mathbf{x}, \mathbf{x}) = 0$, so $\mathbf{x} = \mathbf{0}$. □

Applying the above lemma to the difference $\mathbf{x} - \mathbf{y}$ we get the following

**Corollary 1.5.** *Let* $\mathbf{x}, \mathbf{y}$ *be vectors in an inner product space* $V$*. The equality* $\mathbf{x} = \mathbf{y}$ *holds if and only if*

$$(\mathbf{x}, \mathbf{z}) = (\mathbf{y}, \mathbf{z}) \qquad \forall \mathbf{z} \in V.$$

The following corollary is very simple, but will be used a lot

**Corollary 1.6.** *Suppose two operators* $A, B : X \to Y$ *satisfy*

$$(A\mathbf{x}, \mathbf{y}) = (B\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x} \in X,\ \forall \mathbf{y} \in Y.$$

*Then* $A = B$*.*

**Proof.** By the previous corollary (fix $\mathbf{x}$ and take all possible $\mathbf{y}$'s) we get $A\mathbf{x} = B\mathbf{x}$. Since this is true for all $\mathbf{x} \in X$, the transformations $A$ and $B$ coincide. □

The following property relates the norm and the inner product.

**Theorem 1.7** (Cauchy–Schwarz inequality)**.**

$$|(\mathbf{x}, \mathbf{y})| \le \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

**Proof.** The proof we are going to present, is not the shortest one, but it shows where the main ideas came from.

Let us consider the real case first. If $\mathbf{y} = \mathbf{0}$, the statement is trivial, so we can assume that $\mathbf{y} \neq \mathbf{0}$. By the properties of an inner product, for all scalar $t$

$$0 \le \|\mathbf{x} - t\mathbf{y}\|^2 = (\mathbf{x} - t\mathbf{y}, \mathbf{x} - t\mathbf{y}) = \|\mathbf{x}\|^2 - 2t(\mathbf{x}, \mathbf{y}) + t^2\|\mathbf{y}\|^2.$$

In particular, this inequality should hold for $t = \frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{y}\|^2}$ [1], and for this point the inequality becomes

$$0 \le \|\mathbf{x}\|^2 - 2\frac{(\mathbf{x}, \mathbf{y})^2}{\|\mathbf{y}\|^2} + \frac{(\mathbf{x}, \mathbf{y})^2}{\|\mathbf{y}\|^2} = \|\mathbf{x}\|^2 - \frac{(\mathbf{x}, \mathbf{y})^2}{\|\mathbf{y}\|^2},$$

which is exactly the inequality we need to prove.

There are several possible ways to treat the complex case. One is to replace $\mathbf{x}$ by $\alpha\mathbf{x}$, where $\alpha$ is a complex constant, $|\alpha| = 1$ such that $(\alpha\mathbf{x}, \mathbf{y})$ is real, and then repeat the proof for the real case.

The other possibility is again to consider

$$\begin{aligned} 0 \le \|\mathbf{x} - t\mathbf{y}\|^2 &= (\mathbf{x} - t\mathbf{y}, \mathbf{x} - t\mathbf{y}) = (\mathbf{x}, \mathbf{x} - t\mathbf{y}) - t(\mathbf{y}, \mathbf{x} - t\mathbf{y}) \\ &= \|\mathbf{x}\|^2 - t(\mathbf{y}, \mathbf{x}) - \bar{t}(\mathbf{x}, \mathbf{y}) + |t|^2\|\mathbf{y}\|^2. \end{aligned}$$

[1] That is the point where the above quadratic polynomial has a minimum: it can be computed, for example by taking the derivative in $t$ and equating it to 0

Substituting $t = \frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{y}\|^2} = \frac{\overline{(\mathbf{y},\mathbf{x})}}{\|\mathbf{y}\|^2}$ into this inequality, we get

$$0 \leq \|\mathbf{x}\|^2 - \frac{|(\mathbf{x},\mathbf{y})|^2}{\|\mathbf{y}\|^2}$$

which is the inequality we need.

Note, that the above paragraph is in fact a complete formal proof of the theorem. The reasoning before that was only to explain why do we need to pick this particular value of $t$. □

An immediate Corollary of the Cauchy–Schwarz Inequality is the following lemma.

**Lemma 1.8** (Triangle inequality)**.** *For any vectors* $\mathbf{x}$*,* $\mathbf{y}$ *in an inner product space*

$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

**Proof.**

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

□

The following *polarization identities* allow one to reconstruct the inner product from the norm:

**Lemma 1.9** (Polarization identities)**.** *For* $\mathbf{x},\mathbf{y} \in V$

$$(\mathbf{x},\mathbf{y}) = \frac{1}{4}\left(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\right)$$

*if* $V$ *is a real inner product space, and*

$$(\mathbf{x},\mathbf{y}) = \frac{1}{4}\sum_{\alpha=\pm1,\pm i} \alpha\|\mathbf{x}+\alpha\mathbf{y}\|^2$$

*if* $V$ *is a complex inner product space.*

The lemma is proved by direct computation. We leave the proof as an exercise for the reader.

Another important property of the norm in an inner product space can be also checked by direct calculation.

**Lemma 1.10** (Parallelogram Identity)**.** *For any vectors* $\mathbf{u},\mathbf{v}$

$$\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2).$$

In 2-dimensional space this lemma relates sides of a parallelogram with its diagonals, which explains the name. It is a well-known fact from planar geometry.

**1.5. Norm. Normed spaces.** We have proved before that the norm $\|\mathbf{v}\|$ satisfies the following properties:

1. Homogeneity: $\|\alpha\mathbf{v}\| = |\alpha| \cdot \|\mathbf{v}\|$ for all vectors $\mathbf{v}$ and all scalars $\alpha$.
2. Triangle inequality: $\|\mathbf{u} + \mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|$.
3. Non-negativity: $\|\mathbf{v}\| \ge 0$ for all vectors $\mathbf{v}$.
4. Non-degeneracy: $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

Suppose in a vector space $V$ we assigned to each vector $\mathbf{v}$ a number $\|\mathbf{v}\|$ such that above properties 1–4 are satisfied. Then we say that the function $\mathbf{v} \mapsto \|\mathbf{v}\|$ is a norm. A vector space $V$ equipped with a norm is called a *normed space*.

Any inner product space is a normed space, because the norm $\|\mathbf{v}\| = \sqrt{(\mathbf{v}, \mathbf{v})}$ satisfies the above properties 1–4. However, there are many other normed spaces. For example, given $p$, $1 \le p < \infty$ one can define the norm $\|\cdot\|_p$ on $\mathbb{R}^n$ or $\mathbb{C}^n$ by

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{1/p} = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p}.$$

One can also define the norm $\|\cdot\|_\infty$ ($p = \infty$) by

$$\|\mathbf{x}\|_\infty = \max\{|x_k| : k = 1, 2, \ldots, n\}.$$

The norm $\|\cdot\|_p$ for $p = 2$ coincides with the regular norm obtained from the inner product.

To check that $\|\cdot\|_p$ is indeed a norm one has to check that it satisfies all the above properties 1–4. Properties 1, 3 and 4 are very easy to check, we leave it as an exercise for the reader. The triangle inequality (property 2) is easy to check for $p = 1$ and $p = \infty$ (and we proved it for $p = 2$).

For all other $p$ the triangle inequality is true, but the proof is not so simple, and we will not present it here. The triangle inequality for $\|\cdot\|_p$ even has special name: its called *Minkowski inequality*, after the German mathematician H. Minkowski.

Note, that the norm $\|\cdot\|_p$ for $p \ne 2$ cannot be obtained from an inner product. It is easy to see that this norm is not obtained from the *standard* inner product in $\mathbb{R}^n$ ($\mathbb{C}^n$). But we claim more! We claim that it is *impossible to introduce* an inner product which gives rise to the norm $\|\cdot\|_p$, $p \ne 2$.

This statement is actually quite easy to prove. By Lemma 1.10 any norm obtained from an inner product must satisfy the Parallelogram Identity. It

is easy to see that the Parallelogram Identity fails for the norm $\|\cdot\|_p$, $p \neq 2$, and one can easily find a counterexample in $\mathbb{R}^2$, which then gives rise to a counterexample in all other spaces.

In fact, the Parallelogram Identity, as the theorem below asserts, completely characterizes norms obtained from an inner product.

**Theorem 1.11.** *A norm in a normed space is obtained from some inner product if and only if it satisfies the Parallelogram Identity*

$$\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2) \qquad \forall \mathbf{u}, \mathbf{v} \in V.$$

Lemma 1.10 asserts that a norm obtained from an inner product satisfies the Parallelogram Identity.

The converse implication is more complicated. If we are given a norm, and this norm came from an inner product, then we do not have any choice; this inner product must be given by the polarization identities, see Lemma 1.9. But, we need to show that $(\mathbf{x}, \mathbf{y})$ which we got from the polarization identities is indeed an inner product, i.e. that it satisfies all the properties.

It is indeed possible to verify that if the norm satisfies the parallelogram identity then the inner product $(\mathbf{x}, \mathbf{y})$ obtained from the polarization identities is indeed an inner product (i.e. satisfies all the properties of an inner product). However, the proof is a bit too involved, so we do not present it here.

**Exercises.**

**1.1.** Compute

$$(3+2i)(5-3i), \qquad \frac{2-3i}{1-2i}, \qquad \operatorname{Re}\left(\frac{2-3i}{1-2i}\right), \qquad (1+2i)^3, \qquad \operatorname{Im}((1+2i)^3).$$

**1.2.** For vectors $\mathbf{x} = (1, 2i, 1+i)^T$ and $\mathbf{y} = (i, 2-i, 3)^T$ compute

a) $(\mathbf{x}, \mathbf{y})$, $\|\mathbf{x}\|^2$, $\|\mathbf{y}\|^2$, $\|\mathbf{y}\|$;

b) $(3\mathbf{x}, 2i\mathbf{y})$, $(2\mathbf{x}, i\mathbf{x} + 2\mathbf{y})$;

c) $\|\mathbf{x} + 2\mathbf{y}\|$.

**Remark:** After you have done part a), you can do parts b) and c) without actually computing all vectors involved, just by using the properties of inner product.

**1.3.** Let $\|\mathbf{u}\| = 2$, $\|\mathbf{v}\| = 3$, $(\mathbf{u}, \mathbf{v}) = 2 + i$. Compute

$$\|\mathbf{u}+\mathbf{v}\|^2, \qquad \|\mathbf{u}-\mathbf{v}\|^2, \qquad (\mathbf{u}+\mathbf{v}, \mathbf{u}-i\mathbf{v}), \qquad (\mathbf{u}+3i\mathbf{v}, 4i\mathbf{u}).$$

**1.4.** Prove that for vectors in a inner product space

$$\|\mathbf{x} \pm \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \pm 2\operatorname{Re}(\mathbf{x}, \mathbf{y})$$

Recall that $\operatorname{Re} z = \frac{1}{2}(z + \overline{z})$

**1.5.** Explain why each of the following *is not* an inner product on a given vector space:

a) $(\mathbf{x}, \mathbf{y}) = x_1 y_1 - x_2 y_2$ on $\mathbb{R}^2$;

b) $(A, B) = \operatorname{trace}(A + B)$ on the space of real $2 \times 2$ matrices'

c) $(f, g) = \int_0^1 f'(t)\overline{g(t)}dt$ on the space of polynomials; $f'(t)$ denotes derivative.

**1.6** (Equality in Cauchy–Schwarz inequality)**.** Prove that

$$|(\mathbf{x}, \mathbf{y})| = \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

if and only if one of the vectors is a multiple of the other. **Hint:** Analyze the proof of the Cauchy–Schwarz inequality.

**1.7.** Prove the parallelogram identity for an inner product space $V$,

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

**1.8.** Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a spanning set (in particular, a basis) in an inner product space $V$. Prove that

a) If $(\mathbf{x}, \mathbf{v}) = 0$ for all $\mathbf{v} \in V$, then $\mathbf{x} = \mathbf{0}$;

b) If $(\mathbf{x}, \mathbf{v}_k) = 0$ $\forall k$, then $\mathbf{x} = \mathbf{0}$;

c) If $(\mathbf{x}, \mathbf{v}_k) = (\mathbf{y}, \mathbf{v}_k)$ $\forall k$, then $\mathbf{x} = \mathbf{y}$.

**1.9.** Consider the space $\mathbb{R}^2$ with the norm $\|\cdot\|_p$, introduced in Section 1.5. For $p = 1, 2, \infty$, draw the "unit ball" $B_p$ in the norm $\|\cdot\|_p$

$$B_p := \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_p \le 1\}.$$

Can you guess what the balls $B_p$ for other $p$ look like?

## 2. Orthogonality. Orthogonal and orthonormal bases.

**Definition 2.1.** Two vectors $\mathbf{u}$ and $\mathbf{v}$ are called *orthogonal* (also *perpendicular*) if $(\mathbf{u}, \mathbf{v}) = 0$. We will write $\mathbf{u} \perp \mathbf{v}$ to say that the vectors are orthogonal.

Note, that for orthogonal vectors $\mathbf{u}$ and $\mathbf{v}$ we have the following, so-called Pythagorean identity:

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 \qquad \text{if } \mathbf{u} \perp \mathbf{v}.$$

The proof is straightforward computation,

$$\|\mathbf{u} + \mathbf{v}\|^2 = (\mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v}) = (\mathbf{u}, \mathbf{u}) + (\mathbf{v}, \mathbf{v}) + (\mathbf{u}, \mathbf{v}) + (\mathbf{v}, \mathbf{u}) = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$

($(\mathbf{u}, \mathbf{v}) = (\mathbf{v}, \mathbf{u}) = 0$ because of orthogonality).

**Definition 2.2.** We say that a vector $\mathbf{v}$ is orthogonal to a subspace $E$ if $\mathbf{v}$ is orthogonal to all vectors $\mathbf{w}$ in $E$.

We say that subspaces $E$ and $F$ are orthogonal if all vectors in $E$ are orthogonal to $F$, i.e. all vectors in $E$ are orthogonal to all vectors in $F$

The following lemma shows how to check that a vector is orthogonal to a subspace.

**Lemma 2.3.** *Let $E$ be spanned by vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$. Then $\mathbf{v} \perp E$ if and only if*

$$\mathbf{v} \perp \mathbf{v}_k, \qquad \forall k = 1, 2, \dots, r.$$

**Proof.** By the definition, if $\mathbf{v} \perp E$ then $\mathbf{v}$ is orthogonal to all vectors in $E$. In particular, $\mathbf{v} \perp \mathbf{v}_k$, $k = 1, 2, \dots, r$.

On the other hand, let $\mathbf{v} \perp \mathbf{v}_k$, $k = 1, 2, \dots, r$. Since the vectors $\mathbf{v}_k$ span $E$, any vector $\mathbf{w} \in E$ can be represented as a linear combination $\sum_{k=1}^{r} \alpha_k \mathbf{v}_k$. Then

$$(\mathbf{v}, \mathbf{w}) = \sum_{k=1}^{r} \alpha_k (\mathbf{v}, \mathbf{v}_k) = 0,$$

so $\mathbf{v} \perp \mathbf{w}$. □

**Definition 2.4.** A system of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is called *orthogonal* if any two vectors are orthogonal to each other (i.e. if $(\mathbf{v}_j, \mathbf{v}_k) = 0$ for $j \neq k$).

If, in addition $\|\mathbf{v}_k\| = 1$ for all $k$, we call the system *orthonormal*.

**Lemma 2.5** (Generalized Pythagorean identity)**.** *Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be an orthogonal system. Then*

$$\left\| \sum_{k=1}^{n} \alpha_k \mathbf{v}_k \right\|^2 = \sum_{k=1}^{n} |\alpha_k|^2 \|\mathbf{v}_k\|^2$$

This formula looks particularly simple for orthonormal systems, where $\|\mathbf{v}_k\| = 1$.

**Proof of the Lemma.**

$$\left\| \sum_{k=1}^{n} \alpha_k \mathbf{v}_k \right\|^2 = \Big( \sum_{k=1}^{n} \alpha_k \mathbf{v}_k, \sum_{j=1}^{n} \alpha_j \mathbf{v}_j \Big) = \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \overline{\alpha}_j (\mathbf{v}_k, \mathbf{v}_j).$$

Because of orthogonality $(\mathbf{v}_k, \mathbf{v}_j) = 0$ if $j \neq k$. Therefore we only need to sum the terms with $j = k$, which gives exactly

$$\sum_{k=1}^{n} |\alpha_k|^2 (\mathbf{v}_k, \mathbf{v}_k) = \sum_{k=1}^{n} |\alpha_k|^2 \|\mathbf{v}_k\|^2.$$

□

**Corollary 2.6.** *Any orthogonal system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ of non-zero vectors is linearly independent.*

**Proof.** Suppose for some $\alpha_1, \alpha_2, \dots, \alpha_n$ we have $\sum_{k=1}^{n} \alpha_k \mathbf{v}_k = \mathbf{0}$. Then by the Generalized Pythagorean identity (Lemma 2.5)

$$0 = \|\mathbf{0}\|^2 = \sum_{k=1}^{n} |\alpha_k|^2 \|\mathbf{v}_k\|^2.$$

Since $\|\mathbf{v}_k\| \neq \mathbf{0}$ ($\mathbf{v}_k \neq \mathbf{0}$) we conclude that

$$\alpha_k = 0 \qquad \forall k,$$

so only the trivial linear combination gives $\mathbf{0}$. □

**Remark.** In what follows we will usually mean by an orthogonal system an orthogonal system of non-zero vectors. Since the zero vector $\mathbf{0}$ is orthogonal to everything, it always can be added to any orthogonal system, but it is really not interesting to consider orthogonal systems with zero vectors.

### 2.1. Orthogonal and orthonormal bases.

**Definition 2.7.** An orthogonal (orthonormal) system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ which is also a basis is called an orthogonal (orthonormal) basis.

It is clear that in $\dim V = n$ then any orthogonal system of $n$ non-zero vectors is an orthogonal basis.

As we studied before, to find coordinates of a vector in a basis one needs to solve a linear system. However, for an orthogonal basis finding coordinates of a vector is much easier. Namely, suppose $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is an orthogonal basis, and let

$$\mathbf{x} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \ldots + \alpha_n\mathbf{v}_n = \sum_{j=1}^{n} \alpha_j\mathbf{v}_j.$$

Taking inner product of both sides of the equation with $\mathbf{v}_1$ we get

$$(\mathbf{x}, \mathbf{v}_1) = \sum_{j=1}^{n} \alpha_j(\mathbf{v}_j, \mathbf{v}_1) = \alpha_1(\mathbf{v}_1, \mathbf{v}_1) = \alpha_1\|\mathbf{v}_1\|^2$$

(all inner products $(\mathbf{v}_j, \mathbf{v}_1) = 0$ if $j \neq 1$), so

$$\alpha_1 = \frac{(\mathbf{x}, \mathbf{v}_1)}{\|\mathbf{v}_1\|^2}.$$

Similarly, multiplying both sides by $\mathbf{v}_k$ we get

$$(\mathbf{x}, \mathbf{v}_k) = \sum_{j=1}^{n} \alpha_j(\mathbf{v}_j, \mathbf{v}_k) = \alpha_k(\mathbf{v}_k, \mathbf{v}_k) = \alpha_k\|\mathbf{v}_k\|^2$$

so

$$\alpha_k = \frac{(\mathbf{x}, \mathbf{v}_k)}{\|\mathbf{v}_k\|^2}. \tag{2.1}$$

Therefore,

| to find coordinates of a vector in an orthogonal basis one does not need to solve a linear system, the coordinates are determined by the formula (2.1). |
|---|

This formula is especially simple for orthonormal bases, when $\|\mathbf{v}_k\| = 1$.

Namely, if $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthonormal basis, any vector $\mathbf{v}$ can be represented as

$$\mathbf{v} = \sum_{k=1}^{n} (\mathbf{v}, \mathbf{v}_k)\mathbf{v}_k. \tag{2.2}$$

This formula is sometimes called (a baby version of) the *abstract orthogonal Fourier decomposition*. The classical (non-abstract) Fourier decomposition deals with a concrete orthonormal system (sines and cosines or complex exponentials). We call this formula a *baby version* because the real Fourier decomposition deals with infinite orthonormal systems.

This is a very important remark allowing one to translate any statement about the standard inner product space $\mathbb{F}^n$ to an inner product space with an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$

**Remark 2.8.** The importance of orthonormal bases is that if we fix an orthonormal basis in an inner product space $V$, we can work with coordinates in this basis the same way we work with vectors in $\mathbb{F}^n$. Namely, as it was discussed in the very beginning of the book, see Remark 2.4 in Chapter 1, if we have a vector space $V$ (over a field $\mathbb{F}$) with a basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, then we can perform the standard vector operations (vector addition and multiplication by a scalar) by working with the columns of coordinates in the basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in absolutely the same way we work with vectors in the standard coordinate space $\mathbb{F}^n$.

Exercise 2.3 below shows that if we have an *orthonormal* basis in an inner product space $V$, we can compute the inner product of 2 vectors in $V$ by taking columns of their coordinates in this orthonormal basis and computing the standard inner product (in $\mathbb{C}^n$ or $\mathbb{R}^n$) of these columns.

As it will be shown below in Section 3 any finite-dimensional inner product space has an orthonormal basis. Thus, the standard inner product spaces $\mathbb{C}^n$ (or $\mathbb{R}^n$ in the case of real spaces) are essentially the only examples of a finite-dimensional inner product spaces.

**Exercises.**

**2.1.** Find all vectors in $\mathbb{R}^4$ orthogonal to vectors $(1,1,1,1)^T$ and $(1,2,3,4)^T$.

**2.2.** Let $A$ be a real $m \times n$ matrix. Describe $(\operatorname{Ran} A^T)^\perp$, $(\operatorname{Ran} A)^\perp$

**2.3.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be an orthonormal basis in $V$.

a) Prove that for any $\mathbf{x} = \sum_{k=1}^n \alpha_k \mathbf{v}_k$, $\mathbf{y} = \sum_{k=1}^n \beta_k \mathbf{v}_k$

$$(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} \alpha_k \overline{\beta}_k.$$

b) Deduce from this the *Parseval's identity*

$$(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} (\mathbf{x}, \mathbf{v}_k)\overline{(\mathbf{y}, \mathbf{v}_k)}$$

c) Assume now that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is only an orthogonal basis, not an orthonormal one. Can you write down Parseval's identity in this case?

This problem shows that if we have an orthonormal basis, we can use the coordinates in this basis absolutely the same way we use the standard coordinates in $\mathbb{C}^n$ (or $\mathbb{R}^n$).

The problem below shows that we can define an inner product by declaring a basis to be an orthonormal one.

**2.4.** Let $V$ be a vector space and let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be a basis in $V$. For $\mathbf{x} = \sum_{k=1}^n \alpha_k \mathbf{v}_k$, $\mathbf{y} = \sum_{k=1}^n \beta_k \mathbf{v}_k$ define $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{k=1}^n \alpha_k \bar{\beta}_k$.

Prove that $\langle \mathbf{x}, \mathbf{y} \rangle$ defines an inner product in $V$.

**2.5.** Let $A$ be a real $m \times n$ matrix. Describe the set of all vectors in $\mathbb{F}^m$ orthogonal to to $\operatorname{Ran} A$.

## 3. Orthogonal projection and Gram-Schmidt orthogonalization

Recalling the definition of orthogonal projection from the classical planar (2-dimensional) geometry, one can introduce the following definition. Let $E$ be a subspace of an inner product space $V$.

**Definition 3.1.** For a vector $\mathbf{v}$ its orthogonal projection $P_E\mathbf{v}$ onto the subspace $E$ is a vector $\mathbf{w}$ such that

1. $\mathbf{w} \in E$ ;
2. $\mathbf{v} - \mathbf{w} \perp E$.

We will use notation $\mathbf{w} = P_E\mathbf{v}$ for the orthogonal projection.

After introducing an object, it is natural to ask:

1. Does the object exist?
2. Is the object unique?
3. How does one find it?

We will show first that the projection is unique. Then we present a method of finding the projection, proving its existence.

The following theorem shows why the orthogonal projection is important and also proves that it is unique.

**Theorem 3.2.** *The orthogonal projection $\mathbf{w} = P_E\mathbf{v}$ minimizes the distance from $\mathbf{v}$ to $E$, i.e. for all $\mathbf{x} \in E$*

$$\|\mathbf{v} - \mathbf{w}\| \le \|\mathbf{v} - \mathbf{x}\|.$$

*Moreover, if for some $\mathbf{x} \in E$*

$$\|\mathbf{v} - \mathbf{w}\| = \|\mathbf{v} - \mathbf{x}\|,$$

*then* $\mathbf{x} = \mathbf{w}$.

**Proof.** Let $\mathbf{y} = \mathbf{w} - \mathbf{x}$. Then

$$\mathbf{v} - \mathbf{x} = \mathbf{v} - \mathbf{w} + \mathbf{w} - \mathbf{x} = \mathbf{v} - \mathbf{w} + \mathbf{y}.$$

Since $\mathbf{v} - \mathbf{w} \perp E$ we have $\mathbf{y} \perp \mathbf{v} - \mathbf{w}$ and so by Pythagorean Theorem

$$\|\mathbf{v} - \mathbf{x}\|^2 = \|\mathbf{v} - \mathbf{w}\|^2 + \|\mathbf{y}\|^2 \geq \|\mathbf{v} - \mathbf{w}\|^2.$$

Note that equality happens only if $\mathbf{y} = \mathbf{0}$ i.e. if $\mathbf{x} = \mathbf{w}$. □

The following proposition shows how to find an orthogonal projection if we know an orthogonal basis in $E$.

**Proposition 3.3.** *Let* $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ *be an orthogonal basis in* $E$. *Then the orthogonal projection* $P_E\mathbf{v}$ *of a vector* $\mathbf{v}$ *is given by the formula*

$$P_E\mathbf{v} = \sum_{k=1}^{r} \alpha_k \mathbf{v}_k, \qquad \textit{where} \qquad \alpha_k = \frac{(\mathbf{v}, \mathbf{v}_k)}{\|\mathbf{v}_k\|^2}.$$

*In other words*

$$P_E\mathbf{v} = \sum_{k=1}^{r} \frac{(\mathbf{v}, \mathbf{v}_k)}{\|\mathbf{v}_k\|^2} \mathbf{v}_k. \tag{3.1}$$

Note that the formula for $\alpha_k$ coincides with (2.1), i.e. this formula applied to an orthogonal system (not a basis) gives us a projection onto its span.

**Remark 3.4.** It is easy to see now from formula (3.1) that the orthogonal projection $P_E$ is a linear transformation.

One can also see linearity of $P_E$ directly, from the definition and uniqueness of the orthogonal projection. Indeed, it is easy to check that for any $\mathbf{x}$ and $\mathbf{y}$ the vector $\alpha\mathbf{x} + \beta\mathbf{y} - (\alpha P_E\mathbf{x} - \beta P_E\mathbf{y})$ is orthogonal to any vector in $E$, so by the definition $P_E(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha P_E\mathbf{x} + \beta P_E\mathbf{y}$.

**Remark 3.5.** Recalling the definition of inner product in $\mathbb{C}^n$ and $\mathbb{R}^n$ one can get from the above formula (3.1) the matrix of the orthogonal projection $P_E$ onto $E$ in $\mathbb{C}^n$ ($\mathbb{R}^n$) is given by

$$P_E = \sum_{k=1}^{r} \frac{1}{\|\mathbf{v}_k\|^2} \mathbf{v}_k \mathbf{v}_k^* \tag{3.2}$$

where columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ form an orthogonal basis in $E$.

**Proof of Proposition 3.3.** Let

$$\mathbf{w} := \sum_{k=1}^{r} \alpha_k \mathbf{v}_k, \qquad \text{where} \qquad \alpha_k = \frac{(\mathbf{v}, \mathbf{v}_k)}{\|\mathbf{v}_k\|^2}.$$

We want to show that $\mathbf{v}-\mathbf{w}\perp E$. By Lemma 2.3 it is sufficient to show that $\mathbf{v}-\mathbf{w}\perp\mathbf{v}_k$, $k=1,2,\dots,n$. Computing the inner product we get for $k=1,2,\dots,r$

$$\begin{aligned}(\mathbf{v}-\mathbf{w},\mathbf{v}_k)&=(\mathbf{v},\mathbf{v}_k)-(\mathbf{w},\mathbf{v}_k)=(\mathbf{v},\mathbf{v}_k)-\sum_{j=1}^{r}\alpha_j(\mathbf{v}_j,\mathbf{v}_k)\\&=(\mathbf{v},\mathbf{v}_k)-\alpha_k(\mathbf{v}_k,\mathbf{v}_k)=(\mathbf{v},\mathbf{v}_k)-\frac{(\mathbf{v},\mathbf{v}_k)}{\|\mathbf{v}_k\|^2}\|\mathbf{v}_k\|^2=0.\end{aligned}$$

□

So, if we know an orthogonal basis in $E$ we can find the orthogonal projection onto $E$. In particular, since any system consisting of one vector is an orthogonal system, we know how to perform orthogonal projection onto one-dimensional spaces.

But how do we find an orthogonal projection if we are only given a basis in $E$? Fortunately, there exists a simple algorithm allowing one to get an orthogonal basis from a basis.

**3.1. Gram-Schmidt orthogonalization algorithm.** Suppose we have a linearly independent system $\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n$. The Gram-Schmidt method constructs from this system an orthogonal system $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_n$ such that

$$\operatorname{span}\{\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n\}=\operatorname{span}\{\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_n\}.$$

Moreover, for all $r\le n$ we get

$$\operatorname{span}\{\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_r\}=\operatorname{span}\{\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_r\}$$

Now let us describe the algorithm.

**Step 1.** Put $\mathbf{v}_1:=\mathbf{x}_1$. Denote by $E_1:=\operatorname{span}\{\mathbf{x}_1\}=\operatorname{span}\{\mathbf{v}_1\}$.

**Step 2.** Define $\mathbf{v}_2$ by

$$\mathbf{v}_2=\mathbf{x}_2-P_{E_1}\mathbf{x}_2=\mathbf{x}_2-\frac{(\mathbf{x}_2,\mathbf{v}_1)}{\|\mathbf{v}_1\|^2}\mathbf{v}_1.$$

Define $E_2=\operatorname{span}\{\mathbf{v}_1,\mathbf{v}_2\}$. Note that $\operatorname{span}\{\mathbf{x}_1,\mathbf{x}_2\}=E_2$.

**Step 3.** Define $\mathbf{v}_3$ by

$$\mathbf{v}_3:=\mathbf{x}_3-P_{E_2}\mathbf{x}_3=\mathbf{x}_3-\frac{(\mathbf{x}_3,\mathbf{v}_1)}{\|\mathbf{v}_1\|^2}\mathbf{v}_1-\frac{(\mathbf{x}_3,\mathbf{v}_2)}{\|\mathbf{v}_2\|^2}\mathbf{v}_2$$

Put $E_3:=\operatorname{span}\{\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3\}$. Note that $\operatorname{span}\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}=E_3$. Note also that $\mathbf{x}_3\notin E_2$ so $\mathbf{v}_3\ne\mathbf{0}$.

...

**Step** $r+1$**.** Suppose that we already made $r$ steps of the process, constructing an orthogonal system (consisting of non-zero vectors) $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ such that $E_r := \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$. Define

$$\mathbf{v}_{r+1} := \mathbf{x}_{r+1} - P_{E_r}\mathbf{x}_{r+1} = \mathbf{x}_{r+1} - \sum_{k=1}^{r} \frac{(\mathbf{x}_{r+1}, \mathbf{v}_k)}{\|\mathbf{v}_k\|^2}\mathbf{v}_k$$

Note,that $\mathbf{x}_{r+1} \notin E_r$ so $\mathbf{v}_{r+1} \neq \mathbf{0}$.

...

Continuing this algorithm we get an orthogonal system $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.

**3.2. An example.** Suppose we are given vectors

$$\mathbf{x}_1 = (1,1,1)^T, \qquad \mathbf{x}_2 = (0,1,2)^T, \qquad \mathbf{x}_3 = (1,0,2)^T,$$

and we want to orthogonalize it by Gram-Schmidt. On the first step define

$$\mathbf{v}_1 = \mathbf{x}_1 = (1,1,1)^T.$$

On the second step we get

$$\mathbf{v}_2 = \mathbf{x}_2 - P_{E_1}\mathbf{x}_2 = \mathbf{x}_2 - \frac{(\mathbf{x}_2, \mathbf{v}_1)}{\|\mathbf{v}_1\|^2}\mathbf{v}_1.$$

Computing

$$(\mathbf{x}_2, \mathbf{v}_1) = \left(\begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right) = 3, \quad \|\mathbf{v}_1\|^2 = 3,$$

we get

$$\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} - \frac{3}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

Finally, define

$$\mathbf{v}_3 = \mathbf{x}_3 - P_{E_2}\mathbf{x}_3 = \mathbf{x}_3 - \frac{(\mathbf{x}_3, \mathbf{v}_1)}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 - \frac{(\mathbf{x}_3, \mathbf{v}_2)}{\|\mathbf{v}_2\|^2}\mathbf{v}_2.$$

Computing

$$\left(\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right) = 3, \quad \left(\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\right) = 1, \ \|\mathbf{v}_1\|^2 = 3, \ \|\mathbf{v}_2\|^2 = 2$$

($\|\mathbf{v}_1\|^2$ was already computed before) we get

$$\mathbf{v}_3 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} - \frac{3}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{pmatrix}$$

**Remark.** Since the multiplication by a scalar does not change the orthogonality, one can multiply vectors $\mathbf{v}_k$ obtained by Gram-Schmidt by any non-zero numbers.

In particular, in many theoretical constructions one *normalizes* vectors $\mathbf{v}_k$ by dividing them by their respective norms $\|\mathbf{v}_k\|$. Then the resulting system will be orthonormal, and the formulas will look simpler.

On the other hand, when performing the computations one may want to avoid fractional entries by multiplying a vector by the least common denominator of its entries. Thus one may want to replace the vector $\mathbf{v}_3$ from the above example by $(1, -2, 1)^T$.

### 3.3. Orthogonal complement. Decomposition $V = E \oplus E^\perp$.

**Definition.** For a subspace $E$ its *orthogonal complement* $E^\perp$ is the set of all vectors orthogonal to $E$,

$$E^\perp := \{\mathbf{x} : \mathbf{x} \perp E\}.$$

If $\mathbf{x}, \mathbf{y} \perp E$ then for any linear combination $\alpha\mathbf{x} + \beta\mathbf{y} \perp E$ (can you see why?). Therefore $E^\perp$ is a subspace.

By the definition of orthogonal projection any vector in an inner product space $V$ admits a unique representation

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2, \qquad \mathbf{v}_1 \in E,\ \mathbf{v}_2 \perp E \text{ (eqv. } \mathbf{v}_2 \in E^\perp)$$

(where clearly $\mathbf{v}_1 = P_E\mathbf{v}$).

This statement is often symbolically written as $V = E \oplus E^\perp$, which mean exactly that any vector admits the unique decomposition above.

The following proposition gives an important property of the orthogonal complement.

**Proposition 3.6.** *For a subspace $E$*

$$(E^\perp)^\perp = E.$$

The proof is left as an exercise, see Exercise 3.12 below.

**Exercises.**

**3.1.** Apply Gram–Schmidt orthogonalization to the system of vectors $(1, 2, -2)^T$, $(1, -1, 4)^T$, $(2, 1, 1)^T$.

**3.2.** Apply Gram–Schmidt orthogonalization to the system of vectors $(1, 2, 3)^T$, $(1, 3, 1)^T$. Write the matrix of the orthogonal projection onto 2-dimensional subspace spanned by these vectors.

**3.3.** Complete an orthogonal system obtained in the previous problem to an orthogonal basis in $\mathbb{R}^3$, i.e. add to the system some vectors (how many?) to get an orthogonal basis.

Can you describe how to complete an orthogonal system to an orthogonal basis in general situation of $\mathbb{R}^n$ or $\mathbb{C}^n$?

**3.4.** Find the distance from a vector $(2,3,1)^T$ to the subspace spanned by the vectors $(1,2,3)^T$, $(1,3,1)^T$. Note, that I am only asking to find the distance to the subspace, not the orthogonal projection.

**3.5.** Find the orthogonal projection of a vector $(1,1,1,1)^T$ onto the subspace spanned by the vectors $\mathbf{v}_1 = (1,3,1,1)^T$ and $\mathbf{v}_2 = (2,-1,1,0)^T$ (note that $\mathbf{v}_1 \perp \mathbf{v}_2$).

**3.6.** Find the distance from a vector $(1,2,3,4)$ to the subspace spanned by the vectors $\mathbf{v}_1 = (1,-1,1,0)^T$ and $\mathbf{v}_2 = (1,2,1,1)^T$ (note that $\mathbf{v}_1 \perp \mathbf{v}_2$). Can you find the distance without actually computing the projection? That would simplify the calculations.

**3.7.** True or false: if $E$ is a subspace of $V$, then $\dim E + \dim(E^\perp) = \dim V$? Justify.

**3.8.** Let $P$ be the orthogonal projection onto a subspace $E$ of an inner product space $V$, $\dim V = n$, $\dim E = r$. Find the eigenvalues and the eigenvectors (eigenspaces). Find the algebraic and geometric multiplicities of each eigenvalue.

**3.9.** (Using eigenvalues to compute determinants).

a) Find the matrix of the orthogonal projection onto the one-dimensional subspace in $\mathbb{R}^n$ spanned by the vector $(1,1,\dots,1)^T$;

b) Let $A$ be the $n \times n$ matrix with all entries equal 1. Compute its eigenvalues and their multiplicities (use the previous problem);

c) Compute eigenvalues (and multiplicities) of the matrix $A - I$, i.e. of the matrix with zeroes on the main diagonal and ones everywhere else;

d) Compute $\det(A - I)$.

**3.10** (Legendre's polynomials:)**.** Let an inner product on the space of polynomials be defined by $(f,g) = \int_{-1}^{1} f(t)\overline{g(t)}dt$. Apply Gram-Schmidt orthogonalization to the system $1, t, t^2, t^3$.

Legendre's polynomials are particular case of the so-called orthogonal polynomials, which play an important role in many branches of mathematics.

**3.11.** Let $P = P_E$ be the matrix of an orthogonal projection onto a subspace $E$. Show that

a) The matrix $P$ is *self-adjoint*, meaning that $P^* = P$.

b) $P^2 = P$.

**Remark:** The above 2 properties completely characterize orthogonal projection, i.e. any matrix $P$ satisfying these properties is the matrix of some orthogonal projection. We will discuss this some time later.

**3.12.** Show that for a subspace $E$ we have $(E^\perp)^\perp = E$. **Hint:** It is easy to see that $E$ is orthogonal to $E^\perp$ (why?). To show that *any* vector $\mathbf{x}$ orthogonal to $E^\perp$ belongs to $E$ use the decomposition $V = E \oplus E^\perp$ from Section 3.3 above.

**3.13.** Suppose $P$ is the orthogonal projection onto a subspace $E$, and $Q$ is the orthogonal projection onto the orthogonal complement $E^\perp$.

a) What are $P + Q$ and $PQ$?

b) Show that $P - Q$ is its own inverse.

## 4. Least square solution. Formula for the orthogonal projection

As it was discussed before in Chapter 2, the equation

$$A\mathbf{x} = \mathbf{b}$$

has a solution if and only if $\mathbf{b} \in \operatorname{Ran} A$. But what do we do to solve an equation that does not have a solution?

This seems to be a silly question, because if there is no solution, then there is no solution. But, situations when we want to solve an equation that does not have a solution can appear naturally, for example, if we obtained the equation from an experiment. If we do not have any errors, the right side $\mathbf{b}$ belongs to the column space $\operatorname{Ran} A$, and equation is consistent. But, in real life it is impossible to avoid errors in measurements, so it is possible that an equation that in theory should be consistent, does not have a solution. So, what can one do in this situation?

**4.1. Least square solution.** The simplest idea is to write down the error

$$\|A\mathbf{x} - \mathbf{b}\|$$

and try to find $\mathbf{x}$ minimizing it. If we can find $\mathbf{x}$ such that the error is 0, the system is consistent and we have exact solution. Otherwise, we get the so-called *least square* solution.

The term *least square* arises from the fact that minimizing $\|A\mathbf{x} - \mathbf{b}\|$ is equivalent to minimizing

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_{k=1}^{m} |(A\mathbf{x})_k - b_k|^2 = \sum_{k=1}^{m} \Big| \sum_{j=1}^{n} A_{k,j} x_j - b_k \Big|^2$$

i.e. to minimizing the sum of squares of linear functions.

There are several ways to find the least square solution. If we are in $\mathbb{R}^n$, and everything is real, we can forget about absolute values. Then we can just take partial derivatives with respect to $x_j$ and find the where all of them are 0, which gives us the minimum.

4.1.1. *Geometric approach.* However, there is a simpler way of finding the minimum. Namely, if we take all possible vectors $\mathbf{x}$, then $A\mathbf{x}$ gives us all possible vectors in $\operatorname{Ran} A$, so minimum of $\|A\mathbf{x} - \mathbf{b}\|$ is exactly the distance from $\mathbf{b}$ to $\operatorname{Ran} A$. Therefore the value of $\|A\mathbf{x} - \mathbf{b}\|$ is minimal if and only if $A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}$, where $P_{\operatorname{Ran} A}$ stands for the orthogonal projection onto the column space $\operatorname{Ran} A$.

So, to find the least square solution we simply need to solve the equation

$$A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}.$$

If we know an orthogonal basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in $\operatorname{Ran} A$, we can find vector $P_{\operatorname{Ran} A}\mathbf{b}$ by the formula

$$P_{\operatorname{Ran} A}\mathbf{b} = \sum_{k=1}^{n} \frac{(\mathbf{b}, \mathbf{v}_k)}{\|\mathbf{v}_k\|^2}\mathbf{v}_k.$$

If we only know a basis in $\operatorname{Ran} A$, we need to use the Gram–Schmidt orthogonalization to obtain an orthogonal basis from it.

So, theoretically, the problem is solved, but the solution is not very simple: it involves Gram–Schmidt orthogonalization, which can be computationally intensive. Fortunately, there exists a simpler solution.

4.1.2. *Normal equation.* Namely, $A\mathbf{x}$ is the orthogonal projection $P_{\operatorname{Ran} A}\mathbf{b}$ if and only if $\mathbf{b} - A\mathbf{x} \perp \operatorname{Ran} A$ ($A\mathbf{x} \in \operatorname{Ran} A$ for all $\mathbf{x}$).

If $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ are columns of $A$, then the condition $A\mathbf{x} \perp \operatorname{Ran} A$ can be rewritten as

$$\mathbf{b} - A\mathbf{x} \perp \mathbf{a}_k, \qquad \forall k = 1, 2, \dots, n.$$

That means

$$0 = (\mathbf{b} - A\mathbf{x}, \mathbf{a}_k) = \mathbf{a}_k^*(\mathbf{b} - A\mathbf{x}) \qquad \forall k = 1, 2, \dots, n.$$

Joining rows $\mathbf{a}_k^*$ together we get that these equations are equivalent to

$$A^*(\mathbf{b} - A\mathbf{x}) = \mathbf{0},$$

which in turn is equivalent to the so-called *normal equation*

$$A^*A\mathbf{x} = A^*\mathbf{b}.$$

A solution of this equation gives us the least square solution of $A\mathbf{x} = \mathbf{b}$.

Note, that the least square solution is unique if and only if $A^*A$ is invertible.

**4.2. Formula for the orthogonal projection.** As we already discussed above, if $\mathbf{x}$ is a solution of the *normal equation* $A^*A\mathbf{x} = A^*\mathbf{b}$ (i.e. a least square solution of $A\mathbf{x} = \mathbf{b}$), then $A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}$. So, to find the orthogonal projection of $\mathbf{b}$ onto the column space $\operatorname{Ran} A$ we need to solve the normal equation $A^*A\mathbf{x} = A^*\mathbf{b}$, and then multiply the solution by $A$.

If the operator $A^*A$ is invertible, the solution of the normal equation $A^*A\mathbf{x} = A^*\mathbf{b}$ is given by $\mathbf{x} = (A^*A)^{-1}A^*\mathbf{b}$, so the orthogonal projection $P_{\operatorname{Ran} A}\mathbf{b}$ can be computed as

$$P_{\operatorname{Ran} A}\mathbf{b} = A(A^*A)^{-1}A^*\mathbf{b}.$$

Since this is true for all $\mathbf{b}$,

$$P_{\operatorname{Ran} A} = A(A^*A)^{-1}A^*$$

is the formula for the matrix of the orthogonal projection onto $\operatorname{Ran} A$.

The following theorem implies that for an $m \times n$ matrix $A$ the matrix $A^*A$ is invertible if and only if $\operatorname{rank} A = n$.

**Theorem 4.1.** *For an $m \times n$ matrix $A$*

$$\operatorname{Ker} A = \operatorname{Ker}(A^*A).$$

Indeed, according to the rank theorem $\operatorname{Ker} A = \{\mathbf{0}\}$ if and only if rank $A$ is $n$. Therefore $\operatorname{Ker}(A^*A) = \{\mathbf{0}\}$ if and only if $\operatorname{rank} A = n$. Since the matrix $A^*A$ is square, it is invertible if and only if $\operatorname{rank} A = n$.

We leave the proof of the theorem as an exercise. To prove the equality $\operatorname{Ker} A = \operatorname{Ker}(A^*A)$ one needs to prove two inclusions $\operatorname{Ker}(A^*A) \subset \operatorname{Ker} A$ and $\operatorname{Ker} A \subset \operatorname{Ker}(A^*A)$. One of the inclusions is trivial, for the other one use the fact that

$$\|A\mathbf{x}\|^2 = (A\mathbf{x}, A\mathbf{x}) = (A^*A\mathbf{x}, \mathbf{x}).$$

**4.3. An example: line fitting.** Let us introduce a few examples where the least square solution appears naturally. Suppose that we know that two quantities $x$ and $y$ are related by the law $y = a + bx$. The coefficients $a$ and $b$ are unknown, and we would like to find them from experimental data.

Suppose we run the experiment $n$ times, and we get $n$ pairs $(x_k, y_k)$, $k = 1, 2, \dots, n$. Ideally, all the points $(x_k, y_k)$ should be on a straight line, but because of errors in measurements, it usually does not happen: the point are usually close to some line, but not exactly on it. That is where the least square solution helps!

Ideally, the coefficients $a$ and $b$ should satisfy the equations

$$a + bx_k = y_k, \qquad k = 1, 2, \dots, n$$

(note that here, $x_k$ and $y_k$ are some fixed numbers, and the unknowns are $a$ and $b$). If it is possible to find such $a$ and $b$ we are lucky. If not, the standard thing to do is to minimize the total quadratic error

$$\sum_{k=1}^{n} |a + bx_k - y_k|^2.$$

But, minimizing this error is exactly finding the least square solution of the system

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

(recall that $x_k$ $y_k$ are some given numbers, and the unknowns are $a$ and $b$).

4.3.1. *An example.* Suppose our data $(x_k, y_k)$ consist of pairs

$$(-2,4),\ (-1,2),\ (0,1),\ (2,1),\ (3,1).$$

Then we need to find the least square solution of

$$\begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Then

$$A^*A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ 2 & 18 \end{pmatrix}$$

and

$$A^*\mathbf{b} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 2 & 3 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ -5 \end{pmatrix}$$

so the normal equation $A^*A\mathbf{x} = A^*\mathbf{b}$ is rewritten as

$$\begin{pmatrix} 5 & 2 \\ 2 & 18 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 9 \\ -5 \end{pmatrix}.$$

The solution of this equation is

$$a = 2, b = -1/2,$$

so the best fitting straight line is

$$y = 2 - 1/2x.$$

**4.4. Other examples: curves and planes.** The least square method is not limited to the line fitting. It can also be applied to more general curves, as well as to surfaces in higher dimensions.

The only constraint here is that the parameters we want to find be involved linearly. The general algorithm is as follows:

1. Find the equations that your data should satisfy if there is exact fit;
2. Write these equations as a linear system, where unknowns are the parameters you want to find. Note, that the system need not to be consistent (and usually is not);
3. Find the least square solution of the system.

4.4.1. *An example: curve fitting.* For example, suppose we know that the relation between $x$ and $y$ is given by the quadratic law $y = a + bx + cx^2$, so we want to fit a parabola $y = a + bx + cx^2$ to the data. Then our unknowns $a$, $b$, $c$ should satisfy the equations

$$a + bx_k + cx_k^2 = y_k, \qquad k = 1, 2, \ldots, n$$

or, in matrix form

$$\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

For example, for the data from the previous example we need to find the least square solution of

$$\begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Then

$$A^*A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 2 & 3 \\ 4 & 1 & 0 & 4 & 9 \end{pmatrix} \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 18 \\ 2 & 18 & 26 \\ 18 & 26 & 114 \end{pmatrix}$$

and

$$A^*\mathbf{b} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 2 & 3 \\ 4 & 1 & 0 & 4 & 9 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ -5 \\ 31 \end{pmatrix}.$$

Therefore the normal equation $A^*A\mathbf{x} = A^*\mathbf{b}$ is

$$\begin{pmatrix} 5 & 2 & 18 \\ 2 & 18 & 26 \\ 18 & 26 & 114 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 9 \\ -5 \\ 31 \end{pmatrix}$$

which has the unique solution

$$a = 86/77, \qquad b = -62/77, \qquad c = 43/154.$$

Therefore,

$$y = 86/77 - 62x/77 + 43x^2/154$$

is the best fitting parabola.

4.4.2. *Plane fitting.* As another example, let us fit a plane $z = a + bx + cy$ to the data

$$(x_k, y_k, z_k) \in \mathbb{R}^3, \qquad k = 1, 2, \dots n.$$

The equations we should have in the case of exact fit are

$$a + bx_k + cy_k = z_k, \qquad k = 1, 2, \dots, n,$$

or, in the matrix form

$$\begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

So, to find the best fitting plane, we need to find the best square solution of this system (the unknowns are $a$, $b$, $c$).

**Exercises.**

**4.1.** Find the least square solution of the system

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

**4.2.** Find the matrix of the orthogonal projection $P$ onto the column space of

$$\begin{pmatrix} 1 & 1 \\ 2 & -1 \\ -2 & 4 \end{pmatrix}.$$

Use two methods: Gram–Schmidt orthogonalization and formula for the projection.

Compare the results.

**4.3.** Find the best straight line fit (least square solution) to the points $(-2, 4)$, $(-1, 3)$, $(0, 1)$, $(2, 0)$.

**4.4.** Fit a plane $z = a + bx + cy$ to four points $(1, 1, 3)$, $(0, 3, 6)$, $(2, 1, 5)$, $(0, 0, 0)$.

To do that

a) Find 4 equations with 3 unknowns $a, b, c$ such that the plane pass through all 4 points (this system does not have to have a solution);

b) Find the least square solution of the system.

**4.5.** *Minimal norm solution.* let an equation $A\mathbf{x} = \mathbf{b}$ has a solution, and let $A$ has non-trivial kernel (so the solution is not unique). Prove that

a) There exist a unique solution $\mathbf{x}_0$ of $A\mathbf{x} = \mathbf{b}$ minimizing the norm $\|\mathbf{x}\|$, i.e. that there exists unique $\mathbf{x}_0$ such that $A\mathbf{x}_0 = \mathbf{b}$ and $\|\mathbf{x}_0\| \le \|\mathbf{x}\|$ for any $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{b}$.

b) $\mathbf{x}_0 = P_{(\operatorname{Ker} A)^\perp} \mathbf{x}$ for any $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{b}$.

**4.6.** *Minimal norm least square solution.* Applying previous problem to the equation $A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}$ show that

a) There exists a unique least square solution $\mathbf{x}_0$ of $A\mathbf{x} = \mathbf{b}$ minimizing the norm $\|\mathbf{x}\|$.

b) $\mathbf{x}_0 = P_{(\operatorname{Ker} A)^\perp}\mathbf{x}$ for any least square solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$.

## 5. Adjoint of a linear transformation. Fundamental subspaces revisited.

**5.1. Adjoint matrices and adjoint operators.** Let as recall that for an $m \times n$ matrix $A$ its *Hermitian adjoint* (or simply *adjoint*) $A^*$ is defined by $A^* := \overline{A^T}$. In other words, the matrix $A^*$ is obtained from the transposed matrix $A^T$ by taking complex conjugate of each entry.

The following identity is the main property of adjoint matrix:

$$\boxed{(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A^*\mathbf{y}) \qquad \forall \mathbf{x} \in \mathbb{C}^n,\ \forall \mathbf{y} \in \mathbb{C}^m.}$$

Before proving this identity, let us introduce some useful formulas. Let us recall that for transposed matrices we have the identity $(AB)^T = B^T A^T$. Since for complex numbers $z$ and $w$ we have $\overline{zw} = \overline{z}\,\overline{w}$, the identity

$$(AB)^* = B^* A^*$$

holds for the adjoint.

Also, since $(A^T)^T = A$ and $\overline{\overline{z}} = z$,

$$(A^*)^* = A.$$

Now, we are ready to prove the main identity:

$$(A\mathbf{x}, \mathbf{y}) = \mathbf{y}^* A\mathbf{x} = (A^*\mathbf{y})^*\mathbf{x} = (\mathbf{x}, A^*\mathbf{y});$$

the first and the last equalities here follow from the definition of inner product in $\mathbb{F}^n$, and the middle one follows from the fact that

$$(A^*\mathbf{x})^* = \mathbf{x}^*(A^*)^* = \mathbf{x}^* A.$$

5.1.1. *Uniqueness of the adjoint.* The above main identity $(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A^*\mathbf{y})$ is often used as the definition of the adjoint operator. Let us first notice that the adjoint operator is unique: if a matrix $B$ satisfies

$$(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, B\mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y},$$

then $B = A^*$. Indeed, by the definition of $A^*$ for a given $\mathbf{y}$ we have

$$(\mathbf{x}, A^*\mathbf{y}) = (\mathbf{x}, B\mathbf{y}) \qquad \forall \mathbf{x},$$

and therefore by Corollary 1.5 $A^*\mathbf{y} = B\mathbf{y}$. Since it is true for all $\mathbf{y}$, the linear transformations, and therefore the matrices $A^*$ and $B$ coincide.

5.1.2. *Adjoint transformation in abstract setting.* The above main identity $(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A^*\mathbf{y})$ can be used to define the adjoint operator in abstract setting, where $A : V \to W$ is an operator acting from one inner product space to another. Namely, we define $A^* : W \to V$ to be the operator satisfying

$$(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A^*\mathbf{y}) \qquad \forall \mathbf{x} \in V,\ \forall \mathbf{y} \in W.$$

Why does such an operator exists? We can simply construct it: consider orthonormal bases $\mathcal{A} = \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in $V$ and $\mathcal{B} = \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$ in $W$. If $[A]_{\mathcal{B}\mathcal{A}}$ is the matrix of $A$ with respect to these bases, we define the operator $A^*$ by defining its matrix $[A^*]_{\mathcal{A}\mathcal{B}}$ as

$$[A^*]_{\mathcal{A}\mathcal{B}} = ([A]_{\mathcal{B}\mathcal{A}})^*.$$

We leave the proof that this indeed gives the adjoint operator as an exercise for the reader.

Note, that the reasoning in the above Sect. 5.1.1 implies that the adjoint operator is unique.

5.1.3. *Useful formulas.* Below we present the properties of the adjoint operators (matrices) we will use a lot. We leave the proofs as an exercise for the reader.

1. $(A + B)^* = A^* + B^*$;
2. $(\alpha A)^* = \overline{\alpha} A^*$;
3. $(AB)^* = B^* A^*$;
4. $(A^*)^* = A$;
5. $(\mathbf{y}, A\mathbf{x}) = (A^*\mathbf{y}, \mathbf{x})$.

## 5.2. Relation between fundamental subspaces.

**Theorem 5.1.** *Let $A : V \to W$ be an operator acting from one inner product space to another. Then*

1. $\operatorname{Ker} A^* = (\operatorname{Ran} A)^\perp$;
2. $\operatorname{Ker} A = (\operatorname{Ran} A^*)^\perp$;
3. $\operatorname{Ran} A = (\operatorname{Ker} A^*)^\perp$;
4. $\operatorname{Ran} A^* = (\operatorname{Ker} A)^\perp$.

**Remark.** Earlier in Section 7 of Chapter 2 the fundamental subspaces were defined (as it is often done in the literature) using $A^T$ instead of $A^*$. Of course, there is no difference for real matrices, so in the real case the above theorem gives the geometric description of the fundamentals subspaces defined there.

Geometric interpretation of the fundamental subspaces defined using $A^T$ is presented in Chapter 8 below, see Section 3 there (Theorem 3.7). The

formulas in this theorem are essentially the same as in Theorem 5.1 here, only the interpretation is a bit different.

**Proof of Theorem 5.1.** First of all, let us notice, that since for a subspace $E$ we have $(E^\perp)^\perp = E$, the statements 1 and 3 are equivalent. Similarly, for the same reason, the statements 2 and 4 are equivalent as well. Finally, statement 2 is exactly statement 1 applied to the operator $A^*$ (here we use the fact that $(A^*)^* = A$).

So, to prove the theorem we only need to prove statement 1.

We will present 2 proofs of this statement: a "matrix" proof, and an "invariant", or "coordinate-free" one.

In the "matrix" proof, we assume that $A$ is an $m \times n$ matrix, i.e. that $A : \mathbb{F}^n \to \mathbb{F}^m$. The general case can be always reduced to this one by picking orthonormal bases in $V$ and $W$, and considering the matrix of $A$ in this bases.

Let $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ be the columns of $A$. Note, that $\mathbf{x} \in (\operatorname{Ran} A)^\perp$ if and only if $\mathbf{x} \perp \mathbf{a}_k$ (i.e. $(\mathbf{x}, \mathbf{a}_k) = 0$) $\forall k = 1, 2, \dots, n$.

By the definition of the inner product in $\mathbb{F}^n$, that means

$$0 = (\mathbf{x}, \mathbf{a}_k) = \mathbf{a}_k^* \mathbf{x} \qquad \forall k = 1, 2, \dots, n.$$

Since $\mathbf{a}_k^*$ is the row number $k$ of $A^*$, the above $n$ equalities are equivalent to the equation

$$A^*\mathbf{x} = \mathbf{0}.$$

So, we proved that $\mathbf{x} \in (\operatorname{Ran} A)^\perp$ if and only if $A^*\mathbf{x} = \mathbf{0}$, and that is exactly the statement 1.

Now, let us present the "coordinate-free" proof. The inclusion $\mathbf{x} \in (\operatorname{Ran} A)^\perp$ means that $\mathbf{x}$ is orthogonal to all vectors of the form $A\mathbf{y}$, i.e. that

$$(\mathbf{x}, A\mathbf{y}) = 0 \qquad \forall \mathbf{y}.$$

Since $(\mathbf{x}, A\mathbf{y}) = (A^*\mathbf{x}, \mathbf{y})$, this identity is equivalent to

$$(A^*\mathbf{x}, \mathbf{y}) = 0 \qquad \forall \mathbf{y},$$

and by Lemma 1.4 this happens if and only if $A^*\mathbf{x} = \mathbf{0}$. So we proved that $x \in (\operatorname{Ran} A)^\perp$ if and only if $A^*\mathbf{x} = \mathbf{0}$, which is exactly the statement 1 of the theorem. □

**5.3. The "essential" part of a linear transformation.** The above theorem makes the structure of the operator $A$ and the geometry of fundamental subspaces much more transparent. It follows from this theorem that the operator $A$ can be represented as a composition of orthogonal projection onto $\operatorname{Ran} A^*$ and an isomorphism from $\operatorname{Ran} A^*$ to $\operatorname{Ran} A$.

Indeed, let $\widetilde{A} : \operatorname{Ran} A^* \to \operatorname{Ran} A$ be the restriction of $A$ to the domain $\operatorname{Ran} A^*$ and the target space $\operatorname{Ran} A$,

$$\widetilde{A}\mathbf{x} = A\mathbf{x}, \qquad \forall \mathbf{x} \in \operatorname{Ran} A^*.$$

Since $\operatorname{Ker} A = (\operatorname{Ran} A^*)^\perp$, we have

$$Ax = AP_{\operatorname{Ran} A^* x} = \widetilde{A}P_{\operatorname{Ran} A^*} \qquad \forall \mathbf{x} \in X;$$

the fact that $\mathbf{x} - P_{\operatorname{Ran} A^*}\mathbf{x} \in (\operatorname{Ran} A^*)^\perp = \operatorname{Ker} A$ is used here. Therefore we can write

$$A\mathbf{x} = \widetilde{A}P_{\operatorname{Ran} A^*}\mathbf{x} \qquad \forall \mathbf{x} \in X, \tag{5.1}$$

or, equivalently, $A = \widetilde{A}P_{\operatorname{Ran} A^*}$.

Note also that $\widetilde{A} : \operatorname{Ran} A^* \to \operatorname{Ran} A$ is an invertible transformation. First we notice that $\operatorname{Ker} \widetilde{A} = \{\mathbf{0}\}$: if $\mathbf{x} \in \operatorname{Ran} A^*$ is such that $\widetilde{A}\mathbf{x} = A\mathbf{x} = \mathbf{0}$, then $\mathbf{x} \in \operatorname{Ker} A = (\operatorname{Ran} A^*)^\perp$, so $\mathbf{x} \in \operatorname{Ran} A^* \cap (\operatorname{Ran} A^*)^\perp$, thus $\mathbf{x} = \mathbf{0}$. Then to see that $\widetilde{A}$ is invertible, it is sufficient to she that $\widetilde{A}$ is onto (surjective). But this immediately follows from (5.1):

$$\operatorname{Ran} \widetilde{A} = \widetilde{A} \operatorname{Ran} A^* = AP_{\operatorname{Ran} A^*} X = AX = \operatorname{Ran} A.$$

The isomorphism $\widetilde{A}$ is sometimes called the "essential part" of the operator $A$ (a non-standard terminology).

The fact the "essential part" $\widetilde{A} : \operatorname{Ran} A^* \to \operatorname{Ran} A$ of $A$ is an isomorphism implies the following "complex" rank theorem: $\operatorname{rank} A = \operatorname{rank} A^*$. But, of course, this theorem also follows from an elementary observation that complex conjugation does not change rank of a matrix, $\operatorname{rank} A = \operatorname{rank} \overline{A}$.

**Exercises.**

**5.1.** Show that for a square matrix $A$ the equality $\det(A^*) = \overline{\det(A)}$ holds.

**5.2.** Find matrices of orthogonal projections onto all 4 fundamental subspaces of the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 2 & 4 & 3 \end{pmatrix}.$$

Note, that really you need only to compute 2 of the projections. If you pick an appropriate 2, the other 2 are easy to obtain from them (recall, how the projections onto $E$ and $E^\perp$ are related).

**5.3.** Let $A$ be an $m \times n$ matrix. Show that $\operatorname{Ker} A = \operatorname{Ker}(A^*A)$.

To do that you need to prove 2 inclusions, $\operatorname{Ker}(A^*A) \subset \operatorname{Ker} A$ and $\operatorname{Ker} A \subset \operatorname{Ker}(A^*A)$. One of the inclusions is trivial, for the other one use the fact that

$$\|A\mathbf{x}\|^2 = (A\mathbf{x}, A\mathbf{x}) = (A^*A\mathbf{x}, \mathbf{x}).$$

**5.4.** Use the equality $\operatorname{Ker} A = \operatorname{Ker}(A^*A)$ to prove that

a) $\operatorname{rank} A = \operatorname{rank}(A^*A)$;

b) If $A\mathbf{x} = \mathbf{0}$ has only the trivial solution, $A$ is left invertible. (You can just write a formula for a left inverse).

**5.5.** Suppose, that for a matrix $A$ the matrix $A^*A$ is invertible, so the orthogonal projection onto $\operatorname{Ran} A$ is given by the formula $A(A^*A)^{-1}A^*$. Can you write formulas for the orthogonal projections onto the other 3 fundamental subspaces ($\operatorname{Ker} A$, $\operatorname{Ker} A^*$, $\operatorname{Ran} A^*$)?

**5.6.** Let a matrix $P$ be self-adjoint ($P^* = P$) and let $P^2 = P$. Show that $P$ is the matrix of an orthogonal projection. **Hint:** consider the decomposition $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, $\mathbf{x}_1 \in \operatorname{Ran} P$, $\mathbf{x}_2 \perp \operatorname{Ran} P$ and show that $P\mathbf{x}_1 = \mathbf{x}_1$, $P\mathbf{x}_2 = \mathbf{0}$. For one of the equalities you will need self-adjointness, for the other one the property $P^2 = P$.

## 6. Isometries and unitary operators. Unitary and orthogonal matrices.

### 6.1. Main definitions.

**Definition.** An operator $U : X \to Y$ is called an *isometry*, if it preserves the norm,

$$\|U\mathbf{x}\| = \|\mathbf{x}\| \qquad \forall \mathbf{x} \in X.$$

The following theorem shows that an isometry preserves the inner product

**Theorem 6.1.** *An operator $U : X \to Y$ is an isometry if and only if it preserves the inner product, i.e if and only if*

$$(\mathbf{x}, \mathbf{y}) = (U\mathbf{x}, U\mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y} \in X.$$

**Proof.** The proof uses the polarization identities (Lemma 1.9). For example, if $X$ is a complex space

$$\begin{aligned}(U\mathbf{x}, U\mathbf{y}) &= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|U\mathbf{x} + \alpha U\mathbf{y}\|^2 \\ &= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|U(\mathbf{x} + \alpha\mathbf{y})\|^2 \\ &= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|\mathbf{x} + \alpha\mathbf{y}\|^2 = (\mathbf{x}, \mathbf{y}).\end{aligned}$$

Similarly, for a real space $X$

$$\begin{aligned}(U\mathbf{x}, U\mathbf{y}) &= \frac{1}{4}\left(\|U\mathbf{x}+U\mathbf{y}\|^2 - \|U\mathbf{x}-U\mathbf{y}\|^2\right)\\ &= \frac{1}{4}\left(\|U(\mathbf{x}+\mathbf{y})\|^2 - \|U(\mathbf{x}-\mathbf{y})\|^2\right)\\ &= \frac{1}{4}\left(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\right) = (\mathbf{x},\mathbf{y}).\end{aligned}$$

□

**Lemma 6.2.** *An operator $U : X \to Y$ is an isometry if and only if $U^*U = I$.*

**Proof.** If $U^*U = I$, then by the definition of adjoint operator

$$(\mathbf{x},\mathbf{x}) = (U^*U\mathbf{x},\mathbf{x}) = (U\mathbf{x},U\mathbf{x}) \qquad \forall \mathbf{x} \in X.$$

Therefore $\|\mathbf{x}\| = \|U\mathbf{x}\|$, and so $U$ is an isometry.

On the other hand, if $U$ is an isometry, then by the definition of adjoint operator and by Theorem 6.1 we have for all $\mathbf{x} \in X$

$$(U^*U\mathbf{x},\mathbf{y}) = (U\mathbf{x},U\mathbf{y}) = (\mathbf{x},\mathbf{y}) \qquad \forall \mathbf{y} \in X,$$

and therefore by Corollary 1.5 $U^*U\mathbf{x} = \mathbf{x}$. Since it is true for all $\mathbf{x} \in X$, we have $U^*U = I$. □

The above lemma implies that an isometry is always left invertible ($U^*$ being a left inverse).

**Definition.** An isometry $U : X \to Y$ is called a *unitary* operator if it is invertible.

**Proposition 6.3.** *An isometry $U : X \to Y$ is a unitary operator if and only if* $\dim X = \dim Y$.

**Proof.** Since $U$ is an isometry, it is left invertible, and since $\dim X = \dim Y$, it is invertible (a left invertible square matrix is invertible).

On the other hand, if $U : X \to Y$ is invertible, $\dim X = \dim Y$ (only square matrices are invertible, isomorphic spaces have equal dimensions). □

A square matrix $U$ is called *unitary* if $U^*U = I$, i.e. a unitary matrix is a matrix of a unitary operator acting in $\mathbb{F}^n$.

A unitary matrix with real entries is called an *orthogonal* matrix. An orthogonal matrix can be interpreted a matrix of a unitary operator acting in the real space $\mathbb{R}^n$.

Few properties of unitary operators:

1. For a unitary transformation $U$, $U^{-1} = U^*$;

2. If $U$ is unitary, $U^* = U^{-1}$ is also unitary;
3. If $U$ is a isometry, and $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthonormal basis, then $U\mathbf{v}_1, U\mathbf{v}_2, \dots, U\mathbf{v}_n$ is an orthonormal system. Moreover, if $U$ is unitary, $U\mathbf{v}_1, U\mathbf{v}_2, \dots, U\mathbf{v}_n$ is an orthonormal basis.
4. A product of unitary operators is a unitary operator as well.

**6.2. Examples.** First of all, let us notice, that

| a matrix $U$ is an isometry if and only if its columns form an orthonormal system. |
|---|

This statement can be checked directly by computing the product $U^*U$.

It is easy to check that the columns of the rotation matrix

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

are orthogonal to each other, and that each column has norm 1. Therefore, the rotation matrix is an isometry, and since it is square, it is unitary. Since all entries of the rotation matrix are real, it is an orthogonal matrix.

The next example is more abstract. Let $X$ and $Y$ be inner product spaces, $\dim X = \dim Y = n$, and let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ be orthonormal bases in $X$ and $Y$ respectively. Define an operator $U : X \to Y$ by

$$U\mathbf{x}_k = \mathbf{y}_k, \qquad k = 1, 2, \dots, n.$$

Since for a vector $\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \dots + c_n\mathbf{x}_n$

$$\|\mathbf{x}\|^2 = |c_1|^2 + |c_2|^2 + \dots + |c_n|^2$$

and

$$\|U\mathbf{x}\|^2 = \|U(\sum_{k=1}^{n} c_k\mathbf{x}_k)\|^2 = \|\sum_{k=1}^{n} c_k\mathbf{y}_k\|^2 = \sum_{k=1}^{n} |c_k|^2,$$

one can conclude that $\|U\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in X$, so $U$ is a unitary operator.

**6.3. Properties of unitary operators.**

**Proposition 6.4.** *Let $U$ be a unitary matrix. Then*

1. $|\det U| = 1$. *In particular, for an orthogonal matrix* $\det U = \pm 1$;
2. *If $\lambda$ is an eigenvalue of $U$, then $|\lambda| = 1$*

**Remark.** Note, that for an orthogonal matrix, an eigenvalue (unlike the determinant) does not have to be real. Our old friend, the rotation matrix gives an example.

**Proof of Proposition 6.4.** Let $\det U = z$. Since $\det(U^*) = \overline{\det(U)}$, see Problem 5.1, we have

$$|z|^2 = \overline{z}z = \det(U^*U) = \det I = 1,$$

so $|\det U| = |z| = 1$. Statement 1 is proved.

To prove statement 2 let us notice that if $U\mathbf{x} = \lambda\mathbf{x}$ then

$$\|U\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|,$$

so $|\lambda| = 1$. □

### 6.4. Unitary equivalent operators.

**Definition.** Operators (matrices) $A$ and $B$ are called *unitarily equivalent* if there exists a unitary operator $U$ such that $A = UBU^*$.

Since for a unitary $U$ we have $U^{-1} = U^*$, any two unitary equivalent matrices are similar as well.

The converse is not true, it is easy to construct a pair of similar matrices, which are not unitarily equivalent.

The following proposition gives a way to construct a counterexample.

**Proposition 6.5.** *A matrix $A$ is unitarily equivalent to a diagonal one if and only if it has an orthogonal (orthonormal) basis of eigenvectors.*

**Proof.** Let $A = UBU^*$ and let $B\mathbf{x} = \lambda\mathbf{x}$. Then $AU\mathbf{x} = UBU^*U\mathbf{x} = UB\mathbf{x} = U(\lambda\mathbf{x}) = \lambda U\mathbf{x}$, i.e. $U\mathbf{x}$ is an eigenvector of $A$.

So, let $A$ be unitarily equivalent to a diagonal matrix $D$, i.e. let $A = UDU^*$. The vectors $\mathbf{e}_k$ of the standard basis are eigenvectors of $D$, so the vectors $U\mathbf{e}_k$ are eigenvectors of $A$. Since $U$ is unitary, the system $U\mathbf{e}_1, U\mathbf{e}_2, \dots, U\mathbf{e}_n$ is an orthonormal basis.

Now let $A$ has an orthogonal basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ of eigenvectors. Dividing each vector $\mathbf{u}_k$ by its norm if necessary, we can always assume that the system $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ is an *orthonormal* basis. Let $D$ be the matrix of $A$ in the basis $\mathcal{B} = \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$. Clearly, $D$ is a diagonal matrix.

Denote by $U$ the matrix with columns $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$. Since the columns form an orthonormal basis, $U$ is unitary. The standard change of coordinate formula implies

$$A = [A]_{\mathcal{S}\mathcal{S}} = [I]_{\mathcal{S}\mathcal{B}}[A]_{\mathcal{B}\mathcal{B}}[I]_{\mathcal{B}\mathcal{S}} = UDU^{-1}$$

and since $U$ is unitary, $A = UDU^*$. □

**Exercises.**

**6.1.** Orthogonally diagonalize the following matrices,

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}$$

i.e. for each matrix $A$ find a unitary matrix $U$ and a diagonal matrix $D$ such that $A = UDU^*$

**6.2.** True or false: a matrix is unitarily equivalent to a diagonal one if and only if it has an orthogonal basis of eigenvectors.

**6.3.** Prove the polarization identities

$$(A\mathbf{x}, \mathbf{y}) = \frac{1}{4}\big[(A(\mathbf{x}+\mathbf{y}), \mathbf{x}+\mathbf{y}) - (A(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{y})\big] \qquad \text{(real case, } A = A^*\text{)},$$

and

$$(A\mathbf{x}, \mathbf{y}) = \frac{1}{4}\sum_{\alpha=\pm 1, \pm i} \alpha(A(\mathbf{x}+\alpha\mathbf{y}), \mathbf{x}+\alpha\mathbf{y}) \qquad \text{(complex case, } A \text{ is arbitrary)}.$$

**6.4.** Show that a product of unitary (orthogonal) matrices is unitary (orthogonal) as well.

**6.5.** Let $U : X \to X$ be a linear transformation on a finite-dimensional inner product space. True or false:

a) If $\|U\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in X$, then $U$ is unitary.

b) If $\|U\mathbf{e}_k\| = \|\mathbf{e}_k\|$, $k = 1, 2 \ldots, n$ for some orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, then $U$ is unitary.

Justify your answers with a proof or a counterexample.

**6.6.** Let $A$ and $B$ be unitarily equivalent $n \times n$ matrices.

a) Prove that $\operatorname{trace}(A^*A) = \operatorname{trace}(B^*B)$.

b) Use a) to prove that

$$\sum_{j,k=1}^{n} |A_{j,k}|^2 = \sum_{j,k=1}^{n} |B_{j,k}|^2.$$

c) Use b) to prove that the matrices

$$\begin{pmatrix} 1 & 2 \\ 2 & i \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} i & 4 \\ 1 & 1 \end{pmatrix}$$

are not unitarily equivalent.

**6.7.** Which of the following pairs of matrices are unitarily equivalent:

a) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

b) $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}$.

c) $\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

d) $\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 0 & -i & 0 \\ 0 & 0 & i \end{pmatrix}$.

e) $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$.

**Hint:** It is easy to eliminate matrices that are not unitarily equivalent: remember, that unitarily equivalent matrices are similar, and trace, determinant and eigenvalues of similar matrices coincide.

Also, the previous problem helps in eliminating non unitarily equivalent matrices.

Finally, a matrix is unitarily equivalent to a diagonal one if and only if it has an orthogonal basis of eigenvectors.

**6.8.** Let $U$ be a $2 \times 2$ orthogonal matrix with $\det U = 1$. Prove that $U$ is a rotation matrix.

**6.9.** Let $U$ be a $3 \times 3$ orthogonal matrix with $\det U = 1$. Prove that

a) 1 is an eigenvalue of $U$.

b) If $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is an orthonormal basis, such that $U\mathbf{v}_1 = \mathbf{v}_1$ (remember, that 1 is an eigenvalue), then in this basis the matrix of $U$ is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix},$$

where $\alpha$ is some angle.

**Hint:** Show, that since $\mathbf{v}_1$ is an eigenvector of $U$, all entries below 1 must be zero, and since $\mathbf{v}_1$ is also an eigenvector of $U^*$ (why?), all entries right of 1 also must be zero. Then show that the lower right $2 \times 2$ matrix is an orthogonal one with determinant 1, and use the previous problem.

## 7. Rigid motions in $\mathbb{R}^n$

A rigid motion in an inner product space $V$ is a transformation $f : V \to V$ preserving the distance between point, i.e. such that

$$\|f(\mathbf{x}) - f(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in V.$$

Note, that in the definition we do not assume that the transformation $f$ is linear.

Clearly, any unitary transformation is a rigid motion. Another example of a rigid motion is a translation (shift) by $\mathbf{a} \in V$, $f(\mathbf{x}) = \mathbf{x} + \mathbf{a}$.

The main result of this section is the following theorem, stating that any rigid motion in a *real* inner product space is a composition of an orthogonal transformation and a translation.

**Theorem 7.1.** *Let $f$ be a rigid motion in a real inner product space $X$, and let $T(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{0})$. Then $T$ is an orthogonal transformation.*

To prove this theorem we need the following simple lemma.

**Lemma 7.2.** *Let $T$ be as defined in Theorem 7.1. Then for all $\mathbf{x}, \mathbf{y} \in X$*

1. $\|T\mathbf{x}\| = \|\mathbf{x}\|$;
2. $\|T(\mathbf{x}) - T(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|$;
3. $(T(\mathbf{x}), T(\mathbf{y})) = (\mathbf{x}, \mathbf{y})$.

**Proof.** To prove statement 1 notice that

$$\|T(\mathbf{x})\| = \|f(\mathbf{x}) - f(\mathbf{0})\| = \|\mathbf{x} - \mathbf{0}\| = \|\mathbf{x}\|.$$

Statement 2 follows from the following chain of identities:

$$\begin{aligned}\|T(\mathbf{x}) - T(\mathbf{y})\| &= \|(f(\mathbf{x}) - f(\mathbf{0})) - (f(\mathbf{y}) - f(\mathbf{0}))\| \\ &= \|f(\mathbf{x}) - f(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|.\end{aligned}$$

An alternative explanation would that $T$ is a composition of 2 rigid motions ($f$ followed by the translation by $\mathbf{a} = -f(\mathbf{0})$), and one can easily see that a composition of rigid motions is a rigid motion. Since $T(\mathbf{0}) = \mathbf{0}$, and so $\|T(\mathbf{x})\| = \|T(\mathbf{x}) - T(\mathbf{0})\|$, statement 1 can be treated as a particular case of statement 2.

To prove statement 3, let us notice that in a real inner product space

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

and

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

Recalling that $\|T(\mathbf{x}) - T(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|$ and $\|T(\mathbf{x})\| = \|\mathbf{x}\|$, $\|T(\mathbf{y})\| = \|\mathbf{y}\|$, we immediately get the desired conclusion. □

**Proof of Theorem 7.1.** First of all notice that for all $\mathbf{x} \in X$

$$\|T(\mathbf{x})\| = \|f(\mathbf{x}) - f(\mathbf{0})\| = \|\mathbf{x} - \mathbf{0}\| = \|\mathbf{x}\|,$$

so $T$ preserves the norm, $\|T\mathbf{x}\| = \|\mathbf{x}\|$.

We would like to say that the identyty $\|T\mathbf{x}\| = \|\mathbf{x}\|$ means $T$ is an isometry, but to be able to say that we need to prove that $T$ is a linear transformation.

To do that, let us fix an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ in $X$, and let $\mathbf{b}_k := T(\mathbf{e}_k)$, $k = 1, 2, \ldots, n$. Since $T$ preserves the inner product (statement

3 of Lemma 7.2), we can conclude that $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is an orthonormal system. In fact, since $\dim X = n$ (because basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ consists of $n$ vectors), we can conclude that $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is an orthonormal *basis*.

Let $\mathbf{x} = \sum_{k=1}^{n} \alpha_k \mathbf{e}_k$. Recall that by the abstract orthogonal Fourier decomposition (2.2) we have that $\alpha_k = (\mathbf{x}, \mathbf{e}_k)$. Applying the abstract orthogonal Fourier decomposition (2.2) to $T(\mathbf{x})$ and the orthonormal basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ we get

$$T(\mathbf{x}) = \sum_{k=1}^{n} (T(\mathbf{x}), \mathbf{b}_k)\mathbf{b}_k.$$

Since

$$(T(\mathbf{x}), \mathbf{b}_k) = (T(\mathbf{x}), T(\mathbf{e}_k)) = (\mathbf{x}, \mathbf{e}_k) = \alpha_k,$$

we get that

$$T\Big(\sum_{k=1}^{n} \alpha_k \mathbf{e}_k\Big) = \sum_{k=1}^{n} \alpha_k \mathbf{b}_k.$$

This means that $T$ is a linear transformation whose matrix with respect to the bases $\mathcal{S} := \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ and $\mathcal{B} := \mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is identity matrix, $[T]_{\mathcal{B},\mathcal{S}} = I$.

An alternative way to show that $T$ is a linear transformation is the following direct calculation

$$\begin{aligned}
&\|T(\mathbf{x} + \alpha\mathbf{y}) - (T(\mathbf{x}) + \alpha T(\mathbf{y}))\|^2 = \|(T(\mathbf{x} + \alpha\mathbf{y}) - T(\mathbf{x})) - \alpha T(\mathbf{y})\|^2 \\
&= \|T(\mathbf{x} + \alpha\mathbf{y}) - T(\mathbf{x})\|^2 + \alpha^2\|T(\mathbf{y})\|^2 - 2\alpha(T(\mathbf{x} + \alpha\mathbf{y}) - T(\mathbf{x}), T(\mathbf{y})) \\
&= \|\mathbf{x} + \alpha\mathbf{y} - \mathbf{x}\|^2 + \alpha^2\|\mathbf{y}\|^2 - 2\alpha(T(\mathbf{x} + \alpha\mathbf{y}), T(\mathbf{y})) + 2\alpha(T(\mathbf{x}), T(\mathbf{y})) \\
&= \alpha^2\|\mathbf{y}\|^2 + \alpha^2\|\mathbf{y}\|^2 - 2\alpha(\mathbf{x} + \alpha\mathbf{y}, \mathbf{y}) + 2\alpha(\mathbf{x}, \mathbf{y}) \\
&= 2\alpha^2\|\mathbf{y}\|^2 - 2\alpha(\mathbf{x}, \mathbf{y}) - 2\alpha^2(\mathbf{y}, \mathbf{y}) + 2\alpha(\mathbf{x}, \mathbf{y}) = 0
\end{aligned}$$

Therefore

$$T(\mathbf{x} - \alpha\mathbf{y}) = T(\mathbf{x}) + \alpha T(\mathbf{y}),$$

which implies that $T$ is linear (taking $\mathbf{x} = \mathbf{0}$ or $\alpha = 1$ we get two properties from the definition of a linear transformation).

So, $T$ is a linear transformation satisfying $\|T\mathbf{x}\| = \|\mathbf{x}\|$, i.e. $T$ is an isometry. Since $T : X \to X$, $T$ is unitary transformation (see Proposition 6.3). That completes the proof, since an orthogonal transformation is simply a unitary transformation in a real inner product space. □

**Exercises.**

**7.1.** Give an example of a rigid motion $T$ in $\mathbb{C}^n$, $T(\mathbf{0}) = \mathbf{0}$, which is not a linear transformation.

## 8. Complexification and decomplexification

This section is probably a bit more abstract than the rest of the chapter, and can be skipped at the first reading.

### 8.1. Decomplexification.

8.1.1. *Decomplexification of a vector space.* Any complex vector space can be interpreted a real vector space: we just need to forget that we can multiply vectors by complex numbers and act as only multiplication by real numbers is allowed.

For example, the space $\mathbb{C}^n$ is *canonically* identified with the real space $\mathbb{R}^{2n}$: each complex coordinate $z_k = x_k + iy_k$ gives us 2 real ones $x_k$ and $y_k$.

"Canonically" here means that this is a standard, most natural way of identifying $\mathbb{C}^n$ and $\mathbb{R}^{2n}$. Note, that while the above definition gives us a canonical way to get real coordinates from complex ones, it does not say anything about ordering real coordinates.

In fact, there are two standard ways to order the coordinates $x_k$, $y_k$. One way is to take first the real parts and then the imaginary parts, so the ordering is $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n$. The other standard alternative is the ordering $x_1, y_1, x_2, y_2, \dots, x_n, y_n$. The material of this section does not depend on the choice of ordering of coordinates, so the reader does not have to worry about picking an ordering.

8.1.2. *Decomplexification of an inner product.* It turns out that if we are given a complex inner product (in a complex space), we can in a canonical way get a real inner product from it. To see how we can do that, let as first consider the above example of $\mathbb{C}^n$ canonically identifies with $\mathbb{R}^{2n}$. Let $(\mathbf{x}, \mathbf{y})_{\mathbb{C}}$ denote the standard inner product in $\mathbb{C}^n$, and $(\mathbf{x}, \mathbf{y})_{\mathbb{R}}$ be the standard inner product in $\mathbb{R}^{2n}$ (note that the standard inner product in $\mathbb{R}^n$ does not depend on the ordering of coordinates). Then (see Exercise 8.1 below)

$$(\mathbf{x}, \mathbf{y})_{\mathbb{R}} = \operatorname{Re}(\mathbf{x}, \mathbf{y})_{\mathbb{C}} \tag{8.1}$$

This formula can be used to canonically define a real inner product from the complex one in general situation. Namely, it is an easy exercise to show that if $(\mathbf{x}, \mathbf{y})_{\mathbb{C}}$ is an inner product in a complex inner product space, then $(\mathbf{x}, \mathbf{y})_{\mathbb{R}}$ defined by (8.1) is a real inner product (on the corresponding real space).

Summarizing we can say that

> To decomplexify a complex inner product space we simply "forget" that we can multiply by complex numbers, i.e. we only allow multiplication by reals. The canonical real inner product in the decomplexified space is given by formula (8.1)

**Remark.** Any (complex) linear linear transformation on $\mathbb{C}^n$ (or, more generally, on a complex vector space) gives as a real linear transformation: it is simply the fact that if $T(\alpha\mathbf{x}+\beta\mathbf{y}) = \alpha T\mathbf{x} + \beta T\mathbf{y}$ holds for $\alpha, \beta \in \mathbb{C}$, then it of course holds for $\alpha, \beta \in \mathbb{R}$.

The converse is not true, i.e. a (real) linear transformation on the decomplexification $\mathbb{R}^{2n}$ of $\mathbb{C}^n$ does not always give a (complex) linear transformation of $\mathbb{C}^n$ (the same in the abstract settings).

For example, if one considers the case $n = 1$, then the multiplication by a complex number $z$ (general form of a linear transformation in $\mathbb{C}^1$) treated as a linear transformation in $\mathbb{R}^2$ has a very specific structure (can you describe it?).

**8.2. Complexification.** We can also do a converse, namely get a complex inner product space from a real one: in fact, you probably already did it before, without paying much attention to it.

Namely, given a real inner product space $\mathbb{R}^n$ we can obtain a complex space $\mathbb{C}^n$ out of it by allowing complex coordinates (with the standard inner product in both cases). The space $\mathbb{R}^n$ in this case will be a *real*[2]subspace of $\mathbb{C}^n$ consisting of vectors with real coordinates.

Abstractly, this construction can be described as follows: given a real vector space $X$ we can define its complexification $X_{\mathbb{C}}$ as the collection of all pairs $[\mathbf{x}_1, \mathbf{x}_2]$, $\mathbf{x}_1, \mathbf{x}_2 \in X$ with the addition and multiplication by a real number $\alpha$ are defined coordinate-wise,

$$[\mathbf{x}_1, \mathbf{x}_2] + [\mathbf{y}_1, \mathbf{y}_2] = [\mathbf{x}_1 + \mathbf{y}_1, \mathbf{x}_2 + \mathbf{y}_2], \qquad \alpha[\mathbf{x}_1, \mathbf{x}_2] = [\alpha\mathbf{x}_1, \alpha\mathbf{x}_2].$$

If $X = \mathbb{R}^n$ then the vector $\mathbf{x}_1$ consists of real parts of complex coordinates of $\mathbb{C}^n$ and the vector $\mathbf{x}_2$ of the imaginary parts. Thus informally one can write the pair $[\mathbf{x}_1, \mathbf{x}_2]$ as $\mathbf{x}_1 + i\mathbf{x}_2$.

To define multiplication by complex numbers we define multiplication by $i$ as

$$i[\mathbf{x}_1, \mathbf{x}_2] = [-\mathbf{x}_2, \mathbf{x}_1]$$

(writing $[\mathbf{x}_1, \mathbf{x}_2]$ as $\mathbf{x}_2 + i\mathbf{x}_2$ we can see that it must be defined this way) and define multiplication by arbitrary complex numbers using second distributive property $(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$.

If, in addition, $X$ is an inner product space we can extend the inner product to $X_{\mathbb{C}}$ by

$$\big([\mathbf{x}_1, \mathbf{x}_2], [\mathbf{y}_1, \mathbf{y}_2]\big)_{X_{\mathbb{C}}} = (\mathbf{x}_1, \mathbf{y}_1)_X + (\mathbf{x}_2, \mathbf{y}_2)_X;$$

The easiest way to see that everything is well defined, is to fix a basis (an orthonormal basis in the case of a real inner product space) and see what

[2]Real subspace mean the set closed with respect to sum and multiplication by real numbers

this construction gives us in coordinates. Then we can see that if we treat vector $\mathbf{x}_1$ as the vector consisting of the real parts of complex coordinates, and vector $\mathbf{x}_2$ as the vector consisting of imaginary parts of coordinates, then this construction is exactly the standard complexification of $\mathbb{R}^n$ (by allowing complex coordinates) described above.

The fact that we can express this construction in coordinate-free way, without picking a basis and working with coordinates, means that the result does not depend on the choice of a basis.

So, the easiest way to think about complexification is probably as follows:

> To construct a complexification of a real vector space $X$ we can pick a basis (an orthonormal basis if $X$ is a real inner product space) and then work with coordinates, allowing the complex ones. The resulting space does not depend on the choice of a basis; we can get from one coordinates to the others by the standard change of coordinate formula.

Note, that any linear transformation $T$ in the real space $X$ gives rise to a linear transformation $T_{\mathbb{C}}$ in the complexification $X_{\mathbb{C}}$.

The easiest way to see that is to fix a basis in $X$ (an orthonormal basis if $X$ is a real inner product space) and to work in a coordinate representation: in this case $T_{\mathbb{C}}$ has the same matrix as $T$. In the abstract representation we can write

$$T_{\mathbb{C}}[\mathbf{x}_1, \mathbf{x}_2] = [T\mathbf{x}_1, T\mathbf{x}_2].$$

On the other hand, not all linear transformations in $X_{\mathbb{C}}$ can be obtained from the transformations in $X$; if we do complexification in coordinates, only the transformations with real matrices work.

Note, that this is completely opposite to the situation in the case of decomplexification, described in Section 8.1.

An attentive reader probably already noticed that the operations of complexification and decomplixification are not the inverse of each other. First, the space and its complexification have the same dimension, while the decomplixification of an $n$-dimensional space has dimension $2n$. Moreover, as we just discussed, the relation between real and complex linear transformations is completely opposite in these cases.

In the next section we discuss the operation, inverse in some sense to decomplexification.

**8.3. Introducing complex structure to a real space.** The construction described in this section works only for real spaces of even dimension.

8.3.1. *An elementary way to introduce a complex structure.* Let $X$ be a real inner product space of dimension $2n$. We want invert the decomplexification

procedure to introduce a complex structure on $X$, i.e. to identify this space with a complex space such that its decomplexification (see Section 8.1) give us the original space $X$. The simplest idea is to fix an orthonormal basis in $X$ and then split the coordinates in this basis into two equal parts.

We than treat one half of the coordinates (say coordinates $x_1, x_2, \dots, x_n$) as real parts of complex coordinates, and treat the rest as the imaginary parts. Then we have to join real and imaginary parts together: for example if we treat $x_1, x_2, \dots, x_n$ as real parts and $x_{n+1}, x_{n+2}, \dots, x_{2n}$ as imaginary parts, we can define complex coordinates $z_k = x_k + ix_{n+k}$.

Of course, the result will generally depend on the choice of the orthonormal basis and on the way we split the real coordinates into real and imaginary parts and on how we join them.

One can also see from the decomplexification construction described in Section 8.1 that all complex structures on a real inner product space $X$ can be obtained in this way.

8.3.2. *From elementary to abstract construction of complex structure.* The above construction can be described in an abstract, coordinate-free way. Namely, let us split the space $X$ as $X = E \oplus E^\perp$, where $E$ is a subspace, $\dim E = n$ (so $\dim E^\perp = n$ as well), and let $U_0 : E \to E^\perp$ be a unitary (more precisely, orthogonal, since our spaces are real) transformation.

Note, that if $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthonormal basis in $E$, then the system $U_0\mathbf{v}_1, U_0\mathbf{v}_2, \dots, U_0\mathbf{v}_n$ is an orthonormal basis in $E^\perp$, so

$$\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n, U_0\mathbf{v}_1, U_0\mathbf{v}_2, \dots, U_0\mathbf{v}_n$$

is an orthonormal basis in the whole space $X$.

If $x_1, x_2, \dots, x_{2n}$ are coordinates of a vector $\mathbf{x}$ in this basis, and we treat $x_k + ix_{n+k}$, $k = 1, 2, \dots, n$ as complex coordinates of $\mathbf{x}$, then the multiplication by $i$ is represented by the orthogonal transformation $U$ which is given in the orthogonal basis of subspaces $E, E^\perp$ by the block matrix

$$U = \begin{pmatrix} \mathbf{0} & -U_0^* \\ U_0 & \mathbf{0} \end{pmatrix}.$$

This means that

$$i \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = U \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} & -U_0^* \\ U_0 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$$

$\mathbf{x}_1 \in E$, $\mathbf{x}_2 \in E^\perp$.

Clearly, $U$ is an orthogonal transformation such that $U^2 = -I$. Therefore, any complex structure on $X$ is given by an orthogonal transformation $U$, satisfying $U^2 = -I$; the transformation $U$ gives us the multiplication by the imaginary unit $i$.

The converse is also true, namely any orthogonal transformation $U$ satisfying $U^2 = -I$ defines a complex structure on a real inner product space $X$. Let us explain how.

8.3.3. *An abstract construction of complex structure.* Let us first consider an abstract explanation. To define a complex structure, we need to define the multiplication of vectors by complex numbers (initially we only can multiply by real numbers). In fact we need only to define the multiplication by $i$, the rest will follow from linearity in the original real space. And the multiplication by $i$ is given by the orthogonal transformation $U$ satisfying $U^2 = -I$.

Namely, if the multiplication by $i$ is given by $U$, $i\mathbf{x} = U\mathbf{x}$, then the complex multiplication must be defined by

$$(\alpha + \beta i)\mathbf{x} := \alpha\mathbf{x} + \beta U\mathbf{x} = (\alpha I + \beta U)\mathbf{x}, \qquad \alpha, \beta \in \mathbb{R}, \quad \mathbf{x} \in X. \tag{8.2}$$

We will use this formula now as the definition of complex multiplication.

It is not hard to check that for the complex multiplication defined above by (8.2) all axioms of comples wector space are satisfied. One can see that, for example by using linearity in the real space $X$ and noticing that that with respect to algebraic operations (addition and multiplication) the linear transformations of form

$$\alpha I = \beta U, \qquad \alpha, \beta \in \mathbb{R},$$

behave absolutely the same way as complex numbers $\alpha + \beta i$, i.e such transformations give us a *representation* of the field of complex numbers $\mathbb{C}$.

This means that first, a sum and a product of transformations of the form $\alpha I + \beta U$ is a transformation of the same form, and to get the coeeficients $\alpha$, $\beta$ of the result we can perform the operation on the corresponding complex numbers and take the real and imaginary parts of the result. Note, that here we need the identity $U^2 = -I$, but we do not need the fact that $U$ is an orthogonal transformation.

Thus, we got the structure of a complex vector space. To get a complex *inner product* space we need to introduce complex inner product, such that the original real inner product is the real part of it.

We really do not have nay choice here: noticing that for a complex inner product

$$\operatorname{Im}(\mathbf{x}, \mathbf{y}) = \operatorname{Re}\big[-i(\mathbf{x}, \mathbf{y})_{\mathbb{R}}\big] = \operatorname{Re}(\mathbf{x}, i\mathbf{y})_{\mathbb{R}},$$

we see that the only way to define the complex inner product is

$$(\mathbf{x}, \mathbf{y})_{\mathbb{C}} := (\mathbf{x}, \mathbf{y})_{\mathbb{R}} + i(\mathbf{x}, U\mathbf{y})_{\mathbb{R}}. \tag{8.3}$$

Let us show that this is indeed an inner product. We will need the fact that $U^* = -U$, see Exercise 8.4 below (by $U^*$ here we mean the adjoint with respect to the original real inner product).

To show that $(\mathbf{y},\mathbf{x})_{\mathbb{C}} = \overline{(\mathbf{x},\mathbf{y})_{\mathbb{C}}}$ we use the udentity $U^* = -U$ and symmetry of the real inner product:

$$\begin{aligned}(\mathbf{y},\mathbf{x})_{\mathbb{C}} &= (\mathbf{y},\mathbf{x})_{\mathbb{R}} + i(\mathbf{y},U\mathbf{x}) \\ &= (\mathbf{x},\mathbf{y})_{\mathbb{R}} + i(U\mathbf{x},\mathbf{y})_{\mathbb{R}} \\ &= (\mathbf{x},\mathbf{y})_{\mathbb{R}} - i(\mathbf{x},U\mathbf{y})_{\mathbb{R}} \\ &= \overline{(\mathbf{x},\mathbf{y})_{\mathbb{R}} + i(\mathbf{x},U\mathbf{y})_{\mathbb{R}}} \\ &= \overline{(\mathbf{x},\mathbf{y})_{\mathbb{C}}}.\end{aligned}$$

To prove the linearity of the complex inner product, let us first notice that $(\mathbf{x},\mathbf{y})_{\mathbb{C}}$ is *real linear* in the first (in fact in each) argument, i.e. that $(\alpha\mathbf{x}+\beta\mathbf{y},\mathbf{z})_{\mathbb{C}} = \alpha(\mathbf{x},\mathbf{z})_{\mathbb{C}} + \beta(\mathbf{y},\mathbf{z})_{\mathbb{C}}$ for $\alpha,\beta\in\mathbb{R}$; this is true because each summand in the right side of (8.3) is real linear in argument.

Using real linearity of $(\mathbf{x},\mathbf{y})_{\mathbb{C}}$ and the identity $U^* = -U$ (which implies that $(U\mathbf{x},\mathbf{y})_{\mathbb{R}} = -(\mathbf{x},U\mathbf{y})_{\mathbb{R}}$) together with the orthogonality of $U$, we get the following chain of equalities

$$\begin{aligned}((\alpha I+\beta U)\mathbf{x},\mathbf{y})_{\mathbb{C}} &= \alpha(\mathbf{x},\mathbf{y})_{\mathbb{C}} + \beta(U\mathbf{x},\mathbf{y})_{\mathbb{C}} \\ &= \alpha(\mathbf{x},\mathbf{y})_{\mathbb{C}} + \beta\left[(U\mathbf{x},\mathbf{y})_{\mathbb{R}} + i(U\mathbf{x},U\mathbf{y})_{\mathbb{R}}\right] \\ &= \alpha(\mathbf{x},\mathbf{y})_{\mathbb{C}} + \beta\left[-(\mathbf{x},U\mathbf{y})_{\mathbb{R}} + i(\mathbf{x},\mathbf{y})_{\mathbb{R}}\right] \\ &= \alpha(\mathbf{x},\mathbf{y})_{\mathbb{C}} + \beta i\left[(\mathbf{x},\mathbf{y})_{\mathbb{R}} + i(\mathbf{x},U\mathbf{y})_{\mathbb{R}}\right] \\ &= \alpha(\mathbf{x},\mathbf{y})_{\mathbb{C}} + \beta i(\mathbf{x},\mathbf{y})_{\mathbb{C}} = (\alpha+\beta i)(\mathbf{x},\mathbf{y})_{\mathbb{C}},\end{aligned}$$

which proves *complex* linearity.

Finally, to prove non-negativity of $(\mathbf{x},\mathbf{x})_{\mathbb{C}}$ let us notice (see Exercise 8.3 below) that $(\mathbf{x},U\mathbf{x})_{\mathbb{R}} = 0$, so

$$(\mathbf{x},\mathbf{x})_{\mathbb{C}} = (\mathbf{x},\mathbf{x})_{\mathbb{R}} = \|\mathbf{x}\|^2 \geq 0.$$

8.3.4. *The abstract construction via the elementary one.* For a reader who is not comfortable with such "high brow" and abstract proof, there is another, more hands on, explanation.

Namely, it can be shown, see Exercise 8.5 below, that there exists a subspace $E$, $\dim E = n$ (recall that $\dim X = 2n$), such that the matrix of $U$ with respect to the decomposition $X = E\oplus E^\perp$ is given by

$$U = \begin{pmatrix} \mathbf{0} & -U_0^* \\ U_0 & \mathbf{0}\end{pmatrix},$$

where $U_0 : E \to E^\perp$ is some orthogonal transformation.

Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be an orthonormal basis in $E$. Then the system $U_0\mathbf{v}_1, U_0\mathbf{v}_2, \dots, U_0\mathbf{v}_n$ is an orthonormal basis in $E^\perp$, so

$$\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n, U_0\mathbf{v}_1, U_0\mathbf{v}_2, \dots, U_0\mathbf{v}_n$$

is an orthonormal basis in the whole space $X$. Considering the coordinates $x_1, x_2, \dots, x_{2n}$ in this basis and treating $x_k + ix_{n+k}$ as complex coordinates, we get an elementary, "coordinate" way of defining complex structure, which was already described above. But if we look carefully, we see that multiplication by $i$ is given by the transformation $U$: it is trivial for $\mathbf{x} \in E$ and for $\mathbf{y} \in E^\perp$, and so it is true for all *real* linear combinations of $\alpha\mathbf{x} + \beta\mathbf{y}$, i.e. for all vectors in $X$.

But that means that the abstract introduction of complex structure and the corresponding elementary approach give us the same result! And since the elementary approach clearly gives us the a complex structure, the abstract approach gives us the same complex structure.

**Exercises.**

**8.1.** Prove formula (8.1). Namely, show that if

$$\mathbf{x} = (z_1, z_2, \dots, z_n)^T, \qquad \mathbf{y} = (w_1, w_2, \dots, w_n)^T,$$

$z_k = x_k + iy_k$, $w_k = u_k + iv_k$, $x_k, y_k, u_k, v_k \in \mathbb{R}$, then

$$\operatorname{Re}\Big(\sum_{k=1}^{n} z_k \overline{w}_k\Big) = \sum_{k=1}^{n} x_k u_k + \sum_{k=1}^{n} y_k v_k.$$

**8.2.** Show that if $(\mathbf{x}, \mathbf{y})_{\mathbb{C}}$ is an inner product in a complex inner product space, then $(\mathbf{x}, \mathbf{y})_{\mathbb{R}}$ defined by (8.1) is a real inner product space.

**8.3.** Let $U$ be an orthogonal transformation (in a real inner product space $X$), satisfying $U^2 = -I$. Prove that for all $\mathbf{x} \in X$

$$U\mathbf{x} \perp \mathbf{x}.$$

**8.4.** Show, that if $U$ is an orthogonal transformation satisfying $U^2 = -I$, then $U^* = -U$.

**8.5.** Let $U$ be an orthogonal transformation in a real inner product space, satisfying $U^2 = -I$. Show that in this case $\dim X = 2n$, and that there exists a subspace $E \subset X$, $\dim E = n$, and an orthogonal transformation $U_0 : E \to E^\perp$ such that $U$ in the decomposition $X = E \oplus E^\perp$ is given by the block matrix

$$U = \begin{pmatrix} \mathbf{0} & -U_0^* \\ U_0 & \mathbf{0} \end{pmatrix}.$$

This statement can be easily obtained from Theorem 5.1 of Chapter 6, if one notes that the only rotation $R_\alpha$ in $\mathbb{R}^2$ satisfying $R_\alpha^2 = -I$ are rotations through $\alpha = \pm\pi/2$.

However, one can find an elementary proof here, not using this theorem. For example, the statement is trivial if $\dim X = 2$: in this case we can take for $E$ any one-dimensional subspace, see Exercise 8.3.

Then it is not hard to show, that such operator $U$ does not exists in $\mathbb{R}^2$, and one can use induction in $\dim X$ to complete the proof.

*Chapter 6*

# Structure of operators in inner product spaces.

In this chapter we are again assuming that all spaces are finite-dimensional. Again, we are dealing only with complex or real spaces, theory of inner product spaces does not apply to spaces over general fields. When we are not mentioning what space are we in, everything work for both complex and real spaces.

To avoid writing essentially the same formulas twice we will use the notation for the complex case: in the real case it give correct, although sometimes a bit more complicated, formulas.

## 1. Upper triangular (Schur) representation of an operator.

**Theorem 1.1.** *Let $A : X \to X$ be an operator acting in a complex inner product space. There exists an orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ in $X$ such that the matrix of $A$ in this basis is upper triangular.*

*In other words, any $n \times n$ matrix $A$ can be represented as $A = UTU^*$, where $U$ is a unitary, and $T$ is an upper triangular matrix.*

**Proof.** We prove the theorem using the induction in $\dim X$. If $\dim X = 1$ the theorem is trivial, since any $1 \times 1$ matrix is upper triangular.

Suppose we proved that the theorem is true if $\dim X = n - 1$, and we want to prove it for $\dim X = n$.

Let $\lambda_1$ be an eigenvalue of $A$, and let $\mathbf{u}_1$, $\|\mathbf{u}_1\| = 1$ be a corresponding eigenvector, $A\mathbf{u}_1 = \lambda_1\mathbf{u}_1$. Denote $E = \mathbf{u}_1^\perp$, and let $\mathbf{v}_2, \dots, \mathbf{v}_n$ be some orthonormal basis in $E$ (clearly, $\dim E = \dim X - 1 = n-1$), so $\mathbf{u}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthonormal basis in $X$. In this basis the matrix of $A$ has the form

$$\left(\begin{array}{c|c} \lambda_1 & * \\ \hline 0 & \\ \vdots & A_1 \\ 0 & \end{array}\right); \tag{1.1}$$

here all entries below $\lambda_1$ are zeroes, and $*$ means that we do not care what entries are in the first row right of $\lambda_1$.

We do care enough about the lower right $(n-1)\times(n-1)$ block, to give it name: we denote it as $A_1$.

Note, that $A_1$ defines a linear transformation in $E$, and since $\dim E = n-1$, the induction hypothesis implies that there exists an orthonormal basis (let us denote it as $\mathbf{u}_2, \dots, \mathbf{u}_n$) in which the matrix of $A_1$ is upper triangular.

So, matrix of $A$ in the orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ has the form (1.1), where matrix $A_1$ is upper triangular. Therefore, the matrix of $A$ in this basis is upper triangular as well. □

**Remark.** Note, that the subspace $E = \mathbf{u}_1^\perp$ introduced in the proof is not invariant under $A$, i.e. the inclusion $AE \subset E$ does not necessarily hold. That means that $A_1$ is not a part of $A$, it is some operator constructed from $A$.

Note also, that $AE \subset E$ if and only if all entries denoted by $*$ (i.e. all entries in the first row, except $\lambda_1$) are zero.

**Remark.** Note, that even if we start from a real matrix $A$, the matrices $U$ and $T$ can have complex entries. The rotation matrix

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \qquad \alpha \neq k\pi, k \in \mathbb{Z}$$

is not unitarily equivalent (not even similar) to a real upper triangular matrix. Indeed, eigenvalues of this matrix are complex, and the eigenvalues of an upper triangular matrix are its diagonal entries.

**Remark.** An analogue of Theorem 1.1 can be stated and proved for an arbitrary vector space, without requiring it to have an inner product. In this case the theorem claims that any operator have an upper triangular form in some basis. A proof can be modeled after the proof of Theorem 1.1. An alternative way is to equip $V$ with an inner product by fixing a basis in $V$ and declaring it to be an orthonormal one, see Problem 2.4 in Chapter 5.

Note, that the version for inner product spaces (Theorem 1.1) is stronger than the one for the vector spaces, because it says that we always can find an orthonormal basis, not just a basis.

The following theorem is a real-valued version of Theorem 1.1

**Theorem 1.2.** *Let $A : X \to X$ be an operator acting in a* real *inner product space. Suppose that all eigenvalues of $A$ are real (meaning that $A$ has exactly $n = \dim X$ real eigenvalues, counting multiplicities). Then there exists an orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ in $X$ such that the matrix of $A$ in this basis is upper triangular.*

*In other words, any real $n \times n$ matrix $A$ with all real eigenvalues can be represented as $T = UTU^* = UTU^T$, where $U$ is an orthogonal, and $T$ is a real upper triangular matrices.*

**Proof.** To prove the theorem we just need to analyze the proof of Theorem 1.1. Let us assume (we can always do this without loss of generality) that the operator (matrix) $A$ acts in $\mathbb{R}^n$.

Suppose, the theorem is true for $(n-1) \times (n-1)$ matrices. As in the proof of Theorem 1.1 let $\lambda_1$ be a real eigenvalue of $A$, $\mathbf{u}_1 \in \mathbb{R}^n$, $\|\mathbf{u}_1\| = 1$ be a corresponding eigenvector, and let $\mathbf{v}_2, \dots, \mathbf{v}_n$ be on orthonormal system (in $\mathbb{R}^n$) such that $\mathbf{u}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthonormal basis in $\mathbb{R}^n$.

The matrix of $A$ in this basis has form (1.1), where $A_1$ is some real matrix.

If we can prove that matrix $A_1$ has only real eigenvalues, then we are done. Indeed, then by the induction hypothesis there exists an orthonormal basis $\mathbf{u}_2, \dots, \mathbf{u}_n$ in $E = \mathbf{u}_1^\perp$ such that the matrix of $A_1$ in this basis is upper triangular, so the matrix of $A$ in the basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ is also upper triangular.

To show that $A_1$ has only real eigenvalues, let us notice that

$$\det(A - \lambda I) = (\lambda_1 - \lambda) \det(A_1 - \lambda)$$

(take the cofactor expansion in the first row, for example), and so any eigenvalue of $A_1$ is also an eigenvalue of $A$. But $A$ has only real eigenvalues! □

**Exercises.**

**1.1.** Use the upper triangular representation of an operator to give an alternative proof of the fact that determinant is the product and the trace is the sum of eigenvalues counting multiplicities.

## 2. Spectral theorem for self-adjoint and normal operators.

In this section we deal with matrices (operators) which are unitarily equivalent to diagonal matrices.

Let us recall that an operator is called self-adjoint if $A = A^*$. A matrix of a self-adjoint operator (in some orthonormal basis), i.e. a matrix satisfying $A^* = A$ is called a Hermitian matrix.

The terms *self-adjoint* and *Hermitian* essentially mean the same. Usually people say self-adjoint when speaking about operators (transformations), and Hermitian when speaking about matrices. We will try to follow this convention, but since we often do not distinguish between operators and their matrices, we will sometimes mix both terms.

**Theorem 2.1.** *Let $A = A^*$ be a self-adjoint operator in an inner product space $X$ (the space can be complex or real). Then all eigenvalues of $A$ are real, and there exists and orthonormal basis of eigenvectors of $A$ in $X$.*

This theorem can be restated in matrix form as follows

**Theorem 2.2.** *Let $A = A^*$ be a self-adjoint (and therefore square) matrix. Then $A$ can be represented as*

$$A = UDU^*,$$

*where $U$ is a unitary matrix and $D$ is a diagonal matrix with real entries.*

*Moreover, if the matrix $A$ is real, matrix $U$ can be chosen to be real (i.e. orthogonal).*

**Proof.** To prove Theorems 2.1 and 2.2 let us first apply Theorem 1.1 (Theorem 1.2 if $X$ is a real space) to find an orthonormal basis in $X$ such that the matrix of $A$ in this basis is upper triangular. Now let us ask ourself a question: What upper triangular matrices are self-adjoint?

The answer is immediate: an upper triangular matrix is self-adjoint if and only if it is a diagonal matrix with real entries. Theorem 2.1 (and so Theorem 2.2) is proved. □

**Remark.** In many textbooks only real matrices are considered, and Theorem 2.2 is often called the "Spectral Theorem for *symmetric* matrices". However, we should emphasize that the conclusion of Theorem 2.2 fails for *complex* symmetric matrices: the theorem holds for Hermitian matrices, and in particular for *real* symmetric matrices.

Let us give an independent proof to the fact that eigenvalues of a self-adjoint operators are real. Let $A = A^*$ and $A\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$. Then

$$(A\mathbf{x}, \mathbf{x}) = (\lambda\mathbf{x}, \mathbf{x}) = \lambda(\mathbf{x}, \mathbf{x}) = \lambda\|\mathbf{x}\|^2.$$

On the other hand,

$$(A\mathbf{x}, \mathbf{x}) = (\mathbf{x}, A^*\mathbf{x}) = (\mathbf{x}, A\mathbf{x}) = (\mathbf{x}, \lambda\mathbf{x}) = \overline{\lambda}(\mathbf{x}, \mathbf{x}) = \overline{\lambda}\|\mathbf{x}\|^2,$$

so $\lambda\|\mathbf{x}\|^2 = \overline{\lambda}\|\mathbf{x}\|^2$. Since $\|\mathbf{x}\| \neq 0$ ($\mathbf{x} \neq \mathbf{0}$), we can conclude $\lambda = \overline{\lambda}$, so $\lambda$ is real.

It also follows from Theorem 2.1 that eigenspaces of a self-adjoint operator are orthogonal. Let us give an alternative proof of this result.

**Proposition 2.3.** *Let $A = A^*$ be a self-adjoint operator, and let $\mathbf{u}, \mathbf{v}$ be its eigenvectors, $A\mathbf{u} = \lambda\mathbf{u}$, $A\mathbf{v} = \mu\mathbf{v}$. Then, if $\lambda \neq \mu$, the eigenvectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal.*

**Proof.** This proposition follows from the spectral theorem (Theorem 1.1), but here we are giving a direct proof. Namely,

$$(A\mathbf{u}, \mathbf{v}) = (\lambda\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v}).$$

On the other hand

$$(A\mathbf{u}, \mathbf{v}) = (\mathbf{u}, A^*\mathbf{v}) = (\mathbf{u}, A\mathbf{v}) = (\mathbf{u}, \mu\mathbf{v}) = \overline{\mu}(\mathbf{u}, \mathbf{v}) = \mu(\mathbf{u}, \mathbf{v})$$

(the last equality holds because eigenvalues of a self-adjoint operator are real), so $\lambda(\mathbf{u}, \mathbf{v}) = \mu(\mathbf{u}, \mathbf{v})$. If $\lambda \neq \mu$ it is possible only if $(\mathbf{u}, \mathbf{v}) = 0$. □

Now let us try to find what matrices are unitarily equivalent to a diagonal one. It is easy to check that for a diagonal matrix $D$

$$D^*D = DD^*.$$

Therefore $A^*A = AA^*$ if the matrix of $A$ in some orthonormal basis is diagonal.

**Definition.** An operator (matrix) $N$ is called *normal* if $N^*N = NN^*$.

Clearly, any self-adjoint operator ($A^*A = AA^*$) is normal. Also, any unitary operator $U : X \to X$ is normal since $U^*U = UU^* = I$.

Note, that a normal operator is an operator acting in one space, not from one space to another. So, if $U$ is a unitary operator acting from one space to another, we cannot say that $U$ is normal.

**Theorem 2.4.** *Any normal operator $N$ in a complex vector space has an orthonormal basis of eigenvectors.*

*In other words, any matrix $N$ satisfying $N^*N = NN^*$ can be represented as*

$$N = UDU^*,$$

*where $U$ is a unitary matrix, and $D$ is a diagonal one.*

**Remark.** Note, that in the above theorem even if $N$ is a real matrix, we did not claim that matrices $U$ and $D$ are real. Moreover, it can be easily shown, that if $D$ is real, $N$ must be self-adjoint.

**Proof of Theorem 2.4.** To prove Theorem 2.4 we apply Theorem 1.1 to get an orthonormal basis, such that the matrix of $N$ in this basis is upper triangular. To complete the proof of the theorem we only need to show that an upper triangular normal matrix must be diagonal.

We will prove this using induction in the dimension of matrix. The case of $1 \times 1$ matrix is trivial, since any $1 \times 1$ matrix is diagonal.

Suppose we have proved that any $(n-1) \times (n-1)$ upper triangular normal matrix is diagonal, and we want to prove it for $n \times n$ matrices. Let $N$ be $n \times n$ upper triangular normal matrix. We can write it as

$$N = \left(\begin{array}{c|ccc} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ \hline 0 & & & \\ \vdots & & N_1 & \\ 0 & & & \end{array}\right)$$

where $N_1$ is an upper triangular $(n-1) \times (n-1)$ matrix.

Let us compare upper left entries (first row first column) of $N^*N$ and $NN^*$. Direct computation shows that that

$$(N^*N)_{1,1} = \overline{a}_{1,1}a_{1,1} = |a_{1,1}|^2$$

and

$$(NN^*)_{1,1} = |a_{1,1}|^2 + |a_{1,2}|^2 + \ldots + |a_{1,n}|^2.$$

So, $(N^*N)_{1,1} = (NN^*)_{1,1}$ if and only if $a_{1,2} = \ldots = a_{1,n} = 0$. Therefore, the matrix $N$ has the form

$$N = \left(\begin{array}{c|ccc} a_{1,1} & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & N_1 & \\ 0 & & & \end{array}\right)$$

It follows from the above representation that

$$N^*N = \left(\begin{array}{c|ccc} |a_{1,1}|^2 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & N_1^*N_1 & \\ 0 & & & \end{array}\right), \qquad NN^* = \left(\begin{array}{c|ccc} |a_{1,1}|^2 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & N_1N_1^* & \\ 0 & & & \end{array}\right)$$

so $N_1^*N_1 = N_1N_1^*$. That means the matrix $N_1$ is also normal, and by the induction hypothesis it is diagonal. So the matrix $N$ is also diagonal. □

The following proposition gives a very useful characterization of normal operators.

**Proposition 2.5.** *An operator $N : X \to X$ is normal if and only if*

$$\|N\mathbf{x}\| = \|N^*\mathbf{x}\| \qquad \forall \mathbf{x} \in X.$$

**Proof.** Let $N$ be normal, $N^*N = NN^*$. Then

$$\|N\mathbf{x}\|^2 = (N\mathbf{x}, N\mathbf{x}) = (N^*N\mathbf{x}, \mathbf{x}) = (NN^*\mathbf{x}, \mathbf{x}) = (N^*\mathbf{x}, N^*\mathbf{x}) = \|N^*\mathbf{x}\|^2$$

so $\|N\mathbf{x}\| = \|N^*\mathbf{x}\|$.

Now let

$$\|N\mathbf{x}\| = \|N^*\mathbf{x}\| \qquad \forall \mathbf{x} \in X.$$

The Polarization Identities (Lemma 1.9 in Chapter 5) imply that for all $\mathbf{x}, \mathbf{y} \in X$

$$\begin{aligned}
(N^*N\mathbf{x}, \mathbf{y}) = (N\mathbf{x}, N\mathbf{y}) &= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|N\mathbf{x} + \alpha N\mathbf{y}\|^2 \\
&= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|N(\mathbf{x} + \alpha\mathbf{y})\|^2 \\
&= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|N^*(\mathbf{x} + \alpha\mathbf{y})\|^2 \\
&= \frac{1}{4} \sum_{\alpha=\pm 1, \pm i} \alpha \|N^*\mathbf{x} + \alpha N^*\mathbf{y}\|^2 \\
&= (N^*\mathbf{x}, N^*\mathbf{y}) = (NN^*\mathbf{x}, \mathbf{y})
\end{aligned}$$

and therefore (see Corollary 1.6) $N^*N = NN^*$. □

**Exercises.**

**2.1.** True or false:

a) Every unitary operator $U : X \to X$ is normal.

b) A matrix is unitary if and only if it is invertible.

c) If two matrices are unitarily equivalent, then they are also similar.

d) The sum of self-adjoint operators is self-adjoint.

e) The adjoint of a unitary operator is unitary.

f) The adjoint of a normal operator is normal.

g) If all eigenvalues of a linear operator are 1, then the operator must be unitary or orthogonal.

h) If all eigenvalues of a normal operator are 1, then the operator is identity.

i) A linear operator may preserve norm but not the inner product.

**2.2.** True or false: The sum of normal operators is normal? Justify your conclusion.

**2.3.** Show that an operator unitarily equivalent to a diagonal one is normal.

**2.4.** Orthogonally diagonalize the matrix,

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

Find all square roots of $A$, i.e. find all matrices $B$ such that $B^2 = A$.

Note, that all square roots of $A$ are self-adjoint.

**2.5.** True or false: any self-adjoint matrix has a self-adjoint square root. Justify.

**2.6.** Orthogonally diagonalize the matrix,

$$A = \begin{pmatrix} 7 & 2 \\ 2 & 4 \end{pmatrix},$$

i.e. represent it as $A = UDU^*$, where $D$ is diagonal and $U$ is unitary.

Among all square roots of $A$, i.e. among all matrices $B$ such that $B^2 = A$, find one that has positive eigenvalues. You can leave $B$ as a product.

**2.7.** True or false:

a) A product of two self-adjoint matrices is self-adjoint.

b) If $A$ is self-adjoint, then $A^k$ is self-adjoint.

Justify your conclusions

**2.8.** Let $A$ be $m \times n$ matrix. Prove that

a) $A^*A$ is self-adjoint.

b) All eigenvalues of $A^*A$ are non-negative.

c) $A^*A + I$ is invertible.

**2.9.** Give a proof if the statement is true, or give a counterexample if it is false:

a) If $A = A^*$ then $A + iI$ is invertible.

b) If $U$ is unitary, $U + \frac{3}{4}I$ is invertible.

c) If a matrix $A$ is real, $A - iI$ is invertible.

**2.10.** Orthogonally diagonalize the rotation matrix

$$R_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix},$$

where $\alpha$ is not a multiple of $\pi$. Note, that you will get complex eigenvalues in this case.

**2.11.** Orthogonally diagonalize the matrix

$$A = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}.$$

**Hints:** You will get real eigenvalues in this case. Also, the trigonometric identities $\sin 2x = 2\sin x \cos x$, $\sin^2 x = (1 - \cos 2x)/2$, $\cos^2 x = (1 + \cos 2x)/2$ (applied to $x = \alpha/2$) will help to simplify expressions for eigenvectors.

**2.12.** Can you describe the linear transformation with matrix $A$ from the previous problem geometrically? Is has a very simple geometric interpretation.

**2.13.** Prove that a normal operator with unimodular eigenvalues (i.e. with all eigenvalues satisfying $|\lambda_k| = 1$) is unitary. **Hint:** Consider diagonalization

**2.14.** Prove that a normal operator with real eigenvalues is self-adjoint.

**2.15.** Show by example that conclusion of Theorem 2.2 fails for *complex* symmetric matrices. Namely

a) construct a (diagonalizable) $2\times 2$ complex symmetric matrix not admitting an orthogonal basis of eigenvectors;

b) construct a $2\times 2$ complex symmetric matrix which cannot be diagonalized.

## 3. Polar and singular value decompositions.

### 3.1. Positive definite operators. Square roots.

**Definition.** A self adjoint operator $A : X \to X$ is called *positive definite* if

$$(A\mathbf{x}, \mathbf{x}) > 0 \qquad \forall \mathbf{x} \neq \mathbf{0},$$

and it is called *positive semidefinite* if

$$(A\mathbf{x}, \mathbf{x}) \geq 0 \qquad \forall \mathbf{x} \in X.$$

We will use the notation $A > 0$ for positive definite operators, and $A \geq 0$ for positive semi-definite.

The following theorem describes positive definite and semidefinite operators.

**Theorem 3.1.** *Let $A = A^*$. Then*

1. *$A > 0$ if and only if all eigenvalues of $A$ are positive.*
2. *$A \geq 0$ if and only if all eigenvalues of $A$ are non-negative.*

**Proof.** Pick an orthonormal basis such that matrix of $A$ in this basis is diagonal (see Theorem 2.1). To finish the proof it remains to notice that a diagonal matrix is positive definite (positive semidefinite) if and only if all its diagonal entries are positive (non-negative). □

**Corollary 3.2.** *Let $A = A^* \geq 0$ be a positive semidefinite operator. There exists a unique positive semidefinite operator $B$ such that $B^2 = A$*

Such $B$ is called (positive) square root of $A$ and is denoted as $\sqrt{A}$ or $A^{1/2}$.

**Proof.** Let us prove that $\sqrt{A}$ exists. Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be an orthonormal basis of eigenvectors of $A$, and let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the corresponding eigenvalues. Note, that since $A \geq 0$, all $\lambda_k \geq 0$.

In the basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ the matrix of $A$ is a diagonal matrix $\operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ with entries $\lambda_1, \lambda_2, \dots, \lambda_n$ on the diagonal. Define the matrix of $B$ in the same basis as $\operatorname{diag}\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n}\}$.

Clearly, $B = B^* \geq 0$ and $B^2 = A$.

To prove that such $B$ is unique, let us suppose that there exists an operator $C = C^* \geq 0$ such that $C^2 = A$. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be an orthonormal basis of eigenvectors of $C$, and let $\mu_1, \mu_2, \ldots, \mu_n$ be the corresponding eigenvalues (note that $\mu_k \geq 0\ \forall k$). The matrix of $C$ in the basis $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is a diagonal one $\operatorname{diag}\{\mu_1, \mu_2, \ldots, \mu_n\}$, and therefore the matrix of $A = C^2$ in the same basis is $\operatorname{diag}\{\mu_1^2, \mu_2^2, \ldots, \mu_n^2\}$. This implies that any eigenvalue $\lambda$ of $A$ is of form $\mu_k^2$, and, moreover, if $A\mathbf{x} = \lambda\mathbf{x}$, then $C\mathbf{x} = \sqrt{\lambda}\mathbf{x}$.

Therefore in the basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ above, the matrix of $C$ has the diagonal form $\operatorname{diag}\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\}$, i.e. $B = C$. □

**3.2. Modulus of an operator. Singular values.** Consider an operator $A : X \to Y$. Its *Hermitian square* $A^*A$ is a positive semidefinite operator acting in $X$. Indeed,

$$(A^*A)^* = A^*A^{**} = A^*A$$

and

$$(A^*A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) = \|A\mathbf{x}\|^2 \geq 0 \qquad \forall \mathbf{x} \in X.$$

Therefore, there exists a (unique) positive-semidefinite square root $R = \sqrt{A^*A}$. This operator $R$ is called the *modulus* of the operator $A$, and is often denoted as $|A|$.

The modulus of $A$ shows how "big" the operator $A$ is:

**Proposition 3.3.** *For a linear operator $A : X \to Y$*

$$\||A|\mathbf{x}\| = \|A\mathbf{x}\| \qquad \forall \mathbf{x} \in X.$$

**Proof.** For any $\mathbf{x} \in X$

$$\begin{aligned}\||A|\mathbf{x}\|^2 = (|A|\mathbf{x}, |A|\mathbf{x}) = (|A|^*|A|\mathbf{x}, \mathbf{x}) = (|A|^2\mathbf{x}, \mathbf{x}) \\ = (A^*A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) = \|A\mathbf{x}\|^2\end{aligned}$$

□

**Corollary 3.4.**

$$\operatorname{Ker} A = \operatorname{Ker} |A| = (\operatorname{Ran} |A|)^\perp.$$

**Proof.** The first equality follows immediately from Proposition 3.3, the second one follows from the identity $\operatorname{Ker} T = (\operatorname{Ran} T^*)^\perp$ ($|A|$ is self-adjoint). □

**Theorem 3.5** (Polar decomposition of an operator)**.** *Let $A : X \to X$ be an operator (square matrix). Then $A$ can be represented as*

$$A = U|A|,$$

*where $U$ is a unitary operator.*

**Remark.** The unitary operator $U$ is generally not unique. As one will see from the proof of the theorem, $U$ is unique if and only if $A$ is invertible.

**Remark.** The polar decomposition $A = U|A|$ also holds for operators $A : X \to Y$ acting from one space to another. But in this case we can only guarantee that $U$ is an *isometry* from $\operatorname{Ran}|A| = (\operatorname{Ker} A)^\perp$ to $Y$.

If $\dim X \le \dim Y$ this isometry can be extended to an isometry from the whole $X$ to $Y$ (if $\dim X = \dim Y$ this will be a unitary operator).

**Proof of Theorem 3.5.** Consider a vector $\mathbf{x} \in \operatorname{Ran}|A|$. Then vector $\mathbf{x}$ can be represented as $\mathbf{x} = |A|\mathbf{v}$ for some vector $\mathbf{v} \in X$.

Define $U_0\mathbf{x} := A\mathbf{v}$. By Proposition 3.3

$$\|U_0\mathbf{x}\| = \|A\mathbf{v}\| = \||A|\mathbf{v}\| = \|\mathbf{x}\|$$

so it looks like $U$ is an isometry from $\operatorname{Ran}|A|$ to $X$.

But first we need to prove that $U_0$ is well defined. Let $\mathbf{v}_1$ be another vector such that $\mathbf{x} = |A|\mathbf{v}_1$. But $\mathbf{x} = |A|\mathbf{v} = |A|\mathbf{v}_1$ means that $\mathbf{v} - \mathbf{v}_1 \in \operatorname{Ker}|A| = \operatorname{Ker} A$ (cf Corollary 3.4), so $A\mathbf{v} = A\mathbf{v}_1$, meaning that $U_0\mathbf{x}$ is well defined.

By the construction $A = U_0|A|$. We leave as an exercise for the reader to check that $U_0$ is a linear transformation.

To extend $U_0$ to a unitary operator $U$, let us find some unitary transformation $U_1 : \operatorname{Ker} A \to (\operatorname{Ran} A)^\perp = \operatorname{Ker} A^*$. It is always possible to do this, since for square matrices $\dim \operatorname{Ker} A = \dim \operatorname{Ker} A^*$ (the Rank Theorem).

It is easy to check that $U = U_0 + U_1$ is a unitary operator, and that $A = U|A|$. □

### 3.3. Singular values. Schmidt decomposition.

**Definition.** Eigenvalues of $|A|$ are called the *singular values* of $A$. In other words, if $\lambda_1, \lambda_2, \dots, \lambda_n$ are eigenvalues of $A^*A$ then $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n}$ are singular values of $A$.

**Remark.** Very often in the literature the singular values are defined as the non-negative square roots of the eigenvalues of $A^*A$, without any reference to the modulus $|A|$.

I consider the notion of the modulus of an operator to be an important one, so it was introduced above. However, the notion of the modulus of an operator is not required for what follows (defining the Schur and singular value decompositions). Moreover, as it will be shown below, the modulus of $A$ can be easily constructed from the singular value decomposition.

Consider an operator $A : X \to Y$, and let $\sigma_1, \sigma_2, \dots, \sigma_n$ be the singular values of $A$ counting multiplicities. Assume also that $\sigma_1, \sigma_2, \dots, \sigma_r$ are the *non-zero* singular values of $A$, counting multiplicities. This means, in particular, that $\sigma_k = 0$ for $k > r$.

By the definition of singular values the numbers $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ are eigenvalues of $A^*A$. Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be an orthonormal basis[1] of eigenvectors of $A^*A$, $A^*A\mathbf{v}_k = \sigma_k^2 \mathbf{v}_k$.

**Proposition 3.6.** *The system*

$$\mathbf{w}_k := \frac{1}{\sigma_k} A\mathbf{v}_k, \qquad k = 1, 2, \dots, r$$

*is an orthonormal system.*

**Proof.**

$$(A\mathbf{v}_j, A\mathbf{v}_k) = (A^*A\mathbf{v}_j, \mathbf{v}_k) = (\sigma_j^2 \mathbf{v}_j, \mathbf{v}_k) = \sigma_j^2 (\mathbf{v}_j, \mathbf{v}_k) = \begin{cases} 0, & j \neq k \\ \sigma_j^2, & j = k \end{cases}$$

since $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ is an orthonormal system. □

In the notation of the above proposition, the operator $A$ can be represented as

$$A = \sum_{k=1}^{r} \sigma_k \mathbf{w}_k \mathbf{v}_k^*, \tag{3.1}$$

or, equivalently

$$A\mathbf{x} = \sum_{k=1}^{r} \sigma_k (\mathbf{x}, \mathbf{v}_k) \mathbf{w}_k. \tag{3.2}$$

Indeed, we know that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthonormal basis in $X$. Then substituting $\mathbf{x} = \mathbf{v}_j$ into the right side of (3.2) we get

$$\sum_{k=1}^{r} \sigma_k (\mathbf{v}_j, \mathbf{v}_k) \mathbf{w}_k = \sigma_j (\mathbf{v}_j, \mathbf{v}_j) \mathbf{w}_j = \sigma_j \mathbf{w}_j = A\mathbf{v}_j \qquad \text{if } j = 1, 2, \dots, r,$$

and

$$\sum_{k=1}^{r} \sigma_k (\mathbf{v}_k^* \mathbf{v}_j) \mathbf{w}_k = \mathbf{0} = A\mathbf{v}_j \qquad \text{for } j > r.$$

So the operators in the left and right sides of (3.1) coincide on the basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, so they are equal.

**Definition.** The above decomposition (3.1) (or (3.2)) is called the *Schmidt decomposition* of the operator $A$.

**Remark.** Shmidt decomposition of an operator is not unique. Why?

[1] We know, that for a self-adjoint operator ($A^*A$ in our case) there exists an orthonormal basis of eigenvectors.

**Lemma 3.7.** *Let $A$ can be represented as*

$$A = \sum_{k=1}^{r} \sigma_k \mathbf{w}_k \mathbf{v}_k^*$$

*where $\sigma_k > 0$ and $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$, $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ are some orthonormal systems.*

*Then this representation gives a Schmidt decomposition of $A$.*

**Proof.** We only need to show that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ are eigenvectors of $A^*A$, $A^*A\mathbf{v}_k = \sigma_k^2 \mathbf{v}_k$. Since $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ is an orthonormal system,

$$\mathbf{w}_k^* \mathbf{w}_j = (\mathbf{w}_j, \mathbf{w}_k) = \delta_{k,j} := \begin{cases} 0, & j \neq k \\ 1, & j = k, \end{cases}$$

and therefore

$$A^*A = \sum_{k=1}^{r} \sigma_k^2 \mathbf{v}_k \mathbf{v}_k^*.$$

Since $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ is an orthonormal system

$$A^*A\mathbf{v}_j = \sum_{k=1}^{r} \sigma_k^2 \mathbf{v}_k \mathbf{v}_k^* \mathbf{v}_j = \sigma_j^2 \mathbf{v}_j$$

thus $\mathbf{v}_k$ are eigenvectors of $A^*A$. □

**Corollary 3.8.** *Let*

$$A = \sum_{k=1}^{r} \sigma_k \mathbf{w}_k \mathbf{v}_k^*$$

*be a Schmidt decomposition of $A$. Then*

$$A^* = \sum_{k=1}^{r} \sigma_k \mathbf{v}_k \mathbf{w}_k^*$$

*is a Scmidt decomposition of $A^*$*

**3.4. Matrix representation of the Shmidt decomposition. Singular value decomposition.** The Shmidt decomposition can be written in a nice matrix form. Namely, let us assume that $A : \mathbb{F}^n \to \mathbb{F}^m$, where $\mathbb{F}$ is either $\mathbb{C}$ or $\mathbb{R}$ (we can always do that by fixing orthonormal bases in $X$ and $Y$ and working with coordinates in these bases). Let $\sigma_1, \sigma_2, \dots, \sigma_r$ be non-zero singular values of $A$, and let

$$A = \sum_{k=1}^{r} \sigma_k \mathbf{w}_k \mathbf{v}_k^*$$

be a Scmidt decomposition of $A$.

As one can easily see, this equality can be rewritten as

$$A = \widetilde{W}\widetilde{\Sigma}\widetilde{V}^*, \tag{3.3}$$

where $\widetilde{\Sigma} = \operatorname{diag}\{\sigma_1, \sigma_2, \dots, \sigma_r\}$ and $\widetilde{V}$ and $\widetilde{W}$ are matrices with columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ respectively. (Can you tell what is the size of each matrix?)

Note, that since $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ are orthonormal systems, the matrices $\widetilde{V}$ and $\widetilde{W}$ are isometries. Note also that $r = \operatorname{rank} A$, see Exercise 3.1 below.

If the matrix $A$ is invertible, then $m = n = r$, the matrices $\widetilde{V}$, $\widetilde{W}$ are unitary and $\widetilde{\Sigma}$ is an invertible diagonal matrix.

It turns out that it is always possible to write a representation similar (3.3) with unitary $V$ and $W$ instead of $\widetilde{V}$ and $\widetilde{W}$, and in many situations it is more convenient to work with such a representation. To write this representation one needs first to complete the systems $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$ to orthogonal bases in $\mathbb{F}^n$ and $\mathbb{F}^m$ respectively.

Recall, that to complete, say $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ to an orthonormal basis in $\mathbb{F}^n$ one just needs to find and orthonormal basis $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ in $\operatorname{Ker} V^*$; then the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ will be an orthonormal basis in $\mathbb{F}^n$. And one can always get an orthonormal basis from an arbitrary one using Gram–Schmidt orthogonalization.

Then $A$ can be represented as

$$A = W\Sigma V^*, \tag{3.4}$$

where $V \in M_{n\times n}^{\mathbb{F}}$ and $W \in M_{m\times m}^{\mathbb{F}}$ are unitary matrices with columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$ respectively, and $\Sigma$ is a "diagonal" $m \times n$ matrix

$$\Sigma_{j,k} = \begin{cases} \sigma_k & j = k \le r : \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

In other words, to get the matrix $\Sigma$ one has to take the diagonal matrix $\operatorname{diag}\{\sigma_1, \sigma_2, \dots, \sigma_r\}$ and make it to an $m \times n$ matrix by adding extra zeroes "south and east".

**Definition 3.9.** For a matrix $A \in M_{m\times n}^{\mathbb{F}}$ (recall that here $\mathbb{F}$ is always $\mathbb{C}$ or $\mathbb{R}$) its *singular value decomposition* (SVD) is a decomposition of form (3.4), i.e. a decomposition $A = W\Sigma V^*$, where $W \in M_{n\times n}^{\mathbb{F}}$, $V \in M_{m\times m}^{\mathbb{F}}$ are unitary matrices and $\Sigma \in M_{m\times n}^{\mathbb{R}_+}$ is a "diagonal" one (meaning that $\sigma_{k,k} \ge 0$ for all $k = 1, 2, \dots, \min\{m, n\}$, and $\sigma_{j,k} = 0$ for all $j \ne k$).

The representation (3.3) is often called the *reduced* or *compact* SVD. More precisely the reduced SVD is a representation $A = \widetilde{W}\widetilde{\Sigma}\widetilde{V}^*$, where $\widetilde{\Sigma} \in M_{r\times r}^{\mathbb{R}_+}$, $r \le \min\{m, n\}$ is a diagonal matrix with strictly positive diagonal

entries, and $\widetilde{W} \in M^{\mathbb{F}}_{n\times r}$, $\widetilde{V} \in M^{\mathbb{F}}_{m\times r}$ are isometries; moreover, we require that at least one of the matrices $\widetilde{W}$ and $\widetilde{V}$ is not square.

**Remark 3.10.** It is easy to see that if $A = W\Sigma V^*$ is a singular value decomposition of $A$, then $\sigma_k := \sigma_{k,k}$ are singular values of $A$, i.e. $\sigma_k^2$ are eigenvalues of $A^*A$. Moreover, the columns $\mathbf{v}_k$ of $V$ are the corresponding eigenvectors of $A^*A$, $A^*A\mathbf{v}_k = \sigma_k^2\mathbf{v}_k$. Note also that if $\sigma_k \neq 0$ then $\mathbf{w}_k = \frac{1}{\sigma_k}A\mathbf{v}_k$.

All that means that any singular value decomposition $A = W\Sigma V^*$ can be obtained from a Schmidt decomposition (3.2) by the construction described above in this section.

The reduced singular value decomposition can be interpreted as a matrix form of the Schmidt decomposition (3.2) for a non-invertible matrix $A$. For an invertible matrix $A$ the matrix form of the Schmidt decomposition gives the singular value decomposition.

**Remark 3.11.** An alternative way to interpret the singular value decomposition $A = W\Sigma V^*$ is to say that $\Sigma$ is the matrix of $A$ in the (orthonormal) bases $\mathcal{A} = \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ and $\mathcal{B} := \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$, i.e that $\Sigma = [A]_{\mathcal{B},\mathcal{A}}$.

We will use this interpretation later.

3.4.1. *From singular value decomposition to the polar decomposition.* Note, that if we know the singular value decomposition $A = W\Sigma V^*$ of a square matrix $A$, we can write a polar decomposition of $A$:

$$A = W\Sigma V^* = (WV^*)(V\Sigma V^*) = U|A| \tag{3.6}$$

where $|A| = V\Sigma V^*$ and $U = WV^*$.

To see that this indeed give us a polar decomposition let us notice that $V\Sigma V^*$ is a self-adjoint, positive semidefinite operator and that

$$A^*A = V\Sigma W^*W\Sigma V^* = V\Sigma\Sigma V^* = V\Sigma V^*V\Sigma V^* = (V\Sigma V^*)^2.$$

So by the definition of $|A|$ as the unique positive semidefinite square root of $A^*A$, we can see that $|A| = V\Sigma V^*$. The transformation $WV^*$ is clearly unitary, as a product of two unitary transformations, so (3.6) indeed gives us a polar decomposition of $A$.

Note, that this reasoning only works for square matrices, because if $A$ is not square, then the product $V\Sigma$ is not defined (dimensions do not match, can you see how?).

**Exercises.**

**3.1.** Show that the number of non-zero singular values of a matrix $A$ coincides with its rank.

**3.2.** Find Schmidt decompositions $A = \sum_{k=1}^{r} s_k \mathbf{w}_k \mathbf{v}_k^*$ for the following matrices $A$:

$$\begin{pmatrix} 2 & 3 \\ 0 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 7 & 1 \\ 0 & 0 \\ 5 & 5 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

**3.3.** Let $A$ be an invertible matrix, and let $A = W\Sigma V^*$ be its singular value decomposition. Find a singular value decomposition for $A^*$ and $A^{-1}$.

**3.4.** Find singular value decomposition $A = W\Sigma V^*$ where $V$ and $W$ are unitary matrices for the following matrices:

a) $A = \begin{pmatrix} -3 & 1 \\ 6 & -2 \\ 6 & -2 \end{pmatrix}$;

b) $A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$.

**3.5.** Find singular value decomposition of the matrix

$$A = \begin{pmatrix} 2 & 3 \\ 0 & 2 \end{pmatrix}$$

Use it to find

a) $\max_{\|\mathbf{x}\|\le 1} \|A\mathbf{x}\|$ and the vectors where the maximum is attained;

b) $\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ and the vectors where the minimum is attained;

c) the image $A(B)$ of the closed unit ball in $\mathbb{R}^2$, $B = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \le 1\}$. Describe $A(B)$ geometrically.

**3.6.** Show that for a square matrix $A$, $|\det A| = \det |A|$.

**3.7.** True or false

a) Singular values of a matrix are also eigenvalues of the matrix.

b) Singular values of a matrix $A$ are eigenvalues of $A^*A$.

c) Is $s$ is a singular value of a matrix $A$ and $c$ is a scalar, then $|c|s$ is a singular value of $cA$.

d) The singular values of any linear operator are non-negative.

e) Singular values of a self-adjoint matrix coincide with its eigenvalues.

**3.8.** Let $A$ be an $m \times n$ matrix. Prove that *non-zero* eigenvalues of the matrices $A^*A$ and $AA^*$ (counting multiplicities) coincide.

Can you say when zero eigenvalue of $A^*A$ and zero eigenvalue of $AA^*$ have the same multiplicity?

**3.9.** Let $s$ be the largest singular value of an operator $A$, and let $\lambda$ be the eigenvalue of $A$ with largest absolute value. Show that $|\lambda| \le s$.

**3.10.** Show that the rank of a matrix is the number of its non-zero singular values (counting multiplicities).

**3.11.** Show that the operator norm of a matrix $A$ coincides with its Frobenius norm if and only if the matrix has rank one. **Hint:** The previous problem might help.

**3.12.** For the matrix $A$

$$A = \begin{pmatrix} 2 & -3 \\ 0 & 2 \end{pmatrix},$$

describe the inverse image of the unit ball, i.e. the set of all $\mathbf{x} \in \mathbb{R}^2$ such that $\|A\mathbf{x}\| \leq 1$. Use singular value decomposition.

## 4. Applications of the singular value decomposition.

As we discussed above, the singular value decomposition is simply diagonalization with respect to two different orthonormal bases. Since we have two different bases here, we cannot say much about spectral properties of an operator from its singular value decomposition. For example, the diagonal entries of $\Sigma$ in the singular value decomposition (3.5) are not the eigenvalues of $A$. Note, that for $A = W\Sigma V^*$ as in (3.5) we generally have $A^n \neq W\Sigma^n V^*$, so this diagonalization does not help us in computing functions of a matrix.

However, as the examples below show, singular values tell us a lot about so-called *metric properties* of a linear transformation.

Final remark: performing singular value decomposition requires finding eigenvalues and eigenvectors of the Hermitian (self-adjoint) matrix $A^*A$. To find eigenvalues we usually computed characteristic polynomial, found its roots, and so on... This looks like quite a complicated process, especially if one takes into account that there is no formula for finding roots of polynomials of degree 5 and higher.

However, there are very effective numerical methods of find eigenvalues and eigenvectors of a hermitian matrix up to any given precision. These methods do not involve computing the characteristic polynomial and finding its roots. They compute approximate eigenvalues and eigenvectors directly by an iterative procedure. Because a Hermitian matrix has an orthogonal basis of eigenvectors, these methods work extremely well.

We will not discuss these methods here, it goes beyond the scope of this book. However, you should believe me that there are very effective numerical methods for computing eigenvalues and eigenvectors of a Hermitian matrix and for finding the singular value decomposition. These methods are extremely effective, and just a little more computationally intensive than solving a linear system.

**4.1. Image of the unit ball.** Consider for example the following problem: let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation, and let $B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \le 1\}$ be the closed *unit ball* in $\mathbb{R}^n$. We want to describe $A(B)$, i.e. we want to find out how the unit ball is transformed under the linear transformation.

Let us first consider the simplest case when $A$ is a diagonal matrix $A = \operatorname{diag}\{\sigma_1, \sigma_2, \dots, \sigma_n\}$, $\sigma_k > 0$, $k = 1, 2, \dots, n$. Then for $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ and $(y_1, y_2, \dots, y_n)^T = \mathbf{y} = A\mathbf{x}$ we have $y_k = \sigma_k x_k$ (equivalently, $x_k = y_k/\sigma_k$) for $k = 1, 2, \dots, n$, so

$$\mathbf{y} = (y_1, y_2, \dots, y_n)^T = A\mathbf{x} \qquad \text{for } \|\mathbf{x}\| \le 1,$$

if and only if the coordinates $y_1, y_2, \dots, y_n$ satisfy the inequality

$$\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2} + \cdots + \frac{y_n^2}{\sigma_n^2} = \sum_{k=1}^{n} \frac{y_k^2}{\sigma_k^2} \le 1$$

(this is simply the inequality $\|\mathbf{x}\|^2 = \sum_k |x_k|^2 \le 1$).

The set of points in $\mathbb{R}^n$ satisfying the above inequalities is called an ellipsoid. If $n = 2$ it is an ellipse with half-axes $\sigma_1$ and $\sigma_2$, for $n = 3$ it is an ellipsoid with half-axes $\sigma_1, \sigma_2$ and $\sigma_2$. In $\mathbb{R}^n$ the geometry of this set is also easy to visualize, and we call that set an ellipsoid with half axes $\sigma_1, \sigma_2, \dots, \sigma_n$. The vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ or, more precisely the corresponding lines are called the *principal axes* of the ellipsoid.

The singular value decomposition essentially says that any operator in an inner product space is diagonal with respect to a pair of orthonormal bases, see Remark 3.11. Namely, consider the orthogonal bases $\mathcal{A} = \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ and $\mathcal{B} = \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ from the singular value decomposition (3.1). Then the matrix of $A$ in these bases is diagonal

$$[A]_{\mathcal{B},\mathcal{A}} = \operatorname{diag}\{\sigma_n : n = 1, 2, \dots, n\}.$$

Assuming that all $\sigma_k > 0$ and essentially repeating the above reasoning, it is easy to show that any point $\mathbf{y} = A\mathbf{x} \in A(B)$ if and only if it satisfies the inequality

$$\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2} + \cdots + \frac{y_n^2}{\sigma_n^2} = \sum_{k=1}^{n} \frac{y_k^2}{\sigma_k^2} \le 1.$$

where $y_1, y_2, \dots, y_n$ are coordinates of $\mathbf{y}$ in the orthonormal basis $\mathcal{B} = \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$, not in the standard one. Similarly, $(x_1, x_2, \dots, x_n)^T = [\mathbf{x}]_{\mathcal{A}}$.

But that is essentially the same ellipsoid as before, only "rotated" (with different but still orthogonal principal axes)!

There is also an alternative explanation which is presented below.

Consider the general case, when the matrix $A$ is not necessarily square, and (or) not all singular values are non-zero. Consider first the case of a "diagonal" matrix $\Sigma$ of form (3.5). It is easy to see that the image $\Sigma B$ of

the unit ball $B$ is the ellipsoid (not in the whole space but in the $\operatorname{Ran}\Sigma$) with half axes $\sigma_1, \sigma_2, \dots, \sigma_r$.

Consider now the general case, $A = W\Sigma V^*$, where $V$, $W$ are unitary operators. Unitary transformations do not change the unit ball (because they preserve norm), so $V^*(B) = B$. We know that $\Sigma(B)$ is an ellipsoid in $\operatorname{Ran}\Sigma$ with half-axes $\sigma_1, \sigma_2, \dots, \sigma_r$. Unitary transformations do not change geometry of objects, so $W(\Sigma(B))$ is also an ellipsoid with the same half-axes. It is not hard to see from the decomposition $A = W\Sigma V^*$ (using the fact that both $W$ and $V^*$ are invertible) that $W$ transforms $\operatorname{Ran}\Sigma$ to $\operatorname{Ran}A$, so we can conclude:

| |
|---|
| the image $A(B)$ of the closed unit ball $B$ is an ellipsoid in $\operatorname{Ran}A$ with half axes $\sigma_1, \sigma_2, \dots, \sigma_r$. Here $r$ is the number of non-zero singular values, i.e. the rank of $A$. |

**4.2. Operator norm of a linear transformation.** Given a linear transformation $A : X \to Y$ let us consider the following optimization problem: find the maximum of $\|A\mathbf{x}\|$ on the closed unit ball $B = \{\mathbf{x} \in X : \|\mathbf{x}\| \le 1\}$.

Again, singular value decomposition allows us to solve the problem. For a diagonal matrix $A$ with non-negative entries the maximum is exactly maximal diagonal entry. Indeed, let $s_1, s_2, \dots, s_r$ be non-zero diagonal entries of $A$ and let $s_1$ be the maximal one. Since for $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$

$$A\mathbf{x} = \sum_{k=1}^{r} s_k x_k \mathbf{e}_k, \tag{4.1}$$

we can conclude that

$$\|A\mathbf{x}\|^2 = \sum_{k=1}^{r} s_k^2 |x_k|^2 \le s_1^2 \sum_{k=1}^{r} |x_k|^2 = s_1^2 \cdot \|\mathbf{x}\|^2,$$

so $\|A\mathbf{x}\| \le s_1\|\mathbf{x}\|$. On the other hand, $\|A\mathbf{e}_1\| = \|s_1\mathbf{e}_1\| = s_1\|\mathbf{e}_1\|$, so indeed $s_1$ is the maximum of $\|A\mathbf{x}\|$ on the closed unit ball $B$. Note, that in the above reasoning we did not assume that the matrix $A$ is square; we only assumed that all entries outside the "main diagonal" are 0, so formula (4.1) holds.

To treat the general case let us consider the singular value decomposition (3.5), $A = W\Sigma V$, where $W$, $V$ are unitary operators, and $\Sigma$ is the diagonal matrix with non-negative entries. Since unitary transformations do not change the norm, one can conclude that the maximum of $\|A\mathbf{x}\|$ on the unit ball $B$ is the maximal diagonal entry of $\Sigma$ i.e. that

| |
|---|
| the maximum of $\|A\mathbf{x}\|$ on the unit ball $B$ is the maximal singular value of $A$. |

**Definition.** The quantity $\max\{\|A\mathbf{x}\| : \mathbf{x} \in X, \|\mathbf{x}\| \leq 1\}$ is called the operator norm of $A$ and denoted $\|A\|$.

It is an easy exercise to see that $\|A\|$ satisfies all properties of the norm:

1. $\|\alpha A\| = |\alpha| \cdot \|A\|$;
2. $\|A + B\| \leq \|A\| + \|B\|$;
3. $\|A\| \geq 0$ for all $A$;
4. $\|A\| = 0$ if and only if $A = \mathbf{0}$,

so it is indeed a norm on a space of linear transformations from from $X$ to $Y$.

One of the main properties of the operator norm is the inequality

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

which follows easily from the homogeneity of the norm $\|\mathbf{x}\|$.

In fact, it can be shown that the operator norm $\|A\|$ is the best (smallest) number $C \geq 0$ such that

$$\|A\mathbf{x}\| \leq C\|\mathbf{x}\| \qquad \forall \mathbf{x} \in X.$$

This is often used as a definition of the operator norm.

On the space of linear transformations we already have one norm, the Frobenius, or Hilbert-Schmidt norm $\|A\|_2$,

$$\|A\|_2^2 = \operatorname{trace}(A^*A).$$

So, let us investigate how these two norms compare.

Let $s_1, s_2, \ldots, s_r$ be non-zero singular values of $A$ (counting multiplicities), and let $s_1$ be the largest eigenvalues. Then $s_1^2, s_2^2, \ldots, s_r^2$ are non-zero eigenvalues of $A^*A$ (again counting multiplicities). Recalling that the trace equals the sum of the eigenvalues we conclude that

$$\|A\|_2^2 = \operatorname{trace}(A^*A) = \sum_{k=1}^{r} s_k^2.$$

On the other hand we know that the operator norm of $A$ equals its largest singular value, i.e. $\|A\| = s_1$. So we can conclude that $\|A\| \leq \|A\|_2$, i.e. that

| the operator norm of a matrix cannot be more than its Frobenius norm. |
|---|

This statement also admits a direct proof using the Cauchy–Schwarz inequality, and such a proof is presented in some textbooks. The beauty of the proof we presented here is that it does not require any computations and illuminates the reasons behind the inequality.

**4.3. Condition number of a matrix.** Suppose we have an invertible matrix $A$ and we want to solve the equation $A\mathbf{x} = \mathbf{b}$. The solution, of course, is given by $\mathbf{x} = A^{-1}\mathbf{b}$, but we want to investigate what happens if we know the data only approximately.

That happens in the real life, when the data is obtained, for example by some experiments. But even if we have exact data, round-off errors during computations by a computer may have the same effect of distorting the data.

Let us consider the simplest model, suppose there is a small error in the right side of the equation. That means, instead of the equation $A\mathbf{x} = \mathbf{b}$ we are solving

$$A\mathbf{x} = \mathbf{b} + \Delta\mathbf{b},$$

where $\Delta\mathbf{b}$ is a small perturbation of the right side $\mathbf{b}$.

So, instead of the exact solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$ we get the approximate solution $\mathbf{x}+\Delta\mathbf{x}$ of $A(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}+\Delta\mathbf{b}$. We are assuming that $A$ is invertible, so $\Delta\mathbf{x} = A^{-1}\Delta\mathbf{b}$.

We want to know how big is the relative error in the solution $\|\Delta\mathbf{x}\|/\|\mathbf{x}\|$ in comparison with the relative error in the right side $\|\Delta\mathbf{b}\|/\|\mathbf{b}\|$. It is easy to see that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|A^{-1}\Delta\mathbf{b}\|}{\|\mathbf{x}\|} = \frac{\|A^{-1}\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\frac{\|\mathbf{b}\|}{\|\mathbf{x}\|} = \frac{\|A^{-1}\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Since $\|A^{-1}\Delta\mathbf{b}\| \le \|A^{-1}\| \cdot \|\Delta\mathbf{b}\|$ and $\|A\mathbf{x}\| \le \|A\| \cdot \|\mathbf{x}\|$ we can conclude that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

The quantity $\|A\|\cdot\|A^{-1}\|$ is called the *condition number* of the matrix $A$. It estimates how the relative error in the solution $\mathbf{x}$ depends on the relative error in the right side $\mathbf{b}$.

Let us see how this quantity is related to singular values. Let the numbers $s_1, s_2, \dots, s_n$ be the singular values of $A$, and let us assume that $s_1$ is the largest singular value and $s_n$ is the smallest. We know that the (operator) norm of an operator equals its largest singular value, so

$$\|A\| = s_1, \qquad \|A^{-1}\| = \frac{1}{s_n},$$

so

$$\|A\| \cdot \|A^{-1}\| = \frac{s_1}{s_n}.$$

In other words

| The condition number of a matrix equals to the ratio of the largest and the smallest singular values. |
|---|

We deduced above that $\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$. It is not hard to see that this estimate is sharp, i.e. that it is possible to pick the right side $\mathbf{b}$ and the error $\Delta\mathbf{b}$ such that we have equality

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} = \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

We just put $\mathbf{b} = \mathbf{w}_1$ and $\Delta\mathbf{b} = \alpha\mathbf{w}_n$, where $\mathbf{w}_1$ and $\mathbf{w}_n$ are respectively the first and the last column of the matrix $W$ in the singular value decomposition $A = W\Sigma V^*$, and $\alpha \neq 0$ is an arbitrary scalar. Here, as usual, the singular values are assumed to be in non-increasing order $s_1 \ge s_2 \ge \ldots \ge s_n$, so $s_1$ is the largest and $s_n$ is the smallest eigenvalue.

We leave the details as an exercise for the reader.

A matrix is called *well conditioned* if its condition number is not too big. If the condition number is big, the matrix is called ill conditioned. What is "big" here depends on the problem: with what precision you can find your right side, what precision is required for the solution, etc.

**4.4. Effective rank of a matrix.** Theoretically, the rank of a matrix is easy to compute: one just needs to row reduce matrix and count pivots. However, in practical applications not everything is so easy. The main reason is that very often we do not know the exact matrix, we only know its approximation up to some precision.

Moreover, even if we know the exact matrix, most computer programs introduce round-off errors in the computations, so effectively we cannot distinguish between a zero pivot and a very small pivot.

A simple naïve idea of working with round-off errors is as follows. When computing the rank (and other objects related to it, like column space, kernel, etc) one simply sets up a tolerance (some small number) and if the pivot is smaller than the tolerance, count it as zero. The advantage of this approach is its simplicity, since it is very easy to program. However, the main disadvantage is that is is impossible to see what the tolerance is responsible for. For example, what do we lose is we set the tolerance equal to $10^{-6}$? How much better will $10^{-8}$ be?

While the above approach works well for well conditioned matrices, it is not very reliable in the general case.

A better approach is to use singular values. It requires more computations, but gives much better results, which are easier to interpret. In this approach we also set up some small number as a tolerance, and then perform singular value decomposition. Then we simply treat singular values smaller than the tolerance as zero. The advantage of this approach is that we can see what we are doing. The singular values are the half-axes of the

ellipsoid $A(B)$ ($B$ is the closed unit ball), so by setting up the tolerance we just deciding how "thin" the ellipsoid should be to be considered "flat".

**4.5. Moore–Penrose (pseudo)inverse.** As we discussed in Section 4 of Chapter 5 above, the least square solution gives us, in the case when an equation $A\mathbf{x} = \mathbf{b}$ does not have a solutions, the "next best thing" (and gives us the solution of $A\mathbf{x} = \mathbf{b}$ when it exists).

Note, that the question of uniqueness is not addressed by the least square solution: a solution of the normal equation $A^*A\mathbf{x} = A^*\mathbf{b}$ does not have to be unique. A natural distinguished solution would be a solution of minimal norm; such a solution is indeed unique, and can be obtained by taking an arbitrary solution and then taking its projection onto $(\operatorname{Ker} A^*A)^\perp = (\operatorname{Ker} A)^\perp$, see problems 4.5 and 4.6 in Chapter 5.

It is not hard to see that if $A = \widetilde{W}\widetilde{\Sigma}\widetilde{V}^*$ is a *reduced* singular value decomposition of $A$, then the minimal norm least square solution $\mathbf{x}_0$ is given by

$$\mathbf{x}_0 = \widetilde{V}\widetilde{\Sigma}^{-1}\widetilde{W}^*\mathbf{b}. \tag{4.2}$$

Indeed, $\mathbf{x}_0$ is a least square solution of $A\mathbf{x} = \mathbf{b}$ (i.e. a solution of $A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}$):

$$A\mathbf{x}_0 = \widetilde{W}\widetilde{\Sigma}\widetilde{V}\widetilde{V}^*\widetilde{\Sigma}^{-1}\widetilde{W}^*\mathbf{b} = \widetilde{W}\widetilde{\Sigma}\widetilde{\Sigma}^{-1}\widetilde{W}^*\mathbf{b} = \widetilde{W}\widetilde{W}^*\mathbf{b} = P_{\operatorname{Ran} A}\mathbf{b};$$

in the last equality in the chain we used the fact that $\widetilde{W}\widetilde{W}^* = P_{\operatorname{Ran}\widetilde{W}}$ ($P_{\operatorname{Ran}\widetilde{W}} = \widetilde{W}(\widetilde{W}^*\widetilde{W})^{-1}\widetilde{W}^* = \widetilde{W}\widetilde{W}^*$) and that $\operatorname{Ran}\widetilde{W} = \operatorname{Ran} A$ (see Problem 4.4 below).

The general solution of $A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}$ is given by

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{y}, \qquad y \in \operatorname{Ker} A,$$

so $\mathbf{x}_0$ is indeed a unique minimal norm solution of $A\mathbf{x} = P_{\operatorname{Ran} A}\mathbf{b}$, or equivalently, the minimal norm least square solution of $A\mathbf{x} = \mathbf{b}$.

**Definition 4.1.** The operator $A^+ := \widetilde{V}\widetilde{\Sigma}^{-1}\widetilde{W}^*$, where $A = \widetilde{W}\widetilde{\Sigma}\widetilde{V}^*$ is a *reduced* singular value decomposition of $A$, is called the *Moore–Penrose inverse* (or *Moore–Penrose pseudoinverse*) of the operator $A$. In other words, the *Moore–Penrose inverse* is the operator giving the unique least square solution of $A\mathbf{x} = \mathbf{b}$.

**Remark 4.2.** In the literature the Moore–Penrose inverse is usually defined as a matrix $A^+$ such that

1. $AA^+A = A$;
2. $A^+AA^+ = A^+$;
3. $(AA^+)^* = AA^+$;

4. $(A^+A)^* = A^+A$.

It is very easy to check that the operator $A^+ := \widetilde{V}\widetilde{\Sigma}^{-1}\widetilde{W}^*$ satisfies properties 1–4 above.

It is also possible (although a bit harder) to show that an operator $A^+$ satisfying properties 1–4 is unique. Indeed, right and left multiplying identity 1 by $A^+$, we get that $(A^+A)^2 = A^+A$ and $(AA^+)^2 = AA^+$; together with properties 3 and 4 this means that $A^+A$ and $AA^+$ ore orthogonal projections (see Problem 5.6 in Capter 5).

Trivially, $\operatorname{Ker} A \subset \operatorname{Ker} A^+A$. On the other hand, identity 1 implies that $\operatorname{Ker} A^+A \subset \operatorname{Ker} A$ (why?), so $\operatorname{Ker} A^+A = \operatorname{Ker} A$. But this means that $A^+A$ is the orthogonal projection onto $(\operatorname{Ker} A)^\perp = \operatorname{Ran} A^*$,

$$A^+A = P_{\operatorname{Ran} A^*}.$$

Property 1 also implies that $AA^+\mathbf{y} = \mathbf{y}$ for all $\mathbf{y} \in \operatorname{Ran} A$. Since $AA^+$ is an orthogonal projection, we conclude that $\operatorname{Ran} A \subset \operatorname{Ran} AA^+$. The opposite inclusion $\operatorname{Ran} AA^+ \subset \operatorname{Ran} A$ is trivial, so $AA^+$ is the orthogonal projection onto $\operatorname{Ran} A$,

$$AA^+ = P_{\operatorname{Ran} A}.$$

Knowing $A^+A$ and $AA^+$ we can rewrite property 2 as

$$P_{\operatorname{Ran} A^*}A^+ = A^+ \qquad \text{or} \qquad A^+P_{\operatorname{Ran} A} = A^+.$$

Combining the above identities we get

$$P_{\operatorname{Ran} A^*}A^+P_{\operatorname{Ran} A} = A^+.$$

Finally, for any $\mathbf{b}$ in the target space of $A$

$$\mathbf{x}_0 := A^+\mathbf{b} = P_{\operatorname{Ran} A^*}A^+\mathbf{b} \in \operatorname{Ran} A^*$$

and

$$A\mathbf{x}_0 = AA^+\mathbf{b} = P_{\operatorname{Ran} A}\mathbf{b},$$

i.e. $\mathbf{x}_0$ is a least square solution of $A\mathbf{x} = \mathbf{b}$. Since $\mathbf{x}_0 \in \operatorname{Ran} A^* = (\operatorname{Ker} A)^\perp$, $\mathbf{x}_0$ is, as we discussed above, the least square solution of minimal norm. But, as we had shown before, such minimal norm solution is given by (4.2), so $A^+ = \widetilde{V}\widetilde{\Sigma}^{-1}\widetilde{W}^*$. □

**Exercises.**

**4.1.** Find norms and condition numbers for the following matrices:

a) $A = \begin{pmatrix} 4 & 0 \\ 1 & 3 \end{pmatrix}$.

b) $A = \begin{pmatrix} 5 & 3 \\ -3 & 3 \end{pmatrix}$.

For the matrix $A$ from part a) present an example of the right side $\mathbf{b}$ and the error $\Delta\mathbf{b}$ such that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|};$$

here $A\mathbf{x} = \mathbf{b}$ and $A\Delta\mathbf{x} = \Delta\mathbf{b}$.

**4.2.** Let $A$ be a normal operator, and let $\lambda_1, \lambda_2, \dots, \lambda_n$ be its eigenvalues (counting multiplicities). Show that singular values of $A$ are $|\lambda_1|, |\lambda_2|, \dots, |\lambda_n|$.

**4.3.** Find singular values, norm and condition number of the matrix

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

You can do this problem practically without any computations, if you use the previous problem and can answer the following questions:

a) What are singular values (eigenvalues) of an orthogonal projection $P_E$ onto some subspace $E$?

b) What is the matrix of the orthogonal projection onto the subspace spanned by the vector $(1, 1, 1)^T$?

c) How the eigenvalues of the operators $T$ and $aT + bI$, where $a$ and $b$ are scalars, are related?

Of course, you can also just honestly do the computations.

**4.4.** Let $A = \widetilde{W}\widetilde{\Sigma}\widetilde{V}^*$ be a *reduced* singular value decomposition of $A$. Show that $\operatorname{Ran} A = \operatorname{Ran} \widetilde{W}$, and then by taking adjoint that $\operatorname{Ran} A^* = \operatorname{Ran} \widetilde{V}$.

**4.5.** Write a formula for the Moore–Penrose inverse $A^+$ in terms of the singular value decomposition $A = W\Sigma V^*$.

**4.6.** *Tychonov's regularization:* Prove that the Moore–Penrose inverse $A^+$ can be computed as the limits

$$A^+ = \lim_{\varepsilon\to 0+} (A^*A + \varepsilon I)^{-1}A^* = \lim_{\varepsilon\to 0+} A^*(AA^* + \varepsilon I)^{-1}.$$

## 5. Structure of orthogonal matrices

An orthogonal matrix $U$ with $\det U = 1$ is often called a *rotation*. The theorem below explains this name.

**Theorem 5.1.** *Let $U$ be an orthogonal operator in $\mathbb{R}^n$ and let $\det U = 1$.*[2] *Then there exists an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ such that the matrix*

[2] For an orthogonal matrix $U$ $\det U = \pm 1$.

*of $U$ in this basis has the block diagonal form*

$$\begin{pmatrix} R_{\varphi_1} & & & & 0 \\ & R_{\varphi_2} & & & \\ & & \ddots & & \\ & & & R_{\varphi_k} & \\ 0 & & & & I_{n-2k} \end{pmatrix},$$

*where $R_{\varphi_k}$ are 2-dimensional rotations,*

$$R_{\varphi_k} = \begin{pmatrix} \cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & \cos\varphi_k \end{pmatrix}$$

*and $I_{n-2k}$ stands for the identity matrix of size $(n-2k)\times(n-2k)$.*

**Proof.** We know that if $p$ is a polynomial with real coefficient and $\lambda$ is its complex root, $p(\lambda)=0$, then $\overline{\lambda}$ is a root of $p$ as well, $p(\overline{\lambda})=0$ (this can easily be checked by plugging $\overline{\lambda}$ into $p(z)=\sum_{k=0}^{n} a_k z^k$).

Therefore, all complex eigenvalues of a real matrix $A$ can be split into pairs $\lambda_k$, $\overline{\lambda}_k$.

We know, that eigenvalues of a unitary matrix have absolute value 1, so all complex eigenvalues of $A$ can be written as $\lambda_k = \cos\alpha_k + i\sin\alpha_k$, $\overline{\lambda}_k = \cos\alpha_k - i\sin\alpha_k$.

Fix a pair of complex eigenvalues $\lambda$ and $\overline{\lambda}$, and let $\mathbf{u}\in\mathbb{C}^n$ be the eigenvector of $U$, $U\mathbf{u}=\lambda\mathbf{u}$. Then $U\overline{\mathbf{u}}=\overline{\lambda}\overline{\mathbf{u}}$. Now, split $\mathbf{u}$ into real and imaginary parts, i.e. define

$$\mathbf{x} := \operatorname{Re}\mathbf{u} = (\mathbf{u}+\overline{\mathbf{u}})/2, \qquad \mathbf{y} = \operatorname{Im}\mathbf{u} = (\mathbf{u}-\overline{\mathbf{u}})/(2i),$$

so $\mathbf{u}=\mathbf{x}+i\mathbf{y}$ (note, that $\mathbf{x},\mathbf{y}$ are real vectors, i.e. vectors with real entries). Then

$$U\mathbf{x} = U\frac{1}{2}(\mathbf{u}+\overline{\mathbf{u}}) = \frac{1}{2}(\lambda\mathbf{u}+\overline{\lambda}\overline{\mathbf{u}}) = \operatorname{Re}(\lambda\mathbf{u}).$$

Similarly,

$$U\mathbf{y} = \frac{1}{2i}U(\mathbf{u}-\overline{\mathbf{u}}) = \frac{1}{2i}(\lambda\mathbf{u}-\overline{\lambda}\overline{\mathbf{u}}) = \operatorname{Im}(\lambda\mathbf{u}).$$

Since $\lambda=\cos\alpha+i\sin\alpha$, we have

$$\lambda\mathbf{u} = (\cos\alpha+i\sin\alpha)(\mathbf{x}+i\mathbf{y}) = ((\cos\alpha)\mathbf{x}-(\sin\alpha)\mathbf{y})+i((\cos\alpha)\mathbf{y}+(\sin\alpha)\mathbf{x}).$$

so

$$U\mathbf{x} = \operatorname{Re}(\lambda\mathbf{u}) = (\cos\alpha)\mathbf{x}-(\sin\alpha)\mathbf{y}, \qquad U\mathbf{y} = \operatorname{Im}(\lambda\mathbf{u}) = (\cos\alpha)\mathbf{y}+(\sin\alpha)\mathbf{x}.$$

In other word, $U$ leaves the 2-dimensional subspace $E_\lambda$ spanned by the vectors $\mathbf{x},\mathbf{y}$ invariant and the matrix of the restriction of $U$ onto this subspace

is the rotation matrix

$$R_{-\alpha} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}.$$

Note, that the vectors $\mathbf{u}$ and $\overline{\mathbf{u}}$ (eigenvectors of a unitary matrix, corresponding to different eigenvalues) are orthogonal, so by the Pythagorean Theorem

$$\|\mathbf{x}\| = \|\mathbf{y}\| = \frac{\sqrt{2}}{2}\|\mathbf{u}\|.$$

It is easy to check that $\mathbf{x} \perp \mathbf{y}$, so $\mathbf{x}, \mathbf{y}$ is an orthogonal basis in $E_\lambda$. If we multiply each vector in the basis $\mathbf{x}, \mathbf{y}$ by the same non-zero number, we do not change matrices of linear transformations, so without loss of generality we can assume that $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ i.e. that $\mathbf{x}, \mathbf{y}$ is an orthogonal basis in $E_\lambda$.

Let us complete the orthonormal system $\mathbf{v}_1 = \mathbf{x}, \mathbf{v}_2 = \mathbf{y}$ to an orthonormal basis in $\mathbb{R}^n$. Since $UE_\lambda \subset E_\lambda$, i.e. $E_\lambda$ is an invariant subspace of $U$, the matrix of $U$ in this basis has the block triangular form

$$\left(\begin{array}{c|c} R_{-\alpha} & * \\ \hline \mathbf{0} & U_1 \end{array}\right)$$

where $\mathbf{0}$ stands for the $(n-2) \times 2$ block of zeroes.

Since the rotation matrix $R_{-\alpha}$ is invertible, we have $UE_\lambda = E_\lambda$. Therefore

$$U^* E_\lambda = U^{-1} E_\lambda = E_\lambda,$$

so the matrix of $U$ in the basis we constructed is in fact block diagonal,

$$\left(\begin{array}{c|c} R_{-\alpha} & \mathbf{0} \\ \hline \mathbf{0} & U_1 \end{array}\right).$$

Since $U$ is unitary

$$I = U^*U = \left(\begin{array}{c|c} I & \mathbf{0} \\ \hline \mathbf{0} & U_1^* U_1 \end{array}\right),$$

so, since $U_1$ is square, it is also unitary.

If $U_1$ has complex eigenvalues we can apply the same procedure to decrease its size by 2 until we are left with a block that has only real eigenvalues. Real eigenvalues can be only $+1$ or $-1$, so in some orthonormal basis

the matrix of $U$ has the form

$$\begin{pmatrix} R_{-\alpha_1} & & & & & 0 \\ & R_{-\alpha_2} & & & & \\ & & \ddots & & & \\ & & & R_{-\alpha_d} & & \\ & & & & -I_r & \\ 0 & & & & & I_l \end{pmatrix};$$

here $I_r$ and $I_l$ are identity matrices of size $r \times r$ and $l \times l$ respectively. Since $\det U = 1$, the multiplicity of the eigenvalue $-1$ (i.e. $r$) must be even.

Note, that the $2 \times 2$ matrix $-I_2$ can be interpreted as the rotation through the angle $\pi$. Therefore, the above matrix has the form given in the conclusion of the theorem with $\varphi_k = -\alpha_k$ or $\varphi_k = \pi$ □

Let us give a different interpretation of Theorem 5.1. Define $T_j$ to be a rotation thorough $\varphi_j$ in the plane spanned by the vectors $\mathbf{v}_j$, $\mathbf{v}_{j+1}$. Then Theorem 5.1 simply says that $U$ is the composition of the rotations $T_j$, $j = 1, 2, \dots, k$. Note, that because the rotations $T_j$ act in mutually orthogonal planes, they commute, i.e. it does not matter in what order we take the composition. So, the theorem can be interpreted as follows:

| Any rotation in $\mathbb{R}^n$ can be represented as a composition of at most $n/2$ commuting planar rotations. |
|---|

If an orthogonal matrix has determinant $-1$, its structure is described by the following theorem.

**Theorem 5.2.** *Let $U$ be an orthogonal operator in $\mathbb{R}^n$, and let $\det U = -1$. Then there exists an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ such that the matrix of $U$ in this basis has block diagonal form*

$$\begin{pmatrix} R_{\varphi_1} & & & & & 0 \\ & R_{\varphi_2} & & & & \\ & & \ddots & & & \\ & & & R_{\varphi_k} & & \\ & & & & I_r & \\ 0 & & & & & -1 \end{pmatrix},$$

*where $r = n - 2k - 1$ and $R_{\varphi_k}$ are 2-dimensional rotations,*

$$R_{\varphi_k} = \begin{pmatrix} \cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & \cos\varphi_k \end{pmatrix}$$

*and $I_{n-2k}$ stands for the identity matrix of size $(n-2k) \times (n-2k)$.*

We leave the proof as an exercise for the reader. The modification that one should make to the proof of Theorem 5.1 are pretty obvious.

Note, that it follows from the above theorem that an orthogonal $2 \times 2$ matrix $U$ with determinant $-1$ is always a reflection.

Let us now fix an orthonormal basis, say the standard basis in $\mathbb{R}^n$. We call an *elementary rotation*[3] a rotation in the $x_j$-$x_k$ plane, i.e. a linear transformation which changes only the coordinates $x_j$ and $x_k$, and it acts on these two coordinates as a plane rotation.

**Theorem 5.3.** *Any rotation $U$ (i.e. an orthogonal transformation $U$ with $\det U = 1$) can be represented as a product at most $n(n-1)/2$ elementary rotations.*

To prove the theorem we will need the following simple lemmas.

**Lemma 5.4.** *Let $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$. There exists a rotation $R_\alpha$ of $\mathbb{R}^2$ which moves the vector $\mathbf{x}$ to the vector $(a, 0)^T$, where $a = \sqrt{x_1^1 + x_2^2}$.*

The proof is elementary, and we leave it as an exercise for the reader. One can just draw a picture or/and write a formula for $R_\alpha$.

**Lemma 5.5.** *Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$. There exist $n-1$ elementary rotations $R_1, R_2, \dots, R_{n-1}$ such that $R_{n-1} \dots, R_2 R_1 \mathbf{x} = (a, 0, 0, \dots, 0)^T$, where $a = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$.*

**Proof.** The idea of the proof of the lemma is very simple. We use an elementary rotation $R_1$ in the $x_{n-1}$-$x_n$ plane to "kill" the last coordinate of $\mathbf{x}$ (Lemma 5.4 guarantees that such rotation exists). Then use an elementary rotation $R_2$ in $x_{n-2}$-$x_{n-1}$ plane to "kill" the coordinate number $n-1$ of $R_1\mathbf{x}$ (the rotation $R_2$ does not change the last coordinate, so the last coordinate of $R_2 R_1 \mathbf{x}$ remains zero), and so on...

For a formal proof we will use induction in $n$. The case $n = 1$ is trivial, since any vector in $\mathbb{R}^1$ has the desired form. The case $n = 2$ is treated by Lemma 5.4.

Assuming now that Lemma is true for $n-1$, let us prove it for $n$. By Lemma 5.4 there exists a $2 \times 2$ rotation matrix $R_\alpha$ such that

$$R_\alpha \begin{pmatrix} x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} a_{n-1} \\ 0 \end{pmatrix},$$

where $a_{n-1} = \sqrt{x_{n-1}^2 + x_n^2}$. So if we define the $n \times n$ elementary rotation $R_1$ by

$$R_1 = \begin{pmatrix} I_{n-2} & \mathbf{0} \\ \mathbf{0} & R_\alpha \end{pmatrix}$$

[3] This term is not widely accepted.

($I_{n-2}$ is $(n-2)\times(n-2)$ identity matrix), then

$$R_1\mathbf{x} = (x_1, x_2, \ldots, x_{n-2}, a_{n-1}, 0)^T.$$

We assumed that the conclusion of the lemma holds for $n-1$, so there exist $n-2$ elementary rotations (let us call them $R_2, R_3, \ldots, R_{n-1}$) in $\mathbb{R}^{n-1}$ which transform the vector $(x_1, x_2, \ldots, x_{n-1}, a_{n-1})^T \in \mathbb{R}^{n-1}$ to the vector $(a, 0, \ldots, 0)^T \in \mathbb{R}^{n-1}$. In other words

$$R_{n-1}\ldots R_3R_2(x_1, x_2, \ldots, x_{n-1}, a_{n-1})^T = (a, 0, \ldots, 0)^T.$$

We can always assume that the elementary rotations $R_2, R_3, \ldots, R_{n-1}$ act in $\mathbb{R}^n$, simply by assuming that they do not change the last coordinate. Then

$$R_{n-1}\ldots R_3R_2R_1\mathbf{x} = (a, 0, \ldots, 0)^T \in \mathbb{R}^n.$$

Let us now show that $a = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$. It can be easily checked directly, but we apply the following indirect reasoning. We know that orthogonal transformations preserve the norm, and we know that $a \geq 0$. But, then we do not have any choice, the only possibility for $a$ is $a = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$. □

**Lemma 5.6.** *Let $A$ be an $n\times n$ matrix with real entries. There exist elementary rotations $R_1, R_2, \ldots, R_N$, $N \leq n(n-1)/2$ such that the matrix $B = R_N\ldots R_2R_1A$ is upper triangular, and, moreover, all its diagonal entries except the last one $B_{n,n}$ are non-negative.*

**Proof.** We will use induction in $n$. The case $n = 1$ is trivial, since we can say that any $1\times 1$ matrix is of desired form.

Let us consider the case $n = 2$. Let $\mathbf{a}_1$ be the first column of $A$. By Lemma 5.4 there exists a rotation $R$ which "kills" the second coordinate of $\mathbf{a}_1$, making the first coordinate non-negative. Then the matrix $B = RA$ is of desired form.

Let us now assume that lemma holds for $(n-1)\times(n-1)$ matrices, and we want to prove it for $n\times n$ matrices. For the $n\times n$ matrix $A$ let $\mathbf{a}_1$ be its first column. By Lemma 5.5 we can find $n-1$ elementary rotations (say $R_1, R_2, \ldots, R_{n-1}$ which transform $\mathbf{a}_1$ into $(a, 0, \ldots, 0)^T$. So, the matrix $R_{n-1}\ldots R_2R_1A$ has the following block triangular form

$$R_{n-1}\ldots R_2R_1A = \begin{pmatrix} a & * \\ \mathbf{0} & A_1 \end{pmatrix},$$

where $A_1$ is an $(n-1)\times(n-1)$ block.

We assumed that lemma holds for $n-1$, so $A_1$ can be transformed by at most $(n-1)(n-2)/2$ rotations into the desired upper triangular form. Note, that these rotations act in $\mathbb{R}^{n-1}$ (only on the coordinates $x_2, x_3, \ldots, x_n$), but

we can always assume that they act on the whole $\mathbb{R}^n$ simply by assuming that they do not change the first coordinate. Then, these rotations do not change the vector $(a, 0, \dots, 0)^T$ (the first column of $R_{n-1} \dots R_2 R_1 A$), so the matrix $A$ can be transformed into the desired upper triangular form by at most $n - 1 + (n-1)(n-2)/2 = n(n-1)/2$ elementary rotations. □

**Proof of Theorem 5.3.** By Lemma 5.5 there exist elementary rotations $R_1, R_2, \dots, R_N$ such that the matrix $U_1 = R_N \dots R_2 R_2 U$ is upper triangular, and all diagonal entries, except maybe the last one, are non-negative.

Note, that the matrix $U_1$ is orthogonal. Any orthogonal matrix is normal, and we know that an upper triangular matrix can be normal only if it is diagonal. Therefore, $U_1$ is a diagonal matrix.

We know that an eigenvalue of an orthogonal matrix can either be 1 or $-1$, so we can have only 1 or $-1$ on the diagonal of $U_1$. But, we know that all diagonal entries of $U_1$, except may be the last one, are non-negative, so all the diagonal entries of $U_1$, except may be the last one, are 1. The last diagonal entry can be $\pm 1$.

Since elementary rotations have determinant 1, we can conclude that $\det U_1 = \det U = 1$, so the last diagonal entry also must be 1. So $U_1 = I$, and therefore $U$ can be represented as a product of elementary rotations $U = R_1^{-1} R_2^{-1} \dots R_N^{-1}$. Here we use the fact that the inverse of an elementary rotation is an elementary rotation as well. □

## 6. Orientation

**6.1. Motivation.** In Figures 1, 2 below we see 3 orthonormal bases in $\mathbb{R}^2$ and $\mathbb{R}^3$ respectively. In each figure, the basis b) can be obtained from the standard basis a) by a rotation, while it is impossible to rotate the standard basis to get the basis c) (so that $\mathbf{e}_k$ goes to $\mathbf{v}_k$ $\forall k$).

You have probably heard the word "orientation" before, and you probably know that bases a) and b) have positive orientation, and orientation of the bases c) is negative. You also probably know some rules to determine

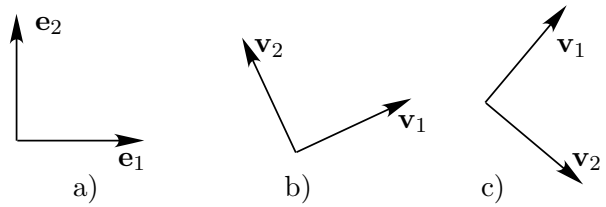


**Figure 1.** Orientation in $\mathbb{R}^2$

the orientation, like the right hand rule from physics. So, if you can *see* a basis, say in $\mathbb{R}^3$, you probably can say what orientation it has.

But what if you only given coordinates of the vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$? Of course, you can try to draw a picture to visualize the vectors, and then to see what the orientation is. But this is not always easy. Moreover, how do you "explain" this to a computer?

It turns out that there is an easier way. Let us explain it. We need to check whether it is possible to get a basis $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ in $\mathbb{R}^3$ by rotating the standard basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_n$. There is unique linear transformation $U$ such that

$$U\mathbf{e}_k = \mathbf{v}_k, \qquad k = 1, 2, 3;$$

its matrix (in the standard basis) is the matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. It is an orthogonal matrix (because it transforms an orthonormal basis to an orthonormal basis), so we need to see when it is rotation. Theorems 5.1 and 5.2 give us the answer: the matrix $U$ is a rotation if and only if $\det U = 1$. Note, that (for $3 \times 3$ matrices) if $\det U = -1$, then $U$ is the composition of a rotation about some axis and a reflection in the plane of rotation, i.e. in the plane orthogonal to this axis.

This gives us a motivation for the formal definition below.

**6.2. Formal definition.** Let $\mathcal{A}$ and $\mathcal{B}$ be two bases in a *real* vector space $X$. We say that the bases $\mathcal{A}$ and $\mathcal{B}$ have the same orientation, if the change of coordinates matrix $[I]_{\mathcal{B},\mathcal{A}}$ has positive determinant, and say that they have different orientations if the determinant of $[I]_{\mathcal{B},\mathcal{A}}$ is negative.

Note, that since $[I]_{\mathcal{A},\mathcal{B}} = [I]^{-1}_{\mathcal{B},\mathcal{A}}$, one can use the matrix $[I]_{\mathcal{A},\mathcal{B}}$ in the definition.

We usually assume that the standard basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ in $\mathbb{R}^n$ has positive orientation. In an abstract space one just needs to fix a basis and declare that its orientation is positive.

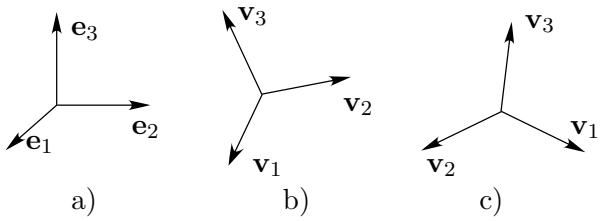


**Figure 2.** Orientation in $\mathbb{R}^3$

If an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ in $\mathbb{R}^n$ has positive orientation (i.e. the same orientation as the standard basis) Theorems 5.1 and 5.2 say that the basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is obtained from the standard basis by a rotation.

### 6.3. Continuous transformations of bases and orientation.

**Definition.** We say that a basis $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$ can be continuously transformed to a basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ if there exists a continuous family of bases $\mathcal{V}(t) = \{\mathbf{v}_1(t), \mathbf{v}_2(t), \dots, \mathbf{v}_n(t)\}$, $t \in [a, b]$ such that

$$\mathbf{v}_k(a) = \mathbf{a}_k, \quad \mathbf{v}_k(b) = \mathbf{b}_k, \qquad k = 1, 2 \dots, n.$$

"Continuous family of bases" mean that the vector-functions $\mathbf{v}_k(t)$ are continuous (their coordinates in some bases are continuous functions) and, which is essential, the system $\mathbf{v}_1(t), \mathbf{v}_2(t), \dots, \mathbf{v}_n(t)$ is a basis for all $t \in [a, b]$.

Note, that performing a change of variables, we can always assume, if necessary that $[a, b] = [0, 1]$.

**Theorem 6.1.** *Two bases $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$ and $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ have the same orientation, if and only if one of the bases can be continuously transformed to the other.*

**Proof.** Suppose the basis $\mathcal{A}$ can be continuously transformed to the basis $\mathcal{B}$, and let $\mathcal{V}(t)$, $t \in [a, b]$ be a continuous family of bases, performing this transformation. Consider a matrix-function $V(t)$ whose columns are the coordinate vectors $[\mathbf{v}_k(t)]_{\mathcal{A}}$ of $\mathbf{v}_k(t)$ in the basis $\mathcal{A}$.

Clearly, the entries of $V(t)$ are continuous functions and $V(a) = I$, $V(b) = [I]_{\mathcal{A},\mathcal{B}}$. Note, that because $\mathcal{V}(t)$ is always a basis, $\det V(t)$ is never zero. Then, the Intermediate Value Theorem asserts that $\det V(a)$ and $\det V(b)$ has the same sign. Since $\det V(a) = \det I = 1$, we can conclude that

$$\det[I]_{\mathcal{A},\mathcal{B}} = \det V(b) > 0,$$

so the bases $\mathcal{A}$ and $\mathcal{B}$ have the same orientation.

To prove the opposite implication, i.e. the "only if" part of the theorem, one needs to show that the identity matrix $I$ can be continuously transformed through invertible matrices to any matrix $B$ satisfying $\det B > 0$. In other words, that there exists a continuous matrix-function $V(t)$ on an interval $[a, b]$ such that for all $t \in [a, b]$ the matrix $V(t)$ is invertible and such that

$$V(a) = I, \qquad V(b) = B.$$

We leave the proof of this fact as an exercise for the reader. There are several ways to prove that, on of which is outlined in Problems 6.2—6.5 below. □

**Exercises.**

**6.1.** Let $R_\alpha$ be the rotation through $\alpha$, so its matrix in the standard basis is

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Find the matrix of $R_\alpha$ in the basis $\mathbf{v}_1, \mathbf{v}_2$, where $\mathbf{v}_1 = \mathbf{e}_2$, $\mathbf{v}_2 = \mathbf{e}_1$.

**6.2.** Let $R_\alpha$ be the rotation matrix

$$R_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Show that the $2 \times 2$ identity matrix $I_2$ can be continuously transformed through invertible matrices into $R_\alpha$.

**6.3.** Let $U$ be an $n \times n$ orthogonal matrix, and let $\det U > 0$. Show that the $n \times n$ identity matrix $I_n$ can be continuously transformed through invertible matrices into $U$. **Hint:** Use the previous problem and representation of a rotation in $\mathbb{R}^n$ as a product of planar rotations, see Section 5.

**6.4.** Let $A$ be an $n \times n$ positive definite Hermitian matrix, $A = A^* > \mathbf{0}$. Show that the $n \times n$ identity matrix $I_n$ can be continuously transformed through invertible matrices into $A$. **Hint:** What about diagonal matrices?

**6.5.** Using polar decomposition and Problems 6.3, 6.4 above, complete the proof of the "only if" part of Theorem 6.3

*Chapter 7*

# Bilinear and quadratic forms

While the study of *real* quadratic forms (i.e. real homogeneous polynomials of degree 2) was probably the initial motivation for the subject of this chapter, complex quadratic forms $(A\mathbf{x}, \mathbf{x})$, $\mathbf{x} \in \mathbb{C}^n$, $A = A^*$ are also of significant interest. So, unless otherwise specified, result and calculations hold in both real and complex case.

To avoid writing twice essentially the same formulas, we use the notation adapted to the complex case: in particular, in the real case the notation $A^*$ is used instead of $A^T$.

## 1. Main definition

**1.1. Bilinear forms on $\mathbb{R}^n$.** A bilinear form on $\mathbb{R}^n$ is a function $L = L(\mathbf{x}, \mathbf{y})$ of two arguments $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ which is linear in each argument, i.e. such that

1. $L(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2, \mathbf{y}) = \alpha L(\mathbf{x}_1, \mathbf{y}) + \beta L(\mathbf{x}_2, \mathbf{y})$;
2. $L(\mathbf{x}, \alpha\mathbf{y}_1 + \beta\mathbf{y}_2) = \alpha L(\mathbf{x}, \mathbf{y}_1) + \beta L(\mathbf{x}, \mathbf{y}_2)$.

One can consider bilinear form whose values belong to an arbitrary vector space, but in this book we only consider forms that take real values.

If $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, a bilinear form can be written as

$$L(\mathbf{x}, \mathbf{y}) = \sum_{j,k=1}^{n} a_{j,k} x_k y_j,$$

or in matrix form

$$L(\mathbf{x},\mathbf{y}) = (A\mathbf{x},\mathbf{y})$$

where

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix}.$$

The matrix $A$ is determined uniquely by the bilinear form $L$.

**1.2. Quadratic forms on $\mathbb{R}^n$.** There are several equivalent definition of a quadratic form.

One can say that a quadratic form on $\mathbb{R}^n$ is the *"diagonal"* of a bilinear form $L$, i.e. that any quadratic form $Q$ is defined by $Q[\mathbf{x}] = L(\mathbf{x},\mathbf{x}) = (A\mathbf{x},\mathbf{x})$.

Another, more algebraic way, is to say that a quadratic form is a *homogeneous polynomial of degree 2*, i.e. that $Q[\mathbf{x}]$ is a polynomial of $n$ variables $x_1, x_2, \dots, x_n$ having only terms of degree 2. That means that only terms $ax_k^2$ and $cx_jx_k$ are allowed.

There many ways (in fact, infinitely many) to write a quadratic form $Q[\mathbf{x}]$ as $Q[\mathbf{x}] = (A\mathbf{x},\mathbf{x})$. For example, the quadratic form $Q[\mathbf{x}] = x_1^2 + x_2^2 - 4x_1x_2$ on $\mathbb{R}^2$ can be represented as $(A\mathbf{x},\mathbf{x})$ where $A$ can be any of the matrices

$$\begin{pmatrix} 1 & -4 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ -4 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}.$$

In fact, any matrix $A$ of form

$$\begin{pmatrix} 1 & a-4 \\ -a & 1 \end{pmatrix}$$

will work.

But if we require the matrix $A$ to be symmetric, then such a matrix is unique:

| Any quadratic form $Q[\mathbf{x}]$ on $\mathbb{R}^n$ admits unique representation $Q[\mathbf{x}] = (A\mathbf{x},\mathbf{x})$ where $A$ is a (real) symmetric matrix. |
|---|

For example, for the quadratic form

$$Q[\mathbf{x}] = x_1^2 + 3x_2^2 + 5x_3^2 + 4x_1x_2 - 16x_1x_3 + 7x_2x_3$$

on $\mathbb{R}^3$, the corresponding symmetric matrix $A$ is

$$\begin{pmatrix} 1 & 2 & -8 \\ 2 & 3 & 3.5 \\ -8 & 3.5 & 5 \end{pmatrix}.$$

**1.3. Quadratic forms on $\mathbb{C}^n$.** One can also define a *quadratic form* on $\mathbb{C}^n$ (or any complex inner product space) by taking a self-adjoint transformation $A = A^*$ and defining $Q$ by $Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$. While our main examples will be in $\mathbb{R}^n$, all the theorems are true in the setting of $\mathbb{C}^n$ as well. Bearing this in mind, we will always use $A^*$ instead of $A^T$

The only essential difference with the real case is that in the complex case we do not have any freedom of choice: if the quadratic form is real, the corresponding matrix has to be Hermitian (self-adjoint).

Note that if $A = A^*$ then

$$(A\mathbf{x}, \mathbf{x}) = (\mathbf{x}, A^*\mathbf{x}) = (\mathbf{x}, A\mathbf{x}) = \overline{(A\mathbf{x}, \mathbf{x})},$$

so $(A\mathbf{x}, \mathbf{x}) \in \mathbb{R}$.

The converse is also true.

**Lemma 1.1.** *Let $(A\mathbf{x}, \mathbf{x})$ be real for all $\mathbf{x} \in \mathbb{C}^n$. Then $A = A^*$.*

We leave the proof as an exercise for the reader, see Problem 1.4 below

**Exercises.**

**1.1.** Find the matrix of the bilinear form $L$ on $\mathbb{R}^3$,

$L(\mathbf{x}, \mathbf{y}) = x_1y_1 + 2x_1y_2 + 14x_1y_3 - 5x_2y_1 + 2x_2y_2 - 3x_2y_3 + 8x_3y_1 + 19x_3y_2 - 2x_3y_3$.

**1.2.** Define the bilinear form $L$ on $\mathbb{R}^2$ by

$$L(\mathbf{x}, \mathbf{y}) = \det[\mathbf{x}, \mathbf{y}],$$

i.e. to compute $L(\mathbf{x}, \mathbf{y})$ we form a $2 \times 2$ matrix with columns $\mathbf{x}, \mathbf{y}$ and compute its determinant.

Find the matrix of $L$.

**1.3.** Find the matrix of the quadratic form $Q$ on $\mathbb{R}^3$

$$Q[\mathbf{x}] = x_1^2 + 2x_1x_2 - 3x_1x_3 - 9x_2^2 + 6x_2x_3 + 13x_3^2.$$

**1.4.** Prove Lemma 1.1 above.

**Hint:** Consider the expression $(A(\mathbf{x} + z\mathbf{y}), \mathbf{x} + z\mathbf{y})$ and show that if it is real for all $z \in \mathbb{C}$ then $(A\mathbf{x}, \mathbf{y}) = \overline{(\mathbf{y}, A^*\mathbf{x})}$

## 2. Diagonalization of quadratic forms

You have probably met quadratic forms before, when you studied second order curves in the plane. Maybe you even studied the second order surfaces in $\mathbb{R}^3$.

We want to present a unified approach to classification of such objects. Suppose we are given a set in $\mathbb{R}^n$ defined by the equation $Q[\mathbf{x}] = 1$, where $Q$ is some quadratic form. If $Q$ has some simple form, for example if the corresponding matrix is diagonal, i.e. if $Q[x] = a_1x_1^2 + a_2x_2^2 + \ldots + a_nx_n^2$, we

can easily visualize this set, especially if $n = 2, 3$. In higher dimensions, it is also possible, if not to visualize, then to understand the structure of the set very well.

So, if we are given a general, complicated quadratic form, we want to simplify it as much as possible, for example to make it diagonal. The standard way of doing that is the change of variables.

**2.1. Orthogonal diagonalization.** Let us have a quadratic form $Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$ in $\mathbb{F}^n$ ($\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$). Introduce new variables $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in \mathbb{F}^n$, with $\mathbf{y} = S^{-1}\mathbf{x}$, where $S$ is some invertible $n \times n$ matrix, so $\mathbf{x} = S\mathbf{y}$.

Then,

$$Q[\mathbf{x}] = Q[S\mathbf{y}] = (AS\mathbf{y}, S\mathbf{y}) = (S^*AS\mathbf{y}, \mathbf{y}),$$

so in the new variables $\mathbf{y}$ the quadratic form has matrix $S^*AS$.

So, we want to find an invertible matrix $S$ such that the matrix $S^*AS$ is diagonal. Note, that it is different from the diagonalization of matrices we had before: we tried to represent a matrix $A$ as $A = SDS^{-1}$, so the matrix $D = S^{-1}AS$ was diagonal. However, for unitary matrices $U$, we have $U^* = U^{-1}$, and we can orthogonally diagonalize symmetric matrices. So we can apply the orthogonal diagonalization we studied before to the quadratic forms.

Namely, we can represent the matrix $A$ as $A = UDU^* = UDU^{-1}$. Recall, that $D$ is a diagonal matrix with eigenvalues of $A$ on the diagonal, and $U$ is the matrix of eigenvectors (we need to pick an orthogonal basis of eigenvectors). Then $D = U^*AU$, so in the variables $\mathbf{y} = U^{-1}\mathbf{x}$ the quadratic form has diagonal matrix.

Let us analyze the geometric meaning of the orthogonal diagonalization. The columns $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ of the unitary matrix $U$ form an orthonormal basis in $\mathbb{F}^n$, let us call this basis $\mathcal{B}$. The change of coordinate matrix $[I]_{\mathcal{S},\mathcal{B}}$ from this basis to the standard one is exactly $U$. We know that $\mathbf{y} = (y_1, y_2, \dots, y_n)^T = U^{-1}\mathbf{x}$, so the coordinates $y_1, y_2, \dots, y_n$ can be interpreted as coordinates of the vector $\mathbf{x}$ in the new basis $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$.

So, orthogonal diagonalization allows us to visualize very well the set $Q[\mathbf{x}] = 1$, or a similar one, as long as we can visualize it for diagonal matrices.

**Example.** Consider the quadratic form of two variables (i.e. quadratic form on $\mathbb{R}^2$), $Q(x, y) = 2x^2 + 2y^2 + 2xy$. Let us describe the set of points $(x, y)^T \in \mathbb{R}^2$ satisfying $Q(x, y) = 1$.

The matrix $A$ of $Q$ is

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Orthogonally diagonalizing this matrix we can represent it as

$$A = U \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} U^*, \qquad \text{where} \quad U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

or, equivalently

$$U^* A U = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} =: D.$$

The set $\{\mathbf{y} : (D\mathbf{y}, \mathbf{y}) = 1\}$ is the ellipse with half-axes $1/\sqrt{3}$ and 1. Therefore the set $\{\mathbf{x} \in \mathbb{R}^2 : (A\mathbf{x}, \mathbf{x}) = 1\}$, is the same ellipse only in the basis $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$, $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$, or, equivalently, the same ellipse, rotated $\pi/4$.

**2.2. Non-orthogonal diagonalization.** Orthogonal diagonalization involves computing eigenvalues and eigenvectors, so it may be difficult to do without computers for large $n$. On the other hand, the non-orthogonal diagonalization, i.e. finding an invertible $S$ (without requiring $S^{-1} = S^*$) such that $D = S^* A S$ is diagonal, is much easier computationally and can be done using only algebraic operations (addition, subtraction, multiplication, division).

Below we present two most used methods of non-orthogonal diagonalization.

2.2.1. *Diagonalization by completion of squares.* The first methods is based on completion of squares. We will illustrate this method on real quadratic forms (forms on $\mathbb{R}^n$). After simple modifications this method could be used in the complex case, but we will not discuss it here. If necessary, an interested reader should be able to to make the appropriate modifications.

Let us again consider the quadratic form of two variables, $Q[\mathbf{x}] = 2x_1^2 + 2x_1x_2 + 2x_2^2$ (it is the same quadratic form as in the above example, only here we call variables not $x, y$ but $x_1, x_2$). Since

$$2\left(x_1 + \frac{1}{2}x_2\right)^2 = 2\left(x_1^2 + 2x_1\frac{1}{2}x_2 + \frac{1}{4}x_2^2\right) = 2x_1^2 + 2x_1x_2 + \frac{1}{2}x_2^2$$

(note, that the first two terms coincide with the first two terms of $Q$), we get

$$2x_1^2 + 2x_1x_2 + 2x_2^2 = 2\left(x_1 + \frac{1}{2}x_2\right)^2 + \frac{3}{2}x_2^2 = 2y_1^2 + \frac{3}{2}y_2^2,$$

where $y_1 = x_1 + \frac{1}{2}x_2$ and $y_2 = x_2$.

The same method can be applied to quadratic form of more than 2 variables. Let us consider, for example, a form $Q[\mathbf{x}]$ in $\mathbb{R}^3$,

$$Q[\mathbf{x}] = x_1^2 - 6x_1x_2 + 4x_1x_3 - 6x_2x_3 + 8x_2^2 - 3x_3^2.$$

Considering all terms involving the first variable $x_1$ (the first 3 terms in this case), let us pick a full square or a multiple of a full square which has exactly these terms (plus some other terms).

Since

$$(x_1 - 3x_2 + 2x_3)^2 = x_1^2 - 6x_1x_2 + 4x_1x_3 - 12x_2x_3 + 9x_2^2 + 4x_3^2$$

we can rewrite the quadratic form as

$$(x_1 - 3x_2 + 2x_3)^2 - x_2^2 + 6x_2x_3 - 7x_3^2.$$

Note, that the expression $-x_2^2 + 6x_2x_3 - 7x_3^2$ involves only variables $x_2$ and $x_3$. Since

$$-(x_2 - 3x_3)^2 = -(x_2^2 - 6x_2x_3 + 9x_3^2) = -x_2^2 + 6x_2x_3 - 9x_3^2$$

we have

$$-x_2^2 + 6x_2x_3 - 7x_3^2 = -(x_2 - 3x_3)^2 + 2x_3^2.$$

Thus we can write the quadratic form $Q$ as

$$Q[\mathbf{x}] = (x_1 - 3x_2 + 2x_3)^2 - (x_2 - 3x_3)^2 + 2x_3^2 = y_1^2 - y_2^2 + 2y_3^2$$

where

$$y_1 = x_1 - 3x_2 + 2x_3, \qquad y_2 = x_2 - 3x_3, \qquad y_3 = x_3.$$

Finally, let us address the question that an attentive reader is probably already asking: what to do if at some point we do have a product of two variables, but no corresponding squares? For example, how to diagonalize the form $x_1x_2$? The answer follows immediately from the identity

$$4x_1x_2 = (x_1 + x_2)^2 - (x_1 - x_2)^2, \tag{2.1}$$

which gives us the representation

$$Q[\mathbf{x}] = y_1^2 - y_2^2, \qquad y_1 = (x_1 + x_2)/2,\ y_2 = (x_1 - x_2)/2.$$

2.2.2. *Diagonalization using row/column operations.* There is another way of performing non-orthogonal diagonalization of a quadratic form. The idea is to perform row operations on the matrix $A$ of the quadratic form. The difference with the row reduction (Gauss–Jordan elimination) is that after each row operation we need to perform the same column operation, the reason for that being that we want to make the matrix $S^*AS$ diagonal.

Let us explain how everything works on an example. Suppose we want to diagonalize a quadratic form with matrix

$$A = \begin{pmatrix} 1 & -1 & 3 \\ -1 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix}.$$

We augment the matrix $A$ by the identity matrix, and perform on the augmented matrix $(A|I)$ row/column operations. After each row operation we have to perform on the matrix $A$ the same column operation.[1] We get

$$\left(\begin{array}{ccc|ccc} 1 & -1 & 3 & 1 & 0 & 0 \\ -1 & 2 & 1 & 0 & 1 & 0 \\ 3 & 1 & 1 & 0 & 0 & 1 \end{array}\right)\begin{array}{l}\\ +R_1 \\ \\ \end{array} \sim \left(\begin{array}{ccc|ccc} 1 & -1 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 3 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 3 & 4 & 1 & 0 & 0 & 1 \end{array}\right)\begin{array}{l}\\ \\ -3R_1 \end{array} \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 0 & 4 & -8 & -3 & 0 & 1 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 0 & 4 & -8 & -3 & 0 & 1 \end{array}\right)\begin{array}{l}\\ \\ -4R_2 \end{array} \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 0 & 0 & -24 & -7 & -4 & 1 \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & -24 & -7 & -4 & 1 \end{array}\right).$$

Note, that we perform column operations only on the left side of the augmented matrix

We get the diagonal $D$ matrix on the left, and the matrix $S^*$ on the right, so $D = S^*AS$,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -24 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -7 & -4 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 3 \\ -1 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -7 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{pmatrix}.$$

Let us explain why the method works. A row operation is a left multiplication by an elementary matrix. The corresponding column operation is the right multiplication by the transposed elementary matrix. Therefore, performing row operations $E_1, E_2, \dots, E_N$ and the same column operations we transform the matrix $A$ to

$$E_N \dots E_2 E_1 A E_1^* E_2^* \dots E_N^* = EAE^*. \tag{2.2}$$

As for the identity matrix in the right side, we performed only row operations on it, so the identity matrix is transformed to

$$E_N \dots E_2 E_1 I = EI = E.$$

If we now denote $E^* = S$ we get that $S^*AS$ is a diagonal matrix, and the matrix $E = S^*$ is the right half of the transformed augmented matrix.

In the above example we got lucky, because we did not need to interchange two rows. This operation is a bit tricker to perform. It is quite

[1] In the case of complex Hermitian matrices we perform for each row operation the conjugate o the corresponding column operation, see Remark 2.1 below

simple if you know what to do, but it may be hard to guess the correct row operations. Let us consider, for example, a quadratic form with the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

If we want to diagonalize it by row and column operations, the simplest idea would be to interchange rows 1 and 2. But we also must to perform the same column operation, i.e. interchange columns 1 and 2, so we will end up with the same matrix.

So, we need something more non-trivial. The identity (2.1), for example, can be used to diagonalize this quadratic form. However, a simpler idea also works: use row operations to get a non-zero entry on the diagonal! For example, if we start with making $a_{1,1}$ non-zero, the following series of row (and the corresponding column) operations is one of the possible choices:

$$\left(\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right)^{+\frac{1}{2}R_2} \sim \left(\begin{array}{cc|cc} 1/2 & 1 & 1 & 1/2 \\ 1 & 0 & 0 & 1 \end{array}\right) \sim$$
$$\left(\begin{array}{cc|cc} 1 & 1 & 1 & 1/2 \\ 1 & 0 & 0 & 1 \end{array}\right)_{-R_1} \sim \left(\begin{array}{cc|cc} 1 & 1 & 1 & 1/2 \\ 0 & -1 & -1 & 1/2 \end{array}\right) \sim$$
$$\left(\begin{array}{cc|cc} 1 & 0 & 1 & 1/2 \\ 0 & -1 & -1 & 1/2 \end{array}\right).$$

**Remark.** Non-orthogonal diagonalization gives us a simple description of a set $Q[\mathbf{x}] = 1$ in a non-orthogonal basis. It is harder to visualize, than the representation given by the orthogonal diagonalization. However, if we are not interested in the details, for example if it is sufficient for us just to know that the set is an ellipsoid (or hyperboloid, etc), then the non-orthogonal diagonalization is an easier way to get the answer.

**Remark 2.1.** For quadratic forms with complex entries (i.e. for forms $(A\mathbf{x}, \mathbf{x})$, $A = A^*$), the non-orthogonal diagonalization works the same way as in the real case, with the only difference, that the corresponding "column operations" have the complex conjugate coefficients.

The reason for that is that if a row operation is given by left multiplication by an elementary matrix $E_k$, then the corresponding column operation is given by the right multiplication by $E_k^*$, see (2.2).

Note that formula (2.2) works in both complex and reals cases: in real case we could write $E_k^T$ instead of $E_k^*$, but using Hermitian adjoint allows us to have the same formula in both cases.

**Exercises.**

**2.1.** Diagonalize the quadratic form with the matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}.$$

Use two methods: completion of squares and row operations. Which one do you like better?

Can you say if the matrix $A$ is positive definite or not?

**2.2.** For the matrix $A$

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

orthogonally diagonalize the corresponding quadratic form, i.e. find a diagonal matrix $D$ and a unitary matrix $U$ such that $D = U^*AU$.

## 3. Silvester's Law of Inertia

As we discussed above, there are many ways to diagonalize a quadratic form. Note, that a resulting diagonal matrix is not unique. For example, if we got a diagonal matrix

$$D = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\},$$

we can take a diagonal matrix

$$S = \operatorname{diag}\{s_1, s_2, \dots, s_n\}, \qquad s_k \in \mathbb{R}, \quad s_k \neq 0$$

and transform $D$ to

$$S^*DS = \operatorname{diag}\{s_1^2\lambda_1, s_2^2\lambda_2, \dots, s_n^2\lambda_n\}.$$

This transformation changes the diagonal entries of $D$. However, it does not change the *signs* of the diagonal entries. And this is always the case!

Namely, the famous Silvester's Law of Inertia states that:

> For a Hermitian matrix $A$ (i.e. for a quadratic form $Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$) and any of its diagonalization $D = S^*AS$, the number of positive (negative, zero) diagonal entries of $D$ depends only on $A$, but not on a particular choice of diagonalization.

Here we of course assume that $S$ is an invertible matrix, and $D$ is a diagonal one.

The idea of the proof of the Silvester's Law of Inertia is to express the number of positive (negative, zero) diagonal entries of a diagonalization $D = S^*AS$ in terms of $A$, not involving $S$ or $D$.

We will need the following definition.

**Definition.** Given an $n \times n$ Hermitian matrix $A = A^*$ (a quadratic form $Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$ on $\mathbb{F}^n$) we call a subspace $E \subset \mathbb{F}^n$ *positive* (resp. *negative*, resp. *neutral*) if

$$(A\mathbf{x}, \mathbf{x}) > 0 \qquad (\text{resp. } (A\mathbf{x}, \mathbf{x}) < 0, \quad \text{resp. } (A\mathbf{x}, \mathbf{x}) = 0)$$

for all $\mathbf{x} \in E$, $\mathbf{x} \neq \mathbf{0}$.

Sometimes, to emphasize the role of $A$ we will say $A$-positive ($A$ negative, $A$-neutral).

**Theorem 3.1.** *Let $A$ be an $n \times n$ Hermitian matrix, and let $D = S^*AS$ be its diagonalization by an invertible matrix $S$. Then the number of positive (resp. negative) diagonal entries of $D$ coincides with the maximal dimension of an $A$-positive (resp. $A$-negative) subspace.*

The above theorem says that if $r_+$ is the number of positive diagonal entries of $D$, then there exists an $A$-positive subspace $E$ of dimension $r_+$, but it is impossible to find a positive subspace $E$ with $\dim E > r_+$.

We will need the following lemma, which can be considered a particular case of the above theorem.

**Lemma 3.2.** *Let $D$ be a diagonal matrix $D = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. Then the number of positive (resp. negative) diagonal entries of $D$ coincides with the maximal dimension of a $D$-positive (resp. $D$-negative) subspace.*

**Proof.** By rearranging the standard basis in $\mathbb{F}^n$ (changing the numeration) we can always assume without loss of generality that the positive diagonal entries of $D$ are the first $r_+$ diagonal entries.

Consider the subspace $E_+$ spanned by the first $r_+$ coordinate vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{r_+}$. Clearly $E_+$ is a $D$-positive subspace, and $\dim E_+ = r_+$.

Let us now show that for any other $D$-positive subspace $E$ we have $\dim E \leq r_+$. Consider the orthogonal projection $P = P_{E_+}$,

$$P\mathbf{x} = (x_1, x_2, \dots, x_{r_+}, 0 \dots, 0)^T, \qquad \mathbf{x} = (x_1, x_2, \dots, x_n)^T.$$

For a $D$-positive subspace $E$ define an operator $T : E \to E_+$ by

$$T\mathbf{x} = P\mathbf{x}, \qquad \forall \mathbf{x} \in E.$$

In other words, $T$ is the *restriction* of the projection $P$: $P$ is defined on the whole space, but we restricted its domain to $E$ and target space to $E_+$. We got an operator acting from $E$ to $E_+$, and we use a different letter to distinguish it from $P$.

Note, that $\operatorname{Ker} T = \{\mathbf{0}\}$. Indeed, let for $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in E$ we have $T\mathbf{x} = P\mathbf{x} = \mathbf{0}$. Then, by the definition of $P$

$$x_1 = x_2 = \ldots = x_{r_+} = 0,$$

and therefore

$$(D\mathbf{x},\mathbf{x}) = \sum_{k=r_++1}^{n} \lambda_k x_k^2 \le 0 \qquad (\lambda_k \le 0 \text{ for } k > r_+).$$

But $\mathbf{x}$ belongs to a $D$-positive subspace $E$, so the inequality $(D\mathbf{x},\mathbf{x}) \le 0$ holds only for $\mathbf{x} = \mathbf{0}$.

Let us now apply the Rank Theorem (Theorem 7.1 from Chapter 2). First of all, $\operatorname{rank} T = \dim \operatorname{Ran} T \le \dim E_+ = r_+$ because $\operatorname{Ran} T \subset E_+$. By the Rank Theorem, $\dim \operatorname{Ker} T + \operatorname{rank} T = \dim E$. But we just proved that $\operatorname{Ker} T = \{\mathbf{0}\}$, i.e. that $\dim \operatorname{Ker} T = 0$, so

$$\dim E = \operatorname{rank} T \le \dim E_+ = r_+.$$

To prove the statement about negative entries, we just apply the above reasoning to the matrix $-D$. □

**Proof of Theorem 3.1.** Let $D = S^*AS$ be a diagonalization of $A$. Since

$$(D\mathbf{x},\mathbf{x}) = (S^*AS\mathbf{x},\mathbf{x}) = (AS\mathbf{x},S\mathbf{x})$$

it follows that for any $D$-positive subspace $E$, the subspace $SE$ is an $A$-positive subspace. The same identity implies that for any $A$-positive subspace $F$ the subspace $S^{-1}F$ is $D$-positive.

Since $S$ and $S^{-1}$ are invertible transformations, $\dim E = \dim SE$ and $\dim F = \dim S^{-1}F$. Therefore, for any $D$ positive subspace $E$ we can find an $A$-positive subspace (namely $SE$) of the same dimension, and vice versa: for any $A$-positive subspace $F$ we can find a $D$-positive subspace (namely $S^{-1}F$) of the same dimension. Therefore the maximal possible dimensions of a $A$-positive and a $D$-positive subspace coincide, and the theorem is proved.

The case of negative diagonal entries treated similarly, we leave the details as an exercise to the reader. □

## 4. Positive definite forms. Minimax characterization of eigenvalues and the Silvester's criterion of positivity

**Definition.** A quadratic form $Q$ is called

- Positive definite if $Q[\mathbf{x}] > 0$ for all $\mathbf{x} \ne \mathbf{0}$.
- Positive semidefinite if $Q[\mathbf{x}] \ge 0$ for all $\mathbf{x}$.
- Negative definite if $Q[\mathbf{x}] < 0$ for all $\mathbf{x} \ne \mathbf{0}$.
- Negative semidefinite if $Q[\mathbf{x}] \le 0$ for all $\mathbf{x}$.
- Indefinite if it take both positive and negative values, i.e. if there exist vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $Q[\mathbf{x}_1] > 0$ and $Q[\mathbf{x}_2] < 0$.

**Definition.** A Hermitian matrix $A = A^*$ is called positive definite (negative definite, etc...) if the corresponding quadratic form $Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$ is positive definite (negative definite, etc...).

**Theorem 4.1.** *Let $A = A^*$. Then*

1. *$A$ is positive definite iff all eigenvalues of $A$ are positive.*
2. *$A$ is positive semidefinite iff all eigenvalues of $A$ are non-negative.*
3. *$A$ is negative definite iff all eigenvalues of $A$ are negative.*
4. *$A$ is negative semidefinite iff all eigenvalues of $A$ are non-positive.*
5. *$A$ is indefinite iff it has both positive and negative eigenvalues.*

**Proof.** The proof follows trivially from the orthogonal diagonalization. Indeed, there is an orthonormal basis in which matrix of $A$ is diagonal, and for diagonal matrices the theorem is trivial. □

**Remark.** Note, that to find whether a matrix (a quadratic form) is positive definite (negative definite, etc) one does not have to compute eigenvalues. By Silvester's Law of Inertia it is sufficient to perform an arbitrary, not necessarily orthogonal diagonalization $D = S^*AS$ and look at the diagonal entries of $D$.

**4.1. Silvester's criterion of positivity.** It is an easy exercise to see that a $2 \times 2$ matrix

$$A = \begin{pmatrix} a & b \\ \overline{b} & c \end{pmatrix}$$

is positive definite if and only if

$$a > 0 \qquad \text{and } \det A = ac - |b|^2 > 0 \tag{4.1}$$

Indeed, if $a > 0$ and $\det A = ac - |b|^2 > 0$, then $c > 0$, so $\operatorname{trace} A = a + c > 0$. So we know that if $\lambda_1, \lambda_2$ are eigenvalues of $A$ then $\lambda_1\lambda_2 > 0$ ($\det A > 0$) and $\lambda_1 + \lambda_2 = \operatorname{trace} A > 0$. But that only possible if both eigenvalues are positive. So we have proved that conditions (4.1) imply that $A$ is positive definite. The opposite implication is quite simple, we leave it as an exercise for the reader.

This result can be generalized to the case of $n \times n$ matrices. Namely, for a matrix $A$

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix}$$

let us consider its all *upper left submatrices*

$$A_1 = (a_{1,1}),\ A_2 = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix},\ A_3 = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix}, \dots, A_n = A$$

**Theorem 4.2** (Silvester's Criterion of Positivity)**.** *A matrix $A = A^*$ is positive definite if and only if*

$$\det A_k > 0 \qquad \text{for all } k = 1, 2, \dots, n.$$

First of all let us notice that if $A > 0$ then $A_k > 0$ also (can you explain why?). Therefore, since all eigenvalues of a positive definite matrix are positive, see Theorem 4.1, $\det A_k > 0$ for all $k$.

One can show that if $\det A_k > 0\ \forall k$ then all eigenvalues of $A$ are positive by analyzing diagonalization of a quadratic form using row and column operations, which was described in Section 2.2. The key here is the observation that if we perform row/column operations in natural order (i.e. first subtracting the first row/column from all other rows/columns, then subtracting the second row/columns from the rows/columns $3, 4, \dots, n$, and so on...), and if we are not doing any row interchanges, then we automatically diagonalize quadratic forms $A_k$ as well. Namely, after we subtract first and second rows and columns, we get diagonalization of $A_2$; after we subtract the third row/column we get the diagonalization of $A_2$, and so on.

Since we are performing only row replacement we do not change the determinant. Moreover, since we are not performing row exchanges and performing the operations in the correct order, we preserve determinants of $A_k$. Therefore, the condition $\det A_k > 0$ guarantees that each new entry in the diagonal is positive.

Of course, one has to be sure that we can use only row replacements, and perform the operations in the correct order, i.e. that we do not encounter any pathological situation. If one analyzes the algorithm, one can see that the only bad situation that can happen is the situation where at some step we have zero in the pivot place. In other words, if after we subtracted the first $k$ rows and columns and obtained a diagonalization of $A_k$, the entry in the $k+1$st row and $k+1$st column is 0. We leave it as an exercise for the reader to show that this is impossible. □

The proof we outlined above is quite simple. However, let us present, in more detail, another one, which can be found in more advanced textbooks. I personally prefer this second proof, for it demonstrates some important connections.

We will need the following characterization of eigenvalues of a hermitian matrix.

**4.2. Minimax characterization of eigenvalues.** Let us recall that the *codimension* of a subspace $E \subset X$ is by the definition the dimension of its orthogonal complement, $\operatorname{codim} E = \dim(E^\perp)$. Since for a subspace $E \subset X$, $\dim X = n$ we have $\dim E + \dim E^\perp = n$, we can see that $\operatorname{codim} E = \dim X - \dim E$.

Recall that the trivial subspace $\{\mathbf{0}\}$ has dimension zero, so the whole space $X$ has codimension 0.

**Theorem 4.3** (Minimax characterization of eigenvalues)**.** *Let $A = A^*$ be an $n \times n$ matrix, and let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be its eigenvalues taken in the decreasing order. Then*

$$\lambda_k = \max_{\substack{E:\\ \dim E = k}} \min_{\substack{\mathbf{x}\in E\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) = \min_{\substack{F:\\ \operatorname{codim} F = k-1}} \max_{\substack{\mathbf{x}\in F\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}).$$

Let us explain in more details what the expressions like $\max\min$ and $\min\max$ mean. To compute the first one, we need to consider all subspaces $E$ of dimension $k$. For each such subspace $E$ we consider the set of all $\mathbf{x} \in E$ of norm 1, and find the minimum of $(A\mathbf{x}, \mathbf{x})$ over all such $\mathbf{x}$. Thus for each subspace we obtain a number, and we need to pick a subspace $E$ such that the number is maximal. That is the $\max\min$.

The $\min\max$ is defined similarly.

**Remark.** A sophisticated reader may notice a problem here: why do the maxima and minima exist? It is well known, that maximum and minimum have a nasty habit of not existing: for example, the function $f(x) = x$ has neither maximum nor minimum on the open interval $(0, 1)$.

However, in this case maximum and minimum do exist. There are two possible explanations of the fact that $(A\mathbf{x}, \mathbf{x})$ attains maximum and minimum. The first one requires some familiarity with basic notions of analysis: one should just say that the unit sphere in $E$, i.e. the set $\{\mathbf{x} \in E : \|\mathbf{x}\| = 1\}$ is compact, and that a continuous function ($Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$ in our case) on a compact set attains its maximum and minimum.

Another explanation will be to notice that the function $Q[\mathbf{x}] = (A\mathbf{x}, \mathbf{x})$, $\mathbf{x} \in E$ is a quadratic form on $E$. It is not difficult to compute the matrix of this form in some orthonormal basis in $E$, but let us only note that this matrix is not $A$: it has to be a $k \times k$ matrix, where $k = \dim E$.

It is easy to see that for a quadratic form the maximum and minimum over a unit sphere is the maximal and minimal eigenvalues of its matrix.

As for optimizing over all subspaces, we will prove below that the maximum and minimum do exist.

**Proof of Theorem 4.3.** First of all, by picking an appropriate orthonormal basis, we can assume without loss of generality that the matrix $A$ is diagonal, $A = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$.

Pick subspaces $E$ and $F$, $\dim E = k$, $\operatorname{codim} F = k-1$, i.e. $\dim E = n-k+1$. Since $\dim E + \dim F > n$, there exists a non-zero vector $\mathbf{x}_0 \in E \cap F$. By normalizing it we can assume without loss of generality that $\|\mathbf{x}_0\| = 1$. We can always arrange the eigenvalues in decreasing order, so let us assume that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

Since $\mathbf{x}$ belongs to the both subspaces $E$ and $F$

$$\min_{\substack{\mathbf{x}\in E\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) \leq (A\mathbf{x}_0, \mathbf{x}_0) \leq \max_{\substack{\mathbf{x}\in F\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}).$$

We did not assume anything except dimensions about the subspaces $E$ and $F$, so the above inequality

$$\min_{\substack{\mathbf{x}\in E\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) \leq \max_{\substack{\mathbf{x}\in F\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}). \tag{4.2}$$

holds for all pairs of $E$ and $F$ of appropriate dimensions.

Define

$$E_0 := \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k\}, \qquad F_0 := \operatorname{span}\{\mathbf{e}_k, \mathbf{e}_{k+1}, \mathbf{e}_{k+2}, \dots, \mathbf{e}_n\}.$$

Since for a self-adjoint matrix $B$, the maximum and minimum of $(B\mathbf{x}, \mathbf{x})$ over the unit sphere $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ are the maximal and the minimal eigenvalue respectively (easy to check on diagonal matrices), we get that

$$\min_{\substack{\mathbf{x}\in E_0\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) = \max_{\substack{\mathbf{x}\in F_0\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) = \lambda_k.$$

It follows from (4.2) that for any subspace $E$, $\dim E = k$

$$\min_{\substack{\mathbf{x}\in E\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) \leq \max_{\substack{\mathbf{x}\in F_0\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) = \lambda_k$$

and similarly, for any subspace $F$ of codimension $k-1$,

$$\max_{\substack{\mathbf{x}\in F\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) \geq \min_{\substack{\mathbf{x}\in E_0\\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) = \lambda_k.$$

But on subspaces $E_0$ and $F_0$ both maximum and minimum are $\lambda_k$, so $\min\max = \max\min = \lambda_k$. □

**Corollary 4.4** (Intertwining of eigenvalues)**.** *Let $A = A^* = \{a_{j,k}\}_{j,k=1}^n$ be a self-adjoint matrix, and let $\widetilde{A} = \{a_{j,k}\}_{j,k=1}^{n-1}$ be its submatrix of size $(n-1)\times(n-1)$. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ and $\mu_1, \mu_2, \dots, \mu_{n-1}$ be the eigenvalues of $A$ and $\widetilde{A}$ respectively, taken in decreasing order. Then*

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \ldots \geq \lambda_{n-1} \geq \mu_{n-1} \geq \lambda_n.$$

*i.e.*

$$\lambda_k \ge \mu_k \ge \lambda_{k+1}, \qquad k = 1, 2, \dots, n-1$$

**Proof.** Let $\widetilde{X} \subset \mathbb{F}^n$ be the subspace spanned by the first $n-1$ basis vectors, $\widetilde{X} = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{n-1}\}$. Since $(\widetilde{A}\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, \mathbf{x})$ for all $\mathbf{x} \in \widetilde{X}$, Theorem 4.3 implies that

$$\mu_k = \max_{\substack{E \subset \widetilde{X} \\ \dim E = k}} \min_{\substack{\mathbf{x} \in E \\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}).$$

To get $\lambda_k$ we need to get maximum over the set of all subspaces $E$ of $\mathbb{F}^n$, $\dim E = k$, i.e. take maximum over a bigger set (any subspace of $\widetilde{X}$ is a subspace of $\mathbb{F}^n$). Therefore

$$\mu_k \le \lambda_k.$$

(the maximum can only increase, if we increase the set).

On the other hand, any subspace $E \subset \widetilde{X}$ of codimension $k-1$ (here we mean codimension in $\widetilde{X}$) has dimension $n-1-(k-1) = n-k$, so its codimension in $\mathbb{F}^n$ is $k$. Therefore

$$\mu_k = \min_{\substack{E \subset \widetilde{X} \\ \dim E = n-k}} \max_{\substack{\mathbf{x} \in E \\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) \le \min_{\substack{E \subset \mathbb{F}^n \\ \dim E = n-k}} \max_{\substack{\mathbf{x} \in E \\ \|\mathbf{x}\|=1}} (A\mathbf{x}, \mathbf{x}) = \lambda_{k+1}$$

(minimum over a bigger set can only be smaller). □

**Proof of Theorem 4.2.** If $A > 0$, then $A_k > 0$ for $k = 1, 2, \dots, n$ as well (can you explain why?). Since all eigenvalues of a positive definite matrix are positive (see Theorem 4.1), $\det A_k > 0$ for all $k = 1, 2, \dots, n$.

Let us now prove the other implication. Let $\det A_k > 0$ for all $k$. We will show, using induction in $k$, that all $A_k$ (and so $A = A_n$) are positive definite.

Clearly $A_1$ is positive definite (it is $1 \times 1$ matrix, so $A_1 = \det A_1$). Assuming that $A_{k-1} > 0$ (and $\det A_k > 0$) let us show that $A_k$ is positive definite. Let $\lambda_1, \lambda_2, \dots, \lambda_k$ and $\mu_1, \mu_2, \dots, \mu_{k-1}$ be eigenvalues of $A_k$ and $A_{k-1}$ respectively. By Corollary 4.4

$$\lambda_j \ge \mu_j > 0 \qquad \text{for } j = 1, 2, \dots, k-1.$$

Since $\det A_k = \lambda_1 \lambda_2 \dots \lambda_{k-1} \lambda_k > 0$, the last eigenvalue $\lambda_k$ must also be positive. Therefore, since all its eigenvalues are positive, the matrix $A_k$ is positive definite. □

**4.3. Some remarks.** First of all notice, that Silvester Criterion of Positivity does not generalize to positive semidefinite matrices if $n \ge 3$, meaning that for $n \times n$ matrices, $n \ge 3$, the conditions $\det A_k \ge 0$ do not imply that $A$ is positive semidefinite, see Problem 4.4 below.

For $2 \times 2$ matrices, however, the conditions $\det A_k \geq 0$ imply that $A$ is positive semidefinite, see Problem 4.3 below. This sometimes leads to the wrong conclusion about $n \times n$ matrices.

Finally, we should should say couple words about negative definite matrices. It is a typical students' mistake to say that the condition $\det A_k < 0$ implies that $A$ is negative definite. But that is wrong!

To check if the matrix $A$ is negative definite, one just have to check that the matrix $-A$ is positive definite. Applying Silvester's Criterion of Positivity to $-A$ one can see that $A$ is negative definite if and only if $(-1)^k \det A_k > 0$ for all $k = 1, 2, \dots, n$.

**Exercises.**

**4.1.** Using Silvester's Criterion of Positivity check if the matrices

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 3 & -1 \\ 1 & -1 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 3 & -1 & 2 \\ -1 & 4 & -2 \\ 2 & -2 & 1 \end{pmatrix}$$

are positive definite or not.

Are the matrices $-A$, $A^3$ and $A^{-1}$, $A + B^{-1}$, $A + B$, $A - B$ positive definite?

**4.2.** True or false:

a) If $A$ is positive definite, then $A^5$ is positive definite.

b) If $A$ is negative definite, then $A^8$ is negative definite.

c) If $A$ is negative definite, then $A^{12}$ is positive definite.

d) If $A$ is positive definite and $B$ is negative semidefinite, then $A - B$ is positive definite.

e) If $A$ is indefinite, and $B$ is positive definite, then $A + B$ is indefinite.

**4.3.** Let $A$ be a $2 \times 2$ Hermitian matrix, such that $a_{1,1} \geq 0$, $\det A \geq 0$. Prove that $A$ is positive semidefinite.

**4.4.** Find a real symmetric $n \times n$ matrix $A$ such that $\det A_k \geq 0$ for all $k = 1, 2, \dots, n$, but the matrix $A$ is not positive semidefinite. Note that $n$ should be at least 3, see Problem 4.3 above.

## 5. Positive definite forms and inner products

Let $V$ be an inner product space and let $\mathcal{B} = \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be a basis (not necessarily orthogonal) in $V$. Let $G = \{g_{j,k}\}_{j.k=1}^n$ be the matrix defined by

$$g_{j,k} = (\mathbf{v}_k, \mathbf{v}_j).$$

If $\mathbf{x} = \sum_k x_k \mathbf{v}_k$ and $\mathbf{y} = \sum_k y_k \mathbf{v}_k$, then

$$(\mathbf{x}, \mathbf{y}) = \left( \sum_k x_k \mathbf{v}_k, \sum_j y_j \mathbf{v}_j \right) = \sum_{k,j=1}^{n} x_k \overline{y}_j (\mathbf{v}_k, \mathbf{v}_j)$$
$$= \sum_{j=1}^{n} \sum_{k=1}^{n} g_{j,k} x_k \overline{y}_j = (G[\mathbf{x}]_{\mathcal{B}}, [\mathbf{y}]_{\mathcal{B}})_{\mathbb{C}^n} ,$$

where $(\,\cdot\,,\,\cdot\,)_{\mathbb{C}^n}$ stands for the standard inner product in $\mathbb{C}^n$. One can immediately see that $G$ is a positive definite matrix (why?).

So, when one works with coordinates in an arbitrary (not necessarily orthogonal) basis in an inner product space, the inner product (in terms of coordinates) is not computed as the standard inner product in $\mathbb{C}^n$, but with the help of a positive definite matrix $G$ as described above.

Note, that this $G$-inner product coincides with the standard inner product in $\mathbb{C}^n$ if and only if $G = I$, which happens if and only if the basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is orthonormal.

Conversely, given a positive definite matrix $G$ one can define a non-standard inner product ($G$-inner product) in $\mathbb{C}^n$ by

$$(\mathbf{x}, \mathbf{y})_G := (G\mathbf{x}, \mathbf{y})_{\mathbb{C}^n}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{C}^n.$$

One can easily check that $(\mathbf{x}, \mathbf{y})_G$ is indeed an inner product, i.e. that properties 1–4 from Section 1.3 of Chapter 5 are satisfied.

*Chapter 8*

# Dual spaces and tensors

All vector spaces in this chapter are finite-dimensional.

## 1. Dual spaces

### 1.1. Linear functionals and the dual space. Change of coordinates in the dual space.

**Definition 1.1.** A linear functional on a vector space $V$ (over a field $\mathbb{F}$) is a linear transformation $L : V \to \mathbb{F}$.

This special class of linear transformation sufficiently important to deserve a separate name.

If one thinks of vectors as of some physical objects, like force or velocity, then one can think of a linear functional as a (linear) measurement, that gives you some a scalar quantity as the result: think about force or velocity in a given direction.

**Definition 1.2.** A collection of all linear functionals on a finite-dimensional[1] vector space $V$ is called the *dual* of $V$ and is usually denoted as $V'$ or $V^*$

As it was discussed earlier in Section 4 of Chapter 1, the collection $\mathcal{L}(V, W)$ of all linear transformations acting from $V$ to $W$ is a vector space

[1]We consider here only finite-dimensional spaces because for infinite-dimensional spaces the dual space consists not of all but only of the so-called *bounded* linear functionals. Without giving the precise definition, let us only mention than in the finite-dimensional case (both the domain and the target space are finite-dimensional) all linear transformations are bounded, and we do not need to mention the word *bounded*

(with naturally defined operations of addition and multiplication by a scalar. So, the dual space $V' = \mathcal{L}(V, \mathbb{F})$ is a vector space.

Let us consider an example. Let the space $V$ be $\mathbb{R}^n$, what is its dual? As we know, a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is represented by an $m \times n$ matrix, so a linear functional on $\mathbb{R}^n$ (i.e. a linear transformation $L : \mathbb{R}^n \to \mathbb{R}$ is given by an $1 \times n$ matrix (row), let us denote it by $[L]$. The collection of all such rows is isomorphic to $\mathbb{R}^n$ (isomorphism is given by taking the transpose $[L] \to [L]^T$).

So, the dual of $\mathbb{R}^n$ is $\mathbb{R}^n$ itself. The same holds true for $\mathbb{C}^n$, of course, as well as for $\mathbb{F}^n$, where $\mathbb{F}$ is an arbitrary field. Since the space $V$ over a field $\mathbb{F}$ (here we mostly interested in the case $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$) of dimension $n$ is isomorphic to $\mathbb{F}^n$, and the dual to $\mathbb{F}^n$ is isomorphic to $\mathbb{F}^n$, we can conclude that the dual $V'$ is isomorphic to $V$

Thus, the definition of the dual space is starting to look a bit silly, since it does not appear to give us anything new.

However, that is not the case! If we look carefully, we can see that the dual space is indeed a new object. To see that, let us analyze how the entries of the matrix $[L]$ (which we can call the coordinates of $L$) change when we change the basis in $V$.

1.1.1. *Change of coordinates formula.* Let

$$\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}, \qquad \mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$$

be two bases in $V$, and let $[L]_{\mathcal{A}} = [L]_{\mathcal{S},\mathcal{A}}$ and $[L]_{\mathcal{B}} = [L]_{\mathcal{S},\mathcal{B}}$ be the matrices of $L$ in the bases $\mathcal{A}$ and $\mathcal{B}$ respectively (we suppose that the basis in the target space of scalars is always the standard one, so we can skip the subscript $\mathcal{S}$ in the notation). Then recalling the change of coordinate rule from Section 8.4 in Chapter 2 we get that

$$[L]_{\mathcal{B}} = [L]_{\mathcal{A}}[I]_{\mathcal{A},\mathcal{B}}.$$

Recall that for a vector $\mathbf{v} \in V$ its coordinates in different bases are related by the formula

$$[\mathbf{v}]_{\mathcal{B}} = [I]_{\mathcal{B},\mathcal{A}}[\mathbf{v}]_{\mathcal{A}},$$

and that

$$[I]_{\mathcal{A},\mathcal{B}} = [I]^{-1}_{\mathcal{B},\mathcal{A}}.$$

If we denote $S := [I]_{\mathcal{B},\mathcal{A}}$, so $[\mathbf{v}]_{\mathcal{B}} = S[\mathbf{v}]_{\mathcal{A}}$, the entries of the vectors $[L]^T_{\mathcal{B}}$ and $[L]^T_{\mathcal{A}}$ are related by the formula

$$[L]^T_{\mathcal{B}} = (S^{-1})^T [L]^T_{\mathcal{A}} \tag{1.1}$$

(since we usually represent a vector as a column of its coordinates, we use $[L]^T_{\mathcal{A}}$ and $[L]^T_{\mathcal{B}}$ instead of $[L]_{\mathcal{A}}$ and $[L]_{\mathcal{B}}$)

Saying it in words

If $S$ is the change of coordinate matrix (from old coordinates to the new ones) in $X$, then the change of coordinate matrix in the dual space $X'$ is $(S^{-1})^T$.

So, the dual space $V'$ of $V$ while isomorphic to $V$ is indeed a different object: the difference is in how the coordinates in $V$ and $V'$ change when one changes the basis in $V$.

**Remark.** One can ask: why can't we pick a basis in $X$ and some completely unrelated basis in the dual $X'$? Of course, we can do that, but imagine, what would it take to compute $L(\mathbf{x})$, knowing coordinates of $\mathbf{x}$ in some basis and coordinates of $L$ in some completely unrelated basis.

So, if we want (knowing the coordinates of a vector $\mathbf{x}$ in some basis) to compute the action of a linear functional $L$ using the standard rules of matrix algebra, i.e. to multiply a row (the functional) by a column (the vector), we have no choice: the "coordinates" of the linear functional $L$ should be the entries of its matrix (in the same basis).

As we can see later, see Section 1.3 below, the entries ("coordinates") of a linear functional are indeed the coordinates in some basis (the so-called *dual* basis.

1.1.2. *A uniqueness theorem.*

**Lemma 1.3.** *Let $\mathbf{v} \in V$. If $L(\mathbf{v}) = 0$ for all $L \in V'$ then $\mathbf{v} = 0$. As a corollary, if $L(\mathbf{v}_1) = L(\mathbf{v}_2)$ for all $L \in V'$, then $\mathbf{v}_1 = \mathbf{v}_2$*

**Proof.** Fix a basis $\mathcal{B}$ in $V$. Then

$$L(\mathbf{v}) = [L]_{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}}.$$

Picking different matrices (i.e. different $L$) we can easily see that $[\mathbf{v}]_{\mathcal{B}} = \mathbf{0}$. Indeed, if

$$L_k = [0, \dots, 0, \underset{k}{1}, 0, \dots, 0]$$

then the equality

$$L_k[\mathbf{v}]_{\mathcal{B}} = 0$$

implies that $k$th coordinate of $[\mathbf{v}]_{\mathcal{B}}$ is 0.

Using this equality for all $k$ we conclude that $[\mathbf{v}]_{\mathcal{B}} = \mathbf{0}$, so $\mathbf{v} = \mathbf{0}$. □

**1.2. Second dual.** As we discussed above, the dual space $V'$ is a vector space, so one can consider its dual $V'' = (V')'$. It looks like one that can consider the dual $V'''$ of $V''$ and so on... However, the fun stops with $V''$ because

The second dual $V''$ is *canonically* (i.e. in a natural way) isomorphic to $V$

Let us decipher this statement. Any vector $\mathbf{v} \in V$ canonically defines a linear functional $L_{\mathbf{v}}$ on $V'$ (i.e. an element of the second dual $V''$ by the rule

$$L_{\mathbf{v}}(f) = f(\mathbf{v}) \qquad \forall f \in V'$$

It is easy to check that the mapping $T : V \to V''$, $T\mathbf{v} = L_{\mathbf{v}}$ is a linear transformation.

Note, that $\operatorname{Ker} T = \{\mathbf{0}\}$. Indeed, if $T\mathbf{v} = \mathbf{0}$, then

$$f(\mathbf{v}) = 0 \qquad \forall f \in V',$$

and by Lemma 1.3 above we have $\mathbf{v} = \mathbf{0}$.

Since $\dim V'' = \dim V' = \dim V$, the condition $\operatorname{Ker} T = \{\mathbf{0}\}$ implies that $T$ is an invertible transformation (isomorphism).

The isomorphism $T$ is very natural, (at least for a mathematician). In particular, it was defined without using a basis, so it does not depend on the choice of basis. So, informally we say that $V''$ is canonically isomorphic to $V$: the rigorous statement is that the map $T$ described above (which we consider to be a natural and canonical) is an isomorphism from $V$ to $V''$.

**1.3. Dual, a.k.a. biorthogonal bases.** In the previous sections, we several times referred to the entries of the matrix of a linear functional $L$ as coordinates. But *coordinates* in this book usually means the coordinates in some basis. Are the "coordinates' of a linear functional really coordinates in some basis? Turns out the answer is "yes", so the terminology remains consistent.

Let us find the basis corresponding to the coordinates of $L \in V'$. Let $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ be a basis in $V$. For $L \in V'$, let $[L]_{\mathcal{B}} = [L_1, L_2, \dots, L_n]$ be its matrix (row) in the basis $\mathcal{B}$. Consider linear functionals $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n \in V'$ defined by

$$\mathbf{b}'_k(\mathbf{b}_j) = \delta_{k,j} \tag{1.2}$$

where $\delta_{k,j}$ is the *Kroneker delta*,

$$\delta_{k.j} = \begin{cases} 1, & j = k \\ 0 & j \neq k \end{cases}$$

Recall, that a linear transformation is defined by its action on a basis, so the functionals $\mathbf{b}'_k$ are well defined.

As one can easily see, the functional $L$ can be represented as

$$L = \sum L_k \mathbf{b}'_k.$$

Indeed, take an arbitrary $\mathbf{v} = \sum_k \alpha_k \mathbf{b}_k \in V$, so $[\mathbf{v}]_{\mathcal{B}} = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$. By linearity and definition of $\mathbf{b}'_k$

$$\mathbf{b}'_k(\mathbf{v}) = \mathbf{b}'_k \left( \sum_j \alpha_j \mathbf{b}_j \right) = \sum_j \alpha_j \mathbf{b}'_k(\mathbf{b}_j) = \alpha_k.$$

Therefore

$$L\mathbf{v} = [L]_{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}} = \sum_k L_k \alpha_k = \sum_k L_k \mathbf{b}'_k(\mathbf{v}).$$

Since this identity holds for all $\mathbf{v} \in V$, we conclude that $L = \sum_k L_k \mathbf{b}'_k$.

Since we did not assume anything about $L \in V'$, we have just shown that any linear functional $L$ can be represented as a linear combination of $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$, so the system $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ is generating.

Let us show that this system is linearly independent (and so it is a basis). Let $\mathbf{0} = \sum_k L_k \mathbf{b}'_k$. Then for an arbitrary $j = 1, 2, \dots, n$

$$0 = \mathbf{0}\mathbf{b}_j = \left( \sum_k L_k \mathbf{b}'_k \right) (\mathbf{b}_j) = \sum_k L_k \mathbf{b}'_k(\mathbf{b}_j) = L_j$$

so $L_j = 0$. Therefore, all $L_k$ are 0 and the system is linearly independent.

So, the system $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ is indeed a basis in the dual space $V'$ and the entries of $[L]_{\mathcal{B}}$ are coordinates of $L$ with respect to the basis $\mathcal{B}$.

**Definition 1.4.** Let $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ be a basis in $V$. The system of vectors

$$\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n \in V',$$

uniquely defined by the equation (1.2) is called the dual (or biorthogonal) basis to $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$.

Note that we have shown that the dual system to a basis is a basis. Note also that in $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ is the dual system to a basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$, then $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is the dual to the basis $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$

1.3.1. *Abstract non-orthogonal Fourier decomposition.* The dual system can be used for computing the coordinates in the basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$. Let $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ be the biorthogonal system to $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$, and let $\mathbf{v} = \sum_k \alpha_k \mathbf{b}_k$. Then, as it was shown before

$$\mathbf{b}'_j(\mathbf{v}) = \mathbf{b}_j \left( \sum_k \alpha_k \mathbf{b}_k \right) = \sum_k \alpha_k \mathbf{b}_j(\mathbf{b}_k) = \alpha_j \mathbf{b}'_j(\mathbf{b}_j) = \alpha_j,$$

so $\alpha_k = \mathbf{b}'_k(\mathbf{v})$. Then we can write

$$\mathbf{v} = \sum_k \mathbf{b}'_k(\mathbf{v}) \mathbf{b}_k. \tag{1.3}$$

In other words,

The $k$th coordinate of a vector $\mathbf{v}$ in a basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ is $\mathbf{b}'_k(\mathbf{v})\}$, where $\mathcal{B}' = \{\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n\}$ is the dual basis.

This formula is called (a baby version of) the *abstract non-orthogonal Fourier decomposition* of $\mathbf{v}$ (in the basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$). The reason for this name will be clear later in Section 2.3.

**Remark 1.5.** Let $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$ and $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_m\}$ be bases in $X$ and $Y$ respectively, and let $\mathcal{B}' = \{\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_m\}$ be the dual basis to $\mathcal{B}$. Then the matrix $[T]_{\mathcal{B},\mathcal{A}} =: A = \{a_{k,j}\}_{k=1}^{m}{}_{j=1}^{n}$ of the transformation $T$ in the bases $\mathcal{A}$, $\mathcal{B}$ is given by

$$a_{k,j} = \mathbf{b}'_k(T\mathbf{a}_j), \qquad j = 1, 2, \dots, n, \quad k = 1, 2, \dots, m.$$

**1.4. Examples of dual systems.** The first example we consider is a trivial one. Let $V$ be $\mathbb{R}^n$ (or $\mathbb{C}^n$) and let $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ be the standard basis there. The dual space will be the space of $n$-dimensional row vectors, which is isomorphic to $\mathbb{R}^n$ (or $\mathbb{C}^n$ in the complex case), and the standard basis there is the dual to $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. The standard basis in $(\mathbb{R}^n)'$ (or in $(\mathbb{C}^n)'$ is $\mathbf{e}_1^T, \mathbf{e}_2^T, \dots, \mathbf{e}_n^T$) obtained from $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ by transposition.

1.4.1. *Taylor formula.* The next example is more interesting. Let us consider the space $\mathbb{P}_n$ of polynomials of degree at most $n$. As we know, the powers $\{\mathbf{e}_k\}_{k=0}^n$, $\mathbf{e}(t) = t^n$ form the standard basis in this space. What is the dual to this basis?

The answer might be tricky to guess, but it is very easy to check when you know it. Namely, consider the linear functionals $\mathbf{e}'_k \in (\mathbb{P}_n)'$, $k = 0, 1, \dots, n$, acting on polynomials as follows:

$$\mathbf{e}'_k(p) := \frac{1}{k!}\frac{d^k}{dt^k}p(t)\,\Big|_{t=0} = \frac{1}{k!}p^{(k)}(0);$$

here we use the usual agreement that $0! = 1$ and $d^0 f/dt^0 = f$.

Since

$$\frac{d^k}{dt^k}j^j = \begin{cases} j(j-1)\dots(j-k+1)t^{j-k}, & k \le j \\ 0 & k > j \end{cases}$$

we can easily see that the system $\{\mathbf{e}'_k\}_{k=0}^n$ is the dual to the system of powers $\{\mathbf{e}_k\}_{k=0}^n$.

Applying (1.3) to the above system $\{\mathbf{e}_k\}_{k=0}^n$ and its dual we get that any polynomial $p$ of degree at most $n$ can be represented as

$$p(t) = \sum_{k=0}^{n} \frac{p^{(k)}(0)}{k!}t^k \tag{1.4}$$

This formula is well-known in Calculus as the Taylor formula for polynomials. More precisely, this is a particular case of the Taylor formula, the

so-called Maclaurin formula. The general Taylor formula

$$p(t) = \sum_{k=0}^{n} \frac{p^{(k)}(a)}{k!}(t-a)^k$$

can be obtained from (1.4) by applying it to the polynomial $p(\tau - a)$ and then denoting $t := \tau - a$. It also can be obtained by considering powers $(t-a)^k$, $k = 0, 1, \dots, n$ and finding the dual system the same way we did it for $t^k$.[2]

1.4.2. *Lagrange interpolation.* Our next example deals with the so-called Lagrange interpolating formula. Let $a_1, a_2, \dots, a_{n+1}$ be distinct points (in $\mathbb{R}$ or $\mathbb{C}$), and let $\mathbb{P}_n$ be the space of polynomials of degree at most $n$. Define functionals $\mathbf{f}_k \in \mathbb{P}_n'$ by

$$\mathbf{f}_k(p) = p(a_k) \qquad \forall p \in \mathbb{P}_n.$$

What is the dual of this system of functionals? Note, that while it is not hard to show that the functionals $\mathbf{f}_k$ are linearly independent, and so (since $\dim(\mathbb{P}_n)' = \dim \mathbb{P}_n = n+1$) form a basis in $(\mathbb{P}_n)'$, we do not need that. We will construct the dual system directly, and then will be able to see that the system $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{n+1}$ is indeed a basis.

Namely, let us define the polynomials $p_k$, $k = 1, 2, \dots, n+1$ as

$$p_k(t) = \prod_{j:j\neq k} (t - a_j) \Big/ \prod_{j:j\neq k} (a_k - a_j)$$

where $j$ in the products runs from 1 to $n+1$. Clearly $p_k(a_k) = 1$ and $p_k(a_j) = 0$ if $j \neq k$, so indeed the system $p_1, p_2, \dots, p_{n+1}$ is dual to the system $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{n+1}$.

There is a little detail here, since the notion of a dual system was defined only for a basis, and we did not prove that either of the systems is one. But one can immediately see that the system $p_1, p_2, \dots, p_{n+1}$ is linearly independent (can you explain why?), and since it contains $n+1 = \dim \mathbb{P}_n$ vectors, it is a basis. Therefore, the system of functionals $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{n+1}$ is also a basis in the dual space $(\mathbb{P}_n)'$.

**Remark.** Note, that we did not just got lucky here, this is a general phenomenon. Namely, as Problem 1.1 below asserts, any system of vectors having a "'dual" one must be linearly independent. So, constructing a dual system is a way of proving linear independence (and an easy one, if you can do it easily as in the above example).

[2] Note, that the general Taylor formula says more than the formula for polynomials obtained here: it says that any $n$ times differentiable function can be approximated near the point $a$ by its Taylor polynomial. Moreover, if the function is $n+1$ times differentiable, it allows us to estimate the error. The above formula for polynomials serves as a motivation and a starting point for the general case

Applying formula (1.3) to the above example one can see that the (unique) polynomial $p$, $\deg p \le n$ satisfying

$$p(a_k) = y_k, \qquad k = 1, 2, \dots, n+1 \tag{1.5}$$

can be reconstructed by the formula

$$p(t) = \sum_{k=1}^{n+1} y_k p_k(t). \tag{1.6}$$

This formula is well -known in mathematics as the "Lagrange interpolation formula".

**Exercises.**

**1.1.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ be a system of vectors in $X$ such that there exists a system $\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_r$ of linear functionals such that

$$\mathbf{v}'_k(\mathbf{v}_j) = \begin{cases} 1, & j = k \\ 0 & j \neq k \end{cases}$$

a) Show that the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ is linearly independent.

b) Show that if the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ is not generating, then the "biorthogonal" system $\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_r$ is not unique. **Hint:** Probably the easiest way to prove that is to complete the system $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ to a basis, see Proposition 5.4 from Chapter 2

**1.2.** Prove that given distinct points $a_1, a_2, \dots, a_{n+1}$ and values $y_1, y_2, \dots, y_{n+1}$ (not necessarily distinct) the polynomial $p$, $\deg p \le n$ satisfying (1.5) is unique. Try to prove it using the ideas from linear algebra, and not what you know about polynomials.

## 2. Dual of an inner product space

Let us recall that there is no inner product space over an arbitrary field, that all our inner product spaces are either real or complex.

### 2.1. Riesz representation theorem.

**Theorem 2.1** (Riesz representation theorem)**.** *Let $H$ be an inner product space. Given a linear functional $L$ on $H$ there exists a unique vector $\mathbf{y} \in H$ such that*

$$L(\mathbf{v}) = (\mathbf{v}, \mathbf{y}) \qquad \forall \mathbf{v} \in H. \tag{2.1}$$

**Proof.** Fix an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ in $H$, and let

$$[L] = [L_1, L_2, \dots, L_n]$$

be the matrix of $L$ in this basis. Define vector $\mathbf{y}$ by

$$\mathbf{y} := \sum_k \overline{L}_k \mathbf{e}_k, \tag{2.2}$$

where $\overline{L}_k$ denotes the complex conjugate of $L_k$. In the case of a real space the conjugation does nothing and can be simply ignored.

We claim that $\mathbf{y}$ satisfies (2.1).

Indeed, take an arbitrary vector $\mathbf{v} = \sum_k \alpha_k \mathbf{e}_k$. Then

$$[\mathbf{v}] = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$$

and

$$L(\mathbf{v}) = [L][\mathbf{v}] = \sum_k L_k \alpha_k.$$

On the other hand

$$(\mathbf{v}, \mathbf{y}) = \sum_k \alpha_k \overline{\overline{L}}_k = \sum_k \alpha_k L_k$$

so (2.1) holds.

Recall that if we know coordinates of 2 vectors in an orthonormal basis, we can compute the inner product by taking these coordinate and computing the standard inner product in $\mathbb{C}^n$ (or $\mathbb{R}^n$).

To show that the vector $\mathbf{y}$ is unique, let us assume that $\mathbf{y}$ satisfies (2.1). Then for $k = 1, 2, \dots, n$

$$(\mathbf{e}_k, \mathbf{y}) = L(\mathbf{e}_k) = L_k,$$

so $(\mathbf{y}, \mathbf{e}_k) = \overline{L}_k$. Then, using the formula for the decomposition in the orthonormal basis, see Section 2.1 of Chapter 5 we get

$$\mathbf{y} = \sum_k (\mathbf{y}, \mathbf{e}_k)\mathbf{e}_k = \sum_k \overline{L}_k \mathbf{e}_k$$

which means that any vector satisfying (2.1) must be represented by (2.2). □

**Remark.** While the statement of the theorem does not require a basis, the proof presented above utilizes an orthonormal basis in $H$, although the resulting vector $\mathbf{y}$ does not depend on the choice of the basis[3]. An advantage of this proof is that it gives a formula for computing the representing vector $\mathbf{y}$.

**2.2. Is an inner product space a dual to itself?** For a vector $\mathbf{y}$ in an inner product space $H$ one can define a linear functional $L_{\mathbf{y}}$,

$$L_{\mathbf{y}}(\mathbf{v}) := (\mathbf{v}, \mathbf{y}).$$

It is easy to see that the mapping $\mathbf{y} \mapsto L_{\mathbf{y}}$ is an injective mapping from $H$ to sits dual $H^*$. The above Theorem 2.1 asserts that this mapping is a surjection (onto), so one is tempted to say that the dual of an inner product space $H$ is (canonically isomorphic to) the space $H$ itself, with the canonical isomorphism given by $\mathbf{y} \mapsto L_{\mathbf{y}}$.

[3] An alternative proof that does need a basis is also possible. This alternative proof, that works in infinite-dimensional case, uses strong convexity of the unit ball in the inner product space together with the idea of completeness from analysis.

This is indeed the case if $H$ is a *real* inner product space and it is easy to show that the map $\mathbf{y} \mapsto L_{\mathbf{y}}$ is a *linear* transformation. We already discussed that the map is injective and surjective, so it is an invertible linear transformations, i.e. an *isomorphism*.

However if $H$ is a complex space, one needs to be a bit more careful. Namely, the mapping $\mathbf{y} \mapsto L_{\mathbf{y}}$ that that maps a vector $\mathbf{y} \in H$ to the linear functional $L_{\mathbf{y}}$ as in Theorem 2.1 ($L_{\mathbf{y}}(\mathbf{v}) = (\mathbf{v}, \mathbf{y})$) is not a linear one.

More precisely, while it is easy to show that

$$L_{\mathbf{y}_1+\mathbf{y}_2} = L_{\mathbf{y}_1} + L_{\mathbf{y}_2}, \tag{2.3}$$

it follows from the definition of $L_{\mathbf{y}}$ and properties of inner product that

$$L_{\alpha\mathbf{y}}(\mathbf{v}) = (\mathbf{v}, \alpha\mathbf{y}) = \overline{\alpha}(\mathbf{v}, \mathbf{y}) = \overline{\alpha} L_{\mathbf{y}}(\mathbf{v}), \tag{2.4}$$

so $L_{\alpha\mathbf{y}} = \overline{\alpha} L_{\mathbf{y}}$.

In other words, one can say that the dual of a complex inner product space is the space itself but with the *different linear structure*: adding 2 vectors is equivalent to adding corresponding linear functionals, but multiplying a vector by $\alpha$ is equivalent to multiplying the corresponding functional by $\overline{\alpha}$.

A transformation $T$ satisfying $T(\alpha\mathbf{x} + \beta\mathbf{y}) = \overline{\alpha}T\mathbf{x} + \overline{\beta}T\mathbf{y}$ is sometimes called a *conjugate linear* transformation.

So, for a complex inner product space $H$ its dual can be canonically identified with $H$ by a conjugate linear isomorphism (i.e. invertible conjugate linear transformation)

Of course, for a real inner product space the complex conjugation can be simply ignored (because $\alpha$ is real), so the map $\mathbf{y} \mapsto L_{\mathbf{y}}$ is a linear one. In this case we can, indeed say that the dual of an inner product space $H$ is the space itself.

In both, real and complex cases, we nevertheless can say that the dual of an inner product space can be canonically identified with the space itself.

### 2.3. Biorthogonal systems and orthonormal bases.

**Definition 2.2.** Let $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ be a basis in an inner product space $H$. The unique system $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ in $H$ defined by

$$(\mathbf{b}_j, \mathbf{b}'_k) = \delta_{j,k},$$

where $\delta_{j,k}$ is the Kroneker delta, is called the *biorthogonal* or *dual* to the basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$.

This definition clearly agrees with Definition 1.4, if one identifies the dual $H'$ with $H$ as it was discussed above. Then it follows immediately

from the discussion in Section 1.3 that the dual system $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ to a basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ is uniquely defined and forms a basis, and that the dual to $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$ is $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$.

The abstract non-orthogonal Fourier decomposition formula (1.3) can be rewritten as

$$\mathbf{v} = \sum_{k=1}^{n} (\mathbf{v}, \mathbf{b}'_k)\mathbf{b}_k$$

Note, that an orthonormal basis is dual to itself. So, if $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is an orthonormal basis, the above formula is rewritten as

$$\mathbf{v} = \sum_{k=1}^{n} (\mathbf{v}, \mathbf{e}_k)\mathbf{e}_k$$

which is the classical (orthogonal) abstract Fourier decomposition, see formula (2.2) in Section 2.1 of Chapter 5.

## 3. Adjoint (dual) transformations and transpose. Fundamental subspace revisited (once more)

By analogy with the case of an inner product spaces, see Theorem 2.1, it is customary to write $L(\mathbf{v})$, where $L$ is a linear functional (i.e. $L \in V'$, $\mathbf{v} \in V$) in the form resembling inner product

$$L(\mathbf{v}) = \langle \mathbf{v}, L \rangle$$

Note, that the expression $\langle \mathbf{v}, L \rangle$ is linear in both arguments, unlike the inner product which in the case of a complex space is linear in the first argument and conjugate linear in the second. So, to distinguish it from the inner product, we use the angular brackets.[4]

Note also, that while in the inner product both vectors belong to the same space, $\mathbf{v}$ and $L$ above belong to different spaces: in particular, we cannot add them.

### 3.1. Dual (adjoint) transformation.

**Definition 3.1.** Let $A : X \to Y$ be a linear transformation. The transformation $A' : Y' \to X'$ ($X'$ and $Y'$ are dual spaces for $X$ and $Y$ respectively) such that

$$\langle A\mathbf{x}, \mathbf{y}' \rangle = \langle \mathbf{x}, A'\mathbf{y}' \rangle \qquad \forall \mathbf{x} \in X, \mathbf{y}' \in Y'$$

is called the adjoint (dual) to $A$.

[4]This notation, while widely used, is far from the standard. Sometimes $(\mathbf{v}, L)$ is used, sometimes the angular brackets are used for the inner product. So, encountering expression like that in the text, one has to be very careful to distinguish inner product from the action of a linear functional.

Of course, it is not a priori clear why the transformation $A'$ exists. Below we will show that indeed such transformation exists, and moreover, it is unique.

3.1.1. *Dual transformation for the case $A : \mathbb{F}^n \to \mathbb{F}^m$.* Let us first consider the case when $X = \mathbb{F}^n$, $Y = \mathbb{F}^m$ ($\mathbb{F}$ here is, as usual, either $\mathbb{R}$ or $\mathbb{C}$, but everything works for the case of arbitrary fields)

As usual, we identify a vector $\mathbf{v}$ in $\mathbb{F}^n$ with the column of its coordinates, and a linear transformation with it matrix (in the standard basis).

The dual of $\mathbb{F}^n$ is, as it was discussed above, the space of rows of size $n$, so we can identify its with $\mathbb{F}^n$. Again, we will treat an element of $(\mathbb{F}^n)'$ as a column vector of its coordinates.

Under these agreements we have for $\mathbf{x} \in \mathbb{F}^n$ and $\mathbf{x}' \in (\mathbb{F}^n)'$

$$\mathbf{x}'(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}' \rangle = (\mathbf{x}')^T \mathbf{x}$$

where the right side is the product of matrices (or a row and a column). Then, for arbitrary $\mathbf{x} \in X = \mathbb{F}^n$ and $\mathbf{y}' \in Y' = (\mathbb{F}^m)'$

$$\langle A\mathbf{x}, \mathbf{y}' \rangle = (\mathbf{y}')^T A\mathbf{x} = (A^T\mathbf{y}')^T \mathbf{x} = \langle \mathbf{x}, A^T\mathbf{y} \rangle$$

(the expressions in the middle are products of matrices).

So we have proved that the adjoint transformation exists. let us show that it is unique. Assume that for some transformation $B$

$$\langle A\mathbf{x}, \mathbf{y}' \rangle = \langle \mathbf{x}, B\mathbf{y}' \rangle \qquad \forall \mathbf{x} \in \mathbb{F}^n, \forall \mathbf{y}' \in (\mathbb{F}^m)'.$$

That means that for arbitrary

$$\langle \mathbf{x}, (A^T - B)\mathbf{y}' \rangle = 0, \qquad \forall \mathbf{x} \in \mathbb{F}^n, \forall \mathbf{y}' \in (\mathbb{F}^m)'$$

By taking for $\mathbf{x}$ and $\mathbf{y}'$ the vectors from the standard bases in $\mathbb{F}^n$ and $(\mathbb{F}^m)' \cong \mathbb{F}^m$ respectively we get that the matrices $B$ and $A^T$ coincide. □

So, for $X = \mathbb{F}^n$, $Y = \mathbb{F}^m$

| The dual transformation $A'$ exists, and is unique. Moreover, its matrix (in the standard bases) equals $A^T$ (the transpose of the matrix of $A$) |
|---|

3.1.2. *Dual transformation in the abstract setting.* Now, let us consider the general case. In fact, we do not need to do much, since everything can be reduced to the case of spaces $\mathbb{F}^n$.

Namely, let us fix bases $\mathcal{A} = \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ in $X$, and $\mathcal{B} = \mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m$ in $Y$, and let $\mathcal{A}' = \mathbf{a}'_1, \mathbf{a}'_2, \ldots, \mathbf{a}'_n$ and $\mathcal{B} = \mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_m$ be their dual bases (in $X'$ and $Y'$ respectively). For a vector $\mathbf{v}$ (from a space or its dual) we as usual denote by $[\mathbf{v}]_{\mathcal{B}}$ the column of its coordinates in the basis $\mathcal{B}$. Then

$$\langle \mathbf{x}, \mathbf{x}' \rangle = ([\mathbf{x}']_{\mathcal{A}'})^T [\mathbf{x}]_{\mathcal{A}}, \qquad \forall \mathbf{x} \in X \quad \forall \mathbf{x}' \in X',$$

i.e. instead of working with $\mathbf{x} \in X$ and $\mathbf{x}' \in X'$ we can work with columns their coordinates (in the dual bases $\mathcal{A}$ and $\mathcal{A}'$ respectively) absolutely the same way we do in in the case of $\mathbb{F}^n$. Of course, the same works for $Y$, so working with columns of coordinates and then translating everything back to the abstract setting we get that the dual transformation exists in unique in this case as well.

Moreover, using the fact (which we just proved) that for $A : \mathbb{F}^n \to \mathbb{F}^m$ the matrix of $A'$ is $A^T$ we get

$$[A']_{\mathcal{A}',\mathcal{B}'} = ([A]_{\mathcal{B},\mathcal{A}})^T, \tag{3.1}$$

or in plain English

| The matrix of the dual transformation in the dual basis is the transpose of the matrix of the transformation in the original bases. |
|---|

**Remark 3.2.** Note, that while we used basis to construct the dual transformation, the resulting transformation does not depend on the choice of a basis.

3.1.3. *A coordinate-free way to define the dual transformation.* Let us now present another, more "high brow" way of defining the dual of a linear transformation. Namely, for $\mathbf{x} \in X$, $\mathbf{y}' \in Y$ let us fix for a moment $\mathbf{y}'$ and treat the expression $\langle A\mathbf{x}, \mathbf{y}' \rangle = \mathbf{y}'(A\mathbf{x})$ as a function of $\mathbf{x}$. It is easy to see that this is a composition of two linear transformations (which ones?) and so it is a linear function of $\mathbf{x}$, i.e. a linear functional on $X$, i.e. an element of $X'$.

Let us call this linear functional $B(\mathbf{y}')$ to emphasize the fact that it depends on $\mathbf{y}'$. Since we can do this for every $\mathbf{y}' \in Y'$, we can define the transformation $B : Y' \to X'$ such that

$$\langle A\mathbf{x}, \mathbf{y}' \rangle = \langle \mathbf{x}, B(\mathbf{y}') \rangle$$

Our next step is to show that $B$ is a linear transformation. Note, that since the transformation $B$ was defined in rather indirect way, we cannot see immediately from the definition that it is linear. To show the linearity of $B$ let us take $\mathbf{y}'_1, \mathbf{y}'_2 \in Y'$. For $\mathbf{x} \in X$

$$\begin{aligned}
\langle \mathbf{x}, B(\alpha\mathbf{y}'_1 + \beta\mathbf{y}'_2) \rangle &= \langle A\mathbf{x}, \alpha\mathbf{y}'_1 + \beta\mathbf{y}'_2 \rangle && \text{by the definition of } B \\
&= \alpha\langle A\mathbf{x}, \mathbf{y}'_1 \rangle + \beta\langle A\mathbf{x}, \mathbf{y}'_2 \rangle && \text{by linearity} \\
&= \alpha\langle \mathbf{x}, B(\mathbf{y}'_1) \rangle + \beta\langle \mathbf{x}, B(\mathbf{y}'_2) \rangle && \text{by the definition of } B \\
&= \langle \mathbf{x}, \alpha B(\mathbf{y}'_1) + \beta B(\mathbf{y}'_2) \rangle && \text{by linearity}
\end{aligned}$$

Since this identity is true for all $\mathbf{x}$, we conclude that $B(\alpha\mathbf{y}'_1 + \beta\mathbf{y}'_2) = \alpha B(\mathbf{y}'_1) + \beta B(\mathbf{y}'_2)$, i.e. that $B$ is linear.

The main advantage of this approach that it does not require a basis, so it can be (and is) used in the infinite-dimensional situation. However, the proof that we presented above in Sections 3.1.1, 3.1.2 gives a constructive way to compute the dual transformation, so we used that proof instead of more general coordinate-free one.

**Remark 3.3.** Note, that the above coordinate-free approach can be used to define the Hermitian adjoint of as operator in an inner product space. The only addition to the reasoning presented above will be the use of the Riesz Representation Theorem (Theorem 2.1). We leave the details as an exercise to the reader, see Problem 3.2 below.

### 3.2. Annihilators and relations between fundamental subspaces.

**Definition 3.4.** Let $X$ be a vector space and let $E \subset X$. The *annihilator* of $E$, denoted by $E^\perp$ is the set of all $\mathbf{x}' \in X'$ such that $\langle \mathbf{x}, \mathbf{x}' \rangle = 0$ for all $\mathbf{x} \in E$.

Using the fact that $X''$ is canonically isomorphic to $X$ (see Section 1.2) we say that for $E \subset X'$ its *annihilator* $E^\perp$ consists of all vectors $\mathbf{x} \in X$ such that $\langle \mathbf{x}, \mathbf{x}' \rangle = 0$ for all $\mathbf{x}' \in E$.

**Remark 3.5.** Formally speaking, for $E \subset X'$ the set $E^\perp$ should be defined as the set of all $\mathbf{x}'' \in X''$ such that $\langle \mathbf{x}', \mathbf{x}'' \rangle = 0$ for all $\mathbf{x}' \in E$; the symbol $E_\perp$ is often used for the annihilator from the second part of Definition 3.4. However, because of the natural isomorphism of $X''$ and $X$ there is no real difference between these two cases, so we will always use $E^\perp$.

Distinguishing the cases $E \subset X$ and $E \subset X'$ makes a lot of sense in the infinite-dimensional situation, where $X''$ is not always canonically isomorphic to $X$.

The spaces such that $X''$ canonically isomorphic to $X$ have a special name: they are called *reflexive* spaces.

**Proposition 3.6.** *Let $E$ be a subspace of $X$. Then $(E^\perp)^\perp = E$*

This proposition looks absolutely like Proposition 3.6 from Chapter 5. However its proof is a bit more complicated, since the suggested proof of Proposition 3.6 from Chapter 5 heavily used the inner product space structure: it used the decomposition $X = E \oplus E^\perp$, which is not true in our situation because, for example, $E$ and $E^\perp$ are in different spaces.

**Proof.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ be a basis in $E$ (recall that all spaces in this chapter are assumed to be finite-dimensional), so $E = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$.

By Proposition 5.4 from Chapter 2 the system can be extended to a basis in all $X$, i.e. one can find vectors $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ ($n = \dim X$) such that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis in $X$.

Let $\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_n$ be the dual basis to $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$. By Problem 3.3 $E^\perp = \operatorname{span}\{\mathbf{v}'_{r+1}, \dots, \mathbf{v}'_n\}$. Applying again this problem to $E^\perp$ we get that

$$(E^\perp)^\perp = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} = E.$$

□

The following theorem is analogous to Theorem 5.1 from Chapter 5

**Theorem 3.7.** *Let $A : X \to Y$ be an operator acting from one vector space to another. Then*

a) $\operatorname{Ker} A' = (\operatorname{Ran} A)^\perp$;
b) $\operatorname{Ker} A = (\operatorname{Ran} A')^\perp$;
c) $\operatorname{Ran} A = (\operatorname{Ker} A')^\perp$;
d) $\operatorname{Ran} A' = (\operatorname{Ker} A)^\perp$.

**Proof.** First of all, let us notice, that since for a subspace $E$ we have $(E^\perp)^\perp = E$, the statements 1 and 3 are equivalent. Similarly, for the same reason, the statements 2 and 4 are equivalent as well. Finally, statement 2 is exactly statement 1 applied to the operator $A'$ (here we use the trivial fact fact that $(A')' = A$, which is true, for example, because of the corresponding fact for the transpose).

So, to prove the theorem we only need to prove statement 1.

Recall that $A' : Y' \to X'$. The inclusion $\mathbf{y}' \in (\operatorname{Ran} A)^\perp$ means that $\mathbf{y}'$ annihilates all vectors of the form $A\mathbf{x}$, i.e. that

$$\langle A\mathbf{x}, \mathbf{y}' \rangle = 0 \qquad \forall \mathbf{x} \in X.$$

Since $\langle A\mathbf{x}, \mathbf{y}' \rangle = \langle \mathbf{x}, A'\mathbf{y}' \rangle$, the last identity is equivalent to

$$\langle \mathbf{x}, A'\mathbf{y}' \rangle = 0 \qquad \forall \mathbf{x} \in X.$$

But that means that $A'\mathbf{y}' = \mathbf{0}$ ($A'\mathbf{y}'$ is a zero functional).

So we have proved that $\mathbf{y}' \in (\operatorname{Ran} A)^\perp$ iff $A'\mathbf{y}' = \mathbf{0}$, or equivalently iff $\mathbf{y}' \in \operatorname{Ker} A'$. □

**Exercises.**

**3.1.** Prove that if for linear transformations $T, T_1 : X \to Y$

$$\langle T\mathbf{x}, \mathbf{y}' \rangle = \langle T_1\mathbf{x}, \mathbf{y}' \rangle$$

for all $\mathbf{x} \in X$ and for all $\mathbf{y}' \in Y'$, then $T = T_1$.

Probably one of the easiest ways of proving this is to use Lemma 1.3.

**3.2.** Combine the Riesz Representation Theorem (Theorem 2.1) with the reasoning in Section 3.1.3 above to present a coordinate-free definition of the Hermitian adjoint of an operator in an inner product space.

The next problem gives a way to prove Proposition 3.6

**3.3.** Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be a basis in $X$ and let $\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_n$ be its dual basis. Let $E := \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$, $r < n$. Prove that $E^\perp = \operatorname{span}\{\mathbf{v}'_{r+1}, \dots, \mathbf{v}'_n\}$.

**3.4.** Use the previous problem to prove that for a subspace $E \subset X$

$$\dim E + \dim E^\perp = \dim X.$$

## 4. What is the difference between a space and its dual?

We know that the dual space $X'$ has the same dimension as $X$, so the space and its dual are isomorphic. So one can think that really there is no difference between the space and its dual. However, as we discussed above in Section 1.1, when we change basis in the space $X$ the coordinates in $X$ and in $X'$ change according to different rules, see formula (1.1) above.

On the other hand, using the natural isomorphism of $X$ and $X''$ we can say that $X$ is the dual of $X'$. From this point of view, there is no difference between $X$ and $X'$: we can start from $X$ and say that $X'$ is its dual, or we can do it the other way around and start from $X'$.

We already used this point of view above, for example in the proof of Theorem 3.7.

Note also, that the change of coordinate formula (1.1) (see also the boxed statement below it) agrees with this point of view: if $\widetilde{S} := (S^{-1})^T$, then $(\widetilde{S}^{-1})^T = S$, so we get the change of coordinate formula in $X$ from the one in $X'$ by the same rule!

**4.1. Isomorphisms between $X$ and $X'$.** There are infinitely many possibilities to define an isomorphism between $X$ and $X'$.

If $X = \mathbb{F}^n$ then the most natural way to identify $X$ and $X'$ is to identify the standard basis in $\mathbb{F}^n$ with the one in $(\mathbb{F}^n)'$. In this case the action of a linear functional will be given by the "inner product type" expression

$$\langle \mathbf{v}, \mathbf{v}' \rangle = (\mathbf{v}')^T \mathbf{v}.$$

To generalize this to the general case one has to fix a basis $\mathcal{B} = \mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ in $X$ and consider the dual basis $\mathcal{B}' = \mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$, and define an isomorphism $T : X \to X'$ by $T\mathbf{b}_k = \mathbf{b}'_k$, $k = 1, 2, \dots, n$.

This isomorphism is natural in some sense, but it depends on the choice of the basis, so in general there is no natural way to identify $X$ and $X'$.

The exception to this is case when $X$ is a real inner product space: the Riesz representation theorem (Theorem 2.1) gives a natural way to identify a linear functional with a vector in $X$. Note that this approach works only for *real* inner product spaces. In the complex case, the Riesz representation

theorem gives a natural identification of $X$ and $X'$, but this identification is not linear but *conjugate linear*.

**4.2. An example: velocities (differential operators) and differential forms as vectors and linear functionals.** To illustrate the relations between vectors and linear functional, let us consider an example from multivariable calculus, which gives rise to important ideas like tangent and cotangent bundles in differential geometry.

Let us recall the notion of the path integral (of the second kind) from the calculus. Recall that a path $\gamma$ in $\mathbb{R}^n$ is defined by its parameterization, i.e. by a function

$$t \mapsto \mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$$

acting from an interval $[a, b]$ to $\mathbb{R}^n$. If $\omega$ is the so-called *differential form* (differential 1-form),

$$\omega = f_1(\mathbf{x})dx_1 + f_2(\mathbf{x})dx_2 + \ldots + f_n(\mathbf{x})dx_n,$$

the *path integral*

$$\int_\gamma \omega = \int_\gamma f_1 dx_1 + f_2 dx_2 + \ldots + f_n dx_n$$

is computed by substituting $\mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$ in the expression, i.e. $\int_\gamma \omega$ is computed as

$$\int_a^b \left( f_1(\mathbf{x}(t))\frac{dx_1(t)}{dt} + f_2(\mathbf{x}(t))\frac{dx_2(t)}{dt} + \ldots + f_n(\mathbf{x}(t))\frac{dx_n(t)}{dt} \right) dt.$$

In other words, at each moment $t$ we have to evaluate the velocity

$$\mathbf{v} = \frac{d\mathbf{x}(t)}{dt} = \left( \frac{dx_1(t)}{dt}, \frac{dx_2(t)}{dt}, \dots, \frac{dx_n(t)}{dt} \right)^T,$$

apply to it the linear functional $\mathbf{f} = (f_1, f_2, \dots, f_n)$, $\mathbf{f}(\mathbf{v}) = \sum_{k=1}^n f_k v_k$ (here $f_k = f_k(\mathbf{x}(t))$ but for a fixed $t$ each $f_k$ is just a number, so we simply write $f_k$), and then integrate the result (which depends on $t$) with respect to $t$.

4.2.1. *Velocities as vectors.* Let us fix $t$ and analyze $\mathbf{f}(\mathbf{v})$. We will show that according to the rules of Calculus, the coordinates of $\mathbf{v}$ change as coordinates of a vector, and the coordinates of $\mathbf{f}$ as the coordinates of a linear functional (covector). Let us assume as it is customary in Calculus, that $x_k$ are the coordinates in the standard basis in $\mathbb{R}^n$, and let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ be a different basis in $\mathbb{R}^n$. We will use notation $\widetilde{x}_k$ to denote the coordinates of a vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, i.e. $[\mathbf{x}]_\mathcal{B} = (\widetilde{x}_1, \widetilde{x}_2, \dots, \widetilde{x}_n)^T$.

Let $A = \{a_{k,j}\}_{k,j=1}^n$ be the change of coordinates matrix, $A = [I]_{\mathcal{B},\mathcal{S}}$, so the new coordinates $\widetilde{x}_k$ are expressed in terms of the old ones as

$$\widetilde{x}_k = \sum_{j=1}^n a_{k,j} x_j, \qquad k = 1, 2, \dots, n.$$

So the new coordinates $\widetilde{v}_k$ of the vector $\mathbf{v}$ are obtained from its old coordinates $v_k$ as

$$\widetilde{v}_k = \sum_{j=1}^n a_{k,j} v_j, \qquad k = 1, 2, \dots, n.$$

4.2.2. *Differential forms as linear functionals (covectors).* Let us now calculate the differential form

$$\omega = \sum_{k=1}^n f_k dx_k \tag{4.1}$$

in terms of new coordinates $\widetilde{x}_k$. The change of coordinates matrix from the new to the old ones is $A^{-1}$. Let $A^{-1} = \{\widetilde{a}_{k,j}\}_{k,j=1}^n$, so

$$x_k = \sum_{j=1}^n \widetilde{a}_{k,j} \widetilde{x}_j, \quad \text{and} \quad dx_k = \sum_{j=1}^n \widetilde{a}_{k,j} d\widetilde{x}_j, \qquad k = 1, 2, \dots, n.$$

Substituting this into (4.1) we get

$$\begin{aligned} \omega &= \sum_{k=1}^n f_k \sum_{j=1}^n \widetilde{a}_{k,j} d\widetilde{x}_j \\ &= \sum_{j=1}^n \left( \sum_{k=1}^n \widetilde{a}_{k,j} f_k \right) d\widetilde{x}_j \\ &= \sum_{j=1}^n \widetilde{f}_j d\widetilde{x}_j \end{aligned}$$

where

$$\widetilde{f}_j = \sum_{k=1}^n \widetilde{a}_{k,j} f_k.$$

But that is exactly the change of coordinate rule for the dual space! So

| according to the rules of Calculus, the coefficients of a differential 1-form change by the same rule as coordinates in the dual space |
|---|

So, according to the accepted rules of Calculus, the coordinates of velocity $\mathbf{v}$ change as coordinates of a vector and coefficients (coordinates) of a differential 1-form change as the entries of a linear functional. In the differential the set of all velocities is called the *tangent space*, and the set of all differential 1 forms is its dual and is called the *cotangent space*.

4.2.3. *Differential operators as vectors.* As we discussed above, in differential geometry vectors are represented by velocities, i.e. by the derivatives $d\mathbf{x}(t)/dt$. This is a simple and intuitively clear point of view, but sometimes it is viewed as a bit naïve.

More "highbrow" point of view, also used in differential geometry (although in more advanced texts) is that vectors are represented by a differential operators

$$D = \sum_k v_k \frac{\partial}{\partial x_k}. \tag{4.2}$$

The informal reason for that is the following. Suppose we want to compute the derivative of a function $\Phi$ along the path given by the function $t \mapsto \mathbf{x}(t)$, i.e. the derivative

$$\frac{d\Phi(\mathbf{x}(t))}{dt}.$$

By the Chain Rule, at a given time $t$

$$\frac{d\Phi(\mathbf{x}(t))}{dt} = \sum_{k=1}^{n} \left( \frac{\partial \Phi}{\partial x_k} \Big|_{\mathbf{x}=\mathbf{x}(t)} \right) x_k'(t) = D\Phi \Big|_{\mathbf{x}=\mathbf{x}(t)},$$

where the differential operator $D$ is given by (4.2) with $v_k = x_k'(t)$.

Of course, we need to show that the coefficient $v_k$ of a differential form change according to the change of coordinate rule for vectors. This is intuitively clear, and can be easily shown by using the multivariable Chain Rule. We leave this as an exercise for the reader, see Problem 4.1 below.

**4.3. The case of a real inner product space.** As we already discussed above, it follows from the Riesz Representation Theorem (Theorem 2.1) that a real inner product space $X$ and its dual $X'$ are canonically isomorphic. Thus we can say that vectors and functionals live in the same space which makes things both simpler and more confusing.

**Remark.** First of all let us note, that if the change of coordinates matrix $S$ is orthogonal ($S^{-1} = S^T$), then $(S^{-1})^T = S$. Therefore, for an orthogonal change of coordinate matrix the coordinates of a vector and of a linear functional change according to the same rule, so one cannot really see a difference between a vector and a functional.

The change of coordinate matrix is orthogonal, for example, if we change from one orthonormal basis to another.

4.3.1. *Einstein notation, metric tensor.* Let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ be a basis in a real inner product space $X$ and let $\mathcal{B}' = \{\mathbf{b}_1', \mathbf{b}_2', \dots, \mathbf{b}_n'\}$ be the dual basis (we identify the dual space $X'$ with $X$ via Riesz Representation Theorem, so $\mathbf{b}_k'$ can be assumed to be in $X$).

Here we present the standard in differential geometry notation (the so-called Einstein notation) for working with coordinates in these bases. Since we will only be working with coordinates, we can assume that we are working in the space $\mathbb{R}^n$ with the non-standard inner product $(\,\cdot\,,\,\cdot\,)_G$ defined by the positive definite matrix $G = \{g_{j,k}\}_{j,k=1}^n$, $g_{j,k} = (\mathbf{b}_k, \mathbf{b}_j)_X$, which is often called the *metric tensor*,

$$(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})_G = \sum_{j=1}^{n}\sum_{k=1}^{n} g_{j,k} x_j y_k, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \tag{4.3}$$

(see Section 5 in Chapter 7).

To distinguish between vectors and linear functionals (co-vectors) it is agreed to write the coordinates of a vector with indices as superscripts and the coordinates a a linear functional with indices as subscripts: thus $x^j$, $j = 1, 2, \dots, n$ denotes the coordinates of a vector $\mathbf{x}$ and $f_k$, $k = 1, 2, \dots, n$ denotes the coordinates of a linear functional $\mathbf{f}$.

**Remark.** Putting indices as superscripts can be confusing, since one will need to distinguish it from the power. However, this is a standard and widely used notation, so we need to get acquainted with it. While I personally, like a lot of mathematicians, prefer using coordinate-free notation, all final computations are done in coordinates, so the coordinate notation has to be used. And as far as coordinate notations go, you will see that this notation is quite convenient to work with.

Another convention in the Einstein notation is that whenever in a product the same index appear in the subscript and superscript, it means one needs to sum up in this index. Thus $x^j f_j$ means $\sum_j x^j f_j$, so we can write $\mathbf{f}(\mathbf{x}) = x^j f_j$. The same convention holds when we have more than one index of summation, so (4.3) can be rewritten in this notation as

$$(\mathbf{x}, \mathbf{y}) = g_{j,k} x^k y^j, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \tag{4.4}$$

(mathematicians are lazy and are always trying to avoid writing extra symbols, whenever they can).

Finally, the last convention in the Einstein notation is the *preservation of the position* of the indices: if we do not sum over an index, it remains in the same position (subscript or superscript) as it was before. Thus we can write $y^j = a^j_k x^k$, but not $f_j = a^j_k x^k$, because the index $j$ must remain as a superscript.

Note, that to compute the inner product of 2 vectors, knowing their coordinates is not sufficient. One also needs to know the matrix $G$ (which is often called the *metric tensor*). This agrees with the Einstein notation: if we try to write $(\mathbf{x}, \mathbf{y})$ as the standard inner product, the expression $x_k y_k$

means just the product of coordinates, since for the summation we need the same index both as the subscript and the superscript. The expression (4.4), on the other hand, fit this convention perfectly.

4.3.2. *Covariant and contravariant coordinates. Lovering and raising the indices.* Let us recall that we have a basis $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ in a real inner product space, and that $\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n$, $\mathbf{b}'_k \in X$ is its dual basis (we identify $X$ with its dual $X'$ via Riesz Representation Theorem, so $\mathbf{b}'_k$ are in $X$).

Given a vector $\mathbf{x} \in X$ it can be represented as

$$\mathbf{x} = \sum_{k=1}^{n} (\mathbf{x}, \mathbf{b}'_k)\mathbf{b}_k =: \sum_{k=1}^{n} x^k \mathbf{b}_k, \qquad \text{and as} \tag{4.5}$$

$$\mathbf{x} = \sum_{k=1}^{n} (\mathbf{x}, \mathbf{b}_k)\mathbf{b}'_k =: \sum_{k=1}^{n} x_k \mathbf{b}'_k. \tag{4.6}$$

The coordinates $x_k$ are called the *covariant* coordinates of the vector $\mathbf{x}$ and the coordinates $x^k$ are called the *contravariant* coordinates.

Now let us ask ourselves a question: how can one get covariant coordinates of a vector from the contravariant ones?

According to the Einstein notation, we use the contravariant coordinates working with vectors, and covariant ones for linear functionals (i.e. when we interpret a vector $\mathbf{x} \in X$ as a linear functional). We know (see (4.6)) that $x_k = (\mathbf{x}, \mathbf{b}_k)$, so

$$x_k = (\mathbf{x}, \mathbf{b}_k) = \Big(\sum_j x^j \mathbf{b}_j, \mathbf{b}_k\Big) = \sum_j x^j (\mathbf{b}_j, \mathbf{b}_k) = \sum_j g_{k,j} x^j,$$

or in the Einstein notation

$$x_k = g_{k,j} x^j.$$

In other words,

| the metric tensor $G$ is the change of coordinates matrix from contravariant coordinates $x^k$ to the covariant ones $x_k$. |
|---|

The operation of getting from contravariant coordinates to covariant is called *lowering of the indices*.

Note the following interpretation of the formula (4.4) for the inner product: as we know for the vector $\mathbf{x}$ we get his covariant coordinate as $x_j = g_{j,k} x^k$, so $(\mathbf{x}, \mathbf{y}) = x_j y^j$. Similarly, because $G$ is symmetric, we can say that $y_k = g_{j,k} y^k$ and that $(\mathbf{x}, \mathbf{y}) = x^k y_k$. In other words

| To compute the inner product of two vectors, one first needs to use the metric tensor $G$ to lower indices of one vector, and then, treating this vector as a functional compute its value on the other vector. |
|---|

Of course, we can also change from covariant coordinates $x_j$ to contravariant ones $x^j$ (raise the indices). Since

$$(x_1, x_2, \dots, x_n)^T = G(x^1, x^2, \dots, x^n)^T,$$

we get that

$$(x^1, x^2, \dots, x^n)^T = G^{-1}(x_1, x_2, \dots, x_n)^T$$

so the change of coordinate matrix in this case is $G^{-1}$.

$G^{-1}$ is the metric tensor in covariant coordinates.

Since, as we know, the change of coordinate matrix is the metric tensor, we can immediately conclude that $G^{-1}$ is the metric tensor in covariant coordinates, i.e. that if $G^{-1} = \{g^{k,j}\}_{k,j=1}^n$ then

$$(\mathbf{x}, \mathbf{y}) = g^{k,j} x_j y_k.$$

**Remark.** Note, that if one looks at the big picture, the covariant and contravariant coordinates are completely interchangeable. It is just the matter of which one of the bases in the dual pair $\mathcal{B}$ and $\mathcal{B}'$ we assign to be the "primary" one and which one to be the dual.

What to chose as a "primary" object, and what as the "dual" one depends mostly on accepted conventions.

**Remark 4.1.** Einstein notation is usually used in differential, and especially Riemannian geometry, where vectors are identified with velocities and covectors (linear functionals) with the differential 1-forms, see Section 4.2 above. Vectors and covectors here are clearly different objects and form what is called *tangent* and *cotangent* spaces respectively.

In Riemannian geometry one then introduces inner product (i.e. the metric tensor, if one thinks in terms of coordinates) on the tangent space, which allows us identify vectors and covectors (linear functionals). In coordinate representation this identification is done by lowering/raising indices, as described above.

**4.4. Conclusions.** Let us summarize the above discussion on whether or not a space is different from its dual.

In short, the answer is "Yes", they are different objects. Although in the finite-dimensional case, which is treated in this book, they are isomorphic, nothing is usually gained from the identification of a space and its dual.

Even in the simplest case of $\mathbb{F}^n$ it is useful to think that the elements of $\mathbb{F}^n$ are columns and the elements of its dual are rows (even though, when doing manipulations with the elements of the dual space we often put the rows vertically). More striking examples are ones considered in Sections 1.4.1 and 1.4.2 dealing with Taylor formula and Lagrange interpolation. One can clearly see there that the linear functionals are indeed completely

different objects than polynomials, and that hardly anything can be gained by identifying functionals with the polynomials.

For inner product spaces the situation is different, because such spaces can be *canonically* identified with their duals. This identification is linear for real inner product spaces, so a real inner product space is canonically isomorphic to its dual. In the case of complex spaces, this identification is only *conjugate linear*, but it is nevertheless very helpful to identify a linear functional with a vector and use the inner product space structure and ideas like orthogonality, self-adjointness, orthogonal projections, etc.

However, sometimes even in the case of real inner product spaces, it is more natural to consider the space and its dual as different objects. For example, in Riemannian geometry, see Remark 4.1 above vector and covectors come from different objects, velocities and differential 1-forms respectively. Even though the introduction of the metric tensor allows us to identify vectors and covectors, it is sometimes more convenient to remember their origins think of them as of different objects.

**Exercises.**

**4.1.** Let $D$ be a differential operator

$$D = \sum_{k=1}^{n} v_k \frac{\partial}{\partial x_k}.$$

Show, using the chain rule, that if we change a basis and write $D$ in new coordinates, its coefficients $v_k$ change according to the change of coordinates rule for vectors.

## 5. Multilinear functions. Tensors

### 5.1. Multilinear functions.

**Definition 5.1.** Let $V_1, V_2, \dots, V_p, V$ be vector spaces (over the same field $\mathbb{F}$). A *multilinear* ($p$-linear) map with values in $V$ is a function $F$ of $p$ vector variables $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$, $\mathbf{v}_k \in V_k$, with the target space $V$, which is linear in each variable $\mathbf{v}_k$. In other words, it means that if we fix all variables except $\mathbf{v}_k$ we get a linear map, and this should be true for all $k = 1, 2, \dots, p$. We will use the symbol $L(V_1, V_2, \dots, V_p; V)$ for the set of all such multilinear functions.

If the target space $V$ is the field of scalars $\mathbb{F}$, we call $F$ a *multilinear functional*, or *tensor*. The number $p$ is called the *valency* of the multilinear functional (tensor). Thus, tensor of valency 1 is a linear functional, tensor of valency 2 is called a *bilinear* form.

**Example.** Let $\mathbf{f}_k \in (V_k)'$. Define a polylinear functional $F = \mathbf{f}_1 \otimes \mathbf{f}_2 \otimes \ldots \otimes \mathbf{f}_p$ by multiplying the functionals $\mathbf{f}_k$,

$$\mathbf{f}_1 \otimes \mathbf{f}_2 \otimes \ldots \otimes \mathbf{f}_p(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = \mathbf{f}_1(\mathbf{v}_1)\mathbf{f}_2(\mathbf{v}_2)\ldots\mathbf{f}_p(\mathbf{v}_p), \tag{5.1}$$

for $\mathbf{v}_k \in V_k$, $k = 1, 2, \ldots, p$. The polylinear functional $\mathbf{f}_1 \otimes \mathbf{f}_2 \otimes \ldots \otimes \mathbf{f}_p$ is called the *tensor product* of functionals $\mathbf{f}_k$.

5.1.1. *Multilinear functions form vector space.* Notice, that in the space $L(V_1, V_2, \ldots, V_p; V)$ one can introduce the natural operations of addition and multiplication by a scalar,

$$(F_1 + F_2)(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) := F_1(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) + F_2(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p),$$
$$(\alpha F_1)(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) := \alpha F_1(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p),$$

where $F_1, F_2 \in L(V_1, V_2, \ldots, V_p; V)$, $\alpha \in \mathbb{F}$.

Equipped with these operations, the space $L(V_1, V_2, \ldots, V_p; V)$ is a vector space.

To see that we first need to show that $F_1 + F_2$ and $\alpha F_1$ are multilinear functions. Since "multilinear" means that it is linear in each argument separately (with all the other variables fixed), this follows from the corresponding fact about linear transformation; namely from the fact that the sum of linear transformations and a scalar multiple of a linear transformation are linear transformations, cf. Section 4 of Chapter 1.

Then it is easy to show that $L(V_1, V_2, \ldots, V_p; V)$ satisfies all axioms of vector space; one just need to use the fact that $V$ satisfies these axioms. We leave the details as an exercise for the reader. He/she can look at Section 4 of Chapter 1, where it was shown that the set of linear transformations satisfies axiom 7. Literally the same proof work for multilinear functions; the proof that all other axioms are also satisfied is very similar.

5.1.2. *Dimension of* $L(V_1, V_2, \ldots, V_p; V)$. Let $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_p$ be bases in the spaces $V_1, V_2, \ldots, V_p$ respectively. Since a linear transformation is defined by its action on a basis, a multilinear function $F \in L(V_1, V_2, \ldots, V_p; V)$ is defined by its values on all tuples

$$\mathbf{b}^1_{j_1}, \mathbf{b}^2_{j_2}, \ldots, \mathbf{b}^p_{j_p}, \qquad \mathbf{b}^k_{j_k} \in \mathcal{B}_k.$$

Since there are exactly

$$(\dim V_1)(\dim V_2)\ldots(\dim V_p)$$

such tuples, and each $F(\mathbf{b}^1_{j_1}, \mathbf{b}^2_{j_2}, \ldots, \mathbf{b}^p_{j_p})$ is determined by $\dim V$ coordinates (in some basis in $V$). we can conclude that $F \in L(V_1, V_2, \ldots, V_p; V)$ is determined by $(\dim V_1)(\dim V_2)\ldots(\dim V_p)(\dim V)$ entries. In other words

$$\dim L(V_1, V_2, \ldots, V_p; V) = (\dim V_1)(\dim V_2)\ldots(\dim V_p)(\dim V).$$

in particular, if the target space is the field of scalars $\mathbb{F}$ (i.e. if we are dealing with multilinear functionals)

$$\dim L(V_1, V_2, \dots, V_p; \mathbb{F}) = (\dim V_1)(\dim V_2)\dots(\dim V_p).$$

It is easy to find a basis in $L(V_1, V_2, \dots, V_p; \mathbb{F})$. Namely, let for $k = 1, 2, \dots, p$ the system $\mathcal{B}_k = \{\mathbf{b}_j^k\}_{j=1}^{\dim V_k}$ be a basis in $V_k$ and let $\mathcal{B}' = \{\widetilde{\mathbf{b}}_j^k\}_{j=1}^{\dim V_k}$ be its dual system, $\widetilde{\mathbf{b}}_j^k \in V_k'$.

**Proposition 5.2.** *The system*

$$\widetilde{\mathbf{b}}_{j_1}^1 \otimes \widetilde{\mathbf{b}}_{j_2}^2 \otimes \dots \otimes \widetilde{\mathbf{b}}_{j_p}^p, \qquad 1 \le j_k \le \dim V_k, \quad k = 1, 2, \dots, p,$$

*is a basis in the space* $L(V_1, V_2, \dots, V_p; \mathbb{F})$.

*Here* $\widetilde{\mathbf{b}}_{j_1}^1 \otimes \widetilde{\mathbf{b}}_{j_2}^2 \otimes \dots \otimes \widetilde{\mathbf{b}}_{j_p}^p$ *is the tensor product of functionals, as defined in* (5.1).

**Proof.** We want to represent $F$ as

$$F = \sum_{j_1, j_2, \dots, j_p} \alpha_{j_1, j_2, \dots, j_p} \widetilde{\mathbf{b}}_{j_1}^1 \otimes \widetilde{\mathbf{b}}_{j_2}^2 \otimes \dots \otimes \widetilde{\mathbf{b}}_{j_p}^p \tag{5.2}$$

Since $\widetilde{\mathbf{b}}_j(\mathbf{b}_l) = \delta_{j,l}$, we have

$$\widetilde{\mathbf{b}}_{j_1}^1 \otimes \widetilde{\mathbf{b}}_{j_2}^2 \otimes \dots \otimes \widetilde{\mathbf{b}}_{j_p}^p(\mathbf{b}_{j_1}^1, \mathbf{b}_{j_2}^2, \dots, \mathbf{b}_{j_p}^p) = 1 \qquad \text{and} \tag{5.3}$$

$$\widetilde{\mathbf{b}}_{j_1}^1 \otimes \widetilde{\mathbf{b}}_{j_2}^2 \otimes \dots \otimes \widetilde{\mathbf{b}}_{j_p}^p(\mathbf{b}_{j_1'}^1, \mathbf{b}_{j_2'}^2, \dots, \mathbf{b}_{j_p'}^p) = 0 \tag{5.4}$$

for any collection of indices $j_1', j_2', \dots, j_p'$ different from $j_1, j_2, \dots, j_p$.

Therefore, applying (5.2) to the tuple $\mathbf{b}_{j_1}^1, \mathbf{b}_{j_2}^2, \dots, \mathbf{b}_{j_p}^p$ we get

$$\alpha_{j_1, j_2, \dots, j_p} = F(\mathbf{b}_{j_1}^1, \mathbf{b}_{j_2}^2, \dots, \mathbf{b}_{j_p}^p),$$

so the representation (5.2) is unique (if exists).

On the other hand, defining $\alpha_{j_1, j_2, \dots, j_p} := F(\mathbf{b}_{j_1}^1, \mathbf{b}_{j_2}^2, \dots, \mathbf{b}_{j_p}^p)$ and using (5.3) and (5.4), we can see that the equality (5.2) holds on all tuples of form $\mathbf{b}_{j_1}^1, \mathbf{b}_{j_2}^2, \dots, \mathbf{b}_{j_p}^p$. So decomposition (5.2) holds, so we indeed have a basis. □

### 5.2. Tensor Products.

**Definition.** Let $V_1, V_2, \dots, V_p$ be vector spaces. The *tensor product*

$$V_1 \otimes V_2 \otimes \dots \otimes V_p$$

of spaces $V_k$ is simply the set $L(V_1', V_2', \dots, V_p'; \mathbb{F})$ of multilinear functionals; here $V_k'$ is the dual of $V_k$.

**Remark 5.3.** By Proposition 5.2 we get that if $\mathcal{B}_k = \{\mathbf{b}_j^k\}_{j=1}^{\dim V_k}$ is a basis in $V_k$ for $k = 1, 2, \dots, p$, then the system

$$\mathbf{b}_{j_1}^1 \otimes \mathbf{b}_{j_2}^2 \otimes \dots \otimes \mathbf{b}_{j_p}^p, \qquad 1 \le j_k \le \dim V_k, \quad k = 1, 2, \dots, p, \tag{5.5}$$

is a basis in $V_1 \otimes V_2 \otimes \dots \otimes V_p$.

Here we treat a vector $\mathbf{v}_k \in V_k$ as a linear functional on $V_k'$; the tensor product of vectors $\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p$ is the defined according to (5.1).

**Remark.** The tensor product $\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p$ of vectors is clearly linear in each argument $\mathbf{v}_k$. In other words, the map $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p) \mapsto \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p$ is a multilinear functional with values in $V_1 \otimes V_2 \otimes \dots \otimes V_p$. We leave the proof as an exercise for a reader, see Problem 5.1 below

**Remark.** Note, that the set $\{\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p \ : \ \mathbf{v}_k \in V_k\}$ of tensor products of vectors is strictly less than $V_1 \otimes V_2 \otimes \dots \otimes V_p$, see Problem 5.2 below.

5.2.1. *Lifting a multilinear function to a linear transformation on the tensor product.*

**Proposition 5.4.** *For any multilinear function $F \in L(V_1, V_2, \dots, V_p; V)$ there exists a unique linear transformation $T : V_1 \otimes V_2 \otimes \dots \otimes V_p \to V$ extending $F$, i.e. such that*

$$F(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p) = T\, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p, \tag{5.6}$$

*for all choices of vectors $\mathbf{v}_k \in V_k$, $1 \le k \le p$.*

**Remark.** If $T : V_1 \otimes V_2 \otimes \dots \otimes V_p \to V$ is a linear transformation, then trivially the function $F$,

$$F(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p) := T\, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p,$$

is a multilinear function in $L(V_1, V_2, \dots, V_p; V)$. This follows immediately from the fact that the expression $\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p$ is linear in each variable $\mathbf{v}_k$.

**Proof of Proposition 5.4.** Define $T$ on the basis (5.5) by

$$T\mathbf{b}_{j_1}^1 \otimes \mathbf{b}_{j_2}^2 \otimes \dots \otimes \mathbf{b}_{j_p}^p = F(\mathbf{b}_{j_1}^1, \mathbf{b}_{j_2}^2, \dots, \mathbf{b}_{j_p}^p)$$

and then extend it by linearity to all space $V_1 \otimes V_2 \otimes \dots \otimes V_p$. To complete the proof we need to show that (5.6) holds for all choices of vectors $\mathbf{v}_k \in V_k$, $1 \le k \le p$ (we now know that only when each $\mathbf{v}_k$ is one of the vectors $\mathbf{b}_{j_k}^k$).

To prove that, let us decompose $\mathbf{v}_k$ as

$$\mathbf{v}_k = \sum_{j_k} \alpha_{j_k}^k \mathbf{b}_{j_k}^k, \qquad k = 1, 2, \dots, p.$$

Using linearity in each variable $\mathbf{v}_k$ we get

$$\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \ldots \otimes \mathbf{v}_p = \sum_{j_1,j_2,\ldots,j_p} \alpha^1_{j_1}\alpha^2_{j_2},\ldots,\alpha^p_{j_p}\mathbf{b}^1_{j_1} \otimes \mathbf{b}^2_{j_2} \otimes \ldots \otimes \mathbf{b}^p_{j_p},$$

$$F(\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_p) = \sum_{j_1,j_2,\ldots,j_p} \alpha^1_{j_1}\alpha^2_{j_2},\ldots,\alpha^p_{j_p}F(\mathbf{b}^1_{j_1},\mathbf{b}^2_{j_2},\ldots,\mathbf{b}^p_{j_p})$$

so by the definition of $T$ identity (5.6) holds. □

5.2.2. *Dual of a tensor product.* As one can easily see, the dual of the tensor product $V_1 \otimes V_2 \otimes \ldots \otimes V_p$ is the tensor product of dual spaces $V_1' \otimes V_2' \otimes \ldots \otimes V_p'$.

Indeed, by Proposition 5.4 and remark after it, there is a natural one-to-one correspondence between multilinear functionals in $L(V_1, V_2, \ldots, V_p, \mathbb{F})$ (i.e. the elements of $V_1' \otimes V_2' \otimes \ldots \otimes V_n'$) and the linear transformations $T : V_1 \otimes V_2 \otimes \ldots \otimes V_p \to \mathbb{F}$ (i.e. with the elements of the dual of $V_1 \otimes V_2 \otimes \ldots \otimes V_p$).

Note, that the bases from Remark 5.3 and Proposition 5.2 are the dual bases (in $V_1 \otimes V_2 \otimes \ldots \otimes V_p$ and $V_1' \otimes V_2' \otimes \ldots \otimes V_n'$ respectively). Knowing the dual bases allows us easily calculate the *duality* between the spaces $V_1 \otimes V_2 \otimes \ldots \otimes V_p$ and $V_1' \otimes V_2' \otimes \ldots \otimes V_p'$, i.e. the expression $\langle \mathbf{x}, \mathbf{x}' \rangle$, $\mathbf{x} \in V_1 \otimes V_2 \otimes \ldots \otimes V_p$, $\mathbf{x}' \in V_1' \otimes V_2' \otimes \ldots \otimes V_p'$

**5.3. Covariant and contravariant tensors.** Let $X_1, X_2, \ldots, X_p$ be vector spaces, and let $V_k$ be either $X_k$ or $X_k'$, $k = 1, 2, \ldots, p$. For a multilinear function $F \in L(V_1, V_2, \ldots, V_p; V)$ we say that that it is *covariant* in variable $\mathbf{v}_k \in V_k$ if $V_k = X_k$ and *contravariant* in this variable if $V_k = X_k'$.

If a multilinear function is covariant (contravariant) in all variables, we say that the multilinear function is covariant (contravariant). In general, if a function is covariant in $r$ variables and contravariant in $s$ variables, we say that the multilinear function is $r$-covariant $s$-contravariant (or simply $(r, s)$ multilinear function, or that its valency is $(r, s)$).

Thus, a linear functional can be interpreted as 1-covariant tensor (recall, that we use the word *tensor* for the case of *functionals*, i.e. when the target space is the field of scalars $\mathbb{F}$). By duality, a vector can be interpreted as 1-contravariant tensor.

**Remark.** At first the terminology might look a bit confusing: if a variable is a vector (not a functional), it is a covariant variable but a contravariant object. But notice, that we did not say here a "covariant variable": we said that if $\mathbf{v}_k \in X_k$ then the *mulitilinear function is covariant in the variable* $\mathbf{v}_k$. So, the covariant object is not $\mathbf{v}_k$, but the "slot" in the tensor where we put it!

So there is no contradiction, we put the contravariant objects into covariant slots and vice versa.

Sometimes, slightly abusing the language, people talk about covariant (contravariant) variables or arguments. But it is usually meant that the corresponding "slots" in the tensor are covariant (contravariant), and not the variables as objects.

5.3.1. *Linear transformations as tensors.* A linear transformation $T : X_1 \to X_2$ can be interpreted as 1-covariant 1-contravariant tensor. Namely, the bilinear functional $F$,

$$F(\mathbf{x}_1, \mathbf{x}'_2) := \langle T\mathbf{x}_1, \mathbf{x}'_2\rangle, \qquad \mathbf{x}_1 \in X_1, \mathbf{x}'_2 \in X'_2$$

is covariant in the first variable $\mathbf{x}_1$ and contravariant in the second one $\mathbf{x}'_2$.

Conversely,

**Proposition 5.5.** *Given a* 1-1 *tensor* $F \in L(X_1, X'_2; \mathbb{F})$*, there exists a unique linear transformation* $T : X_1 \to X_2$ *such that*

$$F(\mathbf{x}_1, \mathbf{x}'_2) := \langle T\mathbf{x}_1, \mathbf{x}'_2\rangle, \tag{5.7}$$

*for all* $\mathbf{x}_1 \in X_2$, $\mathbf{x}'_2 \in X'_2$.

**Proof.** First of all note, that the uniqueness is a trivial corollary of Lemma 1.3, cf. Problem 3.1 above. So we only need to prove existence of $T$.

Let $\mathcal{B}_k = \{\mathbf{b}^k_j\}_{j=1}^{\dim X_k}$ be a basis in $X_k$, and let $\mathcal{B}'_k = \{\widetilde{\mathbf{b}}^k_j\}_{j=1}^{\dim X_k}$ be the dual basis in $X'_k$, $k = 1, 2$. Then define the matrix $A = \{a_{k,j}\}_{k=1}^{\dim X_2}{}_{j=1}^{\dim X_1}$ by

$$a_{k,j} = F(\mathbf{b}^1_j, \widetilde{\mathbf{b}}^2_k).$$

Define $T$ to be the operator with matrix $[T]_{\mathcal{B}_2,\mathcal{B}_1} = A$. Clearly (see Remark 1.5)

$$\langle T\mathbf{b}^1_j, \widetilde{\mathbf{b}}^2_k\rangle = a_{k,j} = F(\mathbf{b}^1_j, \widetilde{\mathbf{b}}^2_k) \tag{5.8}$$

which implies the equality (5.7). This can be easily seen by decomposing $\mathbf{x}_1 = \sum_j \alpha_j \mathbf{b}_j$ and $\mathbf{x}'_2 = \sum_k \beta_k \mathbf{b}'_k$ and using linearity in each argument.

Another, more high brow explanation is that the tensors in left and the right sides of (5.7) coincide on a basis in $X_1 \otimes X'_2$ (see Remark 5.3 about the basis), so they coincide. To be more precise, one should lift the bilinear forms to the linear transformations (functionals) $X_1 \otimes X'_2 \to \mathbb{F}$ (see Proposition 5.4), and since the transformations coincide on a basis, they are equal.

One can also give an alternative, coordinate-free proof of existence of $T$, along the lines of the coordinate-free definition of the dual space (see Section 3.1.3). Namely, if we fix $\mathbf{x}_1$, the function $F(\mathbf{x}_1, \mathbf{x}'_2)$ is a linear in $\mathbf{x}'_2$, so it is a linear functional on $X'_2$, i.e. a vector in $X_2$.

Let us call this vector $T(\mathbf{x}_1)$. So we defined a transformation $T : X_1 \to X_2$. One can easily show that $T$ is a linear transformation by essentially

repeating the reasoning from Section 3.1.3. The equality (5.7) follows authomatically from the definition of $T$. □

**Remark.** Note that we also can say that the function $F$ from Proposition 5.5 defines not the transformation $T$, but its adjoint. Apriori, without assuming anything (like order of variables and its interpretation) we cannot distinguish between a transformation and its adjoint.

**Remark.** Note, that if we would like to follow the Einstein notation, the entries $a_{j,k}$ of the matrix $A = [T]_{\mathcal{B}_2,\mathcal{B}_1}$ of the transformation $T$ should be written as $a^j_k$. Then if $x^k$, $k = 1, 2, \dots, \dim X_1$ are the coordinates of the vector $\mathbf{x} \in X_1$, the $j$th coordinate of $\mathbf{y} = T\mathbf{x}$ is given by

$$y^j = a^j_k x^k.$$

Recall the here we skip the sign of summation, but we mean the sum over $k$. Note also, that we preserve positions of the indices, so the index $j$ stays upstairs. The index $k$ does not appear in the left side of the equation because we sum over this index in the right side, and its got "killed".

Similarly, if $x_j$, $j = 1, 2, \dots, \dim X_2$ are the coordinates of the vector $\mathbf{x}' \in X_2'$, then $k$th coordinate of $\mathbf{y}' := T'\mathbf{x}'$ is given by

$$y_k = a^j_k x_j$$

(again, skipping the sign of summation over $j$). Again, since we preserve the position of the indices, so the index $k$ in $y_k$ is a subscript.

Note, that since $\mathbf{x} \in X_1$ and $\mathbf{y} = T\mathbf{x} \in X_2$ are vectors, according to the conventions of the Einstein notation, the indices in their coordinates indeed should be written as superscripts.

Similarly, $\mathbf{x}' \in X_2'$ and $\mathbf{y}' = T'\mathbf{x}' \in X_1'$ are *covectors*, so indices in their coordinates should be written as subscripts.

The Einstein notation emphasizes the fact mentioned in the previous remark, that a 1-covariant 1-contravariant tensor gives us both a linear transformation and its adjoint: the expression $a^j_k x^k$ gives the action of $T$, and $a^j_k x_j$ gives the action of its adjoint $T'$.

5.3.2. *Polylinear transformations as tensors.* More generally, any polylinear transformation can be interpreted as a tensor. Namely, given a polylinear transformation $F \in L(V_1, V_2, \dots, V_p; V)$ one can define the tensor $\widetilde{F} \in L(V_1, V_2, \dots, V_p, V'; \mathbb{F})$ by

$$\widetilde{F}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p, \mathbf{v}') = \langle F(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p), \mathbf{v}' \rangle, \qquad \mathbf{v}_k \in V_k, \mathbf{v}' \in V'. \tag{5.9}$$

Conversely,

**Proposition 5.6.** *Given a tensor $\widetilde{F} \in L(V_1, V_2, \dots, V_p, V'; \mathbb{F})$ there exists a unique polylinear transformation $F \in L(V_1, V_2, \dots, V_p; V)$ such that (5.9) is satisfied.*

**Proof.** By Proposition 5.4 the tensor $\widetilde{F}$ can be extended to a linear transformation (functional) $\widetilde{T} : V_1 \otimes V_2 \otimes \dots \otimes V_p \otimes V' \to \mathbb{F}$ such that

$$\widetilde{F}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p, \mathbf{v}') = \widetilde{T}(\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p \otimes \mathbf{v}')$$

for all $\mathbf{v}_k \in V_k$, $\mathbf{v}' \in V'$.

If $\mathbf{w} \in W := V_1 \otimes V_2 \otimes \dots \otimes V_p$ and $\mathbf{v}' \in V'$, then

$$\mathbf{w} \otimes \mathbf{v}' \in V_1 \otimes V_2 \otimes \dots \otimes V_p \otimes V'.$$

So, we can define a bilinear functional (tensor) $G \in L(W, V'; \mathbb{F})$ by

$$G(\mathbf{w}, \mathbf{v}') := \widetilde{T}(\mathbf{w} \otimes \mathbf{v}).$$

By Proposition 5.5, $G$ gives rise to a linear transformation, i.e. there exists a unique linear transformation $T : W \to V$ such that

$$G(\mathbf{w}, \mathbf{v}') = \langle T\mathbf{w}, \mathbf{v}' \rangle \qquad \forall \mathbf{w} \in W, \quad \forall \mathbf{v}' \in V'.$$

And the linear transformation $T$ gives us the polylinear map

$$F \in L(V_1, V_2, \dots, V_p; V)$$

by

$$F(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p) = T(\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p),$$

see Remark after Proposition 5.4.

The uniqueness of the transformation $F$, is, as in Proposition 5.5, is a trivial corollary of Lemma 1.3. We leave the details as an exercise for the reader. □

This section shows that

| tensors are universal objects in polylinear algebra, since any polylinear transformation can be interpreted as a tensor and vice versa. |
|---|

**Exercises.**

**5.1.** Show that the tensor product $\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p$ of vectors is linear in each argument $\mathbf{v}_k$.

**5.2.** Show that the set $\{\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_p \,:\, \mathbf{v}_k \in V_k\}$ of tensor products of vectors is strictly less than $V_1 \otimes V_2 \otimes \dots \otimes V_p$.

**5.3.** Prove that the transformation $F$ from Proposition 5.6 is unique.

## 6. Change of coordinates formula for tensors.

The main reason for the differentiation of covariant and contravariant variables is that under the change of bases, their coordinates change according to different rules. Thus, the entries of covariant and contravariant vectors change according to different rules as well.

In this section we going to investigate this in details. Note, that coordinate representations are extremely important, since, for example, all numerical computations (unlike the theoretical investigations) are performed using some coordinate system.

**6.1. Coordinate representation of a tensor.** Let $F$ be an $r$-covariant $s$-contravariant tensor, $r+s=p$. Let $\mathbf{x}_1,\dots,\mathbf{x}_r$ be covariant variables ($\mathbf{x}_k \in X_k$), and $\mathbf{f}_1,\dots,\mathbf{f}_s$ be the contravariant ones ($\mathbf{f}_k \in X_k'$). Let us write the covariant variables first, so the the tensor will be written as $F(\mathbf{x}_1,\dots,\mathbf{x}_r,\mathbf{f}_1,\dots,\mathbf{f}_s)$. For $k=1,2,\dots,p$ fix a basis $\mathcal{B}_k = \{\mathbf{b}_j^{(k)}\}_{j=1}^{\dim X_k}$ in $X_k$, and let $\mathcal{B}_k' = \{\widetilde{\mathbf{b}}_j^{(k)}\}_{j=1}^{\dim X_k}$ be the dual basis in $X_k'$.

For a vector $\mathbf{x}_k \in X_k$ let $x_{(k)}^j$, $j=1,2,\dots,\dim X_k$ be its coordinates in the basis $\mathcal{B}_k$, and similarly, if $\mathbf{f}_k \in X_k'$ let $f_j^{(k)}$, $j=1,2,\dots,\dim X_k$ be its coordinates in the dual basis $\mathcal{B}_k'$ (note that in agreement with the Einstein notation, the coordinates of the vector are indexed by a superscript, and the coordinate of a covector re indexed by a subscript).

**Proposition 6.1.** *Denote*

$$\varphi_{j_1,\dots,j_r}^{k_1,\dots,k_s} := F(\mathbf{b}_{j_1}^{(1)},\dots,\mathbf{b}_{j_r}^{(r)},\widetilde{\mathbf{b}}_{k_1}^{(r+1)},\dots,\widetilde{\mathbf{b}}_{k_s}^{(r+s)}). \tag{6.1}$$

*Then, in the Einstein notation*

$$F(\mathbf{x}_1,\dots,\mathbf{x}_r,\mathbf{f}_1,\dots,\mathbf{f}_s) = \varphi_{j_1,\dots,j_r}^{k_1,\dots,k_s} x_{(1)}^{j_1}\dots x_{(r)}^{j_r} f_{k_1}^{(1)}\dots f_{k_s}^{(s)} \tag{6.2}$$

*(the summation here is over the indices $j_1,\dots,j_r$ and $k_1,\dots,k_s$).*

Note that we use the notation $(1),\dots,(r)$ and $(1),\dots,(s)$ to emphasize that these are not the indices: the numbers in parenthesis just show the order of argument. Thus, right side of (6.2) does not have any indices left (all indices were used in summation), so it is just a number (for fixed $\mathbf{x}_k$s and $\mathbf{f}_k$s).

**Proof of Proposition 6.1.** To show that (6.1) implies (6.2) we first notice that (6.1) means that (6.2) hods when $\mathbf{x}_j$s and $\mathbf{f}_k$s are the elements of the corresponding bases. Decomposing each argument $\mathbf{x}_j$ and $\mathbf{f}_k$ in the corresponding basis and using linearity in each argument we can easily get (6.2). The computation is rather simple, but because there are a lot of indices, the formulas could be quite big and could look quite frightening.

To avoid writing too many huge formulas, we leave this computation to the reader as an exercise.

We do not want the reader to feel cheated, so we present a different, more "high brow" (abstract) explanation, which does not require any computations! Namely, let us notice that the expressions in the left and the right side of (6.2) define tensors. By Proposition 5.4 they can be lifted to linear functionals on the tensor product $X_1 \otimes \ldots \otimes X_r \otimes X'_{r+1} \otimes \ldots \otimes X'_{r+s}$.

Rephrasing what we discussed in the beginning of the proof, we can say that (6.1) means that the functional coincide on all vectors

$$\mathbf{b}^{(1)}_{j_1} \otimes \ldots \otimes \mathbf{b}^{(r)}_{j_r} \otimes \widetilde{\mathbf{b}}^{(r+1)}_{k_1} \otimes \ldots \otimes \widetilde{\mathbf{b}}^{(r+s)}_{k_s}$$

of a basis in the tensor product, so the functionals (and therefore the tensors) are equal. □

The entries $\varphi^{k_1,\ldots,k_s}_{j_1,\ldots,j_r}$ are called the entries of the tensor $F$ in the bases $\mathcal{B}_k$, $k = 1, 2, \ldots, p$.

Now, let for $k = 1, 2, \ldots p$, $\mathcal{A}_k$ be a basis in $X_k$ (and $\mathcal{A}'_k$ be the dual basis in $X'_k$). We want to investigate how the entries of the tensor $F$ change when we change the bases from $\mathcal{B}_k$ to $\mathcal{A}_k$.

**6.2. Change of coordinate formulas in Einstein notation.** Let us first consider the familiar cases of vectors and linear functionals, considered above in Section 1.1.1 but write everything down using the Einstein notation. Let we have in $X$ two bases, $\mathcal{B}$ and $\mathcal{A}$ and let

$$A = [A]_{\mathcal{A},\mathcal{B}}$$

be the change of coordinates matrix from $\mathcal{B}$ to $\mathcal{A}$. For a vector $\mathbf{x} \in X$ let $x^k$ be its coordinates in the basis $\mathcal{B}$ and $\widetilde{\mathbf{x}}^k$ be the coordinates in the basis $\mathcal{A}$. Similarly, for $\mathbf{f} \in X'$ let $f_k$ denote the coordinates in the basis $\mathcal{B}'$ and $\widetilde{f}_k$–the coordinates in the basis $\mathcal{A}'$ ($\mathcal{B}'$ and $\mathcal{A}'$ are the dual bases to $\mathcal{B}$ and $\mathcal{A}$ respectively).

Denote by $(A)^j_k$ the entries of the matrix $A$: to be consistent with the Einstein notation the superscript $j$ denotes the number of the row. Then we can write the change of coordinate formula as

$$\widetilde{x}^j = (A)^j_k x^k. \tag{6.3}$$

Similarly, let $(A^{-1})^k_j$ be the entries of $A^{-1}$: again superscript is used to denote the number of the row. Then we can write the change of coordinate formula for the dual space as

$$\widetilde{f}_j = (A^{-1})^k_j f_k; \tag{6.4}$$

the summation here is over the index $k$ (i.e. along the columns of $A^{-1}$), so the change of coordinate matrix in this case is indeed $(A^{-1})^T$.

Let us emphasize that we did not prove anything here: we only rewrote formula (1.1) from Section 1.1.1 using the Einstein notation.

**Remark.** While it is not needed in what follows, let us play a bit more with the Einstein notation. Namely, the equations

$$A^{-1}A = I \qquad \text{and} \qquad AA^{-1} = I$$

can be rewritten in the Einstein notation as

$$(A)^j_k (A^{-1})^k_l = \delta_{j,l} \qquad \text{and} \qquad (A^{-1})^k_j (A)^j_l = \delta_{k,l}$$

respectively.

**6.3. Change of coordinates formula for tensors.** Now we are ready to give the change of coordinate formula for general tensors.

For $k = 1, 2, \dots, p := r + s$ let $A_k := [I]_{\mathcal{A},\mathcal{B}}$ be the change of coordinates matrices, and let $A_k^{-1}$ be their inverses.

As in Section 6.2 we denote by $(A)^j_k$ the entries of a matrix $A$, with the agreement that superscript gives the number of the column.

**Proposition 6.2.** *Given an $r$-covariant $s$-contravariant tensor $F$ let*

$$\varphi^{k_1,\dots,k_s}_{j_1,\dots,j_r} \qquad \text{and} \qquad \widetilde{\varphi}^{k_1,\dots,k_s}_{j_1,\dots,j_r}$$

*be its entries in the bases $\mathcal{B}_k$ (the old ones) and $\mathcal{A}_k$ (the new ones) respectively. In the above notation*

$$\widetilde{\varphi}^{k_1,\dots,k_s}_{j_1,\dots,j_r} = \varphi^{k'_1,\dots,k'_s}_{j'_1,\dots,j'_r} (A_1^{-1})^{j'_1}_{j_1} \dots (A_r^{-1})^{j'_r}_{j_r} (A_{r+1})^{k_1}_{k'_1} \dots (A_{r+s})^{k_s}_{k'_s}$$

*(the summation here is in the indices $j'_1, \dots, j'_r$ and $k'_1, \dots, k_s$).*

Because of many indices, the formula in this proposition looks very complicated. However if one understands the main idea, the formula will turn out to be quite simple and easy to memorize.

To explain the main idea let us, sightly abusing the language, express this formula "in plain English". namely, we can say, that

> To express the "new" tensor entries $\widetilde{\varphi}^{k_1,\dots,k_s}_{j_1,\dots,j_r}$ in terms of the "old" ones $\varphi^{k_1,\dots,k_s}_{j_1,\dots,j_r}$, one needs for each *covariant* index (subscript) apply the covariant rule (6.4), and for each *contravariant* index (superscript) apply the contravariant rule (6.3)

**Proof of Proposition 6.2.** Informally, the idea of the proof is very simple: we just change the bases one at a time, applying each time the change

of coordinate formulas (6.3) or (6.4), depending on whether the tensor is covariant or contravariant in the corresponding variable.

To write the rigorous formal proof we will use the induction in $r$ ans $s$ (the number of covariant and contravariant arguments of the tensor). Proposition is true for $r = 1$, $s = 0$ and for $r = 0$, $s = 1$, see (6.4) or (6.3) respectively.

Assuming now that the proposition is proved for some $p$ and $s$, let us prove it for $r + 1$, $s$ and for $r$, $s + 1$.

Let us do the latter case, the other one is done similarly. The maid idea is that we first change $p = r + s$ bases and use the induction hypothesis; then we change the last one and use (6.3).

Namely, let $\widehat{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_{s+1}}$ be the entries of an $(r, s+1)$ tensor $F$ in the bases $\mathcal{A}_1, \dots, \mathcal{A}_p, \mathcal{B}_{p+1}$, $p = r + s$.

Let us fix the index $k_{s+1}$ and consider the $r$-covariant $s$-contravariant tensor $F(\mathbf{x}_1, \dots, \mathbf{x}_r, \mathbf{f}_1, \dots, \mathbf{f}_s, \widetilde{\mathbf{b}}_{s+1}^{(r+s+1)})$, where $\mathbf{x}_1, \dots, \mathbf{x}_r, \mathbf{f}_1, \dots, \mathbf{f}_s$ are the variables. Clearly

$$\varphi_{j_1,\dots,j_r}^{k_1,\dots,k_s,k_{s+1}} \qquad \text{and} \qquad \widehat{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_s,k_{s+1}}$$

are its entries in the bases $\mathcal{B}_1, \dots, \mathcal{B}_p$ and $\mathcal{A}_1, \dots, \mathcal{A}_p$ respectively (can you see why?) Recall, that the index $k_{s+1}$ here is fixed.

By the induction hypothesis

$$\widehat{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_s,k_{s+1}} = \varphi_{j'_1,\dots,j'_r}^{k'_1,\dots,k'_s,k_{s+1}} (A_1^{-1})_{j_1}^{j'_1} \dots (A_r^{-1})_{j_r}^{j'_r} (A_{r+1})_{k'_1}^{k_1} \dots (A_{r+s})_{k'_s}^{k_s}. \tag{6.5}$$

Note, that we did not assume anything about the index $k_{s+1}$, so (6.5) holds for all $k_{s+1}$.

Now let us fix indices $j_1, \dots, j_r, k_1, \dots, k_s$ and consider 1-contravariant tensor

$$F(\mathbf{a}_{j_1}^{(1)}, \dots, \mathbf{a}_{j_r}^{(r)}, \widetilde{\mathbf{a}}_{k_1}^{(r+1)}, \dots, \widetilde{\mathbf{a}}_{k_s}^{(r+s)}, \mathbf{f}_{s+1})$$

of the variable $\mathbf{f}_{s+1}$. Here $\mathbf{a}_j^{(k)}$ are the vectors in the basis $\mathcal{A}_k$ and $\widetilde{\mathbf{a}}_j^{(k)}$ are the vectors in the dual basis $\mathcal{A}'_k$.

It is again easy to see that

$$\widehat{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_s,k_{s+1}} \qquad \text{and} \qquad \widetilde{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_s,k_{s+1}},$$

$j_{s+1} = 1, 2, \dots, \dim X_{p+1}$, are the indices of this functional in the bases $\mathcal{B}_{p+1}$ and $\mathcal{A}_{p+1}$ respectively. According to (6.3)

$$\widetilde{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_s,k_{s+1}} = \widehat{\varphi}_{j_1,\dots,j_r}^{k_1,\dots,k_s,k'_{s+1}} (A_{p+1})_{k'_{s+1}}^{k_{s+1}},$$

and since we did not assume anything about the indices $j_1, \dots, j_r, k_1, \dots, k_s$, the above identity holds for all their combinations. Combining this with (6.5) we get that the proposition holds for tensors of valency $(r, s + 1)$.

The case of valency $(r+1, s)$ is treated absolutely the same way: the only difference is that in the end we get a 1-covariant tensor and use (6.4) instead of (6.3). □

*Chapter 9*

# Advanced spectral theory

## 1. Cayley–Hamilton Theorem

**Theorem 1.1** (Cayley–Hamilton)**.** *Let $A$ be a square matrix, and let $p(\lambda) = \det(A - \lambda I)$ be its characteristic polynomial. Then $p(A) = \mathbf{0}$.*

**A wrong proof.** The proof looks ridiculously simple: plugging $A$ instead of $\lambda$ in the definition of the characteristic polynomial we get

$$p(A) = \det(A - AI) = \det \mathbf{0} = 0.$$

□

But this is a wrong proof! To see why, let us analyze what the theorem states. It states, that if we compute the characteristic polynomial

$$\det(A - \lambda I) = p(\lambda) = \sum_{k=0}^{n} c_k \lambda^k$$

and *then* plug matrix $A$ instead of $\lambda$ to get

$$p(A) := \sum_{k=0}^{n} c_k A^k = c_0 I + c_1 A + \ldots + c_n A^n$$

then the result will be zero *matrix*.

It is not clear why we get the same result if we just plug $A$ instead of $\lambda$ in the determinant $\det(A - \lambda I)$. Moreover, it is easy to see that with the exception of trivial case of $1 \times 1$ matrices we will get a different object. Namely, $A - AI$ is zero *matrix*, and its determinant is just the *number* 0.

But $p(A)$ is a matrix, and the theorem claims that this matrix is the zero *matrix*. Thus we are comparing apples and oranges. Even though in both cases we got zero, these are different zeroes: he number zero and the zero matrix!

Let us present another proof, which is based on some ideas from analysis.

This proof illustrates an important idea that often it is sufficient to consider only a typical, generic situation. It is going beyond the scope of the book, but let us mention, without going into details, that a generic (i.e. typical) matrix is diagonalizable.

**A "continuous" proof.** The proof is based on several observations. First of all, the theorem is trivial for diagonal matrices, and so for matrices similar to diagonal (i.e. for diagonalizable matrices), see Problem 1.1 below.

The second observation is that any matrix can be approximated (as close as we want) by diagonalizable matrices. Since any operator has an upper triangular matrix in some orthonormal basis (see Theorem 1.1 in Chapter 6), we can assume without loss of generality that $A$ is an upper triangular matrix.

We can perturb diagonal entries of $A$ (as little as we want), to make them all different, so the perturbed matrix $\widetilde{A}$ is diagonalizable (eigenvalues of a a triangular matrix are its diagonal entries, see Section 1.7 in Chapter 4, and by Corollary 2.3 in Chapter 4 an $n \times n$ matrix with $n$ distinct eigenvalues is diagonalizable).

As I just mentioned, we can perturb the diagonal entries of $A$ as little as we want, so Frobenius norm $\|A - \widetilde{A}\|_2$ is as small as we want. Therefore one can find a sequence of diagonalizable matrices $A_k$ such that $A_k \to A$ as $k \to \infty$ for example such that $\|A_k - A\|_2 \to 0$ as $k \to \infty$). It can be shown that the characteristic polynomials $p_k(\lambda) = \det(A_k - \lambda I)$ converge to the characteristic polynomial $p(\lambda) = \det(A - \lambda I)$ of $A$. Therefore

$$p(A) = \lim_{k\to\infty} p_k(A_k).$$

But as we just discussed above the Cayley–Hamilton Theorem is trivial for diagonalizable matrices, so $p_k(A_k) = \mathbf{0}$. Therefore $p(A) = \lim_{k\to\infty} \mathbf{0} = \mathbf{0}$. □

This proof is intended for a reader who is comfortable with such ideas from analysis as continuity and convergence[1]. Such a reader should be able to fill in all the details, and for him/her this proof should look extremely easy and natural.

However, for others, who are not comfortable yet with these ideas, the proof definitely may look strange. It may even look like some kind of cheating, although, let me repeat that it is an absolutely correct and rigorous proof (modulo some standard facts in analysis). So, let us present another,

[1]Here I mean *analysis*, i.e. a rigorous treatment of continuity, convergence, etc, and not calculus, which, as it is taught now, is simply a collection of recipes.

proof of the theorem which is one of the "standard" proofs from linear algebra textbooks.

**A "standard" proof.** We know, see Theorem 6.1.1 from Chapter 6, that any square matrix is unitary equivalent to an upper triangular one. Since for any polynomial $p$ we have $p(UAU^{-1}) = Up(A)U^{-1}$, and the characteristic polynomials of unitarily equivalent matrices coincide, it is sufficient to prove the theorem only for upper triangular matrices.

So, let $A$ be an upper triangular matrix. We know that diagonal entries of a triangular matrix coincide with it eigenvalues, so let $\lambda_1, \lambda_2, \dots, \lambda_n$ be eigenvalues of $A$ ordered as they appear on the diagonal, so

$$A = \begin{pmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_n \end{pmatrix}.$$

The characteristic polynomial $p(z) = \det(A - zI)$ of $A$ can be represented as $p(z) = (\lambda_1 - z)(\lambda_2 - z)\dots(\lambda_n - z) = (-1)^n(z - \lambda_1)(z - \lambda_2)\dots(z - \lambda_n)$, so

$$p(A) = (-1)^n(A - \lambda_1 I)(A - \lambda_2 I)\dots(A - \lambda_n I).$$

Define subspaces $E_k := \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k\}$, where $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is the standard basis in $\mathbb{C}^n$. Since the matrix of $A$ is upper triangular, the subspaces $E_k$ are so-called *invariant* subspaces of the operator $A$, i.e. $AE_k \subset E_k$ (meaning that $A\mathbf{v} \in E_k$ for all $\mathbf{v} \in E_k$). Moreover, since for any $\mathbf{v} \in E_k$ and any $\lambda$

$$(A - \lambda I)\mathbf{v} = A\mathbf{v} - \lambda\mathbf{v} \in E_k,$$

because both $A\mathbf{v}$ and $\lambda\mathbf{v}$ are in $E_k$. Thus $(A - \lambda I)E_k \subset E_k$, i.e. $E_k$ is an invariant subspace of $A - \lambda I$.

We can say even more about the the subspace $(A - \lambda_k I)E_k$. Namely, $(A - \lambda_k I)\mathbf{e}_k \in \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{k-1}\}$, because only the first $k-1$ entries of the $k$th column of the matrix of $A - \lambda_k I$ can be non-zero. On the other hand, for $j < k$ we have $(A - \lambda_k)\mathbf{e}_j \in E_j \subset E_k$ (because $E_j$ is an invariant subspace of $A - \lambda_k I$).

Take any vector $\mathbf{v} \in E_k$. By the definition of $E_k$ it can be represented as a linear combination of the vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$. Since all vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$ are transformed by $A - \lambda_k I$ to some vectors in $E_{k-1}$, we can conclude that

$$(A - \lambda_k I)\mathbf{v} \in E_{k-1} \qquad \forall \mathbf{v} \in E_k. \tag{1.1}$$

Take an arbitrary vector $\mathbf{x} \in \mathbb{C}^n = E_n$. Applying (1.1) inductively with $k = n, n-1, \dots 1$ we get

$$\begin{aligned}
&\mathbf{x}_1 := (A - \lambda_n I)\mathbf{x} \in E_{n-1},\\
&\mathbf{x}_2 := (A - \lambda_{n-1} I)\mathbf{x}_1 = (A - \lambda_{n-1} I)(A - \lambda_n I)\mathbf{x} \in E_{n-2},\\
&\quad \dots\\
&\mathbf{x}_n := (A - \lambda_2 I)\mathbf{x}_{n-1} = (A - \lambda_2 I)\dots(A - \lambda_{n-1} I)(A - \lambda_n I)\mathbf{x} \in E_1.
\end{aligned}$$

The last inclusion mean that $\mathbf{x}_n = \alpha \mathbf{e}_1$. But $(A - \lambda_1 I)\mathbf{e}_1 = \mathbf{0}$, so

$$\mathbf{0} = (A - \lambda_1 I)\mathbf{x}_n = (A - \lambda_1 I)(A - \lambda_2 I)\dots(A - \lambda_n I)\mathbf{x}.$$

Therefore $p(A)\mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in \mathbb{C}^n$, which means exactly that $p(A) = \mathbf{0}$. □

**Exercises.**

**1.1** (Cayley–Hamilton Theorem for diagonalizable matrices)**.** As discussed in the above section, the Cayley–Hamilton theorem states that if $A$ is a square matrix, and

$$p(\lambda) = \det(A - \lambda I) = \sum_{k=0}^{n} c_k \lambda^k$$

is its characteristic polynomial, them $p(A) := \sum_{k=0}^{n} c_k A^k = \mathbf{0}$ (we assuming, that by definition $A^0 = I$).

Prove this theorem for the special case when $A$ is similar to a diagonal matrix, $A = SDS^{-1}$.

**Hint:** If $D = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ and $p$ is any polynomial, can you compute $p(D)$? What about $p(A)$?

## 2. Spectral Mapping Theorem

**2.1. Polynomials of operators.** Let us also recall that for a square matrix (an operator) $A$ and for a polynomial $p(z) = \sum_{k=1}^{N} a_k z^k$ the operator $p(A)$ is defined by substituting $A$ instead of the independent variable,

$$p(A) := \sum_{k=1}^{N} a_k A^k = a_0 I + a_1 A + a_2 A^2 + \ldots + a_N A^N;$$

here we agree that $A^0 = I$.

We know that generally matrix multiplication is not commutative, i.e. generally $AB \neq BA$ so the order is essential. However

$$A^k A^j = A^j A^k = A^{k+j},$$

and from here it is easy to show that for arbitrary polynomials $p$ and $q$

$$p(A)q(A) = q(A)p(A) = R(A)$$

where $R(z) = p(z)q(z)$.

That means that when dealing only with polynomials of an operator $A$, one does not need to worry about non-commutativity, and act like $A$ is simply an independent (scalar) variable. In particular, if a polynomial $p(z)$ can be represented as a product of monomials

$$p(z) = a(z - z_1)(z - z_2)\ldots(z - z_N),$$

where $z_1, z_2, \ldots, z_N$ are the roots of $p$, then $p(A)$ can be represented as

$$p(A) = a(A - z_1 I)(A - z_2 I)\ldots(A - z_N I)$$

**2.2. Spectral Mapping Theorem.** Let us recall that the spectrum $\sigma(A)$ of a square matrix (an operator) $A$ is the set of all eigenvalues of $A$ (not counting multiplicities).

**Theorem 2.1** (Spectral Mapping Theorem)**.** *For a square matrix $A$ and an arbitrary polynomial $p$*

$$\sigma(p(A)) = p(\sigma(A)).$$

*In other words, $\mu$ is an eigenvalue of $p(A)$ if and only if $\mu = p(\lambda)$ for some eigenvalue $\lambda$ of $A$.*

Note, that as stated, this theorem does not say anything about multiplicities of the eigenvalues.

**Remark.** Note, that one inclusion is trivial. Namely, if $\lambda$ is an eigenvalue of $A$, $A\mathbf{x} = \lambda\mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$, then $A^k\mathbf{x} = \lambda^k\mathbf{x}$, and $p(A)\mathbf{x} = p(\lambda)\mathbf{x}$, so $p(\lambda)$ is an eigenvalue of $p(A)$. That means that the inclusion $p(\sigma(A)) \subset \sigma(p(A))$ is trivial.

If we consider a particular case $\mu = 0$ of the above theorem, we get the following corollary.

**Corollary 2.2.** *Let $A$ be a square matrix with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ and let $p$ be a polynomial. Then $p(A)$ is invertible if and only if*

$$p(\lambda_k) \neq 0 \qquad \forall k = 1, 2, \dots, n.$$

**Proof of Theorem 2.1.** As it was discussed above, the inclusion

$$p(\sigma(A)) \subset \sigma(p(A))$$

is trivial.

To prove the opposite inclusion $\sigma(p(A)) \subset p(\sigma(A))$ take a point $\mu \in \sigma(p(A))$. Denote $q(z) = p(z) - \mu$, so $q(A) = p(A) - \mu I$. Since $\mu \in \sigma(p(A))$ the operator $q(A) = p(A) - \mu I$ is not invertible.

Let us represent the polynomial $q(z)$ as a product of monomials,

$$q(z) = a(z - z_1)(z - z_2) \dots (z - z_N).$$

Then, as it was discussed above in Section 2.1, we can represent

$$q(A) = a(A - z_1 I)(A - z_2 I) \dots (A - z_N I).$$

The operator $q(A)$ is not invertible, so one of the terms $A - z_k I$ must be not invertible (because a product of invertible transformations is always invertible). That means $z_k \in \sigma(A)$.

On the other hand $z_k$ is a root of $q$, so

$$0 = q(z_k) = p(z_k) - \mu$$

and therefore $\mu = p(z_k)$. So we have proved the inclusion $\sigma(p(A)) \subset p(\sigma(A))$. □

**Exercises.**

**2.1.** An operator $A$ is called *nilpotent* if $A^k = \mathbf{0}$ for some $k$. Prove that if $A$ is nilpotent, then $\sigma(A) = \{0\}$ (i.e. that 0 is the only eigenvalue of $A$).

Can you do it without using the spectral mapping theorem?

## 3. Generalized eigenspaces. Geometric meaning of algebraic multiplicity

### 3.1. Invariant subspaces.

**Definition.** Let $A : V \to V$ be an operator (linear transformation) in a vector space $V$. A subspace $E$ of the vector space $V$ is called an invariant subspace of the operator $A$ (or, shortly, $A$-invariant) if $AE \subset E$, i.e. if $A\mathbf{v} \in E$ for all $\mathbf{v} \in E$.

If $E$ is $A$-invariant, then

$$A^2E = A(AE) \subset AE \subset E,$$

i.e. $E$ is $A^2$-invariant.

Similarly one can show (using induction, for example), that if $AE \subset E$ then

$$A^kE \subset E \qquad \forall k \geq 1.$$

This implies that $P(A)E \subset E$ for any polynomial $p$, i.e. that:

| any $A$-invariant subspace $E$ is an invariant subspace of $p(A)$. |
|---|

If $E$ is an $A$-invariant subspace, then for all $\mathbf{v} \in E$ the result $A\mathbf{v}$ also belongs to $E$. Therefore we can treat $A$ as an operator acting on $E$, not on the whole space $V$.

Formally, for an $A$-invariant subspace $E$ we define the so-called restriction $A|_E : E \to E$ of $A$ onto $E$ by

$$(A|_E)\mathbf{v} = A\mathbf{v} \qquad \forall \mathbf{v} \in E.$$

Here we changed domain and target space of the operator, but the rule assigning value to the argument remains the same.

We will need the following simple lemma

**Lemma 3.1.** *Let $p$ be a polynomial, and let $E$ be an $A$-invariant subspace. Then*

$$p(A|_E) = p(A)|_E.$$

**Proof.** The proof is trivial □

If $E_1, E_2, \dots, E_r$ a basis of $A$-invariant subspaces, and $A_k := A|_{E_k}$ are the corresponding restrictions, then, since $AE_k = A_kE_k \subset E_k$, the operators $A_k$ act independently of each other (do not interact), and to analyze action of $A$ we can analyze operators $A_k$ separately.

In particular, if we pick a basis in each subspace $E_k$ and join them to get a basis in $V$ (see Theorem 2.6 from Chapter 4) then the operator $A$ will have in this basis the following block-diagonal form

$$A = \begin{pmatrix} A_1 & & & \mathbf{0} \\ & A_2 & & \\ & & \ddots & \\ \mathbf{0} & & & A_r \end{pmatrix}$$

(of course, here we have the correct ordering of the basis in $V$, first we take a basis in $E_1$,then in $E_2$ and so on).

Our goal now is to pick a basis of invariant subspaces $E_1, E_2, \dots, E_r$ such that the restrictions $A_k$ have a simple structure. In this case we will get a basis in which the matrix of $A$ has a simple structure.

The eigenspaces $\operatorname{Ker}(A - \lambda_k I)$ would be good candidates, because the restriction of $A$ to the eigenspace $\operatorname{Ker}(A-\lambda_k I)$ is simply $\lambda_k I$. Unfortunately, as we know eigenspaces do not always form a basis (they form a basis if and only if $A$ can be diagonalized, cf Theorem 2.1 in Chapter 4.

However, the so-called *generalized eigenspaces* will work.

### 3.2. Generalized eigenspaces.

**Definition 3.2.** A vector $\mathbf{v}$ is called a generalized eigenvector (corresponding to an eigenvalue $\lambda$) if $(A - \lambda I)^k \mathbf{v} = \mathbf{0}$ for some $k \geq 1$.

The collection $E_\lambda$ of all generalized eigenvectors, together with $\mathbf{0}$ is called the generalized eigenspace (corresponding to the eigenvalue $\lambda$.

In other words one can represent the generalized eigenspace $E_\lambda$ as

$$E_\lambda = \bigcup_{k \geq 1} \operatorname{Ker}(A - \lambda I)^k. \tag{3.1}$$

The sequence $\operatorname{Ker}(A - \lambda I)^k$, $k = 1, 2, 3, \dots$ is an increasing sequence of subspaces, i.e.

$$\operatorname{Ker}(A - \lambda I)^k \subset \operatorname{Ker}(A - \lambda I)^{k+1} \qquad \forall k \geq 1.$$

The representation (3.1) does not look very simple, for it involves an infinite union. However, the sequence of the subspaces $\operatorname{Ker}(A-\lambda I)^k$ stabilizes, i.e.

$$\operatorname{Ker}(A - \lambda I)^k = \operatorname{Ker}(A - \lambda I)^{k+1} \qquad \forall k \geq k_\lambda,$$

so, in fact one can take the finite union.

To show that the sequence of kernels stabilizes, let us notice that if for finite-dimensional subspaces $E$ and $F$ we have $E \subsetneqq F$ (symbol $E \subsetneqq F$ means that $E \subset F$ but $E \neq F$), then $\dim E < \dim F$.

Since $\dim \operatorname{Ker}(A-\lambda I)^k \le \dim V < \infty$, it cannot grow to infinity, so at some point

$$\operatorname{Ker}(A-\lambda I)^k = \operatorname{Ker}(A-\lambda I)^{k+1}.$$

The rest follows from the lemma below.

**Lemma 3.3.** *Let for some $k$*

$$\operatorname{Ker}(A-\lambda I)^k = \operatorname{Ker}(A-\lambda I)^{k+1}.$$

*Then*

$$\operatorname{Ker}(A-\lambda I)^{k+r} = \operatorname{Ker}(A-\lambda I)^{k+r+1} \qquad \forall r \ge 0.$$

**Proof.** Let $\mathbf{v} \in \operatorname{Ker}(A-\lambda I)^{k+r+1}$, i.e. $(A-\lambda I)^{k+r+1}\mathbf{v} = \mathbf{0}$. Then

$$\mathbf{w} := (A-\lambda I)^r \in \operatorname{Ker}(A-\lambda I)^{k+1}.$$

But we know that $\operatorname{Ker}(A-\lambda I)^k = \operatorname{Ker}(A-\lambda I)^{k+1}$ so $\mathbf{w} \in \operatorname{Ker}(A-\lambda I)^k$, which means $(A-\lambda I)^k\mathbf{w} = \mathbf{0}$. Recalling the definition of $\mathbf{w}$ we get that

$$(A-\lambda I)^{k+r}\mathbf{v} = (A-\lambda I)^k\mathbf{w} = \mathbf{0}$$

so $\mathbf{v} \in \operatorname{Ker}(A-\lambda I)^{k+r}$. We proved that $\operatorname{Ker}(A-\lambda I)^{k+r+1} \subset \operatorname{Ker}(A-\lambda I)^{k+r}$. The opposite inclusion is trivial. □

**Definition.** The number $d = d(\lambda)$ on which the sequence $\operatorname{Ker}(A-\lambda I)^k$ stabilizes, i.e. the number $d$ such that

$$\operatorname{Ker}(A-\lambda I)^{d-1} \subsetneqq \operatorname{Ker}(A-\lambda I)^d = \operatorname{Ker}(A-\lambda I)^{d+1}$$

is called the depth of the eigenvalue $\lambda$.

It follows from the definition of the depth, that for the generalized eigenspace $E_\lambda$

$$(A-\lambda I)^{d(\lambda)}\mathbf{v} = \mathbf{0} \qquad \forall \mathbf{v} \in E_\lambda. \tag{3.2}$$

Now let us summarize, what we know about generalized eigenspaces.

a) $E_\lambda$ is an invariant subspace of $A$, $AE_\lambda \subset E_\lambda$.

b) If $d(\lambda)$ is the depth of the eigenvalue $\lambda$, then

$$((A-\lambda I)|_{E_\lambda})^{d(\lambda)} = (A|_{E_\lambda} - \lambda I_{E_\lambda})^{d(\lambda)} = \mathbf{0}.$$

(this is just another way of writing (3.2))

c) $\sigma(A|_{E_\lambda}) = \{\lambda\}$, because the operator $A|_{E_\lambda} - \lambda I_{E_\lambda}$, is nilpotent, see 2, and the spectrum of nilpotent operator consists of one point 0, see Problem 2.1

Now we are ready to state the main result of this section. Let $A : V \to V$.

**Theorem 3.4.** *Let $\sigma(A)$ consists of $r$ points $\lambda_1, \lambda_2, \dots, \lambda_r$, and let $E_k := E_{\lambda_k}$ be the corresponding generalized eigenspaces. Then the system of subspace $E_1, E_2, \dots, E_r$ is a basis of subspaces in $V$.*

**Remark 3.5.** If we join the bases in all generalized eigenspaces $E_k$, then by Theorem 2.6 from Chapter 4 we will get a basis in the whole space. In this basis the matrix of the operator $A$ has the block diagonal form $A = \operatorname{diag}\{A_1, A_2, \dots, A_r\}$, where $A_k := A|_{E_k}$, $E_k = E_{\lambda_k}$. It is also easy to see, see (3.2) that the operators $N_k := A_k - \lambda_k I_{E_k}$ are nilpotent, $N_k^{d_k} = \mathbf{0}$.

**Proof of Theorem 3.4.** Let $m_k$ be the multiplicity of the eigenvalue $\lambda_k$, so $p(z) = \prod_{k=1}^{r}(z-\lambda_k)^{m_k}$ is the characteristic polynomial of $A$. Define

$$p_k(z) = p(z)/(z-\lambda_k)^{m_k} = \prod_{j\neq k}(z-\lambda_j)^{m_j}.$$

**Lemma 3.6.**

$$(A-\lambda_k I)^{m_k}|_{E_k} = \mathbf{0}, \tag{3.3}$$

**Proof.** There are 2 possible simple proofs. The first one is to notice that $m_k \geq d_k$, where $d_k$ is the depth of the eigenvalue $\lambda_k$ and use the fact that

$$(A-\lambda_k I)^{d_k}|_{E_k} = (A_k - \lambda_k I_{E_k})^{m_k} = \mathbf{0},$$

where $A_k := A|_{E_k}$ (property 2 of the generalized eigenspaces).

The second possibility is to notice that according to the Spectral Mapping Theorem, see Corollary 2.2, the operator $P_k(A)|_{E_k} = p_k(A_k)$ is invertible. By the Cayley–Hamilton Theorem (Theorem 1.1)

$$\mathbf{0} = p(A) = (A-\lambda_k I)^{m_k} p_k(A),$$

and restriction all operators to $E_k$ we get

$$\mathbf{0} = p(A_k) = (A_k - \lambda_k I_{E_k})^{m_k} p_k(A_k),$$

so

$$(A_k - \lambda_k I_{E_k})^{m_k} = p(A_k)p_k(A_k)^{-1} = \mathbf{0}\, p_k(A_k)^{-1} = \mathbf{0}.$$

□

To prove the theorem define

$$q(z) = \sum_{k=1}^{r} p_k(z).$$

Since $p_k(\lambda_j) = 0$ for $j \neq k$ and $p_k(\lambda_k) \neq 0$, we can conclude that $q(\lambda_k) \neq 0$ for all $k$. Therefore, by the Spectral Mapping Theorem, see Corollary 2.2, the operator

$$B = q(A)$$

is invertible.

Note that $BE_k \subset E_k$ (any $A$-invariant subspace is also $p(A)$-invariant). Since $B$ is an invertible operator, $\dim(BE_k) = \dim E_k$, which together with $BE_k \subset E_k$ implies $BE_k = E_k$. Multiplying the last identity by $B^{-1}$ we get that $B^{-1}E_k = E_k$, i.e. that $E_k$ is an invariant subspace of $B^{-1}$.

Note also, that it follows from (3.3) that

$$p_k(A)|_{E_j} = \mathbf{0} \qquad \forall j \neq k,$$

because $p_k(A)|_{E_j} = p_k(A_j)$ and $p_k(A_j)$ contains the factor $(A_j - \lambda_j I_{E_j})^{m_j} = \mathbf{0}$.

Define the operators $\mathcal{P}_k$ by

$$\mathcal{P}_k = B^{-1} p_k(A).$$

**Lemma 3.7.** *For the operators $\mathcal{P}_k$ defined above*

a) $\mathcal{P}_1 + \mathcal{P}_2 + \ldots + \mathcal{P}_r = I$;

b) $\mathcal{P}_k|_{E_j} = \mathbf{0}$ *for* $j \neq k$;

c) $\operatorname{Ran} \mathcal{P}_k \subset E_k$;

d) *moreover,* $\mathcal{P}_k \mathbf{v} = \mathbf{v}$ $\forall \mathbf{v} \in E_k$, *so, in fact* $\operatorname{Ran} \mathcal{P}_k = E_k$.

**Proof.** Property 1 is trivial:

$$\sum_{k=1}^{r} \mathcal{P}_k = B^{-1} \sum_{k=1}^{r} \mathcal{P}_k p_k(A) = B^{-1} B = I.$$

Property 2 follows from (3.3). Indeed, $p_k(A)$ contains the factor $(A - \lambda_j)^{m_j}$, restriction of which to $E_j$ is zero. Therefore $p_k(A)|_{E_j} = \mathbf{0}$ and thus $\mathcal{P}_k|_{E_j} = B^{-1} p_k(A)|_{E_j} = \mathbf{0}$.

To prove property 3, recall that according to Cayley–Hamilton Theorem $p(A) = \mathbf{0}$. Since $p(z) = (z - \lambda_k)^{m_k} p_k(z)$, we have for $\mathbf{w} = p_k(A)\mathbf{v}$

$$(A - \lambda_k I)^{m_k} \mathbf{w} = (A - \lambda_k I)^{m_k} p_k(A)\mathbf{v} = p(A)\mathbf{v} = \mathbf{0}.$$

That means, any vector $\mathbf{w}$ in $\operatorname{Ran} p_k(A)$ is annihilated by some power of $(A - \lambda_k I)$, which by definition means that $\operatorname{Ran} p_k(A) \subset E_k$.

To prove the last property, let us notice that it follows from (3.3) that for $\mathbf{v} \in E_k$

$$p_k(A)\mathbf{v} = \sum_{j=1}^{r} p_j(A)\mathbf{v} = B\mathbf{v},$$

which implies $\mathcal{P}_k \mathbf{v} = B^{-1} B \mathbf{v} = \mathbf{v}$. □

Now we are ready to complete the proof of the theorem. Take $\mathbf{v} \in V$ and define $\mathbf{v}_k = \mathcal{P}_k \mathbf{v}$. Then according to Statement c) of Lemma 3.7, $\mathbf{v}_k \in E_k$,

and by Statement a),

$$\mathbf{v} = \sum_{k=1}^{r} \mathbf{v}_k,$$

so $\mathbf{v}$ admits a representation as a linear combination.

To show that this representation is unique, we can just note, that if $\mathbf{v}$ is represented as $\mathbf{v} = \sum_{k=1}^{r} \mathbf{v}_k$, $\mathbf{v}_k \in E_k$, then it follows from the Statements b) and d) of Lemma 3.7 that

$$\mathcal{P}_k \mathbf{v} = \mathcal{P}_k(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_r) = \mathcal{P}_k \mathbf{v}_k = \mathbf{v}_k.$$

□

### 3.3. Geometric meaning of algebraic multiplicity.

**Proposition 3.8.** *Algebraic multiplicity of an eigenvalue equals the dimension of the corresponding generalized eigenspace.*

**Proof.** According to Remark 3.5, if we joint bases in generalized eigenspaces $E_k = E_{\lambda_k}$ to get a basis in the whole space, the matrix of $A$ in any such basis has a block-diagonal form $\operatorname{diag}\{A_1, A_2, \ldots, A_r\}$, where $A_k := A|_{E_k}$. Operators $N_k = A_k - \lambda_k I_{E_k}$ are nilpotent, so $\sigma(N_k) = \{0\}$. Therefore, the spectrum of the operator $A_k$ (recall that $A_k = N_k - \lambda_k I$) consists of one eigenvalue $\lambda_k$ of (algebraic) multiplicity $n_k = \dim E_k$. The multiplicity equals $n_k$ because an operator in a finite-dimensional space $V$ has exactly $\dim V$ eigenvalues counting multiplicities, and $A_k$ has only one eigenvalue.

Note that we are free to pick bases in $E_k$, so let us pick them in such a way that the corresponding blocks $A_k$ are upper triangular. Then

$$\det(A - \lambda I) = \prod_{k=1}^{r} \det(A_k - \lambda I_{E_k}) = \prod_{k=1}^{r} (\lambda_k - \lambda)^{n_k}.$$

But this means that the algebraic multiplicity of the eigenvalue $\lambda_k$ is $n_k = \dim E_{\lambda_k}$. □

**3.4. An important application.** The following corollary is very important for differential equations.

**Corollary 3.9.** *Any operator $A$ in $V$ can be represented as $A = D + N$, where $D$ is diagonalizable (i.e. diagonal in some basis) and $N$ is nilpotent ($N^m = \mathbf{0}$ for some $m$), and $DN = ND$.*

**Proof.** As we discussed above, see Remark 3.5, if we join the bases in $E_k$ to get a basis in $V$, then in this basis $A$ has the block diagonal form $A = \operatorname{diag}\{A_1, A_2, \ldots, A_r\}$, where $A_k := A|_{E_k}$, $E_k = E_{\lambda_k}$. The operators $N_k := A_k - \lambda_k I_{E_k}$ are nilpotent, and the operator $D = \operatorname{diag}\{\lambda_1 I_{E_1}, \lambda_2 I_{E_2} \ldots, \lambda_r I_{E_r}\}$

is diagonal (in this basis). Notice also that $\lambda_k I_{E_k} N_k = N_k \lambda_k I_{E_k}$ (identity operator commutes with any operator), so the block diagonal operator $N = \operatorname{diag}\{N_1, N_2, \ldots, N_r\}$ commutes with $D$, $DN = ND$. Therefore, defining $N$ as the block diagonal operator $N = \operatorname{diag}\{N_1, N_2, \ldots, N_r\}$ we get the desired decomposition. □

This corollary allows us to compute functions of operators. Let us recall that if $p$ is a polynomial of degree $d$, then $p(a + x)$ can be computed with the help of Taylor's formula

$$p(a + x) = \sum_{k=0}^{d} \frac{p^{(k)}(a)}{k!} x^k$$

This formula is an algebraic identity, meaning that for each polynomial $p$ we can check that the formula is true using formal algebraic manipulations with $a$ and $x$ and not caring about their nature.

Since operators $D$ and $N$ commute, $DN = ND$, the same rules as for usual (scalar) variables apply to them, and we can write (by plugging $D$ instead of $a$ and $N$ instead of $x$

$$p(A) = p(D + N) = \sum_{k=0}^{d} \frac{p^{(k)}(D)}{k!} N^k.$$

Here, to compute the derivative $p^{(k)}(D)$ we first compute the $k$th derivative of the polynomial $p(x)$ (using the usual rules from calculus), and then plug $D$ instead of $x$.

But since $N$ is nilpotent, $N^m = \mathbf{0}$ for some $m$, only first $m$ terms can be non-zero, so

$$p(A) = p(D + N) = \sum_{k=0}^{m-1} \frac{f^{(k)}(D)}{k!} N^k.$$

In $m$ is much smaller than $d$, this formula makes computation of $p(A)$ much easier.

The same approach works if $p$ is not a polynomial, but an infinite power series. For general power series we have to be careful about convergence of all the series involved, so we cannot say that the formula is true for an arbitrary power series $p(x)$. However, if the radius of convergence of the power series is $\infty$, then everything works fine. In particular, if $p(x) = e^x$, then, using the fact that $(e^x)' = e^x$ we get.

$$e^A = \sum_{k=0}^{m-1} \frac{e^D}{k!} N^k = e^D \sum_{k=0}^{m-1} \frac{1}{k!} N^k$$

This formula has important applications in differential equation.

Note, that the fact that $ND = DN$ is essential here!

## 4. Structure of nilpotent operators

Recall, that an operator $A$ in a vector space $V$ is called *nilpotent* if $A^k = 0$ for some exponent $k$.

In the previous section we have proved, see Remark 3.5, that if we join the bases in all generalized eigenspaces $E_k = E_{\lambda_k}$ to get a basis in the whole space, then the operator $A$ has in this basis a block diagonal form $\operatorname{diag}\{A_1, A_2, \dots, A_r\}$ and operators $A_k$ ca be represented as $A_k = \lambda_k I + N_k$, where $N_k$ are nilpotent operators.

In each generalized eigenspace $E_k$ we want to pick up a basis such that the matrix of $A_k$ in this basis has the simplest possible form. Since matrix (in any basis) of the identity operator is the identity matrix, we need to find a basis in which the nilpotent operator $N_k$ has a simple form.

Since we can deal with each $N_k$ separately, we will need to consider the following problem:

> For a nilpotent operator $A$ find a basis such that the matrix of $A$ in this basis is simple.

Let see, what does it mean for a matrix to have a simple form. It is easy to see that the matrix

$$\begin{pmatrix} 0 & 1 & & & 0 \\ & 0 & 1 & & \\ & & 0 & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & 0 \end{pmatrix} \tag{4.1}$$

is nilpotent.

These matrices (together with $1 \times 1$ zero matrices) will be our "building blocks". Namely, we will show that for any nilpotent operator one can find a basis such that the matrix of the operator in this basis has the block diagonal form $\operatorname{diag}\{A_1, A_2, \dots, A_r\}$, where each $A_k$ is either a block of form (4.1) or a $1 \times 1$ zero block.

Let us see what we should be looking for. Suppose the matrix of an operator $A$ has in a basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ the form (4.1). Then

$$A\mathbf{v}_1 = \mathbf{0} \tag{4.2}$$

and

$$A\mathbf{v}_{k+1} = \mathbf{v}_k, \qquad k = 1, 2, \dots, p-1. \tag{4.3}$$

Thus we have to be looking for the chains of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ satisfying the above relations (4.2), (4.3).

### 4.1. Cycles of generalized eigenvectors.

**Definition.** Let $A$ be a nilpotent operator. A chain of non-zero vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ satisfying relations (4.2), (4.3) is called *a cycle of generalized eigenvectors* of $A$. The vector $\mathbf{v}_1$ is called the *initial vector* of the cycle, the vector $\mathbf{v}_p$ is called the *end vector* of the cycle, and the number $p$ is called the *length* of the cycle.

**Remark.** A similar definition can be made for an arbitrary operator. Then all vectors $\mathbf{v}_k$ must belong to the same generalized eigenspace $E_\lambda$, and they must satisfy the identities

$$(A - \lambda I)\mathbf{v}_1 = \mathbf{0}, \qquad (A - \lambda I)\mathbf{v}_{k+1} = \mathbf{v}_k, \quad k = 1, 2, \dots, p-1,$$

**Theorem 4.1.** *Let $A$ be a nilpotent operator, and let $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_r$ be cycles of its generalized eigenvectors, $\mathcal{C}_k = \mathbf{v}_1^k, \mathbf{v}_2^k, \dots, \mathbf{v}_{p_k}^k$, $p_k$ being the length of the cycle $\mathcal{C}_k$. Assume that the initial vectors $\mathbf{v}_1^1, \mathbf{v}_1^2, \dots, \mathbf{v}_1^r$ are linearly independent. Then no vector belongs to two cycles, and the union of all the vectors from all the cycles is a linearly independent.*

**Proof.** Let $n = p_1 + p_2 + \ldots + p_r$ be the total number of vectors in all the cycles[2]. We will use induction in $n$. If $n = 1$ the theorem is trivial.

Let us now assume, that the theorem is true for all operators and for all collection of cycles, as long as the total number of vectors in all the cycles is strictly less than $n$.

Without loss of generality we can assume that the vectors $\mathbf{v}_j^k$ span the whole space $V$, because, otherwise we can consider instead of the operator $A$ its restriction onto the invariant subspace $\operatorname{span}\{\mathbf{v}_j^k : k = 1, 2, \dots, r,\ 1 \le j \le p_k\}$.

Consider the subspace $\operatorname{Ran} A$. It follows from the relations (4.2), (4.3) that vectors $\mathbf{v}_j^k : k = 1, 2, \dots, r,\ 1 \le j \le p_k - 1$ span $\operatorname{Ran} A$. Note that if $p_k > 1$ then the system $\mathbf{v}_1^k, \mathbf{v}_2^k, \dots, \mathbf{v}_{p_k-1}^k$ is a cycle, and that $A$ annihilates any cycle of length 1.

Therefore, we have finitely many cycles, and initial vectors of these cycles are linearly independent, so the induction hypothesis applies, and the vectors $\mathbf{v}_j^k : k = 1, 2, \dots, r,\ 1 \le j \le p_k - 1$ are linearly independent. Since these vectors also span $\operatorname{Ran} A$, we have a basis there. Therefore,

$$\operatorname{rank} A = \dim \operatorname{Ran} A = n - r$$

[2]Here we just count vectors in each cycle, and add all the numbers. We do not care if some cycles have a common vector, we count this vector in each cycle it belongs to (of course, according to the theorem, it is impossible, but initially we cannot assume that)

(we had $n$ vectors, and we removed one vector $\mathbf{v}^k_{p_k}$ from each cycle $\mathcal{C}_k$, $k = 1, 2, \dots, r$, so we have $n - r$ vectors in the basis $\mathbf{v}^k_j : k = 1, 2, \dots, r,\ 1 \le j \le p_k - 1$ ). On the other hand $A\mathbf{v}^k_1 = 0$ for $k = 1, 2, \dots, r$, and since these vectors are linearly independent $\dim \operatorname{Ker} A \ge r$. By the Rank Theorem (Theorem 7.1 from Chapter 2)

$$\dim V = \operatorname{rank} A + \dim \operatorname{Ker} A = (n - r) + \dim \operatorname{Ker} A \ge (n - r) + r = n$$

so $\dim V \ge n$.

On the other hand $V$ is spanned by $n$ vectors, therefore the vectors $\mathbf{v}^k_j : k = 1, 2, \dots, r,\ 1 \le j \le p_k$, form a basis, so they are linearly independent $\square$

### 4.2. Jordan canonical form of a nilpotent operator.

**Theorem 4.2.** *Let $A : V \to V$ be a nilpotent operator. Then $V$ has a basis consisting of union of cycles of generalized eigenvectors of the operator $A$.*

**Proof.** We will use induction in $n$ where $n = \dim V$. For $n = 1$ the theorem is trivial.

Assume that the theorem is true for any operator acting in a space of dimension strictly less than $n$.

Consider the subspace $X = \operatorname{Ran} A$. $X$ is an invariant subspace of the operator $A$, so we can consider the restriction $A|_X$.

Since $A$ is not invertible, $\dim \operatorname{Ran} A < \dim V$, so by the induction hypothesis there exist cycles $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_r$ of generalized eigenvectors such that their union is a basis in $X$. Let $\mathcal{C}_k = \mathbf{v}^k_1, \mathbf{v}^k_2, \dots, \mathbf{v}^k_{p_k}$, where $\mathbf{v}^k_1$ is the initial vector of the cycle.

Since the end vector $\mathbf{v}^k_{p_k}$ belong to $\operatorname{Ran} A$, one can find a vector $\mathbf{v}^k_{p_k+1}$ such that $A\mathbf{v}_{p_k+1} = \mathbf{v}^k_{p_k}$. So we can extend each cycle $\mathcal{C}_k$ to a bigger cycle $\widetilde{\mathcal{C}}_k = \mathbf{v}^k_1, \mathbf{v}^k_2, \dots, \mathbf{v}^k_{p_k}, \mathbf{v}^k_{p_k+1}$. Since the initial vectors $\mathbf{v}^k_1$ of cycles $\widetilde{\mathcal{C}}_k$, $k = 1, 2, \dots, r$ are linearly independent, the above Theorem 4.1 implies that the union of these cycles is a linearly independent system.

By the definition of the cycle we have $\mathbf{v}^k_1 \in \operatorname{Ker} A$, and we assumed that the initial vectors $\mathbf{v}^k_1$, $k = 1, 2, \dots, r$ are linearly independent. Let us complete this system to a basis in $\operatorname{Ker} A$, i.e. let find vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q$ such that the system $\mathbf{v}^1_1, \mathbf{v}^2_1, \dots, \mathbf{v}^r_1, \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q$ is a basis in $\operatorname{Ker} A$ (it may happen that the system $\mathbf{v}^k_1$, $k = 1, 2, \dots, r$ is already a basis in $\operatorname{Ker} A$, in which case we put $q = 0$ and add nothing).

The vector $\mathbf{u}_j$ can be treated as a cycle of length 1, so we have a collection of cycles $\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2, \dots, \widetilde{\mathcal{C}}_r, \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q$, whose initial vectors are linearly independent. So, we can apply Theorem 4.1 to get that the union of all these cycles is a linearly independent system.

To show that it is a basis, let us count the dimensions. We know that the cycles $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_r$ have $\dim \operatorname{Ran} A = \operatorname{rank} A$ vectors total. Each cycle $\widetilde{\mathcal{C}}_k$ was obtained from $\mathcal{C}_k$ by adding 1 vector to it, so the total number of vectors in all the cycles $\widetilde{\mathcal{C}}_k$ is $\operatorname{rank} A + r$.

We know that $\dim \operatorname{Ker} A = r + q$ (because $\mathbf{v}_1^1, \mathbf{v}_1^2, \dots, \mathbf{v}_1^r, \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q$ is a basis there). We added to the cycles $\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2, \dots, \widetilde{\mathcal{C}}_r$ additional $q$ vectors, so we got

$$\operatorname{rank} A + r + q = \operatorname{rank} A + \dim \operatorname{Ker} A = \dim V$$

linearly independent vectors. But $\dim V$ linearly independent vectors is a basis. □

**Definition.** A basis consisting of a union of cycles of generalized eigenvectors of a nilpotent operator $A$ (existence of which is guaranteed by the Theorem 4.2) is called a Jordan canonical basis for $A$.

Note, that such basis is not unique.

**Corollary 4.3.** *Let $A$ be a nilpotent operator. There exists a basis (a Jordan canonical basis) such that the matrix of $A$ in this basis is a block diagonal* $\operatorname{diag}\{A_1, A_2, \dots, A_r\}$, *where all $A_k$ (except may be one) are blocks of form (4.1), and one of the blocks $A_k$ can be zero.*

The matrix of $A$ in a Jordan canonical basis is called the Jordan canonical form of the operator $A$. We will see later that the Jordan canonical form is unique, if we agree on how to order the blocks (i.e. on how to order the vectors in the basis).

**Proof of Corollary 4.3.** According to Theorem 4.2 one can find a basis consisting of a union of cycles of generalized eigenvectors. A cycle of size $p$ gives rise to a $p \times p$ diagonal block of form (4.1), and a cycle of length 1 correspond to a $1 \times 1$ zero block. We can join these $1 \times 1$ zero blocks in one large zero block (because off-diagonal entries are 0). □

**4.3. Dot diagrams. Uniqueness of the Jordan canonical form.** There is a good way of visualizing Theorem 4.2 and Corollary 4.3, the so-called dot diagrams. This methods also allows us to answer many natural questions, like "is the block diagonal representation given by Corollary 4.3 unique?"

Of course, if we treat this question literally, the answer is "no", for we always can change the order of the blocks. But, if we exclude such trivial possibilities, for example by agreeing on some order of blocks (say, if we put all non-zero blocks in decreasing order, and then put the zero block), is the representation unique, or not?

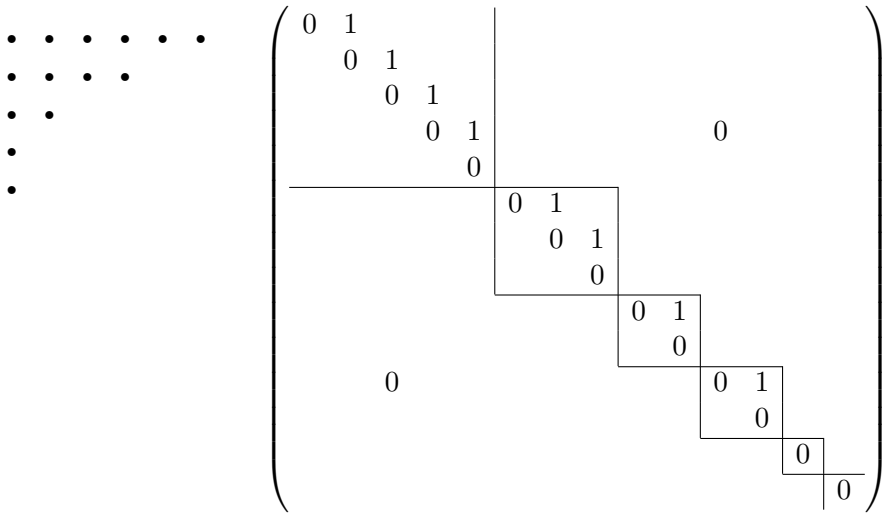


**Figure 1.** Dot diagram and corresponding Jordan canonical form of a nilpotent operator

To better understand the structure of nilpotent operators, described in the Section 4.1, let us draw the so-called dot diagram. Namely, suppose we have a basis, which is a union of cycles of generalized eigenvalues. Let us represent the basis by an array of dots, so that each column represents a cycle. The first row consists of initial vectors of cycles, and we arrange the columns (cycles) by their length, putting the longest one to the left.

On the figure 1 we have the dot diagram of a nilpotent operator, as well as its Jordan canonical form. This dot diagram shows, that the basis has 1 cycle of length 5, one cycle of length 3, two cycles of length 2, and 2 cycles of length 1. The cycle of length 5 corresponds to the $5 \times 5$ block of the matrix, the cycle of length 3 correspond to 3 non-zero block, and two cycles of length 2 correspond to two $2 \times 2$ blocks. Three cycles of length 1 correspond to two zero entries on the diagonal. Here in each block we only giving the main diagonal and the diagonal above it; all other entries of the matrix are zero.

If we agree on the ordering of the blocks, there is a one-to-one correspondence between dot diagrams and Jordan canonical forms (for nilpotent operators). So, the question about uniqueness of the Jordan canonical form is equivalent to the question about uniqueness of the dot diagram.

To answer this question, let us analyze, how the operator $A$ transforms the dot diagram. Since the operator $A$ annihilates initial vectors of the cycles, and moves vector $\mathbf{v}_{k+1}$ of a cycle to the vector $\mathbf{v}_k$, we can see that

the operator $A$ acts on its dot diagram by deleting the first (top) row of the diagram.

The new dot diagram corresponds to a Jordan canonical basis in $\operatorname{Ran} A$, and allows us to write down the Jordan canonical form for the restriction $A|_{\operatorname{Ran} A}$.

Similarly, it is not hard to see that the operator $A^k$ removes the first $k$ rows of the dot diagram. Therefore, if for all $k$ we know the dimensions $\dim \operatorname{Ker}(A^k)$, we know the dot diagram of the operator $A$. Namely, the number of dots in the first row is $\dim \operatorname{Ker} A$, the number of dots in the second row is

$$\dim \operatorname{Ker}(A^2) - \dim \operatorname{Ker} A,$$

and the number of dots in the $k$th row is

$$\dim \operatorname{Ker}(A^k) - \dim \operatorname{Ker}(A^{k+1}).$$

But this means that the dot diagram, which was initially defined using a Jordan canonical basis, does not depend on a particular choice of such a basis. Therefore, the dot diagram, is unique! This implies that if we agree on the order of the blocks, then the Jordan canonical form is unique.

**4.4. Computing a Jordan canonical basis.** Let us say few words about computing a Jordan canonical basis for a nilpotent operator. Let $p_1$ be the largest integer such that $A^{p_1} \neq \mathbf{0}$ (so $A^{p_1+1} = \mathbf{0}$). One can see from the above analysis of dot diagrams, that $p_1$ is the length of the longest cycle.

Computing operators $A^k$, $k = 1, 2, \dots, p_1$, and counting $\dim \operatorname{Ker}(A^k)$ we can construct the dot diagram of $A$. Now we want to put vectors instead of dots and find a basis which is a union of cycles.

We start by finding the longest cycles (because we know the dot diagram, we know how many cycles should be there, and what is the length of each cycle). Consider a basis in the column space $\operatorname{Ran}(A^{p_1})$. Name the vectors in this basis $\mathbf{v}_1^1, \mathbf{v}_1^2, \dots, \mathbf{v}_1^{r_1}$, these will be the initial vectors of the cycles. Then we find the end vectors of the cycles $\mathbf{v}_{p_1}^1, \mathbf{v}_{p_1}^2, \dots, \mathbf{v}_{p_1}^{r_1}$ by solving the equations

$$A^{p_1}\mathbf{v}_{p_1}^k = \mathbf{v}_1^k, \qquad k = 1, 2, \dots, r_1.$$

Applying consecutively the operator $A$ to the end vector $\mathbf{v}_{p_1}^k$, we get all the vectors $\mathbf{v}_j^k$ in the cycle. Thus, we have constructed all cycles of maximal length.

Let $p_2$ be the length of a maximal cycle among those that are left to find. Consider the subspace $\operatorname{Ran}(A^{p_2})$, and let $\dim \operatorname{Ran}(A^{p_2}) = r_2$. Since $\operatorname{Ran}(A^{p_1}) \subset \operatorname{Ran}(A^{p_2})$, we can complete the basis $\mathbf{v}_1^1, \mathbf{v}_1^2, \dots, \mathbf{v}_1^{r_1}$ to a basis

$\mathbf{v}_1^1, \mathbf{v}_1^2, \dots, \mathbf{v}_1^{r_1}, \mathbf{v}_1^{r_1+1}, \dots, \mathbf{v}_1^{r_2}$ in $\operatorname{Ran}(A^{r_2})$. Then we find end vectors of the cycles $\mathcal{C}_{r_1+1}, \dots, \mathcal{C}_{r_2}$ by solving (for $\mathbf{v}_{p_2}^k$) the equations

$$A^{p_1}\mathbf{v}_{p_2}^k = \mathbf{v}_1^k, \qquad k = r_1+1, r_1+2, \dots, r_2,$$

thus constructing th cycles of length $p_2$.

Let $p_3$ denote the length of a maximal cycle among ones left. Then, completing the basis $\mathbf{v}_1^1, \mathbf{v}_1^2, \dots, \mathbf{v}_1^{r_2}$ in $\operatorname{Ker}(A^{p_2})$ to a basis in $\operatorname{Ker}(A^{p_3}$ we construct the cycles of length $p_3$, and so on...

One final remark: as we discussed above, if we know the dot diagram, we know the canonical form, so after we have found a Jordan canonical basis, we do not need to compute the matrix of $A$ in this basis: we already know it!

## 5. Jordan decomposition theorem

**Theorem 5.1.** *Given an operator $A$ there exist a basis (Jordan canonical basis) such that the matrix of $A$ in this basis has a block diagonal form with blocks of form*

$$\begin{pmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \lambda & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix} \tag{5.1}$$

*where $\lambda$ is an eigenvalue of $A$. Here we assume that the block of size 1 is just $\lambda$.*

The block diagonal form from Theorem 5.1 is called the *Jordan canonical form* of the operator $A$. The corresponding basis is called a *Jordan canonical basis* for an operator $A$.

**Proof of Theorem 5.1.** According to Theorem 3.4 and Remark 3.5, if we join bases in the generalized eigenspaces $E_k = E_{\lambda_k}$ to get a basis in the whole space, the matrix of $A$ in this basis has a block diagonal form $\operatorname{diag}\{A_1, A_2, \dots, A_r\}$, where $A_k = A|_{E_k}$. The operators $N_k = A_k - \lambda_k I_{E_k}$ are nilpotent, so by Theorem 4.2 (more precisely, by Corollary 4.3) one can find a basis in $E_k$ such that the matrix of $N_k$ in this basis is the Jordan canonical form of $N_k$. To get the matrix of $A_k$ in this basis one just puts $\lambda_k$ instead of 0 on the main diagonal. □

**5.1. Remarks about computing Jordan canonical basis.** First of all let us recall that the computing of eigenvalues is the hardest part, but here we do not discuss this part, and assume that eigenvalues are already computed.

For each eigenvalue $\lambda$ we compute subspaces $\operatorname{Ker}(A-\lambda I)^k$, $k=1,2,\dots$ until the sequence of the subspaces stabilizes. In fact, since we have an increasing sequence of subspaces $(\operatorname{Ker}(A-\lambda I)^k \subset \operatorname{Ker}(A-\lambda I)^{k+1})$, then it is sufficient only to keep track of their dimension (or ranks of the operators $(A-\lambda I)^k$). For an eigenvalue $\lambda$ let $m=m_\lambda$ be the number where the sequence $\operatorname{Ker}(A-\lambda I)^k$ stabilizes, i.e. $m$ satisfies

$$\dim \operatorname{Ker}(A-\lambda I)^{m-1} < \dim \operatorname{Ker}(A-\lambda I)^m = \dim \operatorname{Ker}(A-\lambda I)^{m+1}.$$

Then $E_\lambda = \operatorname{Ker}(A-\lambda I)^m$ is the generalized eigenspace corresponding to the eigenvalue $\lambda$.

After we computed all the generalized eigenspaces there are two possible ways of action. The first way is to find a basis in each generalized eigenspace, so the matrix of the operator $A$ in this basis has the block-diagonal form $\operatorname{diag}\{A_1, A_2, \dots, A_r\}$, where $A_k = A|_{E_{\lambda_k}}$. Then we can deal with each matrix $A_k$ separately. The operators $N_k = A_k - \lambda_k I$ are nilpotent, so applying the algorithm described in Section 4.4 we get the Jordan canonical representation for $N_k$, and putting $\lambda_k$ instead of 0 on the main diagonal, we get the Jordan canonical representation for the block $A_k$. The advantage of this approach is that we are working with smaller blocks. But we need to find the matrix of the operator in a new basis, which involves inverting a matrix and matrix multiplication.

Another way is to find a Jordan canonical basis in each of the generalized eigenspaces $E_{\lambda_k}$ by working directly with the operator $A$, without splitting it first into the blocks. Again, the algorithm we outlined above in Section 4.4 works with a slight modification. Namely, when computing a Jordan canonical basis for a generalized eigenspace $E_{\lambda_k}$, instead of considering subspaces $\operatorname{Ran}(A_k-\lambda_k I)^j$, which we would need to consider when working with the block $A_k$ separately, we consider the subspaces $(A-\lambda_k I)^j E_{\lambda_k}$.

# Index